PHD

Online drill-system parameter estimation and hazardous event detection

Burrows, Daniel

Award date:
2023

Awarding institution:
University of Bath

Online drill-system parameter estimation and hazardous event detection

Daniel Burrows

A thesis submitted for the degree of Doctor of Philosophy

University of Bath
Department of Mathematical Sciences
June 2023

Copyright notice

Access to this thesis is restricted until .....

Signed on behalf of the Doctoral College .....

Declaration of any previous submission of the work

The material presented here for examination for the award of a higher degree by research has not been incorporated into a submission for another degree.

Candidate's signature .....

Declaration of authorship

I am the author of this thesis, and the work described therein was carried out by myself personally, with the exception of Chapter 3, where the work was carried out jointly between myself and Dr. Kari Heine. The original idea and proofs were formulated by Dr. Heine and subsequently developed and tested jointly. Chapter 2 is an original collation of established theory that is necessary to include in order to describe the subsequent material in a rigorous way.

Candidate's signature .....

Acknowledgements

I have something of a track record of writing overly sentimental acknowledgements in each of the theses submitted within my time in academia, but on this occasion the sentiment seems to be supplemented by a "liner notes in a debut album" style of tone. In the spirit of these ideals and in accordance with tradition, there are a few people I wish to thank.

Firstly, I'd like to thank my lead supervisor Kari Heine for being a profoundly influential teacher and collaborator throughout this Ph.D. project. From detailed, technical nuances to a general research methodology, I'm grateful to have learnt something from you in every one of these facets.

Additionally, I'd like to thank EPSRC and Schlumberger for providing the support to make this project possible. From SLB in particular I'd like to thank Can Evren Yarman, without whose backing I would not have been given this opportunity, and Inês Cecílio, whose advice was influential on our approach. From the University of Bath, I'm similarly grateful to Mark Opmeer for the direction provided when the project was still in its infancy, to Ivan Graham and Euan Spence for the role they played in the interview process, and to Susie Douglas for being a constant source of advice and support that has kept me connected to the department, even when physical attendance has not been possible.

On a more personal note, thank you Kyle Strickland; you've been with me on every step of this journey and, to quote the lyrics of the great John Frusciante, 'you wouldn't have made it without her'. Remaining on the topic of music, thank you Concrete Prairie; you've not only kept me sane these past 5 years but you've become nothing short of a musical family to me. Thank you Fergus Bright, Ana Soria, James Bradford, Vivien Maertens, Martyn Lewis Moss, Jasmine Tokyo Daisy Scott and Sean Gavin Smith; without the respective South West pints, gigs and laughs we've shared, life would have been dull.

Lastly, I'd like to thank my family for always trusting me in my convictions and believing in my ability to achieve them. None of this would be possible without you.

In loving memory of Calum Brereton, Auntie Joy and Grandad Michael Burrows; each of whom left this world during the time which the project was conducted, but whose enduring memory has been with me throughout.

Abstract

The use of on-line, noisy, partial observations as a means of inferring obscured system parameters is one that arises in a number of applications. Hydrocarbon drilling operations are one such example, in which the presence of hazardous events and general subsurface activity must often be inferred using measurements of the drill equipment taken at the surface. Adopting a Bayesian viewpoint, the task is then to estimate or — where possible — exactly compute the probability distribution of the parameters as the corresponding measurements are assimilated.

If the latent parameters are assumed to form a Markov process then the described statistical model has sufficient structure for the implementation of a number of inference methods. In applications where the underlying mappings are non-linear, the distributions are generally intractable and sequential Monte Carlo (SMC) methods offer a means of constructing estimates. For their relative robustness and versatility, SMC methods are often advantageous over other approaches. However, their construction requires repeated generation of weights based on the map between the parameter space and the observation space. If this weight generation cost is expensive then the repeated computation of these terms can limit the usefulness of the method in applications such as the drilling operation.

In this thesis we present a novel SMC method called the multilevel bootstrap particle filter (MLBPF) that arises from applying the approach of multilevel Monte Carlo (MLMC) to the weight computation step. The aim of this adaptation is to translate the efficiency savings of MLMC to the SMC setting, thus enabling more accurate estimates to be constructed with the same computational budget. Despite the relative simplicity of the idea, the implications it has on the method are profound, since in general the weights are no longer guaranteed to be non-negative. In addition to specifying an operational MLBPF algorithm, we prove a strong law of large numbers and central limit theorem result, before numerically testing the MLBPF on a number of models comparable to the drill-system. In addition to attaining empirical accuracy gains in these experiments, we derive a general approach to particle filtering on drill-system type models that provides a means of estimating the hidden parameter distributions in the absence of exact knowledge of the underlying system state.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Hydrocarbon drilling operations are complex procedures that apply specialised machinery and engineering techniques in often volatile conditions. This scale and complexity comes with a risk of hazardous events (HE) that compromise the safety of the field engineers and the surrounding ecosystem. Furthermore, to continue to meet demand, oilfield services are drilling in riskier conditions such as offshore deepwater, ultra-high pressure and high temperature wells, which intensifies the already substantial risks to the environment and human life [4, 60, 66].

While large-scale environmental catastrophes such as the 1969 Santa Barbara and 2010 Deepwater Horizon oil spills are rare, operational time loss as a result of HE is a prevalent and enduring issue. As of 2018, this non-productive time accounted for between 20-25% of the total drilling time [26]. With the various costs of operating a drilling rig being extremely high, early detection of HE is also a means by which the associated operational costs can be significantly reduced.

From a safety, environmental and economic perspective it is therefore desirable to detect HE at the earliest possible opportunity to minimise the extent of the potential resulting damage. More generally, accurately estimating parameters that describe the expected behaviour of the system enables greater operational efficiency which in turn minimises any negative impacts of the drilling. Improving the capabilities to which this can be achieved is a key objective not only to facilitating a safer transition to reaching net zero carbon emissions, but also because much of this intelligence is transferable to emerging sustainable approaches such as geothermal well applications and

carbon removal.

The repertoire of methods for parameter estimation and HE detection is constantly expanding and draws from numerous fields in science and technology. Managed pressure drilling is a technique that adaptively calibrates the drilling equipment to the environment to enable safe drilling in settings where the formation has a narrow pressure window [2]. The recent telemetry hardware innovation *wired drill pipe technology* delivers data up to a rate $10,000$ times faster than the traditional mud pulsing method [33] and enables measurements of the drilling equipment to be taken locally subsurface instead of being globally averaged at the surface. These measurements facilitate a greater insight into subsurface activity that enables better fault detection in managed pressure drilling [26] and real-time vertical seismic profiling by using the drill-bit as an imaging source [58]. Real-time, large scale performance is becoming increasingly more accessible through efficient cloud-based technology and greater availability of high performance computers, thus enabling solvers and simulations to be run alongside a drilling operation.

Each of these advancements open the possibility for the application of mathematical methods that have previously been infeasible. However, the physical limitations of the drilling procedure are such that improving upon the current level of performance is non-trivial for several reasons. The drill system is predominantly obscured from direct observation, meaning knowledge of the subsurface activity has to be inferred indirectly using measurements that themselves are subject to observation noise. This raises the possibility of a certain set of measurements suggesting the occurrence of one or several HE, or alternatively being a harmless artefact of the drilling. In the case of the former, prior knowledge and the historical behaviour of the measurements must be used to classify the HE for it to be remediated with minimum disruption to the operation [79]. On the other hand, if changes in the measurement data are wrongly attributed as symptoms of an HE then the risk of raising a false alarm is increased, which can also threaten the safety and efficiency of well control [73].

Another complicating factor is the occurrence of unforeseen events during the operation. Prior to drilling, knowledge of the surrounding geology is acquired using seismic imaging at the proposed location of the well. Part of the process of generating these images is to use synthetic data simulated

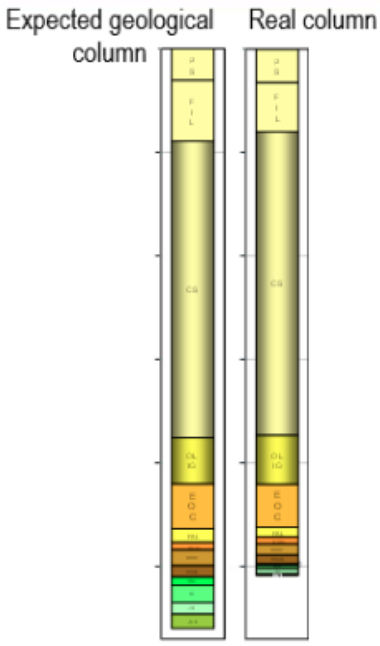

Figure 1.1: A schematic of the discrepancy that can arise between a geological column estimate and its true composition.

from an erroneous model to adjust the parameters of a global model [74]. This error can lead to discrepancies between the estimated and true geological column that the drilling device excavates to reach the target hydrocarbon reservoir; see Figure 1.1. In turn, this can lead to the drilling equipment being incorrectly calibrated to its environment, resulting in either sub-optimal performance or burnout from being overworked. Furthermore, the surroundings are themselves unpredictable and can damage the equipment, which can also malfunction of its own accord.

1.1 Description of the drill system

While the components of drilling rigs may vary slightly between onshore and offshore applications, their basic apparatus is broadly the same. A drilling device comprising of two main components; a hollow *drill string* and a *bottomhole assembly* is supported by a *derrick* or *mast* at the designated point of drilling. A multipurpose fluid known as *mud* is pumped at a known and fixed volumetric flow rate through the center of the drill string, which can

Figure 1.2: Left: The drilling device is rotated and applied with a downward force, while the mud flows through its center and up to the surface via the resulting annulus. Right: A close up of the drill bit, which is used for both cutting the rock and flushing the mud via the bit nozzles.

be constructed to be several kilometres long to access the target hydrocarbon reservoir. At surface level a *rotary table* provides power for rotation of the drill string measured in rotations-per-minute (RPM), and a *weight-on-bit* (WOB) is applied to provide a downward force to the device. The purpose of these manoeuvres is to provide a mechanical power to the drilling process that is measured in the rate of penetration (ROP). It is also at the surface that a *standpipe* is used to pressurise the mud as it enters the drill string. Together with the RPM and WOB, this pressure level is an observable input quantity that is in part controllable by the operator but is also responsive to events in the well. As such, they form observable inlet measurements that are used to infer knowledge of the drill system.

At the other end of the drilling device is the bottomhole assembly. This is a collection of machinery for which the main components of interest from an HE perspective are the *mud motor* and the *bit*. The purpose of the mud

motor is to use the passing fluid as a hydraulic energy source as it flows through the motor to provide extra drive to the bit. In applications where directional drilling is required the motor is also steerable; allowing drilling to be performed at different angles and rotation of the bit independently of the drill string. The role of the bit at the very end of the drilling device is to cut and crush the rock. Throughout the course of an operation the bit is likely to require changing to remain effective. To do so a so-called *tripping* manoeuvre is performed on the drill string, in which it is extracted and subsequently reinserted into the bore hole after the change or repair has been made [2].

Once passing through the mud motor the mud flows out of the drilling device into the surrounding well through small *bit nozzles*. Due to the high pressure at these depths the fluid is ejected out of the bit at a high velocity that aides flushing the cavity of cuttings for their transport to the surface via the mud. Ensuring the mud is kept in an appropriate pressure window in relation to the formation as it moves up along the annulus is vital to preserving the integrity of the well. Doing so prevents borehole collapse from an insufficiently low mud pressure and rock fracture from one that is excessively high [12, 46, 82]. If wired drill pipe technology is used then the sensors are placed on the outer wall of the drill string where they can be used to perform tasks such as monitoring the location profiles of the passing cuttings [24]. Once at the surface the cuttings are extracted from the mud and monitored for the presence of hydrocarbon content and their geological properties by an engineer, while the mud is retreated and recycled to be sent through the system again. The mud exits the well at its known atmospheric pressure and an observable volumetric flow rate that can be used as an outlet measurement for downhole inference.

1.2 Common hazardous events

There are a number of HE that can occur during a drilling operation and their formal categorisation and diagnosis is a field in its own right. For the purposes of further describing the model application, as well as some of the underlying statistical challenges, we state some of the most prominent HE here.

- *Kick/blowout*: if the well pressure is lower than the formation pore

pressure at points in the well then an influx of the higher pressure formation fluid occurs [66, 80]. A blowout is the event in which a kick becomes uncontrolled. Blowouts are the cause of some of the worst oil and gas drilling disasters and many preventative measures both before and after kicks have been devised to prevent them escalating to blowout [70]. Early detection of kick is therefore crucial and has motivated many approaches; cf. [35]. Some potential indicators of kick are discrepancies between the inlet and outlet flow rates due to the influx of the formation fluid, a sudden increase in the ROP, and a change in the inlet mud pressure [80], although these are just a few of the simpler methods in circulation; cf. [25] for more detail.

- *Circulation loss*: conversely to kick, circulation loss is well leakage due to the pressure of the mud being excessively high. There are different levels of severity of loss — from seepage losses to catastrophic losses — and this instructs the extent of the measures taken to rectify the situation. If the loss is excessive it may lead to insufficient pressure in the well, at which point an influx could again occur [79]. Circulation loss is therefore characterised in its moderate state by a decreased outlet volumetric flow rate and by the symptoms of kick as the losses become catastrophic.

- *Washout*: the appearance of a hole in the wall of the drill string due to erosion that allows mud to flow from inside the drill string to the well without passing through the bottomhole assembly. This bypassing compromises the effectiveness of the mud and hence the efficiency of the operation, as the mud is no longer being used to its full potential in the bottomhole assembly. In the worst case scenario *pipe twistoff* can occur, typically inducing 3 to 12 days of non-productive time [51]. Washout typically manifests itself as a decrease in the observable inlet mud pressure level.

- *Pack off*: the partial or complete blockage of the annulus due to an accumulation of cuttings. The consequences of pack off are numerous and are listed in [24], one of the most serious of which is *stuck drillpipe*. Stuck drillpipe remains one of the most expensive complications in drilling, costing the industry more than $250 million a year as of 2020 [22]. Since pack off is a restriction of the flow, its occurrence can

be identified by a lower outlet volumetric flow rate than expected, anomalies in the drill string rotary speed and WOB, and an increase in the inlet pressure.

- *Bit nozzles plugging*: obstruction of the bit nozzles by small cuttings, reducing the effectiveness of the cleansing role of the mud in the cavity. This compromises the efficiency of the drilling as the transport of the cuttings out of the way of the drill is no longer optimal. Plugging is observed as an increase in the inlet mud pressure but does not affect the pressure in the well. This makes the event less serious in practice but important to detect and distinguish from the more serious pack off [80].

We observe that each of the listed HE manifest themselves as anomalies in the drilling equipment which may not be exclusive to one particular event. These anomalies are then used as partial observations of events in the obscured part of the drill system.

In addition to estimating parameters that describe HE, it is of interest to determine the optimal configuration of the equipment. For example, given some environment, tuning the location of drill pipe sensors or the combination of RPM and WOB may improve the performance of the sensor signal quality or ROP respectively. Since the system measurements are the means by which the condition of the system is inferred, they can also be used to estimate regions of parameter values that elicit better performance from the equipment. This is somewhat of a reversal of the HE detection problem, in which some features of the system are now being systematically calibrated to induce desired changes in others, but in essence is the same problem of hidden parameter estimation. To this end the task at hand may more generally be viewed as one in which we have a collection of events of interest that are only partially observable via changes in the drilling equipment, and it is from these changes that we wish to infer information about the underlying dependencies.

1.3 Modelling assumptions

With this generalisation in mind, we consider the drilling application in the more general paradigm of Bayesian parameter estimation. During the oper-

ation, partial observations of the parameters are systematically generated in an online fashion, meaning the estimation problem is sequential by nature. The task is then to use the noisy observations — for example, the drill equipment measurements — to infer the probability distribution of the parameter at the corresponding moment in time.

Let $n \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$. For each $\mathbb{Y}$-valued observation Y_n, we denote the corresponding $\mathbb{X}$-valued parameter by X_n. To use the observations for inference, a model is required that relates each Y_n to X_n. We view this model as the composition of two functions: a deterministic map $h : \mathbb{X} \to \mathbb{Y}$ that describes the dependency of Y_n on X_n in the absence of noise, and the application of a random component to $h(x_n)$ that maps $h(x_n)$ to $y_n \in \mathbb{Y}$.

A class of functions h that describe a wide range of dependencies are those for which $h(x_n)$ is a component of a solution to a differential equation with respect to the realised state $X_n = x_n$. For example, in the context of drilling, one of the primary media for connecting the observations to the obscured parameter is the mud. To suitably describe the dynamics of the mud for the purposes of inference thus requires a fluids model based on the principles of conservation of mass, momentum and energy [72, Ch. 1]. We note that, due to the physical limitations of how quickly waves can be transferred in a fluid, there is an unavoidable lag between the observation and the parameter state it corresponds to. As such, inferential estimates will always pertain to state distributions that occurred in the past, and the state may have changed substantially within the time taken to receive the most recent observation.

If the fluid viscosity is neglected and the solutions are sufficiently smooth [14, Ch. 1] then, in the one-dimensional case, they can be expressed as the solution to a partial differential equation (PDE) of the form

$$\partial_t w + \partial_z f(w) = S(z, t, w) \tag{1.3.1}$$

where:

- $z \in \mathbb{R}$ and $t \in [0, \infty)$ are the space and time variables respectively;
- $w = w(z, t) = (\rho(z, t), J(z, t), E(z, t)) \in \mathbb{R}^3$ is the solution of (1.3.1) comprising of the density ρ, momentum J and energy E;
- $\partial_{(\,\cdot\,)}$ is the partial derivative operator with respect to the variable $(\,\cdot\,)$;

- $f : \mathbb{R}^3 \to \mathbb{R}^3$ is a flux function that describes the spatial dynamics;
- $S : \mathbb{R} \times [0, \infty) \times \mathbb{R}^3 \to \mathbb{R}^3$ is a source term function describing external forces acting on the fluid.

When the latent parameter X_n is also factored into proceedings, the solution and source term in (1.3.1) inherit an additional dependence on X_n. For example, suppose X_n describes the area of a washout hole and p is the mud pressure level derived from ρ via an Equation of State $p = F(\rho)$ satisfying $F'(\rho) > 0$ [14, Ch. 1]. Then the source term is modelled as a point mass at the location of the hole and, in a manner consistent with the physical behaviour of the operation, the inlet pressure component derived from the mathematical solution $w(z, t; X_n)$ decreases as X_n increases. We note by the monotonicity assumption of the equation of state that this is equivalent to a decrease in the density solution evaluated at the inlet.

If the fluid is compressible, the equations that form (1.3.1) are known as the Euler equations. For applications such as the drilling operation in which the mud is subjected to extremely high pressures, this compressibility assumption is a realistic one. The Euler equations neglect second order terms modelling fluid viscosity and heat conduction that are present in the Navier-Stokes equations. Consequently, (1.3.1) is first-order and therefore hyperbolic, meaning information is propagated at a finite speed and that discontinuous solutions can arise even when the initial data is continuous [47, Ch. 1]. In contrast, the PDE specified by the Navier-Stokes equations is parabolic and hence possesses solutions that are continuous solutions for all times [47, Ch. 14]. We note, by modifying f and S, that (1.3.1) can also be used to model other applications such as the shallow water equations or, with the inclusion of a second order term, the convection-diffusion equation. While keeping the drilling model in mind, we therefore consider the general form (1.3.1) rather than focusing only on the specific w, f and S that specify the Euler equations.

In general, there are no closed form solutions to PDEs of the type (1.3.1); instead, approximate numerical solutions are required. From a numerical point of view, both the Euler and Navier-Stokes equations are challenging problems in their own right that are regularly used as benchmarks for contemporary numerical methods; for the Euler equations see [3, 42, 72] and Navier-Stokes equations see [34, 69, 81], for example. Typically in the hy-

perbolic case, finite element and finite volume methods are favoured over finite difference methods, which typically are not well suited to handling the discontinuities that can arise [47, Ch 1].

Irrespective of the method, numerical solutions to (1.3.1) are generally expensive to generate and impose a computational bottleneck to applications such as drilling operations, which in general require accurate solutions on a frequent basis. Moreover, (1.3.1) models the simplest case in which both the spatial domain and the solution components are one-dimensional. In more advanced models these quantities can be vector-valued, in which case the computational complexity of generating solutions increases by potentially numerous orders of magnitude.

Heuristically, if approximate solutions to (1.3.1) can be obtained with respect to $N > 0$ "proposed" parameter states $(\xi_n^i)_{i=1}^N$, then the credibility of each ξ_n^i against the true x_n can be ascertained by comparing the synthetic observation y_n^i based on $w(z, t; \xi_n^i)$ to the realised observation y_n of x_n. Under some non-restrictive assumptions about the latent parameter process and the observations that we make precise in Chapter 2, this approach can be repeatedly applied in a rigorous way to construct estimates of the probability distribution of X_n at each iteration. This is the general approach of Sequential Monte Carlo (SMC) methods [19] that we use as a foundation for the parameter estimation method presented in this thesis. In particular, we identify the overall cost of generating the y_n^i as a target for cost savings. In turn, this enables the generation of more proposed states ξ_n^i that provide more accurate estimates of the parameter distributions in the same computational time. In the context of drilling operations, such gains translate to a better insight of the subsurface developments and hence more ability to take reactive steps to improve operational safety and efficiency.

1.4 Literature review

The methods by which HE detection is performed may broadly be classified as data-based or model-based [26]. The distinction between the two approaches is based primarily on the fact that, although a model-based approach may use data to estimate parameters, the relationship between the two quantities is specified explicitly using a model of some or all of the components of the drill system. In contrast, data-based methods do not

explicitly factor in any underlying physical structure and rely solely on the data to form estimates.

The argument for data-based methods is that, by taking a model-free approach, the estimates avoid any potential bias of performing inference via a model that does not reflect the true physical behaviour of the system. They have recently grown in popularity due to the greater availability of real-time drilling data, but both approaches are still widely used and the preference of one over the other can depend on the features and requirements of the task. Examples of data-based methods include the use of a slow feature analysis to detect faults in a manner that limits false alarms [26]; the application of a random forest to estimate a rock property coefficient [65]; applying an artificial neural network hybrid method to fluid measurement data to estimate rheological properties of the mud [1]; and the study of an artificial neural network to estimate the in-situ stress of a reservoir section using wireline log data [21].

The model-based approach uses the data as a means of inferring HE but does so with the addition of a model that maps the measurements to the HE. Since the mud is one of the primary means by which this mapping takes place, these models are often fluid based. Depending on the application or the equipment, the models can greatly vary and incorporate bespoke auxiliary modifications to specifically study features of interest. Some common examples are friction, heave and flow models that augment a 1-D hydraulic transmission line [44], a BHP model based on multiphase flow [83], and modelling the annular flow velocity by a wave frequency based on the Doppler principle [25]. In other scenarios it may be the case that a simplified modelling approach sufficiently captures a feature while reducing the complexity of the problem. In [38] a simplified hydraulics model was used to estimate the downhole pressure profile, while in [24] a fluids model was applied to capture the dynamics of passing cuttings and quantify uncertainty in the transport process via the ensemble Kalman filter.

In this thesis we take the model-based approach and pursue a statistical solution methodology for drill-system type problems. In particular we use the framework of SMC methods based on random samples and on-line measurements to estimate the sequence of probability distributions of the corresponding evolving parameter states. We do so because the mappings within models such as those based on drill-systems are generally non-linear

and SMC methods are more robust to these features than alternative linearisation methods. Moreover, SMC methods possess convergence properties that alternative methods do not, and in practice they continue to become more powerful with the increased accessibility to high-performance computing. The novelty of our approach is that we combine the interacting particle mechanism of SMC with multilevel Monte Carlo (MLMC) [29] to produce a multilevel SMC methodology that seeks to further broaden the current repertoire of methods.

Specifically, we remain in the problem of estimating a flow of probability distributions that are related to one another via a Markovian mapping, but at each iteration apply the multilevel decomposition within the costly weight assignment step. One profound consequence of this is that, due to the telescoping sum, the weights are now based on the difference of two distinct likelihood terms and are no longer guaranteed to be positive as they are in traditional SMC methods. As such, the notion of resampling a particle with probability proportional to its importance weight no longer has a well-defined meaning and an alternative resampling scheme is required for the method to be practically feasible. We construct one such scheme, in which resampling is instead performed according to the total variation filter measure and additionally the sign of the weight is retained alongside each corresponding resampled particle. The resulting particle approximations are therefore also signed measures. However, we prove that almost sure convergence to the exact respective probability measures and a central limit theorem hold.

The integration of multilevel methods into Bayesian estimation problems has been explored in a variety of different contexts. In [7] the approach was implemented within the framework of SMC samplers, but these problems are quite different to the SMC estimation problem that we consider. SMC samplers proceed by sampling from a sequence of probability distributions, with the aim of composing a particle ensemble that resembles one drawn from a target distribution that may otherwise be difficult to sample from; see [56]. In [7] this sequence was constructed according to a multilevel scheme that iteratively refines a sample throughout the levels towards one that achieves equivalent accuracy to its closest competitor in fewer floating point operations. However, since there is no transience in the target distribution, the algorithm avoids the issues arising from mutating particles to estimate probability distributions of a latent Markov process. In particular, the nature

of the problem means the method does not have to deal with the issues of weight negativity that we face in multilevel SMC.

Within the SMC paradigm a multilevel scheme has previously been constructed in [36] to leverage a multilevel particle filter, but both the application setting and the method are different to that which we consider. Firstly, the setting is based specifically on estimating the probability distribution of a partially observable diffusion process at a collection of discrete time measurements. Consequently, the multilevel aspect is introduced using the established framework of Euler-Maruyama type methods as used in [29], based on a hierarchy of nested Brownian increment solver steps. Secondly, since the estimates are targeted at the diffusion discrepancies; i.e. the distribution of the difference in the diffusion process at times $n+1$ and n, the multilevel scheme is applied with respect to different step sizes over this time interval. The sequentiality in the method then arises as a result of considering the discrepancy distribution at the next iteration, which essentially extends the "payoff time" estimation problem in [29] to a sequence of termination times. This specific interpretation enables a coupled resampling method to be applied, which again circumvents any weight negativity. In contrast, we do not assume any such structure in our models which, while enabling more general applications, leads to a method which possesses different weighting and resampling steps.

A similar diffusion-based filtering problem is considered in [23], this time in the context of random weight particle filtering. In such methods negative weights can occur with probability greater than zero as a result of instead drawing importance weight samples from an unbiased estimator of the Radon-Nikodym derivative as opposed to evaluating the derivative itself. In the presence of a negative weight, each of the weight samples are incremented by one from a newly generated sample set until no negativity remains. Such an approach is justified by Wald's identity for martingales, which ensures that the expected value of the resulting importance sampling estimator is proportional to the original random weight estimator. However, while in this method the presence of a negative weight is undesirable, in our algorithm the negativity is an essential part of the approach that captures the effect of decomposing the filter distribution into a telescoping sum of signed measures. To this end we do not wish to transform them in the manner used in [23]; rather, they are required to facilitate a mapping of the

target probability measure into the space of signed measures in a rigorous way. Indeed, the only time the weight negativity becomes an issue is in the resampling step. One advantage of our simple total variation measure resampling approach is that it does not require the generation of any more weights, which in the applications we consider are expensive. In contrast, while the random weight approach provides a well-understood alternative means of accommodating the weight negativity, the repeated generation of a new collection of weights is infeasible within the assumptions of the application. Furthermore, it is not clear in our method that iteratively increasing a negative weight towards one that is positive preserves the asymptotics that we have established, nor that it is even desirable from an estimate accuracy perspective.

Applying the multilevel approach to expensive and complex applications such as PDE-based observation models has been explored before in the context of uncertainty quantification of groundwater flow [15], but this was in a setting that was both non-sequential and with respect to one fixed expectation of interest. Extending this approach to sequences of probability measures, in particular those based on solutions to PDEs, seems relatively unexplored. On the other hand, Bayesian cost reduction methods — specifically approximate Bayesian computation (ABC) algorithms [55, 57, 59, 71] — provide a means of inference when evaluations of the likelihood function are infeasible due to either being too costly or simply intractable. Such methods proceed by simulating a collection of proposed observations that serve as a substitute for the absent likelihood function based on their proximity to the actual observation. In [71] an SMC sampler approach (ABC-SMC) was used to further improve the efficiency of the Markov chain Monte Carlo based method in [52]. Both approaches are based on the property that the acceptance rate of synthetic observations can be low when the prior distribution is a poor approximation to the posterior, and that this prior can be iteratively refined to improve the acceptance rate. In [55] the complexity of the ABC-SMC method was reduced from one that is quadratic in its samples to one that is linear by using an adaptive sampling approach, while [59] utilises a hierarchical acceptance structure to improve the efficiency of the ABC approach and achieve performance gains. This work is particularly relevant to our own due to the appearance once again of negative weights, ruling out a conventional integration of the SMC sampling approach to the method. We

note however that, in general, ABC methods are likelihood-free, whereas in our problem we seek to make efficiency gains in the sustained presence of the likelihood function.

This thesis is organised as follows. In Chapter 2 we review and discuss topics from SMC and multilevel Monte Carlo literature that are required to provide a thorough theoretical foundation to the subsequent material. In particular, we discuss the exact filtering equations in Section 2.1 and show that these equations inherit the predict and update recursions from the more general Feynman-Kac formulae [54]. This type of analysis will reappear later in Chapter 3 when we formally derive our own SMC method. In Section 2.2 we then discuss how particle filters are formally obtained by replacing the exact measures in the Feynman-Kac formulae with their particle-based estimates, before considering some of the more practical issues of particle filters that are also relevant to our own method. Section 2.3 discusses the work of MLMC [15, 29], which we present in our more general notation to facilitate a comparison between MLMC and our own multilevel-based method. In Chapter 3 we define the multilevel bootstrap particle filter (MLBPF) that is the core method of this thesis. In particular we provide a formal derivation in Section 3.1 before proving a strong law of large numbers result and a central limit theorem in Sections 3.2 and 3.3 respectively. In Chapter 4 we conduct numerical experiments with the MLBPF on a variety of models. In Section 4.1 we first describe the mathematical properties of the models, their physical interpretations, and how the PDE models in particular can be interpreted as surrogate models for the drill-system application. Sections 4.2 and 4.3 then present the results of the numerical experiments that demonstrate comparative empirical accuracy gains from the MLBPF over the equivalent benchmark SMC method. The codes for the experiments conducted in these sections can be found at `https://github.com/dwb26/mlbpf_steady_state_swe_final` and `https://github.com/dwb26/mlbpf_convection_diffusion` respectively. In Section 4.3 we also discuss some further PDE-based applications, the modelling challenges they pose, and some potential ways in which they can be made feasible. Finally, in Chapter 5 we review the work and propose possible future research directions that could be used to further build on the contributions of this thesis.

Chapter 2

Sequential and multilevel Monte Carlo

2.1 Non-linear filtering

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and, for $n \in \mathbb{N}_0$, consider the time homogeneous Markov process (X_n) with each X_n taking values in the measurable space $(\mathbb{X}, \mathcal{X})$. We adopt the convention in SMC literature of referring to (X_n) as a *signal process*. Additionally, we denote the set of bounded and measurable functions on $\mathbb{X}$ by $\mathscr{B}(\mathbb{X})$ and the set of probability measures on $\mathcal{X}$ by $\mathscr{P}(\mathcal{X})$.

The signal process is described by specifying a prior distribution π_0 for X_0 and a mapping $K : \mathbb{X} \times \mathcal{X} \to [0, 1]$ describing the transition between states, where

$$K(x_n, dx_{n+1}) := \mathbb{P}(X_{n+1} \in dx_{n+1} \mid X_n = x_n) \tag{2.1.1}$$

and dx_{n+1} denotes an infinitesimal neighbourhood of x_{n+1}. For every $x \in \mathbb{X}$, $A \in \mathcal{X}$, we note that

$$\begin{aligned} x \in \mathbb{X} &\mapsto K(x, A) \in \mathscr{B}(\mathbb{X}), \\ A \in \mathcal{X} &\mapsto K(x, A) \in \mathscr{P}(\mathcal{X}), \end{aligned}$$

making K a Markov kernel (cf. [63, Ch. 8], for example). Consequently, K induces the following integral operators, yielding a function $K(\varphi) \in \mathscr{B}(\mathbb{X})$

and probability measure $\nu K \in \mathscr{P}(\mathcal{X})$ respectively:

$$K(\varphi)(\cdot) := \int_{\mathbb{X}} \varphi(x_{n+1}) K(\,\cdot\,, dx_{n+1}), \quad \varphi \in \mathscr{B}(\mathbb{X}), \tag{2.1.2}$$

$$(\nu K)(\cdot) := \int_{\mathbb{X}} \nu(dx_n) K(x_n, \,\cdot\,), \quad \nu \in \mathscr{P}(\mathcal{X}). \tag{2.1.3}$$

In particular, for $n > 0$ the distribution $\nu_n = \mathbb{P}(X_n \in \,\cdot\,)$ may be exactly described in terms of π_0 and K by first noting the recursion

$$\nu_n = \int_{\mathbb{X}} \nu_{n-1}(dx_{n-1}) K(x_{n-1}, \,\cdot\,) = \nu_{n-1} K \tag{2.1.4}$$

from which we deduce $\nu_n = \pi_0 K^n$, where $K^n := K \circ K \ldots \circ K$ is the n-fold integral operator that results from applying K a total of n times. We assume K satisfies the Feller property as stated in [19, Ch. 2], so that $K(\varphi) \in \mathscr{B}(\mathbb{X})$ is continuous whenever $\varphi \in \mathscr{B}(\mathbb{X})$ is continuous.

We denote the partial observations of (X_n) by the sequence of random variables (Y_n), each taking value in the measurable space $(\mathbb{Y}, \mathcal{Y})$, and frequently refer to (Y_n) as the *observation process*. Given the signal (X_n), the (Y_n) are assumed to be conditionally independent, with marginal conditional probabilities distributed according to

$$Y_n \mid X_n = x_n \sim G(x_n, \,\cdot\,)$$

for a probability kernel $G : \mathbb{X} \times \mathcal{Y} \to [0, 1]$. For each x_n we assume G admits a probability density g with respect to some σ-finite measure, and for a fixed observation $Y_n = y_n$ we write $g_n(\,\cdot\,) := g(\,\cdot\,, y_n)$.

The non-linear filtering problem is the task of computing the *filter* and *prediction* probability measures given respectively by

$$\widehat{\pi}_n(\,\cdot\,) := \mathbb{P}(X_n \in \,\cdot\, \mid Y_0 = y_0, \ldots, Y_n = y_n), \quad n \geq 0, \tag{2.1.5}$$

$$\pi_{n+1}(\,\cdot\,) := \mathbb{P}(X_{n+1} \in \,\cdot\, \mid Y_0 = y_0, \ldots, Y_n = y_n), \quad n \geq 0. \tag{2.1.6}$$

In practice, these measures describe exactly the distribution of the latent state random variable in question, given the knowledge provided by the collection of partial observations. For example, if in the drilling application we define X_n to be the size of a washout hole at a certain time, then the partial observations $(Y_m)_{m=0}^n$ could be defined to be the collection of inlet

pressure measurements up to the present time, and the filter measure (2.1.5) is the probability distribution of the hidden washout hole size at time n given the knowledge leveraged by these observations (and analogously for the predictive measure (2.1.6)).

We will also frequently use the corresponding weak forms of (2.1.5)–(2.1.6) considered over all $\varphi \in \mathscr{B}(\mathbb{X})$ and given by

$$\widehat{\pi}_n(\varphi) = \int_{\mathbb{X}} \varphi(x_n)\widehat{\pi}_n(dx_n), \quad \pi_n(\varphi) = \int_{\mathbb{X}} \varphi(x_n)\pi_n(dx_n). \tag{2.1.7}$$

Integrals of the form (2.1.7) are of common theoretical and practical interest because they describe not only the underlying probability measure but also a large class of expectations. They are in fact examples of the more general Feynman-Kac formulae [54, Ch. 2]

$$\widehat{\mu}_n(\varphi) := \mathbb{E}\left[\varphi(X_n)\prod_{p=0}^{n} V_p(X_p)\right], \quad \mu_n(\varphi) = \mathbb{E}\left[\varphi(X_n)\prod_{p=0}^{n-1} V_p(X_p)\right], \tag{2.1.8}$$

$$\widehat{\eta}_n(\varphi) = \widehat{\mu}_n(\varphi)/\widehat{\mu}_n(1), \quad \eta_n(\varphi) = \mu_n(\varphi)/\mu_n(1), \tag{2.1.9}$$

with potential functions $V_p : \mathbb{X} \to [0, \infty)$ and expectations taken with respect to the joint path measure

$$\mathbb{P}(X_{0:n} \in dx_{0:n}) = \pi_0(dx_0)K(x_0, dx_1)\dots K(x_{n-1}, dx_n).$$

Consequently, $\widehat{\pi}_n$ and π_n inherit the structure that relates the Feynman-Kac measures (2.1.8)–(2.1.9) to one another through a recursive procedure of prediction and update. To see this connection, we first note that

$$\begin{aligned}
\mu_n(\varphi) &:= \int_{\mathbb{X}^{n+1}} \varphi(x_n)\left[\prod_{p=0}^{n-1} V_p(x_p)\right]\pi_0(dx_0)K(x_0, dx_1)\dots K(x_{n-1}, dx_n) \\
&= \int_{\mathbb{X}^n}\int_{\mathbb{X}} \varphi(x_n)K(x_{n-1}, dx_n)\left[\prod_{p=0}^{n-1} V_p(x_p)\right]\pi_0(dx_0)\dots K(x_{n-2}, dx_{n-1}) \\
&= \widehat{\mu}_{n-1}(K(\varphi)).
\end{aligned} \tag{2.1.10}$$

By the property $K(x, \cdot) \in \mathscr{P}(\mathcal{X})$, it follows immediately from (2.1.10) that

$$\mu_n(1) = \int_{\mathbb{X}^n} \left[\prod_{p=0}^{n-1} V_p(x_p) \right] \pi_0(dx_0) \dots K(x_{n-2}, dx_{n-1}) = \widehat{\mu}_{n-1}(1).$$

Therefore, by (2.1.9),

$$\eta_n(\varphi) = \widehat{\mu}_{n-1}(K(\varphi))/\widehat{\mu}_{n-1}(1) = \widehat{\eta}_{n-1}(K(\varphi)). \tag{2.1.11}$$

This is the so-called *mutation step*, which demonstrates how the normalised prediction Feynman-Kac measure η_n is obtained from the normalised updated Feynman-Kac measure $\widehat{\eta}_{n-1}$ by integrating the Markov kernel conditioned on the state x_{n-1} with respect to $\widehat{\eta}_{n-1}$.

The other part of the recursion is the *update step*, which describes how the new information provided in the form of V_n is used to update η_n to $\widehat{\eta}_n$. By the definitions (2.1.8), we first note the simple relation $\widehat{\mu}_n(\varphi) = \mu_n(V_n\varphi)$. Normalising $\widehat{\mu}_n$, it then follows that

$$\widehat{\eta}_n(\varphi) = \mu_n(V_n\varphi)/\mu_n(V_n) = \eta_n(V_n\varphi)/\eta_n(V_n) =: \Psi_n(\eta_n)(\varphi), \tag{2.1.12}$$

where $\Psi : \eta \in \mathscr{P}(\mathcal{X}) \mapsto \Psi(\eta) \in \mathscr{P}(\mathcal{X})$ is the Boltzmann-Gibbs operator

$$\Psi_n(\eta)(dx) = \frac{1}{\eta(V_n)} V_n(x)\eta(dx). \tag{2.1.13}$$

Combining (2.1.11) and (2.1.12) gives the following recursive formulae for the Feynman-Kac probability measures

$$\eta_n = \Psi_{n-1}(\eta_{n-1})K_{n-1}, \tag{2.1.14}$$

$$\widehat{\eta}_n = \Psi_n(\widehat{\eta}_{n-1}K_{n-1}). \tag{2.1.15}$$

Equation (2.1.14) describes how η_n is obtained by first updating η_{n-1} via the Boltzmann-Gibbs operator in light of V_{n-1}, and then mutating the resulting $\widehat{\eta}_{n-1}$ with respect to the Markov kernel. Similarly, (2.1.15) describes how $\widehat{\eta}_n$ is obtained by mutating $\widehat{\eta}_{n-1}$ and updating the resulting η_n in light of V_n, again using the Boltzmann-Gibbs operator.

To relate the Feynman-Kac mutation and update formulae to the filter and prediction distributions (2.1.5)–(2.1.6), by Bayes' theorem and (2.1.4)

we have the following recursion

$$\begin{aligned}\widehat{\pi}_n(\varphi) &\propto \int_{\mathbb{X}} \varphi(x_n) g_n(x_n) \pi_n(dx_n) \\ &= \int_{\mathbb{X}} \int_{\mathbb{X}} \varphi(x_n) g_n(x_n) \widehat{\pi}_{n-1}(dx_{n-1}) K(x_{n-1}, dx_n) \\ &= \widehat{\pi}_{n-1}(g_n K(\varphi)).\end{aligned}$$

From this we deduce the formula

$$\widehat{\pi}_n(\varphi) \propto \widehat{\pi}_0\left(K^n(\varphi) \prod_{p=1}^{n} g_p\right) = \pi_0\left(K^n(\varphi) \prod_{p=0}^{n} g_p\right).$$

However, up to proportionality, the last expression is simply $\widehat{\eta}_n(\varphi)$ as given in (2.1.8)–(2.1.9) in which $V_n = g_n$.

Consequently, the filter and prediction distributions (2.1.5)–(2.1.6) are specific cases of the Feynman-Kac measures (2.1.8)–(2.1.9) and thus inherit the same predict and update recursive structure. However, exact computation of π_n and $\widehat{\pi}_n$ is only possible under restrictive modelling assumptions about the signal and observation processes. This is due to the presence of integrals in each of the formulas that are tractable in only a few cases. For example, if the signal and observation processes take values in Euclidean space and K and G both admit Gaussian density functions, then for all n the distributions π_n and $\widehat{\pi}_n$ are also Gaussian, and the Kalman filter parametrises the exact solution via the respective mean and covariance terms [39]. Alternatively, if $\mathbb{X}$ consists of a finite number of values then the integrals in π_n and $\widehat{\pi}_n$ are finite sums over $\mathbb{X}$ and hence tractable, in which case the solution is found using grid-based methods; cf. [5].

The Kalman filter and grid-based methods offer optimal solutions to special cases of the filtering problem but are sub-optimal whenever the integrals in the computation of π_n and $\widehat{\pi}_n$ are intractable. A number of methods have been derived that seek to approximate solutions to such problems, including the extended Kalman filter for signal and observation processes in Euclidean space [40, 41, 39], and approximate grid-based methods when $\mathbb{X}$ is infinite but can be well approximated by a finite grid [5]. The extended Kalman filter furnishes Gaussian approximations to π_n and $\widehat{\pi}_n$ by performing a linearisation about point estimates of the predictive and posterior means and applying the operations from the Kalman filter to the resulting matrices. However,

these approximations will have limited accuracy if the target distributions are heavily-skewed or bimodal, and in certain scenarios the estimates may diverge. The unscented Kalman filter seeks to improve on this drawback by constructing a Gaussian approximation based on a collection of specified sigma points [37, 77], but for the same reasons struggles in the presence of heavily-skewed or bimodal target distributions. The approximate-grid based method approximates π_n and $\widehat{\pi}_n$ by an empirical measure on a finite grid which, in addition to suffering from the curse of dimensionality, requires truncation of $\mathbb{X}$ if it is unbounded [5].

An alternative means for approximating solutions to (2.1.5)–(2.1.6) is the Monte Carlo (MC) method of constructing estimates of π_n and $\widehat{\pi}_n$ based on $N \in \mathbb{N}$ random samples, known as *particles*, taken from an appropriate distribution. The resulting estimates are formed by constructing empirical distributions π_n^N and $\widehat{\pi}_n^N$ that have as support the particles, each of which are assigned a *weight* that represents its probability. The evolution of π_n^N and $\widehat{\pi}_n^N$ from one iterate to the next is then performed by transforming the empirical distributions in exactly the manner described by (2.1.11) and (2.1.12). Such an approach enables discrete estimates of the exact distributions that possess convergence properties other sub-optimal methods do not, and are robust to non-linear, non-Gaussian features in the signal and observation models.

2.2 Sequential Monte Carlo

2.2.1 Formulation

The purpose of SMC methods is to construct sample-based estimates of (2.1.5)–(2.1.6). Since these estimating measures are discrete — in the sense that they are based on a finite number of samples — there are a number of theoretical and practical issues to consider in their design.

At each iteration $n \in \mathbb{N}_0$ we denote by $(\xi_n^i)_{i=1}^N$ the collection of $\mathbb{X}$-valued particles and by $(w_n^i)_{i=1}^N$ their corresponding non-negative weights. Together, $(\xi_n^i)_{i=1}^N$ and $(w_n^i)_{i=1}^N$ form the empirical estimate π_n^N of π_n, given by

$$\pi_n^N := \sum_{i=1}^N w_n^i \delta_{\xi_n^i}, \tag{2.2.1}$$

where δ_x denotes the Dirac-delta measure with point mass at x. The anal-

ogous weak-form MC estimator of π_n is obtained by integrating $\varphi \in \mathscr{B}(\mathbb{X})$ with respect to (2.2.1) to give

$$\pi_n^N(\varphi) = \int_{\mathbb{X}} \varphi(x_n)\pi_n^N(dx_n) = \sum_{i=1}^N w_n^i \varphi(\xi_n^i). \tag{2.2.2}$$

If the ξ_n^i are independently, identically distributed (i.i.d.) according to π_n, denoted by $\xi_n^i \overset{\text{i.i.d.}}{\sim} \pi_n$, then $w_n^i = N^{-1}$ for all i, and by elementary operations it follows that (2.2.2) is unbiased. The construction of π_n^N using i.i.d. samples from π_n implies that its standard deviation $\sigma(\pi_n^N(\varphi))$ is found via

$$\sigma^2(\pi_n^N(\varphi)) = \sigma^2\left(N^{-1}\sum_{i=1}^N \varphi(\xi_n^i)\right) = N^{-1}\sigma^2(\varphi), \tag{2.2.3}$$

so that $\sigma(\pi_n^N(\varphi)) = \sigma(\varphi)/\sqrt{N}$ whenever the variance $\sigma^2(\varphi) := \mathrm{Var}(\varphi(X_n))$ is finite. Asymptotically, the strong law of large numbers implies that $\pi_n^N(\varphi)$ converges almost surely to $\pi_n(\varphi)$ for all $\varphi \in \mathscr{B}(\mathbb{X})$, while a central limit theorem also applies whenever $\sigma^2(\varphi) < \infty$; see, for example, [63, Ch. 4] and [19, Ch. 1] respectively.

A central issue to SMC methods is how to form convergent estimators when it is not possible to sample directly from π_n, and how to evolve the empirical measure in a way such that the asymptotic properties are preserved. This is accomplished by applying the exact recursive formula (2.1.14) to the empirical measure π_n^N, or equivalently, applying (2.1.15) to the estimate $\widehat{\pi}_n^N$ of $\widehat{\pi}_n$.

Considering the case for π_n^N, suppose that ξ_n^i and w_n^i are known (we note that for $n = 0$, these quantities are computed using standard MC from the prior π_0). Using (2.1.14), the next iteration empirical prediction is specified by

$$\pi_{n+1}^N = \Psi_n(\pi_n^N)K. \tag{2.2.4}$$

In particular, the application of the Boltzmann-Gibbs operator Ψ_n updates π_n^N to $\widehat{\pi}_n^N$ via

$$\widehat{\pi}_n^N := \Psi_n(\pi_n^N) = \sum_{i=1}^N \frac{g_n(\xi_n^i)w_n^i}{\sum_{j=1}^N g_n(\xi_n^j)}\delta_{\xi_n^i} =: \sum_{i=1}^N \widetilde{w}_n^i \delta_{\xi_n^i}. \tag{2.2.5}$$

Then, by applying K:

$$\pi_{n+1}^N(\,\cdot\,) = \widehat{\pi}_n^N K(\,\cdot\,) = \int_{\mathbb{X}} \widehat{\pi}_n^N(dx_n) K(x_n,\,\cdot\,) = \sum_{i=1}^N \widetilde{w}_n^i K(\xi_n^i,\,\cdot\,) \tag{2.2.6}$$

In other words, given π_n^N, the next iteration empirical prediction measure π_{n+1}^N is given by the weighted sum of the probability measures induced by conditioning the Markov kernel on each ξ_n^i, in which the observational information provided by the update step is encoded into the weights. By the normalisation of the $\widetilde{w}_n^i$ it follows that (2.2.6) integrates to unity and is therefore a well-defined probability distribution; to obtain a new empirical prediction measure and thus close the recursion we then sample $\xi_{n+1}^i \sim K(\xi_n^i,\,\cdot\,)$ and set $w_{n+1}^i = \widetilde{w}_n^i$ to give (2.2.1), with n replaced by $n+1$.

2.2.2 Importance sampling

By formulating a discrete representation of the exact predict and update steps, the formulae (2.2.5)–(2.2.6) specify a particle filter algorithm that produces empirical estimates of the intractable measures π_n and $\widehat{\pi}_n$ at each n. It is an extension of importance sampling [28] to the sequential setting, which is based on the idea of sampling from an importance distribution q when sampling from a probability distribution π is not possible, such that $\operatorname{supp}\pi \subseteq \operatorname{supp} q$. This approach is justified by the Radon-Nikodym formula

$$\pi(\varphi) \propto q\left(\varphi \frac{d\pi}{dq}\right), \tag{2.2.7}$$

see [64, Ch. 2.6.]. Based on (2.2.7), an estimate of π is given by instead sampling $\xi^i \overset{\text{i.i.d.}}{\sim} q$ and forming weights based on evaluations of $d\pi/dq$ that correct for the bias of sampling from q. Due to the normalisation term in (2.2.7), the resulting Monte Carlo estimate is based on the ratio of two estimators and is therefore biased for any N. However, under some mild assumptions about π and φ, it also exhibits almost sure convergence and has a central limit theorem; see [28].

In the context of SMC, importance sampling features by considering the distribution of $X_0, \ldots, X_n \mid Y_0 = y_0, \ldots, Y_n = y_n$ at some $n > 0$ and computing its Radon-Nikodym derivative with respect to an importance distribution q. The recursive property of SMC is preserved for a general choice of impor-

tance distribution by virtue of the fact that this Radon-Nikodym derivative can be expressed by constructing the potential function V_n accordingly; thus remaining in the predict-update framework of Feynman-Kac theory. To this end, we note that the choice $V_n = g_n$ represents one particular choice of importance distribution — namely that for which q is chosen to be the Markov kernel — but that one advantage of the Feynman-Kac interpretation is that it encompasses a more general class of importance distributions in the process.

To utilise the Markov property and the information provided by the observations, a general choice of q is often designed to have the form

$$q(\,\cdot\,) = q(X_n \in \,\cdot\, \mid X_{n-1} = x_{n-1}, Y_0 = y_0, \ldots, Y_n = y_n). \tag{2.2.8}$$

One of the principles behind conditioning on the observations in (2.2.8) is that knowledge of the latest observation y_n can be incorporated to "retrospectively" propagate the particles $\xi_n^i \sim q$ towards a neighbourhood of the point in $\mathbb{X}$ that produced it, should such a mapping be known. The idea behind this approach is that the observation guides the particles towards a region that is more likely to be close to the latent state value than by naively mutating conditioned only on an estimate of the previous state. Without loss of generality we continue to assume $V_n = g_n$ in our work, but note that all of the forthcoming discussion also applies to a more general class of importance distributions of the form (2.2.8), the design of which is a non-trivial problem in its own right. For a deeper discussion on sequential importance sampling and design of importance distributions we refer to [5, 16], for example.

2.2.3 Resampling

While the estimators π_n^N and $\widehat{\pi}_n^N$ are theoretically plausible, in practice the majority of the weights $\widehat{w}_n^i \propto g_n(\xi_n^i) w_n^i$ (and hence also w_n^i) are likely to be close to zero after a few iterations. It has been proven that the variance of the weights increases over time [20] and that this issue is therefore unavoidable. This is the well-known *degeneracy problem* that, until the introduction of resampling, inhibited the widespread use of particle filters.

Consequently, the update step described by the Boltzmann-Gibbs operator is typically supplemented by an additional resampling/selection operator $S(x, \,\cdot\,)$ that mitigates the particle degeneracy. For example, if multinomial resampling is performed — in which the resampled particle $\widehat{\xi}_n^i$ is set equal

to ξ_n^j with probability $\widehat{w}_n^j$ — then

$$S_{\pi_n^N}(\xi_n^i, \cdot\,) = \Psi_n(\pi_n^N). \tag{2.2.9}$$

For $\widehat{\xi}_n^i \sim S_{\pi_n^N}(\xi_n^i, \cdot\,)$ we then have $\widehat{\pi}_n^N = N^{-1}\sum_i \delta_{\widehat{\xi}_n^i}$, and the prediction step (2.2.6) proceeds with $\widetilde{w}_n^i = \widehat{w}_n^i = N^{-1}$.

The operator (2.2.9) is part of a larger class of mixture-type models in which resampling is only performed with probability depending on ξ_n^i; see [54, Ch. 3]. These models are somewhat analogous to the application of an accept/reject step to proposed samples in MCMC [62]. Furthermore, there exists a rich collection of resampling schemes other than multinomial sampling such as stratified, residual and deterministic, to name a few [48].

While resampling provides a solution to the degeneracy problem, it has several drawbacks. By drawing the resampled particles $\widehat{\xi}_n^i$ from $\widehat{\pi}_n^N$, additional statistical error is introduced into the resulting filter estimator. For this reason, the pre-resample distribution is preferable in practice for the purpose of forming estimates, while the post-resample distribution is preferred for proving convergence properties, since any result will also hold for its pre-resample counterpart.

Another drawback is that, by requiring the particles to interact, the ability to parallelise the particle filter is compromised. Though the cost of resampling is cheap and dimension-free ($\mathcal{O}(N)$), the inability to easily partition the particles into independent patches makes the typically-more costly mutation/weight computation steps more problematic. More recently, methods such as the island particle filter [75] and butterfly resampling [32] have provided ways to mitigate this problem. Another related issue is that of *sample impoverishment* [5], in which repeated resampling of a small subset of particles leads to a lack of diversity in the generated offspring. This issue is particularly prevalent when the signal-to-noise ratio is large, since fewer particles tend to be assigned a greater share of the total weight density, thus increasing their likelihood of being replicated a large number of times [11].

A popular approach to limiting the drawbacks of resampling is to only do so when the particle diversity falls below a tolerable threshold. The theoretical diagnostic by which this is measured is the effective sample size

of the particles, given by

$$N_{\text{ESS}} := \frac{N}{1 + \text{Var}(\widehat{w}_n^{*i})}, \tag{2.2.10}$$

where $\widehat{w}_n^{*i}$ is the "true weight" of ξ_n^i; see [43, 49]. Although $\widehat{w}_n^{*i}$ cannot generally be computed exactly, a widely used estimator of (2.2.10) is

$$\widehat{N}_{\text{ESS}} := \frac{1}{\sum_{i=1}^N (\widehat{w}_n^i)^2}. \tag{2.2.11}$$

By its construction and the features of the weights, the quantity $\widehat{N}_{\text{ESS}}$ is bounded via $1 \leq \widehat{N}_{\text{ESS}} \leq N$, with the upper and lower bounds corresponding to the best and worst case degeneracy scenarios respectively. To see this, by Minkowski's inequality and the normalisation of the weights,

$$\sum_{i=1}^N (\widehat{w}_n^i)^2 \leq \left(\sum_{i=1}^N \widehat{w}_n^i \right)^2 = 1.$$

Furthermore, writing each weight as $\widehat{w}_n^i = N^{-1} + \epsilon_i$ we obtain the following lower bound

$$\sum_{i=1}^N (\widehat{w}_n^i)^2 = \sum_{i=1}^N \left(\frac{1}{N} + \epsilon_i \right)^2 = \frac{1}{N} + \sum_{i=1}^N \epsilon_i^2 \geq \frac{1}{N}, \tag{2.2.12}$$

where we have used the fact that $\sum_i \epsilon_i = 0$. Intuitively, the optimal scenario in which $\widehat{w}_n^i = N^{-1}$ for all i corresponds to standard MC sampling, where N i.i.d. samples are taken from the true distribution; i.e. where sampling from $\widehat{\pi}_n^N$ is equivalent to sampling from $\widehat{\pi}_n$. At the other extreme, using a similar argument to (2.2.12), the value $N_{\text{ESS}} = 1$ is obtained if and only if all weights but one are zero, in which case inference is essentially being performed using one particle.

The estimator (2.2.11) can therefore be incorporated into a particle filtering method by setting a degeneracy tolerance $1 \leq \tau \leq N$ and resampling whenever $\widehat{N}_{\text{ESS}} \leq \tau$. This adaptive approach can also be extended to particle filters with certain constrained resampling schemes, in which interaction of only a subset of particles needs be performed in order to achieve time uniform convergence [78]. In the special case where $\tau \equiv N$, we note that resampling is performed at every step.

2.2.4 Bootstrap particle filter

With the choice of potential function $V_n = g_n$ and the inclusion of a resampling step, the resulting interacting particle system is the *bootstrap particle filter* (BPF). The BPF is a popular choice of particle filter for its ease of

Algorithm 1 Bootstrap Particle Filter (BPF)

```
% Initialisation
for i = 1, ..., N do
    ξ_n^i ~ π_0 and w_0^i = 1/N.
for n ≥ 0 do
    % Calculate weights
    for i = 1, ..., N do
        w̃_n^i = g_n(ξ_n^i) w_n^i
    % Resampling
    for i = 1, ..., N do
        ξ̂_n^i ~ (Σ_{i=1}^N w̃_n^i δ_{ξ_n^i}) / (Σ_{i=1}^N w̃_n^i) and ŵ_n^i = 1/N
    % Mutation
    for i = 1, ..., N do
        ξ_{n+1}^i ~ K(ξ̂_n^i, ·) and w_{n+1}^i = ŵ_n^i
```

implementation, since its weight computation step involves only computations involving g_n. For further ease of implementation, resampling is often performed at every iteration since it essentially negates the need to store the weights from the previous iteration. Such an algorithm is presented in Algorithm 1. It is, however, just one example of a much broader class of particle filters, with practically every step applicable for modification, often in ways that obtain better performance (see [19, Ch. 13-14], for example). Nonetheless, the BPF captures in an intuitive way each of the key processes that make up particle filtering, and the alternative approaches taken in other versions are essentially generalisations of these core principles. For this reason, we use the BPF as a template for our own particle filtering method.

For $\widehat{\pi}_n^i$ and $\widehat{w}_n^i$ computed as in Algorithm 1, let $\widehat{\pi}_n^N$ denote the resulting empirical filter measure. The following strong law of large numbers and central limit theorems provide a theoretical justification for the use of particle filtering methods.

Theorem 1 (Crisan and Doucet [16]). *For a Markov kernel K satisfying*

the Feller property and for g_n such that $g_n > 0$ is bounded and continuous for all n,

$$\widehat{\pi}_n^N(\varphi) \xrightarrow[N\to\infty]{a.s.} \widehat{\pi}_n(\varphi).$$

Theorem 2 (Chopin [13])**.** *For all bounded and measurable $\varphi : \mathbb{X} \mapsto \mathbb{R}$,*

$$\sqrt{N}\left(\widehat{\pi}_n^N(\varphi) - \widehat{\pi}_n(\varphi)\right) \xrightarrow[N\to\infty]{\mathrm{D}} \mathcal{N}(0, \widehat{\sigma}_n^2(\varphi))$$

for some $\widehat{\sigma}_n^2(\varphi) \in (0, \infty)$.

We note that, in light of the previous discussion, each of the convergence results also apply to the measure $\widehat{\pi}_n$ prior to resampling, in which the weights $\widetilde{w}_n^i$ are instead proportional to $g_n(\xi_n^i)$.

2.3 Multilevel Monte Carlo

Multilevel Monte Carlo (MLMC) methods are applicable to expectation estimation problems in which there is an associated numerical cost to generating each sample. Given this cost, it is assumed that samples based on a cheaper but less accurate solution can be generated in a way that satisfies some general assumptions. In this case, it can be shown that MLMC can exploit its multilevel solution hierarchy in such a way that it provides equivalent accuracy to standard MC in fewer floating point operations. The strength of the computational gains in works such as [15, 29] is such that we use the approach as a starting point for our own adaption of multilevel methods to SMC. Consequently, in this section we briefly review the general principles behind the method while remaining in the Feynman-Kac framework ahead of deriving our multilevel SMC method.

Formally, we pause our consideration of sequential problems and instead consider the class of problems for which estimates are sought for

$$\eta_{\varphi(X_T)} := \mathbb{E}_\eta[\varphi(X_T)], \tag{2.3.1}$$

where $X_t = X(t, \omega)$ is an $\mathbb{X}$-valued stochastic process, X_T is a random variable defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ at some fixed $T \in \mathbb{R}$, and $\eta(\,\cdot\,) := \mathbb{P}(X_T \in \,\cdot\,)$. A crucial feature of problems in the MLMC setting is that φ is generally assumed *fixed*, meaning the integral $\eta_{\varphi(X_T)}$ is an unknown

scalar instead of the weak form probability measure considered in the SMC setting. To emphasise this distinction we place $\varphi(X_T)$ as a subscript whenever the expectation of $\varphi(X_T)$ is taken with respect to the distribution η of X_T.

Problems of the type (2.3.1) frequently arise in applications involving stochastic differential equations such as mathematical finance [29], in which $\eta_{\varphi(X_T)}$ is the expected value of a portfolio option at an exercise time T; and uncertainty quantification in groundwater flow [15], in which $\eta_{\varphi(X_T)}$ is the expected value of a functional of an elliptic PDE solution with random coefficients. Due to exact samples $\xi_T^i \sim \eta$ of X_T typically being unavailable, the direct MC estimate of (2.3.1) is therefore also infeasible and a different strategy is required.

Instead, a common approach is to first approximate X_t by an $\mathbb{R}^M$-valued random vector X_t^M, for which the component corresponding to $t = T$ represents an approximation to the random variable X_T. The resulting numerical approximation to $\eta_{\varphi(X_T)}$ is then given by

$$\eta_{\varphi(X_T^M)} := \mathbb{E}_\eta[\varphi(X_T^M)]. \tag{2.3.2}$$

Equation (2.3.2) is useful because it provides a pathway to estimating (2.3.1) via samples based on numerical approximations of X_t. Under certain assumptions about the convergence of (2.3.2) to (2.3.1) as $M \to \infty$, an MC estimate of (2.3.1) of arbitrary accuracy can in theory be constructed by using a large enough number of samples to estimate (2.3.2) with a sufficiently-large value of M. For clarity of notation, we use the abbreviation $\eta_\varphi^M := \eta_{\varphi(X_T^M)}$ of the scalar-valued estimator of $\eta_{\varphi(X_T)}$, which we also abbreviate to η_φ.

As per [15], we write $a \lesssim b$ whenever a/b is uniformly bounded independently of any parameters (the sample size or approximation resolution, for example) and write $a \simeq b$ whenever $a \lesssim b$ and $b \lesssim a$. Furthermore, we assume $\eta_\varphi^M \to \eta_\varphi$ in mean as $M \to \infty$ with order of convergence $\alpha > 0$ so that, for any M, the bias is bounded in the following sense

$$|\eta_\varphi^M - \eta_\varphi| \lesssim M^{-\alpha}. \tag{2.3.3}$$

For a fixed $M \in \mathbb{N}$, we denote a generic unbiased MC estimator of η_φ^M comprising of N samples $\varphi(\xi_T^{M,i})$ by $\eta_\varphi^{M,N}$. To measure the accuracy of $\eta_\varphi^{M,N}$ from the true expectation η_φ requires a metric, which is typically chosen to

be the root mean squared error (RMSE)

$$\varepsilon(\eta_\varphi^{M,N}) := \mathbb{E}_{\eta^M}[(\eta_\varphi^{M,N} - \eta_\varphi)^2]^{1/2}. \tag{2.3.4}$$

In addition to being a practical choice, (2.3.4) is a useful tool for analysis on account of the fact that, for any unbiased estimator $\eta_\varphi^{M,N}$ of η_φ^M,

$$\varepsilon^2(\eta_\varphi^{M,N}) = \sigma^2\big(\eta_\varphi^{M,N}\big) + \big(\eta_\varphi^M - \eta_\varphi\big)^2, \tag{2.3.5}$$

thus decomposing $\varepsilon^2(\eta_\varphi^{M,N})$ in terms of its variance and numerical error (for a derivation of (2.3.5), see [15], for example). In particular, the numerical error $|\eta_\varphi^M - \eta_\varphi|$ can only be reduced by increasing the resolution parameter M, while the variance term is a function of both M and the sample size N. The legacy of MLMC is that the influence of these two parameters can be exploited to obtain a specified RMSE threshold in fewer floating point operations than standard MC.

To quantify the cost of an estimator, let $\mathcal{C}(\eta_\varphi^{M,N})$ denote the number of floating point operations performed with $\eta_\varphi^{M,N}$, and for a tolerance $\epsilon > 0$, let $\mathcal{C}_\epsilon(\eta_\varphi^{M,N})$ denote the cost of obtaining an RMSE of ϵ with $\eta_\varphi^{M,N}$. For some $\gamma > 0$, we assume the cost of generating each $\varphi(\xi_T^{M,i})$ is $\mathcal{C}(\varphi(\xi_T^{M,i})) \lesssim M^\gamma$, and hence for N samples

$$\mathcal{C}(\eta_\varphi^{M,N}) \lesssim NM^\gamma. \tag{2.3.6}$$

If $\eta_\varphi^{M,N}$ is chosen to be the standard MC estimator

$$\eta_{\varphi,\mathrm{MC}}^{M,N} := N^{-1}\sum_{i=1}^N \varphi(\xi_T^{M,i}), \qquad \xi_T^{M,i} \overset{\text{i.i.d.}}{\sim} \eta_\varphi^M$$

then by (2.3.5),

$$\varepsilon^2(\eta_{\varphi,\mathrm{MC}}^{M,N}) = N^{-1}\sigma_M^2(\varphi) + (\eta_\varphi^M - \eta_\varphi)^2, \tag{2.3.7}$$

where $\sigma_M^2(\varphi) = \mathrm{Var}(\varphi(X_T^M))$. In the presence of no numerical error, we note that (2.3.7) reduces to the exact same standard MC variance (2.2.3).

A sufficient condition to attain an RMSE of order ϵ with $\eta_{\varphi,\mathrm{MC}}^{M,N}$ is that each of the terms in (2.3.7) are of order ϵ^2. This implies $N \gtrsim \epsilon^{-2}$ and, by the numerical bias assumption (2.3.3), that $M \gtrsim \epsilon^{-1/\alpha}$, so that according to

(2.3.6) we have

$$\mathcal{C}_\epsilon(\eta^{M,N}_{\varphi,\mathrm{MC}}) \lesssim \epsilon^{-2-\gamma/\alpha}.$$

In particular, the ϵ-cost of $\eta^{M,N}_{\varphi,\mathrm{MC}}$ is governed by the ratio γ/α; if the application is expensive and hence γ/α is large, this correlates to the rate of convergence of the estimator being slow compared to the cost of generating the approximate solutions $\xi^{M,i}$.

In the multilevel approach, a sequence of $L+1$ *levels* $(M_\ell)_{\ell=0}^L$ is constructed such that $M_\ell > M_{\ell-1}$ for all ℓ and $M_L = M$. The purpose of the levels is to parametrise the complexity of the corresponding numerical approximation $X_T^{M_\ell}$. Therefore by the assumptions (2.3.3) and (2.3.6), samples produced on the lower levels exhibit greater bias but are cheaper to generate, while for increasing ℓ, the bias decreases at the expense of a higher solution generation cost.

A key principle of MLMC is that the integral η^M_φ can be written as the telescoping sum

$$\eta^M_\varphi = \sum_{\ell=0}^{L} \eta^\ell_{\Delta\varphi}, \tag{2.3.8}$$

where

$$\eta^\ell_{\Delta\varphi} := \begin{cases} \mathbb{E}_\eta[\varphi(X_T^{M_\ell}) - \varphi(X_T^{M_{\ell-1}})] & \text{for } \ell > 0, \\ \mathbb{E}_\eta(\varphi(X_T^{M_\ell})) & \text{for } \ell = 0, \end{cases} \tag{2.3.9}$$

meaning that η^M_φ can be expressed exactly as the expectation of $\varphi(X_T^{M_0})$ plus a sum of terms that correct the resulting bias from the previous level.

Given the exact expression (2.3.8), the multilevel approach is to construct independent, unbiased estimators $\eta^{M_\ell,N_\ell}_{\Delta\varphi}$ of each of the $\eta^\ell_{\Delta\varphi}$ using N_ℓ samples. In particular, if the standard MC estimator based on i.i.d. samples is used on each level, then the MLMC estimator of η^M_φ is

$$\eta^{M_\ell,N}_{\Delta\varphi,\mathrm{ML}} := \sum_{\ell=0}^{L} \eta^{M_\ell,N_\ell}_{\Delta\varphi} = \sum_{\ell=0}^{L} N_\ell^{-1} \sum_{i\in P_\ell} \left[\varphi(\xi_T^{M_\ell,i}) - \varphi(\xi_T^{M_{\ell-1},i})\right], \tag{2.3.10}$$

where N is interpreted as the total number of samples, and where we define P_ℓ to be the set of indices corresponding to the samples assigned to the MC estimate of $\eta^\ell_{\Delta\varphi}$. While in the context of MLMC this notation is somewhat superfluous, when applying the multilevel approach in an SMC setting it is required to handle in a rigorous way the effects of resampling particles over

all levels. It is also important to note that, for each $i \in P_\ell$ and fixed $\ell > 0$, the solutions $\xi^{M_\ell,i}$ and $\xi^{M_{\ell-1},i}$ must be generated using the same underlying sample $\omega^i \in \Omega$ to ensure that $\eta_{\Delta\varphi}^{M_\ell,N_\ell}$ is a consistent estimator of $\eta_{\Delta\varphi}^{\ell}$.

Applying the RMSE decomposition formula (2.3.5) to the MLMC estimator gives

$$\varepsilon^2(\eta_{\varphi,\mathrm{ML}}^{M,N}) = \sum_{\ell=0}^{L} N_\ell^{-1}\Delta\sigma_\ell^2(\varphi) + (\eta_\varphi^M - \eta_\varphi)^2. \tag{2.3.11}$$

where $\Delta\sigma_\ell^2(\varphi)$ is the variance of the level-specific φ difference terms in (2.3.9). Thus, the effect of formulating the MLMC estimator is to generalise the unilevel variance in (2.3.7) to the sum of the level-wise estimator variances, while leaving the bias term unchanged. By (2.3.6), the cost of generating (2.3.11) is

$$\mathcal{C}(\eta_{\mathrm{ML}}^M) \lesssim \sum_{\ell=0}^{L} N_\ell \mathcal{C}_\ell,$$

where $\mathcal{C}_\ell$ is the cost of generating a single sample according to (2.3.9). To attain ϵ-RMSE accuracy, it therefore once again suffices to choose $M \gtrsim \epsilon^{-1/\alpha}$. To sufficiently attenuate the remaining error requires the sum in (2.3.11) to be of order ϵ^2 which, for reasons given in [15], can generally be achieved with fewer floating point operations than standard MC. One intuition is that if $\Delta\sigma_\ell^2(\varphi)$ is large then the cost of at least one of $X^{M_{\ell-1}}$ is likely to be comparatively low, meaning N_ℓ can be made large without placing substantial demands on the computational budget. In particular, the main theorem in [15] provides a methodology for computing approximately-optimal values of each N_ℓ by basing them on empirical estimates of $\Delta\sigma_\ell^2(\varphi)$. Since these estimates arise as a natural by-product of the sampling, they therefore enable the estimate accuracy to be improved at a negligible cost to the overall efficiency.

To facilitate later comparisons with our multilevel method we conclude this chapter by stating the MLMC algorithm of [15], which explicitly details how to compute the approximately-optimal number of level specific samples N_ℓ to leverage the efficiency gains of MLMC.

For a fixed L, Algorithm 2 proceeds by iterating through all levels up to L. In particular, for each $\ell = 0, \ldots, L$ an initial collection of samples are generated which are then used to estimate the optimal number of samples N_ℓ. The formula for estimating N_ℓ is given in terms of the level-specific

Algorithm 2 Multilevel Monte Carlo (MLMC)

```
% Initialisation
```
Set $L = 0$ and generate an initial number of samples $\xi^{M_L,i}$.
```
% Estimate the variance
```
Construct the estimate $\widehat{\sigma}_L^2$ of $\sigma_{M_L}^2(\varphi)$
for $\ell = 0, \ldots, L$ **do**
```
    % Estimate the optimal sample sizes
```
    Set $N_\ell \simeq \sqrt{\widehat{\sigma}_\ell^2 / \mathcal{C}_\ell}$
```
    % Evaluate extra samples
```
    Evaluate extra samples $\xi^{M_\ell,i}$ as needed for the computed N_ℓ
```
% Test for convergence
```
if $L \geq 1$ **then**
    if $\widehat{\sigma}_L^2 \simeq M^{-\alpha}$ **then**
        return η_{MLMC}^M
```
% Repeat
```
Set $L = L + 1$ and go back to the variance estimation step.

standard deviation and cost-per-sample $\widehat{\sigma}_\ell$ and $\mathcal{C}_\ell$ respectively, which are a natural by-product of the generated collection of samples. On each level, the derived estimate of N_ℓ then informs the user whether more samples are required. The multilevel MC estimation is then a simple case of assigning the N_ℓ samples to each level-specific MC estimator and testing for convergence using the specified convergence criteria, which can also be estimated using the numerical bias assumption (2.3.3). If the convergence criterion is not satisfied then L is incremented and the process is repeated; by increasing L, this in effect "inserts" another M_ℓ in between M_0 and M_L, since we always require $M_L = M$ in order for the MLMC estimator to converge to η_φ^M.

Chapter 3

Multilevel bootstrap particle filter

3.1 Method

While SMC methods such as the BPF are credible theoretical approaches to hidden parameter estimation problems, in practice they can be infeasible when the cost of generating the weights is high. The drilling application described in Chapter 1 is one such example, in which each weight entails the solution of a system of PDEs. Even in applications where the weight generation cost is not high, a reduction in this cost is still desirable since it enables the configuration of more particles and hence asymptotically-more accurate estimators.

We recall that, for $n \geq 0$, the filter distribution $\widehat{\pi}_n$ is obtained from π_n by Bayes' formula

$$\widehat{\pi}_n(\varphi) = \frac{\pi_n(g_n\varphi)}{\pi_n(g_n)}, \qquad \varphi \in \mathscr{B}(\mathbb{X}). \tag{3.1.1}$$

In applications where the exact form of π_n is unknown, we saw in Section 2.2 that SMC methods use (3.1.1) to update an empirical estimate π_n^N of π_n to $\widehat{\pi}_n^N$ by replacing the exact measures by the empirical ones and defining weights based on evaluations of g_n. In light of the related asymptotics, (3.1.1) is both the limiting case of the empirical measures and the means by which the cost of an empirical measure formed using weights based on g_n can be quantified. To analyse the weight computation cost, we therefore consider

this exact formula.

In addition to the statistical estimation, it is often the case that the likelihood evaluations $g_n(\xi_n^i)$ are based on equations that have no exact solution and require numerical approximation. Most relevantly to our application this approximate solution could be based on a finite element or finite volume PDE method, for example, but equally could be based on a numerical integration method or an iterative solution to an optimisation problem. In Section 2.3 we saw how MLMC exploits a cost-accuracy trade-off that arises when a multilevel approach is applied to numerical approximations of a particular functional of interest; by constructing the solver resolution and level-specific sample sizes in an optimal way, an equivalent level of accuracy to standard MC is achieved at a lower computational complexity [15, 29].

In our method we seek to emulate these gains in the SMC setting by considering an analogous multilevel approach for the likelihood function g_n and exploiting the linearity of the integral operators in the numerator and denominator of (3.1.1). The resulting challenges of this task differ from MLMC in two significant ways. Firstly, MLMC is designed to be applied to problems in which the expectation of only one functional is of interest and thus the exact solution is scalar-valued. By contrast, in filtering problems the expectation is considered over *all* $\varphi \in \mathscr{B}(\mathbb{X})$ to describe the weak form of a probability measure. Moreover, in this setting we are in fact estimating sequences of probability measures that are constructed using the results of non-linear filtering. It is therefore non-trivial that the multilevel-based estimates converge to their target distributions and that these asymptotics are preserved at each iteration. With these challenges in mind, we remark that our approach extends MLMC in two senses: from one particular integral to an integral functional, and from the non-sequential setting to the sequential one.

3.1.1 Algorithm

Let $L \in \mathbb{N}_0$ and, given the full accuracy likelihood function $g_n^L := g_n$, construct a sequence of level-based likelihoods $(g_n^\ell)_{\ell=0}^L$ such that both the computational cost and accuracy of g_n^ℓ with respect to some metric increases with ℓ. We note that the generality of these assumptions allows a broad range of applications to be considered. Each ℓ could, for example, denote the number of spatial discretisation points M_ℓ in a PDE solver, or denote a

sparse matrix based on a subset of entries from a covariance matrix; see [31].

Considering the numerator of (3.1.1), we invoke the multilevel telescoping sum and linearity technique to get

$$\pi_n(g_n\varphi) = \sum_{\ell=0}^{L} \pi_n(\Delta g_n^\ell \varphi),$$

where $\Delta g_n^\ell := g_n^\ell - g_n^{\ell-1}$ and $g_n^{-1} \equiv 0$. By also applying this formula to the normalisation term for which $\varphi \equiv 1$, the filter distribution can be expressed in terms of level-specific filters $\widehat{\pi}_{n,\ell}$ to give the multilevel update step

$$\widehat{\pi}_n(\varphi) = \sum_{\ell=0}^{L} p_{n,\ell}\widehat{\pi}_{n,\ell}(\varphi), \tag{3.1.2}$$

where

$$p_{n,\ell} := \frac{\pi_n(\Delta g_n^\ell)}{\sum_{\ell'=0}^{L} \pi_n(\Delta g_n^{\ell'})}, \qquad \widehat{\pi}_{n,\ell}(\varphi) = \frac{\pi_n(\Delta g_n^\ell \varphi)}{\pi_n(\Delta g_n^\ell)}. \tag{3.1.3}$$

By taking this approach, the standard update step is reformulated as the sum of $L+1$ update steps based instead on level-specific corrections $\Delta g_n^\ell \varphi$, with coefficients $p_{n,\ell}$ that are tractable once the normalisation terms $\pi_n(\Delta g_n^\ell)$ are known over all levels. By virtue of the introduction of the g_n^ℓ, an added flexibility is provided by the $\widehat{\pi}_{n,\ell}$, which are cheaper to compute than the full cost/accuracy $\widehat{\pi}_n$ at the trade-off of varying degrees of level-dependent bias.

Since the corrective difference functions $\Delta g_n^\ell \varphi$ are in no way guaranteed to be non-negative, the level-specific measures $\widehat{\pi}_{n,\ell}$ are in general *signed measures* that combine to provide the formal probability measure $\widehat{\pi}_n$. This extension from classes of probability measures to classes of signed measures has profound consequences on the resulting filtering process, since the weights generated according to Δg_n^ℓ are also no longer guaranteed to be non-negative. Consequently, any prospective multilevel particle filter must accommodate these features correctly to produce a workable algorithm.

The form taken by (3.1.2)–(3.1.3) suggests that a multilevel particle filter can be constructed by applying the standard SMC approach to each of the level-specific update steps (3.1.3). To this end, let $(c_\ell)_{\ell=0}^{L}$ be a sequence of positive integers such that $c_L = 1$. For $N \in \mathbb{N}$, we denote the sample size

of the ℓ-th level particle estimate by $N_\ell = c_\ell N$, so that the sample size over all levels is $S(N) := \sum_{\ell=0}^{L} N_\ell$. Moreover, we assume the particles assigned to each level are indexed according to the partitioning sets

$$P_\ell^N := \{I_\ell(N) + 1, \dots, I_{\ell+1}(N)\}, \qquad 0 \le \ell \le L,$$

where $I_\ell(N) := \sum_{k=0}^{\ell-1} N_k$ and $\sum_a^b(\,\cdot\,) = 0$ whenever $a > b$. The *multilevel bootstrap particle filter* (MLBPF) is defined in Algorithm 3.

Algorithm 3 Multilevel Bootstrap Particle Filter (MLBPF)

```
% Initialisation
```
for $i = 1, \dots, S(N)$ **do**
  $\xi_n^i \sim \pi_0$ and $w_0^i = 1$.
for $n \ge 0$ **do**
```
  % Calculate weights for each level
```
  for $0 \le \ell \le L$ **do**
    for $i \in P_\ell^N$ **do**
      $\widetilde{w}_n^i = N_\ell^{-1}(g_n^\ell(\xi_n^i) - g_n^{\ell-1}(\xi_n^i))w_n^i$
```
  % Signed resampling
```
  for $i = 1, \dots, S(N)$ **do**
    $\widehat{\xi}_n^i \sim \dfrac{\sum_{i=1}^{S(N)} |\widetilde{w}_n^i| \delta_{\xi_n^i}}{\sum_{i=1}^{S(N)} |\widetilde{w}_n^i|}$ and $\widehat{w}_n^i = \operatorname{sgn}\left(\sum_{i=1}^{S(N)} \widetilde{w}_n^i \mathbb{I}[\widehat{\xi}_n^i = \xi_n^i]\right)$
```
  % Mutation
```
  for $i = 1, \dots, S(N)$ **do**
    $\xi_{n+1}^i \sim K(\widehat{\xi}_n^i, \cdot\,)$ and $w_{n+1}^i = \widehat{w}_n^i$

For $L = 0$, we see that the MLBPF reduces exactly to the BPF; given the predictive weights w_n^i on level 0, the unnormalised, pre-resampled weights $\widetilde{w}_n^i$ coincide exactly with the non-negative weights in the BPF, up to a constant that is cancelled out during the normalisation step. In particular, in this unilevel case we remark that the signed resampling step in Algorithm 3 reduces to the conventional multinomial resampling as defined in the BPF.

For $L > 0$, the situation is radically altered by the fact that both the filter and predictive weights can take negative values. This feature is a natural expression of forming weighted MC estimates of the signed measures in (3.1.2), but it means that resampling can no longer be performed with respect to a probability proportional to each weight, since a negative number has no meaningful probabilistic interpretation. Before discussing the signed

resampling step we've implemented in Algorithm 3 that handles this weight negativity in a rigorous way, we first describe the two scenarios in which a negative weight can arise.

<u>$g_n^\ell(\xi_n^i) < g_n^{\ell-1}(\xi_n^i)$ and $w_n^i > 0$:</u>

In this situation a larger likelihood value is assigned to ξ_n^i under the less accurate approximation than the more accurate one, while the predictive weight w_n^i is positive. By construction the w_n^i are all initialised to be positive and, due to the recursive nature of Algorithm 3, will remain positive until at least one particle induces the inequality $g_n^\ell(\xi_n^i) < g_n^{\ell-1}(\xi_n^i)$. Consequently, this scenario is the origin of all weight negativity.

Due to the fact that $g_n^{-1} \equiv 0$, we also note that this current scenario can only occur on levels for which $\ell > 0$. In practice what is happening here is that, due to the bias of the lower accuracy likelihood, regions of $\mathbb{X}$ that in reality do not align with the realised observation are incorrectly deemed as doing so, to an extent that exceeds the higher accuracy likelihood. Additionally, since $w_n^i > 0$, a particle belonging to this less credible region was mutated from one at the previous iteration for which this was not the case (assuming $n > 0$). While this scenario is the origin of all weight negativity, in contrast the following case can only occur after this current one has done so at some iteration.

<u>$g_n^\ell(\xi_n^i) > g_n^{\ell-1}(\xi_n^i)$ and $w_n^i < 0$:</u>

By construction of Algorithm 3, in this scenario we necessarily have $n > 0$, and $w_n^i = \widehat{w}_{n-1}^i < 0$ is therefore the result of a particle with a negative weight being resampled at the previous iteration to give

$$\widehat{w}_n^i = \operatorname{sgn}(\widetilde{w}_n^i(\widehat{\xi}_n^i)),$$

where $\operatorname{sgn}(x) := \mathbb{1}[x > 0] - \mathbb{1}[x < 0]$. For $\ell > 0$, a negative weight arises as the result of another accuracy "switch", for which ξ_{n-1}^i was this time located in a region of weight negativity, was resampled, and then mutated into a more credible region in light of the new observation. We note this is not the situation for $\ell = 0$, in which $\widetilde{w}_n^i < 0$ whenever $\widehat{w}_{n-1}^i < 0$ due to $g_n^0 > 0$. Consequently, this scenario is the only one in which a level 0 particle can ever be assigned a negative weight, since in the previous one the condition $g_n^\ell(\xi_n^i) < g_n^{\ell-1}(\xi_n^i)$ can never be true when $\ell = 0$.

We therefore see from both of the above cases that, for $\ell > 0$, a negative weight arises when a particle jumps between regions of low and high relative credibility between iterations, while for $\ell = 0$ the negativity is inherited solely from resampling a particle with a negative weight. Beyond describing how a negative weight can arise, the practical intuition of why a "switch" happens and what it means is somewhat difficult to grasp, other than asserting that Algorithm 3 is what we get when constructing a convergent MLBPF method. It is not currently clear for example, whether one of the described weight negativity scenarios indicates more accurate estimates than the other, or whether there is a desirable proportion of weight negativity required to elicit an optimal performance from the algorithm. While these questions could have valuable answers that provide greater insight into how to calibrate the MLBPF, they may also be superfluous and simply distract from the fact that the negativity scenarios take the form they do precisely because that's how convergent estimators of sequences of signed measures are leveraged.

3.1.2 Total variation resampling

In contrast to the BPF, resampling in the MLBPF is performed across all levels with respect to a distribution proportional to the *total variation measure*

$$\sum_{i=1}^{S(N)} |\widetilde{w}_n^i| \delta_{\xi_n^i}. \tag{3.1.4}$$

We immediately see that the weights $\widetilde{w}_n^i$ used in conventional multinomial resampling are replaced by those proportional to their absolute value $|\widetilde{w}_n^i|$, while consideration over the entire particle ensemble as used in conventional multinomial resampling is retained. One consequence of this global mixing property which we will shortly prove is that the level-specific signed measures $\widehat{\pi}_{n,\ell}^N$ in fact converge to the *global* posterior $\widehat{\pi}_n$ rather than the level-specific signed measure $\widehat{\pi}_{n,\ell}$ as defined in (3.1.3). Furthermore, we see from Algorithm 3 that in addition to a particle being resampled according to (3.1.4) that its sign is also paired in this process. The inclusion of these terms is necessary to ensuring the correct asymptotics of the resulting signed measures and they can be easily interpreted as a method-specific predictive weight as we have done in Algorithm 3. Consequently, weight negativity can be transferred from any one level to another, in particular according to the manner

described previously in Subsection 3.1.1.

The non-negative weights in (3.1.4) are those that arise when the Radon-Nikodym derivative is instead considered with respect to the total variation measure. While multinomial resampling may not be the only means by which resampling could be performed (such as stratified or systematic, for example), the approach we have taken is crucial to producing convergent estimators. In contrast, it would not be valid in our setting to modify negative weights in a way that can be done in random-weight importance sampling such as [23], since they are crucial in constructing empirical measures that asymptotically correct the error from the previous level. Furthermore, in the presence of a resampling scheme based on (3.1.4) the concept of a properly weighted sample as in [50] holds, since our measure approximations are still using the principles of proper weighting, even if they are approximating signed measures instead of conventional positive measures.

However, by resampling according to (3.1.4), we adopt a different perspective to the standard one in a way that accommodates the features of our multilevel approach. Firstly, for a level 0 particle the conventional intuition of resampling with probability proportional to its Radon-Nikodym derivative is retained: particles that are assigned larger likelihood values are more likely to be resampled. For $\ell > 0$ this intuition is replaced with a different one. In this case the resulting $|\widetilde{w}_n^i|$ instead quantify the magnitude of the *numerical error* between levels. To this extent the size of the term $g_n^\ell(\xi_n^i)$ becomes redundant; rather, the magnitude from which it differs from $g_n^{\ell-1}(\xi_n^i)$ determines the probability with which ξ_n^i is resampled. Intuitively, the particles that exhibit greater level-wise solution error are assigned a higher importance for resampling, since they are located in regions in which the resulting level-specific measures are most distinct from one another. In other words, if these particles were to somehow be ignored, then the loss of their contribution to correcting the telescoping sum would be more significant than by selecting particles in regions where the solution discrepancy is small. This would be undesirable, so therefore these particles are assigned a higher probability of survival. In this context the sign of $\widetilde{w}_n^i$ is also somewhat irrelevant compared to its magnitude, which in turn is captured by (3.1.4).

While facilitating a well-defined, operational particle filter, the implications of resampling according to (3.1.4) can be more subtle than they may first appear. For example, consider the two-level case, i.e. where $L = 1$.

For regions of $\mathbb{X}$ in which g_n^0 is very accurate (with respect to g_n^1), the level 1 particles $\xi_n^i \in P_1^N$ located in these regions will have a small probability of survival due to the correction terms begin small, while those in level 0 $\xi_n^i \in P_0^N$ will have a relatively large one. Conversely, level 0 particles in regions for which g_n^0 is not accurate will have a small probability of survival, while those in level 1 a relatively large one. We can reverse this logic to say that, given a resampled particle, if it came from level 1 then it is more likely to have been resampled from a region in which g_n^0 is inaccurate, while if it came from level 0 then it more likely came from a region in which g_n^0 is accurate. At first thought, the latter part of that statement seems to contradict the former: why would we for one level wish to resample according to the accuracy of g_n^0, while on another according to its inaccuracy? The key point is that, for the special case of $\widehat{\pi}_{n,0}$, we *want* to sample from this distribution as much as we can, all the time that it resembles $\widehat{\pi}_{n,1}$. As soon as this isn't the case, our priorities change and we want to then sample from regions that correct for this difference in a way that is consistent with the density of the approximating total variation measure. Indeed, to resample particles from regions in which $g_n^1 - g_n^0$ is small would be somewhat inefficient, since these regions have already been accounted for in the level 0 component of the resampling. This intuition extends naturally when considering multiple levels and is again captured by (3.1.4) in an exact sense.

From a practical point of view, the remaining problem then becomes how to select the sample sizes $(N_\ell)_{\ell=0}^L$ with respect to the solver hierarchy $(g_n^\ell)_{\ell=0}^L$ in a way that elicits optimal performance from the algorithm. While we have made no assumptions that $(N_\ell)_{\ell=0}^L$ is a decreasing sequence, in practice this is typically the optimal approach and is one that is most consistent with the ideology of MLMC methods. If implemented, the lower level particle allocations are the most likely to inherit a negative weight via the resampling according to (3.1.4), since they occupy a greater proportion of the total sample set. Given a subset of particles with negative weights this means that, for a fixed computational budget, the capacity for a negative weight to "spread" through the system is maximised if the maximum amount of computational budget is placed on level 0, since this provides the most number of trials for a particle with a negative weight to be resampled. However, such a perspective overlooks the fact that, for large N_0, the probability of obtaining a negative weight in the first place is decreased, due to the discussion in

Subsection 3.1.1. The ratio of negative weights to positive are of interest since, in practice, they are often symptomatic of poor performance of the algorithm. The signs are more than just a diagnostic tool however; they asymptotically correct the bias from the previous level estimate and are an essential component of the algorithm.

Using Algorithm 3, an approximation of $\widehat{\pi}_n$ can be formulated via

$$\widehat{\pi}_n^N = \frac{\sum_{i=1}^{S(N)} \widehat{w}_n^i \delta_{\xi_n^i}}{\sum_{i=1}^{S(N)} \widehat{w}_n^i}.$$

We note this is the empirical estimator *post* resampling which, due to the added sampling error, is in general less accurate than the one prior to resampling. It therefore follows that the subsequent convergence theorems hold for the pre-resampled estimator. Furthermore, the results also hold for the prediction filter

$$\pi_n^N = \frac{\sum_{i=1}^{S(N)} w_n^i \delta_{\xi_n^i}}{\sum_{i=1}^{S(N)} w_n^i}$$

as a trivial by-product of our analysis.

Before stating and proving the strong law of large numbers and central limit convergence theorems, we make the following mild assumptions about the Markov kernel and the likelihood functions and assume they hold throughout.

Assumption 1. *For all $x \in \mathbb{X}$, the measure $K(x, \cdot)$ admits a strictly positive density with respect to a σ-finite measure on $\mathcal{X}$.*

Assumption 2. *For all $n \in \mathbb{N}$, $g_n > 0$.*

3.2 Strong law of large numbers

Theorem 3. *For all bounded and measurable $\varphi : \mathbb{X} \mapsto \mathbb{R}$ and all $n \geq 0$,*

$$\widehat{\pi}_n^N(\varphi) \xrightarrow[N\to\infty]{\text{a.s.}} \widehat{\pi}_n(\varphi).$$

The proof of Theorem 3 proceeds as follows. In Proposition 1 we prove the almost-sure convergence for all $\varphi \in \mathscr{B}(\mathbb{X})$ of the level-specific unnormalised

prediction and total-variation prediction measures given by

$$\gamma_{n,\ell}^N(\varphi) := \sum_{i\in P_\ell^N} w_n^i \varphi(\xi_n^i), \qquad |\gamma_{n,\ell}^N|(\varphi) = \sum_{i\in P_\ell^N} \varphi(\xi_n^i) \tag{3.2.1}$$

to some integral forms $\gamma_n(\varphi)$ and $\eta_n(\varphi)$ respectively that each satisfy a recursive formula. This is proved by induction, an application of Lemma 3 (which we later prove) and an application of Lemma 1, which is assumed true by the induction statement. Along with the level-specific unnormalised filter measure

$$\widehat{\gamma}_{n,\ell}^N(\varphi) := \sum_{i\in P_\ell^N} \widehat{w}_n^i \varphi(\widehat{\xi}_n^i),$$

the empirical measures (3.2.1) are the building blocks of our analysis and comprise the normalised level-specific filter and prediction measures in the following sense

$$\widehat{\pi}_{n,\ell}^N(\varphi) = \frac{\widehat{\gamma}_{n,\ell}^N(\varphi)}{\widehat{\gamma}_{n,\ell}^N(1)}, \qquad \pi_{n,\ell}^N(\varphi) = \frac{\gamma_{n,\ell}^N(\varphi)}{\gamma_{n,\ell}^N(1)}.$$

Once Proposition 1 is established, it is then invoked in Lemma 1 to prove that the almost-sure convergence of $\gamma_{n,\ell}^N$ and $|\gamma_{n,\ell}^N|$ implies the almost sure convergence of $\widehat{\gamma}_{n,\ell}^N$ and $|\widehat{\gamma}_{n,\ell}^N|$ to unnormalised measures also specified in Proposition 1. Lemma 1 is then used to prove the following key result of Lemma 2

$$\widehat{\pi}_{n,\ell}^N(\varphi) \xrightarrow[N\to\infty]{\text{a.s.}} \widehat{\pi}_n(\varphi)$$

which says that the limit of each of the level-specific normalised empirical filter measures is in fact the *full*, level-free filter measure $\widehat{\pi}_n$ and not $\widehat{\pi}_{n,\ell}$. The reason for this somewhat unexpected result is due to the resampling step (3.1.4), which is performed with respect to the global particle set in which all levels are mixed. An equivalent explanation for this result is that, by taking absolute values in (3.2.2), we see that the total variation measure (3.1.4) is analogous to the multilevel decomposition of the filter measure (3.1.2)–(3.1.3).

Lastly in Lemma 3 we show that the established asymptotics are preserved in the mutation step, thus proving that convergence holds at each

stage of the iterative process. Theorem 3 then follows first by noting that

$$\begin{aligned}\widehat{\pi}_n^N(\varphi) = \frac{\sum_{i=1}^{S(N)} \widehat{w}_n^i \varphi(\widehat{\xi}_n^i)}{\sum_{i=1}^{S(N)} \widehat{w}_n^i} &= \sum_{\ell=0}^{L} \frac{\sum_{i\in P_\ell^N} \widehat{w}_n^i \varphi(\widehat{\xi}_n^i)}{\sum_{i=1}^{S(N)} \widehat{w}_n^i} \\ &= \sum_{\ell=0}^{L} \left(\frac{\sum_{i\in P_\ell^N} \widehat{w}_n^i \varphi(\widehat{\xi}_n^i)}{\sum_{i\in P_\ell^N} \widehat{w}_n^i} \cdot \frac{\sum_{i\in P_\ell^N} \widehat{w}_n^i}{\sum_{i=1}^{S(N)} \widehat{w}_n^i} \right) \\ &= \sum_{\ell=0}^{L} \widehat{\pi}_{n,\ell}^N(\varphi) \frac{\widehat{\gamma}_{n,\ell}^N(1)}{\sum_{\ell'=0}^{L} \widehat{\gamma}_{n,\ell'}^N(1)},\end{aligned} \tag{3.2.2}$$

to which Lemmas 1 and 2 can be applied to conclude the result.

Proposition 1. *For all $n \geq 0$, $\varphi \in \mathscr{B}(\mathbb{X})$ and $0 \leq \ell \leq L$*

$$\gamma_{n,\ell}^N(\varphi) \xrightarrow[N\to\infty]{\text{a.s.}} \gamma_n(\varphi) \quad \textit{and} \quad |\gamma_{n,\ell}^N|(\varphi) \xrightarrow[N\to\infty]{\text{a.s.}} \eta_n(\varphi), \tag{3.2.3}$$

where

$$\gamma_{n+1}(\varphi) = \pi_{n+1}(\varphi)\widehat{\gamma}_n(1) \quad \text{and} \quad \eta_{n+1}(\varphi) = \widehat{\eta}_n(K(\varphi)) \tag{3.2.4}$$

and

$$\widehat{\gamma}_n(\varphi) = \frac{\sum_{\ell=0}^{L} \gamma_n(\Delta g_n^\ell \varphi)}{\sum_{\ell=0}^{L} \eta_n(|\Delta g_n^\ell|)} \quad \textit{and} \quad \widehat{\eta}_n(\varphi) = \frac{\sum_{\ell=0}^{L} \eta_n(|\Delta g_n^\ell|\varphi)}{\sum_{\ell=0}^{L} \eta_n(|\Delta g_n^\ell|)}. \tag{3.2.5}$$

The statement of Proposition 1 has introduced two measure sequences $(\eta_n)_{n\geq 0}$ and $(\widehat{\eta}_n)_{n\geq 0}$ that are non-standard in SMC literature. These measures have no real world interpretation and instead are theoretical constructs that are necessary for establishing the main convergence result. They are what arise as the result of replacing g_n by the total variation of the telescoping sum $\sum_{\ell=0}^{L} |\Delta g_n^\ell|$ in the prediction and filter measures. In particular, we see in (3.2.3) that η_n is the weak limit of the unnormalised total variation prediction measure $|\gamma_{n,\ell}^N|$, which in turn is expressed in terms of $\widehat{\eta}_n$.

We also have the following corollary of Proposition 1 which, in addition to being of interest due to its insight into the relationship between η_n and γ_n, is required in our proof of the central limit theorem.

Corollary 1. *For all $n > 0$, $\eta_n - \gamma_n$ and $\widehat{\eta}_n - \widehat{\gamma}_n$ are positive measures.*

The two following lemmas each assume the convergence results of the prediction measures (3.2.3) hold in order to prove convergence of the respective

unnormalised filter measures. We note that due to the resampling according to the total variation measure (3.1.4), the convergence result (3.2.7) must also be established, since in general weak convergence of a measure does not imply the weak convergence of its total variation measure (see, for example, [9, Corollary 8.4.8]).

Lemma 1. *If Proposition 1 holds for some $n \geq 0$ then*

$$\widehat{\gamma}^N_{n,\ell}(\varphi) \xrightarrow[N\to\infty]{\text{a.s.}} \widehat{\gamma}_n(\varphi) \tag{3.2.6}$$

$$|\widehat{\gamma}^N_{n,\ell}|(\varphi) \xrightarrow[N\to\infty]{\text{a.s.}} \widehat{\eta}_n(\varphi) \tag{3.2.7}$$

for all $\varphi \in \mathscr{B}(\mathbb{X})$ and $0 \leq \ell \leq L$.

Lemma 2. *If Proposition 1 holds for some $n \geq 0$ then*

$$\widehat{\pi}^N_{n,\ell}(\varphi) - \widehat{\pi}_n(\varphi) \xrightarrow[N\to\infty]{\text{a.s.}} 0 \tag{3.2.8}$$

for all $\varphi \in \mathscr{B}(\mathbb{X})$ and $0 \leq \ell \leq L$.

Once Lemmas 1 and 2 are proved, the remaining result required to complete the proof of Theorem 3 is the following Lemma 3. Due to the mutation step in the MLBPF being unchanged from that used in the BPF, Lemma 3 does not require Proposition 1 as an assumption, since the asymptotics of the result are not affected by the asymptotics of the resampled particles $(\widehat{\xi}^i_n)_{i=1}^{S(N)}$. In light of the previous results, this then proves that convergence holds at every step of the multilevel filtering process and hence completes the recursion.

Lemma 3. *For all $n \geq 0$, $\varphi \in \mathscr{B}(\mathbb{X})$ and $0 \leq \ell \leq L$*

$$\frac{1}{c_\ell N} \sum_{i \in P^N_\ell} \operatorname{sgn}(\widetilde{w}_n(\widehat{\xi}^i_n))(\varphi(\xi^i_{n+1}) - K(\varphi)(\widehat{\xi}^i_n)) \xrightarrow[N\to\infty]{\text{a.s.}} 0 \tag{3.2.9}$$

$$\frac{1}{c_\ell N} \sum_{i \in P^N_\ell} (\varphi(\xi^i_{n+1}) - K(\varphi)(\widehat{\xi}^i_n)) \xrightarrow[N\to\infty]{\text{a.s.}} 0. \tag{3.2.10}$$

Proof of Proposition 1. The proof proceeds by induction. For $n = 0$, the result for $\gamma^N_{n,\ell}$ and $|\gamma^N_{n,\ell}|$ holds by repeating the proof of (3.2.10) in Lemma 3 with ξ^i_{n+1} replaced by ξ^i_0 and $K(\varphi)(\widehat{\xi}^i_n)$ with $\pi_0(\varphi)$.

For the induction step, we assume the results hold for some fixed $n \geq 0$. By definition we have

$$\gamma^N_{n+1,\ell}(\varphi) = \frac{1}{c_\ell N} \sum_{i \in P^N_\ell} \operatorname{sgn}(\widetilde{w}_n(\widehat{\xi}^i_n))\varphi(\xi^i_{n+1}).$$

By (3.2.9) in Lemma 3 it therefore suffices to show that

$$\frac{1}{c_\ell N} \sum_{i \in P^N_\ell} \operatorname{sgn}(\widetilde{w}_n(\widehat{\xi}^i_n))K(\varphi)(\widehat{\xi}^i_n) \xrightarrow[N\to\infty]{\text{a.s.}} \gamma_{n+1}(\varphi) = \pi_{n+1}(\varphi)\widehat{\gamma}_n(1).$$

But by the definition of $\widehat{\gamma}^N_{n,\ell}$ and the mutation formula $\pi_{n+1}(\varphi) = \widehat{\pi}_n(K(\varphi))$, this is equivalent to showing

$$\widehat{\gamma}^N_{n,\ell}(K(\varphi)) \xrightarrow[N\to\infty]{\text{a.s.}} \widehat{\pi}_n(K(\varphi))\widehat{\gamma}_n(1),$$

which is ensured by (3.2.6) in Lemma 1 (which holds by the induction assumption) and the fact that $\widehat{\gamma}_n(\varphi) = \widehat{\pi}_n(\varphi)\widehat{\gamma}_n(1)$. Similarly for $|\gamma^N_{n,\ell}|$, applying the definition for $n+1$ and utilising (3.2.9) in Lemma 3, it suffices to show

$$\frac{1}{c_\ell N} \sum_{i \in P^N_\ell} K(\varphi)(\widehat{\xi}^i_n) \xrightarrow[N\to\infty]{\text{a.s.}} \eta_{n+1}(\varphi),$$

i.e.

$$|\widehat{\gamma}^N_{n,\ell}|(K(\varphi)) = \widehat{\eta}_n(K(\varphi)),$$

which follows from (3.2.7) in Lemma 1. □

Proof of Corollary 1. By (3.2.5), the linearity of γ_n and (3.2.4), we have the following recursion in $\widehat{\gamma}_n(1)$

$$\widehat{\gamma}_n(1) = \frac{\sum_{\ell=0}^L \gamma_n(\Delta g^\ell_n)}{\sum_{\ell=0}^L \eta_n(|\Delta g^\ell_n|)} = \frac{\gamma_n(g_n)}{\sum_{\ell=0}^L \eta_n(|\Delta g^\ell_n|)} = \frac{\pi_n(g_n)}{\sum_{\ell=0}^L \eta_n(|\Delta g^\ell_n|)}\widehat{\gamma}_{n-1}(1).$$

Hence

$$\widehat{\gamma}_n(1) = \prod_{q=0}^n \left(\frac{\pi_q(g_q)}{\sum_{\ell_q=0}^L \eta_q(|\Delta g^{\ell_q}_q|)} \right). \tag{3.2.11}$$

By (2.1.14), the predictive distribution π_{n+1} can also be decomposed into

the prior π_0 as

$$\pi_{n+1}(\varphi) = \frac{\pi_n(g_n K(\varphi))}{\pi_n(g_n)} = \frac{\pi_0\Big(K^n\Big(\prod_{q=0}^n g_q K(\varphi)\Big)\Big)}{\prod_{q=0}^n \pi_q(g_q)},$$

where K^n is the n-fold integral operator we encountered in the Feynman-Kac discussion in Section 2.1. Therefore, by (3.2.4) and the result (3.2.11),

$$\begin{aligned}
\gamma_{n+1}(\varphi) &= \pi_{n+1}(\varphi)\widehat{\gamma}_n(1) && (3.2.12)\\
&= \frac{\pi_0\Big(K^n\Big(\prod_{q=0}^n g_q K(\varphi)\Big)\Big)}{\sum_{\ell_q=0}^L \eta_q(|\Delta g_q^{\ell_q}|)} \\
&= \frac{\pi_0\Big(K^n\Big(\sum_{\ell_0}^L \cdots \sum_{\ell_q}^L \prod_{q=0}^n \Delta g_q^{\ell_q} K(\varphi)\Big)\Big)}{\sum_{\ell_q=0}^L \eta_q(|\Delta g_q^{\ell_q}|)}, && (3.2.13)
\end{aligned}$$

where in the last equality we have used $g_q = \sum_{\ell=0}^L \Delta g_q^\ell$, and the interchange of summation and multiplication is justified by the fact that $\sum_{\ell=0}^{L-1} \Delta g_q^\ell = 0$.

A similar decomposition for η_{n+1} can be obtained by repeatedly applying (3.2.4) in the following manner

$$\begin{aligned}
\eta_{n+1}(\varphi) = \widehat{\eta}_n(K(\varphi)) &\propto \sum_{\ell_n=0}^L \eta_n(|\Delta g_n^{\ell_n}| K(\varphi)) \\
&\propto \sum_{\ell_n=0}^L \sum_{\ell_{n-1}=0}^L \eta_{n-1}\Big(K^n\Big(|\Delta g_{n-1}^{\ell_{n-1}}||\Delta g_n^{\ell_n}| K(\varphi)\Big)\Big) \\
&= \eta_{n-1}\Big(K^n\Big(\sum_{\ell_{n-1}=0}^L \sum_{\ell_n=0}^L |\Delta g_{n-1}^{\ell_{n-1}}||\Delta g_n^{\ell_n}| K(\varphi)\Big)\Big)
\end{aligned}$$

and hence

$$\eta_{n+1}(\varphi) = \frac{\pi_0\Big(K^n\Big(\sum_{\ell_0=0}^L \cdots \sum_{\ell_n=0}^L \prod_{q=0}^n |\Delta g_q^{\ell_q}| K(\varphi)\Big)\Big)}{\prod_{q=0}^n \sum_{\ell_q=0}^L \eta_q(|\Delta g_q^{\ell_q}|)}. \qquad (3.2.14)$$

Subtracting (3.2.13) from (3.2.14), to determine the positivity of $\eta_n - \gamma_n$ for $n > 0$ it suffices to consider the sign of

$$\pi_0\Big(K^n\Big(\sum_{\ell_0=0}^L \cdots \sum_{\ell_n=0}^L \prod_{q=0}^n \Big(|\Delta g_q^{\ell_q}| - \Delta g_q^{\ell_q}\Big) K(\varphi)\Big)\Big).$$

However, by the definition of modulus and by Assumption 1, this measure is positive. For $\widehat{\eta}_n - \widehat{\gamma}_n$ we note by (3.2.5) that these measures are simply a sum of $\eta_n(|\Delta g_n^\ell|) - \gamma_n(\Delta g_n^\ell)$ and that, for a positive $\varphi \in \mathscr{B}(\mathbb{X})$,

$$\eta_n(|\Delta g_n^\ell|\varphi) - \gamma_n(\Delta g_n^\ell \varphi) \geq \eta_n(\Delta g_n^\ell \varphi) - \gamma_n(\Delta g_n^\ell \varphi) \geq 0.$$

In both cases we note that the measures are either strictly positive or identically zero if and only if the measures are equal. □

Proof of Lemma 1. The proof proceeds as follows. Using the Burkholder-Davis-Gundy theorem for convex functions of martingales [10], we show

$$\frac{1}{c_\ell N} \sum_{i \in P_\ell^N} \varphi(\widehat{\xi}_n^i) - \frac{\sum_{j=1}^{S(N)} |\widetilde{w}_n^j| \varphi(\xi_n^i)}{\sum_{j=1}^{S(N)} |\widetilde{w}_n^j|} \xrightarrow[N\to\infty]{\text{a.s.}} 0. \tag{3.2.15}$$

for $|\widehat{\gamma}_{n,\ell}^N|(\varphi)$ and derive a similar result for $\widehat{\gamma}_{n,\ell}^N$ by making appropriate alterations. To show (3.2.15) we construct a triangular martingale array $(\widehat{U}_{\rho,\ell}^N, \mathcal{G}_\rho^N)_{0\leq\rho\leq N, 0\leq\ell\leq L}$, i.e. a collection of pairs satisfying the following criteria:

- $(\mathcal{G}_\rho^N)_{0\leq\rho\leq N}$ is a non-decreasing sequence of σ-algebras.
- $\widehat{U}_{\rho,\ell}^N$ is $\mathcal{G}_\rho^N$-measurable for all ρ and ℓ.
- $\mathbb{E}[\widehat{U}_{0,\ell}^N] = 0$ and $\mathbb{E}[\widehat{U}_{\rho,\ell} \mid \mathcal{G}_{\rho-1}^N] = 0$ almost surely for all ℓ and all $\rho > 0$.

This then furnishes a bound that proves the convergence result (3.2.15) by an application of Markov's inequality and the Borel-Cantelli lemma. Finally, to prove the main convergence results for $|\widehat{\gamma}_{n,\ell}^N|$ and $\widehat{\gamma}_{n,\ell}^N$, we use the following results implied by Proposition 1:

$$\sum_{i=1}^{S(N)} |\widetilde{w}_n^i| \varphi(\xi_n^i) = \sum_{\ell=0}^{L} |\gamma_{n,\ell}^N|(|\Delta g_n^\ell|\varphi) \xrightarrow[N\to\infty]{\text{a.s.}} \sum_{\ell=0}^{L} \eta_n(|\Delta g_n^\ell|\varphi) \tag{3.2.16}$$

$$\sum_{i=1}^{S(N)} \widetilde{w}_n^i \varphi(\xi_n^i) = \sum_{\ell=0}^{L} \gamma_{n,\ell}^N(\Delta g_n^\ell \varphi) \xrightarrow[N\to\infty]{\text{a.s.}} \sum_{\ell=0}^{L} \gamma_n(\Delta g_n^\ell \varphi). \tag{3.2.17}$$

The results for $|\widehat{\gamma}_{n,\ell}^N|$, $\widehat{\gamma}_{n,\ell}^N$ will then follow by combining each of the established convergence results with the definitions of $\widehat{\eta}_n$ and $\widehat{\gamma}_n$ respectively.

For $0 \leq \ell \leq L$ set $\widehat{U}_{0,\ell} = 0$ and for $1 \leq \rho \leq N$ define

$$\widehat{U}^N_{\rho,\ell} := \frac{1}{c_\ell \sqrt{N}} \sum_{i \in \mathcal{I}^N_\ell(\rho)} \left(\varphi(\widehat{\xi}^i_n) - \frac{\sum_{j=1}^{S(N)} |\widetilde{w}^j_n| \varphi(\xi^j_n)}{\sum_{j=1}^{S(N)} |\widetilde{w}^j_n|} \right)$$

where $\mathcal{I}^N_\ell(\rho)$ denotes the level-specific index subset

$$\mathcal{I}^N_\ell(\rho) := \{I_\ell(N) + (\rho - 1)c_\ell + 1, \ldots, I_\ell(N) + \rho c_\ell\}.$$

We note that $\{\mathcal{I}^N_\ell(\rho)\}_{\rho=1}^N$ partitions $\mathcal{P}^N_\ell$ into N subsets of size $c_\ell \geq 1$ and that $\{|\mathcal{I}^N_\ell(\rho)|\}_{\ell=0}^L$ is a decreasing sequence. Define the σ-algebras $\mathcal{G}^N_\rho$ by

$$\mathcal{G}^N_0 = \mathcal{F}^N_n, \qquad \mathcal{G}^N_\rho = \mathcal{G}^N_{\rho-1} \vee \bigvee_{0 \leq \ell \leq L} \bigvee_{i \in \mathcal{I}^N_\ell(\rho)} \sigma(\widehat{\xi}^i_n), \quad 1 \leq \rho \leq N,$$

where $\mathcal{F}^N_n$ is the σ-algebra generated by ξ^i_q and $\widehat{\xi}^i_p$ for $0 \leq q, p \leq n$ and $1 \leq i \leq S(N)$. Hence for $\rho > 0$, $\mathcal{G}^N_\rho$ is the σ-algebra generated by all of the historical killed and resampled particles and the resampled particles across all levels of the current iterate ρ-th level-specific index subsets. By construction, $(\mathcal{G}^N_\rho)$ is a non-decreasing sequence of σ-algebras and $\widehat{U}^N_{\rho,\ell}$ is $\mathcal{G}^N_\rho$-measurable. Clearly, $\mathbb{E}[\widehat{U}^N_{0,\ell}] = 0$, while for $\rho > 0$ the fact that $\mathbb{E}[\widehat{U}_{\rho,\ell} \mid \mathcal{G}_{\rho-1}] = 0$ is a straightforward consequence of the fact that the resampled particles $\widehat{\xi}^i_n$ are distributed according to the normalised total variation measure as in Algorithm 3, and hence

$$\mathbb{E}[\varphi(\widehat{\xi}^i_n) \mid \mathcal{G}^N_{\rho-1}] = \frac{\sum_{j=1}^{S(N)} |\widetilde{w}^j_n| \varphi(\xi^j_n)}{\sum_{j=1}^{S(N)} |\widetilde{w}^j_n|}. \tag{3.2.18}$$

Noting that $|\widehat{U}_{\rho,\ell}| \leq 2\|\varphi\|_\infty / \sqrt{N}$, the Burkholder-Davis-Gundy theorem can therefore be applied for all $1 \leq r \leq \infty$ to deduce

$$\begin{aligned} \mathbb{E}\left[\left| \frac{1}{c_\ell N} \sum_{i \in \mathcal{P}^N_\ell} \varphi(\widehat{\xi}^i_n) - \frac{\sum_{j=1}^{S(N)} |\widetilde{w}^j_n| \varphi(\xi^j_n)}{\sum_{j=1}^{S(N)} |\widetilde{w}^j_n|} \right|^r \middle| \mathcal{G}^N_0 \right] &= \frac{1}{N^{r/2}} \mathbb{E}\left[\left| \sum_{\rho=1}^N \widehat{U}^N_{\rho,\ell} \right|^r \middle| \mathcal{G}^N_0 \right] \\ &\leq \frac{B_r \|\varphi\|_\infty}{N^{r/2}}, \end{aligned}$$

where B_r is some constant that depends only on r. By Markov's inequality and the Borel-Cantelli lemma this implies (3.2.15), and by (3.2.16) and

(3.2.5) we have

$$|\widehat{\gamma}^N_{n,\ell}|(\varphi) \xrightarrow[N\to\infty]{\text{a.s.}} \widehat{\eta}_n(\varphi),$$

which concludes the section of the proof for $|\widehat{\gamma}^N_{n,\ell}|$.

For $\widehat{\gamma}^N_{n,\ell}$ we instead construct

$$\breve{U}^N_{\rho,\ell} := \frac{1}{c_\ell\sqrt{N}} \sum_{i\in\mathcal{I}^N_\ell(\rho)} \left(\operatorname{sgn}(\widetilde{w}_n(\widehat{\xi}^i_n))\varphi(\widehat{\xi}^i_n) - \frac{\sum_{j=1}^{S(N)} \widetilde{w}^j_n\varphi(\xi^j_n)}{\sum_{j=1}^{S(N)} |\widetilde{w}^j_n|}\right),$$

which is simply $\widehat{U}^N_{\rho,\ell}$ with $\widetilde{w}_n(\widehat{\xi}^i_n))\varphi(\widehat{\xi}^i_n)$ in place of $\varphi(\widehat{\xi}^i_n)$ and the identity $\widetilde{w}^j_n = |\widetilde{w}^j_n|\operatorname{sgn}(\widetilde{w}^j_n)$ applied. Using the the same methodology this implies the analogous result of (3.2.15) in light of the altered φ. The result

$$\widehat{\gamma}_{n,\ell}(\varphi) \xrightarrow[N\to\infty]{\text{a.s.}} \widehat{\gamma}_n(\varphi)$$

then follows from applying (3.2.16), (3.2.17) and finally the definition (3.2.5).

□

Proof of Lemma 2. Let $(\mathcal{G}^N_\rho)_{0\le\rho\le N}$ and $\{\mathcal{I}^N_\ell(\rho)\}^N_{\rho=1}$ be the same collections of σ-algebras and level-specific index subsets respectively as constructed in Lemma 1. For $0 \le \ell \le L$ define $U^N_{0,\ell} = 0$ and for $1 \le \rho \le N$ define

$$U^N_{\rho,\ell} = \frac{1}{c_\ell\sqrt{N}} \sum_{i\in\mathcal{I}^N_\ell(\rho)} \operatorname{sgn}(\widetilde{w}_n(\widehat{\xi}^i_n))\left(\varphi(\widehat{\xi}^i_n) - \frac{\sum_{j=1}^{S(N)} \widetilde{w}^j_n\varphi(\xi^j_n)}{\sum_{j=1}^{S(N)} \widetilde{w}^j_n}\right) \tag{3.2.19}$$

for some $\varphi \in \mathscr{B}(\mathbb{X})$. By construction, each $U^N_{\rho,\ell}$ is $\mathcal{G}^N_\rho$-measurable and $\mathbb{E}[U^N_{0,\ell}] = 0$. To apply the Burkholder-Davis-Gundy theorem to $U^N_{\rho,\ell}$ it remains to show $\mathbb{E}[U^N_{\rho,\ell} \mid \mathcal{G}^N_{\rho-1}] = 0$ for $\rho > 0$ and to bound $|U^N_{\rho,\ell}|$ from above. Applying the expectation result (3.2.18) to $\mathbb{E}[U^N_{\rho,\ell} \mid \mathcal{G}^N_{\rho-1}]$ gives

$$\begin{aligned} c_\ell\sqrt{N}\mathbb{E}[U^N_{\rho,\ell} \mid \mathcal{G}^N_{\rho-1}] &= \mathbb{E}\Bigg[\sum_{i\in\mathcal{I}^N_\ell(\rho)} \widehat{w}^i_n\left(\varphi(\widehat{\xi}^i_n) - \frac{\sum_{j=1}^{S(N)} \widetilde{w}^j_n\varphi(\xi^j_n)}{\sum_{j=1}^{S(N)} \widetilde{w}^j_n}\right)\Bigg|\mathcal{G}^N_{\rho-1}\Bigg] \\ &= \frac{c_\ell}{\sum_{i=1}^{S(N)} |\widetilde{w}^i_n|}\mathbb{E}\Bigg[\sum_{i=1}^{S(N)} \widetilde{w}^i_n\left(\varphi(\xi^i_n) - \frac{\sum_{j=1}^{S(N)} \widetilde{w}^j_n\varphi(\xi^j_n)}{\sum_{j=1}^{S(N)} \widetilde{w}^j_n}\right)\Bigg] \end{aligned}$$

where we have used $|\mathcal{I}_\ell^N(\rho)| = c_\ell$ and $|\widetilde{w}_n^i|\mathrm{sgn}(\widetilde{w}_n(\xi_n^i)) = \widetilde{w}_n^i$. Therefore

$$\mathbb{E}[U_{\rho,\ell}^N \mid \mathcal{G}_{\rho-1}^N] = \frac{1}{\sqrt{N}\sum_{i=1}^{S(N)}|\widetilde{w}_n^i|}\sum_{i=1}^{S(N)}\left(\widetilde{w}_n^i\varphi(\xi_n^i) - \widetilde{w}_n^i\frac{\sum_{j=1}^{S(N)}\widetilde{w}_n^j\varphi(\xi_n^j)}{\sum_{j=1}^{S(N)}\widetilde{w}_n^j}\right) = 0$$

almost surely, and $(U_{\rho,\ell}^N, \mathcal{G}_\rho^N)_{0\le\rho\le N, 0\le\ell\le L}$ is a triangular martingale difference array. To bound $|U_{\rho,\ell}^N|$ we apply (3.2.17) from the proof of Lemma 1 to deduce that

$$\frac{\sum_{j=1}^{S(N)}\widetilde{w}_n^j\varphi(\xi_n^j)}{\sum_{j=1}^{S(N)}\widetilde{w}_n^j} \xrightarrow[N\to\infty]{\text{a.s.}} \frac{\sum_{\ell=0}^{L}\gamma_n(\Delta g_n^\ell\varphi)}{\sum_{\ell=0}^{L}\gamma_n(\Delta g_n^\ell)} = \frac{\pi_n(\varphi g_n)}{\pi_n(g_n)}.$$

Therefore for any $\delta > 0$ there exists almost surely a positive integer N_δ such that for all $N > N_\delta$,

$$\left|\frac{\sum_{j=1}^{S(N)}\widetilde{w}_n^j\varphi(\xi_n^j)}{\sum_{j=1}^{S(N)}\widetilde{w}_n^j}\right| < \left|\frac{\pi_n(\varphi g_n)}{\pi_n(g_n)}\right| + \delta$$

and hence

$$|U_{\rho,\ell}^N| \le \frac{C_{\varphi,\delta}}{\sqrt{N}}, \qquad \text{where } C_{\varphi,\delta} = \|\varphi\|_\infty + \left|\frac{\pi_n(\varphi g_n)}{\pi_n(g_n)}\right| + \delta.$$

By the Burkholder-Davis-Gundy theorem,

$$\frac{1}{N^{r/2}}\mathbb{E}\left[\left|\sum_{\rho=1}^{N}U_{\rho,\ell}^N\right|^r \middle| \mathcal{G}_0^N\right] \le \frac{B_r C_{\varphi,\delta}^r}{N^{r/2}} \tag{3.2.20}$$

for $1 \le r \le \infty$, where B_r is some constant that depends only on r. Scaling and summing $U_{\rho,\ell}^N$ over ρ, (3.2.19) can be rearranged to get

$$\begin{aligned}
\frac{1}{\sqrt{N}}\sum_{\rho=0}^{N}U_{\rho,\ell}^N &= \frac{1}{c_\ell N}\sum_{i\in P_\ell^N}\mathrm{sgn}(\widetilde{w}_n(\widehat{\xi}_n^i))\left(\varphi(\widehat{\xi}_n^i) - \frac{\sum_{j=1}^{S(N)}\widetilde{w}_n^j\varphi(\xi_n^j)}{\sum_{j=1}^{S(N)}\widetilde{w}_n^j}\right)\\
&= \frac{1}{c_\ell N}\sum_{i\in P_\ell^N}\left(\widehat{w}_n^i\varphi(\widehat{\xi}_n^i) - \widehat{w}_n^i\frac{\sum_{j=1}^{S(N)}\widetilde{w}_n^j\varphi(\xi_n^j)}{\sum_{j=1}^{S(N)}\widetilde{w}_n^j}\right)\\
&= \left(\frac{1}{c_\ell N}\sum_{i\in P_\ell^N}\widehat{w}_n^i\right)\left(\frac{\sum_{i\in P_\ell^N}\widehat{w}_n^i\varphi(\widehat{\xi}_n^i)}{\sum_{i\in \mathcal{P}_\ell^N}\widehat{w}_n^i} - \frac{\sum_{j=1}^{S(N)}\widetilde{w}_n^j\varphi(\xi_n^j)}{\sum_{j=1}^{S(N)}\widetilde{w}_n^j}\right)
\end{aligned}$$

$$= \left(\frac{1}{c_\ell N}\sum_{i\in P_\ell^N}\widehat{w}_n^i\right)\left(\widehat{\pi}_{n,\ell}^N(\varphi) - \frac{\sum_{j=1}^{S(N)}\widetilde{w}_n^j\varphi(\xi_n^j)}{\sum_{j=1}^{S(N)}\widetilde{w}_n^j}\right)$$

and hence

$$\widehat{\pi}_{n,\ell}^N(\varphi) - \widehat{\pi}_n(\varphi) = \frac{\frac{1}{\sqrt{N}}\sum_{\rho=0}^N U_{\rho,\ell}^N}{\frac{1}{c_\ell N}\sum_{i\in P_\ell^N}\widehat{w}_n^i} + \frac{\sum_{j=1}^{S(N)}\widetilde{w}_n^j\varphi(\xi_n^j)}{\sum_{j=1}^{S(N)}\widetilde{w}_n^j} - \frac{\pi_n(\varphi g_n)}{\pi_n(g_n)}.$$

By the bound (3.2.20) we can invoke Markov's inequality to make the probability of $N^{1/2}\sum_\rho U_{\rho,\ell}^N$ arbitrarily small. Therefore by the Borel-Cantelli lemma it follows that

$$\frac{1}{\sqrt{N}}\sum_{\rho=0}^N U_{\rho,\ell}^N \xrightarrow[N\to\infty]{\text{a.s.}} 0,$$

while for the remaining terms we apply (3.2.17) to conclude that

$$\frac{\sum_{j=1}^{S(N)}\widetilde{w}_n^j\varphi(\xi_n^j)}{\sum_{j=1}^{S(N)}\widetilde{w}_n^j} - \frac{\pi_n(\varphi g_n)}{\pi_n(g_n)} \xrightarrow[N\to\infty]{\text{a.s.}} 0,$$

which implies the result. □

Proof of Lemma 3. With the level-specific index subsets $\{\mathcal{I}_\ell(\rho)\}_{\rho=1}^N$ defined as in the proof of Lemma 1, we consider the random variables

$$\widetilde{U}_{\rho,\ell}^N = \frac{1}{c_\ell\sqrt{N}}\sum_{i\in\mathcal{I}_\ell^N(\rho)}\operatorname{sgn}(\widetilde{w}_n(\widehat{\xi}_n^i))(\varphi(\xi_{n+1}^i) - K(\varphi)(\widehat{\xi}_n^i))$$

and $\widetilde{U}_{0,\ell}^N = 0$ for $0 \le \ell \le L$. Additionally, define the σ-algebras $(\widetilde{\mathcal{G}}_\rho^N)_{0\le\rho\le N, N>0}$ by

$$\widetilde{\mathcal{G}}_0^N = \widehat{\mathcal{F}}_n^N, \quad \widetilde{\mathcal{G}}_\rho^N = \widetilde{\mathcal{G}}_{\rho-1}^N \vee \bigvee_{0\le\ell\le L}\bigvee_{i\in\mathcal{I}_\ell^N(\rho)}\sigma(\xi_{n+1}^i),$$

where $\widehat{\mathcal{F}}_n^N$ is the σ-algebra generated by ξ_q^i and $\widehat{\xi}_q^i$ for $0 \le q \le n$ and $1 \le i \le S(N)$. Since $\xi_{n+1}^i \sim K(\widehat{\xi}_n^i, \cdot)$, the ξ_{n+1}^i are conditionally independent given $\widetilde{\mathcal{G}}_0^N$, and therefore $\mathbb{E}[\widetilde{U}_{\rho,\ell}^N \mid \widetilde{\mathcal{G}}_{\rho-1}^N] = 0$ almost surely for all $1 \le \rho \le N$. Moreover, $\widetilde{U}_{\rho,\ell}^N$ is $\widetilde{\mathcal{G}}_\rho^N$-measurable for all ρ, making $(\widetilde{U}_{\rho,\ell}^N, \widetilde{\mathcal{G}}_\rho^N)_{0\le\rho\le N, 0\le\ell\le L}$ a triangular martingale difference array. As in Lemma 1, the bound $|\widetilde{U}_{\rho,\ell}^N| \le$

$2\|\varphi\|_\infty/\sqrt{N}$ holds and thus by the Burkholder-Davis-Gundy theorem

$$\mathbb{E}\left[\left|\frac{1}{c_\ell N}\sum_{i\in P_\ell^N}\operatorname{sgn}(\widetilde{w}_n(\widehat{\xi}_n^i))(\varphi(\xi_{n+1}^i)-K(\varphi)(\widehat{\xi}_n^i))\right|^r\middle|\widetilde{\mathcal{G}}_0^N\right]\leq\frac{B_r\|\varphi\|^r}{N^{r/2}}$$

for all $1\leq r\leq\infty$ and some constant B_r depending only on r. The almost-sure convergence follows by applying Markov's inequality with the above expectation bound and the Borel-Cantelli lemma. The analogous result for (3.2.10) follows by repeating the above process but with $\operatorname{sgn}(\widetilde{w}_n(\widehat{\xi}_n^i))$ omitted, which are all equal to 1 in the case of the total variation measure. □

3.3 Central limit theorem

Theorem 4. *For all bounded and measurable $\varphi:\mathbb{X}\mapsto\mathbb{R}$ and all $n\geq 0$,*

$$\sqrt{N}\big(\widehat{\pi}_n^N(\varphi)-\widehat{\pi}_n(\varphi)\big)\xrightarrow[N\to\infty]{\mathrm{D}}\mathcal{N}\big(0,\widehat{\sigma}_n^2(\varphi)\big)$$

for some $\widehat{\sigma}_n^2(\varphi)\in(0,\infty)$.

A well-known corollary of Theorem 3 is that the measure-type convergence also holds in the distribution sense; however, specific knowledge of this measure is unknown. Central limit theorems such as Theorem 4 instead show that, given a $\varphi\in\mathscr{B}(\mathbb{X})$, that the distribution of the random variable $\sqrt{N}\big(\widehat{\pi}_n^N(\varphi)-\widehat{\pi}_n(\varphi)\big)$ is asymptotically normally distributed with a finite variance term depending only n and φ. Such knowledge is beneficial because it enables the well-known properties of the normal distribution to be invoked in order to make quantitative statements about the asymptotic behaviour of $\widehat{\pi}_n^N$.

Recall that the measures $\widehat{\pi}_n^N$ and π_n^N are dependent on the four unnormalised measures in the following set

$$\Gamma_{n,m}=\big\{\gamma_{p,\ell}^N,\,\widehat{\gamma}_{q,\ell}^N,\,|\gamma_{p,\ell}^N|,\,|\widehat{\gamma}_{q,\ell}^N|:\,0\leq\ell\leq L,\,0\leq p\leq n,\,0\leq q\leq m\big\},$$

where $n\in\mathbb{N}$, $m\in\{n-1,n\}$. That is, $\widehat{\pi}_n^N$ and π_n^N are constructed via some functional mapping on $\Gamma_{n,n}$ and $\Gamma_{n,n-1}$. To prove Theorem 4 it is sufficient to show that the measures in these sets are *jointly asymptotically normal*, from which the main result follows by the δ-method (see e.g. [54, Ch. 9]).

Before conducting the proof, we first state the definition of joint asymptotic normality in the context of the current application.

Let $L > 0$, $d = 2(L+1)(n+m+2)$, $\boldsymbol{t} = (t_1, \ldots, t_d)^T \in \mathbb{R}^d$ and $\boldsymbol{\varphi} \in \mathscr{B}(\mathbb{X})^d$. Furthermore, define

$$\begin{aligned}
&\Psi^N_{n,m}(\boldsymbol{t}, \boldsymbol{\varphi}) = \\
&\sum_{\ell=0}^{N} \Bigg[\sum_{p=0}^{n} \Big(t^{(1)}_{p,\ell} \Big[\gamma^N_{p,\ell}\big(\varphi^{(1)}_{p,\ell}\big) - \gamma_p\big(\varphi^{(1)}_{p,\ell}\big) \Big] + t^{(2)}_{p,\ell} \Big[|\gamma^N_{p,\ell}|\big(\varphi^{(2)}_{p,\ell}\big) - \eta_p\big(\varphi^{(2)}_{p,\ell}\big) \Big] \Big) \\
&\quad + \sum_{q=0}^{m} \Big(t^{(3)}_{q,\ell} \Big[\widehat{\gamma}^N_{q,\ell}\big(\varphi^{(3)}_{q,\ell}\big) - \widehat{\gamma}_q\big(\varphi^{(3)}_{q,\ell}\big) \Big] + t^{(4)}_{q,\ell} \Big[|\widehat{\gamma}^N_{q,\ell}|\big(\varphi^{(4)}_{q,\ell}\big) - \widehat{\eta}_q\big(\varphi^{(4)}_{q,\ell}\big) \Big] \Big) \Bigg],
\end{aligned} \tag{3.3.1}$$

where $t^{(k)}_{p,\ell} = t_{\beta(k,p,\ell)}$ and $\varphi^{(k)}_{p,\ell} = \varphi_{\beta(k,p,\ell)}$ for the mapping $\beta : (k, p, \ell) \mapsto (4p + (k-1))(L+1) + \ell + 1$ that converts the three-dimensional indexing over the type of measure $k \in \{1, 2, 3, 4\}$, where k denotes the index of each measure in the tuple $\big(\gamma^N_{p,\ell}, |\gamma^N_{p,\ell}|, \widehat{\gamma}^N_{p,\ell}, |\widehat{\gamma}^N_{p,\ell}|\big)$, the iterate p and the level ℓ respectively, into the set $\{1, \ldots, d\}$. As per [31], we say that the measures in $\Gamma_{n,m}$ satisfy the joint asymptotic normality if for all $\boldsymbol{t} \in \mathbb{R}^d$ and all $\boldsymbol{\varphi} \in \mathscr{B}(\mathbb{X})^d$

$$\sqrt{N}\Psi^N_{n,m}(\boldsymbol{t}, \boldsymbol{\varphi}) \xrightarrow[N\to\infty]{\mathrm{D}} \mathcal{N}\big(\mathbf{0}, \boldsymbol{t}^T \boldsymbol{\Gamma}_{n,m}(\boldsymbol{\varphi}) \boldsymbol{t}\big) \tag{3.3.2}$$

for some symmetric non-negative definite matrix $\boldsymbol{\Gamma}_{n,m}(\boldsymbol{\varphi}) \in \mathbb{R}^{d\times d}$. Then by the Cramér-Wold theorem (see e.g. [8]) this implies that the d individual differences in (3.3.1) are joint asymptotically normal, so that Theorem 4 holds by the δ-method.

The proof proceeds by induction by assuming that (3.3.2) holds at initalisation and that the convergence is preserved within the update and mutation steps. The methods differs from the existing literature predominantly because the triangular martingale difference arrays are based on the MLBPF-specific signed measures in $\Gamma_{n,m}$.

Before proceeding to the proof, we note that the joint asymptotic normality in (3.3.1) is considered both over all levels and all iterations. The consideration of the latter is superfluous for the purposes of proving Theorem 4, but in the process it allows results across multiple iterations to be proved, such as the following central limit theorem for the normalisation

terms across multiple iterations. In particular we note the absence of any dependence on φ in the standard deviation of the limiting measure.

Theorem 5. *There exists $\sigma_Z^2 > 0$ such that*

$$\sqrt{N}\left(\prod_{p=0}^{n-1}\pi_p^N(g_p) - \mathbb{E}\left[\prod_{p=0}^{n-1} g_p(X_p)\right]\right) \xrightarrow[N\to\infty]{\mathrm{D}} \mathcal{N}\big(0,\widehat{\sigma}_Z^2\big).$$

In the proof of Theorem 4 we will require the following well-known auxiliary result, the short proof of which is provided in [31], for example.

Lemma 4. *Let $(A_N)_{N>0}$ and $(B_N)_{N>0}$ be sequences of $\mathbb{X}$-valued random variables, such that for all $N \in \mathbb{N}$, B_N is $\mathcal{G}_N$-measurable,*

$$\sqrt{N}B_N \xrightarrow[N\to\infty]{\mathrm{D}} B \sim \mathcal{N}\Big(0,\sigma_B^2\Big),$$

and

$$\mathbb{E}\Big[\exp\Big(\mathrm{i}u\sqrt{N}A_N\Big)\Big|\mathcal{G}_N\Big] \xrightarrow[N\to\infty]{\mathbb{P}} \exp\Big(-\frac{u^2}{2}\sigma_A^2\Big).$$

Then

$$\sqrt{N}(A_N + B_N) \xrightarrow[N\to\infty]{\mathrm{D}} \mathcal{N}\Big(0,\sigma_A^2+\sigma_B^2\Big).$$

Proof of Theorem 4. The proof is by induction and we start with the update step. This means that we assume the joint asymptotic normality property (3.3.2) holds for $n \in \mathbb{N}$ and $m = n-1$ and that we consider the following augmented coefficient and test function vectors associated to the updated measures over all levels

$$\widehat{\boldsymbol{t}}_n = (t_{n,0}^{(3)},\ldots,t_{n,L}^{(3)},t_{n,0}^{(4)},\ldots,t_{n,L}^{(4)})^T \in \mathbb{R}^{2(L+1)}$$
$$\widehat{\boldsymbol{\varphi}}_n = (\varphi_{n,0}^{(3)},\ldots,\varphi_{n,L}^{(3)},\varphi_{n,0}^{(4)},\ldots,\varphi_{n,L}^{(4)})^T \in \mathscr{B}(\mathbb{X})^{2(L+1)}.$$

We also construct the triangular martingale difference $(\widehat{U}_\rho^N,\mathcal{G}_\rho^N)_{0\le\rho\le N,N>0}$, where

$$\widehat{U}_\rho^N = \sum_{\ell=0}^{L}(t_{n,\ell}^{(3)}\widehat{U}_{\rho,\ell}^N(\bar{\varphi}_{n,\ell}^{(3)}) + t_{n,\ell}^{(4)}\widehat{U}_{\rho,\ell}^N(\varphi_{n,\ell}^{(4)})),$$

where we have defined

$$\bar{\varphi}_{n,\ell}^{(3)}(\xi_n^i) = \mathrm{sgn}(\widetilde{w}_n(\xi_n^i))\varphi_{n,\ell}^{(3)}(\xi_n^i), \quad 1\le i\le S(N)$$

and where $\widehat{U}^N_{\rho,\ell}$ is defined as in Lemma 1, except we now include the dependency on the respective test function explicitly in the notation. By the proof of Lemma 1 we have the following bound

$$\left|\widehat{U}^N_\rho\right| \leq \frac{C_{\bar{\varphi}}}{\sqrt{N}}, \quad \text{where } C_{\bar{\varphi}} = \sum_{\ell=0}^{L}\left(t^{(3)}_{n,\ell}\|\bar{\varphi}^{(3)}_{n,\ell}\| + t^{(4)}_{n,\ell}\|\varphi^{(4)}_{n,\ell}\|\right),$$

and hence by basic integral inequalities and probability-expectation relations,

$$\begin{aligned}\sum_{\rho=1}^{N}\mathbb{E}\left[\left(\widehat{U}^N_\rho\right)^2\mathbb{I}\left[\left|\widehat{U}^N_\rho\right| \geq \epsilon\right]\Big|\mathcal{G}^N_{\rho-1}\right] &\leq \frac{C^2_{\bar{\varphi}}}{N}\sum_{\rho=1}^{N}\mathbb{P}\left[\left|\widehat{U}^N_{\rho,\ell}\right| \geq \epsilon\Big|\mathcal{G}^N_{\rho-1}\right] \\ &\leq \frac{C^2_{\bar{\varphi}}}{N}\sum_{\rho=1}^{N}\mathbb{I}\left[\frac{C_{\bar{\varphi}}}{\sqrt{N}} \geq \epsilon\right] \xrightarrow[N\to\infty]{\text{a.s.}} 0. \quad (3.3.3)\end{aligned}$$

Since the levels are conditionally independent given $\mathcal{G}^N_{\rho-1}$ and for all $0 \leq \ell \leq L$ we have $\mathbb{E}[\widehat{U}^N_{\rho,\ell}(\bar{\varphi}^{(3)}_{n,\ell}) \mid \mathcal{G}^N_{\rho-1}] = \mathbb{E}[\widehat{U}^N_{\rho,\ell}(\varphi^{(4)}_{n,\ell}) \mid \mathcal{G}^N_{\rho-1}] = 0$, then the second moments satisfy

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{U}^N_\rho\right)^2\Big|\mathcal{G}^N_{\rho-1}\right] = \sum_{\ell=0}^{L}\Bigg\{&\left(t^{(3)}_{n,\ell}\right)^2\mathbb{E}\left[\left(\widehat{U}_{\rho,\ell}(\bar{\varphi}^{(3)}_{n,\ell})\right)^2\Big|\mathcal{G}^N_{\rho-1}\right] \\ &+ \left(t^{(4)}_{n,\ell}\right)^2\mathbb{E}\left[\left(\widehat{U}_{\rho,\ell}(\varphi^{(4)}_{n,\ell})\right)^2\Big|\mathcal{G}^N_{\rho-1}\right]\Bigg\} \\ &+ 2\sum_{\ell=0}^{L}\sum_{\ell'=\ell+1}^{L} t^{(3)}_{n,\ell}t^{(4)}_{n,\ell'}\mathbb{E}\left[\widehat{U}^N_{\rho,\ell}(\bar{\varphi}^{(3)}_{n,\ell})\widehat{U}^N_{\rho,\ell}(\varphi^{(4)}_{n,\ell})\Big|\mathcal{G}^N_{\rho-1}\right].\end{aligned}$$

By (3.2.16) and (3.2.17) in Lemma 1 each of the terms converge to the following limits

$$\begin{aligned}\sum_{\rho=1}^{N}\mathbb{E}\left[\left(\widehat{U}^N_{\rho,\ell}(\bar{\varphi}^{(3)}_{n,\ell})\right)^2\Big|\mathcal{G}^N_{\rho-1}\right] &\xrightarrow[N\to\infty]{\text{a.s.}} \frac{1}{c_\ell}\left(\widehat{\eta}_n((\varphi^{(3)}_{n,\ell})^2) - \widehat{\gamma}_n(\varphi^{(3)}_{n,\ell})^2\right) \\ \sum_{\rho=1}^{N}\mathbb{E}\left[\left(\widehat{U}^N_{\rho,\ell}(\varphi^{(4)}_{n,\ell})\right)^2\Big|\mathcal{G}^N_{\rho-1}\right] &\xrightarrow[N\to\infty]{\text{a.s.}} \frac{1}{c_\ell}\left(\widehat{\eta}_n((\varphi^{(4)}_{n,\ell})^2) - \widehat{\eta}_n(\varphi^{(4)}_{n,\ell})^2\right) \\ \sum_{\rho=1}^{N}\mathbb{E}\left[\widehat{U}^N_{\rho,\ell}(\bar{\varphi}^{(3)}_{n,\ell})\widehat{U}^N_{\rho,\ell}(\varphi^{(4)}_{n,\ell})\Big|\mathcal{G}^N_{\rho-1}\right] &\xrightarrow[N\to\infty]{\text{a.s.}} \frac{1}{c_\ell}\left(\widehat{\gamma}_n(\varphi^{(3)}_{n,\ell}\varphi^{(4)}_{n,\ell}) - \widehat{\gamma}_n(\varphi^{(3)}_{n,\ell})\widehat{\eta}_n(\varphi^{(4)}_{n,\ell})\right),\end{aligned}$$

from which we deduce that

$$\sum_{\rho=1}^{N}\mathbb{E}\Big[\Big(\widehat{U}_{\rho}^{N}\Big)^{2}\Big|\mathcal{G}_{\rho-1}^{N}\Big]\xrightarrow[N\to\infty]{\text{a.s.}}\widehat{\boldsymbol{t}}_{n}^{T}\boldsymbol{\Gamma}_{n}'(\widehat{\boldsymbol{\varphi}}_{n})\widehat{\boldsymbol{t}}_{n},\quad\text{where }\boldsymbol{\Gamma}_{n}'(\widehat{\boldsymbol{\varphi}}_{n})=\begin{pmatrix}\mathbf{A}&\mathbf{B}\\\mathbf{B}&\mathbf{C}\end{pmatrix},\tag{3.3.4}$$

where

$$\begin{aligned}\mathbf{A}&=\operatorname{diag}_{0\le\ell\le L}\frac{1}{c_{\ell}}\Big(\widehat{\eta}_{n}((\varphi_{n,\ell}^{(3)})^{2})-\widehat{\gamma}_{n}(\varphi_{n,\ell}^{(3)})^{2}\Big)\\\mathbf{B}&=\operatorname{diag}_{0\le\ell\le L}\frac{1}{c_{\ell}}\Big(\widehat{\gamma}_{n}(\varphi_{n,\ell}^{(3)}\varphi_{n,\ell}^{(4)})-\widehat{\gamma}_{n}(\varphi_{n,\ell}^{(3)})\widehat{\eta}_{n}(\varphi_{n,\ell}^{(4)})\Big)\\\mathbf{C}&=\operatorname{diag}_{0\le\ell\le L}\frac{1}{c_{\ell}}\Big(\widehat{\eta}_{n}((\varphi_{n,\ell}^{(4)})^{2})-\widehat{\eta}_{n}(\varphi_{n,\ell}^{(4)})^{2}\Big),\end{aligned}$$

in which we have used the notation $\operatorname{diag}_{0\le\ell\le L}a_{\ell}=\operatorname{diag}(a_{0},\dots,a_{L})$. We note that, by Corollary 1, the entries of $\boldsymbol{\Gamma}_{n}'(\widehat{\boldsymbol{\varphi}}_{n})$ are ensured to be strictly positive.

To complete the induction for the update step, consider the decomposition

$$\begin{aligned}\Psi_{n,n}^{N}(\boldsymbol{t},\boldsymbol{\varphi})&=\sum_{\ell=0}^{L}t_{n,\ell}^{(3)}\Big(\widehat{\gamma}_{n,\ell}^{N}(\varphi_{n,\ell}^{(3)})-\widehat{\gamma}_{n}(\varphi_{n,\ell}^{(3)})\Big)+\sum_{\ell=0}^{L}t_{n,\ell}^{(4)}\Big(|\widehat{\gamma}_{n,\ell}^{N}|(\varphi_{n,\ell}^{(4)})-\widehat{\eta}_{n}(\varphi_{n,\ell}^{(4)})\Big)\\&\quad+\Psi_{n,n-1}^{N}(\widehat{\boldsymbol{t}}_{0,n},\widehat{\boldsymbol{\varphi}}_{0,n})\\&=\widehat{A}^{N}+\widehat{B}^{N}\end{aligned}$$

where $\widehat{\boldsymbol{t}}_{0,n}$ and $\widehat{\boldsymbol{\varphi}}_{0,n}$ are defined such that $\boldsymbol{t}=(\widehat{\boldsymbol{t}}_{0,n}^{T},\widehat{\boldsymbol{t}}_{n}^{T})^{T}$ and $\boldsymbol{\varphi}=(\widehat{\boldsymbol{\varphi}}_{0,n}^{T},\widehat{\boldsymbol{\varphi}}_{n}^{T})^{T}$. Adding and subtracting the terms implied in the following, we can write $\widehat{A}^{N}$ and $\widehat{B}^{N}$ as

$$\begin{aligned}\widehat{A}^{N}=\sum_{\ell=0}^{L}\Bigg[&t_{n,\ell}^{(3)}\Bigg(\widehat{\gamma}_{n,\ell}^{N}(\varphi_{n,\ell}^{(3)})-\frac{\sum_{j=1}^{S(N)}\widetilde{w}_{n}^{j}\varphi_{n,\ell}^{(3)}(\xi_{n}^{j})}{\sum_{j=1}^{S(N)}|\widetilde{w}_{n}^{j}|}\Bigg)\\&+t_{n,\ell}^{(4)}\Bigg(|\widehat{\gamma}_{n,\ell}^{N}|(\varphi_{n,\ell}^{(4)})-\frac{\sum_{j=1}^{S(N)}|\widetilde{w}_{n}^{j}|\varphi_{n,\ell}^{(4)}(\xi_{n}^{j})}{\sum_{j=1}^{S(N)}|\widetilde{w}_{n}^{j}|}\Bigg)\Bigg]\end{aligned}$$

$$\widehat{B}^{N}=\sum_{\ell=0}^{L}\Bigg[t_{n,\ell}^{(3)}\Bigg(\frac{\sum_{j=1}^{S(N)}\widetilde{w}_{n}^{j}\varphi_{n,\ell}^{(3)}(\xi_{n}^{j})}{\sum_{j=1}^{S(N)}|\widetilde{w}_{n}^{j}|}-\widehat{\gamma}_{n}(\varphi_{n,\ell}^{(3)})\Bigg)$$

$$+ t^{(4)}_{n,\ell}\left(\frac{\sum_{j=1}^{S(N)} |\widetilde{w}^j_n|\varphi^{(4)}_{n,\ell}(\xi^j_n)}{\sum_{j=1}^{S(N)} |\widetilde{w}^j_n|} - \widehat{\eta}_n(\varphi^{(4)}_{n,\ell})\right)\Bigg] + \Psi^N_{n,n-1}(\widehat{\boldsymbol{t}}_{0,n}, \widehat{\boldsymbol{\varphi}}_{0,n}).$$

By the results (3.3.3), (3.3.4) and [18, Theorem A.3] — which we state in Appendix A — we have

$$\mathbb{E}\Big[\exp\Big(\mathrm{i}u\sqrt{N}\widehat{A}^N\Big)\Big|\mathcal{G}^N_0\Big] \xrightarrow[N\to\infty]{\mathbb{P}} \exp\Big(-\frac{u^2}{2}\widehat{\boldsymbol{t}}^T_n\boldsymbol{\Gamma}'_n(\widehat{\boldsymbol{\varphi}}_n)\widehat{\boldsymbol{t}}_n\Big). \tag{3.3.5}$$

Moreover, by the induction assumption that (3.3.2) holds for n and $m = n-1$, we can apply the δ-method to deduce that

$$\sqrt{N}\widehat{B}^N \xrightarrow[N\to\infty]{\mathrm{D}} \mathcal{N}\Big(0, \boldsymbol{t}^T\boldsymbol{\Gamma}'_{n,n-1}(\boldsymbol{\varphi})\boldsymbol{t}\Big) \tag{3.3.6}$$

for some $\boldsymbol{\Gamma}'_{n,n-1}(\boldsymbol{\varphi}) \in \mathbb{R}^{4(L+1)(n+1)\times 4(L+1)(n+1)}$. By Lemma 4 the claim that (3.3.2) holds for n and $m = n$ follows from (3.3.5) and (3.3.6), as we have

$$\sqrt{N}\Psi^N_{n,m}(\boldsymbol{t},\boldsymbol{\varphi}) \xrightarrow[N\to\infty]{\mathrm{D}} \mathcal{N}\Bigg(\boldsymbol{0}, \boldsymbol{t}^T\Bigg(\begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Gamma}'_n(\widehat{\boldsymbol{\varphi}}_n)\end{bmatrix} + \boldsymbol{\Gamma}'_{n,n-1}(\boldsymbol{\varphi})\Bigg)\boldsymbol{t}\Bigg).$$

We now show that the asymptotic normality is also preserved by the mutation step. In this case we assume (3.3.2) holds at $n \in \mathbb{N}$ and $m = n$ and show that this implies the same statement is true for $n+1$ and $m = n$. The measures we work with in $\Gamma_{n,m}$ in this case are those corresponding to $k = 1, 2$ in the earlier assignment. Considering the decomposition

$$\begin{aligned}\Psi^N_{n+1,n}(\boldsymbol{t},\boldsymbol{\varphi}) &= \sum_{\ell=0}^{L} t^{(1)}_{n+1,\ell}\Big(\gamma^N_{n+1,\ell}(\varphi^{(1)}_{n+1,\ell}) - \gamma_{n+1}(\varphi^{(1)}_{n+1,\ell})\Big) \\ &\quad + \sum_{\ell=0}^{L} t^{(2)}_{n+1,\ell}\Big(|\gamma^N_{n+1,\ell}|(\varphi^{(2)}_{n+1,\ell}) - \eta_{n+1}(\varphi^{(2)}_{n+1,\ell})\Big) + \Psi^N_{n,m}(\boldsymbol{t}_{0,n}, \boldsymbol{\varphi}_{0,n}) \\ &= A^N + B^N,\end{aligned}$$

where this time $\boldsymbol{t}_{0,n}$ and $\boldsymbol{\varphi}_{0,n}$ are such that $\boldsymbol{t} = (\boldsymbol{t}^T_{0,n}, \boldsymbol{t}^T_{n+1})^T$ and $\boldsymbol{\varphi} = (\boldsymbol{\varphi}^T_{0,n}, \boldsymbol{\varphi}^T_{n+1})^T$, where

$$\begin{aligned}\boldsymbol{t}_{n+1} &= (t^{(1)}_{n+1,0}, \ldots, t^{(1)}_{n+1,L}, t^{(2)}_{n+1,0}, \ldots, t^{(2)}_{n+1,L})^T \in \mathbb{R}^{2(L+1)} \\ \boldsymbol{\varphi}_{n+1} &= (\varphi^{(1)}_{n+1,0}, \ldots, \varphi^{(1)}_{n+1,L}, \varphi^{(2)}_{n+1,0}, \ldots, \varphi^{(2)}_{n+1,L})^T \in \mathscr{B}(\mathbb{X})^{2(L+1)}\end{aligned}$$

and

$$
\begin{aligned}
A^N &= \sum_{\ell=0}^{L} t^{(1)}_{n+1,\ell}\Big(\gamma^N_{n+1,\ell}(\varphi^{(1)}_{n+1,\ell}) - \widehat{\gamma}^N_{n,\ell}(K(\varphi^{(1)}_{n+1,\ell}))\Big) \\
&\quad + \sum_{\ell=0}^{L} t^{(2)}_{n+1,\ell}\Big(|\gamma^N_{n+1,\ell}|(\varphi^{(2)}_{n+1,\ell}) - |\widehat{\gamma}^N_{n,\ell}|(K(\varphi^{(2)}_{n+1,\ell}))\Big) \\
&= \sum_{\ell=0}^{L}\sum_{i\in P^N_\ell}\Bigg(\frac{t^{(1)}_{n+1,\ell}}{c_\ell N} w^i_{n+1}\Big(\varphi^{(1)}_{n+1,\ell}(\xi^i_{n+1}) - K(\varphi^{(1)}_{n+1,\ell})(\widehat{\xi}^i_n)\Big) \\
&\quad + \frac{t^{(2)}_{n+1,\ell}}{c_\ell N}\Big(\varphi^{(2)}_{n+1,\ell}(\xi^i_{n+1}) - K(\varphi^{(2)}_{n+1,\ell})(\widehat{\xi}^i_n)\Big)\Bigg) \\
B^N &= \sum_{\ell=0}^{L} t^{(1)}_{n+1,\ell}\Big(\widehat{\gamma}^N_{n,\ell}(K(\varphi^{(1)}_{n+1,\ell})) - \gamma_{n+1,\ell}(\varphi^{(1)}_{n+1,\ell})\Big) \\
&\quad + \sum_{\ell=0}^{L} t^{(2)}_{n+1,\ell}\Big(|\widehat{\gamma}^N_{n,\ell}|(K(\varphi^{(2)}_{n+1,\ell})) - \eta_{n+1}(\varphi^{(2)}_{n+1,\ell})\Big) + \Psi^N_{n,m}(\boldsymbol{t}_{0,n}, \boldsymbol{\varphi}_{0,n}).
\end{aligned}
$$

For all $0 \le \ell < L$, $N \in \mathbb{N}$ and $i \in P^N_\ell$ we now define $Z_{i,\ell} = w^i_{n+1} X_{i,\ell} + Y_{i,\ell}$, where

$$
X_{i,\ell} = \frac{t^{(1)}_{n+1,\ell}}{c_\ell}\Big(\varphi^{(1)}_{n+1,\ell} - K(\varphi^{(1)}_{n+1,\ell})(\widehat{\xi}^i_n)\Big), \quad Y_{i,\ell} = \frac{t^{(2)}_{n+1,\ell}}{c_\ell}\Big(\varphi^{(2)}_{n+1,\ell} - K(\varphi^{(2)}_{n+1,\ell})(\widehat{\xi}^i_n)\Big).
$$

Moreover, for all $N \in \mathbb{N}$ we define $\widetilde{U}^N_0 = 0$ and

$$
\widetilde{U}^N_\rho = \frac{1}{\sqrt{N}} \sum_{\ell=0}^{L} \sum_{i\in\mathcal{I}^N_\ell(\rho)} Z_{i,\ell}(\xi^i_{n+1}), \qquad 1 \le \rho \le N,
$$

in which case we have

$$
\sqrt{N} A_N = \sum_{\rho=1}^{N} \widetilde{U}^N_\rho.
$$

It is clear that $\widetilde{U}^N_\rho$ is $\widetilde{\mathcal{G}}^N_\rho$-measurable — where $\widetilde{\mathcal{G}}^N_\rho$ is the same σ-algebra constructed in the proof of Lemma 3 — and that $\mathbb{E}[\widetilde{U}^N_\rho \mid \widetilde{\mathcal{G}}^N_{\rho-1}] = 0$ almost surely, implying that $(\widetilde{U}^N_\rho, \widetilde{\mathcal{G}}^N_\rho)_{0\le\rho\le N, N>0}$ is a triangular martingale

difference array. Moreover,

$$|\widetilde{U}_\rho^N| \leq \frac{C'_\varphi}{\sqrt{N}}, \quad \text{where} \quad C'_\varphi = 2\sum_{\ell=0}^{L}\Big(t_{n+1,\ell}^{(1)}\|\varphi_{n+1,\ell}^{(1)}\| + t_{n+1,\ell}^{(2)}\|\varphi_{n+1,\ell}^{(2)}\|\Big)$$

and therefore by a similar argument to (3.3.3) we have

$$\sum_{\rho=1}^{N}\mathbb{E}\Big[\Big(\widetilde{U}_\rho^N\Big)^2\mathbb{I}\Big[\Big|\widetilde{U}_\rho^N\Big| \geq \epsilon\Big]\Big|\widetilde{\mathcal{G}}_{\rho-1}^N\Big] \xrightarrow[N\to\infty]{\text{a.s.}} 0. \tag{3.3.7}$$

Since for all $0 \leq \ell \leq L$, $1 \leq \rho \leq N$ and $i \in \mathcal{I}_\ell^N(\rho)$ the $Z_{i,\ell}$ are conditionally independent given $\widetilde{\mathcal{G}}_{\rho-1}^N$ and $\mathbb{E}[Z_{i,\ell} \mid \widetilde{\mathcal{G}}_{\rho-1}^N] = 0$, we therefore have

$$\sum_{\rho=1}^{N}\mathbb{E}\Big[\Big(\widetilde{U}_\rho^N\Big)^2 \mid \widetilde{\mathcal{G}}_{\rho-1}^N\Big] = \frac{1}{N}\sum_{\ell=0}^{L}\sum_{i\in P_\ell^N} K(Z_{i,\ell}^2)(\widehat{\xi_n^i}). \tag{3.3.8}$$

The quantity

$$K(Z_{i,\ell}^2)(\widehat{\xi_n^i}) = K((X_{i,\ell} + \operatorname{sgn}(\widehat{w}_n^i)Y_{i,\ell})^2)(\widehat{\xi_n^i})$$

is clearly non-negative, but it remains to determine if it is strictly positive. To do so we need to consider two cases: Case 1: $\varphi_{n+1,\ell}^{(i)}$ is almost surely a constant for all $i \in \{1,2\}$, i.e. for all $i \in \{1,2\}$ and some $a_i \in \mathbb{R}$ we have $\varphi_{n+1,\ell}^{(i)} = a_i$ almost surely with respect to the dominating σ-finite measure of Assumption 1, which we call λ. Then $Z_{i,\ell}^2 = 0$ almost surely for all $0 \leq \ell \leq L$ and $i \in P_\ell^N$. Case 2: (the complement of Case 1): for some $\varepsilon > 0$ we define

$$E_\varepsilon^\pm = \Big\{x \in \mathbb{X} : K\Big((X_{i,\ell} \pm Y_{i,\ell})^2\Big)(x) > \varepsilon\Big\}.$$

By Assumption 1 and the fact that at least one of the functions $\varphi_{n+1,\ell}^{(1)}$ or $\varphi_{n+1,\ell}^{(2)}$ must not be a constant λ-a.s., we know that $\eta_n(E_\varepsilon^+) + \gamma_n(E_\varepsilon^+) > 0$ or $\eta_n(E_\varepsilon^-) - \gamma_n(E_\varepsilon^-) > 0$ or both, for a sufficiently small $\varepsilon > 0$. Therefore

$$\begin{aligned}\frac{1}{N}\sum_{i\in P_\ell^N} K(Z_{\ell,i}^2)(\widehat{\xi_n^i}) &= \frac{1}{N}\sum_{i\in P_\ell^{N+}} K((X_{i,\ell} + Y_{i,\ell})^2)(\widehat{\xi_n^i}) \\ &\quad + \frac{1}{N}\sum_{i\in P_\ell^{N-}} K((X_{i,\ell} - Y_{i,\ell})^2)(\widehat{\xi_n^i})\end{aligned}$$

$$> \frac{\varepsilon}{N}\left(\sum_{i\in P_\ell^{N+}} \mathbb{1}[\widehat{\xi}_n^i \in E_\varepsilon^+] + \sum_{i\in P_\ell^{N-}} \mathbb{1}[\widehat{\xi}_n^i \in E_\varepsilon^-]\right)$$
$$\xrightarrow[N\to\infty]{\text{a.s.}} \frac{\varepsilon c_\ell}{2}(\eta_n(E_+) + \gamma_n(E_+) + \eta_n(E_-) - \gamma_n(E_-)),$$

for which the limit is strictly positive by Corollary 1. This implies that the limit

$$\frac{1}{N}\sum_{i\in P_\ell^N} K(Z_{\ell,i}^2)(\widehat{\xi}_n^i) \xrightarrow[N\to\infty]{\text{a.s.}} c_\ell \varsigma_\ell^2 > 0,$$

where it is straightforward to check that

$$\varsigma_\ell^2 = \left(\frac{t_{n+1,\ell}^{(1)}}{c_\ell}\right)^2 \widehat{\eta}_n(K((\varphi_{n+1,\ell}^{(1)} - K(\varphi_{n+1,\ell}^{(1)}))^2)) + \left(\frac{t_{n+1,\ell}^{(2)}}{c_\ell}\right)^2 \widehat{\eta}_n(K((\varphi_{n+1,\ell}^{(2)} - K(\varphi_{n+1,\ell}^{(2)}))^2))$$
$$+ \frac{t_{n+1,\ell}^{(1)} t_{n+1,\ell}^{(2)}}{c_\ell^2} \widehat{\gamma}_n(K((\varphi_{n+1,\ell}^{(1)} - K(\varphi_{n+1,\ell}^{(1)}))(\varphi_{n+1,\ell}^{(2)} - K(\varphi_{n+1,\ell}^{(2)})))).$$

Using (3.3.8) it follows that

$$\sum_{\rho=1}^N \mathbb{E}\left[\left(\widetilde{U}_\rho^N\right)^2 \middle| \widetilde{\mathcal{G}}_{\rho-1}^N\right] \xrightarrow[N\to\infty]{\text{a.s.}} \sum_{\ell=0}^L c_\ell \varsigma_\ell^2 = \boldsymbol{t}_{n+1}^T \boldsymbol{\Gamma}_{n+1}''(\boldsymbol{\varphi}_{n+1}) \boldsymbol{t}_{n+1},$$
$$\text{where } \boldsymbol{\Gamma}_{n+1}''(\boldsymbol{\varphi}_{n+1}) = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B} & \mathbf{C} \end{pmatrix},$$

and

$$\mathbf{A} = \text{diag}_{0\le\ell\le L} \frac{1}{c_\ell} \widehat{\eta}_n(K((\varphi_{n+1,\ell}^{(1)} - K(\varphi_{n+1,\ell}^{(1)}))^2))$$
$$\mathbf{B} = \text{diag}_{0\le\ell\le L} \frac{1}{c_\ell} \widehat{\gamma}_n(K((\varphi_{n+1,\ell}^{(1)} - K(\varphi_{n+1,\ell}^{(1)}))(\varphi_{n+1,\ell}^{(2)} - K(\varphi_{n+1,\ell}^{(2)}))))$$
$$\mathbf{C} = \text{diag}_{0\le\ell\le L} \frac{1}{c_\ell} \widehat{\eta}_n(K((\varphi_{n+1,\ell}^{(2)} - K(\varphi_{n+1,\ell}^{(2)}))^2)).$$

Therefore, if $\boldsymbol{t}_{n+1}^T \boldsymbol{\Gamma}_{n+1}''(\boldsymbol{\varphi}_{n+1}) \boldsymbol{t}_{n+1} > 0$ — which is true if Case 2 holds for any $0 \le \ell \le L$ — then again by [18, Theorem A.3] it follows that

$$\mathbb{E}\left[\exp\left(\mathrm{i}u\sqrt{N}A^N\right) \middle| \widetilde{\mathcal{G}}_0^N\right] \xrightarrow[N\to\infty]{\mathbb{P}} \exp\left(-\frac{u^2}{2} \boldsymbol{t}_{n+1}^T \boldsymbol{\Gamma}_{n+1}''(\boldsymbol{\varphi}_{n+1}) \boldsymbol{t}_{n+1}\right). \quad (3.3.9)$$

By the simple relations

$$\gamma_{n+1}(\varphi^{(1)}_{n+1,\ell}) = \widehat{\gamma}_n(K(\varphi^{(1)}_{n+1,\ell})) \quad \text{and} \quad \eta_{n+1}(\varphi^{(2)}_{n+1,\ell}) = \widehat{\eta}_n(K(\varphi^{(2)}_{n+1,\ell}))$$

we have by the induction assumption that (3.3.2) hold for n and $m = n$, and the δ-method

$$\sqrt{N}B^N \xrightarrow[N\to\infty]{\mathrm{D}} \mathcal{N}\Big(0, \boldsymbol{t}^T \boldsymbol{\Gamma}''_{n,n}(\boldsymbol{\varphi})\boldsymbol{t}\Big) \tag{3.3.10}$$

for some symmetric and positive semi-definite $\boldsymbol{\Gamma}''_{n,n}(\boldsymbol{\varphi})$. The claim that (3.3.2) also holds for $n+1$ and $m = n$ follows from (3.3.9), (3.3.10) and Lemma 4, since we have shown that

$$\sqrt{N}\Psi^N_{n,m}(\boldsymbol{t}, \boldsymbol{\varphi}) \xrightarrow[N\to\infty]{\mathrm{D}} \mathcal{N}\left(\mathbf{0}, \boldsymbol{t}^T\left(\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Gamma}''_n(\boldsymbol{\varphi}) \end{bmatrix} + \boldsymbol{\Gamma}''_{n,n}(\boldsymbol{\varphi})\right)\boldsymbol{t}\right).$$

Since we know that $\boldsymbol{t}^T_{n+1}\boldsymbol{\Gamma}''_{n+1}(\boldsymbol{\varphi}_{n+1})\boldsymbol{t}_{n+1}$ is non-negative, the only case we still need to consider is that for which $\boldsymbol{t}^T_{n+1}\boldsymbol{\Gamma}''_{n+1}(\boldsymbol{\varphi}_{n+1})\boldsymbol{t}_{n+1} = 0$, which occurs only if for all $0 \le \ell \le L$ we have Case 1. In this degenerate case $A_N = 0$ almost surely and the claim follows immediately with a degenerate limiting distribution with zero variance.

Lastly for the base case, we show that (3.3.2) holds for $n = 0$ and $m = -1$. For this we observe that $\gamma_{0,\ell} = |\gamma_{0,\ell}|$ almost surely and $\gamma_0 = \eta_0 = \pi_0$. An analogous proof to that of the update step yields the result

$$\sqrt{N}\Psi^N_{0,-1}(\boldsymbol{t}, \boldsymbol{\varphi}) \xrightarrow[N\to\infty]{\mathrm{D}} \mathcal{N}\left(\mathbf{0}, \boldsymbol{t}^T\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B} & \mathbf{C} \end{pmatrix}\boldsymbol{t}\right),$$

where $\boldsymbol{t} = (t^{(1)}_0, \ldots, t^{(1)}_L, t^{(2)}_0, \ldots, t^{(2)}_L)$, $\boldsymbol{\varphi} = (\varphi^{(1)}_0, \ldots, \varphi^{(1)}_L, \varphi^{(2)}_0, \ldots, \varphi^{(2)}_L)$, $\Psi^N_{0,-1}$ is as defined in (3.3.1) with the sum over q being equal to zero and

$$\begin{aligned}
\mathbf{A} &= \mathrm{diag}_{0\le\ell\le L}\frac{1}{c_\ell}\pi_0((\varphi^{(1)}_\ell - \pi_0(\varphi^{(1)}_\ell))^2) \\
\mathbf{B} &= \mathrm{diag}_{0\le\ell\le L}\frac{1}{c_\ell}\pi_0((\varphi^{(1)}_\ell - \pi_0(\varphi^{(1)}_\ell))(\varphi^{(2)}_\ell - \pi_0(\varphi^{(2)}_\ell))) \\
\mathbf{C} &= \mathrm{diag}_{0\le\ell\le L}\frac{1}{c_\ell}\pi_0((\varphi^{(2)}_\ell - \pi_0(\varphi^{(2)}_\ell))^2).
\end{aligned}$$

The proof is then completed by the result of Lemma 2, the definitions

$$\frac{\gamma_n(\varphi)}{\gamma_n(1)} = \pi_n(\varphi), \quad \text{and} \quad \frac{\widehat{\gamma}_n(\varphi)}{\widehat{\gamma}_n(1)} = \widehat{\pi}_n(\varphi)$$

and by applying the δ-method.

$\square$

Chapter 4

Model applications

4.1 Description of models

In line with the drilling application, we implement the MLBPF in applications where the observations are based on solutions to differential equations. In particular, we consider models involving ordinary differential equations (ODEs) and partial differential equations (PDEs) that have analogous properties to the compressible Euler equations, as well as practical interpretations in their own right. Though the models that we describe here are differential equation-focused, we emphasise that the potential applications of the MLBPF extend beyond this paradigm to any application that satisfies the cost-accuracy hierarchy assumption in the g_n^ℓ. In addition to the method-specific implementation issues for the MLBPF, there are non-trivial issues for general particle filtering approaches on systems based on PDE solutions which we also discuss.

The two main examples of differential equations we focus on are the shallow water equations and the convection-diffusion equation, both of which are considered in a one-dimensional setting. The shallow water equations are a multi-dimensional hyperbolic system of PDE that are used to model fluid heights and velocity when the vertical behaviour is negligible compared to the horizontal behaviour. Consequently, they are often used in flood and tsunami modelling [27]. The convection-diffusion equation is a parabolic PDE analogous to the conservation of mass equation, with the addition of a second-order term to model the diffusivity of the variable of interest. One of its potential applications is to study the concentration of a pollutant moving

in a river; in such scenarios the pollutant is both transported by the fluid and disperses via its own diffusive properties. In contrast to the shallow water case, the velocity of the surrounding body is assumed to be a known function of space and time, and the unknown solution is the function describing the concentration of the matter.

Both the shallow water and convection-diffusion equations can be supplemented with a source term $S = S(z, t)$ that describes a scenario in which a particular quantity is not conserved within the system. For example, in the convection-diffusion model, S represents the creation or destruction of the matter from a process such as a chemical reaction or some physical process, while in the shallow water case one widely-used application of S is its ability to model the non-flat topography of the channel that the fluid moves in. In addition to the described scenarios, in our experiments we assume that S is parametrised by a signal process that we estimate using solutions to the differential equation. An important principle of taking this approach is that S has both a sufficient dependency on the parameter and exerts a sufficient influence on the solution to the differential equation that it can be inferred from the specified observations. To further complicate matters, in the PDE context there is a natural time transience in the differential equation solutions even in the absence of a source term, meaning the respective influences of the parameter and of time must be discerned from one another. We note that the inclusion of the source terms in the shallow water and convection-diffusion equations is consistent with their classical interpretation, in the sense that $S \equiv 0$ models a perfectly flat seabed and an absence of creation or destruction of matter respectively. Lastly, we emphasise that the scenarios we describe are just two of many possible applications; other interesting examples of source term modelling based on the Euler equations alone include using S to model gravitational terms that arise in applications from weather prediction to astrophysics [42], or as a permeability coefficient when quantifying fluid loss to the surrounding rock formation in a drilling application such as that conducted by Schlumberger.

Though in theory both the shallow water and convection-diffusion equations assume unbounded spatial domains, in practice these spaces must be truncated and a finite time fixed. Once appropriate boundary conditions, initial conditions and hyperparameters are specified, the models are then well-posed for numerical approximations and particle filtering applications.

4.1.1 Shallow water equations

For some $W > 0$, let $[0, W]$ denote the spatial seabed domain over which the shallow water is modelled and consider the time variable $t \in [0, T]$ for some experiment run time $T > 0$. We denote by $\lambda(z, t) \in [0, \infty)$ the height of the water from the seabed at the point $z \in [0, W]$ and time $t \in [0, T]$, and by $u(z, t) \in \mathbb{R}$ the associated velocity. The depth-averaged discharge is then given by the quantity $q = \lambda u$. For a non-flat seabed with topography $Z(z, t)$, the corresponding shallow water equations in their full and explicit form are

$$\partial_t \lambda + \partial_z q = 0 \tag{4.1.1}$$

$$\partial_t q + \partial_z \left(\frac{q^2}{\lambda} + \frac{1}{2} g \lambda^2 \right) = -g\lambda \partial_z Z \tag{4.1.2}$$

where $g = 9.81$ is the acceleration due to gravity. We assume that at the time of the n-th observation the topography Z is parametrised by a latent random variable X_n such that $(X_n)_{n=0}^{\infty}$ is a Markov process governed by a Markov kernel as described in Chapter 2.

Before proceeding further with the description of our shallow water equations application, we return momentarily to the drill-system application and discuss the similarities between estimating parameters based on the model (4.1.1)–(4.1.2) and on the fluid dynamics model governing the drill fluid. We first note that, with some minor alterations, the source term function $S = (0, Z)^T$ can be used to model a downhole event parametrised by the hidden Markov process X_n. For example, in the event of washout caused by a hole of area X_n, the corresponding source term is

$$S(z, t_n; X_n) = (X_n \bar{u} \rho_{z_W} \delta_{z_W}, 0)^T,$$

where t_n is the time corresponding to X_n, $\bar{u} = \bar{u}(\rho)$ is the mean flow velocity through the hole, ρ_{z_W} is the mud density at the hole and δ_{z_W} is a delta function centred at the location of the hole. In this sense, we see that the source term really does capture the applications-specific event of interest, and that switching from one event to another (at least within the same application) can be achieved by suitably configuring the source term. However, the true similarities are in the PDEs themselves. By replacing λ with the mud density ρ, the function $q = \rho u$ then becomes the fluid momentum and (4.1.1) is

simply the equation for the conservation of mass. Under this same interpretation, the first two terms in (4.1.2) are also the same as the conservation of momentum equation in the Euler equations, with only the third term instead substituted for an additional unknown pressure function; see e.g. [14, Ch. 1], [47, Ch. 14]. To close the system in the Euler equations scenario, the conservation of mass and momentum equations are often supplemented by an extra equation based on the conservation of energy, or alternatively simply closed with the two conservation equations by establishing a so-called equation of state relating the pressure to the density. Perhaps most crucially, the shallow water equations and Euler equations are both hyperbolic PDE and therefore possess the same mathematical properties in their solutions. Therefore although the two sets of equations may not be identical, for the purposes of hidden parameter estimation these differences are somewhat inconsequential and the two models are essentially interchangeable. Indeed, we only elected to instead study the shallow water equations due to the greater intuition behind the solutions and the relative availability of literature on how to correctly handle shallow water source terms numerically, which is a challenging problem in its own right.

Returning to (4.1.1)–(4.1.2), as a generic model we assume Z has the following relationship to the probability density function f_Γ of the gamma distribution

$$Z(z, t; X_n) = f_\Gamma(z; k_0, X_n), \tag{4.1.3}$$

with random scale parameter $X_n > 0$ and a known, constant shape parameter $k_0 > 0$. In a practical interpretation, (4.1.3) describes a topography of a smooth trough that is moved further from the origin for increasing values of X_n. Such a random evolution of a seabed could be used as a simple model of a tsunami or the disturbance of underwater ecosystems, for example. Since f_Γ is only valid for positive values, we bound X_n in some positive interval $(a, b) \subset (0, \infty)$ using an appropriately scaled sigmoid function. Taking this kind of approach is generally more preferable than simply "rejecting" values outside of (a, b) that arise from an unmodified random walk, since the trajectory of the path near the asymptotes $x = a, b$ is tempered in a way that is consistent with random walk behaviour in non-bounded spaces.

Given a latent seabed parameter state x_n, we define the partial observa-

tion y_n of x_n to be the following observable height of the water

$$y_n = h(x_n) + \nu = \lambda(W, t; x_n) + Z(W, t; x_n) + \nu \tag{4.1.4}$$

where $\nu \sim \mathcal{N}(0, \sigma_Y^2)$ for some known variance σ_Y^2. The corresponding likelihood function is thus described by

$$g_n(x_n) \propto \exp\left(- \frac{(h(x_n) - y_n)^2}{2\sigma_Y^2} \right). \tag{4.1.5}$$

For the multilevel likelihood function g_n^ℓ, we denote the noiseless approximate solution that induces g_n^ℓ by h^ℓ and analogously whenever a different notation is used in place of h.

In the case where the solution $w = (\lambda, q)$ does not vary with time, then $\partial_t w = 0$ and hence $q = q_0$ for some constant q_0. Under this assumption, (4.1.1)–(4.1.2) reduces to the non-linear ODE

$$\frac{d}{dz}\left(\frac{q_0^2}{\lambda} + \frac{1}{2} g \lambda^2 \right) = -g\lambda \frac{dZ}{dz}, \tag{4.1.6}$$

which is the steady state version of the shallow water equations. If λ is sufficiently smooth and an initial value is assigned at $z = 0$, then the solution to (4.1.6) satisfies the initial value problem

$$\frac{d\lambda}{dz} = \frac{g}{q_0^2/\lambda^3 - g} \frac{dZ}{dz}, \qquad \lambda(0) = \lambda_0. \tag{4.1.7}$$

The steady state solutions that satisfy (4.1.7) are the wave profiles that arise if a shallow water system is subjected to time-constant forces and rebounding effects such as a wave hitting a wall are neglected. It is a useful initial model to first apply the MLBPF and BPF to because, in addition to being an intuitive application, there are none of the complications of the additional underlying state transience that is present in PDE applications.

Despite its simplified form, (4.1.7) can in general only be solved analytically in a few cases. One particularly well-studied case is that for which $q_0 = 0$, leading to the so-called *lake at rest problem* [53] with solution $\lambda = -Z + \lambda_0 + Z_0$. As the name suggests, the absence of any velocity describes the case in which the fluid is completely unperturbed, and therefore the solution is symmetrical to the topography of the seabed. However,

for $q_0 \neq 0$ a numerical method is generally required to find approximate solutions to (4.1.7). For its high order accuracy and ease of implementation we choose the Runge-Kutta RK4 method (cf. [68, Ch. 12], for example). The numerical approximation of y_n given x_n is then obtained by numerically solving (4.1.7) subject to x_n and evaluating the resulting numerical approximation according to (4.1.4).

4.1.2 Convection-diffusion equation

For the convection-diffusion equation modelling an unknown concentration $\rho(z,t)$, we consider the specific scenario in which the known fluid velocity is $v(z,t) = v_0$ for some constant $v_0 > 0$ and the diffusion coefficient of the matter that is the coefficient of ∂_{zz} is $D = 1$. The corresponding one-dimensional equation with source term $S(z,t)$ is then

$$\partial_t \rho + v_0 \partial_z \rho = \partial_{zz} \rho + S, \tag{4.1.8}$$

which we consider over the same type of spatio-temporal domain as the shallow water equations. For (4.1.8) to be well-posed requires the specification of both initial and boundary conditions; for the simpler case of the initial conditions we denote

$$\rho(z,0) := \rho_0(z). \tag{4.1.9}$$

For simplicity we assume ρ_0 is a strong solution of (4.1.8) and is therefore twice continuously differentiable on $(0,W)$; by the presence of the second order term in (4.1.8) this ensures $\rho(z,t)$ is also continuous at all future times $t > 0$.

Since we have assumed without loss of generality that $v_0 > 0$, the transporting fluid implicit in (4.1.8) is modelled as moving from left to right. This means it is only possible to impose boundary conditions at the inlet boundary $z = 0$, since the value at the outlet $z = W$ is dictated by the solution. In the event that $v_0 < 0$ the boundary condition would instead be specified at $z = W$, with the information being propagated in time towards $z = 0$.

For the boundary conditions, in our experiment we assign the influx boundary conditions to be the time-dependent values given by

$$\rho(0,t;X_t) := X_t e^{-t}, \tag{4.1.10}$$

where X_t is the concentration creation rate parameter that forms the random process we seek to estimate. This parameter is also encoded in the source term which we define as

$$S(z, t; X_t) = ze^{X_t} \tag{4.1.11}$$

so that, at all times, the rate at which concentration is created increases linearly in space and exponentially in the parameter value. In a practical sense, (4.1.10)–(4.1.11) models an initial influx of the substance ρ at the left boundary that is proportional to the creation rate X_t, which is then created exponentially as the substance is transported by the surrounding fluid. As with the shallow water equations, we confine the range of X_t to a positive interval (a, b). In this experiment this is only necessary to ensure positivity of the concentration in the boundary condition (4.1.10); in general there is no reason why X_t cannot be negative (i.e. the substance is being destroyed).

To estimate the latent random process X_t we let t_n be the experiment time each partial observation is taken and define

$$y_n = \rho(L, t_n; x_n) + \nu \tag{4.1.12}$$

where $x_n := x_{t_n}$, such that (X_n) is a random walk Markov process and $\nu \sim \mathcal{N}(0, \sigma_Y^2)$ for some known variance σ_Y^2. Under our practical interpretation, (4.1.12) measures the rate at which a diffusive substance is created in a transporting body moving at velocity v_0 by observing the substance concentration at the outlet and accounting for measurement error. Its likelihood function is analogous to (4.1.5), where instead we set $f(x_n) = \rho(L, t_n; x_n)$. We note that, although in theory we have $t_0 < t_1 < \ldots < \infty$, in our numerical experiments we have set y_n according to (4.1.12), and then solved the PDE (4.1.8) over an integration time $t \in [0, T]$ for some fixed $T > 0$. This process is then repeated at the next iteration, with the PDE initialised at time $t = 0$ with initial condition the output solution ρ from the previous iteration. In particular, we note this means that the left boundary condition (4.1.10) is equal to the value of X_t at the start of each iteration and decays exponentially in time for $t \in [0, T]$.

As with the shallow water equations, the PDE defined by (4.1.8)–(4.1.12) in general has no closed form solutions. Assuming momentarily that $S \equiv 0$, then (4.1.8) is the combination of the transport (advection) equation and the diffusion equation; each obtainable from (4.1.8) by setting $D = 0$ and $v_0 = 0$

respectively. In particular for the transport equation, closed form solutions are obtainable using the method of characteristics to deduce

$$\rho(z,t) = \rho_0(z - v_0 t). \tag{4.1.13}$$

The solution (4.1.13) is simply the initial substance concentration profile moving from left to right at speed v_0, and reflects both the absence of any diffusive properties and the constant assumption of the speed. In general once any of these assumptions are lifted (i.e. diffusion is accounted for, a source term is added or velocity is assumed to vary) then a numerical method is required. For its unconditional stability, efficiency and time-accuracy, we use the Crank-Nicolson method [45, Ch. 4], which is an implicit finite difference method suited for diffusion problems such as (4.1.8), in which all strong solutions are guaranteed to be continuous.

In each of the following experiments we generate N_{data} independent signal/observation data sets each of length N_{length}. These sets are obtained by simulating an artificial signal process that models a true latent signal and then computing the observations corresponding to each realised state using the full accuracy solver. Since in each experiment we are considering a numerical differential solution, this accuracy is controlled by a spatial mesh resolution parameter $M_L = M$. In addition to the assumption that the observation noise is normally distributed, in each experiment we assume the Markov kernel $K(x_n, \cdot)$ is normally distributed with mean x_n and standard deviation σ_X, making (X_n) a random walk.

To compare the accuracy of the MLBPF to the BPF, a reference solution is produced by running the BPF on each data set with a large number of particles N_{ref}, from which the pre-resample empirical distribution data is stored for each iterate. The BPF is then run N_{trials} times on each data set with $N_{\text{BPF}} < N_{\text{ref}}$ particles. On each trial run, the accuracy of the resulting BPF estimate is measured against the reference solution using some metric of choice, and the discrepancy is averaged over the iterations. At the end of the BPF simulations, we are thus left with $N_{\text{data}} \times N_{\text{trials}}$ errors, each of which is an average taken over N_{length} iterations.

To test the performance of the MLBPF, the average time T_{BPF} taken by the BPF to complete a run on a data set is computed. Throughout all experiments we consider only the two level case, i.e. where $L = 1$. This

simplifies matters by leaving only one level below the full accuracy level on which a mesh size M_0 is tested. In general, we do not know a priori what the optimal choice of M_0 is, nor do we know an optimal particle allocation (N_0, N_1). Therefore, a configuration of N_{mesh} proposed level 0 meshes $\{M_{0,k}\}_{k=1}^{N_{\text{mesh}}}$ are specified along with N_{alloc} level 1 sample sizes $\{N_{1,j}\}_{j=1}^{N_{\text{alloc}}}$, where we assume $\max_k M_{0,k} < M$ and $\max_j N_{1,j} \leq N_{\text{BPF}}$. For each $M_{0,k}$ and $N_{1,j}$, the corresponding level 0 particle allocation $N_{0,j}$ is computed such that the resulting runtime of the MLBPF is approximately equal to T_{BPF}; tables of these sample sizes are provided in Appendix C, while the values of the hyperparameters run on each experiment are given in Appendix B. The MLBPF is then applied with respect to these computed configurations in the same manner as the BPF, and the length-averaged errors between the two methods are recorded over all of the $N_{\text{data}} \times N_{\text{trials}}$ runs. In each experiment we have used the RMSE based on the point estimates and the reference distribution point estimates; that is,

$$\text{RMSE}(\pi^N) = \sqrt{N_{\text{length}}^{-1} \sum_{n=0}^{N_{\text{length}}} (\widehat{x}_n - \widehat{x}_n^{\text{ref}})^2}, \tag{4.1.14}$$

where π^N is the corresponding filtering method configured with N particles and $\widehat{x}_n = \sum_{i=1}^N w_n^i \xi_n^i$ is the *maximum a posteriori* estimate of x_n with respect to the quadratic loss function [62, Ch. 1].

4.2 ODE-based applications

For their relative simplicity compared to PDE-based applications, we first consider the ODE scenario and use the steady state shallow water equations as a case study. Since there is no transience in the state of the system from one iteration to another, we apply Algorithm 3 in its unmodified state to produce the multilevel estimates. In contrast, for the convection-diffusion application the true state of the concentration profile is unknown at the time of each measurement y_n whenever $n > 0$. To implement the MLBPF on models such as these we need to extend Algorithm 3 to one that accommodates the PDE-specific considerations, which we provide in the next section in Algorithm 4. For other ODE-based applications the implementation of the MLBPF will be model-specific but analogous to those discussed in this

section. See [31] for an example of an Euler-Bernoulli beam experiment, in which noisy deflections of a beam fixed at one end and free at the other were used to infer the location of a load moving on the beam.

The following technique can in fact be implemented not only on differential equations models but on any model that is applicable to the MLBPF. Logically, it makes sense to introduce it here before proceeding further with the results, but we emphasise that as a technique it is not exclusive to ODE (or even PDE)-based applications.

4.2.1 Improved approximations with linear regression

While the asymptotics of the MLBPF ensures the convergence of $\widehat{\pi}_n^N$ to $\widehat{\pi}_n$ irrespective of the extent of the bias in the lower level estimates, in practice if this bias is too severe the accuracy of the MLBPF can be poor. However, information provided by the correction data solutions can be used to rectify this inaccuracy at minimal extra cost to the algorithm.

Considering the Gaussian likelihood model (4.1.5) in the two-level scenario, by the telescoping sum expression $h^1 = h^0 + (h^1 - h^0)$, a level 0 solution $h^0(\xi_n^i)$ can be generated at level 0 cost but improved to level 1 accuracy if $h^1 - h^0$ is known at all x_n. In general this is not possible, since the $h^\ell(\xi_n^i)$ are random variables that depend on a numerical approximation to an intractable problem. However, observations of $h^1 - h^0$ are generated at no extra cost in the level 1 weighting step that can be used to construct an approximation

$$r(x_n) = r(x_n \mid (\xi_n^i)_{i \in P_1^N}) \approx h^1(x_n) - h^0(x_n).$$

This can then be evaluated at each level 0 particle to produce an improved-accuracy level 0 solution $h^0(\xi_n^i) + r(\xi_n^i) \approx h^1(\xi_n^i)$ and hence an improved level 0 likelihood term $\widetilde{g}_n^0(\xi_n^i)$ based on $h^0(\xi_n^i) + r(\xi_n^i)$ instead of $h^0(\xi_n^i)$, while maintaining the cheap cost benefit of the level 0 solution.

A natural method for constructing r from the correction data is to use linear regression to produce a polynomial approximation of the deterministic map $h : x \in \mathbb{X} \mapsto h(x) \in \mathbb{Y}$. For example, in the steady-state shallow water equations model, the N_1 level 1 observed water height solutions that we generate with the particles $\xi_n^i \in P_1^N$ can be used as data to construct a linear regression estimator $r(x_n)$ that approximates $h^1(x_n) - h^0(x_n)$. With $r(x_n)$

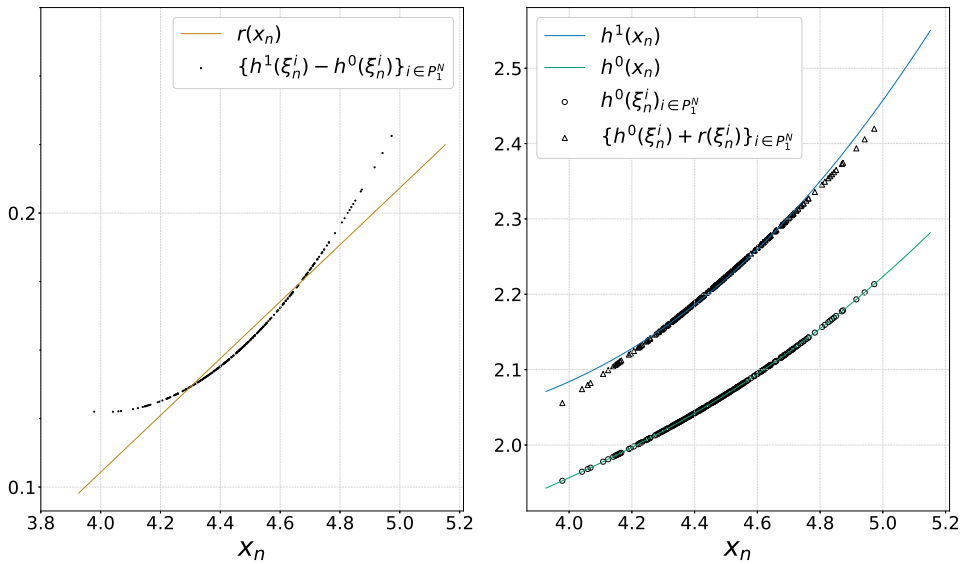

Figure 4.1: Left: Regression approximation $r(x_n)$ of $h^1(x_n) - h^0(x_n)$ in the steady-state shallow water equations using the level 1 correction data for $N_1 = 250$, $M_0 = 40$, $M_1 = 750$ and $n = 1$. Right: The raw h^0 solution samples are improved to samples that are closer to the true h^1 solution samples.

constructed, we then improve the N_0 water height solutions by evaluating r at each of the $\xi_n^i \in P_0^N$ to obtain $\widetilde{h}_0(\xi_n^i) := h_0(\xi_n^i) + r(\xi_n^i)$. In general, $\widetilde{h}_0(\xi_n^i)$ will be a more accurate approximation to $h_1(\xi_n^i)$ than $h_0(\xi_n^i)$, particularly because of the absence of noise in the data the regression curve is trained on. Furthermore, due to the need to evaluate the more expensive level 1 solutions these corrections are obtained at practically no extra cost to the algorithm. An analogous approach has been applied to beneficial effect in [6], in which local linear regression was used to modify parameter samples to produce those which induce sample statistics that are closer to the observed data. The difference between our approach and that of [6] lies in the desired quantities; in our own work we are ultimately interested in modifying the samples to reduce the discrepancy in the solution space, i.e. the image space, while in [6] it is the adjustment of the samples in the domain that is of interest, since parameter samples for which the discrepancy from the observed data is small form an approximate posterior sample.

In the context of filtering the linear regression method that we apply is

favourable over Gaussian processes or a neural network, since we require only a reasonable approximation that is cheap in order to keep the weighting cost down. The task is somewhat different from the conventional setting in which linear regression is typically applied since, for fixed x, the map $h^\ell(x)$ is absent of noise for any ℓ. Nevertheless, overfitting should still be avoided since by assumption the value of the correction curve at all x is unknown. For higher dimensional signal spaces the construction of the approximating hyperplane becomes more computationally cumbersome but the principle remains the same. The situation is somewhat simpler for experiments involving multi-dimensional observation data, in which the regression translation is simply computed and applied element-wise to the components of the observation vector.

One drawback of linear regression is that we do not know in advance the optimal choice of polynomial degree to construct to. Given that h^1 is fixed, the shape of the curve $h^1 - h^0$ is therefore dependent on both the choice of solver and the specified level of accuracy in the level 0 solution. For example, we see in Figure 4.1 that the regions where the linear approximation is at its poorest for the steady-state shallow water equations are those furthest from $\mathbb{E}[h^1(x_n) - h^0(x_n)]$. There is therefore a global curvature to $h^1 - h^0$ that is increasingly prominent the further M_0 decreases from M_1. Similarly, for a solver with lower order accuracy than RK4, this dispersion feature in the tails may be more prominent or feature in a larger proportion of the samples, leading to the same loss of accuracy in the likelihood improvements. In our experiments we found polynomials of degree ≤ 1 to be a suitable choice; while the correction data $\{h^1(\xi_n^i) - h^0(\xi_n^i)\}_{i \in P_1^N}$ is rarely exactly linear, the improvements offered by a degree 1 correction are more than sufficient, are cheaper to compute than higher order polynomials, and avoid the problems that arise from overfitting as the polynomial degree is increased. Consequently, all of the following results correspond to a level 1 regression improvement being applied to the level 0 solution data. In [31] a degree 0 improvement was considered on a big data problem and a degree 1 improvement on an Euler-Bernoulli beam problem; in each of the models considered here we construct the $\widetilde{g}_n^0$ using a polynomial of degree 1 as on the left panel in Figure 4.1.

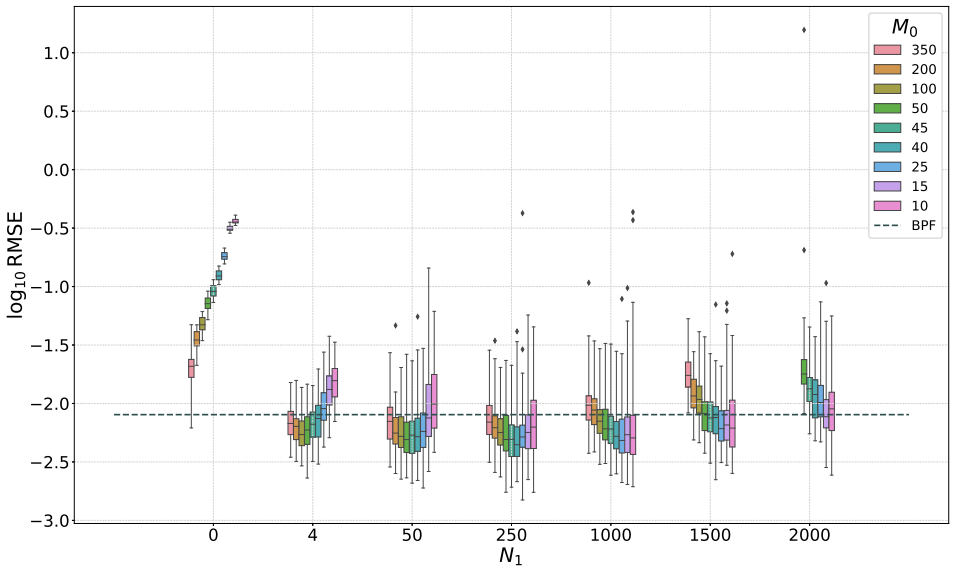

Figure 4.2: Box plots of the MLBPF RMSE over all trials and data sets compared to the median RMSE of the BPF. Note that, for added clarity, we've extended the tails to include data from the standard $\pm 1.5\times$IQR to $\pm 3.0\times$IQR.

4.2.2 Results for ODE model

Figure 4.2 displays box plots of the RMSE (4.1.14) over all of the $N_{\text{data}} \times N_{\text{trials}}$ runs for each particle-mesh allocation considered in the experiment. We see that there are several configurations for which at least 75% of the trials are more accurate than the median BPF RMSE, with $(M_0, N_1) = (40, 250)$ the allocation with both the lowest median and lowest upper quartile. Excluding the case where $N_1 = 0$, in which we are essentially running the BPF with a different ODE solver and hence an uncorrected bias, the configuration where $N_1 = 4$ has the most stable batch of estimates. This is because close to all of the particles are on level 0 and hence, while the resulting estimates may not exhibit the optimal global accuracy, they are essentially all subjected to a stable transformation via the linear regression which is manifested in the RMSE. In contrast, the range of the RMSE for $N_1 = 50$, say, is much larger due to the volatility of more multilevel corrections. The reason for the absence of results for $M_0 \in \{300, 200, 100\}$ at $N_1 = 2000$ is due to the fact that, subject to the level 1 cost, there is

no remaining level 0 budget available in which the MLBPF runtime can be made comparable to the BPF, making any subsequent accuracy comparison invalid. Whilst the results at $N_1 = 2000$ align with the intuition that a mesh-specific optimum is achieved with $N_0(M_0) > N_1$, our assumptions on the sample allocations do not impose this as a necessary condition to run the MLBPF; indeed, Table 2 in Appendix B specifies several instances in which the computed N_0 is such that $N_0 < N_1$ in order to satisfy the empirical time equivalence requirement of the experiment.

While Figure 4.2 shows accuracy gains can be attained using the MLBPF, these gains need to be put into context of the model uncertainty. To do this, we note that the reference standard deviation σ_{ref} averaged over the N_{length} iterations in the N_{data} data sets is 0.227. Selecting $(M_0, N_1) = (40, 250)$ as an optimal configuration, the median BPF RMSE as a proportion of σ_{ref} is approximately 0.0395, while for the MLBPF it is approximately 3.5×10^{-4}. Using the relative gain formula

$$\text{RelGain}(\varepsilon_1, \varepsilon_2) = \left(100\% - \frac{\varepsilon_1 - \varepsilon_2}{\varepsilon_1} \times 100\%\right), \tag{4.2.1}$$

the relative gain in accuracy of the MLBPF with respect to the model uncertainty is approximately 1%. While this is not overly impressive, we note that the results of Figure 4.2 were generated with respect to the BPF configured with 2500 particles and therefore the estimate accuracy is already fairly saturated, particularly for a one-dimensional signal process. However, when tasking the BPF with attaining the same accuracy as the optimal MLBPF configuration it took roughly 11 times longer, requiring approximately 12 times as many particles than what it was originally run with. The practical takeaway from these points is that, for this particular model and configuration, if a high level of precision is desirable then the MLBPF is able to achieve this in significantly less time than the BPF. This is a model-specific feature of the MLBPF and not a general one; for a more significant relative gain on an ODE model we refer the reader to [31].

4.2.3 Numerical and Monte Carlo error

As is synonymous with multilevel methods, there are two sources of error in the MLBPF: the numerical error due to the lower accuracy solver, and the Monte Carlo (MC) error due to the finite number of samples. Each of these

errors are present in the two forms of curvature that we see in the results of Figure 4.2. If we first focus on the "local" curvature, in which N_1 is fixed and the M_0 are iterated over, the explanation for this behaviour is due to both forms of error. Considering the batch corresponding to $N_1 = 4$, as M_0 iterates initially through $\{350, 200, 100\}$, the loss of numerical solution accuracy at each mesh size reduction is compensated for by the gain in the number of available level 0 particles and a subsequent reduction in the MC error. The crucial feature to note is that this gain in N_0 comes with a caveat: if the loss in numerical accuracy becomes too large, then the bias of the level 0 empirical distribution becomes too substantial to be corrected by the level 1 corrections, regardless of how many particles are assigned to level 0. This is why as M_0 iterates through $\{50, 45, \ldots, 10\}$ at $N_1 = 4$ that the MLBPF accuracy gets poorer. This phenomenon is even more prominent without the regression adjustment, for which the bias in the level 0 empirical estimate is frequently too substantial, resulting in likelihood evaluations that are practically all zero. Consequently, inference is being performed with essentially $N_1 < N_{\text{BPF}}$ particles, which naturally produces less accurate estimates than the BPF. If the application at hand has a small signal-to-noise ratio then the biased likelihood evaluations are less likely to be essentially zero, but such scenarios can also render the relative gains in the distribution estimates meaningless.

The other form of curvature in Figure 4.2 is that for which M_0 is fixed and N_1 varies. By viewing the results in this manner, we are observing the aggregate benefits of the statistical gains from having a large level 0 allocation against the bias corrections from the more expensive level 1 evaluations. In general, we see that the more accurate the level 0 solution, the less corrections are required and the mesh-specific optimal choice is likely to be for a smaller N_1. For example, the configuration for which $M_0 = 350$ really only ever "cares" about reducing its MC error, since its numerical solution is accurate enough and it is the absence of samples that is the prime cause of error. This is why the RMSE for $M_0 = 350$ essentially only ever increases in N_1, since it loses more of its level 0 budget as more particles are assigned to the expensive level 1 computations and it is the shortage of level 0 samples that is the most present issue. At the other extreme, $M_0 = 10$ has to contend with a more significant numerical error, meaning that even though it has the largest number of level 0 particles of all of the level 0 mesh choices,

to reach its optimal configuration it is happy to sacrifice some of these particles for some level 1 corrections to reduce the more-prominent numerical error. Eventually the numerical error is corrected enough that the MC error starts to take over and the overall accuracy of the results gets poorer as less particles are assigned to level 0.

While it is clear that this "global" curvature exists in Figure 4.2 under the specified experiment hyperparameters, in many experiments the situation can arise where the global minimum RMSE is at the minimum N_1 and minimum M_0. When this is the case, the regression approximation as in Figure 4.1 is sufficiently accurate even for the minimum M_0 that the numerical error is essentially eradicated and only the MC error remains. Hence, the more particles on level 0, the better. This scenario can occur when the correction data for the M_0 is close to perfectly linear and/or when the signal-to-observation noise ratio is small enough that any error in the regression approximation becomes irrelevant. An example scenario is plotted for the steady state shallow water equations model in Figure 4.3. Despite $r(x_n)$ being globally non-linear, in this example the signal noise is sufficiently small that the domain over which $r(x_n)$ is evaluated means that the curve we consider is close to linear and is therefore practically exactly computed by the linear regression estimator. This results in an optimal configuration at the minimum values of both M_0 and N_1, in which an accurate estimate of the correction curve can be constructed from very few level 1 solution samples. Moreover, if the approximation is correct up to floating point precision then this scenario will hold for practically any choice of observation noise parameter, since the observation noise will not affect the accuracy of the approximation (though we remark that it will still affect the behaviour of both the BPF and the MLBPF).

This raises the question: can we not simply neglect the level 1 multilevel corrections and instead place almost all of the particles on level 0, where the solutions are close to perfect? In some experiments the answer may be yes; however, in theory, if there exists any x_n such that

$$|r(x_n) - (h^1(x_n) - h^0(x_n))| > 0,$$

no matter how small the difference, then asymptotically the estimates based

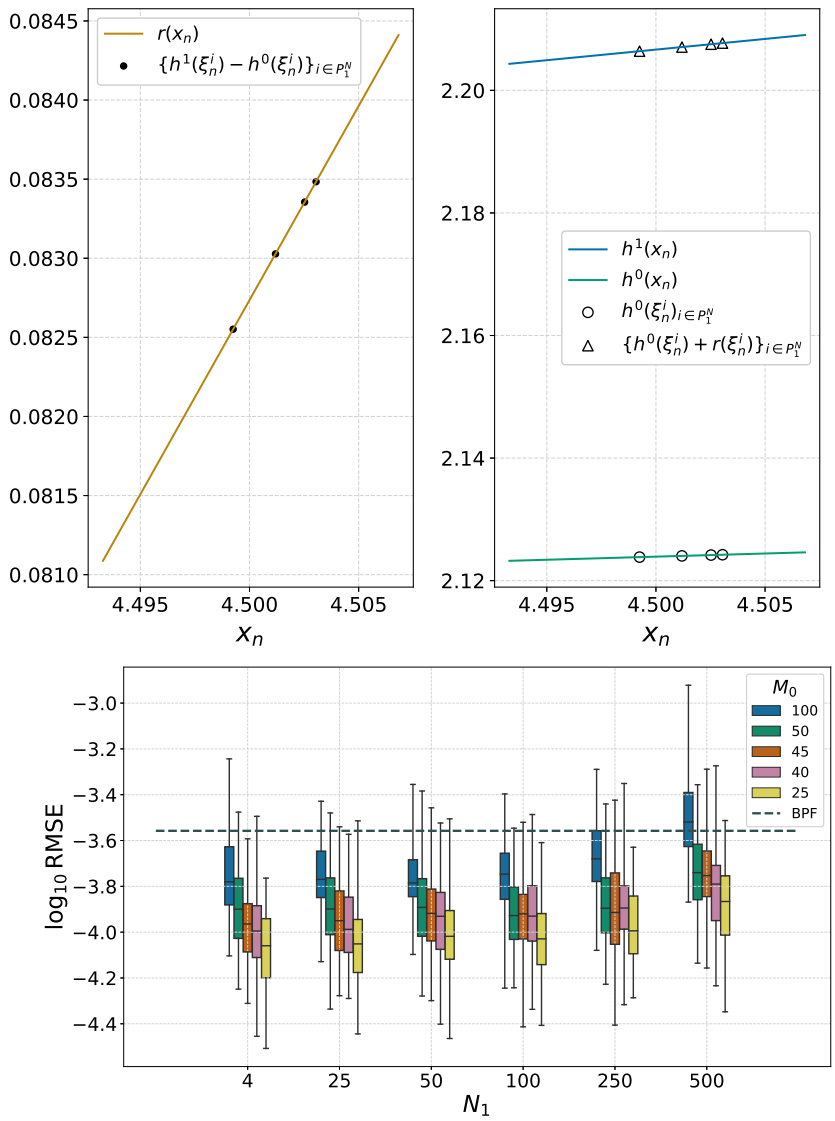

Figure 4.3: Top: One iteration of the overall $N_{\text{length}} = 25$ generated data, in which the level 1 correction data (left) is almost exactly linear and hence essentially perfectly approximated by $N_1 = 4$ particles. Hence the corrected level 1 solutions $g^0(\xi_n^i)$ lie almost exactly on the true level 1 solution curve (right). Bottom: The corresponding experiment results over $N_{\text{data}} = 10$ data sets for $\sigma_X = 0.02$, $\sigma_Y = 0.005$, in which the optimal configuration is at the smallest values of M_0 and N_1.

only on the level 0 improved likelihoods

$$\widetilde{g}_n^0(x_n) \propto \exp\left(- \frac{(h^0(x_n) + r(x_n) - y_n)^2}{2\sigma_Y^2} \right)$$

will not converge to the true solution, since $h^0(x_n) + r(x_n) \neq h^1(x_n)$. Even in practice, it is not out of the realms of possibility that a particular application may require enormous amounts of particles or that forecasts are sought of extreme events in which accurate estimates of the distribution tails are required. In both of these scenarios any discrepancy in the regression approximation again becomes relevant and emphasises the importance of the multilevel correction terms.

4.3 PDE-based applications

Having established the practicalities of ODE based-applications, we now consider the more complicated case of state estimation based on observations that are modelled as the solution of a PDE. This introduces an added element of uncertainty since, unlike the ODE case in which only the signal is dynamic, the surrounding system is now also changing in a way that is both time and signal dependent. On the other hand, accurate knowledge of this system is required in order for it to be a reliable means to estimate the signal. Taking the shallow water equations model (4.1.1)–(4.1.4) for example, the wave height and velocity are governed not only by the conservation laws corresponding to $Z = 0$, but also by the topography of the seabed which is the Markov process we seek to estimate. Consequently, the wave solution at each $n > 0$ is unknown and must be estimated alongside the signal.

In a similar vein to the regression approximation technique discussed in the previous section, the following is a general consideration within the context of particle filtering on PDEs and is not specific to MLBPF. As an acknowledgement to its presence in PDE-based applications, irrespective of the choice of method, we discuss it first.

4.3.1 General evolution of PDE solutions

The task of evolving a numerical approximation of a PDE solution $\rho(z, t)$ within the SMC framework is one that seems relatively unexplored. There are again two sources of error in the PDE estimate: the numerical error

from the discretisation, and the uncertainty of basing this numerical approximation on random estimates. In the same way that the observation noise in general prevents the true distribution of X_n from ever being known, the same is true of $\rho(z,t)$ and therefore of any numerical approximation $\rho_n^M := (\rho_n[j])_{j=1}^M$, where j is the space discretisation parameter and n is the time index corresponding specifically to X_n. For consistency we therefore commute the common notation of a point estimate $\widehat{X}_n$ of X_n to the PDE solution paradigm and use the notation $\widehat{\rho}_n^M$ for a point estimate of ρ_n^M.

Since ρ_n^M depends on the signal values, this solution formally forms part of an augmented signal that is the true Markov process we seek to estimate, namely

$$X_n^{\rho,M} := (X_n, \rho_n^M). \tag{4.3.1}$$

Hence in addition to the original signal estimate, each particle is now assigned an approximate PDE solution that represents a sample of the unknown system state. Under this model, the weighting step therefore extends to quantifying the collective likelihood of these quantities. However, for the spatial resolution of ρ_n^M to be of any meaningful accuracy, the mesh parameter M and hence the dimension of the augmented signal space will generally be large. On the other hand, particle filters are known to generally not perform well in such scenarios due to the curse of dimensionality [17, 61], essentially ruling out the ability to run any such algorithms on this full space. Moreover, in the multilevel scenario the level-specific augmented particles according to (4.3.1) would be defined in altogether different spaces, making resampling over the entire sample set impossible without performing some kind of transformation to the particles; e.g. an interpolating mapping. For large sample sizes, this could impose a computational burden that makes the resulting SMC algorithm infeasible.

A sub-optimal but more practical approach is to construct a point estimate $\widehat{x}_n$ of the true hidden state value x_n. This is then used to form $\widehat{\rho}_n^M$ by running the PDE solver with parameter value $\widehat{x}_n$ and initial condition $\widehat{\rho}_n^M$ to give $\widehat{\rho}_{n+1}^M$. In our experiment we use the empirical mean as the point estimate, which is optimal with respect to the mean loss cost function [62]. While this choice of estimator approach worked for the convection-diffusion experiment, it is unlikely to do so for all applications and could be an area of improvement in future work. Its deficiency arises from the fact that, not

only are we assigning all of the particles ξ_n^i the same initial condition in the manner of (4.3.1), but we are also performing inference on only one component of the true, high-dimensional signal. Both of these features have the potential to induce estimates that diverge.

One possible improvement is to fix a $k \in \mathbb{N}$ and, at some $n \geq k$, instead perform inference on the lag signal

$$X_n^k = (X_{n-k}, \ldots, X_n). \tag{4.3.2}$$

For $n < k$ the particle filter proceeds by solving ξ_n^i for $i = 1, \ldots N$ with respect to $\widehat{\rho}_n^M$ in the manner described above and appending each mutated particle ξ_{n+1}^i to N vectors of length not exceeding that of (4.3.2). Once $n \geq k$, the weight w_n^i of ξ_n^i is computed by solving with respect to $(\xi_p^i, \widehat{\rho}_p^M)$ for $p \in \{n-k, \ldots, n\}$, with the idea that the inclusion of the partial historical path provides more information about the unknown initial condition at time n. At time n the marginal particles and their weights (ξ_n^i, w_n^i) can be used to estimate $\widehat{\pi}_n$, while as more particles are appended as a result of mutation, the oldest particles are removed in order to maintain a stable level of complexity within the algorithm. This lag-based method was implemented to provide a working particle filter on the convection-diffusion model, but seemed to exhibit no immediate benefits over the filter implementation. One possible explanation for this could be that there is minimal transience in the PDE initial condition, meaning that retaining the historical lag data in this instance brings no added benefits to the estimates. However, this could be a feature that is specific to the convection-diffusion model, and the approach could potentially be of more noticeable benefit in other PDE applications.

Another potential improvement could be to design an importance distribution in the style of (2.2.8), in which the latest observation is incorporated to produce mutated particles and hence a point estimate solution $\widehat{\rho}_n^M$ that are collectively more likely under the observed measurement. One challenge of implementing this design in the context of PDE-based observation models is that the presence of time in the system could result in there being multiple solutions to this inverse problem. As a simple example, for a PDE with periodic boundary conditions over $[0, 1]$ and solution $\rho(x, t) = \sin(2\pi xt)$, the same (noise-less) observation is obtained whenever two particles ξ^1, ξ^2 are both integers. Consequently, there are infinitely many regions in $\mathbb{X}$ that the

importance distribution could propagate the particles towards. While in this simple example the previous state estimate could also be utilised to quantify the probability of these regions, in more complicated models such as the full shallow water equations that we discuss in Subsection 4.3.3, determining which regions are likely under these conditions may not be a straightforward task.

A likely reason for the success of the point estimate approach in the convection-diffusion experiment is that the accuracy between the PDE solution estimate and the true state is not overly sensitive to a moderate error between the point estimate and the true signal value. Additionally, the integration time $[0, T] = [0, 0.05]$ over which the PDE is solved at each iteration is small for this experiment. In practice this means that the system is being measured with high frequency, thus preventing too much uncertainty concerning the random PDE solution to develop. This could be a key principle required in order to justify applying the point estimation approach. It is also possible that the true underlying PDE solution is sufficiently intransigent that the model we have configured is essentially behaving like an ODE. Conversely, applications in which there is large estimate uncertainty or in which small errors in $\widehat{x}_n$ induce vastly different likelihood functions are likely to require a different approach in order for an SMC method to be operational.

4.3.2 MLBPF approach and results for PDE model

Another PDE-related consideration this time that is MLBPF-specific is that, in addition to the estimate of the unknown system solution, there is a need to solve a PDE on each of the $L + 1$ levels in order to generate the level-specific likelihood terms. To do so requires $L + 1$ numerical approximations $\widehat{\rho}_n^\ell$ that act as an initial condition on each level. For the current application the most successful way to obtain these $\widehat{\rho}_n^\ell \in \mathbb{R}^\ell$ has been to perform linear interpolation at complexity $\mathcal{O}(L + \ell)$ on the top level solution $\widehat{\rho}_n^L$ once the system estimate has been constructed. For applications in which the spatial dimensions is greater than 1, multilinear interpolation provides a natural generalisation of this method to such settings [76].

In light of the proposed point estimate and interpolation steps, the resulting PDE-based MLBPF algorithm is provided in Algorithm 4. We note that the algorithm captures both the extension of general particle filtering on PDE applications and the considerations required for a multilevel approach.

In particular, the corresponding BPF algorithm is implicit in Algorithm 4 and may be obtained by setting $L = 0$. Similarly, if $(\widehat{\rho}_n^\ell[j])_j^{M_\ell}$ is constant for all n, ℓ, then Algorithm 4 essentially reduces to the original MLBPF algorithm that was applicable to ODEs.

Algorithm 4 MLBPF for PDE-based applications

```
% Initialisation
for i = 1, ..., S(N) do
    ξ_n^i ~ π_0 and w_0^i = 1
for 0 ≤ ℓ ≤ L do
    dz_ℓ = W/(M_ℓ − 1)
    for j = 1, ..., M_ℓ do
        ρ̂_0^ℓ[j] = ρ_0((j − 1) ∗ dz_ℓ)
for n ≥ 0 do
    % Calculate weights for each level
    for 0 ≤ ℓ ≤ L do
        for i ∈ P_ℓ^N do
            w̃_n^i = N_ℓ^{−1}(g_n^ℓ(ξ_n^i; ρ̂_n^ℓ) − g_n^{ℓ−1}(ξ_n^i; ρ̂_n^{ℓ−1}))w_n^i
    % Signed resampling
    for i = 1, ..., S(N) do
        ξ̂_n^i ~ (Σ_{i=1}^{S(N)} |w̃_n^i| δ_{ξ_n^i}) / (Σ_{i=1}^{S(N)} |w̃_n^i|) and ŵ_n^i = sgn(Σ_{i=1}^{S(N)} w̃_n^i I[ξ̂_n^i = ξ_n^i])
    % Compute the hidden state point estimate
    x̂_n = (Σ_{i=1}^{S(N)} ŵ_n^i ξ̂_n^i) / (Σ_{i=1}^{S(N)} ŵ_n^i)
    % Mutation
    for i = 1, ..., S(N) do
        ξ_{n+1}^i ~ K(ξ̂_n^i, ·) and w_{n+1}^i = ŵ_n^i
    ρ̂_{n+1}^L = PDE_SOLVE(ρ̂_n^L, x̂_n)
    if L > 0 then
        for 0 ≤ ℓ ≤ L − 1 do
            ρ̂_{n+1}^ℓ = INTERPOLATE(ρ̂_{n+1}^L)
```

With a slight abuse of notation we have included in Algorithm 4 the dependency on the level specific PDE solution estimate $\widehat{\rho}_n^\ell$ in the likelihood terms g_n^ℓ. In the interpretation of the full signal (4.3.1) this would be unnecessary since this (particle-specific) solution would be encoded into ξ_n^i, though in this case Algorithm 4 would not be valid in its current form due to the need to alter the resampling step. With our current approach, explicitly de-

noting the dependency on $\widehat{\rho}_n^\ell$ seems appropriate, since the construction and application of these estimates is a significant part of the algorithm.

For the convection-diffusion application, we run Algorithm 4 according to the specified numerical experiment and model description in Section 4.1 and the parameters in Appendix B. In the same manner as the steady state shallow water equations experiment, the results are plotted in Figure 4.4. Due to the substantially higher cost of generating a PDE solution with respect to each particle, in this experiment we have run the BPF with a much smaller N_{BPF} than in the steady-state shallow water experiment, meaning that the range of values of N_1 that the MLBPF can take while remaining within the specified computational budget is also smaller. However, we see that the curvature from the previous experiment is still present, though on this occasion we do not have greater accuracy for all choices of level 0. In particular, we see $M_0 = 10$ exhibits too much numerical error for the MLBPF distribution estimate to surpass the BPF in accuracy for any choice of N_1 and the corresponding N_0. However, Figure 4.4 clearly illustrates several configurations for which the length-averaged MLBPF empirical distributions are more accurate in RMSE than the BPF over 75% of the time.

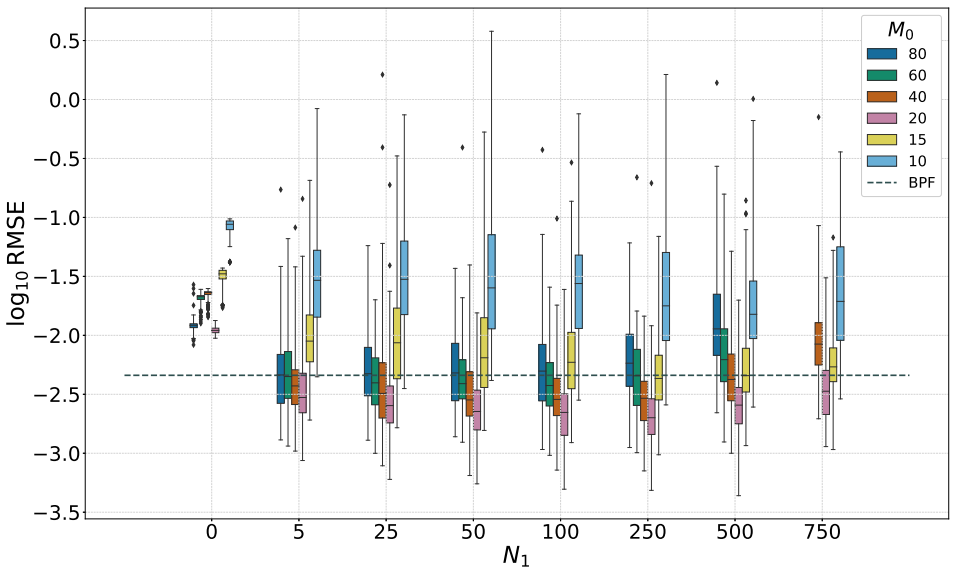

Figure 4.4: MLBPF RMSE boxplots vs. BPF median RMSE for the convection-diffusion experiment, in which the smallest median error is obtained with the configuration $(M_0, N_1) = (20, 250)$.

As was the case in the shallow water experiment, the accuracy of the configurations corresponding to larger values of M_0 only ever worsen with N_1 due to the MC error being the dominant term. There is also clear global curvature for the mid-range choices of M_0, in which the balance between the numerical and MC error is played out as the sample allocations are adjusted. As expected, the highest performing configurations are found within these choices of M_0, with $(M_0, N_1) = (20, 250)$ being that which provides the smallest median RMSE. Applying the relative gain formula (4.2.1) with this configuration there is in this case an 11.6% relative gain in accuracy offered by the MLBPF; a substantial improvement over the steady-state shallow water equations model. This is likely because as the solution complexity of a problem grows, the cost savings of multilevel schemes increases, making the MLBPF likely a particularly favourable choice over the BPF in settings where the spatial component of the underlying PDE is multi-dimensional. For a time comparison, the BPF was unable to reach the RMSE accuracy obtained by the MLBPF on the first data set for the specified number of reference particles ($= 10^5$).

Aside from the application numerical/MC errors that are assimilated into the particle estimates, there are now three additional sources of error arising from implementation considerations in Algorithm 4: the regression approximation improvement, the global point estimate of the underlying PDE solution, and the error that lower level PDE initial conditions inherit from the linear interpolation. Similarly to the motivation behind keeping the regression improvement simple, linear interpolation is not the most advanced method of inferring a coarse solution from a finer one — particularly in the context of polynomial or spline interpolation — yet is sufficiently cheap and accurate enough for the task at hand. Other than more advanced interpolation methods, it is not obvious if there are any alternative ways to approach this task. For example, updating each $\widehat{\rho}_n^\ell$ using the system point estimate is infeasible not only because of the substantially higher cost of solving $L+1$ PDEs with spatial resolution M_ℓ over $[0, T]$, but because $\widehat{x}_n$ is an estimate with respect to the full accuracy solution $\widehat{\rho}_n^L$ and is therefore biased for any $\ell < L$. In practice, this approach again leads to lower level solutions that are so far from the full accuracy solution that they are assigned weights that are essentially zero, hence yielding poor estimates.

In reference to this "drift" property of lower level solutions, the computed

time increment Δt that is specific to a numerical method can be problematic within the PDE-specific MLBPF. Firstly, by the cost-accuracy assumption, it is crucial that Δt does not increase as the space increment Δz is decreased. In the case of a globally stable numerical method like the Crank-Nicolson method satisfying this condition is trivial, since the space and time resolutions can be set independently of one another. However, there are many numerical PDE methods in which the time increment is computed as a function of both Δz and the numerical approximation $(\widehat{\rho}[j])_{j=1}^{M}$ for which this level of control of Δt is no longer in the hands of the user. For example, the well-balanced finite volume method of [53] is an advanced computational fluid dynamics solver that adaptively computes the time increment based on the maximum and minimum wavespeeds over the entire solution. Consequently, it is no longer obvious if less accurate solutions are obtained at cheaper cost, which is a cornerstone of the multilevel methodology. Intuitively, small time increments arise when the solution is particularly "complex" (i.e. the local flux is substantial, or a discontinuity develops), meaning that Δt for the simplest class of constant solutions would be the maximum. However, it is not clear how Δz affects the solution wavespeeds and hence the relationship it has to Δt, nor if such a result is obtainable.

On a slightly related matter, it is also important for consistency in the bias corrections that all of the likelihood terms are based on PDE solutions computed at the same stopping time T, so that the only source of numerical inaccuracy arises in a spatial sense. However, if Δt is level-dependent then this may not be the case, since some schemes may "overshoot" T by a greater margin than others. The well-balanced finite volume solver of [53] is one example of this. One simple way to prevent this is to compute the minimum Δt over all levels and apply this globally as a time increment. Alternatively, it may be more efficient to customise the remaining timestep on each level at the penultimate spatial iteration based on the distance to T, thus ensuring each multilevel PDE solver is terminated exactly at T.

4.3.3 Implementation on additional models

We conclude this section and chapter with a brief discussion on some other PDE applications that are either not suited to the MLBPF or impose general particle filtering challenges that require further investigation.

A simple traffic flow model can be obtained from the convection-diffusion

equation by setting $D = 0$ implicit in (4.1.8), thereby removing the diffusion term and obtaining the transport equation with source term S. If $S \equiv 0$ and the traffic is light enough that its speed is independent of its density, then the equation models traffic density moving from left to right at constant speed v_0 and the solution is (4.1.13). However, if $S(z, t, X_n) := X_n \delta_{z^\perp}(z)$, then the resulting equation

$$\partial_t \rho + v_0 \partial_z \rho = X_n \delta_{z^\perp}(z) \tag{4.3.3}$$

models the same problem but with traffic entering the road with flux X_n at a junction located at $z = z^\perp$, and no longer has a closed-form solution [47, Ch. 11, Ch. 17].

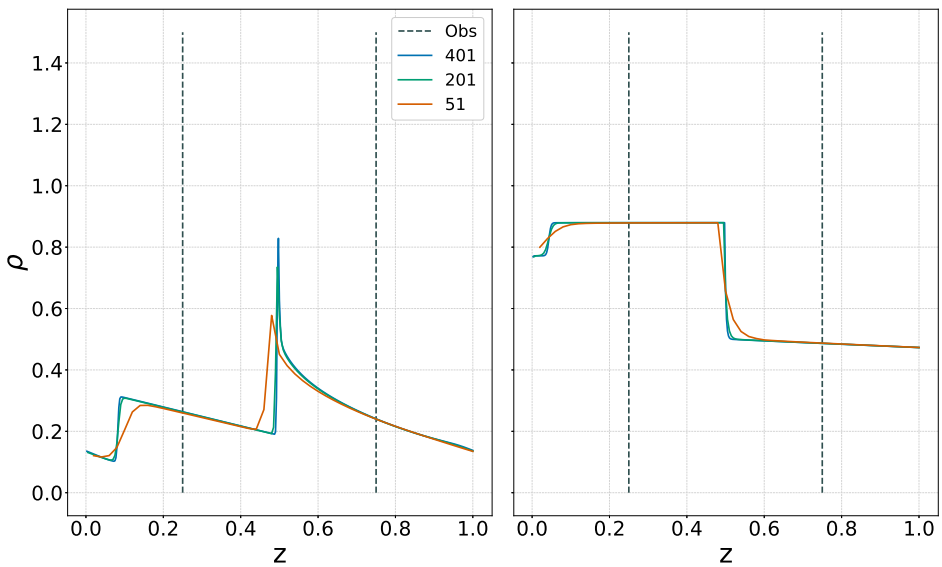

Figure 4.5: Left: The flux term $x_n = 0.0881$ is sufficiently small that a queue is not formed after the junction located at $z = 0.5$. Right: Here $x_n = 0.101$ is too large that the added influx cannot be assimilated into the traffic without increasing the traffic density before the junction.

In Figure 4.5 we plot numerical solutions using the Lax-Friedrichs method [47, Ch. 4] for 2 different values of a random walk signal process X_n such that $z^\perp = 0.5$. On the left panel $x_n = 0.0881$ for $n = 1$ and on the right $x_n = 0.101$ for $n = 14$. The nature of the solutions differ depending on the magnitude of the flux term: for sufficiently small values of X_n the carrying

capacity of the road is not exceeded and the influx of traffic from the junction does not leave a queue in its wake. In contrast, for larger values of X_n the road is unable to assimilate the added load and a traffic jam forms before the junction as a result.

These two situations are described respectively by the traffic density solutions in the left and right panels in Figure 4.5. Because of this solution behaviour, a two-dimensional observation space based on the density before and after the junction is required in order for all values of X_n to be inferred from the data. However, in both scenarios we see that reductions in the solver mesh size does not lead to a loss of accuracy at the locations where the observations are taken, thus leading to a violation of the multilevel cost-accuracy trade-off condition. In fact, non-linearity is only truly imparted on the solution via temporary "shockwaves" that arise as the in/outflux of traffic at the junction is transmitted at a finite speed along the rest of the road; otherwise, the long term solution given a fixed $X_n = x_n$ is essentially a piecewise continuous function that is equally well approximated by a solution based on a small number of mesh points as it is by one on many. This propagation of information at a finite speed is a characteristic property of hyperbolic PDEs that make them suited to modelling wave problems. In the event that an observation was taken as a shockwave passed, a reduction in spatial mesh size would indeed lead to a loss of accuracy and the MLBPF would be applicable. However, mathematically determining when this information arrives requires knowledge of the characteristic lines of (4.3.3), which in general may be difficult or even impossible to compute.

While the traffic model in its current form is not a suitable application for the MLBPF, the full shallow water equations model (4.1.1)–(4.1.2) presents a different kind of challenge to SMC methods in general. For this application, we consider the scenario where the topography is described by the function

$$Z(z, t; X_n) = \max(0, 0.25 - 0.05(z - X_n)^2), \tag{4.3.4}$$

which is a symmetrical seabed "bump" of maximum height 0.25 centered at $X_n = x_n$, where X_n is the random walk signal process we seek to estimate. This is the topography function for the *drain on a non-flat bottom experiment* [53], from which we apply the same initial and boundary conditions. The experiment is given this title because the fluid is allowed to exit the system

at the right-most boundary, while the left boundary is taken to be a solid wall (i.e. has a no-slip boundary condition). Under these conditions and a fixed signal value, this PDE has a steady long-term solution in which the fluid settles at the height of the bump to its left and is emptied according the geometry of the bump to its right. If this situation is allowed to develop then solutions will inherit the same type of simplicity as those in the traffic model. Therefore, to further agitate the system alongside the signal process, we modify the left boundary height condition to include a deterministic, time-dependent oscillation term of the form $\sin(t) + 1$.

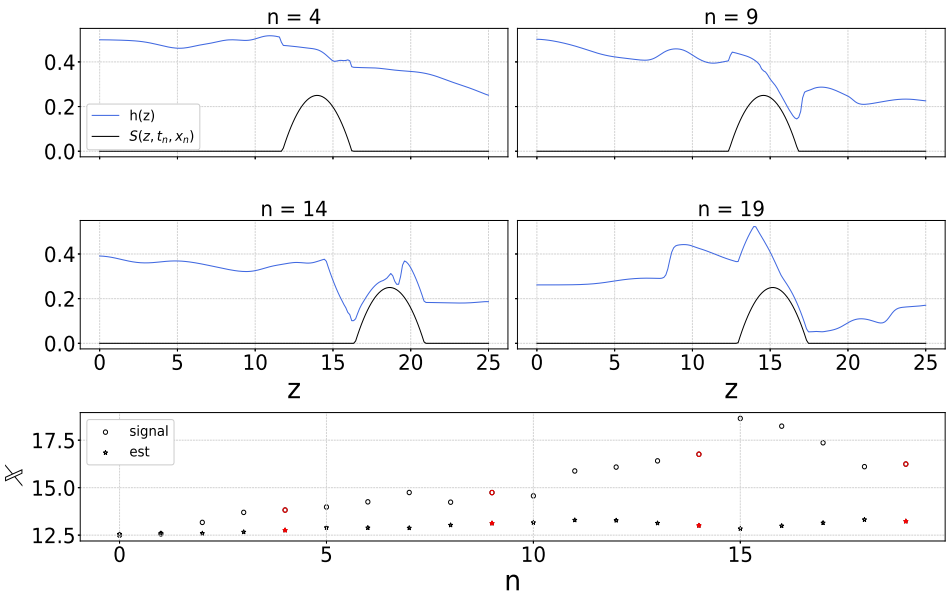

Figure 4.6: Top and middle panels: shallow water equations solution heights and corresponding random walk topography profiles at various sampled iterations. Bottom panel: Generated signal process and corresponding BPF mean estimates, with sampled values highlighted in red. The BPF estimates are poor due to the lack of information provided by the observations.

To approximate solutions of the shallow water equations with respect to the parametrised topography function (4.3.4) we use the well-balanced finite volume solver of [53] with a second-order MUSCL extension and measure the resulting water height at the right outlet boundary according to the observation model (4.1.4). In Figure 4.6 we plot a selection of topography bumps generated from the signal process and the corresponding shallow wa-

ter equation solutions at each of these times. In the lowest panel, we plot mean estimates of the signal generated using the BPF with $N_{\text{BPF}} = 500$ particles, and have highlighted in red the signal values and estimates corresponding to the sampled water height/topography solutions.

The issue with this application is that the BPF is not able to discern between the general chaotic behaviour of the water system and the effect the signal has on the outlet water height via the topography bump. Consequently, the resulting estimates do not track the signal in any meaningful way, since the outlet observations do not leverage enough information about the signal values they model. In the context of our current approach to filtering on PDEs this is particularly problematic, since the PDE solution estimate required at each iteration is based on these signal estimates. Therefore if trends in the signal are not detected then the overall performance of the method quickly deteriorates.

There are several possible ways in which the shallow water equations experiment could be made feasible, such as:

- Modifying the topography function so that changes in the signal process have a more substantial and hence identifiable impact on the wave solution at the point of observation.
- Designing an observation model that is better equipped to detecting the effects of the signal, either by taking observations at more optimal locations or using a more sophisticated method of measurement.
- Refining the boundary conditions so that the system is less chaotic but still complex enough that accurate likelihood evaluations are more expensive to produce than ones that are less accurate.
- Implementing the lag signal extension approach of (4.3.2) to remove some of the uncertainty of the PDE estimate.

It is worth noting that all of these points relate to modelling or general PDE-based particle filtering considerations for this particular shallow water application and are not specific to the MLBPF. Indeed, in light of the accuracy gains we have observed in the convection-diffusion model, it is feasible that the MLBPF has the capacity to be operable in situations where the BPF is not. Furthermore, the traffic model and the non-flat drain shallow water equations applications are in some sense at opposite extremes of

the spectrum in terms of system chaos. The current issue with the traffic model arises essentially because of the lack of agitation in the system, while for the shallow water model the excessive chaos makes the successful implementation any particle filtering method a difficult task. In this sense the modelling considerations of an application have a greater impact on whether the MLBPF and SMC methods in general are feasible. From what we have observed in the numerous experiments we have conducted in this thesis and in [31], it is generally the case that if the BPF can be implemented on a model, then not only is the same true for the MLBPF, but it is able to do so more accurately if it is configured properly. To this end, tasks such as identifying this optimal configuration are those which are MLBPF-specific, while complications of the type arising in this section should be viewed instead as more general SMC/mathematical modelling problems. Furthermore, while for the purposes of the drill-system application we have studied the MLBPF in the context of differential equations, one of the strengths of the method is the flexibility in its cost-accuracy assumption and hence the applications it can be applied to. With this in mind, the empirical benefits and advantages it has offered in the differential equation models we have studied here also have the capacity to transfer to a wide range of other applications.

Chapter 5

Conclusions

The goal of this thesis has been to provide a novel academic contribution to approaching parameter estimation problems of the type that arise in online drill system applications. These parameters vary over time in a way that is essentially random and are only partially observable through measurements that are also subject to uncertainty. A natural approach has therefore been to adopt a Bayesian viewpoint and model the problem using a Markov state-space model of the type that arise frequently in signal processing applications and nonlinear filtering problems [54]. This choice of model is also favourable to applications such as hydrocarbon drilling because the fluid dynamics models that mathematically map the parameters to the observations also fit naturally into the state-space model framework. In the presence of such nonlinear mappings, sequential Monte Carlo methods can be applied to leverage approximate solutions that are often more accurate and reliable than those produced by alternative methods. For this reason we pursued SMC methods as a research direction, and sought to develop a novel approach within this field.

With a view to translating the complexity savings of multilevel Monte Carlo to the SMC setting, we have derived a multilevel bootstrap particle filter that applies the multilevel "telescoping sum/linearity" technique to the integral operators in the filtering update step. The idea behind this has been to reduce the overall cost of computing all of the particle weights at full accuracy by instead assigning portions of the particles to levels that are based on inaccurate solutions but which have a cheaper weight generation cost. The role of the levels in the MLBPF context is to then correct the

resultant bias from the previous level until the combined estimates represent the full accuracy model. One defining property of this approach within the context of SMC is that the level-specific distributions are in general now signed measures, which in practice leads to weights which are potentially negative. Consequently, the MLBPF algorithm has had to be designed to accommodate this negativity in order to provide a method that is practically feasible. In particular, the method of resampling a particle with probability proportional to its weight is no longer possible, which has a profound impact on the rest of the algorithm.

For the empirical filter and prediction measures resulting from the MLBPF we have proved convergence to the exact measures in the sense of a strong law of large numbers and a central limit theorem. These results provide a theoretical basis for the use of the MLBPF with any number of levels, particle allocation or likelihood accuracy hierarchy that are valid within the very general assumptions of the algorithm. This is a weaker result than that presented in [15, 29], in which convergence was already ensured via standard MC theory, and the true contribution was how to apply the MLMC estimator in a way that theoretically guarantees cost savings over the standard MC estimator, independent of the application. The comparison between these theorems and our own is somewhat harsh however, since it is assumed implicitly in MLMC literature that the expectation is taken with respect to one particular functional; a European call option, for example [29]. Therefore,

$$\sum_{i=1}^{N} w^i \varphi(\xi^i) - \int_{\mathbb{X}} \varphi(x) dx$$

is a scalar-valued random variable which has a well-defined meaning when taking the expectation. In contrast,

$$\pi^N(\varphi) - \pi(\varphi), \qquad \varphi \in \mathscr{B}(\mathbb{X})$$

is a weak form probability measure, for which the expectation and hence RMSE has no meaningful interpretation. Instead, a feasible measure-based metric in the case where $\mathbb{X} = \mathbb{R}$ is the Kolmogorov-Smirnov distance (see e.g. [67, Ch. 3]) in which the distance between two distributions is measured by the supremum distance between the respective cumulative distribution functions. For more general forms of $\mathbb{X}$ the maximum mean discrepancy

measures the distance between two probability measures via an embedding into a reproducing kernel Hilbert space [30]. One advantage of using the RMSE in the MLMC derivations, we recall, was that the sample size and resolution parameter could be used to both bound the RMSE and capture the cost of the estimator, thus establishing a connection between the error and the associated number of floating point operations. For an analogous approach to work in the SMC setting would require a similar deconstruction of the Kolmogorov-Smirnov distance or maximum mean discrepancy into measure-based error terms that are controlled by parameters that also capture the cost. This is likely to be a challenging if not intractable task, since the measure distance mappings are more complicated than the RMSE and therefore more difficult to relate to the complexity of the filter. Were such a result to be achieved, it would be a valuable addition to the current progress that has been made as it would enable the type of cost analysis used in MLMC to be applied within SMC.

In the absence of any formal a priori knowledge about the optimal MLBPF configuration, an approximation can be computed by experimentally iterating over a collection of particle allocations/lower level likelihoods and measuring the resulting accuracy with respect to some metric. For all intents and purposes, this essentially restricts our current version of the MLBPF to few (if any) more than two levels, since as a consequence of the "brute force" search we currently deploy the number of degrees of freedom grows rapidly once additional levels are introduced. In the MLMC setting this problem is simplified by virtue of the availability of an optimal level specific sample allocation formula $N_\ell \simeq \sqrt{\widehat{\sigma}_\ell^2/\mathcal{C}_\ell}$ and the ability to test for convergence according to the criterion $\widehat{\sigma}_L^2 \simeq M^{-\alpha}$ as described in Algorithm 2. In particular, the convergence test means that intermediate levels $0 < \ell < L$ need only be introduced where necessary, while the allocation formula allows the level-specific optimal particle allocation to be calculated once each level is defined.

To study the global behaviour of the MLBPF in our experiments we have iterated over the entire "search space"; if the task is instead to locate only the optimal error minimiser, then it is highly plausible that in practice there is a more efficient means of finding it than via a global search. In the two level setting, one possible approach could be to iterate through the points on the lattice $(M_{0,k}, N_{1,j})$ where $k \in \{1, \ldots, N_{\text{mesh}}\}$ and $j \in \{1, \ldots, N_{\text{allocs}}\}$ in

the following way. Assuming that $M_{0,k}$ and $N_{1,j}$ are indexed such that they increase in k and j respectively, and that $N_{1,j} > 0$ for all j, then by initialising at the point $(M_0, N_1) = (M_{0,1}, N_{1,1})$, we guarantee that the MC error is almost surely minimised on account of the fact that N_0 and hence $N_0 + N_1$ is globally maximised. If the experimental errors at the points $(M_{0,2}, N_{1,1})$ and $(M_{0,1}, N_{1,2})$ are both greater than that produced at $(M_{0,1}, N_{1,1})$ then we can already terminate the search, since we know that the gain of any kind of bias reduction — be it from greater level 0 accuracy or a more accurate estimate of the level 1 correction distribution — does not outweigh the overall loss of accuracy induced by increasing the MC error. However, if $(M_{0,1}, N_{1,1})$ does not produce an error less than both of $(M_{0,2}, N_{1,1})$ and $(M_{0,1}, N_{1,2})$, then the configuration giving the minimum error can be selected as the new optimum and the process is repeated with respect to this point. A simple but valuable future research task is to implement this method on the numerical experiments of Chapter 4 and verify that it produces the same lattice-specific optimum as the brute force approach implemented in this thesis. Furthermore, if feasible in the two level case, the method could provide a pathway to practically extending the MLBPF to more than two levels.

We have run the two level MLBPF on numerical experiments in a number of varied applications — two of which are in [31] and two of which are presented in this thesis — and in each case have managed to identify several configurations that elicit a more accurate performance from the MLBPF than the BPF. Where these configurations are located on their respective planes (we note that for one experiment in [31] the only degree of freedom is the particle allocation) depends heavily on the experiment hyperparameters. One influencing factor that is common to all of the applications is how well approximated the true correction curve is by the linear regression accuracy improvement. If it is essentially perfectly approximated for any N_1 on a given level 0 solution, then only the MC error will remain and the optimal particle configuration (N_0, N_1) will be that for which N_0 is largest and hence N_1 is smallest. If this is true for all level 0 solutions, then for the same reason the optimal configuration will be at the cheapest level 0 solution, again at the smallest N_1. Similarly, the behaviour of the MLBPF is also governed by the experiment signal-to-noise ratio. In practice the signal and observation noise terms will be fixed, meaning that the optimal level 0 solution

must be chosen so that its distance from the level 1 solution in the $\mathbb{XY}$-plane accommodates the noise terms in an optimal way. Even with this choice, the optimal particle allocation remains unknown a priori, which in turn depends on the statistical precision of the problem; i.e. the value of N_1 when $N_0 = 0$. All of these considerations serve to highlight the many intricate co-dependencies between the application and the MLBPF and the complexity of how to even experimentally determine an optimal configuration.

While the main contribution of this thesis has been the MLBPF, by directing our focus towards PDE-based applications in particular, we have also devised a general approach to particle filtering on these models. Applications such as the drill-system model are in general more challenging because the PDE solution that is used for inference also depends on the parameters. Formally, this means every component of the approximate PDE solution is an unknown parameter that we seek to estimate, making the problem high-dimensional. With SMC methods generally being ill-suited to such applications, we have instead devised a sub-optimal approach in which an approximate PDE solution is evolved with respect to a point estimate of the hidden parameter value and used as a global initial condition for all of the particles. Along with an interpolation step, this was used to construct a PDE-specific version of the MLBPF that obtained greater accuracy than the BPF on the convection-diffusion model. The likely reason for the functionality of the point estimate approach within this application is that the global PDE approximation is not overly sensitive to the point estimate value it is generated by. In the case of the full shallow water equations application, it is currently unclear whether the point estimate approach is the source of inadequacy or if the model in its current guise is simply ill-posed. We note that the limitations and implementation issues of the point estimate approach to estimating the PDE state are general particle filtering considerations and non-specific to the MLBPF; indeed, it could be that the empirical accuracy gains we've seen from the MLBPF makes the method feasible in applications where the BPF is not. The full shallow water equations model is a particularly good case study for future work because it is mathematically analogous to the Euler equations PDE used in the drilling application, while also exhibiting a physical intuition that facilitates further study of the respective particle filters. If any experimental gains of the type discussed can be made with the MLBPF, then the principles of how this performance was attained

should also be directly transferable to the drill-system application.

A hallmark of particle filtering methods is that they can be virtually endlessly tuned and customised to attain heightened performance, and the MLBPF is no exception. One of the primary components of our algorithm that could be further explored is the resampling step. In its current guise, resampling is performed with respect to a measure proportional to the empirical total variation measure that arises when instead considering the Radon-Nikodym derivative of the total variation measure. Our resulting algorithm with respect to this resampling scheme is well-defined, in the sense that the approximating measures at each iteration are still comprised of properly weighted samples as per [50], except in our case the samples are now components of a signed measure. The intuition behind our choice of resampling method is that (i) by taking the absolute value of the weights, the resampling probabilities again have a well-defined meaning, and (ii) particles are more likely to be resampled from "rich" areas of the total variation measure, i.e. high density, which is analogous to the intuition of multinomial resampling within the BPF. As was discussed in Chapter 3, this leads to a resampling scheme that can be viewed as one that samples according to $|\pi^N_{n,0}|$ as long as $\pi^N_{n,0}$ is an accurate approximation to $\pi_{n,1}$, then favouring particles that correct the level 1 telescoping sum error when $\pi^N_{n,0}$ is not accurate, in a way that approximately reflects the weighted density of the signed correction measure $\pi_{n,1} - \pi_{n,0}$, and so on. Under this scheme, we both resolve the issue of the weight negativity in the context of resampling and obtain an MLBPF algorithm for which the particle approximations converge to their respective measures. While our approach is both operational and theoretically sound, given its relative novelty it seems possible some improvements or an alternative approach could be explored. Any mapping that could somehow resolve the presence of weight negativity would provide a credible alternative to our resampling method, but it would need to be implemented in a way such that the asymptotics of the signed empirical measures are preserved. In particular, simply rejecting particles with negative weights is not valid, while from a practical perspective, iteratively incrementing weights in a manner analogous to [23] is likely to be too expensive in the applications we consider.

Another development of our method is that of implementing a more general importance sampling step, thus extending the MLBPF to a multilevel SMC methodology. Such an extension would allow for greater flexibility in

the design of the algorithm, which in turn could be further beneficial to its performance. For example, the inclusion of an importance distribution that uses the latest observation to retrospectively propagate particles towards a more credible region could in general be a valuable addition, but even more so in the context of our approach to PDE-based observation models, in which an accurate estimate of the underlying PDE state is crucial to the performance of the algorithm. In the case of the former point a more general choice of importance distribution is already theoretically accounted for by virtue of the Feynman-Kac framework we have adopted, and is simply a case of modifying the choice of potential function in a way that continues to satisfy the assumptions we have made. The latter point is not as straightforward, since this involves solving an inverse problem with respect to a PDE operator. Even in the scenario that this problem has a unique solution (which is by no means guaranteed) such a solution could be difficult to determine in the presence of the system uncertainty and — in the presence of the PDE being intractable — too costly to solve efficiently on the fly, particularly in the sequential setting.

Beyond refining the current approaches deployed within the MLBPF, there are also potential options in which the method could be adaptively configured within an application. For example, consider the two-level case in an ODE-type application, and suppose there is a way of inferring the relative contributions of the MC and numerical errors to the overall error. In general we are not restricted to the same choice of level 0 mesh size M_0 at every iteration. Therefore if the numerical error is the more prominent term, then it makes sense to opt for a more pragmatic choice of M_0 as a means of returning an overall higher reduction of the total error. Conversely, a lower value of M_0 will be of greater benefit to the estimate accuracy if the MC error is the dominant term. Another potential research direction is that in which the MLBPF could be "collapsed down" to the BPF either periodically or with respect to some diagnostic analogous to the effective sample size in resampling. Since experiments have shown that the accuracy gains of the MLBPF over the BPF are generally substantial in the short term but tend to diminish over longer time horizons, this suggests that intermittently reverting back to the BPF could provide additional stability to the MLBPF that would prevent this drift effect. However, this procedure is not as straightforward as simply setting $L = 0$ in the MLBPF, since this does not deal with the issue of

what happens to the negative weights. In order to be feasible, the mapping of the weights in such a step would need to be correctly defined in a way that preserves the asymptotics, which on initial inspection is a non-trivial task.

More generally to particle filtering, we have in this thesis made some initial inroads into the problem of SMC methods on PDE-based applications, but these ideas are still embryonic and application-specific in their success. A more rigorous and stable approach to updating the initial PDE condition associated to each particle is a high priority. One option we have proposed is that of instead estimating the lag signal (4.3.2), which may be of more notable influence to models such as the full shallow water equations than the convection-diffusion application, in which the effects were negligible. It also seems likely that the point estimate approach in general is an area that can be improved, though it is not obvious how to find a middle ground between this approach and filtering on the full high-dimensional space. To this end, a dimensionality-reduction technique on the full space may be a fruitful means of making progress. Lastly, while in this thesis and in [31] we have conducted investigations into a diverse collection of models, we have yet to explore applications in which the signal space is multi-dimensional, nor differential equations for which the spatial domain or solution space are multi-dimensional. Given the relative increase in accuracy gains we've seen from the MLBPF between ODE and PDE-based applications, it is worthy of further investigation to see if the advantage of using the MLBPF continues to grow with the complexity of the problems. Far from suggesting any shortcomings, the richness of research directions with which the MLBPF can be further pursued serves to illustrate how our current set of results are likely only a modest insight into the full capabilities of the method.

Bibliography

[1] K. Abdelgawad, S. Elkatatny, T. Moussa, M. Mahmoud, and S. Patil. "Real-time determination of rheological properties of spud drilling fluids using a hybrid artificial intelligence technique". In: *Journal of Energy Resources Technology* 141.3 (2019).

[2] M. Abimbola, F. Khan, N. Khakzad, and S. Butt. "Safety and risk analysis of managed pressure drilling operation using Bayesian network". In: *Safety Science* 76 (2015), pp. 133–144.

[3] W. Aboussi, M. Ziggaf, I. Kissami, and M. Boubekeur. "A highly efficient finite volume method with a diffusion control parameter for hyperbolic problems". In: *Mathematics and Computers in Simulation* (2023).

[4] S.A. Adedigba, O. Oloruntobi, F. Khan, and S. Butt. "Data-driven dynamic risk analysis of offshore drilling operations". In: *Journal of Petroleum Science and Engineering* 165 (2018), pp. 444–452.

[5] M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking". In: *IEEE Transactions on Signal Processing* 50.2 (2002), pp. 174–188.

[6] M.A. Beaumont, W. Zhang, and D.J. Balding. "Approximate Bayesian computation in population genetics". In: *Genetics* 162.4 (2002), pp. 2025–2035.

[7] A. Beskos, A. Jasra, K. Law, R. Tempone, and Y. Zhou. "Multilevel sequential monte carlo samplers". In: *Stochastic Processes and their Applications* 127.5 (2017), pp. 1417–1440.

[8] P. Billingsley. *Convergence of probability measures.* John Wiley & Sons, 2013.

[9] V.I. Bogachev. *Measure theory volume 1 and 2*. Springer, 2007.

[10] D.L. Burkholder, B.J. Davis, and R.F. Gundy. "Integral inequalities for convex functions of operators on martingales". In: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 6.2. University of California Press. 1972, pp. 223–241.

[11] B.P. Carlin, N.G. Polson, and D.S. Stoffer. "A Monte Carlo approach to nonnormal and nonlinear state-space modeling". In: *Journal of the American Statistical Association* 87.418 (1992), pp. 493–500.

[12] X. Chen, C.P. Tan, and C. Detournay. "A study on wellbore stability in fractured rock masses with impact of mud infiltration". In: *Journal of Petroleum Science and Engineering* 38.3-4 (2003), pp. 145–154.

[13] N. Chopin. "Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference". In: *Annals of Statistics* (2004), pp. 2385–2411.

[14] A.J. Chorin and J.E. Marsden. *A mathematical introduction to fluid mechanics*. Vol. 3. Springer, 1990.

[15] K.A. Cliffe, M.B. Giles, R. Scheichl, and A.L. Teckentrup. "Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients". In: *Computing and Visualization in Science* 14.1 (2011), pp. 3–15.

[16] D. Crisan and A. Doucet. "A survey of convergence results on particle filtering methods for practitioners". In: *IEEE Transactions on Signal Processing* (2002), pp. 736–746.

[17] F. Daum and J. Huang. "Curse of dimensionality and particle filters". In: *2003 IEEE Aerospace Conference Proceedings*. Vol. 4. IEEE. 2003.

[18] R. Douc and E. Moulines. "Limit theorems for weighted samples with applications to sequential Monte Carlo methods". In: *ESAIM: Proceedings*. Vol. 19. EDP Sciences. 2007, pp. 101–107.

[19] A. Doucet, N. De Freitas, and N.J. Gordon. *Sequential Monte Carlo methods in practice*. Vol. 1. Springer, 2001.

[20] A. Doucet, S. Godsill, and C. Andrieu. "On sequential Monte Carlo sampling methods for Bayesian filtering". In: *Statistics and Computing* 10.3 (2000), pp. 197–208.

[21] S. Elkatatny. "Application of artificial intelligence techniques to estimate the static Poisson's ratio based on wireline log data". In: *Journal of Energy Resources Technology* 140.7 (2018).

[22] H.H. Elmousalami and M. Elaskary. "Drilling stuck pipe classification and mitigation in the Gulf of Suez oil fields using artificial intelligence". In: *Journal of Petroleum Exploration and Production Technology* 10.5 (2020), pp. 2055–2068.

[23] P. Fearnhead, O. Papaspiliopoulos, G.O. Roberts, and A. Stuart. "Random-weight particle filtering of continuous time processes". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 72.4 (2010), pp. 497–512.

[24] M. Frangos. "Uncertainty quantification for cuttings transport process monitoring while drilling by ensemble Kalman filtering". In: *Journal of Process Control* 53 (2017), pp. 46–56.

[25] J. Fu, Y. Su, W. Jiang, and L. Xu. "Development and testing of kick detection system at mud line in deepwater drilling". In: *Journal of Petroleum Science and Engineering* 135 (2015), pp. 452–460.

[26] X. Gao, H. Li, Y. Wang, T. Chen, X. Zuo, and L. Zhong. "Fault detection in managed pressure drilling using slow feature analysis". In: *IEEE Access* 6 (2018), pp. 34262–34271.

[27] P. Garcia-Navarro, J. Murillo, J. Fernandez-Pato, I. Echeverribar, and M. Morales-Hernandez. "The shallow water equations and their application to realistic cases". In: *Environmental Fluid Mechanics* 19 (2019), pp. 1235–1252.

[28] J. Geweke. "Bayesian inference in econometric models using Monte Carlo integration". In: *Econometrica: Journal of the Econometric Society* (1989), pp. 1317–1339.

[29] M.B. Giles. "Multilevel monte carlo path simulation". In: *Operations Research* 56.3 (2008), pp. 607–617.

[30] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. "A kernel method for the two-sample-problem". In: *Advances in Neural Information Processing Systems* 19 (2006).

[31] K. Heine and D. Burrows. "Multilevel bootstrap particle filter". In: *Bernoulli* 29 (2023), pp. 551–579.

[32] K. Heine, N. Whiteley, and A.T. Cemgil. "Parallelizing particle filters with butterfly interactions". In: *Scandinavian Journal of Statistics* 47.2 (2020), pp. 361–396.

[33] M. Hernandez, D.W. MacNeill, M. Reeves, A.D. Kirkwood, J.P. Ruszka, R. Zaeper, and S.R. Lemke. "High-speed wired drillstring telemetry network delivers increased safety, efficiency, reliability, and productivity to the drilling industry". In: *SPE Indian Oil and Gas Technical Conference and Exhibition*. OnePetro. 2008.

[34] Z. Huang, T. Schneider, M. Li, C. Jiang, D. Zorin, and D. Panozzo. "A large-scale benchmark for the incompressible Navier-Stokes equations". In: *ArXiv Preprint ArXiv:2112.05309* (2021).

[35] M.M. Huque, S. Imtiaz, A. Rahman, and M. Hossain. "Kick detection and remedial action in managed pressure drilling: a review". In: *SN Applied Sciences* 2.7 (2020), pp. 1–29.

[36] A. Jasra, K. Kamatani, K. J. H. Law, and Y. Zhou. "Multilevel particle filters". In: *SIAM Journal on Numerical Analysis* 55.6 (2017), pp. 3068–3096.

[37] S.J. Julier and J.K. Uhlmann. "Unscented filtering and nonlinear estimation". In: *Proceedings of the IEEE* 92.3 (2004), pp. 401–422.

[38] G.O. Kaasa, Ø.N. Stamnes, L. Imsland, and O.M. Aamo. "Simplified hydraulics model used for intelligent estimation of downhole pressure for a managed-pressure-drilling control system". In: *SPE Drilling & Completion* 27.1 (2012), pp. 127–138.

[39] R.E. Kalman. "A new approach to linear filtering and prediction problems". In: *Journal of Basic Engineering* (1960).

[40] R.E. Kalman. "Contributions to the theory of optimal control". In: *Boletin de la Sociedad Matemática Mexicana* 5.2 (1960), pp. 102–119.

[41] R.E. Kalman and R.S. Bucy. "New results in linear filtering and prediction theory". In: *Journal of Basic Engineering* (1961).

[42] R. Käppeli and S. Mishra. "Well-balanced schemes for the Euler equations with gravitation". In: *Journal of Computational Physics* 259 (2014), pp. 199–219.

[43] A. Kong, J.S. Liu, and W.H. Wong. "Sequential imputations and Bayesian missing data problems". In: *Journal of the American Statistical Association* 89.425 (1994), pp. 278–288.

[44] I.S. Landet, A. Pavlov, and O.M. Aamo. "Modeling and control of heave-induced pressure fluctuations in managed pressure drilling". In: *IEEE Transactions on Control Systems Technology* 21.4 (2012), pp. 1340–1351.

[45] L. Lapidus and G.F. Pinder. *Numerical solution of partial differential equations in science and engineering.* John Wiley & Sons, 2011.

[46] H. Lee, S.H. Ong, M. Azeemuddin, and H. Goodman. "A wellbore stability model for formations with anisotropic rock strengths". In: *Journal of Petroleum Science and Engineering* 96 (2012), pp. 109–119.

[47] R.J. LeVeque. *Finite volume methods for hyperbolic problems.* Vol. 31. Cambridge University Press, 2002.

[48] T. Li, M. Bolic, and P.M. Djuric. "Resampling methods for particle filtering: classification, implementation, and strategies". In: *IEEE Signal Processing Magazine* 32.3 (2015), pp. 70–86.

[49] J.S. Liu. "Metropolized independent sampling with comparisons to rejection sampling and importance sampling". In: *Statistics and Computing* 6.2 (1996), pp. 113–119.

[50] J.S. Liu. *Monte Carlo strategies in scientific computing.* Springer, 2001.

[51] K.A. Macdonald and J.V. Bjune. "Failure analysis of drillstrings". In: *Engineering Failure Analysis* 14.8 (2007), pp. 1641–1666.

[52] P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. "Markov chain Monte Carlo without likelihoods". In: *Proceedings of the National Academy of Sciences* 100.26 (2003), pp. 15324–15328.

[53] V. Michel-Dansac, C. Berthon, S. Clain, and F. Foucher. "A well-balanced scheme for the shallow-water equations with topography". In: *Computers & Mathematics with Applications* 72.3 (2016), pp. 568–593.

[54] P. Del Moral. *Feynman-Kac formulae.* Springer, 2004.

[55] P. Del Moral, A. Doucet, and A. Jasra. "An adaptive sequential Monte Carlo method for approximate Bayesian computation". In: *Statistics and Computing* 22 (2012), pp. 1009–1020.

[56] P. Del Moral, A. Doucet, and A. Jasra. "Sequential monte carlo samplers". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 68 (2006), pp. 411–436.

[57] G.W. Peters, Y. Fan, and S.A. Sisson. "On sequential Monte Carlo, partial rejection control and approximate Bayesian computation". In: *Statistics and Computing* 22 (2012), pp. 1209–1222.

[58] F. Poletto, F. Miranda, P. Corubolo, A. Schleifer, and P. Comelli. "Drill-bit seismic monitoring while drilling by downhole wired-pipe telemetry". In: *Geophysical Prospecting* 62.4 (2014), pp. 702–718.

[59] T.P. Prescott and R.E. Baker. "Multifidelity approximate Bayesian computation with sequential Monte Carlo parameter sampling". In: *SIAM/ASA Journal on Uncertainty Quantification* 9.2 (2021), pp. 788–817.

[60] S. Rathnayaka, F. Khan, and P. Amayotte. "Accident modeling and risk assessment framework for safety critical decision-making: application to deepwater drilling operation". In: *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability* 227.1 (2013), pp. 86–105.

[61] P. Rebeschini and R. Van Handel. "Can local particle filters beat the curse of dimensionality?" In: *JSTOR* (2015).

[62] C.P. Robert and G. Casella. *Monte Carlo statistical methods.* Vol. 2. Springer, 2004.

[63] A.N. Shiryaev. *Probability.* Springer, 1995.

[64] A.N. Shiryaev. *Probability-1.* Springer, 2016.

[65] O.M. Siddig, S.F. Al-Afnan, S.M. Elkatatny, and A. Abdulraheem. "Drilling data-based approach to build a continuous static elastic moduli profile utilizing artificial intelligence techniques". In: *Journal of Energy Resources Technology* 144.2 (2022).

[66] J.E. Skogdalen, I.B. Utne, and J.E. Vinnem. "Developing safety indicators for preventing offshore oil and gas deepwater drilling blowouts". In: *Safety Science* 49.8 (2011), pp. 1187–1199.

[67] P. Sprent and N.C. Smeeton. *Applied nonparametric statistical methods*. Vol. 3. CRC Press, 2016.

[68] E. Süli and D.F. Mayers. *An introduction to numerical analysis*. Cambridge University Press, 2003.

[69] H.O. Tahar, M.S.D. Haggar, and M. Mbehou. "On the two-step BDF finite element methods for the incompressible Navier–Stokes problem under boundary conditions of friction type". In: *Results in Applied Mathematics* 17 (2023), p. 100349.

[70] N. Tamim, D.M. Laboureur, A.R. Hasan, and M.S. Mannan. "Developing leading indicators-based decision support algorithms and probabilistic models using Bayesian network to predict kicks while drilling". In: *Process Safety and Environmental Protection* 121 (2019), pp. 239–246.

[71] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M.P.H. Stumpf. "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems". In: *Journal of the Royal Society Interface* 6.31 (2009), pp. 187–202.

[72] F.E. Toro. *Riemann solvers and numerical methods for fluid dynamics: a practical introduction*. Springer, 1999.

[73] S. Unrau, P. Torrione, M. Hibbard, R. Smith, L. Olesen, and J. Watson. "Machine learning algorithms applied to detection of well control events". In: *SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition*. OnePetro. 2017.

[74] A.P. Valentine and J.H. Woodhouse. "Reducing errors in seismic tomography: combined inversion for sources and structure". In: *Geophysical Journal International* 180.2 (2010), pp. 847–857.

[75] C. Vergé, C. Dubarry, P. Del Moral, and E. Moulines. "On parallel implementation of sequential Monte Carlo methods: the island particle model". In: *Statistics and Computing* 25.2 (2015), pp. 243–260.

[76] R. Wagner. "Multi-linear interpolation". In: *Beach Cities Robotics* (2008).

[77] E.A. Wan and R. van der Merwe. "The unscented Kalman filter for nonlinear estimation". In: *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications and Control Symposium*. IEEE. 2000, pp. 153–158.

[78] N. Whiteley, A. Lee, and K. Heine. "On the role of interaction in sequential Monte Carlo algorithms". In: *Bernoulli* 22.1 (2016), pp. 494–529.

[79] A. Willersrud, M. Blanke, L. Imsland, and A. Pavlov. "Fault diagnosis of downhole drilling incidents using adaptive observers and statistical change detection". In: *Journal of Process Control* 30 (2015), pp. 90–103.

[80] A. Willersrud, L. Imsland, M. Blanke, and A. Pavlov. "Early detection and localization of downhole incidents in managed pressure drilling". In: *SPE/IADC Managed Pressure Drilling and Underbalanced Operations Conference & Exhibition*. OnePetro. 2015.

[81] D. Yakoubi. "Enhancing the viscosity-splitting method to solve the time-dependent Navier–Stokes equations". In: *Communications in Nonlinear Science and Numerical Simulation* 123 (2023), p. 107264.

[82] H. Zhang, D. Pan, L. Zhai, Y. Zhang, and C. Chen. "Stability analysis in determining safety drilling fluid pressure windows in ice drilling boreholes". In: *Energies* 11.12 (2018), p. 3378.

[83] J. Zhou and G. Nygaard. "Nonlinear adaptive observer for managed pressure drilling system". In: *2011 6th IEEE Conference on Industrial Electronics and Applications*. IEEE. 2011, pp. 79–84.

Appendices

A Central limit theorem for triangular martingale arrays

Let $(U_{N,\rho})_{1\leq\rho\leq\rho_N^{\max}}$ be a triangular random variable array such that $\mathbb{E}[U_{N,\rho} \mid \mathcal{G}_{N,\rho-1}] = 0$, and let $(\mathcal{G}_{N,\rho})_{0\leq\rho\leq\rho_N^{\max}}$ be a triangular array of sub-σ-algebras of $\mathcal{F}$ of the underlying probability space, such that $\rho_N^{\max}$ is $\mathcal{G}_{N,0}$ measurable, and $\mathcal{G}_{N,\rho-1} \subset \mathcal{G}_{N,\rho}$, and for each N and $1 \leq \rho \leq \rho_N^{\max}$, $U_{N,\rho}$ is $\mathcal{G}_{N,\rho}$-measurable. Then we have the following result:

Theorem 6. *Assume that* $\mathbb{E}[U_{N,\rho}^2 \mid \mathcal{G}_{N,\rho-1}] < \infty$ *for all* $1 \leq \rho \leq \rho_N^{\max}$ *and that*

$$\sum_{\rho=1}^{\rho_N^{\max}} \mathbb{E}[U_{N,\rho}^2 \mathbb{1}[|U_{N,\rho}| \geq \epsilon] \mid \mathcal{G}_{N,\rho-1}] \xrightarrow[N\to\infty]{\mathbb{P}} 0, \qquad \textit{for all } \epsilon > 0,$$

$$\sum_{\rho=1}^{\rho_N^{\max}} \mathbb{E}[U_{N,\rho}^2 \mid \mathcal{G}_{N,\rho-1}] \xrightarrow[N\to\infty]{\mathbb{P}} \sigma^2, \qquad \textit{for some } \sigma^2 > 0.$$

Then for any $u \in \mathbb{R}$

$$\mathbb{E}\left[\exp\left(\mathrm{i}u \sum_{\rho=1}^{\rho_N^{\max}} U_{N,\rho}\right)\middle| \mathcal{G}_{N,0}\right] \xrightarrow[N\to\infty]{\mathbb{P}} \exp\left(-\frac{u^2}{2}\sigma^2\right).$$

B Experiment hyperparameters

Table 1: Experiment hyperparameters for the steady state shallow water equations (left) and the convection-diffusion equation (right).

Parameter	Value
σ_X	2
σ_Y	0.2
W	50
M	750
x_{min}	1.5
x_{max}	9.5
N_{data}	10
N_{length}	25
N_{trials}	10
N_{BPF}	2500
N_{ref}	10^5
k_0	3.5
h_0	2

Parameter	Value
σ_X	1
σ_Y	0.25
W	1
M	100
x_{min}	1.5
x_{max}	8
N_{data}	10
N_{length}	20
N_{trials}	10
N_{BPF}	1000
N_{ref}	10^5
v_0	15
T	0.05

C Computed experiment particle allocations

Table 2: Level 0 sample allocations for the steady state shallow water equations experiment, computed such that the MLBPF runtime is within 0.05 of the BPF runtime.

M_0 \ N_1	0	4	50	250	1000	1500	2000
350	5357	5357	5357	4686	2343	647	0
200	9667	9375	9375	8203	4980	2196	0
100	18750	18750	18750	16405	10546	6444	2269
50	37500	37500	37500	32812	21093	12890	6737
45	41666	41666	41666	36457	23437	14322	7486
40	46875	46875	46875	41015	29296	16112	8422
25	75000	75000	75000	65625	46875	25781	13476
15	128906	125000	125000	109375	78125	42968	22460
10	193359	187500	187500	164062	117187	82031	43944

Table 3: Analogous level 0 sample allocations for the convection-diffusion equation experiment.

M_0 \ N_1	0	5	25	50	100	250	500	750
80	1562	1562	1406	1406	1328	937	292	0
60	2707	2707	2499	2499	2290	1874	937	0
40	5312	5312	5000	5000	4375	3750	1875	585
20	14375	14375	14375	13437	12500	10625	6250	3125
15	19164	19164	19164	17914	16665	14164	8332	4166
10	28750	28750	28750	28750	25000	21250	15000	7500