Author/s:
Mao, Jiadong

Title:
Nonparametric estimation for streaming data

Date:
2020

Persistent Link:
https://hdl.handle.net/11343/237540

Nonparametric estimation for streaming data

Jiadong Mao

ORCID: 0000-0002-3818-1981

Submitted in total fulfilment of the requirements of the degree of

Doctor of Philosophy

School of Mathematics and Statistics
The University of Melbourne

February 2020

Abstract

Streaming data are a type of high-frequency and nonstationary time series data. The collection of streaming data is sequential and potentially never-ending. Examples of streaming data, including data from sensor networks, mobile devices and the Internet, are prevalent in our daily lives. An estimator for streaming data needs to be computationally efficient so that it is relatively easy to update the estimator using newly arrived data. In addition, the estimator has to be adaptive to the nonstationarity of data. These constraints make streaming data analysis more challenging than analysing the conventional non-streaming data sets.

Although streaming data analysis has been discussed in the machine learning community for more than two decades, it has received limited attention from statistical researchers. Estimation methods that are both computationally efficient and theoretically justified are still lacking. In this thesis, we propose nonparametric density and regression estimation methods for streaming data, where the smoothing parameters are chosen in a computationally efficient and fully data-driven way. These methods extend some classical kernel smoothing techniques, such as the kernel density estimator and the Nadaraya–Watson regression estimator, to address the theoretical and computational challenges arising from streaming data analysis. Asymptotic analyses provide these methods with theoretical justification. Numerical studies have shown the superiority of our methods over conventional ones. Through some real-data examples, we show that these methods are potentially useful in modelling real-world problems. Finally, we discuss some directions for future research, including extending these methods to model higher-dimensional streaming data and to streaming data classification.

Declaration

This is to certify that

1. the thesis comprises only my original work towards the PhD except where indicated in the preface,
2. due acknowledgement has been made in the text to all other material used,
3. the thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Jiadong Mao, Feb 2020

Preface

This thesis is submitted to the University of Melbourne in support of my application for admission to the degree of Doctor of Philosophy. No part of it has been submitted in support of an application for another degree or qualification of this or any other institution of learning.

Chapter 2 is a joint work with Professor Aurore Delaigle [1]. Chapter 3 is a joint work with Professor Aurore Delaigle and Dr Félix Camirand-Lemyre [2]. They contain materials from two papers that are about to be submitted for publication.

I acknowledge the financial support from the University of Melbourne in the form of Melbourne International Fee Remission Scholarship and Melbourne International Research Scholarship. My PhD candidature has also been funded by a PhD top-up scholarship from ARC Centre of Excellence for Mathematical and Statistical Frontiers, Professor Maurice Belz Fund and Research Higher Degree Student Travel Fund from the School of Mathematics and Statistics, the University of Melbourne, and a travel fund from the Statistical Society of Australia for attending the Young Statistician Conference 2019 in Canberra.

[1] School of mathematics and Statistics, The University of Melbourne, Parkville, Victoria, Australia.
[2] Department of Mathematics, Université de Sherbrooke, Sherbrooke, Quebec, Canada.

Acknowledgements

Prof Aurore Delaigle is not only my supervisor – she is a skilled teacher, a patient mentor, a helpful colleague, an acute-minded reader and a kind-hearted friend. We've known each other for full six years now – ever since I started my master's study in this beautiful continent. Thanks to her advice, I was able to discover this challenging yet ever motivating research topic and, thanks to her patience over the years, I have been able to grow into a more mature researcher.

I have always enjoyed discussing with Dr Félix Camirand-Lemyre, my co-supervisor and a wonderful friend. His feedback to my research and thesis writing is invaluable. I will also miss the numerous lunches and dinner parties he organised for our research group.

Special thanks to Jiajun Tang, my office mate for more than three years, who helped me a lot in checking some technical details of this thesis.

Guidance and support from other members of my existing and past supervision panels, including Prof Aihua Xia, Dr Zhuosong Zhang, Dr Davide Ferrari, and Prof Jinyuan Chang, are highly appreciated.

I will be forever grateful to Prof Konstantin Borovkov, Prof Aurore Delaigle, Dr Guoqi Qian and Prof Aihua Xia, for their well-prepared and skilfully delivered lectures during my MSc study.

The support from my wife Tian and our parents, Min Sun, Zhe Yu, Chuanhua Mao and Shenghua Yuan, has been a constant source of motivation and a *conditio sine qua non* of this thesis. I also have to thank Tian for being both the *causa materialis* and the *causa efficiens* of our daughter Qingyi (庆一), born on 25 March 2019, a happy 'by-product' of my PhD candidature.

To Tian,

*'Best Image of my self and dearer half' (*Paradise Lost *5.95)*

Content

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Streaming data are a type of temporal data collected sequentially over a potentially infinite time period with high velocity in a nonstationary environment. Different from conventional high-frequency and nonstationary time series, the collection of streaming data may be never-ending and real-time modelling is often required.

The ubiquity of mobile devices and sensor networks has made this type of data very common in our daily lives. A typical example is the Phasor Measurement Units (PMUs) technology used for monitoring large power grids (Zhou et al., 2016). Typically, measurements from PMUs are reported at a standard rate of 60Hz (60 reads per second) and the ability to analyse these data in real time is essential for the management of large power systems. Another example is the NASA satellite *Swift* (Gehrels et al., 2004), which monitors a large number of celestial objects by receiving from them certain kinds of electromagnetic waves, such as gamma-rays and x-rays. For each object, an observed rate (measuring the level of a certain kind of wave) of, say, gamma-ray is generated every few milliseconds. Due to various celestial activities, the distribution of the detected rate is likely to be time-varying.

There are mainly two features that distinguish streaming data from the conventional independent and identically distributed (i.i.d.) data. Firstly, the data arrive at a very high speed and

their collection is never-ending. Hence, storage of all past data is difficult and any estimator has to be online, i.e. it has to be frequently updated using the newly arrived data. Secondly, the data are likely to be nonstationary and hence the estimator has to be adaptive to the time variabilities. For example, the distribution of the measurements from a sensor network can be time-varying, either due to the evolving physical phenomena it monitors, or due to the ageing of sensors (Ditzler et al., 2015).

Because of the above features, conventional nonparametric methods, such as the kernel density estimator (KDE), are often not directly applicable to streaming data. On the one hand, in terms of computation, these methods are often offline. That is, they assume that all data have been observed and stored in the computer memory, so that all data can be accessed for unlimited number of times. In addition, these methods can be computationally intensive due to the method they use for the selection of smoothing parameters. Conventional data-driven procedures for this purpose, such as cross-validation (Hall, 1983; Härdle and Marron, 1985; Hall, 1987) and bootstrap (Taylor, 1989; Faraway and Jhun, 1990; Faraway, 1990; Hall, 1990; Delaigle and Gijbels, 2004), are often too time-consuming for processing streaming data. On the other hand, in terms of adaptivity to nonstationarity, these nonparametric methods are originally designed for stationary data (see e.g. Wand and Jones, 1995; Bosq, 1998), without taking nonstationarity into consideration.

In this thesis, we develop computationally efficient nonparametric density and regression estimation methods for smoothly time-evolving streaming data. The estimators are online and smoothing parameters are selected by some computationally efficient cross-validation procedures. Namely, the streaming cross-validation in §2.4 for density estimation and and the recursive cross-validation in §3.2 for regression. We analyse the asymptotic properties of these estimators and cross-validation procedures within the infill asymptotics framework used in Hall et al. (2006), Vogt (2012) and Zhang and Wu (2015). Superior performances of these methods over competitors are illustrated by simulations. Application of these methods to some real-data examples show their potential use in modelling real-world problems. This thesis shows a possibility of extending classical kernel smoothing techniques to modelling big and complex data, which might

be interesting for researchers in both nonparametric statistics and machine learning.

1.2 Data and models

In this section we first define some different types of streaming data and the corresponding density and regression models. Then we define some other types of data that will be useful for the literature review in §1.3.

1.2.1 Streaming data

We define a data stream $\{Z_i\}_{i=1,2,\dots}$ as an infinite sequence of independent and non-identically distributed (i.n.i.d.) or dependent and non-identically distributed (d.n.i.d.) random variables or random vectors, where each observation Z_i arrives at time

$$t_i = i\Delta t, \text{ for some } \Delta t > 0\,. \tag{1.1}$$

Let n_t denote the number of observations arriving up to a given time $t > 0$, so that

$$t/\Delta t - 1 \leq n_t = \lfloor t/\Delta t \rfloor \leq t/\Delta t\,. \tag{1.2}$$

Assuming equidistant arrival times is realistic since, in practice, many devices (e.g. traffic sensors) make measurements regularly in time. A model assuming equidistant arrivals can readily be applied to data with slightly irregular arrivals.

For the simplicity of theoretical derivations, in this thesis we only focus on the univariate scenario (see Chapter 5 for a discussion on extensions to the higher-dimensional case). That is, in the density estimation case, $Z_i = X_i \sim f(\cdot, t_i)$ denotes a univariate random variable, where $f(\cdot, \cdot)$ is a time-varying probability density function. In the regression case, $Z_i = (X_i, Y_i) \in \mathbb{R}^2$

denotes a random vector from the following time-varying regression model:

$$Y_i = m(X_i, t_i) + \epsilon_i\,, \quad i = 1, 2, \ldots, \tag{1.3}$$

where $m(x, t_i) = \mathrm{E}(Y_i|X_i = x)$ is a time-varying regression function and $\{\epsilon_i\}_{i=1,2,\ldots}$ is an error sequence.

1.2.2 Other types of data

The i.n.i.d. (d.n.i.d.) streaming data defined above are different, both in nature and in the way they can be analysed, from the standard data usually encountered in the literature, which we refer to as offline i.i.d. data, denoted by $\{Z_i\}_{i=1,\ldots,n}$. The data are called offline since they are all available at once and each observation can be accessed for unlimited number of times. The i.n.i.d. (d.n.i.d.) streaming data also differ from what we refer to as online i.i.d. data, where the data $\{Z_i\}_{i=1,2,\ldots}$ are observed sequentially at times $\{t_i\}_{i=1,2,\ldots}$ satisfying (1.1), but where the Z_i's have the same (non-time-varying) distribution. Another type of data encountered in practice is what we refer to as offline i.n.i.d. data, where the data $\{Z_i\}_{i=1,\ldots,n}$ are independent and all available at once, but have been observed at consecutive times $\{t_i\}_{i=1,\ldots,n}$ satisfying (1.1) and are nonstationary. That is, $X_i \sim f(\cdot, t_i)$ and, in the regression problem, the regression function $m(x, t_i) = \mathrm{E}(Y_i|X_i = x)$ is time-varying.

1.2.3 A bibliographic note on definition of streaming data

To the best of our knowledge, there is no consensus in the literature concerning the mathematical definition of streaming data. Some works view online i.i.d. data as a kind of streaming data. For example, Domingos and Hulten (2000) proposed a classification tree method for i.i.d. streaming data and Luo and Song (2020) developed a recursive maximum likelihood estimation procedure, also for i.i.d. streaming data. Assuming stationarity, they only focused on online modelling, i.e. how to efficiently update the tree classifier or the maximum likelihood estimator when more and

more data are available. However, since streaming data are potentially never-ending, it is less realistic to assume that they are stationary. Hence, in line with the majority of existing works on streaming data (see e.g. Ditzler et al., 2015, for a review of classification and clustering methods for nonstationary streaming data), in this thesis we include nonstationarity in the definition of streaming data and consider the i.n.i.d. and the d.n.i.d. case.

1.3 Literature review

Since streaming data analysis first arises in the machine learning community, in this section we first give a brief introduction to the machine learning literature on streaming data. Then, we turn to existing statistical works on streaming data, which are limited in number compared to the machine learning literature. To conclude this section, we review some nonparametric statistics techniques that will be the building blocks of our density and regression estimation methods described in this thesis.

1.3.1 Machine learning literature on streaming data

Streaming data analysis has received attention from the machine learning community as a result of the advances in hardware technologies such as network monitoring, web mining, sensor networks and manufacturing systems (Gama and Gaber, 2007; Sayed-Mouchaweh, 2019). Most of these works focus on modelling tasks such as classification and clustering. See Gama (2010) and Ditzler et al. (2015) for some more comprehensive reviews on the classification and clustering of streaming data. Here we selectively review some methods that directly inspired our works in this thesis.

1.3.1.1 Online learning

As mentioned in §1.1, one of the major challenges for streaming data analysis is that any estimator has to be computationally efficient so that it is relatively easy to be frequently updated.

However, most of the conventional machine learning algorithms are designed for offline data. For streaming data, it is more desirable to use online methods, where the estimators can be frequently updated using only one or a small number of newly arrived data points. Many earlier machine learning works on streaming data focus on designing online versions of classic offline methods for classification, clustering and regression. One of the pioneering works in this direction is the very fast decision tree algorithm proposed by Domingos and Hulten (2000), which constructs classification trees that can be sequentially updated using each newly arrived data point. However, their algorithm is designed for online i.i.d. data and hence is not appropriate for nonstationary streaming data. Moreover, their work is about classification, whilst our focus in this thesis is on nonparametric density estimation and regression.

1.3.1.2 Nonstationarity adaptation

As mentioned in §1.1, streaming data are often nonstationary. In the machine learning literature, nonstationarity is often termed as concept drift and nonstationarity adaptation is one of the most important topics in streaming data analysis (Gama et al., 2014). For an algorithm to be adaptive to the nonstationarity, it has to reduce the influence of older data points to the current estimator. Different approaches of discounting the past in the machine learning literature can be summarised into three categories: windowing, weighting and ensemble learning.

Windowing. Using the notations in §1.2, to construct an estimator at time $t \in \mathbb{R}_+$, the windowing approach uses only the most recent block of data $\{Z_i\}_{i=n_t-w_t+1,\ldots,n_t}$ arriving on time window $(t - w_t\Delta t, t]$, where Δt is defined in (1.1) and w_t is an integer called the window size. This approach is also termed sliding window or rolling window, and is often used for modelling nonstationary time series (e.g. Inoue et al., 2017).

Using the windowing approach, Hulten et al. (2001) modified the very fast decision tree into the concept-adapting very fast decision tree algorithm. This modified algorithm also sequentially updates a classification tree, but at a time $t \in \mathbb{R}_+$ the tree model is computed using only $\{Z_i\}_{i=n_t-w+1,\ldots,n_t}$, i.e. data points arriving on a time window $(t - w\Delta t, t]$, where the fixed

window size w is chosen by hand. In addition, the algorithm periodically checks if any sub-tree (part of the tree structure) has become 'outdated' according to a data-driven criterion computed from $\{Z_i\}_{i=n_t-w+1,\ldots,n_t}$. When a sub-tree fails to satisfy the criterion (hence viewed as outdated), the algorithm starts building an alternative sub-tree. The alternative sub-tree is not used to replace the original sub-tree until the alternative admits lower classification errors than the original. The windowing approach in Hulten et al. (2001) directly inspired the recursive cross-validation procedure in Chapter 3, which is a key part of the nonparametric regression method described therein.

The windowing approach has also been applied to nonparametric density and regression estimation to make the conventional estimators adaptive to nonstationarity, see §1.3.4.1 for more details.

Weighting. For keeping the estimates up-to-date, the weighting approach assigns larger weights to more recent data points and smaller weights to data from more distant past. A simple but popular weighting approach is exponential smoothing, which is often used in the time series literature for the forecast of a univariate time series (Harvey, 1990, pp. 25–26; Gijbels et al., 1999). It recursively computes weights for all past data points, where the weights for data in more distant past decreases exponentially fast.

To illustrate exponential smoothing, we consider the problem of estimating the time-varying sample mean of streaming data. For data stream $\{X_i\}_{i=1,2,\ldots}$ with time-varying means $\{\mu_i\}_{i=1,2,\ldots}$, exponential smoothing estimates the μ_i's using recursive formulas $\hat{\mu}_0 = 0$ and $\hat{\mu}_i = (1-\gamma)\hat{\mu}_{i-1} + \gamma X_i$, for $i = 2, 3, \ldots$. Here $\gamma \in (0, 1)$ is a parameter called the stepsize, controlling how fast the estimator forgets the influence of past data points. Instead of using a fixed γ value, we can also use a sequence of adaptive stepsizes $\{\gamma_i\}_{i=1,2,\ldots}$. This is because, when the μ_i's are changing fast (slowly), it is more appropriate to use a larger (smaller) γ value.

Exponential smoothing, with either fixed or adaptive stepsizes has been extensively used in streaming data analysis. See e.g. Anagnostopoulos et al. (2009), Pavlidis et al. (2011), Anagnostopoulos et al. (2012) , Ross et al. (2012) , Bodenham and Adams (2017) and Noble and

Adams (2018) for some applications of exponential smoothing to classification and changepoint detection.

Exponential smoothing has also been applied to nonparametric density and regression estimation. See §1.3.4.2 for more details.

Ensemble learning. The basic idea of ensemble learning is to build an estimator that combines the strengths of a collection of base learners (Hastie et al., 2009, p. 605). Ensemble learning has been widely used to model offline i.i.d. data. A classic example is the random forest proposed by Breiman (2001). Instead of building only one classification tree using the whole data set, Breiman (2001) proposes to draw a number of bootstrap resamples and build an overfitted classification tree from each resample (as the base learner), thus forming an ensemble of tree classifiers. To classify a new data point, each tree classifier computes a predicted class label and then the class label receiving the highest number of votes will be chosen as the final predicted class label for that data point. One of the keys to the success of ensemble learning is the variance reduction induced by model averaging (such as the voting procedure in the random forest). In the random forest case, due to overfitting, each tree classifier has low bias but high variance. Then the voting can greatly reduce the variance, while only slightly increasing the bias (Hastie et al., 2009, pp. 587–589).

In addition to modelling offline i.i.d. data, ensemble learning has become a popular approach for nonstationarity adaptation in streaming data analysis. See Gomes (2017) for a review of ensemble methods for streaming data classification. At time $t \in \mathbb{R}_+$, we can incorporate information from more recent data by adding base learners computed from these data, and, to forget outdated information, we simply remove base learners computed from older data. To illustrate the flexibility of ensemble learning in coping with nonstationarity, next we review an algorithm proposed by Street and Kim (2001) in some detail. For some more recent examples of ensemble methods for streaming data classification, see Kolter and Maloof (2007), Minku and Yao (2012) and Bertini Jr. and Nicoletti (2002).

The streaming ensemble algorithm proposed by Street and Kim (2001) is one of the earliest

ensemble learning algorithms designed for streaming data classification. It partitions the data stream into consecutive blocks $B_1, B_2, \ldots$, each containing b data points, and iteratively updates an ensemble $\mathscr{E} = \{E_i\}_{i=1,\ldots,G}$ of G base classifiers. The algorithm is iteratively defined as follows. To initialise $\mathscr{E}$, we first build classifiers $C_1, \ldots, C_G$ using the first G data blocks $B_1, \ldots, B_G$, respectively. Then, the initial ensemble is defined by $E_i = C_i$, $i = 1, \ldots, G$, i.e. it contains classifiers trained from the first G data blocks. When a new block B_{G+1} arrives, we use it to build a classifier C_{G+1}. Then we decide whether we replace some existing base classifier $E_j \in \mathscr{E}$ with this new classifier C_{G+1}. For this, we use another new block B_{G+2} as the test set to compute the classification errors of C_{G+1} and all $E_i \in \mathscr{E}$. If C_{G+1} yields lower test error than some existing base classifier $E_j \in \mathscr{E}$, let $E_j \leftarrow C_{G+1}$. Otherwise we discard C_{G+1} and do not update $\mathscr{E}$. The above procedure is repeated to keep $\mathscr{E}$ up-to-date.

With an ensemble $\mathscr{E}$ of base classifiers, the streaming ensemble algorithm classifies a data point by combining the prediction results of all $E_i \in \mathscr{E}$, but not by simple voting as in the random forest case. In the latter case, the data are i.i.d., so each base classifier is given the same weight in voting. In contrast, the streaming ensemble algorithm uses weighted voting. That is, it gives a base classifier E_i higher weight if it yields a lower test error on recent data blocks. This is because the data are nonstationary and classifiers that classify the more recent data points accurately are more useful for real-time classification. The reason that we limit the size of the ensemble $\mathscr{E}$ to be some fixed integer G is because we cannot afford to store all classifiers, when the number of them becomes overly large (recall that streaming data is never-ending so the number of data blocks goes to infinity).

Inspired by the ensemble methods for streaming data classification, we propose to apply the ensemble learning approach to adaptive nonparametric regression estimation for streaming data. Instead of using just one estimator and continuously updating it with new data, we propose to construct an ensemble of estimators, each with a different choice of smoothing parameters. Then, we dynamically choose the best estimator from the ensemble, according to some data-driven criterion computed from a recent block of data. See Chapter 3 for details.

1.3.1.3 Nonparametric estimation for streaming data in machine learning literature

In the machine learning literature, discussions on nonparametric estimation for streaming data mainly focus on the fast computation of the KDE. For example, Gray and Moore (2003), Zhou et al. (2003) and Zheng et al. (2013) considered fastening the computation of the KDE for modelling online stationary data. They proposed different methods to adaptively bin the data and then compute the conventional KDE based on the binned data. Boedihardjo et al. (2008), Heinz and Seeger (2008), Cao et al. (2012) and Qahtan et al. (2017) proposed different methods to compute the binned KDE for nonstationary streaming data. Boedihardjo et al. (2008), Cao et al. (2012) and Qahtan et al. (2017) computed the conventional KDE using binned data in a sliding window, where the window size is fixed and chosen by hand. Heinz and Seeger (2008) used a recursive KDE, where the weights for data points arriving at different times were computed using exponential smoothing (see §1.3.4.2 for details). However, they did not provide any data-driven method for selecting the stepsize. In Chapter 2 we will propose a recursive KDE for streaming data, where the smoothing parameters are adaptively chosen using some theoretically founded data-driven method.

In recent years, some works have emerged for using online kernel-based methods, such as the online version of the support vector machine, for nonparametric regression estimation and the classification of streaming data (Bedi et al., 2019; Shen et al., 2019). These works formulate the nonparametric function estimation as a convex optimisation problem, where the goal is to find a function in a reproducing kernel Hilbert space that minimises some penalised loss function. However, in these works some key tuning parameters, including those for controlling nonstationarity adaptation, are chosen by hand. In addition, theoretical analyses there mainly focus on proving optimisation theoretical properties of the proposed methods, while statistical properties of these methods largely remain unclear.

1.3.2 Statistical literature on streaming data

In contrast to the flourishing research in the machine learning community, streaming data analysis has received much less attention from the statistics community. Moreover, it appears that the majority of existing statistics literature on streaming data focuses on changepoint and anomaly detections. For example, Talagala et al. (2019) considered the real-time monitoring of a large collection of data streams from a sensor network and proposed a framework for the early detection of anomalous streams. To adapt to the nonstationarity, their method employs a sliding window of a user-defined window size. Data in the sliding window are viewed as up-to-date, and then features of these data are extracted as the input to an anomaly detection procedure. Similar works on the real-time monitoring of streaming data include e.g. Chan (2017), Chen (2019), Padilla et al. (2019), Tveten and Glad (2019) and Zhang et al. (2019).

Statistical works on streaming data other than changepoint and anomaly detections have emerged in recent years but are still relatively scarce. Anagnostopoulos (2010) proposed a parametric framework for streaming data analysis based on a recursive maximum likelihood estimation procedure, motivating several other works on streaming data classification (Adams et al., 2010; Pavlidis et al., 2011; Anagnostopoulos et al., 2012). However, to use this framework we need to specify a parametric form for the density of the data, which is something we avoid doing in this thesis, since we are concerned with nonparametric estimation.

Hofmeyr et al. (2016) proposed an online clustering method for high-dimensional nonstationary streaming data, which dynamically partitions the sample space into subspaces, each corresponding to a cluster. This procedure is based on an conventional KDE computed from the one-dimensional projections of the data stream. To find the appropriate one-dimensional representation of the data stream, they employed a recursive method to dynamically compute the time-varying first principal component of the data stream. When a new observation arrives, the first principal component is updated and the new observation is projected onto the updated principal component. To reduce the computational cost for computing a conventional KDE from an increasing number of data, Hofmeyr et al. (2016) proposed to bin the data and the new projected data point is only used to update the bin counts. Then a conventional KDE is computed

from the updated bin counts. However, since their ultimate goal is clustering rather than density estimation, some key tuning parameters are selected in a way such that the KDE produces meaningful clustering results. For example, the bandwidth for the KDE is chosen to ensure that the density estimate has an appropriate number of modes. Hence their method is not very useful for our problem.

Hall and Patil (1994) defined a class of online KDEs for online i.i.d. data and provided theoretical justifications for these estimators. They also proposed an online KDE for i.n.i.d. streaming data, which is computed from a sliding window of a fixed window size w. They suggested that w should be chosen in a way such that every w consecutive data points can be viewed as approximately i.i.d. For example, for some wind speed datasets where one data point is observed every 6 seconds, it is reasonable to take $w = 200$, so that each sliding window contains data observed in a 20 minute time interval. However, they did not provide any general data-driven method to select the window size w.

For nonparametric density estimation for nonstationary streaming data, Caudle and Wegman (2009) and García-Treviño and Barria (2012) proposed orthogonal series density estimators that can be efficiently updated. Both of their estimators of the density $f(x, t)$ at time t can be written as

$$\hat{f}_{\mathrm{OS}}(x, t) = \sum_{i=1}^{N} \hat{b}_{n_t,i} \psi_i(x) \,,$$

where $\{\psi_i\}_{i=1,2,\ldots}$ is a sequence of orthogonal base functions (e.g. the wavelets), $N \geq 1$ is an integer determining how many base functions are used to approximate the density $f(\cdot, t)$ and $\hat{b}_{n_t,i}$ is the coefficient of ψ_i at time t. To update $\hat{b}_{n_t,i}$ while discounting the influence of past data points, Caudle and Wegman (2009) proposed to use the following exponential smoothing formulas:

$$\hat{b}_{n_t,i} = \begin{cases} \psi_i(X_1) \,, & \text{if } n_t = 1 \,, \\ \hat{b}_{n_t,i} = (1-\gamma)\hat{b}_{n_t-1,i} + \gamma\psi_i(X_{n_t}) \,, & \text{if } n_t \geq 2 \,. \end{cases}$$

Motivated by Caudle and Wegman (2009), García-Treviño and Barria (2012) proposed to compute the coefficient $\hat{b}_{n_t,i}$ from a sliding window of size w. That is, for $n_t = w+1, w+2, \ldots,$

$\hat{b}_{n_t,i}$ is updated by

$$\hat{b}_{n_t,i} = \frac{1}{w} \sum_{j=n_t-w+1}^{n_t} \psi_i(X_j) \,.$$

However, Caudle and Wegman (2009) and García-Treviño and Barria (2012) did not provide any general data-driven method to select the stepsize γ or the window size w, which control how fast we forget the influence of past data and the selection of which is crucial in practice. Moreover, both of their works lack theoretical analysis of the orthogonal series estimator.

Hall et al. (2006) proposed a KDE for offline i.n.i.d. data, but the bandwidth selection method is computationally slow, making their method unsuitable for streaming data (see §1.3.4.4). Harvey and Oryshchenko (2012) applied exponential smoothing to estimate the density of offline d.n.i.d. data. Their estimator is recursive, an hence can also be applied to i.n.i.d. (d.n.i.d.) streaming data (see §1.3.4.2). However, their method for selecting smoothing parameters is offline, requiring the storage of all data points, hence cannot be directly applied to streaming data.

In terms of regression, nonparametric methods are relatively abundant in the statistics literature, but few of them are directly applicable to i.n.i.d. (d.n.i.d.) streaming data. Grillenzoni (2000) proposed a kernel regression estimator for offline d.n.i.d. data using exponential smoothing (see §1.3.4.2). When applied to streaming data, his method suffers from the same problem as that of Harvey and Oryshchenko (2012). Namely, although the estimator can be applied to online data since it is efficient to update using new data points, the bandwidth selection method requires the storage of all data points. Mokkadem et al. (2009a), Amiri (2012) and Huang et al. (2014) discussed various recursive and semi-recursive kernel regression estimators for online i.i.d. (d.i.d.) data (see §1.3.4.2). We will see in Chapter 3 how to apply a semi-recursive kernel regression to nonstationary streaming data. Luts et al. (2014) proposed a class of semi-parametric regression estimators for online i.i.d. data under the Bayesian framework. Vogt (2012) and Zhang and Wu (2015) developed kernel regression estimators for offline d.n.i.d. data and investigated their asymptotic properties, but their estimators cannot be efficiently updated in our streaming data context and discussion on practical bandwidth selection is lacking.

1.3.3 Summary

Our focus in this thesis is on nonparametric curve estimation, including density and regression estimation, for streaming data. As discussed in §1.1, any estimator for streaming data should be both online and adaptive to the time variabilities. In the machine learning and statistical works we have reviewed in §1.3.1 and §1.3.2, there are some nonparametric curve estimators satisfying both of these requirements (Hall and Patil, 1994; Grillenzoni, 2000; Hall et al., 2006; Boedihardjo et al., 2008; Heinz and Seeger, 2008; Caudle and Wegman, 2009; Cao et al., 2012; García-Treviño and Barria, 2012; Harvey and Oryshchenko, 2012; Hofmeyr et al., 2016; Qahtan et al., 2017; Bedi et al., 2019; Shen et al., 2019). However, all these works, except Hall et al. (2006), lack theoretical justification for their methods (Hall and Patil, 1994 discussed theoretical properties of their density estimator for online i.i.d. data, but not for their estimator for nonstationary data).

In addition to the lack of theoretical analysis, these works all have some problems concerning the selection of smoothing parameters in our streaming data context. Most of these works, including Hall and Patil (1994), Boedihardjo et al. (2008), Heinz and Seeger (2008), Caudle and Wegman (2009), Cao et al. (2012), García-Treviño and Barria (2012), Bedi et al. (2019) and Shen et al. (2019), did not provide any data-driven way to dynamically select the smoothing parameter controlling how fast we discount the influence of past data, such as the window size and stepsize (see §1.3.1.2). Grillenzoni (2000), Hall et al. (2006) and Harvey and Oryshchenko (2012) proposed some data-driven ways to select the smoothing parameters, which are suitable for offline data and are computationally heavy in our streaming data context. Since selecting smoothing parameters is crucial in practice, these works leave room for improvement.

In view of the lack of suitable nonparametric curve estimation methods for streaming data, in this thesis we develop nonparametric density and regression estimators that are easy to update and designed to handle nonstationarity. Smoothing parameters for these estimators are selected using some computationally efficient cross-validation procedures tailored for streaming data. As the final part of the literature review, next we review some conventional nonparametric statistics techniques that will be the building blocks for our streaming methodology.

1.3.4 Building blocks for our methodology

In this section we review some nonparametric statistics techniques that will be used to build our methodology. Recall the definitions of different types of data in §1.2.

1.3.4.1 Kernel density and regression estimators

For offline i.i.d. data, it is well known that we can consistently estimate the density f nonparametrically by the conventional KDE, defined by

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\,, \tag{1.4}$$

where K is a kernel function satisfying $\int K = 1$, $h = h_n \in \mathbb{R}_+$ is a smoothing parameter called bandwidth, and $K_h(\cdot) = h^{-1}K(\cdot/h)$. Similarly, the regression function $m(x) = \mathrm{E}(Y_i|X_i = x)$ of offline i.i.d. data can be consistently estimated by the Nadaraya–Watson (NW) estimator, defined as

$$\hat{m}(x) = \frac{\hat{r}(x)}{\hat{f}(x)} = \frac{n^{-1}\sum_{i=1}^{n} K_h(x - X_i)Y_i}{n^{-1}\sum_{i=1}^{n} K_h(x - X_i)}\,, \tag{1.5}$$

where $\hat{r}(x)$ can be viewed as an estimator of $(m\cdot f)(x)$. The conventional KDE is first introduced by Rosenblatt (1956) and Parzen (1962). See e.g. Wand and Jones (1995) for discussions of the conventional KDE and the NW estimator. Both estimators can also be applied to offline d.i.d. data (see e.g. Bosq, 1998).

While these estimators can also be computed in the case of online i.i.d. (d.i.d.) data, using, at each time $t \in \mathbb{R}_+$, only the n_t data points available up to time t, a drawback of this approach is that it is not fully recursive. Indeed, for a given data set of size n_t, $\hat{f}$ at (1.4) requires to compute n_t versions of K rescaled by the bandwidth h, where $h = h_{n_t}$ depends (crucially) on n_t. Therefore, when a new observation becomes available, in order to update the estimator of f, taking the new observation into account, we not only need to compute a new rescaled version of K corresponding to the new observation, but we also need to recompute the n_t versions of K corresponding to the previous n_t observations, this time rescaled by a new bandwidth $h = h_{n_t+1}$.

To apply the KDE at (1.4) and the NW estimator at (1.5) to online i.n.i.d. (d.n.i.d.) data, a straightforward approach is to apply these estimators to data in a sliding window (Subramaniam et al., 2006; Tasoulis et al., 2006; Boedihardjo et al., 2008; Lam and Bouillet, 2015). That is, to estimate $f(\cdot, t)$ (respectively $m(\cdot, t)$) for some time $t \in \mathbb{R}_+$, we compute the KDE from data $\{X_i\}_{i=n_t-w+1,\ldots,n_t}$ (respectively $\{(X_i, Y_i)\}_{i=n_t-w+1,\ldots,n_t}$) arriving in $(t - w\Delta t, t]$, where the window size w is a positive integer chosen by hand, as if these data were identically distributed. More formally, the sliding window KDE of $f(x, t)$ used in Subramaniam et al. (2006), Tasoulis et al. (2006) and Boedihardjo et al. (2008) is defined by

$$\hat{f}_w(x, t) = \frac{1}{w} \sum_{i=n_t-w+1}^{n_t} K_{h_{n_t}}(x - X_i)\,. \tag{1.6}$$

Here $\hat{f}_w(x, t)$ is only defined for $t \geq w\Delta t$, since this estimator has to be computed from no less than w data points. Similarly, the sliding window NW estimator of $m(x, t)$ used in Lam and Bouillet (2015) is defined by

$$\hat{m}_w(x, t) = \frac{w^{-1} \sum_{i=n_t-w+1}^{n_t} K_{h_{n_t}}(x - X_i) Y_i}{w^{-1} \sum_{i=n_t-w+1}^{n_t} K_h(x - X_i)}\,. \tag{1.7}$$

The motivation for the sliding window KDE and NW estimator is straightforward. Namely, to estimate $f(x, t)$ and $m(x, t)$, we use only data arriving in a recent time window instead of all n_t data arriving up to time t, since the data are nonstationary. In the simulation studies in Chapter 4, we will compare the performances of these sliding window estimators and the density and regression estimation methods described in Chapters 2 and 3. Our methods showed superior performances than the sliding window estimators in most of the simulations settings. In the remaining settings, the performances of our methods were comparable to those of the sliding window estimators.

1.3.4.2 Recursive and semi-recursive estimators

To estimate the non-time-varying density function f of online i.i.d. data, Wolverton and Wagner (1969) introduced a recursive KDE, which has been studied by Yamato (1971), Davies (1973), Devroye (1979), Wegman and Davies (1979), Hall and Patil (1994), Amiri (2009) and Mokkadem et al. (2009b), among many others. A generalisation of Wolverton and Wagner's estimator, given by Mokkadem et al. (2009b), is defined as follows. For a given time $t \in \mathbb{R}_+$, recalling the definition of n_t at (1.2), let $\boldsymbol{\gamma}_t = \{\gamma_i\}_{i=1,\dots,n_t}$ and $\boldsymbol{h}_t = \{h_i\}_{i=1,\dots,n_t}$ denote two sequences of smoothing parameters, called the stepsize and the bandwidth, respectively, where $(\gamma_i, h_i) \in (0,1) \times \mathbb{R}_+$. Then the recursive KDE has the form

$$\hat{f}(x,t|\boldsymbol{\gamma}_t,\boldsymbol{h}_t) = \begin{cases} K_{h_1}(x - X_1)\,, & \text{if } n_t = 1\,, \\ (1-\gamma_{n_t})\hat{f}(x,t_{n_t-1}|\boldsymbol{\gamma}_{t_{n_t-1}},\boldsymbol{h}_{t_{n_t-1}}) + \gamma_{n_t}K_{h_{n_t}}(x - X_{n_t})\,, & \text{if } n_t \geq 2\,. \end{cases} \tag{1.8}$$

Note from (1.1) that we have $n_t \to \infty$ as $t \to \infty$. On this occasion, $\hat{f}(x,t|\boldsymbol{\gamma}_t,\boldsymbol{h}_t)$ is a consistent estimator of $f(x)$ under some conditions, including $h_{n_t} \to 0$ and $\gamma_{n_t} \to 0$. See Mokkadem et al. (2009b) for details.

To estimate the non-time-varying regression function m of online i.i.d. data, we can use the semi-recursive NW estimator (Györfi et al., 2002, Chapter 24). Motivated by Wolverton and Wagner (1969) and Yamato (1971), the semi-recursive NW estimator was introduced by Ahmad and Lin (1976) and has been studied by Devroye and Wagner (1980), Krzyżak and Pawlak (1984), Greblicki and Pawlak (1987), Krzyżak (1992), Walk (2001), Amiri (2012), Huang et al. (2014) and Slaoui (2016), among many others. A generalisation of Ahmad and Lin's estimator, given by Slaoui (2016), is defined as

$$\hat{m}(x,t|\boldsymbol{\gamma}_t,\boldsymbol{h}_t) = \frac{\hat{r}(x,t|\boldsymbol{\gamma}_t,\boldsymbol{h}_t)}{\hat{f}(x,t|\boldsymbol{\gamma}_t,\boldsymbol{h}_t)}\,, \tag{1.9}$$

where $\boldsymbol{\gamma}_t$ and $\boldsymbol{h}_t$ are defined above (1.8), $\hat{f}(x, t|\boldsymbol{\gamma}_t, \boldsymbol{h}_t)$ is defined at (1.8) and where

$$\hat{r}(x, t|\boldsymbol{\gamma}_t, \boldsymbol{h}_t) = \begin{cases} K_{h_1}(x - X_1)Y_1\,, & \text{if } n_t = 1\,, \\ (1-\gamma_{n_t})\hat{r}(x, t_{n_t-1}|\boldsymbol{\gamma}_{t_{n_t-1}}, \boldsymbol{h}_{t_{n_t-1}}) + \gamma_{n_t}K_{h_{n_t}}(x - X_{n_t})Y_{n_t}\,, & \text{if } n_t \geq 2\,. \end{cases} \tag{1.10}$$

As $t \to \infty$, $\hat{m}(x, t|\boldsymbol{\gamma}_t, \boldsymbol{h}_t)$ is a consistent estimator of $m(x)$ under some conditions, including $h_{n_t} \to 0$ and $\gamma_{n_t} \to 0$. See Györfi et al. (2002, Chapter 24) for details.

The estimator $\hat{m}$ at (1.9) is called semi-recursive since, although it cannot be computed directly recursively, both its numerator $\hat{r}$ and its denominator $\hat{f}$ can. In addition, there is a fully recursive kernel estimator for estimating the non-time-varying regression function m of online i.i.d. data, first introduced by Révész (1973). See Györfi et al. (2002, Chapter 25) for a discussion on the recursive regression estimator and the bibliographic notes therein. A generalisation of Révész' estimator, given by Mokkadem et al. (2009a), is defined as follows. Given parameter sequences $\boldsymbol{\gamma}_t$ and $\boldsymbol{h}_t$, defined above (1.8), the recursive kernel regression estimator is defined as

$$\begin{aligned} \hat{m}_{\text{rec}}(x, t|\boldsymbol{\gamma}_{t_1}, \boldsymbol{h}_{t_1}) &= Y_1\,, \\ \hat{m}_{\text{rec}}(x, t|\boldsymbol{\gamma}_t, \boldsymbol{h}_t) &= \big\{1 - \gamma_{n_t}K_{h_{n_t}}(x - X_{n_t})\big\}\hat{m}(x, t_{n_t-1}|\boldsymbol{\gamma}_{t_{n_t-1}}, \boldsymbol{h}_{t_{n_t-1}}) \\ &\qquad + \gamma_{n_t}K_{h_{n_t}}(x - X_{n_t})Y_{n_t}\,,\ n_t = 2, 3, \ldots \end{aligned}$$

The computation of the above recursive estimator $\hat{m}_{\text{rec}}$ is slightly easier than the semi-recursive NW estimator $\hat{m}$ at (1.9), since $\hat{m}_{\text{rec}}$ does not require computing a density estimator as in (1.8). To the best of our knowledge, although the semi-recursive and the recursive estimators both have the same convergence rate (Mokkadem et al., 2009a; Slaoui, 2016), comparison of these two estimators using simulation or real data sets is still lacking in the literature. After some preliminary simulation studies, we found that $\hat{m}_{\text{rec}}$ was often less well-behaved compared to $\hat{m}$, especially in the areas where the design density (density of the X_i's) is low. Hence we shall use the semi-recursive estimator, instead of the recursive one, to build our regression estimation method for streaming data.

By (1.8) and (1.10), the denominator $\hat{f}$ and the numerator $\hat{r}$ of the semi-recursive NW estimator are recursively defined. Note that both of them can also be expressed as a weighted sum as follows:

$$\hat{r}(x, t_{n_t}|\boldsymbol{\gamma}_t, \boldsymbol{h}_t) = \sum_{i=1}^{n_t} w_{n_t,i} K_{h_i}(x - X_i) Y_i \,, \quad \hat{f}(x, t_{n_t}|\boldsymbol{\gamma}_t, \boldsymbol{h}_t) = \sum_{i=1}^{n_t} w_{n_t,i} K_{h_i}(x - X_i) \,, \tag{1.11}$$

where

$$w_{n_t,i} = \begin{cases} \prod_{j=2}^{n_t}(1-\gamma_j) \,, & \text{for } i = 1 \,, \\ \gamma_i \prod_{j=i+1}^{n_t}(1-\gamma_j) \,, & \text{for } i = 2, \ldots, n_t \,. \end{cases} \tag{1.12}$$

Here and throughout the thesis, we take the convention that for a sequence of real numbers $\{c_j\}$, $\prod_{j=a}^{b} c_j = 1$ if $a > b$.

We conclude this section by making the observation that, with different choice of the stepsize sequence $\boldsymbol{\gamma}_t$, the weights $w_{n_t,i}$ at (1.12) can behave very differently. In the special case where $\gamma_i = 1/i$, we have $w_{n_t,i} = 1/n_t$, i.e. all past data are uniformly weighted. This is more appropriate for online i.i.d. (d.i.d.) data, since density and regression functions there do not vary with time (see §1.2). On the other hand, some works reviewed in §1.3.1 and §1.3.2, including Grillenzoni (2000), Heinz and Seeger (2008) and Harvey and Oryshchenko (2012), used the recursive KDE $\hat{f}(x, t|\boldsymbol{\gamma}_t, \boldsymbol{h}_t)$ or the semi-recursive NW estimator $\hat{m}(x, t|\boldsymbol{\gamma}_t, \boldsymbol{h}_t)$ with the choice $\gamma_i \equiv \gamma \in (0, 1)$ (Grillenzoni, 2000 and Harvey and Oryshchenko, 2012 also took $h_i \equiv h \in \mathbb{R}_+$). This leads to $w_{n_t,1} = (1-\gamma)^{n_t-1}$ and $w_{n_t,i} = \gamma(1-\gamma)^{n_t-i}$. That is, $w_{n_t,i}$ decreases exponentially fast as $n_t - i$ increases. This can be viewed as an application of exponential smoothing (see §1.3.1.2) to kernel functions,

The recursive KDE and the semi-recursive NW estimator with weights computed using exponential smoothing may be applied to modelling streaming data, in cases where the density and regression functions are time-varying. However, considering that the data stream is never-ending, using the same γ or h throughout may be unsatisfactory. In Chapters 2 and 3, we will introduce data driven ways to dynamically select the (γ, h) values for the recursive KDE and the

semi-recursive NW estimator when they are applied to streaming data. Hence our methods will be more flexible compared to the aforementioned ones based on exponential smoothing.

1.3.4.3 Infill asymptotics

In the literature on nonparametric estimation for offline i.n.i.d. (d.n.i.d.) data, asymptotics are often derived in the so-called infill asymptotics setting. See e.g. Hall et al. (2006), Vogt (2012) and Zhang and Wu (2015). For example, consider estimating the time-varying density $f(\cdot, t)$, $t \in [0, \tau]$, based on offline i.n.i.d. data $\{X_i\}_{i=1,\dots,n_\tau}$, where $\tau > 0$ is a fixed time. Instead of keeping Δt fixed and letting $t \to \infty$ as in §1.3.4.2, infill asymptotics assume that $n_\tau \to \infty$, so that, recalling (1.2), the time Δt between two consecutive arrivals is such that $\Delta t \to 0$ (recall that τ is fixed). That is, more and more data are observed in a fixed time interval. In this setting, for any $t \in [0, \tau]$, we have access to more and more data, closer and closer to t, so that we can estimate $f(\cdot, t)$ consistently. The same intuition for consistency also applies to estimating $m(\cdot, t)$.

We will use $\Delta t \to 0$ to denote the infill asymptotics instead of $n_\tau \to \infty$. Letting sample size go to infinity is a more conventional way to talk about asymptotics and is used in Hall et al. (2006), Vogt (2012) and Zhang and Wu (2015). This is because they consider observing data only on a time interval of fixed length, i.e. on $[0, \tau]$ with a fixed $\tau > 0$. However, in this thesis, the time domain $\mathbb{R}_+$ is never ending, since we consider a stream of data. Therefore, we need to be able to refer to different time intervals. Therefore, to avoid confusion, we use the expression $\Delta t \to 0$, which does not depend on a pre-determined value of τ.

As a bibliographic note, we briefly mention another asymptotic setting for nonparametric estimation for i.n.i.d. streaming data called quasi-stationarity, which is used in Rutkowski (1982a), Rutkowski (1982b), Greblicki et al. (1983), Rutkowski (1984), Rutkowski (1989a), Rutkowski (1989b), Rutkowski (2004a), Rutkowski (2004b), Duda, Jaworski et al. (2018) and Duda, Rutkowski, et al. (2018). An i.n.i.d. sequence of random variables $\{Z_i\}_{i=1,2,\dots}$ is said to be quasi-stationary if the distribution of Z_i stabilises as i becomes large. For example, Duda, Rutkowski, et al. (2018) consider nonparametric density estimation for an i.n.i.d. data stream $\{Z_i\}_{i=1,2,\dots}$, where $Z_i \in \mathbb{R}^d$, for some integer $d \geq 1$, has a density function f_i satisfying

$|f_{i+1}(x) - f_i(x)| \to 0$ for all $x \in \mathbb{R}$ as $i \to \infty$. Hence, quasi-stationarity characterises a very specific kind of nonstationarity, where the data become 'ultimately stationary'. Although this setting does not put any constraint on the arrival times of the data stream, such as the equidistant arrival times assumption at (1.1), it may still be too restrictive for some practical problems where the distribution of data does not stabilise in time. In contrast, in the infill asymptotics setting, we only require that the density or the regression function to be estimated varies smoothly in time.

1.3.4.4 KDE and NW estimator equipped with temporal kernels

In the infill asymptotics setting, Hall et al. (2006) considered the estimation of a time-varying density function of offline i.n.i.d. data $\{X_i\}_{i=1,\dots,n_\tau}$. In order to construct a consistent estimator for $f(\cdot, t)$, Hall et al. (2006) used a temporal kernel so as to do smoothing of the data over time. Hence we refer to this method as the temporal KDE. The idea is that densities $f(\cdot, s)$ and $f(\cdot, t)$ with s close to t are close to each other, so we can use data with arrival times close to t to estimate $f(\cdot, t)$. Furthermore, observations with arrival times closer to t should be given larger weights than further observations.

The estimator of Hall et al. (2006) has the form

$$\hat{f}(x,t) = \frac{t}{n_t} \sum_{i=1}^{n_t} K_{T,\lambda_{n_t}}(t - t_i) K_{h_{n_t}}(x - X_i)\,, \quad t \in [0, \tau]\,, \tag{1.13}$$

where K_T is a one-sided second-order temporal kernel supported on $[0, M)$ for some $M > 0$, satisfying $\int_0^M K_T(u)\,\mathrm{d}u = 1$, $\int_0^M u K_T(u)\,\mathrm{d}u = 0$ and $\int_0^M u^2 K_T(u)\,\mathrm{d}u = C \in \mathbb{R}_+$, and where $K_{T,\lambda_{n_t}}(\cdot) = \lambda_{n_t}^{-1} K_T(\cdot/\lambda_{n_t})$, h_{n_t} is the bandwidth and λ_{n_t} is a smoothing parameter called the temporal bandwidth[1]. Since, to compute a density at a time t, their estimator uses only the data

[1]A straightforward adaptation of equation (2.1) of Hall et al. (2006) suggests $\hat{f}(x,t) = 1/n_t \sum_{i=1}^{n_t} K_{T,\lambda_{n_t}}(t - t_i) K_{h_{n_t}}(x - X_i)$. However, the multiplier $1/n_t$ should be replaced by t/n_t, otherwise we will have $\int \hat{f}(x,t)\,\mathrm{d}x \approx 1/t$. Indeed, since $\int K_{h_{n_t}}(x - X_i)\,\mathrm{d}x = 1$, we have $\int \hat{f}(x,t)\,\mathrm{d}x = 1/n_t \sum_{i=1}^{n_t} K_{T,\lambda_{n_t}}(t - t_i)$. Noting from (1.1) and (1.2) that $t_{i+1} - t_i = \Delta t \approx t/n_t$, when n_t is large, and viewing $1/n_t \sum_{i=1}^{n_t} K_{T,\lambda_{n_t}}(t - t_i)$ as a Riemann sum, we have $\int \hat{f}(x,t)\,\mathrm{d}x \approx 1/t \int_0^t K_{T,\lambda_{n_t}}(u)\,\mathrm{d}u = 1/t \int_0^{t/\lambda_{n_t}} K_T(u)\,\mathrm{d}u \approx 1/t$, when t/λ_{n_t} is large, since $\int_0^M K_T(u)\,\mathrm{d}u = 1$.

observed up to t, it can be used for online modelling, where the data arrive sequentially. However, their estimator is not recursive. At each t, it requires computing $K_{h_{n_t}}(x - X_i)$ for each X_i. Moreover, to compute each λ_{n_t}, they use a leave-one-out cross-validation based on all past data $\{X_i\}_{i=1,\dots,n_t}$ with h_{n_t} appropriate for estimating $f(\cdot, t)$ at time t. This makes computations very heavy. In Chapter 4, we will compare the performances of the temporal kernel KDE at (1.13) and our density estimation method described in Chapter 2.

Also in the infill asymptotics setting, Vogt (2012) considered the estimation of a time-varying regression function of offline d.n.i.d. data $\{(X_i, Y_i)\}_{i=1,\dots,n_\tau}$. Similar to (1.13), his estimator also employs a temporal kernel to do the smoothing over time. The estimator of Vogt (2012) has the form

$$\hat{m}(x,t) = \frac{\sum_{i=1}^{n_\tau} K_{T,\lambda}(t - t_i) K_h(x - X_i) Y_i}{\sum_{i=1}^{n_\tau} K_{T,\lambda}(t - t_i) K_h(x - X_i)}, \quad t \in [0, \tau], \tag{1.14}$$

where K_T is a symmetric second-order temporal kernel defined on $\mathbb{R}$ and λ and h, independent from t, are the temporal bandwidth and the bandwidth. Zhang and Wu (2015) modified (1.14) into a local linear estimator, but also using symmetric K_T and time-independent λ and h. This implies that their estimators use data on both sides of t for estimating $m(x, t)$ and hence are not suitable for real-time computation. Besides, recalling that the time domain $\mathbb{R}_+$ of streaming data is never ending, using fixed λ and h for all $t \in \mathbb{R}_+$ seems inappropriate, since appropriate λ and h values may be significantly different for different time periods.

1.4 Outline of the thesis

In Chapter 2, we shall apply the recursive KDE (1.8) to streaming data by incorporating it with a streaming cross-validation (SCV) procedure which adaptively selects appropriate stepsizes and bandwidths. The SCV is motivated by the conventional leave-one-out least-squares cross-validation but is modified to estimate time-varying densities. The SCV updates the γ and h values once in a while and is computationally efficient. We shall prove that the recursive KDE can consistently estimate the time-varying density under infill asymptotics if the bandwidth and stepsize are appropriately chosen. Asymptotic optimality results for the SCV procedure will also

be provided.

In Chapter 3, we will propose a streaming regression algorithm (SRA) which applies the semi-recursive NW estimator to streaming data. Stepsizes and bandwidths for the estimator are selected using a recursive cross-validation (RCV) procedure. The SRA is inspired by the ensemble learning approach for concept drift adaptation (see §1.3.1.2). Instead of using a single semi-recursive NW estimator, the SRA computes an ensemble of them, each with different smoothing parameters. At every time point an estimator is selected using the RCV procedure. The SRA adaptively adjusts the ensemble to include up-to-date estimators and discard the outdated ones. The ensemble always contains a fixed number of estimators, hence keeping the computational cost low. Asymptotic properties of the estimator will be shown.

In Chapter 4 we shall illustrate the finite sample behaviour of our streaming density and regression estimation methods using simulations and real-data examples. Simulation results show the superiority of our methods over the the sliding window estimators introduced in §1.3.4.1. Then we apply our methods to some astronomical and financial datasets.

In Chapter 5 we shall discuss some possible extensions of our streaming density and regression estimation methods to modelling higher-dimensional data streams. These extensions will be potentially useful for building nonparametric classifiers for streaming data. In particular, we shall discuss the possibility of incorporating our density estimation method with naive Bayes to build a computationally efficient classifier for higher-dimensional streaming data.

Chapter 2

Nonparametric density estimation for streaming data

In this chapter, we propose a streaming kernel density estimator (SKDE) for a univariate i.n.i.d. data stream $\{X_i\}_{i=1,2,\dots}$ with equidistant arrival times $\{t_i\}_{i=1,2,\dots}$ defined at (1.1). The SKDE is consistent under infill asymptotics and can be computed recursively. Following Hall et al. (2006), we assume that $\{X_i\}_{i=1,2,\dots}$ is an i.n.i.d. sequence and $X_i \sim f(\cdot, t_i)$, where $f(\cdot, \cdot)$ is a slowly time-varying density. The general idea of the SKDE is to partition the data stream into blocks of observations, and apply the recursive KDE at (1.8) with the same bandwidth and stepsize for all observations within a block.

One of the major motivations for updating the smoothing parameters block by block is for improving the computational efficiency. Although the recursive KDE at (1.8) and the temporal KDE at (1.13), computed with updated smoothing parameters at each t_i, is very flexible, it is very time-consuming to compute the smoothing parameters too often using data-driven methods such as cross-validation and to recompute all terms of these estimators with those new parameters each time a new observation arises. Instead, in the streaming data setting it is more practical to recompute the smoothing parameters once in a while. This is reasonable since, when the underlying density is evolving smoothly in time, we expect that appropriate values of smoothing parameters do not change dramatically over a short period of time. This motivates us to identify

blocks of data such that, within a block, the data are nearly i.i.d. so that it is reasonable to take the same smoothing parameters. Updating tuning parameters block-wise to reduce computational cost is a common practice in the streaming data literature. See, e.g., Qin (1998) and Eliseyev et al. (2017), where the authors used this idea in recursive partial least squares regression.

2.1 Definition of SKDE

Let $\{\tau_i\}_{i=1,2,\dots}$ be a sequence of positive numbers and, for $\ell = 1, 2, \dots$, let $T_\ell = \sum_{i=1}^{\ell} \tau_i$ and take the convention that $T_0 = 0$. Partition the time horizon $\mathbb{R}_+$ into consecutive finite intervals $(0, T_1]$, $(T_1, T_2], \dots$ of respective lengths $\tau_1, \tau_2, \dots$ For $\ell = 1, 2, \dots$, let $B_\ell = \{X_{n_{T_{\ell-1}}+1}, \dots, X_{n_{T_\ell}}\}$ denote the block of data observed in the ℓ-th time interval, where n_{T_ℓ} denotes the number of data observed from time T_0 to time T_ℓ. Let $b_\ell = n_{T_\ell} - n_{T_{\ell-1}}$ be the number of observations in B_ℓ. Then, from (1.2), we have

$$\lfloor \tau_\ell / \Delta t \rfloor \le b_\ell \le \lfloor \tau_\ell / \Delta t \rfloor + 1 \,. \tag{2.1}$$

A dynamic selection of the block sizes b_ℓ will be discussed in §2.4.2. For now, we assume that they are given.

As mentioned above, we update (γ, h) block-wise for reducing computational cost. If b_ℓ is chosen to be too small, then we still need to recompute (γ, h) frequently, so that the goal for reducing computation is lost. Moreover, choosing b_ℓ small implies that we have to use a small number of data in B_ℓ to select $(\gamma_{B_\ell}, h_{B_\ell})$. This may cause numerical instability for data-driven methods such as cross-validation. On the other hand, if b_ℓ is chosen to be too large, then we have to use the same (γ, h) for an overly long time, during which the density may have changed significantly.

For each block B_j, we use a stepsize γ_{B_j} and a bandwidth h_{B_j} (we will see later in §2.4.1 how to select them). Recalling the definition of the recursive KDE $\hat{f}$ at (1.8), we define the SKDE $\check{f}$

at time t by

$$\check{f}(x,t) = \hat{f}(x,t|\check{\boldsymbol{\gamma}}_t, \check{\boldsymbol{h}}_t)\,, \tag{2.2}$$

where $\check{\boldsymbol{\gamma}}_t = \{\gamma_i\}_{i=1,\ldots,n_t}$ and $\check{\boldsymbol{h}}_t = \{h_i\}_{i=1,\ldots,n_t}$ are defined by

$$(\gamma_i, h_i) = (\gamma_{B_{j_i}}, h_{B_{j_i}})\,, \text{ for } i = 1,\ldots,n_t\,, \tag{2.3}$$

with B_{j_i} denoting the block that contains X_i. That is, for streaming data, the SKDE is defined as the recursive KDE at (1.8) except that γ and h are defined blockwise by (2.3). Recall from (1.8) that $\check{f}(x,t)$ can be recursively computed by

$$\check{f}(x,t) = \begin{cases} K_{h_1}(x - X_1)\,, & \text{if } n_t = 1\,, \\ (1-\gamma_{n_t})\check{f}(x, t_{n_t-1}) + \gamma_{n_t} K_{h_{n_t}}(x - X_{n_t})\,, & \text{if } n_t \geq 2\,. \end{cases}$$

For now, we assume that the (γ_i, h_i)'s are given and discuss some theoretical properties of the SKDE.

2.2 Consistency of SKDE

In this section, we demonstrate that, under infill asymptotics (see §1.3.4.3), the SKDE at (2.2) is asymptotically pointwise consistent. We assume that the following conditions hold:

(A1) For $j = 1, 2, \ldots,$ $h_{B_j} \in [h_m, h_M]$, $\gamma_{B_j} \in [\gamma_m, \gamma_M]$, $\gamma_{B_j}/\gamma_{B_{j+1}} - 1 = o(1)$ and $h_{B_j}/h_{B_{j+1}} - 1 = o(1)$, where $h_m = \delta\Delta t^{a_1}$, $h_M = \delta^{-1}\Delta t^{a_1}$, $\gamma_m = \delta\Delta t^{a_2}$, $\gamma_M = \delta^{-1}\Delta t^{a_2}$ and δ, a_1, a_2 are constants satisfying $\delta \in (0,1)$ and $0 < a_1 < a_2 < 1$.

(A2) f and all its partial derivatives with respect to x and t up to order 4 exist and their absolute value is bounded by M for some constant $M > 0$.

(A3) K is symmetric and satisfies $0 \leq K(u) \leq M$ for all $u \in \mathbb{R}$ with M defined at (A2). Furthermore, $\int K(u)\,\mathrm{d}u = 1$, $\int K^2(u)\,\mathrm{d}u < \infty$, $\int uK(u)\,\mathrm{d}u = 0$, $\int u^2K(u)\,\mathrm{d}u < \infty$ and $\int |u|^3K(u)\,\mathrm{d}u < \infty$.

(A4) For any integer $i \geq 1$, $\tau_i \in [\delta\Delta t^{a_3}, \delta^{-1}\Delta t^{a_3}]$ for $a_3 \in (0, 1 - a_2)$, where δ and a_1 are defined in (A1).

Recalling the block-wise definition of (γ_i, h_i) at (2.3), we can see that (A1) implies that the h_i's (the γ_i's) converge to 0 at a rate Δt^{a_1} (Δt^{a_2}) as $\Delta t \to 0$. This is different from requiring $h_k \to 0$ and $\gamma_k \to 0$ at the rate k^{-c_1} and k^{-c_2}, where $c_1 \in (0, c_2)$ and $c_2 \in (1/2, 1]$, as $k \to \infty$, as in Mokkadem et al. (2009b), where the goal is to estimate the density of online i.i.d. data. In the latter case, as $k \to \infty$ we have more and more information about the non-time-varying density, and hence can afford taking growingly smaller smoothing parameters. In the streaming data case, as time $t \to \infty$ (with Δt fixed), locally in time we still have the same amount of information about the underlying density, since the data are nonstationary. Hence (A1) requires that γ_i and h_i converge to 0 as $\Delta t \to 0$, instead of as $i \to \infty$.

Also note that (A1) is in line with the conditions of Theorem 1 in Hall et al. (2006). There, in order to prove the consistency of their temporal KDE at (1.13), they made the simplification that $\lambda_i \equiv \lambda$ and $h_i \equiv h$ and assumed that

$$h \to 0,\ \lambda \to 0 \text{ and } n_\tau^{1-\epsilon} h\lambda \to \infty \text{ for some } \epsilon \in (0, 1). \tag{2.4}$$

Now, if we take $\gamma_i \equiv \gamma$ and $h_i \equiv h$, then the SKDE at (2.2) can be written as a temporal KDE with a temporal kernel K_E and temporal bandwidth $\lambda \asymp \Delta t/\gamma$ (see §2.3 for justification). In view of the latter relation and taking $\gamma \asymp \Delta t^{a_2}$, (2.4) implies that $\Delta t/\gamma = \Delta t^{1-a_2} \to 0$, so that $a_2 < 1$. Furthermore, from (1.2) we have $n_\tau^{1-\epsilon} \asymp \Delta t^{\epsilon-1}$. Plugging this into (2.4), we have $n_\tau^{1-\epsilon} h\lambda \asymp \Delta t^{\epsilon} h/\gamma \asymp \Delta t^{a_1 - a_2 + \epsilon} \to \infty$ for some $\epsilon > 0$, which implies $a_1 < a_2 - \epsilon$. Hence from the above derivations we have $0 < a_1 < a_2 - \epsilon < 1 - \epsilon$. Since ϵ can be taken arbitrarily small, assuming this latter relation is similar to assuming $0 < a_1 < a_2 < 1$ as in (A1).

Furthermore, note that the condition $\gamma_{B_j}/\gamma_{B_{j+1}} - 1 = o(1)$ only implies that neighbouring blocks have similar γ's, but the $o(1)$ term is flexible enough for faraway blocks to have quite different γ values. To see why, note that under (A4), the number of blocks in a time interval of finite length is of order Δt^{-a_3} (reciprocal of the order Δt^{a_3} of the block length τ). Suppose γ_{B_i}

and γ_{B_j} are two γ values for blocks B_{k_1} and B_{k_2}, where $k_1 > k_2$ and $k_1 - k_2 \asymp \Delta t^{-a_3}$. Now, (A1) implies that

$$\frac{\gamma_{B_{k_1}}}{\gamma_{B_{k_2}}} = \frac{\gamma_{B_{k_1}}}{\gamma_{B_{k_1-1}}} \frac{\gamma_{B_{k_1-1}}}{\gamma_{B_{k_1-2}}} \cdots \frac{\gamma_{B_{k_2+1}}}{\gamma_{B_{k_2}}} = \{1 + o(1)\}^{k_1-k_2}, \tag{2.5}$$

which is not the same as requiring, for example, $\gamma_{B_{k_1}}/\gamma_{B_{k_2}} = 1 + o(1)$. Indeed, in general, $\{1 + o(1)\}^{k_1-k_2}$ can be very large, as long as the $o(1)$ term is not too small. For example, if the $o(1)$ term is taken as $(k_1-k_2)^{-1}$, which tends to 0 at a rate Δt^{a_3}, then $\{1+(k_1-k_2)^{-1}\}^{k_1-k_2} \sim e$ as $\Delta t \to 0$. Note too that requiring $\gamma_{B_j}/\gamma_{B_{j+1}}$ and $h_{B_j}/h_{B_{j+1}}$ to be close to 1 is only so that we can obtain the simple bias and variance formulas (2.7) and (2.8). Without these conditions, the first order terms of the bias and variance would depend on $\gamma_{B_{\ell_t-1}}$, $\gamma_{B_{\ell_t}}$, $h_{B_{\ell_t-1}}$ and $h_{B_{\ell_t}}$, instead of just on $(\gamma_{B_{\ell_t}}, h_{B_{\ell_t}})$, making those results harder to interpret.

Condition (A2) implies that the density function is sufficiently smooth in both the spatial and the temporal dimensions. Condition (A3) is standard in the literature on KDE (see, for example, Wand and Jones, 1995, pp. 19–20).

Condition (A4) implies that the block length τ_ℓ goes to zero as $\Delta t \to 0$ but not too fast. In fact, if the goal here was just to prove the consistency of $\check{f}(x,t)$ without caring about practice, then we could relax (A4) to allow $\tau_i \in [\Delta t, \infty]$. However, if $\tau_i \equiv \Delta t$, then we have $b_i \equiv 1$, i.e. each block only contains one observation. This corresponds to the case where we update (γ, h) each time we obtain a new data point, which is too time-consuming in practice. At the other extreme, $\tau_1 = \infty$ would correspond to a situation where we use pre-selected (γ, h) values for all data, including future ones, and do not update them at all, which is unrealistic in practice.

The reason why we use a more stringent version of (A4) is that it makes it possible to derive relatively simple expressions for the bias and variance of $\check{f}(x,t)$, as shown in Proposition 2.1. These expressions clearly show how these quantities depend on the choice of γ and h. On the contrary, if we do not require $a_3 \in (0, 1 - a_2)$, then virtually no meaningful results can be obtained except that the estimator is consistent. Note that Hall et al. (2006), for the same purpose, made the simplification in their theoretical derivations that the smoothing parameters for

all observations are the same and only depend on n_τ, the number of observations up to time τ (recall from §1.3.4.3 that they only consider estimating $f(\cdot, t)$ for $t \in (0, \tau]$). Here we avoid making this simplification since it is contradictory to what we actually do in practice. That is, we update (γ, h) block by block. On the one hand, the block size cannot be too large (e.g. $a_3 = 0$), since updating (γ, h) once in an overly long time is virtually the same as not updating it at all. On the other hand, the block size cannot be too small (e.g. $a_3 \geq 1 - a_2$), since updating (γ, h) too often would be time-consuming.

For a fixed time $t \in \mathbb{R}_+$, let $\ell_t \geq 1$ denote the integer such that $t \in (T_{\ell_t - 1}, T_{\ell_t}]$. That is, t is in the ℓ_t-th time interval corresponding to block B_{ℓ_t}. Note that from (1.2), we have $n_t \to \infty$ as $\Delta t \to 0$. Then, the following proposition gives the asymptotic bias and variance of $\check{f}(x, t)$ at (2.2). Let

$$\mu_{K,2} = \int u^2 K(u)\, \mathrm{d}u,\ R_K = \int K^2(u)\, \mathrm{d}u,\ f_{xx} = \frac{\partial^2 f}{\partial x^2},\ f_t = \frac{\partial f}{\partial t}, \tag{2.6}$$

and $\mathrm{bias}\{\check{f}(x, t)\} = \mathrm{E}\{\check{f}(x, t)\} - f(x, t)$.

Proposition 2.1. *Under (A1)–(A4), for any fixed $(x, t) \in \mathbb{R} \times \mathbb{R}_+$, we have, as $\Delta t \to 0$,*

$$\mathrm{bias}\{\check{f}(x, t)\} = \frac{1}{2} f_{xx}(x, t) \mu_{K,2} h_{B_{\ell_t}}^2 - f_t(x, t) \frac{\Delta t}{\gamma_{B_{\ell_t}}} + o(\Delta t^{2a_1} + \Delta t^{1 - a_2}), \tag{2.7}$$

$$\mathrm{var}\{\check{f}(x, t)\} = \frac{1}{2} R_K f(x, t) \frac{\gamma_{B_{\ell_t}}}{h_{B_{\ell_t}}} + o(\Delta t^{a_2 - a_1}). \tag{2.8}$$

Proposition 2.1 will be proved in §2.A.

Remark 2.1. As a consequence of Proposition 2.1, the mean squared error (MSE) of $\check{f}(x, t)$, defined by $\mathrm{MSE}(x, t) = \mathrm{bias}^2\{\check{f}(x, t)\} + \mathrm{var}\{\check{f}(x, t)\}$, satisfies

$$\mathrm{MSE}(x, t) = \mathrm{AMSE}(x, t) + o\{\Delta t^{4a_1} + \Delta t^{2(1 - a_2)} + \Delta t^{1 + 2a_1 - a_2} + \Delta t^{a_2 - a_1}\}, \tag{2.9}$$

where

$$\begin{aligned}\mathrm{AMSE}(x,t) = \frac{1}{4}f_{xx}^2(x,t)\mu_{K,2}^2 h_{B_{\ell_t}}^4 + f_t^2(x,t)\frac{\Delta t^2}{\gamma_{B_{\ell_t}}^2} \\ - f_{xx}(x,t)f_t(x,t)\mu_{K,2}\frac{\Delta t}{\gamma_{B_{\ell_t}}}h_{B_{\ell_t}}^2 + \frac{1}{2}f(x,t)R_K\frac{\gamma_{B_{\ell_t}}}{h_{B_{\ell_t}}}.\end{aligned} \tag{2.10}$$

Finally, the bandwidth and stepsize that minimise $\mathrm{AMSE}(x,t)$ satisfy $h_{B_{\ell_t}} \asymp \Delta t^{a_1}$ and $\gamma_{B_{\ell_t}} \asymp \Delta t^{a_2}$, with

$$a_1 = 1/7,\ a_2 = 5/7. \tag{2.11}$$

See §2.A.4 for a proof of (2.11).

Remark 2.2. Note that from Proposition 2.1 it is possible to derive a central limit theorem for the SKDE $\check{f}(x,t)$. This we leave for future work.

2.3 Remarks on convergence rate of $\check{f}(x,t)$

Taking any $h_{B_{\ell_t}}$ and $\gamma_{B_{\ell_t}}$ satisfying (2.11), we have $\mathrm{MSE}(x,t) \asymp \Delta t^{4/7}$. This implies, in view of (1.2), that $\mathrm{MSE}(x,t) \asymp n_t^{-4/7}$. This order is slightly slower than the optimal order $\mathrm{MSE}_{\mathrm{TK}}(x,t) \asymp n_t^{-2/3}$ of the temporal KDE $\hat{f}$ of Hall et al. (2006), defined at (1.13). In this section, we will show that the difference in the convergence rates of $\mathrm{MSE}(x,t)$ and $\mathrm{MSE}_{\mathrm{TK}}(x,t)$ is due to the difference in the way $\check{f}$ and $\hat{f}$ do smoothing over time. Then we discuss how to modify $\check{f}$ so that $\mathrm{MSE}(x,t)$ converges to 0 as fast as $\mathrm{MSE}_{\mathrm{TK}}(x,t)$. However, we will argue that although the SKDE $\check{f}$ has slower convergence rate, it should be preferred in the streaming data setting. To simplify the notation, we temporarily assume that $\lambda_{n_t} \equiv \lambda$, $h_{n_t} \equiv h$ in (1.13) and $\gamma_i \equiv \gamma$, $h_i \equiv h$ in (2.2), where λ, γ, h all tend to 0 as $\Delta t \to 0$.

2.3.1 $\check{f}$ as temporal KDE

In this section we illustrate the difference in the way $\check{f}$ and $\hat{f}$ do smoothing over time. For this, we will demonstrate that $\check{f}(x,t)$ is asymptotically equivalent to a temporal KDE equipped with

a first-order temporal kernel K_E. In contrast, recall from §1.3.4.4 that the temporal kernel K_T in (1.13) has order 2. It is precisely this difference in the order of temporal kernels that leads to the difference in the convergence rates of $\mathrm{MSE}(x,t)$ and $\mathrm{MSE}_{\mathrm{TK}}(x,t)$.

Next we show the asymptotic equivalence between $\check{f}$ and a temporal KDE equipped with a first-order temporal kernel K_E. First note that $\check{f}(x,t)$ does smoothing over time t using recursively computed weights $\{w_{n_t,i}\}_{i=1,\ldots,n_t}$. When $\gamma_i \equiv \gamma$, the way we compute $\{w_{n_t,i}\}_{i=1,\ldots,n_t}$ at (1.12) is the same as exponential smoothing (Harvey, 1990). As pointed out by Gijbels et al. (1999) in the context of the forecasting of a univariate time series, if $\gamma \to 0$ as $\Delta t \to 0$, then $\{w_{n_t,i}\}_{i=1,\ldots,n_t}$ generated this way can be written as

$$w_{n_t,i} = \{1 + o(1)\}\frac{t}{n_t\lambda}K_E\left(\frac{t-t_i}{\lambda}\right), \tag{2.12}$$

where λ is a temporal bandwidth satisfying $\lambda \sim \Delta t/\gamma$ and K_E is a one-sided first-order kernel supported on $[0,\infty)$, satisfying $K_E(u) > 0$ for $u \in [0,\infty)$, $\int_0^\infty K_E = 1$ and $\int_0^\infty uK_E(u)\,\mathrm{d}u = 1$. Recall from (1.11) and (2.2) that we have $\check{f}(x,t) = \sum_{i=1}^{n_t} w_{n_t,i}K_h(x - X_i)$. Hence we conclude that

$$\check{f}(x,t) = \{1 + o(1)\}\frac{t}{n_t\lambda}\sum_{i=1}^{n_t} K_E\left(\frac{t-t_i}{\lambda}\right)K_h(x - X_i)\,.$$

That is, asymptotically, the SKDE $\check{f}$ can be viewed as a special case of the temporal KDE.

Recall from (1.13) that the temporal kernel K_T used in $\hat{f}(x,t)$ has order 2. As a result of the difference in the order of K_E and K_T, $\mathrm{bias}\{\check{f}(x,t)\}$ is larger than $\mathrm{bias}\{\hat{f}(x,t)\}$, although $\mathrm{var}\{\hat{f}(x,t)\}$ and $\mathrm{var}\{\check{f}(x,t)\}$ have the same order. Specifically, from Proposition 2.1 and Hall et al. (2006) we have

$$\begin{gathered}\mathrm{var}\{\hat{f}(x,t)\} \asymp \mathrm{var}\{\check{f}(x,t)\} = O\{(n_t h\lambda)^{-1}\} \\ \mathrm{bias}\{\hat{f}(x,t)\} = O(h^2) + O(\lambda^2) \quad \text{and} \quad \mathrm{bias}\{\check{f}(x,t)\} = O(h^2) + O(\lambda)\,,\end{gathered} \tag{2.13}$$

where $O(\lambda)$ is larger than $O(\lambda^2)$ (recall from under (2.12) that $\lambda \asymp \Delta t/\gamma \to 0$ as $\Delta t \to 0$).

From (2.13), we conclude that $\mathrm{MSE}(x,t)$ converges to 0 slower than $\mathrm{MSE}_{\mathrm{TK}}(x,t)$ because $\check{f}(x,t)$ has a larger bias than $\hat{f}(x,t)$. This difference in the bias is caused by the fact that the recursively computed weights $\{w_{n_t,i}\}_{i=1,\ldots,n_t}$ used by $\check{f}(x,t)$ for smoothing over time t mimics those from a first-order temporal kernel K_E, instead of a second-order one. Now a natural question is this: can we recursively compute weights to mimic the behaviour of a second-order temporal kernel, so that, using the new weights to do smoothing over time t, the temporal bias of $\check{f}(x,t)$ has the same order as that of $\hat{f}(x,t)$? Next we show a possible way to do this.

2.3.2 Improving convergence rate of $\check{f}(x,t)$ by double-exponential smoothing

Motivated by Gijbels et al. (1999), we could use the so-called double-exponential smoothing (see, e.g. Harvey, 1990, pp. 27–28) to recursively compute weights $\{\breve{w}_{n_t,i}\}_{i=1,\ldots,n_t}$ in a way that is asymptotically equivalent to using a second-order temporal kernel K_D. That is,

$$\breve{w}_{n_t,i} = \{1+o(1)\}\frac{t}{n_t\lambda}K_D\left(\frac{t-t_i}{\lambda}\right),$$

where K_D is a one-sided second-order kernel.

Gijbels et al. (1999) investigated the properties of the double-exponential smoothing weights $\{\breve{w}_{n_t,i}\}_{i=1,\ldots,n_t}$ in the context of the smoothing of time series. They assumed model $X_i = m(t_i)+\epsilon_i$, where m is a smooth function representing the trend of a time series $\{X_i\}$ and $\{\epsilon_i\}$ is a zero mean error sequence. They considered a smoother $\hat{m}(t) = \hat{m}(t_{n_t})$ which can be recursively computed by

$$\begin{aligned}
&\hat{m}(t_2) = X_2\,, \quad g(t_2) = X_2 - X_1\,,\\
&\hat{m}(t_k) = (1-\gamma)\big\{\hat{m}(t_{k-1}) + g(t_{k-1})\big\} + \gamma X_k\,,\\
&g(t_k) = (1-\gamma)g(t_{k-1}) + \gamma\big\{\hat{m}(t_k) - \hat{m}(t_{k-1})\big\}\,,\ k = 3,\ldots,n_t\,.
\end{aligned}$$

These formulas, introduced by Holt (1957) and Winters (1960), are often used for smoothing time series with a local linear trend (in contrast, the simple exponential smoothing introduced in §1.3.1.2 is for time series with a local constant trend, see e.g. Harvey, 1990, pp. 25–28). The smoother $\hat{m}(t)$ computed this way can also be written as $\hat{m}(t) = \sum_{i=1}^{n_t} \breve{w}_{n_t,i} X_i$. See Gijbels et al. (1999) for the definition of $\{\breve{w}_{n_t,i}\}_{i=1,\ldots,n_t}$.

Instead of smoothing a time series, here we apply the weights $\{\breve{w}_{n_t,i}\}_{i=1,\ldots,n_t}$ to kernel functions $K_h(x - X_i)$ to obtain an estimator of the time-varying density $f(x,t)$ defined by $\breve{f}(x,t) = \sum_{i=1}^{n_t} \breve{w}_{n_t,i} K_h(x - X_i)$, which is asymptotically equivalent to the the following temporal KDE:

$$\frac{t}{n_t \lambda} \sum_{i=1}^{n_t} K_D\left(\frac{t - t_i}{\lambda}\right) K_h(x - X_i)\,. \tag{2.14}$$

Analogously, this estimator can be recursively computed by

$$\begin{aligned}
&\breve{f}(x,t_2) = K_h(x - X_2)\,, \quad g(x,t_2) = K_h(x - X_2) - K_h(x - X_1)\,,\\
&\breve{f}(x,t_k) = (1-\gamma)\big\{\breve{f}(x,t_{k-1}) + g(x,t_{k-1})\big\} + \gamma K_h(x - X_k)\,,\\
&g(x,t_k) = (1-\gamma) g(x,t_{k-1}) + \gamma\big\{\breve{f}(x,t_k) - \breve{f}(x,t_{k-1})\big\}\,,\ k = 2,\ldots,n_t\,.
\end{aligned} \tag{2.15}$$

Since the temporal kernel K_D in (2.14) is of order 2, under certain regularity conditions, we should be able to prove that $\text{bias}\{\breve{f}(x,t)\} = O(h^2) + O(\lambda^2)$. Recalling (2.13), this would imply that $\text{bias}\{\breve{f}(x,t)\} \asymp \text{bias}\{\hat{f}(x,t)\} = o[\text{bias}\{\check{f}(x,t)\}]$, which could motivate us to use $\breve{f}(x,t)$ instead of $\check{f}(x,t)$ for estimating a time-varying density $f(x,t)$. In addition, using (2.15) to compute $\breve{f}(x,t)$ is only slightly more time consuming than computing $\check{f}(x,t)$. Hence the estimator $\breve{f}$ looks promising. However, next we shall give some reasons why we still use $\check{f}(x,t)$ in this thesis.

First recall that we have only temporarily assumed that $(\gamma_i, h_i) \equiv (\gamma, h)$ in this section and in practice we need to update (γ, h) block by block (see §2.1). Hence, in practice, a reliable and computationally efficient method for selecting tuning parameters, including the block size b_ℓ and smoothing parameters (γ, h), for the estimator is far more important than improving its convergence rate. In fact, the weights $\{w_{n_t,i}\}_{i=1,\ldots,n_t}$ at (1.12) are easier to analyse in theory

than the weights $\{\breve{w}_{n_t,i}\}_{i=1,\dots,n_t}$ produced by formulas in (2.15). We will exploit this simplicity further in §2.4 to propose a streaming cross-validation procedure for updating (γ, h) block by block, with automatically selected block sizes b_ℓ. This procedure is justified by theoretical results in §2.4.3.

In addition, note that the weights $\{w_{n_t,i}\}_{i=1,\dots,n_t}$ at (1.12) are always positive, hence the resulting density estimate $\check{f}(x,t)$ at (2.2) is always non-negative, provided that we use a non-negative spatial kernel K. In contrast, some of the weights $\{\breve{w}_{n_t,i}\}_{i=1,2,\dots}$ take negative values (see e.g. Gijbels et al., 1999), hence the resulting density estimate $\breve{f}(x,t)$ could take negative values for some x, even if we use a non-negative spatial kernel K. Indeed, the temporal KDE at (1.13) proposed by Hall et al. (2006) suffers from the same problem. When we computed the temporal KDE using the temporal kernel $K_T(u) = (4-6u)\mathbb{1}\{u \in [0,1]\}$, the resulting density estimates $\hat{f}(x,t)$ sometimes take negative values in some x regions. See §4.1.1 for some illustrations. This is because, for a one-sided kernel K_T to be of order 2, it has to take negative values in some part of its support. This may be seen as the price to pay for the bias reduction induced by using a one-sided second-order kernel (double-exponential smoothing) instead of a one-sided first-order kernel (simple exponential smoothing). For the above reasons, we will still use the estimator $\check{f}$ in this thesis and leave the investigation of $\breve{f}$ to future work.

2.4 Selection of smoothing parameters

To select the smoothing parameters for the SKDE $\check{f}$ defined at (2.2), in this section we propose an algorithm that iteratively partitions the data stream $\{X_i\}_{i=1,2,\dots}$ into consecutive blocks B_ℓ of size b_ℓ, where $\ell = 1, 2, \dots$, and, for each ℓ, selects a stepsize γ_{B_ℓ} and a bandwidth h_{B_ℓ} from a finite set $I^\ell_{\gamma,h}$ for all data in the ℓ-th block.

To start the algorithm, we choose the size b_1 of the first block and the first set $I^\ell_{\gamma,h}$ by hand. The selection of b_1 (e.g. $b_1 = 50$) does not appear to be important since the algorithm employs a data-driven method to dynamically adjust the subsequent block sizes (see §2.4.2), and as the number of observations increases the effect of b_1 disappears quickly. Moreover, we set $I^1_{\gamma,h} =$

$I_\gamma^1 \times I_h^\ell$, where I_γ^ℓ and I_h^ℓ each contain g (an integer chosen by hand, e.g. $g = 10$) candidates for each of γ_{B_ℓ} and h_{B_ℓ}. In §2.4.2 we will discuss how to iteratively select $b_2, b_3, \ldots$ and construct $I_{\gamma,h}^2, I_{\gamma,h}^3, \ldots$ for future blocks. At this stage we assume temporarily that b_ℓ and $I_{\gamma,h}^\ell$ have already been selected for $\ell = 1, 2, \ldots$ and propose a streaming cross-validation (SCV) procedure to select $(\gamma_{B_\ell}, h_{B_\ell}) \in I_{\gamma,h}^\ell$.

2.4.1 Selection of γ_{B_ℓ} and h_{B_ℓ}

To introduce the SCV, we first briefly review a conventional cross-validation method for the conventional KDE defined at (1.4), which is based on the integrated squared error (ISE). Next we define two cross-validation procedures suitable in our setting of streaming data. Both procedures are useful for the selection of γ_{B_ℓ} and h_{B_ℓ} and in §2.4.2 we will show how to use them to select the block size b_ℓ.

2.4.1.1 Least squares cross-validation

The standard least squares cross-validation for the conventional KDE defined at (1.4), computed from offline i.i.d. data, aims at selecting the bandwidth by minimising an estimator of the ISE (see e.g. Hall, 1983), which, for the offline KDE for i.i.d. data, is given by

$$\mathrm{ISE}_{\text{off}}(h) = \int \{\hat{f}(x) - f(x)\}^2 \,\mathrm{d}x = \int \hat{f}^2(x)\,\mathrm{d}x - 2\int \hat{f}(x) f(x)\,\mathrm{d}x + \int f^2(x)\,\mathrm{d}x\,, \quad (2.16)$$

noting that $\hat{f}(x)$, defined at (1.4), depends on h. Since the last integral in (2.16) does not depend on h, minimising $\mathrm{ISE}_{\text{off}}(h)$ is equivalent to minimising $\int \hat{f}^2(x)\,\mathrm{d}x - 2\,\mathrm{E}\{\hat{f}(X)|X_1, \ldots, X_n\}$ where $X \sim f$ is a random variable independent of the X_i's. This can be estimated by

$$\mathrm{CV}_{\text{off}}(h) = \int \hat{f}^2(x)\,\mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i). \quad (2.17)$$

Here, $\hat{f}_{-i}$ denotes the density estimator using a bandwidth h and constructed using all data except X_i.

2.4.1.2 Two cross-validation procedures for streaming data

Now we return to the task of selecting $(\gamma_{B_\ell}, h_{B_\ell})$ for block B_ℓ, $\ell = 1, 2, \ldots$ for computing the estimator $\check{f}(\cdot, t)$ at a time $t \in (T_{\ell-1}, T_\ell]$, when we have already selected (γ_{B_i}, h_{B_i}) for data in blocks B_j, $j = 1, \ldots, \ell - 1$. Since (γ_{B_i}, h_{B_i}) have already been selected for $i = 1, \ldots, \ell - 1$ and we need to select (γ, h) for block B_ℓ, we take $\check{f}$ as at (2.2), but take, on this occasion,

$$(\gamma_i, h_i) = \begin{cases} (\gamma_{B_{j_i}}, h_{B_{j_i}}), & \text{if } i = 1, \ldots, n_{T_{\ell-1}}, \\ (\gamma, h), & \text{if } i = n_{T_{\ell-1}} + 1, \ldots, n_{T_\ell}, \end{cases} \tag{2.18}$$

where B_{j_i} denotes the block that contains X_i. Note that in the case $\ell = 1$ since $T_0 = 0$ (see page 25) so that $n_{T_0} = 0$, (γ_i, h_i) is only defined by the second line above and we have $(\gamma_i, h_i) \equiv (\gamma, h)$. Next we define two cross-validation criteria that are both useful extensions of CV_{off} defined at (2.17) to the setting of streaming data.

Recall from page 25 that B_ℓ is the block of data corresponding to the time interval $(T_{\ell-1}, T_\ell]$. So, ideally, to choose $(\gamma_{B_\ell}, h_{B_\ell})$ we would like to minimise the ISE on $(T_{\ell-1}, T_\ell]$, i.e.

$$\text{ISE}_{(T_{\ell-1}, T_\ell]}(\gamma, h) = \int_{T_{\ell-1}}^{T_\ell} \int \{\check{f}(x, t) - f(x, t)\}^2 \, \mathrm{d}x \, \mathrm{d}t \,. \tag{2.19}$$

This leads to the following average cross-validation (ACV) criterion:

$$\text{ACV}_\ell(\gamma, h) = \frac{\tau_\ell}{b_\ell} \sum_{X_i \in B_\ell} \left\{ \int \check{f}^2(x, t_i) \, \mathrm{d}x - 2\check{f}_{-i}(X_i, t_i) \right\}, \tag{2.20}$$

where $\check{f}_{-i}(x, t_i) = \check{f}(x, t_i) - w_{i,i} K_h(x - X_i)$ is the leave-one-out estimator of $f(x, t_i)$ with $w_{i,i}$ defined as $w_{n_t,i}$ in (1.12), taking $n_t = n_{t_i} = i$, and (γ_j, h_j) defined as in (2.18) for $j \leq i$. See Appendix 2.B for the derivation of the ACV criterion.

In addition to $\mathrm{ISE}_{(T_{\ell-1},T_\ell]}$ defined at (2.19), which measures the average behaviour of $\check{f}$ on the time interval $(T_{\ell-1}, T_\ell]$, we can also consider the following ISE:

$$\mathrm{ISE}_\ell(\gamma, h) = \int \{\check{f}(x, s_\ell) - f(x, s_\ell)\}^2 \, \mathrm{d}x \,, \tag{2.21}$$

where $s_\ell = (T_{\ell-1} + T_\ell)/2$ denotes the centre of time interval $(T_{\ell-1}, T_\ell]$ and where $\check{f}(x, s_\ell)$ is defined as $\check{f}(x, t)$ at (2.2), but replacing there t by s_ℓ and using (γ_i, h_i) as defined by (2.18). Note that (2.21) measures the behaviour of the SKDE at only one time point s_ℓ instead of on the whole time interval $(T_{\ell-1}, T_\ell]$. This is reasonable when $\tau_\ell = T_\ell - T_{\ell-1}$ is not too large so that $f(\cdot, t)$ does not change too much within $(T_{\ell-1}, T_\ell]$.

Now, observe from (2.21) that

$$\mathrm{ISE}_\ell(\gamma, h) = \int \check{f}^2(x, s_\ell) \, \mathrm{d}x - 2 \int \check{f}(x, s_\ell) f(x, s_\ell) \, \mathrm{d}x + \int f^2(x, s_\ell) \, \mathrm{d}x \,, \tag{2.22}$$

where the first integral can be computed directly from data and the third integral does not depend on (γ, h), the parameters of interest. As for the second integral, we have

$$\int \check{f}(x, s_\ell) f(x, s_\ell) \, \mathrm{d}x = \mathrm{E}_{s_\ell}\{\check{f}(X^0_{s_\ell}, s_\ell) \mid X_1, \ldots, X_{n_{s_\ell}}\} \,, \tag{2.23}$$

where n_{s_ℓ} denotes the number of observations arriving up to time s_ℓ and E_{s_ℓ} denotes the expectation with respect to $f(\cdot, s_\ell)$, and where $X^0_{s_\ell} \sim f(\cdot, s_\ell)$ and is independent of $\{X_1, \ldots, X_{n_{s_\ell}}\}$.

We can estimate (2.23) by the empirical mean

$$\frac{1}{b_\ell} \sum_{X_i \in B_\ell} \check{f}_{-i}(X_i, s_\ell) \,, \tag{2.24}$$

where

$$\check{f}_{-i}(x, s_\ell) = \begin{cases} \check{f}(x, s_\ell) - w_{n_{s_\ell},i} K_h(x - X_i), & \text{for } i = n_{T_{\ell-1}} + 1, \ldots, n_{s_\ell}, \\ \check{f}(x, s_\ell), & \text{for } i = n_{s_\ell} + 1, \ldots, n_{T_\ell}, \end{cases} \tag{2.25}$$

$w_{n_{s_\ell},i}$ is defined as $w_{n_t,i}$ in (2.2), but with t substituted by s_ℓ and (γ_j, h_j) defined as in (2.18) for $j \leq n_{s_\ell}$. In other word, like $\hat{f}_{-i}$ at (2.17), $\check{f}_{-i}(x, s_\ell)$ is a leave-one-out estimator of $f(x, s_\ell)$ computed from the data $\{X_1, \ldots, X_{n_{s_\ell}}\} \backslash \{X_i\}$. However, we see from (2.2) that, for $i > n_{s_\ell}$, X_i is not used to compute $\check{f}_{-i}(x, s_\ell)$, so that $\check{f}_{-i}(x, s_\ell) = \check{f}(x, s_\ell)$. We use observations with arrival times on both sides of time s_ℓ at (2.24) to reduce the bias. Now we have derived the following cross-validation criterion:

$$\mathrm{SCV}_\ell(\gamma, h) = \int \check{f}^2(x, s_\ell)\, \mathrm{d}x - \frac{2}{b_\ell} \sum_{X_i \in B_\ell} \check{f}_{-i}(X_i, s_\ell)\,. \tag{2.26}$$

Given the block size b_ℓ and set $I^\ell_{\gamma,h}$, both ACV_ℓ and SCV_ℓ defined in this section are reasonable for selecting $(\gamma_{B_\ell}, h_{B_\ell})$. Moreover, when b_ℓ (or equivalently, τ_ℓ, due to (2.1)) is chosen in a way that the true density f does not vary too much on time interval $(T_{\ell-1}, T_\ell]$, so that the data in B_ℓ are nearly i.i.d., the smoothing parameters selected according to ACV_ℓ and SCV_ℓ should be close to each other. This is because, on this occasion, the behaviour of $\check{f}$ at s_ℓ (measured by ACV_ℓ), the centre of the time interval $(T_{\ell-1}, T_\ell]$, should be able to represent its behaviour on the whole interval (measured by SCV_ℓ). Hence a choice of (γ, h) which minimises $\mathrm{SCV}_\ell(\gamma, h)$ should also yield a small $\mathrm{ACV}_\ell(\gamma, h)$.

However, recall that in this section we only temporarily assume that the block sizes b_ℓ and the sets $I^\ell_{\gamma,h}$, for $\ell = 2, 3, \ldots$, are already given, but a data-driven method for selecting b_ℓ and $I^\ell_{\gamma,h}$ is still lacking. In §2.4.2 we will discuss a method for constructing $I^\ell_{\gamma,h}$, which is motivated by theoretical results in §2.2. Moreover, it turns out that the SCV criterion enjoys some theoretical advantage which we can further exploit for the selection of b_ℓ. In §2.4.2 we will discuss a method for selecting b_ℓ based on some theoretical properties of the SCV in §2.4.3, which makes use of both SCV_ℓ and ACV_ℓ.

2.4.2 Selection of b_ℓ and $I^\ell_{\gamma,h}$

We propose an iterative procedure for selecting the block size b_ℓ and the candidate set $I^\ell_{\gamma,h}$ as follows. Suppose that, for $j = 1, \ldots, \ell - 1$, we have selected γ_j, h_j and b_j. Now we need to select b_ℓ and $I^\ell_{\gamma,h}$, and then can we select $(\gamma_{B_\ell}, h_{B_\ell})$ using either ACV_ℓ at (2.20) or SCV_ℓ at (2.26).

To construct $I^\ell_{\gamma,h}$, first note that, with block sizes $b_{b_\ell - 1}$ and b_ℓ appropriately chosen, we do not expect $(\gamma_{B_{\ell-1}}, h_{B_{\ell-1}}) \in I^{\ell-1}_{\gamma,h}$ and $(\gamma_{B_\ell}, h_{B_\ell}) \in I^\ell_{\gamma,h}$ to be significantly different. Motivated by this fact, we let $I^\ell_\gamma \subset [\gamma_{B_{\ell-1}}/M_1,\, M_1\gamma_{B_{\ell-1}}]$ and $I^\ell_h \subset [h_{B_{\ell-1}}/M_2, M_2 h_{B_{\ell-1}}]$ be two equidistant grids around $\gamma_{B_{\ell-1}}$ and $h_{B_{\ell-1}}$, each containing g elements, and then, let $I^\ell_{\gamma,h} = I^\ell_\gamma \times I^\ell_h$. Here $M_1, M_2 > 1$ are some constants chosen by hand. In practice, we can take, e.g., $M_1 = M_2 = 1.2$. This is flexible enough since, if the initial $I^1_{\gamma,h}$ is inappropriate, consecutive $I^\ell_{\gamma,h}$ can be gradually adjusted to have more appropriate range. Moreover, taking M_1, M_2 too large is likely to cause numerical instability for some cross-validation procedures, since the corresponding cross-validation criteria often have multiple local minima (Hall and Marron, 1991).

The procedure for selecting b_ℓ is motivated by theoretical results about SCV_ℓ in §2.4.3. There, Proposition 2.2 and Theorem 2.1 together imply that minimising SCV_ℓ over γ and h is asymptotically equivalent to minimising ISE_ℓ at (2.21) or $\text{MISE}_\ell(\gamma, h) = \int \text{MSE}(x, s_\ell)\, \mathrm{d}x$ over γ and h. This type of properties is sometimes referred to as the asymptotic optimality of a cross-validation procedure (Hall, 1983), which justifies that using the corresponding cross-validation criterion to select smoothing parameters is asymptotically effective.

In order for the asymptotic optimality of the SCV to hold, Conditions (A1)–(A4) and (B1)–(B3) in §2.4.3 have to be satisfied. In particular, (A4) and (B3) imply that that $\tau_\ell \asymp \Delta t^{a_3}$ with $a_3 \in (4/21, 2/7)$, which, recalling (2.1), implies that $b_\ell \asymp \Delta t^{1-a_3}$. Since a_3 here can only be taken in a relatively narrow interval $(4/21, 2/7)$, the latter results reflect the trade-off involved in selecting b_ℓ. From there, using (A1), (B1) and (B3), we have

$$b_\ell \asymp \gamma_{B_{\ell-1}}^{-(5+\alpha)/5},\ \alpha \in (0, 2/3)\,. \tag{2.27}$$

This result suggests that, if we select an $\alpha \in (0, 2/3)$ (in the numerical examples in Chapter 4, we choose $\alpha = 1/3$), then we can let

$$b_\ell = C_b^\ell \gamma_{B_{\ell-1}}^{-(5+\alpha)/5} \tag{2.28}$$

for some constant $C_b^\ell > 0$. Given α in the right range, it remains to select C_b^ℓ in practice.

Now, with (2.28), we still need to select the constant term C_b^ℓ, whose value can be very important in practice. It is a difficult task since we do not have any analytical form for C_b^ℓ and it is hard to decide what is an appropriate value for it in practice. Here we propose a heuristic solution to this problem, which, although cannot be justified by theory at this stage, is a principled method. The main motivation is that, if the block size is appropriately chosen, then the bandwidths chosen by the SCV at (2.26), i.e.

$$(\gamma_{B_\ell}, h_{B_\ell}) = \underset{(\gamma,h)\in I_{\gamma,h}^\ell}{\arg\min}\ \mathrm{SCV}_\ell(\gamma, h)\,, \tag{2.29}$$

and those chosen by minimising the ACV criterion at (2.20), i.e.

$$(\gamma_{B_\ell}^A, h_{B_\ell}^A) = \underset{(\gamma,h)\in I_{\gamma,h}^\ell}{\arg\min}\ \mathrm{ACV}_\ell(\gamma, h)\,, \tag{2.30}$$

should be relatively close. However, when bandwidths chosen by these two criteria are significantly different, then this suggests that the block size is not right and we should consider adjusting the future block sizes. The next example illustrates an important scenario where the SCV bandwidths and the ACV bandwidths may be very different. We will discuss how to adjust the constant term C_b^ℓ accordingly.

Figure 2.1 illustrates the role of the block size. In both subplots, the true density f has the form $f(x, t) = f_0(x - at)$, where f_0 is a normal mixture with four components and $a > 0$ is a constant. That is, f is a normal mixture density changing its location at a constant speed from left to right on the x axis. The three density curves in different shades of grey in each subplot in Figure 2.1 represent $f(\cdot, T_{\ell-1})$, $f(\cdot, s_\ell)$ and $f(\cdot, T_\ell)$, respectively, where $T_{\ell-1}$ and T_ℓ are defined

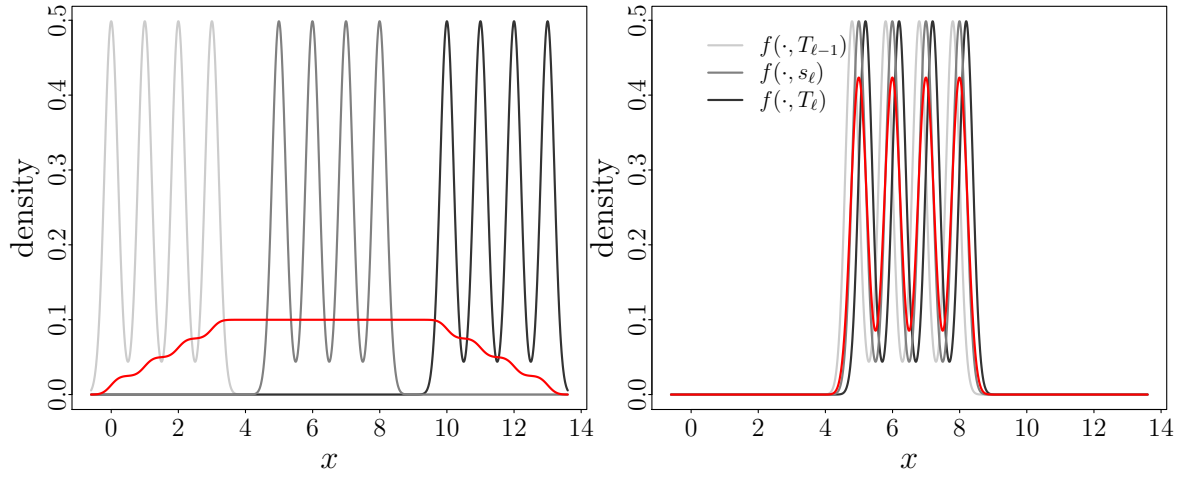

Figure 2.1: Example illustrating the role of block size. In both subplots: normal mixture with four modes moving from left to right in different shades of grey, average density $\int_{T_{\ell-1}}^{T_\ell} f(\cdot, t)\,\mathrm{d}t$ in red. The left subplot corresponds to an over-sized block, the right an appropriately small block.

on page 25 and s_ℓ is defined below (2.21). The left subplot corresponds to a large block length τ_ℓ (hence a large block size b_ℓ) value and the right subplot corresponds to a small block length τ_ℓ (hence a small block size b_ℓ) value.

Recall from §2.4.1 that the SCV at (2.26) measures the behaviour of the estimator $\check{f}$ at the centre of the block, s_ℓ. Also recall that, by using the empirical mean (2.24) to estimate the integral (2.23), the SCV relies on the fact that data $X_i \sim f(\cdot, t_i)$ in B_ℓ are nearly i.i.d., that is, $f(\cdot, t_i) \approx f(\cdot, s_\ell)$, which holds only when the block size b_ℓ is appropriately chosen. In the case illustrated in the left subplot in Figure 2.1, where the block size is too large, the average density $\int_{T_{\ell-1}}^{T_\ell} f(\cdot, t)\,\mathrm{d}t$, represented by the red curve, is rather flat. On this occasion, the density at the centre $f(\cdot, s_\ell)$ (grey curve at the middle), is a very poor representation of the average density. In contrast, in the case illustrated in the right subplot in Figure 2.1, where block size is appropriately small, the average density (red line) is relatively close to $f(\cdot, s_\ell)$.

Now, in the case of Figure 2.1, if the the block size is too large, how will the SCV and the ACV behave? For the SCV, considering that its knowledge about $f(\cdot, s_\ell)$ comes from a very heterogeneous subsample B_ℓ, whose density on average is represented by a flat curve, it tends to choose a larger bandwidth. On the other hand, with the same block size, the ACV at (2.20)

tends to choose a much smaller bandwidth. This is because the ACV measures the behaviour of the estimator $\check{f}$ at different time points in $(T_{\ell-1}, T_\ell]$ instead of focusing just on the centre.

The above example shows how SCV bandwidths can be different from the ACV bandwidths when the true density has large curvature. However, when the true densities are very smooth, the difference is not so significant since, in that case, even the average density over an overly long time interval would look close to the density at the centre of the interval. This suggests that, often, if the SCV bandwidths and the ACV bandwidths are very different, it is because the block size has been chosen too large for the SCV.

Based on the above considerations, we propose the following iterative method to select C_b^ℓ, $\ell = 1, 2, \ldots$ First, to initialise the value of C_b^1, we first choose b_1 by hand as discussed at the beginning of §2.4 and then use the SCV to select (γ_{B_1}, h_{B_1}) as at (2.29). Now, by (2.28), we can take $C_b^1 = b_1 \gamma_{B_1}^{(5+\alpha)/5}$. Then, to adjust future values of C_b^ℓ, we select smoothing parameters by the ACV using (2.30) and let $H^1 = \{h_{B_1} - h_{B_1}^A\}$, where, recalling from (2.29) and (2.30) h_{B_1} and $h_{B_1}^A$ denote the h values for block B_1 selected by SCV and ACV, respectively. For block B_ℓ, $\ell = 2, 3, \ldots$, let $H^\ell = H^{\ell-1} \cup \{h_{B_\ell} - h_{B_\ell}^A\}$. That is, H^ℓ is a set containing the differences between the bandwidths chosen by SCV and ACV.

Whenever H^ℓ contains more than a certain number (e.g. 5) of elements, run a one-sided Wilcoxon signed-rank test (Rey and Neuhäuser, 2011) to see if elements in H^ℓ are significantly larger than 0, for which we calculate the exact p-value. The null hypothesis is that the differences $h_{B_i} - h_{B_i}^A$ are independent and come from a continuous distribution symmetric about zero. In contrast, the null hypothesis of the sign test does not assume symmetry, resulting in smaller statistical power (Rey and Neuhäuser, 2011). Although in general the independence assumption does not hold here, we can view the differences $h_{B_i} - h_{B_i}^A$ as approximately independent, since the bandwidths mainly use data from separate blocks.

Now, if the null hypothesis is rejected under significance level 0.05, we reduce the future value $C_b^{\ell+1}$ by a scale of 1.05, i.e. $C_b^{\ell+1} = C_b^\ell / 1.05$, and then set $H^\ell = \varnothing$. Otherwise we let $C_b^{\ell+1} = C_b^\ell$. When the values of the constant term have remained unchanged for a while, say, for 5 blocks, we increase $C_b^{\ell+1}$ by a scale of 1.05, i.e. let $C_b^{\ell+1} = 1.05 C_b^\ell$, and then set $H^\ell = \varnothing$. The

motivation of this iterative procedure is that, if there is enough evidence showing that bandwidths selected by ACV are smaller than those by SCV, then we slightly reduce the constant term, so that future block sizes may be smaller. Otherwise, we slightly increase it, so that future block sizes may be larger. This is because, ideally we would want to use more data to compute the cross-validation criteria, as long as the distributions of these data are not too different.

Note that when selecting C_b^ℓ we carry out a nonparametric test with sample size as small as 5. On this occasion, the smallest p-value of the exact test can get is 0.03125, smaller than the significance level 0.05, at which the null hypothesis can be rejected.

2.4.3 Some asymptotic results for SCV

In this section, we provide some asymptotic results about the SCV defined at (2.26), under the same setting as in §2.2. These results will be useful to guide our selection of the block sizes b_ℓ, $\ell = 2, 3, \ldots$, assumed given in §2.4.1 (see page 35).

As in §2.2, for a given time $t \in \mathbb{R}_+$, let ℓ_t denote the integer such that $t \in (T_{\ell_t - 1}, T_{\ell_t}]$. Recall from (2.29) that the SCV selects $(\gamma_{B_{\ell_t}}, h_{B_{\ell_t}})$ from a candidate set $I_{\gamma,h}^{\ell_t}$ for block B_{ℓ_t} using the cross-validation criterion at (2.26), which measures the behaviour of the SKDE at the centre s_{ℓ_t} of the time interval $(T_{\ell_t - 1}, T_{\ell_t}]$ corresponding to B_{ℓ_t}. In addition to conditions (A1)–(A4), we assume that the following conditions hold:

(B1) $a_1 = 1/7$ and $a_2 = 5/7$, where a_1 and a_2 are defined in (A1). The cardinality $\#(I_{\gamma,h}^{\ell_t})$ of $I_{\gamma,h}^{\ell_t}$ satisfies $\#(I_{\gamma,h}^{\ell}) \leq C\Delta t^{-a}$ for some constants $C > 0$ and $a > 0$.

(B2) For g being any of the partial derivatives of f with respect to x and t up to order 3, $\int |g(x,t)| \, \mathrm{d}x < M$ for any $t \in (0, T_{\ell_t}]$, where M is defined in (A2).

(B3) $a_3 = (2 - \alpha)/7$ for some $0 < \alpha < 2/3$, where a_3 is defined in (A4).

In addition to (A1), (B1) assumes that (2.11) holds, so that $\mathrm{AMSE}(x,t)$ at (2.10) has the fastest convergence rate $\Delta t^{4/7}$. In addition, (B1) requires that the candidate set $I_{\gamma,h}^{\ell_t}$ is finite but its size

can increase polynomially fast as $\Delta t \to 0$. These are both standard conditions (see e.g. Marron and Härdle, 1986, p. 96).

In addition to (A2), (B2) requires absolute integrability of some density derivatives, that is, it ensures that the functions are all smooth enough.

Conditions (B1) and (B3) imply that $a_3 \in (4/21, 2/7)$. This is stronger than (A4) since, under (B1), (A4) only requires $a_3 \in (0, 2/7)$. That is, by (B1) we further restrict $\tau_i \asymp \Delta t^{a_3}$ to be not too large (a_3 strictly no smaller than $4/21$). Recall from §2.4.1.2 that this is because SCV_{ℓ_t} estimates ISE_{ℓ_t} at (2.21). The latter is only a good measure of the behaviour of $\check{f}$ on the time interval $(T_{\ell_t-1}, T_{\ell_t}]$ when $\tau_{\ell_t} = T_{\ell_t} - T_{\ell_t-1}$ is not too large.

Now we present some standard theoretical results concerning the optimality of the SCV similar to the main theorem in Fan et al. (1996). Recall the definitions of SCV_ℓ at (2.26) and $\mathrm{MSE}(x, s_{\ell_t})$ (2.9). Let $\mathrm{MISE}_{\ell_t}(\gamma, h) = \int \mathrm{MSE}(x, s_{\ell_t})\,\mathrm{d}x$. Recall also the definition of ℓ_t above (B1). Next we give some results regarding the relationships amongst ISE_{ℓ_t}, MISE_{ℓ_t} and SCV_{ℓ_t}.

Proposition 2.2. *Under (A1)–(A4) and (B1)–(B3), we have*

$$\mathrm{ISE}_{\ell_t}(\gamma, h) = \mathrm{MISE}_{\ell_t}(\gamma, h) + o_p(\Delta t^{4/7})\,, \tag{2.31}$$

uniformly $(\gamma, h) \in I^{\ell_t}_{\gamma,h}$ *as* $\Delta t \to 0$.

Theorem 2.1. *Under (A1)–(A4) and (B1)–(B3), we have*

$$\mathrm{SCV}_{\ell_t}(\gamma, h) = \mathrm{ISE}_{\ell_t}(\gamma, h) + \int f^2(x, s_{\ell_t})\,\mathrm{d}x - \frac{2}{b_{\ell_t}} \sum_{X_i \in B_{\ell_t}} f(X_i, s_{\ell_t}) + o_p(\Delta t^{4/7}), \tag{2.32}$$

uniformly in $(\gamma, h) \in I^{\ell_t}_{\gamma,h}$ *as* $\Delta t \to 0$.

See Appendices 2.C and 2.D for the proof of the above results. Here we give the intuition why $\mathrm{SCV}_{\ell_t}(\gamma, h)$ is a consistent estimator of $\mathrm{ISE}_{\ell_t}(\gamma, h) - \int f^2(x, s_{\ell_t})\,\mathrm{d}x$ with $\mathrm{ISE}_{\ell_t}(\gamma, h)$ at (2.22). Note that $X_i \sim f(\cdot, t_i)$ and $f(\cdot, t_i)$ is close to $f(\cdot, s_{\ell_t})$ since, under (A2), (B2) and (B3), t_i is close to s_{ℓ_t} and f is smooth in t. Hence $b_{\ell_t}^{-1} \sum_{X_i \in B_{\ell_t}} f(X_i, s_{\ell_t})$ can be seen as an empirical mean

consistently estimating $\int f^2(x, s_{\ell_t})\,\mathrm{d}x$. This, together with (2.32), implies that $\mathrm{SCV}_{\ell_t}(\gamma, h)$ is indeed a consistent estimator of $\mathrm{ISE}_{\ell_t}(\gamma, h) - \int f^2(x, s_{\ell_t})\,\mathrm{d}x$.

Now we state the major implication of Proposition 2.2 and Theorem 2.1. According to Remark 2.1, $\Delta t^{4/7}$ is the smallest possible order of $\mathrm{MSE}(x, s_{\ell_t})$. Under some additional regularity conditions, we can show that this order is also the smallest possible order of $\mathrm{MISE}_{\ell_t}(\gamma, h)$ (we will prove this in future work). Therefore, Proposition 2.2 implies that the smallest possible order of $\mathrm{ISE}_{\ell_t}(\gamma, h)$ is also $\Delta t^{4/7}$. Furthermore, since the terms $\int f^2(x, s_{\ell_t})\,\mathrm{d}x$ and $2/b_{\ell_t} \sum_{X_i \in B_{\ell_t}} f(X_i, s_{\ell_t})$ in (2.32) do not depend on γ nor h, similar to Fan et al. (1996), we conclude that minimising SCV_{ℓ_t} over γ and h is asymptotically equivalent to minimising ISE_{ℓ_t} over γ and h.

Although SCV is designed for i.n.i.d. streaming data, it may be used to select smoothing parameters for d.n.i.d. streaming data. Indeed, for density estimation for offline i.i.d. data, Hall, Lahiri and Truong (1995) observed that 'even for some strongly dependent data sequences, the asymptotically optimal bandwidth for independent data is a good choice'. This encourages us to investigate the behaviour of SCV for dependent data streams.

For an illustration of the finite-sample behaviour of the SKDE and the SCV, see §4.1.1 and §4.2.1 for some simulation studies and real-data examples.

Appendix

To simplify the notation, we suppress the subscript t of ℓ_t defined above Proposition 2.1 in all appendices to Chapter 2. That is, for a given time $t > 0$, let $\ell = \ell_t$ denote the unique integer satisfying $t \in (T_{\ell-1}, T_\ell]$.

2.A Proof of Proposition 2.1

2.A.1 Notation

First, we present a summary of notations that will be used in this section.

Times and sample sizes. For online i.n.i.d. data $\{X_i\}_{i=1,2,\ldots}$, let $\{t_i\}_{i=1,2,\ldots}$ denote the equidistant arrival times defined in (1.1). The time horizon is partitioned into finite intervals $(T_0, T_1], (T_1, T_2], \ldots$ of respective lengths $\tau_1 = T_1 - T_0, \tau_2 = T_2 - T_1, \ldots$, where $T_0 = 0$. For $k = 1, 2, \ldots$, $B_k = \{X_{n_{T_{k-1}}+1}, \ldots, X_{n_{T_k}}\}$ denotes the block of data observed in the k-th time interval $(T_{k-1}, T_k]$, where n_{T_ℓ} denotes the number of data observed up to time T_ℓ. The number of observations in B_k is denoted by $b_k = n_{T_k} - n_{T_{k-1}}$. For a given time $t > 0$, n_t denotes the number of data observed up to t.

Smoothing parameters. When defining the streaming KDE in (2.2) and (1.12), we used two notations $\{(\gamma_i, h_i)\}_{i=1,\ldots,n_t}$ and $\{(\gamma_{B_i}, h_{B_i})\}_{i=1,\ldots,n_t}$ denoting stepsize and bandwidth used by $\check{f}(x,t)$.

Convention on almost sure relations. We write $Z \leq c$, for some random variable Z and real number c, if Z is bounded by c almost surely.

Estimators. From (1.11), (1.12) and (2.2), we have

$$\check{f}(x,t) = \sum_{i=1}^{n_t} w_{n_t,i} K_{h_i}(x - X_i), \tag{2.A.1}$$

where $w_{n_t,i}$ is defined at (1.12) with blockwise-defined (γ_i, h_i) at (2.3). In addition, let

$$\bar{f}(x,t) = \sum_{i=n_{T_{\ell-2}}+1}^{n_t} w_{n_t,i} K_{h_i}(x - X_i). \tag{2.A.2}$$

2.A.1.1 Technical lemmas

The following formula, known as summation by parts, will be used repeatedly in all appendices.

Lemma 2.A.1. *Let $\{f_k\}$ and $\{g_k\}$ be two sequences of real numbers. Then*

$$\sum_{k=0}^{n} f_k g_k = f_0 \sum_{k=0}^{n} g_k + \sum_{j=0}^{n-1} (f_{j+1} - f_j) \sum_{k=j+1}^{n} g_k. \tag{2.A.3}$$

Using the notation of Proposition 2.1, let $t \in \mathbb{R}_+$ be a given time and let $\ell \geq 1$ denote the integer satisfying $t \in (T_{\ell-1}, T_\ell]$. The following lemma implies that $\bar{f}(x,t)$ at (2.A.2) is asymptotically equivalent to $\check{f}(x,t)$ at (2.2).

Lemma 2.A.2. *Under (A1)–(A4), for any $x \in \mathbb{R}$, we have,*

$$\check{f}(x,t) - \bar{f}(x,t) \leq 2\exp(-3^{-1}\delta^2 \Delta t^{a_2+a_3-1}), \tag{2.A.4}$$

where $\bar{f}(x,t)$ is defined in (2.A.2).

Remark 2.A.1. From (A4), we have $a_2 + a_3 - 1 < 0$. Therefore, (2.A.4) implies that the difference between $\check{f}(x,t)$ and $\bar{f}(x,t)$ decreases to zero exponentially fast as $\Delta t \to 0$.

Proof of Lemma 2.A.2. Throughout the proof we assume that Δt is small enough such that $n_t \geq 2$. Recall from (2.2) that, for any $i = 1, \ldots, n_t$, we have $\gamma_i = \gamma_{B_{j_i}}$ and $h_i = h_{B_{j_i}}$ if X_i is in block B_{j_i}. Then, note that $\check{f}(x,t) = S_1 + S_2$, where

$$S_1 = \prod_{j=2}^{n_t}(1-\gamma_j)K_{h_1}(x - X_1) \text{ and } S_2 = \sum_{i=2}^{n_t}\gamma_i \prod_{j=i+1}^{n_t}(1-\gamma_j)K_{h_i}(x - X_i). \tag{2.A.5}$$

Then we bound S_1 and S_2 as follows.

Bounding S_1. Under (A3), we have $K_{h_i}(x - X_i) \leq M/h_i$ for $i = 1, 2, \ldots$ Under (A1), we have, from (2.A.5),

$$S_1 \leq \frac{M}{h_1}\prod_{j=2}^{n_t}(1-\gamma_j) \leq \frac{M}{h_1}\prod_{j=n_{T_{\ell-2}}+1}^{n_{T_{\ell-1}}}(1-\gamma_j) = \frac{M}{h_1}(1-\gamma_{B_{\ell-1}})^{n_{T_{\ell-1}} - n_{T_{\ell-2}}}. \tag{2.A.6}$$

Now, to bound (2.A.6), first note that it follows from a second-order Taylor expansion of e^x around 0 that

$$e^x = 1 + x + e^{\xi}x^2/2 \tag{2.A.7}$$

for some ξ between 0 and x, so that $0 \leq 1 - \gamma_{B_{\ell-1}} \leq \exp(-\gamma_{B_{\ell-1}})$. Recall from (1.2) that $n_{T_{\ell-1}} = \lfloor T_{\ell-1}/\Delta t \rfloor$ and $n_{T_{\ell-2}} = \lfloor T_{\ell-2}/\Delta t \rfloor$, so that $T_{\ell-1}/\Delta t - 1 \leq n_{T_{\ell-1}} \leq T_{\ell-1}/\Delta t$, $T_{\ell-2}/\Delta t - 1 \leq n_{T_{\ell-2}} \leq T_{\ell-2}/\Delta t$ and $n_{T_{\ell-1}} - n_{T_{\ell-2}} \geq (T_{\ell-1} - T_{\ell-2})/\Delta t - 1 = \tau_{\ell-1}/\Delta t - 1$. Note from (A4) that we have $\tau_{\ell-1}/\Delta t \geq \delta \Delta t^{a_3 - 1}$, and hence $\tau_{\ell-1}/\Delta t - 1 \geq (\delta/2)\Delta t^{a_3-1}$, for Δt small enough, so that $n_{T_{\ell-1}} - n_{T_{\ell-2}} \geq (\delta/2)\Delta t^{a_3-1}$. Then we have, from (A1) and (A4),

$$\begin{aligned}(1-\gamma_{B_{\ell-1}})^{n_{T_{\ell-1}} - n_{T_{\ell-2}}} &\leq \exp\big\{-\gamma_{B_{\ell-1}}(n_{T_{\ell-1}} - n_{T_{\ell-2}})\big\} \\ &\leq \exp\big(-2^{-1}\delta^2 \Delta t^{a_2 + a_3 - 1}\big).\end{aligned} \tag{2.A.8}$$

Note from (A4) that $a_2 + a_3 - 1 < 0$ and hence $\Delta t^{a_2+a_3-1} \to \infty$ as $\Delta t \to 0$. Applying (A1)

and (2.A.8) to (2.A.6), we conclude that,

$$S_1 \leq \frac{M}{\delta}\Delta t^{-a_2}\exp(-2^{-1}\delta^2\Delta t^{a_2+a_3-1}) \leq \exp(-3^{-1}\delta^2\Delta t^{a_2+a_3-1}), \tag{2.A.9}$$

where Δt is small enough.

Bounding S_2. From (2.A.5), we can write $S_2 = S_{21} + S_{22}$, where

$$\begin{aligned} S_{21} &= \sum_{i=2}^{n_{T_{\ell-2}}} \gamma_i \prod_{j=i+1}^{n_t}(1-\gamma_j)K_{h_i}(x-X_i), \\ S_{22} &= \sum_{i=n_{T_{\ell-2}}+1}^{n_t} \gamma_i \prod_{j=i+1}^{n_t}(1-\gamma_j)K_{h_i}(x-X_i). \end{aligned} \tag{2.A.10}$$

Using (A1) and (A3), we have

$$\begin{aligned} S_{21} &\leq \frac{M}{h_m}\gamma_M \sum_{i=2}^{n_{T_{\ell-2}}}(1-\gamma_m)^{n_t-i} \\ &= \frac{M}{h_m}\gamma_M \sum_{j=n_t-n_{T_{\ell-2}}}^{n_t-2}(1-\gamma_m)^{j} \\ &= \frac{M}{h_m}\gamma_M(1-\gamma_m)^{n_t-n_{T_{\ell-2}}}\frac{1-(1-\gamma_m)^{n_{T_{\ell-2}}-1}}{1-(1-\gamma_m)} \\ &= \frac{M}{h_m}\frac{\gamma_M}{\gamma_m}(1-\gamma_m)^{n_t-n_{T_{\ell-2}}}\{1-(1-\gamma_m)^{n_{T_{\ell-2}}-1}\} \\ &\leq \frac{M}{h_m}\frac{\gamma_M}{\gamma_m}(1-\gamma_m)^{n_t-n_{T_{\ell-2}}}. \end{aligned}$$

Then, using (A1), we have $(1-\gamma_m)^{n_t-n_{T_{\ell-2}}} \leq (1-\gamma_m)^{n_{T_{\ell-1}}-n_{T_{\ell-2}}}$. Similarly to (2.A.8), we have $(1-\gamma_m)^{n_t-n_{T_{\ell-2}}} \leq \exp(-2^{-1}\delta^2\Delta t^{a_2+a_3-1})$. Therefore, using (A1), we have,

$$S_{21} \leq \frac{M}{\delta^3}\Delta t^{-a_2}\exp(-2^{-1}\delta^2\Delta t^{a_2+a_3-1}) \leq \exp(-3^{-1}\delta^2\Delta t^{a_2+a_3-1}), \tag{2.A.11}$$

when Δt is small enough. The lemma then follows from (2.A.5), (2.A.9), (2.A.10) and (2.A.11).

□

2.A.2 Proof of (2.7) in Proposition 2.1

First, we show that the bias of $\bar{f}(x,t)$, defined at (2.A.2), satisfies

$$\text{bias}\{\bar{f}(x,t)\} = \frac{1}{2} f_{xx}(x,t)\mu_{K,2} h_{B_\ell}^2 - f_t(x,t)\frac{\Delta t}{\gamma_{B_\ell}} + o(\Delta t^{1-a_2} + \Delta t^{2a_1}). \tag{2.A.12}$$

Then, by Lemma 2.A.2, we have $\text{bias}\{\check{f}(x,t)\} = \text{bias}\{\bar{f}(x,t)\} + o(\Delta t)$, and hence (2.7) follows.

To prove (2.A.12), first note that, from (A2) and (A3), we have

$$\begin{aligned}
&\mathrm{E}\{K_{h_i}(x - X_i)\} \\
&= \frac{1}{h_i}\int K\Big(\frac{x-y}{h_i}\Big) f(y,t_i)\,\mathrm{d}y \\
&= \int K(u) f(x - h_i u, t_i)\,\mathrm{d}u \\
&= \int K(u)\big\{ f(x,t_i) - f_x(x,t_i)h_i u + \frac{1}{2} f_{xx}(x,t_i)h_i^2 u^2 - \frac{1}{6} f_{xxx}(x - \theta_i h_i u, t_i) h_i^3 u^3 \big\}\,\mathrm{d}u \\
&= f(x,t_i) + \frac{1}{2} f_{xx}(x,t_i) h_i^2 \mu_{K,2} - r_i(h_i),
\end{aligned} \tag{2.A.13}$$

where $\theta_i \in [0,1]$ and

$$r_i(h_i) = \frac{1}{6} h_i^3 \int K(u) f_{xxx}(x - \theta h_i u, t_i) u^3\,\mathrm{d}u, \tag{2.A.14}$$

and where we used a second-order Taylor expansion of $f(x - h_i u, t_i)$ around x.

Then, recall that the h_i's and γ_i's are defined blockwise at (1.12). When Δt is small enough, we have $n_t > n_{T_{\ell-2}} \geq 2$. Then, from (1.12) and (1.12), since $t \in (T_{\ell-1}, T_\ell]$, we have

$$w_{n_t,i} = \begin{cases} \gamma_{B_{\ell-1}}(1-\gamma_{B_\ell})^{n_t - n_{T_{\ell-1}}}(1-\gamma_{B_{\ell-1}})^{n_{T_{\ell-1}} - i}, & \text{if } n_{T_{\ell-2}} + 1 \leq i \leq n_{T_{\ell-1}}, \\ \gamma_{B_\ell}(1-\gamma_{B_\ell})^{n_t - i}, & \text{if } n_{T_{\ell-1}} + 1 \leq i \leq n_t, \end{cases} \tag{2.A.15}$$

Therefore, from (2.A.2), (2.A.13) and (2.A.15), we can write

$$\begin{aligned}
\mathrm{E}\{\bar{f}(x,t)\} &= \sum_{i=n_{T_{\ell-2}}+1}^{n_{T_{\ell-1}}} w_{n_t,i}\,\mathrm{E}\{K_{h_i}(x-X_i)\} + \sum_{i=n_{T_{\ell-1}}+1}^{n_t} w_{n_t,i}\,\mathrm{E}\{K_{h_i}(x-X_i)\} \\
&= \gamma_{B_{\ell-1}}(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}} \sum_{i=n_{T_{\ell-2}}+1}^{n_{T_{\ell-1}}} (1-\gamma_{B_{\ell-1}})^{n_{T_{\ell-1}}-i}\,\mathrm{E}\{K_{h_i}(x-X_i)\} \\
&\qquad + \gamma_{B_\ell} \sum_{i=n_{T_{\ell-1}}+1}^{n_t} (1-\gamma_{B_\ell})^{n_t-i}\,\mathrm{E}\{K_{h_i}(x-X_i)\} \\
&= S_1 + S_2 + S_3 + S_4 - S_5 - S_6, \qquad \text{(2.A.16)}
\end{aligned}$$

where

$$S_1 = \gamma_{B_{\ell-1}}(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}} \sum_{i=n_{T_{\ell-2}}+1}^{n_{T_{\ell-1}}} (1-\gamma_{B_{\ell-1}})^{n_{T_{\ell-1}}-i} f(x,t_i), \qquad \text{(2.A.17)}$$

$$S_2 = \gamma_{B_\ell} \sum_{i=n_{T_{\ell-1}}+1}^{n_t} (1-\gamma_{B_\ell})^{n_t-i} f(x,t_i), \qquad \text{(2.A.18)}$$

$$S_3 = \frac{1}{2}\mu_{K,2} h_{B_{\ell-1}}^2 \gamma_{B_{\ell-1}}(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}} \sum_{i=n_{T_{\ell-2}}+1}^{n_{T_{\ell-1}}} (1-\gamma_{B_{\ell-1}})^{n_{T_{\ell-1}}-i} f_{xx}(x,t_i), \qquad \text{(2.A.19)}$$

$$S_4 = \frac{1}{2}\mu_{K,2} h_{B_\ell}^2 \gamma_{B_\ell} \sum_{i=n_{T_{\ell-1}}+1}^{n_t} (1-\gamma_{B_\ell})^{n_t-i} f_{xx}(x,t_i), \qquad \text{(2.A.20)}$$

$$S_5 = \gamma_{B_{\ell-1}}(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}} \sum_{i=n_{T_{\ell-2}}+1}^{n_{T_{\ell-1}}} (1-\gamma_{B_{\ell-1}})^{n_{T_{\ell-1}}-i} r_i(h_{B_{\ell-1}}), \qquad \text{(2.A.21)}$$

$$S_6 = \gamma_{B_\ell} \sum_{i=n_{T_{\ell-1}}+1}^{n_t} (1-\gamma_{B_\ell})^{n_t-i} r_i(h_{B_\ell}). \qquad \text{(2.A.22)}$$

Then, we calculate $S_1, \ldots, S_6$ as follows.

We assume that Δt is small enough to guarantee that $n_{T_{\ell-2}} \geq 2$ and $n_{T_{\ell-1}} - n_{T_{\ell-2}} \geq 2$. For now, we also assume that $n_t - n_{T_{\ell-1}} \geq 2$. The assumption $n_t - n_{T_{\ell-1}} \geq 2$ will be needed below, for example, at the fourth line of (2.A.42), to guarantee the summation by parts formula

is applicable. Then, in §2.A.2.5, we discuss the cases where $n_t - n_{T_{\ell-1}} = 0$ and $n_t - n_{T_{\ell-1}} = 1$.

2.A.2.1 Calculating S_1

Under (A2), it follows from a first-order Taylor expansion of $f(x, t_i)$ around t that, for some $\xi_i \in [t_i, t]$,

$$\begin{aligned} S_1 &= \gamma_{B_{\ell-1}}(1-\gamma_{B_\ell})^{n_t - n_{T_{\ell-1}}} \sum_{i=n_{T_{\ell-2}}+1}^{n_{T_{\ell-1}}} (1-\gamma_{B_{\ell-1}})^{n_{T_{\ell-1}}-i} \\ &\qquad \times \left\{ f(x,t) - (t-t_i) f_t(x,t) + \frac{1}{2}(t-t_i)^2 f_{tt}(x,\xi_i) \right\} \\ &= f(x,t) S_{11} - f_t(x,t) S_{12} + \frac{1}{2} S_{13}, \end{aligned} \tag{2.A.23}$$

where

$$S_{11} = \gamma_{B_{\ell-1}}(1-\gamma_{B_\ell})^{n_t - n_{T_{\ell-1}}} \sum_{i=n_{T_{\ell-2}}+1}^{n_{T_{\ell-1}}} (1-\gamma_{B_{\ell-1}})^{n_{T_{\ell-1}}-i}, \tag{2.A.24}$$

$$S_{12} = \gamma_{B_{\ell-1}}(1-\gamma_{B_\ell})^{n_t - n_{T_{\ell-1}}} \sum_{i=n_{T_{\ell-2}}+1}^{n_{T_{\ell-1}}} (1-\gamma_{B_{\ell-1}})^{n_{T_{\ell-1}}-i}(t-t_i), \tag{2.A.25}$$

$$S_{13} = \gamma_{B_{\ell-1}}(1-\gamma_{B_\ell})^{n_t - n_{T_{\ell-1}}} \sum_{i=n_{T_{\ell-2}}+1}^{n_{T_{\ell-1}}} (1-\gamma_{B_{\ell-1}})^{n_{T_{\ell-1}}-i}(t-t_i)^2 f_{tt}(x,\xi_i). \tag{2.A.26}$$

Then we calculate S_{11}, S_{12} and S_{13} as follows.

Calculating S_{11}. First, from (2.A.8), we have

$$\begin{aligned} \gamma_{B_{\ell-1}} \sum_{i=n_{T_{\ell-2}}+1}^{n_{T_{\ell-1}}} (1-\gamma_{B_{\ell-1}})^{n_{T_{\ell-1}}-i} &= \gamma_{B_{\ell-1}} \sum_{j=0}^{n_{T_{\ell-1}} - n_{T_{\ell-2}} - 1} (1-\gamma_{B_{\ell-1}})^{j} \\ &= 1 - (1-\gamma_{B_{\ell-1}})^{n_{T_{\ell-1}} - n_{T_{\ell-2}}} \\ &= 1 + o(\Delta t). \end{aligned} \tag{2.A.27}$$

Hence, plugging (2.A.27) into (2.A.24), we have

$$S_{11} = (1-\gamma_{B_\ell})^{n_t - n_{T_{\ell-1}}} + o(\Delta t). \tag{2.A.28}$$

Calculating S_{12}. First note from (1.2) that

$$t - \Delta t < n_t \Delta t \leq t, \tag{2.A.29}$$

and that, using (A4) and (2.1), we have

$$1 \leq n_{T_\ell} - n_{T_{\ell-1}} \sim \tau_\ell / \Delta t \asymp \Delta t^{a_3 - 1}. \tag{2.A.30}$$

Then, using (A1) and summation by parts, we have

$$\begin{aligned}
S_{12} &= \gamma_{B_{\ell-1}}(1-\gamma_{B_\ell})^{n_t - n_{T_{\ell-1}}} \sum_{i=n_{T_{\ell-2}}+1}^{n_{T_{\ell-1}}} (1-\gamma_{B_{\ell-1}})^{n_{T_{\ell-1}}-i}\{(n_t - i)\Delta t + (t - n_t \Delta t)\} \\
&= \Delta t\, \gamma_{B_{\ell-1}}(1-\gamma_{B_\ell})^{n_t - n_{T_{\ell-1}}} \sum_{i=n_{T_{\ell-2}}+1}^{n_{T_{\ell-1}}} (n_t - i)(1-\gamma_{B_{\ell-1}})^{n_{T_{\ell-1}}-i} + O(\Delta t S_{11}) \\
&= \Delta t (1-\gamma_{B_\ell})^{n_t - n_{T_{\ell-1}}} \gamma_{B_{\ell-1}} \sum_{j=0}^{n_{T_{\ell-1}} - n_{T_{\ell-2}} - 1} (j + n_t - n_{T_{\ell-1}})(1-\gamma_{B_{\ell-1}})^j + O(\Delta t) \\
&= \Delta t (1-\gamma_{B_\ell})^{n_t - n_{T_{\ell-1}}} (n_t - n_{T_{\ell-1}}) \gamma_{B_{\ell-1}} \sum_{j=0}^{n_{T_{\ell-1}} - n_{T_{\ell-2}} - 1} (1-\gamma_{B_{\ell-1}})^j \\
&\quad + \Delta t (1-\gamma_{B_\ell})^{n_t - n_{T_{\ell-1}}} \gamma_{B_{\ell-1}} \sum_{i=0}^{n_{T_{\ell-1}} - n_{T_{\ell-2}} - 2} \sum_{j=i+1}^{n_{T_{\ell-1}} - n_{T_{\ell-2}} - 1} (1-\gamma_{B_{\ell-1}})^j + O(\Delta t) \\
&= \Delta t (n_t - n_{T_{\ell-1}})(1-\gamma_{B_\ell})^{n_t - n_{T_{\ell-1}}} \\
&\quad + \Delta t (1-\gamma_{B_\ell})^{n_t - n_{T_{\ell-1}}} \sum_{i=0}^{n_{T_{\ell-1}} - n_{T_{\ell-2}} - 2} \{(1-\gamma_{B_{\ell-1}})^{i+1} - (1-\gamma_{B_{\ell-1}})^{n_{T_{\ell-1}} - n_{T_{\ell-2}}}\} \\
&\qquad + O(\Delta t) \\
&= \Delta t (n_t - n_{T_{\ell-1}})(1-\gamma_{B_\ell})^{n_t - n_{T_{\ell-1}}}
\end{aligned}$$

$$+\frac{\Delta t}{\gamma_{B_{\ell-1}}}(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}\{1-\gamma_{B_{\ell-1}}-(1-\gamma_{B_{\ell-1}})^{n_{T_{\ell-1}}-n_{T_{\ell-2}}}\}$$

$$-\Delta t(1-\gamma_{B_\ell})^{n_t-n_{T_\ell}}(n_{T_{\ell-1}}-n_{T_{\ell-2}}-1)(1-\gamma_{B_{\ell-1}})^{n_{T_{\ell-1}}-n_{T_{\ell-2}}}+O(\Delta t)$$

$$=\frac{\Delta t}{\gamma_{B_{\ell-1}}}(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}+\Delta t(n_t-n_{T_{\ell-1}})(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}+O(\Delta t), \tag{2.A.31}$$

where we used (1.1) to get the first equation, (2.A.24) and (2.A.29) to get the second, (2.A.28) to get the third, (2.A.27) for the fifth, and (2.A.8) and (2.A.30) for the last.

Calculating S_{13}. Under (A2), we have

$$|S_{13}|\le M\gamma_{B_{\ell-1}}(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}\sum_{i=n_{T_{\ell-2}}+1}^{n_{T_{\ell-1}}}(1-\gamma_{B_{\ell-1}})^{n_{T_{\ell-1}}-i}(t-t_i)^2$$

$$=M\gamma_{B_{\ell-1}}(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}\sum_{i=n_{T_{\ell-2}}+1}^{n_{T_{\ell-1}}}(1-\gamma_{B_{\ell-1}})^{n_{T_{\ell-1}}-i}\{(n_t-i)\Delta t+(t-n_t\Delta t)\}^2$$

$$\le M\Delta t^2\gamma_{B_{\ell-1}}(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}\sum_{i=n_{T_{\ell-2}}+1}^{n_{T_{\ell-1}}}(1-\gamma_{B_{\ell-1}})^{n_{T_{\ell-1}}-i}(n_t-i+1)^2$$

$$=M\Delta t^2(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}\gamma_{B_{\ell-1}}\sum_{j=0}^{n_{T_{\ell-1}}-n_{T_{\ell-2}}-1}(j+n_t-n_{T_{\ell-1}}+1)^2(1-\gamma_{B_{\ell-1}})^j$$

$$=M\Delta t^2(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}\Bigg[(n_t-n_{T_{\ell-1}}+1)^2\gamma_{B_{\ell-1}}\sum_{j=0}^{n_{T_{\ell-1}}-n_{T_{\ell-2}}-1}(1-\gamma_{B_{\ell-1}})^j$$

$$+\gamma_{B_{\ell-1}}\sum_{i=0}^{n_{T_{\ell-1}}-n_{T_{\ell-2}}-2}\{2(i+n_t-n_{T_{\ell-1}})+3\}\sum_{j=i+1}^{n_{T_{\ell-1}}-n_{T_{\ell-2}}-1}(1-\gamma_{B_{\ell-1}})^j\Bigg]$$

$$=M\Delta t^2(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}\Bigg[(n_t-n_{T_{\ell-1}}+1)^2$$

$$+\sum_{i=0}^{n_{T_{\ell-1}}-n_{T_{\ell-2}}-2}\{2(i+n_t-n_{T_{\ell-1}})+3\}(1-\gamma_{B_{\ell-1}})^{i+1}$$

$$-\sum_{i=0}^{n_{T_{\ell-1}}-n_{T_{\ell-2}}-2}\{2(i+n_t-n_{T_{\ell-1}})+3\}(1-\gamma_{B_{\ell-1}})^{n_{T_{\ell-1}}-n_{T_{\ell-2}}}+O(\Delta t)\Bigg]$$

$$
\begin{aligned}
&\leq M\Delta t^2(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}\Bigg[(n_t-n_{T_{\ell-1}}+1)^2 \\
&\qquad\qquad + 2(n_t-n_{T_{\ell-2}})\sum_{i=0}^{n_{T_{\ell-1}}-n_{T_{\ell-2}}-1}(1-\gamma_{B_{\ell-1}})^{i+1}+O(\Delta t)\Bigg] \\
&= M\Delta t^2(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}\big\{(n_t-n_{T_{\ell-1}}+1)^2+2(n_t-n_{T_{\ell-2}})\gamma_{B_{\ell-1}}^{-1}+O(\Delta t)\big\} \\
&= M\Delta t^2(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}\big\{(n_t-n_{T_{\ell-1}})^2+2(n_t-n_{T_{\ell-1}})+1\big\}+o(\Delta t^{1-a_2}) \\
&= M\Delta t^2(n_t-n_{T_{\ell-1}})^2(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}+o(\Delta t^{1-a_2}), \qquad (2.A.32)
\end{aligned}
$$

where we used (1.1) to get the second line, (2.A.30) and (2.A.29) for the third, (2.A.8) for the eleventh, and where, to get the second last and the last lines, we used the fact that, under (A4), we have, similarly to (2.A.30), $n_t - n_{T_{\ell-2}} \asymp \Delta t^{a_3-1}$.

Shortly, to bound (2.A.32), we will show that

$$
(n_t-n_{T_{\ell-1}})(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}=O(\Delta t^{-a_2}), \qquad (2.A.33)
$$

which, using (2.A.30), (2.A.32) and (2.A.33), implies that

$$
S_{13}=o(\Delta t^{1-a_2}), \qquad (2.A.34)
$$

under (A1) and (A4).

Now, to prove (2.A.33), letting $g_1(x) = x(1-\gamma_{B_\ell})^x$, where $1 \leq x \leq n_{T_\ell} - n_{T_{\ell-1}}$, it suffices to show that $g_1(n_t - n_{T_{\ell-1}}) = O(\Delta t^{-a_2})$. Note that g_1 reaches its global maximum at $x_0 = -1/\log(1-\gamma_{B_\ell})$. Hence, under (A1), we have

$$
g_1(n_t-n_{T_{\ell-1}}) \leq g_1(x_0) \leq -\frac{1}{\log(1-\gamma_{B_\ell})} \sim \frac{1}{\gamma_{B_\ell}} \asymp \Delta t^{-a_2}, \qquad (2.A.35)
$$

where we used a first-order Taylor expansion of $\log(1-\gamma_{B_\ell})$ around 1. Hence (2.A.33) is proved.

Conclusion: Combining (2.A.23), (2.A.28), (2.A.31) and (2.A.34), we deduce that

$$S_1 = f(x,t)(1-\gamma_{B_\ell})^{n_t - n_{T_{\ell-1}}} - \frac{\Delta t}{\gamma_{B_{\ell-1}}} f_t(x,t)(1-\gamma_{B_\ell})^{n_t - n_{T_{\ell-1}}} - f_t(x,t)\Delta t(n_t - n_{T_{\ell-1}})(1-\gamma_{B_\ell})^{n_t - n_{T_{\ell-1}}} + o(\Delta t^{1-a_2}). \tag{2.A.36}$$

2.A.2.2 Calculating S_2

Similarly to (2.A.23), we have

$$S_2 = f(x,t)S_{21} - f_t(x,t)S_{22} + \frac{1}{2}S_{23}, \tag{2.A.37}$$

where

$$S_{21} = \gamma_{B_\ell} \sum_{i=n_{T_{\ell-1}}+1}^{n_t} (1-\gamma_{B_\ell})^{n_t - i}, \tag{2.A.38}$$

$$S_{22} = \gamma_{B_\ell} \sum_{i=n_{T_{\ell-1}}+1}^{n_t} (1-\gamma_{B_\ell})^{n_t - i}(t - t_i), \tag{2.A.39}$$

$$S_{23} = \gamma_{B_\ell} \sum_{i=n_{T_{\ell-1}}+1}^{n_t} (1-\gamma_{B_\ell})^{n_t - i}(t - t_i)^2 f_{tt}(x,\xi_i), \tag{2.A.40}$$

where ξ_i is between t_i and t. Then, we calculate S_{21}, S_{22} and S_{23} separately.

Calculating S_{21}. From (2.A.38), we have

$$S_{21} = \gamma_{B_\ell} \sum_{j=0}^{n_t - n_{T_{\ell-1}} - 1} (1-\gamma_{B_\ell})^j = 1 - (1-\gamma_{B_\ell})^{n_t - n_{T_{\ell-1}}}. \tag{2.A.41}$$

Calculating S_{22}. Applying (1.1), (2.A.29) and (2.A.41) to (2.A.39), we have

$$S_{22} = \gamma_{B_\ell} \sum_{i=n_{T_{\ell-1}}+1}^{n_t} (1-\gamma_{B_\ell})^{n_t - i}\{(n_t - i)\Delta t + (t - n_t\Delta t)\}$$

$$
\begin{aligned}
&= \Delta t\, \gamma_{B_\ell} \sum_{i=n_{T_{\ell-1}}+1}^{n_t} (n_t - i)(1-\gamma_{B_\ell})^{n_t-i} + O(\Delta t S_{21}) \\
&= \Delta t\, \gamma_{B_\ell} \sum_{j=0}^{n_t-n_{T_{\ell-1}}-1} j(1-\gamma_{B_\ell})^{j} + O(\Delta t) \\
&= \Delta t\, \gamma_{B_\ell} \sum_{i=0}^{n_t-n_{T_{\ell-1}}-2} \sum_{j=i+1}^{n_t-n_{T_{\ell-1}}-1} (1-\gamma_{B_\ell})^{j} + O(\Delta t) \\
&= \Delta t \sum_{i=0}^{n_t-n_{T_{\ell-1}}-2} \{(1-\gamma_{B_\ell})^{i+1} - (1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}\} + O(\Delta t) \\
&= \frac{\Delta t}{\gamma_{B_\ell}} \{1 - \gamma_{B_\ell} - (1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}\} \\
&\qquad - \Delta t (n_t - n_{T_{\ell-1}} - 1)(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}} + O(\Delta t) \\
&= \frac{\Delta t}{\gamma_{B_\ell}} \{1 - (1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}\} \\
&\qquad - \Delta t (n_t - n_{T_{\ell-1}})(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}} + O(\Delta t), \qquad (2.A.42)
\end{aligned}
$$

where we used summation by parts and, to get the third line, we used (2.A.41) and the fact that $0 < \gamma_{B_\ell} < 1$.

Calculating S_{23}. Under (A1), (A2) and (A4), from (2.A.40), we have

$$
\begin{aligned}
|S_{23}| &\le M\gamma_{B_\ell} \sum_{i=n_{T_{\ell-1}}+1}^{n_t} (1-\gamma_{B_\ell})^{n_t-i} \{(n_t - i)\Delta t + (t - n_t \Delta t)\}^2 \\
&\le M\Delta t^2 \gamma_{B_\ell} \sum_{i=n_{T_{\ell-1}}+1}^{n_t} (1-\gamma_{B_\ell})^{n_t-i} (n_t - i + 1)^2 \\
&= M\Delta t^2 \gamma_{B_\ell} \sum_{j=0}^{n_t-n_{T_{\ell-1}}-1} (j+1)^2 (1-\gamma_{B_\ell})^{j} \\
&= M\Delta t^2 \left\{ \gamma_{B_\ell} \sum_{j=0}^{n_t-n_{T_{\ell-1}}-1} (1-\gamma_{B_\ell})^{j} + \gamma_{B_\ell} \sum_{i=0}^{n_t-n_{T_{\ell-1}}-2} (2i+3) \sum_{j=i+1}^{n_t-n_{T_{\ell-1}}-1} (1-\gamma_{B_\ell})^{j} \right\} \\
&= M\Delta t^2 \Bigg\{ 1 - (1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}
\end{aligned}
$$

$$+ \sum_{i=0}^{n_t - n_{T_{\ell-1}} - 1} (2i+3)(1-\gamma_{B_\ell})^{i+1} - (1-\gamma_{B_\ell})^{n_t - n_{T_{\ell-1}}} \sum_{i=0}^{n_t - n_{T_{\ell-1}} - 1} (2i+3) \Bigg\}$$

$$\leq M\Delta t^2 \Bigg\{ 1 + \sum_{i=0}^{n_t - n_{T_{\ell-1}} - 1} (2i+3)(1-\gamma_{B_\ell})^{i+1} \Bigg\}$$

$$\leq M\Delta t^2 \{2(n_t - n_{T_{\ell-1}}) + 1\} \sum_{i=0}^{n_t - n_{T_{\ell-1}} - 1} (1-\gamma_{B_\ell})^{i+1} + M\Delta t^2$$

$$\leq M\frac{\Delta t^2}{\gamma_{B_\ell}} \{2(n_t - n_{T_{\ell-1}}) + 1\} + M\Delta t^2 = o(\Delta t^{1-a_2}), \tag{2.A.43}$$

where we used the fact that $0 < \gamma_{B_\ell} < 1$, and where we used (2.A.29) to get the second line, summation by parts to get the fourth line, and (2.A.30) to get the last equation.

Conclusion: Combining (2.A.37), (2.A.41), (2.A.42) and (2.A.43), we deduce that

$$S_2 = f(x,t)\{1 - (1-\gamma_{B_\ell})^{n_t - n_{T_{\ell-1}}}\} - f_t(x,t)\frac{\Delta t}{\gamma_{B_\ell}}\{1 - (1-\gamma_{B_\ell})^{n_t - n_{T_{\ell-1}}}\} + f_t(x,t)\Delta t(n_t - n_{T_{\ell-1}})(1-\gamma_{B_\ell})^{n_t - n_{T_{\ell-1}}} + o(\Delta t^{1-a_2}). \tag{2.A.44}$$

Now, under (A1), combining (2.A.36) and (2.A.44), we have

$$S_1 + S_2 = f(x,t) - f_t(x,t)\frac{\Delta t}{\gamma_{B_\ell}} \Big\{ 1 + \Big(\frac{\gamma_{B_\ell}}{\gamma_{B_{\ell-1}}} - 1 \Big)(1-\gamma_{B_{\ell-1}})^{n_t - n_{T_{\ell-1}}} \Big\} + o(\Delta t^{1-a_2})$$
$$= f(x,t) - f_t(x,t)\frac{\Delta t}{\gamma_{B_\ell}} + o(\Delta t^{1-a_2}). \tag{2.A.45}$$

2.A.2.3 Calculating S_3 and S_4

Observe from (2.A.17)–(2.A.20) that S_3 and S_4 are exactly identical to S_1 and S_2, except that $f(x,t_i)$ in S_1 and S_2 is replaced by $f_{tt}(x,t_i)$ in S_3 and S_4 and that S_3 and S_4 contain an additional multiplicative term ($2^{-1}\mu_{K,2}h_{B_{\ell-1}}^2$ for S_3 and $2^{-1}\mu_{K,2}h_{B_\ell}^2$ for S_4). Therefore, S_3 and

S_4 can be computed in the exact same way as S_1 and S_2, except that we need to use a first-order Taylor expansion of $f_{xx}(x,t_i)$, instead of $f(x,t_i)$, around t at (2.A.23) and (2.A.37), which we can do under (A2). Therefore, under (A1), we can write

$$\begin{aligned} S_3 &= \frac{1}{2}\mu_{K,2}h_{B_{\ell-1}}^2\Big\{f_{xx}(x,t)(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}} - \frac{\Delta t}{\gamma_{B_{\ell-1}}}f_{xxt}(x,t)(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}} \\ &\qquad - f_{xxt}(x,t)\Delta t(n_t-n_{T_{\ell-1}})(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}} + o(\Delta t^{1-a_1})\Big\} \\ &= \frac{1}{2}f_{xx}(x,t)\mu_{K,2}h_{B_{\ell-1}}^2(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}} + o(\Delta t^{2a_1}) \end{aligned} \tag{2.A.46}$$

and

$$\begin{aligned} S_4 &= \frac{1}{2}\mu_{K,2}h_{B_{\ell-1}}^2\Big\{f_{xx}(x,t)\{1-(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}\} \\ &\qquad - f_{xxt}(x,t)\frac{\Delta t}{\gamma_{B_\ell}}\{1-(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}\} \\ &\qquad + f_{xxt}(x,t)\Delta t(n_t-n_{T_{\ell-1}})(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}} + o(\Delta t^{1-a_2})\Big\} \\ &= \frac{1}{2}f_{xx}(x,t)\mu_{K,2}h_{B_{\ell-1}}^2(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}} + o(\Delta t^{2a_1}), \end{aligned} \tag{2.A.47}$$

where we used (2.A.33) and the fact that $0<\gamma_{B_\ell}<1$. Under (A1), we deduce that

$$\begin{aligned} S_3+S_4 &= \frac{1}{2}f_{xx}(x,t)\mu_{K,2}h_{B_\ell}^2\left[1+\left\{\left(\frac{h_{B_{\ell-1}}}{h_{B_\ell}}\right)^2-1\right\}(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}\right]+o(\Delta t^{2a_1}) \\ &= \frac{1}{2}f_{xx}(x,t)\mu_{K,2}h_{B_\ell}^2+o(\Delta t^{2a_1}). \end{aligned} \tag{2.A.48}$$

2.A.2.4 Calculating S_5 and S_6

Under (A1), (A2) and (A3), for $h=h_{B_{\ell-1}}$ and $h=h_{B_\ell}$, we have

$$\begin{aligned} |r_i(h)| &\le \frac{1}{6}h^3\int|K(u)||f_{xxx}(x-\theta hu,t_i)||u|^3\,\mathrm{d}u \\ &\le \frac{M}{6}h^3\int|u|^3K(u)\,\mathrm{d}u \le C_1\Delta t^{3a_1}, \end{aligned} \tag{2.A.49}$$

where $C_1 = M/(6\delta^3) \int |u|^3 K(u)\,\mathrm{d}u$.

Plugging (2.A.49) into (2.A.21) and recalling the definition of S_{11} at (2.A.24), we have

$$S_5 \leq C_1 \Delta t^{3a_1} S_{11} = O(\Delta t^{3a_1}), \tag{2.A.50}$$

where we used (2.A.28) and the fact that $0 \leq \gamma_{B_\ell} \leq 1$. Plugging (2.A.49) into (2.A.22), we also find

$$\begin{aligned} S_6 &\leq C_1 \Delta t^{3a_1} \gamma_{B_\ell} \sum_{i=n_{T_{\ell-1}}+1}^{n_t} (1-\gamma_{B_\ell})^{n_t-i} = C_1 \Delta t^{3a_1} \gamma_{B_\ell} \sum_{j=0}^{n_t-n_{T_{\ell-1}}-1} (1-\gamma_{B_\ell})^j \\ &= C_1 \Delta t^{3a_1} \{1 - (1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}\} \leq C_1 \Delta t^{3a_1}, \end{aligned} \tag{2.A.51}$$

since $0 < \gamma_{B_\ell} < 1$.

2.A.2.5 Conclusion

Combining (2.A.16), (2.A.45), (2.A.48), (2.A.50) and (2.A.51), we have

$$\mathrm{E}\{\bar{f}(x,t)\} = f(x,t) + \frac{1}{2} f_{xx}(x,t) \mu_{K,2} h_{B_\ell}^2 - f_t(x,t) \frac{\Delta t}{\gamma_{B_\ell}} + o(\Delta t^{1-a_2} + \Delta t^{2a_1}),$$

which proves (2.A.16) in the case where $n_t - n_{T_{\ell-1}} \geq 2$ (recall from the assumption we made below (2.A.22) that the above calculations are only valid when $n_t - n_{T_{\ell-1}} \geq 2$).

It remains to prove that (2.A.12) holds also when $n_t - n_{T_{\ell-1}} = 0$ and when $n_t - n_{T_{\ell-1}} = 1$. Note that in our calculations above, the assumption $n_t - n_{T_{\ell-1}} \geq 2$ has only been used for S_2, S_4 and S_6. In particular our calculations for S_1, S_3 and S_5 remain valid and we only need to recompute S_2, S_4 and S_6 when $n_t - n_{T_{\ell-1}} = 0$ or 1.

When $n_t - n_{T_{\ell-1}} = 0$, we have $n_t = n_{T_{\ell-1}}$, and, from (2.A.16), we have $\mathrm{E}\{\bar{f}(x,t)\} = S_1 + S_3 - S_5$, where S_1, S_3 and S_5 are defined at (2.A.17), (2.A.19) and (2.A.21). Hence, from

(2.A.36), (2.A.46) and (2.A.50), we have, using (A1) and (A2)

$$\text{bias}\{\bar{f}(x,t)\} = \frac{1}{2}f_{xx}(x,t)\mu_{K,2}h_{B_{\ell-1}}^2 - f_t(x,t)\frac{\Delta t}{\gamma_{B_{\ell-1}}} + o(\Delta t^{1-a_2} + \Delta t^{2a_1}) \tag{2.A.52}$$

$$= \frac{1}{2}f_{xx}(x,t)\mu_{K,2}h_{B_\ell}^2 - f_t(x,t)\frac{\Delta t}{\gamma_{B_\ell}} + o(\Delta t^{1-a_2} + \Delta t^{2a_1}), \tag{2.A.53}$$

where we used that fact that $\gamma_{B_\ell}/\gamma_{B_{\ell-1}} = 1 + o(1)$ and $h_{B_\ell}/h_{B_{\ell-1}} = 1 + o(1)$ in (A1). This proves that (2.A.12) holds when $n_t - n_{T_{\ell-1}} = 0$.

When $n_t - n_{T_{\ell-1}} = 1$, we have $n_t = n_{T_{\ell-1}} + 1$, so that, using (2.A.38), $S_{21} = \gamma_{B_\ell}$. From (2.A.39), we have $S_{22} = \gamma_{B_\ell}(t - t_{n_t}) = o(\Delta t)$, where we used (A1) and (2.A.29). From (2.A.40), we have $S_{23} = \gamma_{B_\ell}(t - t_{n_t})^2 f_{tt}(x, \xi_{n_t}) = o(\Delta t)$, using (A1), (A2) and (2.A.29). Hence, using (A1) and (2.A.37), we have $S_2 = \gamma_{B_\ell} f(x,t) + o(\Delta t)$. Now, from (2.A.20), we have $S_4 = \mu_{K,2}h_{B_\ell}^2\gamma_{B_\ell}f_{xx}(x,t_{n_t})/2 = O(\Delta t^{a_2+2a_1})$, using (A1) and (A2). From (2.A.22), we also have $S_6 = \gamma_{B_\ell}r_i(h_{B_\ell}) = O(\Delta t^{a_2+3a_1})$, using (A1) and (2.A.49). Now, from (2.A.36), (2.A.46) and (2.A.50), we have $S_1 = f(x,t)(1-\gamma_{B_\ell}) - \Delta t f_t(x,t)/\gamma_{B_{\ell-1}} - f_t(x,t)\Delta t(1-\gamma_{B_\ell}) + o(\Delta t^{1-a_2}) = f(x,t)(1-\gamma_{B_\ell}) - \Delta t f_t(x,t)/\gamma_{B_{\ell-1}} + o(\Delta t^{1-a_2})$, $S_3 = f_{xx}(x,t)\mu_{K,2}h_{B_{\ell-1}}^2/2 + o(\Delta t^{2a_1})$ and $S_5 = O(\Delta t^{3a_1})$. Combining the above results with (2.A.16), we obtain (2.A.52). Then, using (A1) and (A2), we have (2.A.53), which proves that (2.A.12) holds when $n_t - n_{T_{\ell-1}} = 1$.

Combining §2.A.2.1 to §2.A.2.4 and the above calculations, (2.7) is proved.

2.A.3 Proof of (2.8) in Proposition 2.1

Shortly, we shall show that the variance of $\bar{f}(x,t)$ defined at (2.A.2) satisfies

$$\text{var}\{\bar{f}(x,t)\} = \frac{1}{2}f(x,t)R_K\frac{\gamma_{B_\ell}}{h_{B_\ell}} + o(\Delta t^{a_2-a_1}). \tag{2.A.54}$$

Using (2.2), (2.A.2) and Lemma 2.A.2, this proves that

$$\begin{aligned}\text{var}\{\check{f}(x,t)\} &= \text{var}\{\bar{f}(x,t) + \check{f}(x,t) - \bar{f}(x,t)\} \\ &= \text{var}\{\bar{f}(x,t)\} + \text{var}\{\check{f}(x,t) - \bar{f}(x,t)\}\end{aligned}$$

$$= \text{var}\{\bar{f}(x,t)\} + o(\Delta t),$$

where we used the independence of $X_1, \ldots, X_{n_t}$. Then, (2.8) follows from (2.A.54).

Next, we prove (2.A.54). First note that, under (A2) and (A3), we have

$$\begin{aligned}
\mathrm{E}\{K_{h_i}^2(x - X_i)\} &= \frac{1}{h_i^2} \int K^2\Big(\frac{x-y}{h_i}\Big) f(y, t_i)\, \mathrm{d}y \\
&= \frac{1}{h_i} \int K^2(u) f(x - h_i u, t_i)\, \mathrm{d}u \\
&= \frac{1}{h_i} \int K^2(u) \{f(x, t_i) - f_x(x - \theta_i h_i u, t_i) h_i u\}\, \mathrm{d}u \\
&= \frac{R_K}{h_i} f(x, t_i) - \int u K^2(u) f(x - \theta_i h_i u, t_i)\, \mathrm{d}u,
\end{aligned}$$

where $\theta_i \in (0,1)$ and R_K is defined above Proposition 2.1, and where we applied the mean value theorem to $f(\cdot, t_i)$. Using (A2) and (A3), we also have $|\int u K^2(u) f_x(x - \theta_i h_i u, t_i)\, \mathrm{d}u| \leq M^2 \int |u| K(u)\, \mathrm{d}u < \infty$. Hence, we have $\mathrm{E}\{K_{h_i}^2(x - X_i)\} = R_K / h_i f(x, t_i) + O(1)$ and

$$\text{var}\{K_{h_i}(x - X_i)\} = \mathrm{E}\{K_{h_i}^2(x - X_i)\} - [\mathrm{E}\{K_{h_i}(x - X_i)\}]^2 = \frac{R_K}{h_i} f(x, t_i) + O(1), \quad \text{(2.A.55)}$$

where we used the fact that $\mathrm{E}\{K_{h_i}(x - X_i)\} = O(1)$, using (A2), (A3) and (2.A.13).

Recall that, for a given time $t \in (0, \infty)$ and a given Δt, ℓ is the integer for which $t \in (T_{\ell-1}, T_\ell]$. Recall also that $w_{n_t,i}$ is defined at (1.12) and, in particular recalling the blockwise definition of γ_i and h_i at (1.12), we have $w_{n_t,i} = (1 - \gamma_{B_\ell})^{n_t - n_{T_{\ell-1}}} \gamma_{B_{\ell-1}} (1 - \gamma_{B_{\ell-1}})^{n_{T_{\ell-1}} - i}$ for $i = n_{T_{\ell-2}} + 1, \ldots, n_{T_{\ell-1}}$, and $w_{n_t,i} = \gamma_{B_\ell} (1 - \gamma_{B_\ell})^{n_t - i}$ for $i = n_{T_{\ell-1}} + 1, \ldots, n_t$. Now, using (2.A.2), (2.A.55) and the independence of $X_{n_{T_{\ell-2}}+1}, \ldots, X_{n_t}$, we have

$$\begin{aligned}
\text{var}\{\bar{f}(x,t)\} &= \sum_{i=n_{T_{\ell-2}}+1}^{n_t} w_{n_t,i}^2 \,\text{var}\{K_{h_i}(x - X_i)\} \\
&= \sum_{i=n_{T_{\ell-2}}+1}^{n_{T_{\ell-1}}} w_{n_t,i}^2 \,\text{var}\{K_{h_i}(x - X_i)\} + \sum_{i=n_{T_{\ell-1}}+1}^{n_t} w_{n_t,i}^2 \,\text{var}\{K_{h_i}(x - X_i)\}
\end{aligned}$$

$$= \frac{R_K}{h_{B_{\ell-1}}} S_7 + \frac{R_K}{h_{B_\ell}} S_8 + O(S_9 + S_{10}), \tag{2.A.56}$$

where

$$S_7 = (1-\gamma_{B_\ell})^{2(n_t - n_{T_{\ell-1}})} \gamma_{B_{\ell-1}}^2 \sum_{i=n_{T_{\ell-2}}+1}^{n_{T_{\ell-1}}} (1-\gamma_{B_{\ell-1}})^{2(n_{T_{\ell-1}}-i)} f(x,t_i), \tag{2.A.57}$$

$$S_8 = \gamma_{B_\ell}^2 \sum_{i=n_{T_{\ell-1}}+1}^{n_t} (1-\gamma_{B_\ell})^{2(n_t-i)} f(x,t_i), \tag{2.A.58}$$

$$S_9 = (1-\gamma_{B_\ell})^{2(n_t - n_{T_{\ell-1}})} \gamma_{B_{\ell-1}}^2 \sum_{i=n_{T_{\ell-2}}+1}^{n_{T_{\ell-1}}} (1-\gamma_{B_{\ell-1}})^{2(n_{T_{\ell-1}}-i)} \tag{2.A.59}$$

$$S_{10} = \gamma_{B_\ell}^2 \sum_{i=n_{T_{\ell-1}}+1}^{n_t} (1-\gamma_{B_\ell})^{2(n_t-i)}. \tag{2.A.60}$$

Here we have assumed that Δt is small enough to guarantee that $n_{T_{\ell-2}} \geq 2$ and we have also assumed that $n_t - n_{T_{\ell-1}} \geq 1$. In §2.A.3.1 to §2.A.3.3, we calculate $S_7, \ldots, S_{10}$ under these assumptions. All calculations remain valid when $n_t - n_{T_{\ell-1}} = 0$ except that in that case, $S_8 = S_{10} = 0$.

2.A.3.1 Calculating S_7

Under (A2), applying the mean value theorem to $f(x,t_i)$, we have that, with some ξ_i between t_i and t,

$$S_7 = f(x,t) S_{71} - S_{72}, \tag{2.A.61}$$

where

$$S_{71} = (1-\gamma_{B_\ell})^{2(n_t - n_{T_{\ell-1}})} \gamma_{B_{\ell-1}}^2 \sum_{i=n_{T_{\ell-2}}+1}^{n_{T_{\ell-1}}} (1-\gamma_{B_{\ell-1}})^{2(n_{T_{\ell-1}}-i)}, \tag{2.A.62}$$

$$S_{72} = (1-\gamma_{B_\ell})^{2(n_t - n_{T_{\ell-1}})} \gamma_{B_{\ell-1}}^2 \sum_{i=n_{T_{\ell-2}}+1}^{n_{T_{\ell-1}}} (1-\gamma_{B_{\ell-1}})^{2(n_{T_{\ell-1}}-i)} (t-t_i) f(x,\xi_i). \quad (2.A.63)$$

Next we calculate S_{71} and S_{72}.

Calculating S_{71}. We have

$$\begin{aligned}\gamma_{B_{\ell-1}}^2 \sum_{i=n_{T_{\ell-2}}+1}^{n_{T_{\ell-1}}} (1-\gamma_{B_{\ell-1}})^{2(n_{T_{\ell-1}}-i)} &= \gamma_{B_{\ell-1}}^2 \sum_{j=0}^{n_{T_{\ell-1}}-n_{T_{\ell-2}}-1} (1-\gamma_{B_{\ell-1}})^{2j} \\ &= \frac{\gamma_{B_{\ell-1}}}{2-\gamma_{B_{\ell-1}}} \left\{1-(1-\gamma_{B_{\ell-1}})^{2(n_{T_{\ell-1}}-n_{T_{\ell-2}})}\right\} \\ &= \frac{\gamma_{B_{\ell-1}}}{2} + O(\Delta t^{2a_2}), \end{aligned} \quad (2.A.64)$$

where we used the fact that, under (A1), $1/(2-\gamma_{B_{\ell-1}}) = 1/2 + O(\gamma_{B_{\ell-1}})$, and where we used (2.A.8). Then, plugging (2.A.64) into (2.A.62), we find

$$S_{71} = \frac{\gamma_{B_{\ell-1}}}{2}(1-\gamma_{B_\ell})^{2(n_t - n_{T_{\ell-1}})} + O(\Delta t^{2a_2}). \quad (2.A.65)$$

Calculating S_{72}. Recalling the definition of S_{12} at (2.A.25) and the fact that $0 < \gamma_{B_{\ell-1}}, \gamma_{B_\ell} < 1$, using (A1), (A2), (2.A.31), (2.A.33), we have

$$\begin{aligned}|S_{72}| &\le M(1-\gamma_{B_\ell})^{(n_t - n_{T_{\ell-1}})} \gamma_{B_{\ell-1}}^2 \sum_{i=n_{T_{\ell-2}}+1}^{n_{T_{\ell-1}}} (1-\gamma_{B_{\ell-1}})^{(n_{T_{\ell-1}}-i)} (t-t_i) \\ &= M\gamma_{B_{\ell-1}} S_{12} = O(\Delta t) = o(\Delta t^{a_2}). \end{aligned} \quad (2.A.66)$$

Conclusion. Combining (2.A.61), (2.A.65) and (2.A.66), we conclude that

$$S_7 = \frac{\gamma_{B_{\ell-1}}}{2} f(x,t)(1-\gamma_{B_\ell})^{2(n_t - n_{T_{\ell-1}})} + o(\Delta t^{a_2}). \quad (2.A.67)$$

2.A.3.2 Calculating S_8

Using (A2) and (2.A.58), we have, using the mean value theorem for $f(x,\cdot)$,

$$S_8 = f(x,t)S_{81} - S_{82}, \tag{2.A.68}$$

where ξ_i lies between t_i and t, and where

$$S_{81} = \gamma_{B_\ell}^2 \sum_{i=n_{T_{\ell-1}}+1}^{n_t} (1-\gamma_{B_\ell})^{2(n_t-i)} \tag{2.A.69}$$

$$S_{82} = \gamma_{B_\ell}^2 \sum_{i=n_{T_{\ell-1}}+1}^{n_t} (1-\gamma_{B_\ell})^{2(n_t-i)}(t-t_i) f_t(x,\xi_i). \tag{2.A.70}$$

Next, we calculate S_{81} and S_{82}.

Calculating S_{81}. From (2.A.69), under (A1), we have

$$\begin{aligned} S_{81} &= \gamma_{B_\ell}^2 \sum_{j=0}^{n_t-n_{T_{\ell-1}}-1} (1-\gamma_{B_\ell})^{2j} = \frac{\gamma_{B_\ell}}{2-\gamma_{B_\ell}}\{1-(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}\} \\ &= \frac{\gamma_{B_\ell}}{2}\{1-(1-\gamma_{B_\ell})^{n_t-n_{T_{\ell-1}}}\} + O(\Delta t^{2a_2}), \end{aligned} \tag{2.A.71}$$

where we used the fact that, under (A1), $1/(2-\gamma_{B_\ell}) = 1/2 + O(\gamma_{B_\ell})$ and that $0 < \gamma_{B_\ell} < 1$.

Calculating S_{82}. Under (A2), from (2.A.33), (2.A.39), (2.A.42) and (2.A.63), we have, since $0 < \gamma_{B_\ell} < 1$,

$$|S_{82}| \le M\gamma_{B_\ell}^2 \sum_{i=n_{T_{\ell-1}}+1}^{n_t} (1-\gamma_{B_\ell})^{n_t-i}(t-t_i) = M\gamma_{B_\ell} S_{22} = O(\Delta t). \tag{2.A.72}$$

Conclusion. Combining (2.A.68), (2.A.71) and (2.A.72), we deduce, using (A1), that

$$S_8 = \frac{\gamma_{B_\ell}}{2} f(x,t)\big\{1-(1-\gamma_{B_\ell})^{2(n_t-n_{T_{\ell-1}})}\big\} + o(\Delta t^{a_2}). \tag{2.A.73}$$

2.A.3.3 Calculating S_9 and S_{10}

Using (A1) and (2.A.64), we have

$$S_9 = (1-\gamma_{B_\ell})^{2(n_t - n_{T_{\ell-1}})}\left\{\frac{\gamma_{B_{\ell-1}}}{2} + O(\Delta t^{2a_2})\right\} = O(\Delta t^{a_2}), \tag{2.A.74}$$

since $0 < \gamma_{B_\ell} < 1$.

Note that S_{10} at (2.A.59) is equal to S_{81} at (2.A.69), and hence

$$S_{10} = O(\Delta t^{a_2}). \tag{2.A.75}$$

2.A.3.4 Conclusion

Under (A1), combining (2.A.56), (2.A.67), (2.A.73), (2.A.74) and (2.A.75), we find

$$\begin{aligned}\text{var}\{\bar{f}(x,t)\} &= \frac{R_K}{2} f(x,t)\frac{\gamma_{B_\ell}}{h_{B_\ell}}\left\{1 + \left(\frac{\gamma_{B_{\ell-1}} h_{B_\ell}}{\gamma_{B_\ell} h_{B_{\ell-1}}} - 1\right)(1-\gamma_{B_\ell})^{2(n_t - n_{T_{\ell-1}})}\right\} + o(\Delta t^{a_2-a_1}) \\ &= \frac{R_K}{2} f(x,t)\frac{\gamma_{B_\ell}}{h_{B_\ell}} + o(\Delta t^{a_2-a_1}).\end{aligned}$$

As discussed just below (2.A.59), in the special case where $n_t - n_{T_{\ell-1}} = 0$, we have $S_8 = S_{10} = 0$ and, from (2.A.56), (2.A.67) and (2.A.74), we have

$$\begin{aligned}\text{var}\{\bar{f}(x,t)\} &= \frac{R_K}{2} f(x,t)\frac{\gamma_{B_{\ell-1}}}{h_{B_{\ell-1}}}(1-\gamma_{B_\ell})^2 + o(\Delta t^{a_2-a_1}) \\ &= \frac{R_K}{2} f(x,t)\frac{\gamma_{B_\ell}}{h_{B_\ell}} + o(\Delta t^{a_2-a_1}),\end{aligned}$$

where we used (A1).

This completes the proof of (2.A.54).

2.A.4 Proof of (2.11)

Writing $\ell = \ell_t$, we first differentiate AMSE(x,t), defined in (2.10), with respect to h_ℓ and γ_ℓ respectively to obtain

$$\frac{\partial}{\partial h_\ell}\text{AMSE}(x,t) = f_{xx}^2(x,t)\mu_{K,2}^2 h_{B_\ell}^3 - 2f_{xx}(x,t)f_t(x,t)\mu_{K,2}\frac{\Delta t}{\gamma_{B_\ell}}h_{B_\ell} - \frac{1}{2}f(x,t)R_K\frac{\gamma_{B_\ell}}{h_{B_\ell}^2}\,,$$
$$\frac{\partial}{\partial \gamma_\ell}\text{AMSE}(x,t) = -2f_t^2(x,t)\frac{\Delta t^2}{\gamma_{B_\ell}^3} + f_{xx}(x,t)f_t(x,t)\mu_{K,2}\frac{\Delta t}{\gamma_{B_\ell}^2}h_{B_\ell}^2 + \frac{1}{2}f(x,t)R_K\frac{1}{h_{B_\ell}}\,.$$

Equating $\partial\,\text{AMSE}(x,t)/\partial h_\ell$ and $\partial\,\text{AMSE}(x,t)/\partial\gamma_\ell$ to 0, we get

$$\begin{cases} C_1 h_{B_\ell}^3 + C_2\dfrac{\Delta t}{\gamma_{B_\ell}}h_{B_\ell} + C_3\dfrac{\gamma_{B_\ell}}{h_{B_\ell}^2} = 0\,, \\ C_4\dfrac{\Delta t^2}{\gamma_{B_\ell}^3} + C_5\dfrac{\Delta t}{\gamma_{B_\ell}^2}h_{B_\ell}^2 + C_6\dfrac{1}{h_{B_\ell}} = 0\,, \end{cases}$$

where $C_1,\ldots,C_6$ denote some constants.

Using the fact that, under (A1), we have $h_\ell \asymp \Delta t^{a_1}$ and $\gamma_\ell \asymp \Delta t^{a_2}$ where $0 < a_1 < a_2 < 1$, the above systems of equations become

$$\begin{cases} \Delta t^{3a_1} + \Delta t^{1+a_1-a_2} \asymp \Delta t^{a_2-2a_1}\,, \\ \Delta t^{2-3a_2} + \Delta t^{1+2a_1-2a_2} \asymp \Delta t^{-a_1}\,. \end{cases}$$

For the first equation in the above system of equations to hold, we must have either $\Delta t^{3a_1} \asymp \Delta t^{a_2-2a_1}$ or $\Delta t^{1+a_1-a_2} \asymp \Delta t^{a_2-2a_1}$; for the second equation above to hold, we must have either $\Delta t^{2-3a_2} \asymp \Delta t^{-a_1}$ or $\Delta t^{1+2a_1-2a_2} \asymp \Delta t^{-a_1}$. This gives us the following 4 systems of equations:

$$\begin{cases} 3a_1 = a_2 - 2a_1 \\ 2 - 3a_2 = -a_1 \end{cases}, \begin{cases} 3a_1 = a_2 - 2a_1 \\ 1 + 2a_1 - 2a_2 = -a_1 \end{cases}, \begin{cases} 1 + a_1 - a_2 = a_2 - 2a_1 \\ 2 - 3a_2 = -a_1 \end{cases}, \begin{cases} 1 + a_1 - a_2 = a_2 - 2a_1 \\ 1 + 2a_1 - 2a_2 = -a_1 \end{cases},$$

which all give the same solution $a_1 = 1/7$ and $a_2 = 5/7$. Hence we conclude that, to minimise AMSE(x,t) at (2.10), we should take (2.11).

2.B Derivation of ACV

We write $\check{f}(x,t;\gamma,h)$ instead of $\check{f}(x,t)$ to highlight the dependence of $\check{f}(x,t)$ on (γ,h).

To minimise $\text{ISE}_{(T_{\ell-1},T_\ell]}(\gamma,h)$ at (2.19), first note that $\text{ISE}_{(T_{\ell-1},T_\ell]}(\gamma,h) = I_1 - 2I_2 + I_3$, where $I_1 = \int_{T_{\ell-1}}^{T_\ell}\int \check{f}^2(x,t;\gamma,h)\,\mathrm{d}x\,\mathrm{d}t$, $I_2 = \int_{T_{\ell-1}}^{T_\ell}\int \check{f}(x,t;\gamma,h)f(x,t)\,\mathrm{d}x\,\mathrm{d}t$ and $I_3 = \int_{T_{\ell-1}}^{T_\ell}\int f^2(x,t)\,\mathrm{d}x\,\mathrm{d}t$. Here I_1 is known and I_3 does not depend on the parameters of interest (γ,h). To use $\text{ISE}_{(T_{\ell-1},T_\ell]}(\gamma,h)$ to select (γ,h), it suffices to estimate I_2.

For $t \in (T_{\ell-1},T_\ell]$, let $X_t^0 \sim f(\cdot,t)$ denote a variable independent of the X_j's. We approximate I_2 by $\Delta t\sum_{i:X_i\in B_\ell}\int \check{f}(x,t_i;\gamma,h)f(x,t_i)\,\mathrm{d}x = \Delta t\sum_{i:X_i\in B_\ell}\mathrm{E}\{\check{f}(X_{t_i}^0, t_i;\gamma,h)|X_1,\ldots,X_i\}$, which we estimate by $\hat{I}_2 = \tau_\ell/b_\ell\sum_{i:X_i\in B_\ell}\check{f}_{-i}(X_i,t_i;\gamma,h)$, where in the last step we have used the fact that $X_i \sim f(\cdot,t_i)$, $\Delta t/\tau_\ell \sim b_\ell^{-1}$ as $\Delta t \to 0$ by (2.1), and where the leave-one-out estimator $\check{f}_{-i}(x,t_i;\gamma,h)$ is equal to $\check{f}(x,t_i;\gamma,h) - w_{i,i}K_h(x-X_i)$ with w_{ii} as above (2.18), taking $n_t = n_{t_i} = i$.

In practice we can simplify the computation of I_1 as follows. Note that with our choice of b_ℓ, the T_ℓ's always correspond to time points t_i where we have observed a data value X_i. Noting too that for a fixed $x \in \mathbb{R}$ and some $i = 1,2,\ldots,$ $\check{f}(x,t)$ remains constant for $t \in [t_i,t_{i+1})$, where t_i is defined at (1.1). Therefore, we have

$$\begin{aligned} I_1 = \sum_{i:X_i\in B_\ell}\int_{t_{i-1}}^{t_i}\int \check{f}^2(x,t;\gamma,h)\,\mathrm{d}x\,\mathrm{d}t &= \sum_{i:X_i\in B_\ell}\int_{t_{i-1}}^{t_i}\int \check{f}^2(x,t_i;\gamma,h)\,\mathrm{d}x\,\mathrm{d}t \\ &= \tau_\ell\Delta t\sum_{i:X_i\in B_\ell}\int \check{f}^2(x,t_i;\gamma,h)\,\mathrm{d}x\,\mathrm{d}t \\ &\approx b_\ell^{-1}\Delta t\sum_{i:X_i\in B_\ell}\int \check{f}^2(x,t_i;\gamma,h)\,\mathrm{d}x\,\mathrm{d}t \equiv \hat{I}_1\,, \end{aligned}$$

since, using (2.1), we have $\Delta t/\tau_\ell \sim b_\ell^{-1}$ as $\Delta t \to 0$. Therefore, to simplify computations, in practice we approximate I_1 by $\hat{I}_1$.

The above derivations suggest that, to use $\text{ISE}_{(T_{\ell-1},T_\ell]}$ in practice, we can minimise $\text{ACV}_\ell = \hat{I}_1 - 2\hat{I}_2$, which is the same as (2.20).

2.C Proof of Theorem 2.1

The main idea of our proof follows from the proofs of equation (7) and (8) in Fan et al. (1996). However, their problem is density estimation based on i.i.d. data and their estimator is a simple average of kernel functions (observations have the same weight). Here, we have to use slightly different probabilistic tools for the i.n.i.d. data and calculations are more complicated due to the non-uniform weights used by our estimator. First we present some probabilistic results needed in the proofs.

2.C.1 Some probabilistic tools

The first result is the Rosenthal inequality for martingales (Theorem 2.12 from Hall and Heyde , 1980, p. 23). First we give the definition of a martingale (see, for example, Hall and Heyde , 1980, p. 1).

Definition 2.C.1. Let $I \subset \mathbb{Z}$ be an index set and $\{\mathcal{F}_i, i \in I\}$ an increasing sequence of σ-algebras. A sequence of random variables $\{S_i, i \in I\}$ is said to be a martingale with respect to $\{\mathcal{F}_i, i \in I\}$ if: (i) S_i is measurable with respect to $\mathcal{F}_i$, (ii) $\mathrm{E}\,|S_i| < \infty$ and (iii) $\mathrm{E}(S_{i_2}|\mathcal{F}_{i_1}) = S_{i_1}$ for all $i_1, i_2 \in I$ such that $i_1 < i_2$.

Then the following Rosenthal's inequality gives an upper bound for the higher moment of a martingale.

Theorem 2.C.1. *If* $\{S_i = \sum_{j=1}^{i} X_j\}_{i=1,\ldots,n}$ *is a martingale with respect to* $\{\mathcal{F}_i\}_{i=1,\ldots,n}$ *and* $2 \leq p < \infty$*, then there exists a constant* C_p *depending on* p *such that*

$$\mathrm{E}\,|S_n|^p \leq C_p \left[\mathrm{E}\left\{ \sum_{i=1}^{n} \mathrm{E}(X_i^2|\mathcal{F}_{i-1}) \right\}^{p/2} + \sum_{i=1}^{n} \mathrm{E}\,|X_i|^p \right]. \tag{2.C.1}$$

In the special case where $X_1, \ldots, X_n$ are independent zero-mean random variables with $\mathrm{E}\,|X_i| <$

∞ for $i = 1, \ldots, n$, we have (see formula 9.7.c of Lin and Bai, 2010)

$$\mathrm{E}\,|S_n|^p \leq C_p \left\{ \left(\sum_{i=1}^{n} \mathrm{E}\,X_i^2 \right)^{p/2} + \sum_{i=1}^{n} \mathrm{E}\,|X_i|^p \right\}. \tag{2.C.2}$$

According to Hitczenko (1990), the constant C_p in (2.C.1) and (2.C.2) can be taken as

$$C_p = (C_0 p / \log p)^p, \tag{2.C.3}$$

where $C_0 > 0$ is a constant that does not depend on p or the distribution of S_n.

The second result is formula (2.10) on p. 36 of Boucheron et al. (2013), also known as Bernstein's inequality.

Theorem 2.C.2. *Let $X_1, \ldots, X_n$ be independent random variables with finite variances such that $|X_i| \leq A$, $i = 1, \ldots, n$, for some $A > 0$. Let $S_n = \sum_{i=1}^n X_i$ and $\sigma^2 = \sum_{i=1}^n \mathrm{E}(X_i^2)$. Then, for any $\eta > 0$,*

$$\mathrm{P}\{|S_n - \mathrm{E}(S_n)| \geq \eta\} \leq \exp\left\{ -\frac{\eta^2}{2(\sigma^2 + A\eta/3)} \right\}.$$

The third result is Marcus–Zinn inequality for weighted empirical processes of i.n.i.d. random variables (Shorack and Wellner, 1986, p. 820).

Theorem 2.C.3. *Suppose $X_{n1}, \ldots, X_{nn}$ are independent random variables with distribution functions $F_{n1}, \ldots, F_{nn}$. Let $c_{n1}, \ldots, c_{nn}$ be a sequence of constants. For $x \in \mathbb{R}$, let*

$$\mathbb{Z}_n(x) = \frac{1}{\sqrt{\sum_{i=1}^n c_{ni}^2}} \sum_{i=1}^{n} c_{ni} \big\{ \mathbb{1}(X_{ni} \leq x) - F_{ni}(x) \big\}$$

be the weighted empirical process. Then, for any $n \geq 1$ and $\eta > 0$,

$$\mathrm{P}\left\{ \sup_{x \in \mathbb{R}} |\mathbb{Z}_n(x)| \geq \eta \right\} \leq \big(1 + 2\sqrt{2\pi\eta}\big) e^{-\eta^2/8}.$$

Let $F_n(x) = \sum_{i=1}^n c_{ni} \mathbb{1}\{X_{ni} \leq x\}$ and $\bar{F}_n(x) = \sum_{i=1}^n c_{ni} F_{ni}(x)$ be the weighted empirical distribution function and its mean. Then, since for any $\varepsilon > 0$, we can always find $\eta > 0$ large

enough such that $\big(1+2\sqrt{2\pi\eta}\big)e^{-\eta^2/8} \le \varepsilon$, Theorem 2.C.3 implies that for any $n \ge 1$,

$$\mathrm{P}\left\{\sup_{x\in\mathbb{R}}|F_n(x)-\bar{F}_n(x)| \ge \eta\left(\sum_{i=1}^{n} c_{ni}^2\right)^{1/2}\right\} \le \varepsilon,$$

which implies that

$$\sup_{x\in\mathbb{R}}|F_n(x)-\bar{F}_n(x)| = O_p\left(\sum_{i=1}^{n} c_{ni}^2\right)^{1/2}. \tag{2.C.4}$$

2.C.2 Notation and technical results

Recalling the definition of T_ℓ, B_ℓ and b_ℓ at page 25, we denote the data in block B_ℓ by $X_{\ell 1},\ldots,X_{\ell b_\ell}$ with arrival times $t_{\ell 1},\ldots,t_{\ell b_\ell}$. Without loss of generality, we assume that b_ℓ is even and there are $b_\ell/2$ observations within $(T_{\ell-1}, s_\ell]$, the first half of the time interval corresponding to B_ℓ. Then, by (1.1) and (1.2), we have, for $i=1,\ldots,b_\ell$, $X_{\ell i} = X_{n_{s_\ell}-b_\ell/2+i}$, and, recalling that the $t_{\ell i}$'s are equidistant, we have

$$s_\ell - (b_\ell/2-i+1)\Delta t \le t_{\ell i} \le s_\ell - (b_\ell/2-i)\Delta t. \tag{2.C.5}$$

Let $\tilde{f}(x,s_\ell) = \sum_{i=n_{s_\ell}-b_\ell/2+1}^{n_{s_\ell}} w_{n_{s_\ell},i}K_{h_i}(x-X_i)$, where $w_{n_{s_\ell},i}$ is defined as $w_{n_t i}$ at (1.12), but with t substituted by s_ℓ. Using (2.18), we have

$$\tilde{f}(x,s_\ell) = \sum_{j=1}^{b_\ell/2}\tilde{w}_{\ell j}K_h(x-X_{\ell j}), \tag{2.C.6}$$

where $\tilde{w}_{\ell j} = w_{n_{s_\ell},n_{s_\ell}-b_\ell/2+j} = \gamma(1-\gamma)^{b_\ell/2-j}$.

Recall that $\check{f}(x,s_\ell)$ is defined as $\check{f}(x,t)$ at (2.2) with t replaced by s_ℓ. Then, under (A3), (B1) and (B3), we have, when Δt is small enough so that $n_{s_\ell} \ge 2$,

$$\check{f}(x,s_\ell) - \tilde{f}(x,s_\ell) = K_{h_1}(x-X_1)\prod_{j=2}^{n_{s_\ell}}(1-\gamma_j) + \sum_{i=2}^{n_{s_\ell}-b_\ell/2}\gamma_i\prod_{j=i+1}^{n_{s_\ell}}(1-\gamma_j)K_{h_i}(x-X_i)$$

$$
\begin{aligned}
&\leq \frac{M}{h_m}(1-\gamma_m)^{n_{s_\ell}-1} + \frac{M\gamma_M}{h_m}(n_{s_\ell} - b_\ell/2 - 1) \prod_{j=n_{s_\ell}-b_\ell/2+1}^{n_{s_\ell}} (1-\gamma_j) \\
&\leq \frac{M}{h_m}(1-\gamma_m)^{b_\ell/2} + \frac{M\gamma_M}{h_m} n_{s_\ell}(1-\gamma_m)^{b_\ell/2} \\
&= M\left(\frac{1}{h_m} + \frac{n_{s_\ell}\gamma_M}{h_m}\right)(1-\gamma_m)^{b_\ell/2} \\
&\leq M\left(\frac{1}{\delta}\Delta t^{-a_2} + \frac{s_\ell}{\delta^2}\Delta t^{a_2-a_1-1}\right)\exp(-3^{-1}\delta^2\Delta t^{a_2+a_3-1}) \\
&\leq \exp(-4^{-1}\delta^2\Delta t^{a_2+a_3-1}) = \exp(-4^{-1}\delta^2\Delta t^{-\alpha/7}),
\end{aligned} \tag{2.C.7}
$$

where h_m, γ_m, γ_M are defined in (A1) and we used (A1), (A4), (1.2) and the fact that, using (2.1) and (2.A.7), we have

$$
(1-\gamma_m)^{b_\ell/2} \leq \exp(-\gamma_m b_\ell/2) \leq \exp\{-\gamma_m(2^{-1}\tau_\ell\Delta t^{-1} - 1)\} \leq \exp(-3^{-1}\delta^2\Delta t^{a_2+a_3-1}), \tag{2.C.8}
$$

when Δt is small enough, we also used (B1) and (B3) to get the last equation.

For $i = 1, \ldots, b_\ell$, let

$$
\check{f}_{-\ell i}(x, s_\ell) = \check{f}_{-(n_{s_\ell}-b_\ell/2+i)}(x, s_\ell), \tag{2.C.9}
$$

where $\check{f}_{-(n_{s_\ell}-b_\ell/2+i)}(x, s_\ell)$ is defined at (2.25) with i replaced by $n_{s_\ell} - b_\ell/2 + i$ there, that is,

$$
\check{f}_{-(n_{s_\ell}-b_\ell/2+i)}(x, s_\ell) = \sum_{\substack{j=1 \\ j\neq n_{s_\ell}-b_\ell/2+i}}^{n_{s_\ell}} w_{n_{s_\ell},j} K_h(x - X_j).
$$

Then, from (2.26), we have

$$
\mathrm{SCV}_\ell(\gamma, h) = \int \check{f}^2(x, s_\ell)\,\mathrm{d}x - \frac{2}{b_\ell}\sum_{i=1}^{b_\ell} \check{f}_{-\ell i}(X_{\ell i}, s_\ell). \tag{2.C.10}
$$

Furthermore, let

$$\tilde{f}_{-\ell i}(x, s_\ell) = \sum_{j=1, j\neq i}^{b_\ell/2} \tilde{w}_{\ell j} K_h(x - X_{\ell j}), \tag{2.C.11}$$

where $\tilde{w}_{\ell j}$ is defined below (2.C.6). Using the definition of $\tilde{w}_{\ell j}$ and the fact that $X_{\ell j} = X_{n_{s_\ell} - b_\ell/2 + j}$, we have

$$\tilde{f}_{-\ell i}(x, s_\ell) = \sum_{\substack{j=n_{s_\ell}-b_\ell/2+1 \\ j\neq n_{s_\ell}-b_\ell/2+i}}^{n_{s_\ell}} w_{n_{s_\ell},j} K_h(x - X_j),$$

and hence

$$\begin{aligned}\check{f}_{-\ell i}(x, s_\ell) - \tilde{f}_{-\ell i}(x, s_\ell) &= \sum_{j=1}^{n_{s_\ell}-b_\ell/2} w_{n_{s_\ell},j} K_h(x - X_j) \\ &= \check{f}(x, s_\ell) - \tilde{f}(x, s_\ell) \leq \exp(-4^{-1}\delta^2 \Delta t^{-\alpha/7}),\end{aligned} \tag{2.C.12}$$

where we used (2.C.7).

Now, using (2.22) and (2.C.10), we have

$$\begin{aligned}\mathrm{SCV}_\ell(\gamma, h) - \mathrm{ISE}_\ell(\gamma, h) + \int f^2(x, s_\ell)\,\mathrm{d}x = 2J + 2\int \{\check{f}(x, s_\ell) - \tilde{f}(x, s_\ell)\} f(x, s_\ell)\,\mathrm{d}x \\ + \frac{2}{b_\ell}\sum_{i=1}^{b_\ell} \{\tilde{f}_{-\ell i}(X_{\ell i}, s_\ell) - \check{f}_{-\ell i}(X_{\ell i}, s_\ell)\},\end{aligned}$$

where

$$J = \int \tilde{f}(x, s_\ell) f(x, s_\ell)\,\mathrm{d}x - \frac{1}{b_\ell}\sum_{i=1}^{b_\ell} \tilde{f}_{-\ell i}(X_{\ell i}, s_\ell). \tag{2.C.13}$$

Using (2.C.12) and noting that the upper bound $\exp(-4^{-1}\delta^2\Delta t^{-\alpha/7})$ there does not dependent on (γ, h), we have

$$\left|\int \{\check{f}(x, s_\ell) - \tilde{f}(x, s_\ell)\} f(x, s_\ell)\,\mathrm{d}x\right| \leq \exp(-4^{-1}\delta^2\Delta t^{-\alpha/7}) \int f(x, s_\ell)\,\mathrm{d}x = o(\Delta t),$$

uniformly in $(\gamma, h) \in I^{\ell}_{\gamma,h}$, where the last equality is due to (B3) and (2.C.7). Using (2.C.12), (B3) and (2.C.7) again, we also obtain

$$\left| \frac{1}{b_\ell} \sum_{i=1}^{b_\ell} \{ \tilde{f}_{-\ell i}(X_{\ell i}, s_\ell) - \check{f}_{-\ell i}(X_{\ell i}, s_\ell) \} \right| \leq \exp(-4^{-1}\delta^2 \Delta t^{-\alpha/7}) = o(\Delta t),$$

uniformly in $(\gamma, h) \in I^{\ell}_{\gamma,h}$. Hence we conclude that

$$\mathrm{SCV}_\ell(\gamma, h) - \mathrm{ISE}_\ell(\gamma, h) + \int f^2(x, s_\ell)\, \mathrm{d}x = 2J + o(\Delta t), \tag{2.C.14}$$

uniformly in $(\gamma, h) \in I^{\ell}_{\gamma,h}$.

The rest of the proof is devoted to calculating J. First, for $i, j \in \{1, \ldots, b_\ell\}$, let

$$\begin{gathered} U(x, y) = K_h(x - y), \quad U_i(x) = \mathrm{E}_{\ell i}\{U(x, X_{\ell i})\}, \quad U_{ij} = \mathrm{E}_{\ell j}\{U_i(X_{\ell j})\} \\ V_{ij} = U(X_{\ell i}, X_{\ell j}) - U_i(X_{\ell j}) - U_j(X_{\ell i}) + U_{ij}, \end{gathered} \tag{2.C.15}$$

where K_h is defined below (1.4) and $\mathrm{E}_{\ell i}$ denotes the expectation with respect to $X_{\ell i}$. Then, recalling that K is symmetric from (A3), we have

$$U(x, y) = U(y, x), \quad U_{ij} = U_{ji}, \quad V_{ij} = V_{ji}, \quad \mathrm{E}(V_{ij}) = 0. \tag{2.C.16}$$

Using (2.C.11) and (2.C.15), we have

$$\begin{aligned} \sum_{i=1}^{b_\ell} \tilde{f}_{-\ell i}(X_{\ell i}, s_\ell) &= \sum_{i=1}^{b_\ell} \sum_{j=1, j\neq i}^{b_\ell/2} \tilde{w}_{\ell j} U(X_{\ell i}, X_{\ell j}) \\ &= \sum_{i=2}^{b_\ell} \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} \tilde{w}_{\ell j} U(X_{\ell i}, X_{\ell j}) + \sum_{j=2}^{b_\ell/2} \sum_{i=1}^{j-1} \tilde{w}_{\ell j} U(X_{\ell i}, X_{\ell j}) \\ &= \sum_{i=2}^{b_\ell} \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} \tilde{w}_{\ell j} U(X_{\ell i}, X_{\ell j}) + \sum_{i=2}^{b_\ell/2} \sum_{j=1}^{i-1} \tilde{w}_{\ell i} U(X_{\ell j}, X_{\ell i}) \\ &= \sum_{i=2}^{b_\ell} \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j}) U(X_{\ell i}, X_{\ell j}), \end{aligned} \tag{2.C.17}$$

where we used (2.C.16) to get the last equation, taking the convention that $\tilde{w}_{\ell i} = 0$ for $i = b_\ell/2 + 1, \ldots, b_\ell$.

Plugging (2.C.15) and (2.C.17) into (2.C.13), we have

$$J = J_4 - J_1 - J_2 - J_3, \tag{2.C.18}$$

where

$$J_1 = \frac{1}{b_\ell} \sum_{i=2}^{b_\ell} \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j}) V_{ij}, \tag{2.C.19}$$

$$J_2 = \frac{1}{b_\ell} \sum_{i=2}^{b_\ell} \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j}) \{U_i(X_{\ell j}) - U_{ij}\}, \tag{2.C.20}$$

$$J_3 = \frac{1}{b_\ell} \sum_{i=2}^{b_\ell} \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j}) \{U_j(X_{\ell i}) - U_{ij}\}, \tag{2.C.21}$$

$$J_4 = \int \tilde{f}(x, s_\ell) f(x, s_\ell) \, \mathrm{d}x - \frac{1}{b_\ell} \sum_{i=2}^{b_\ell} \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j}) U_{ij}. \tag{2.C.22}$$

Then, we calculate $J_1, \ldots, J_4$ as follows.

2.C.3 Calculating J_1

2.C.3.1 Martingale construction

We first write $b_\ell J_1 = \sum_{i=2}^{b_\ell} Y_i$, where

$$Y_i = \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j}) V_{ij}. \tag{2.C.23}$$

Then we show that $\{\sum_{i=2}^{k} Y_i\}_{k=2,\ldots,b_\ell}$ is a martingale with respect to $\{\mathcal{F}_k\}_{k=2,\ldots,b_\ell}$, where $\mathcal{F}_k$ is the σ-algebra generated by $X_{\ell 1}, \ldots, X_{\ell k}$, so that we can apply Theorem 2.C.1 to $b_\ell J_1$. For this

purpose, we only need to show the following: for $i = 2, \ldots, b_\ell$,

$$Y_i \text{ is measurable with respect to } \mathcal{F}_i,\ \mathrm{E}\,|Y_i| < \infty \text{ and } \mathrm{E}(Y_i|\mathcal{F}_{i-1}) = 0. \tag{2.C.24}$$

Now, we prove that (2.C.24) implies that $\{\sum_{i=2}^{k} Y_i\}_{k=2,\ldots,b_\ell}$ satisfies Definition 2.C.1. First note that, with (2.C.24), $\sum_{i=2}^{k} Y_i$ is measurable with respect to $\mathcal{F}_k$ and $\mathrm{E}\,|\sum_{i=2}^{k} Y_i| \leq \sum_{i=2}^{k} \mathrm{E}\,|Y_i| < \infty$. Then, for any two integers k_1, k_2 satisfying $2 \leq k_1 < k_2 \leq b_\ell$, we have

$$\begin{aligned}
\mathrm{E}\left(\sum_{i=2}^{k_2} Y_i \Big| \mathcal{F}_{k_1}\right) &= \mathrm{E}\left(\sum_{i=2}^{k_1} Y_i \Big| \mathcal{F}_{k_1}\right) + \mathrm{E}\left(\sum_{i=k_1+1}^{k_2} Y_i \Big| \mathcal{F}_{k_1}\right) \\
&= \sum_{i=2}^{k_1} Y_i + \mathrm{E}\left\{ \mathrm{E}\left(\sum_{i=k_1+1}^{k_2} Y_i \Big| \mathcal{F}_{k_1+1}\right) \Big| \mathcal{F}_{k_1} \right\} \\
&= \sum_{i=2}^{k_1} Y_i + \mathrm{E}\left(\sum_{i=k_1+2}^{k_2} Y_i \Big| \mathcal{F}_{k_1}\right) \\
&= \cdots = \sum_{i=2}^{k_1} Y_i.
\end{aligned}$$

Therefore, Definition 2.C.1 is satisfied.

To show (2.C.24), firstly, note from (2.C.15) and (2.C.23) that, for $i = 2, \ldots, b_\ell$, Y_i is a function of $X_{\ell 1}, \ldots, X_{\ell i}$ so it is measurable with respect to $\mathcal{F}_i$. Secondly, by (A3) and (2.C.15), we have

$$U(x, y) = \frac{1}{h} K\Big(\frac{x - y}{h}\Big) \leq \frac{M}{h}, \tag{2.C.25}$$

uniformly in x, y and hence $|V_{ij}| < \infty$ for all i, j, which implies that $|Y_i| < \infty$ and hence $\mathrm{E}\,|Y_i| < \infty$ for all i. Finally, we have

$$\begin{aligned}
\mathrm{E}(Y_i|\mathcal{F}_{i-1}) &= \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j})\, \mathrm{E}(V_{ij}|\mathcal{F}_{i-1}) \\
&= \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j}) \big[\, \mathrm{E}\{U(X_{\ell i}, X_{\ell j})|\mathcal{F}_{i-1}\}
\end{aligned}$$

$$
\begin{aligned}
&\qquad - \mathrm{E}\{U_i(X_{\ell j})|\mathcal{F}_{i-1}\} - \mathrm{E}\{U_j(X_{\ell i})|\mathcal{F}_{i-1}\} + U_{ij}\big] \\
&= \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j})\big[\mathrm{E}\{U(X_{\ell i}, X_{\ell j})|X_{\ell j}\} \\
&\qquad - U_i(X_{\ell j}) - \mathrm{E}\{U_j(X_{\ell i})\} + U_{ij}\big] \\
&= 0, \qquad (2.C.26)
\end{aligned}
$$

where we used the independence of $X_{\ell 1}, \ldots, X_{\ell b_\ell}$ to get the third equation and (2.C.16) to get the last equation.

2.C.3.2 Moment inequalities of $b_\ell J_1$

Using Theorem 2.C.1, we have that for any $p \geq 2$,

$$
\mathrm{E}\,|b_\ell J_1|^p \leq C_p \left[\mathrm{E}\left\{ \sum_{i=2}^{b_\ell} \mathrm{E}(Y_i^2|\mathcal{F}_{i-1}) \right\}^{p/2} + \sum_{i=2}^{b_\ell} \mathrm{E}\,|Y_i|^p \right]. \qquad (2.C.27)
$$

Note that $g_1(x) = x^{p/2}$ is convex on $(0, \infty)$ when $p \geq 2$ and hence, by conditional Jensen's inequality (see, for example, Theorem A29 in McLeish, 2005), we have $g_1\big\{\mathrm{E}(Y_i^2|\mathcal{F}_{i-1})\big\} \leq \mathrm{E}\{g_1(Y_i^2)|\mathcal{F}_{i-1}\}$, which leads to $\mathrm{E}(Y_i^2|\mathcal{F}_{i-1}) \leq \big\{\mathrm{E}\big(|Y_i|^p|\mathcal{F}_{i-1}\big)\big\}^{2/p}$. Hence, we have

$$
\begin{aligned}
\mathrm{E}\left\{ \sum_{i=2}^{b_\ell} \mathrm{E}(Y_i^2|\mathcal{F}_{i-1}) \right\}^{p/2} &\leq \mathrm{E}\left[\sum_{i=2}^{b_\ell} \big\{\mathrm{E}\big(|Y_i|^p|\mathcal{F}_{i-1}\big)\big\}^{2/p} \right]^{p/2} \\
&\leq \left(\sum_{i=2}^{b_\ell} \Big[\mathrm{E}\big\{\mathrm{E}(|Y_i|^p|\mathcal{F}_{i-1})\big\}\Big]^{2/p} \right)^{p/2} \\
&= \left\{ \sum_{i=2}^{b_\ell} (\mathrm{E}\,|Y_i|^p)^{2/p} \right\}^{p/2}, \qquad (2.C.28)
\end{aligned}
$$

where we used conditional Minkowski's inequality (see, for example, Theorem 2.16 in Boucheron et al., 2013) to get the second inequality.

Now, by Minkowski's inequality, we have $\sum_{i=2}^{b_\ell} \mathrm{E}\,|Y_i|^p = \mathrm{E}\big(\sum_{i=2}^{b_\ell} |Y_i|^p\big) \leq \big\{\sum_{i=2}^{b_\ell}$

$\left(\mathrm{E}\,|Y_i|^p\right)^{2/p}\Big\}^{p/2}$ for any $p \geq 2$. Combining (2.C.27) with (2.C.28) and the above result, we have

$$\mathrm{E}\,|b_\ell J_1|^p \leq 2C_p\left\{\sum_{i=2}^{b_\ell}(\mathrm{E}\,|Y_i|^p)^{2/p}\right\}^{p/2}. \tag{2.C.29}$$

Observe from (2.C.29) that, to obtain an upper bound for $\mathrm{E}\,|b_\ell J_1|^p$, it suffices to obtain an upper bound for $\mathrm{E}\,|Y_i|^p$.

To bound $\mathrm{E}\,|Y_i|^p$ for $i = 2, \ldots, b_\ell$, we first rewrite the definition of Y_i at (2.C.23) as $Y_i = \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} V_{ij}^0$, where $V_{ij}^0 = (\tilde{w}_{\ell i} + \tilde{w}_{\ell j})V_{ij}$. Next we show that $\{Y_j\}_{j=2,\ldots,i}$ is a martingale with respect to filtration $\{\mathcal{F}_{ij}\}_{j=1,\ldots,i}$, where, for $i = 1, \ldots, b_\ell/2$, $\mathcal{F}_{ij}$ is the σ-algebra generated by $\{X_{\ell i}; X_{\ell 1}, \ldots, X_{\ell j}\}$. For this, Similar to (2.C.24), it suffices to show that

$$V_{ij}^0 \text{ is measurable with respect to } \mathcal{F}_{ij},\ \mathrm{E}\,|V_{ij}^0| < \infty \text{ and } \mathrm{E}(V_{ij}^0|\mathcal{F}_{i\,j-1}) = 0. \tag{2.C.30}$$

To show (2.C.30), note from (2.C.15) that $V_{ij}^0 = (\tilde{w}_{\ell i} + \tilde{w}_{\ell j})V_{ij}$ is a function of $\{X_{\ell i}; X_{\ell 1}, \ldots, X_{\ell j}\}$ and hence is measurable with respect to $\mathcal{F}_{ij}$. Second, note that we have already showed under (2.C.25) that $|V_{ij}| < 0$ and hence $\mathrm{E}\,|V_{ij}^0| < \infty$ holds trivially. Finally, note from (2.C.15) that, conditioning on $X_{\ell i}$, $V_{i1}, \ldots, V_{i\,(i-1)\wedge(b_\ell/2)}$ are independent random variables and we have

$$\begin{aligned}\mathrm{E}(V_{ij}|X_{\ell i}) &= \mathrm{E}\{U(X_{\ell i}, X_{\ell j})|X_{\ell i}\} - \mathrm{E}\{U_i(X_{\ell j})\} - U_j(X_{\ell i}) + U_{ij} \\ &= U_j(X_{\ell i}) - U_{ij} - U_j(X_{\ell i}) + U_{ij} = 0.\end{aligned}$$

Hence, we have $\mathrm{E}(V_{ij}^0|\mathcal{F}_{i\,j-1}) = \mathrm{E}(V_{ij}^0|X_{\ell i}; X_{\ell 1}, \ldots, X_{\ell\,j-1}) = \mathrm{E}(V_{ij}^0|X_{\ell i}) = 0$. Now (2.C.30) is proved.

Knowing that $\{Y_j\}_{j=2,\ldots,i}$ is a martingale, we apply (2.C.1) and obtain

$$\mathrm{E}\,|Y_i|^p = C_p\left[\mathrm{E}\left\{\sum_{j=1}^{(i-1)\wedge(b_\ell/2)}\mathrm{E}(|V_{ij}^0|^2|\mathcal{F}_{i\,j-1})\right\}^{p/2} + \sum_{j=1}^{(i-1)\wedge(b_\ell/2)}\mathrm{E}\,|V_{ij}^0|^p\right]$$

$$= C_p \left[\mathrm{E} \left\{ \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j})^2 \, \mathrm{E}(V_{ij}^2 | X_{\ell i}) \right\}^{p/2} + \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j})^p \, \mathrm{E} \, |V_{ij}|^p \right]. \tag{2.C.31}$$

Hence, to bound $\mathrm{E}\,|Y_i|^p$, it suffices to bound $\mathrm{E}(V_{ij}^2|X_{\ell i})$ and $\mathrm{E}\,|V_{ij}|^p$.

Bounding $\mathrm{E}(V_{ij}^2|X_{\ell i})$. First, from (2.C.15), we have

$$\begin{aligned}
&\mathrm{E}(V_{ij}^2|X_{\ell i}) \\
&= \mathrm{E}\{U^2(X_{\ell i}, X_{\ell j})|X_{\ell i}\} + \mathrm{E}\{U_i^2(X_{\ell j})\} + U_j^2(X_{\ell i}) + U_{ij}^2 \\
&\quad - 2\,\mathrm{E}\{U_i(X_{\ell j})U(X_{\ell i}, X_{\ell j})|X_{\ell i}\} - 2U_j(X_{\ell i})\,\mathrm{E}\{U(X_{\ell i}, X_{\ell j})|X_{\ell i}\} \\
&\quad + 2U_{ij}\,\mathrm{E}\{U(X_{\ell i}, X_{\ell j})|X_{\ell i}\} + 2U_j(X_{\ell i})\,\mathrm{E}\{U_i(X_{\ell j})\} - 2U_{ij}\,\mathrm{E}\{U_i(X_{\ell j})\} - 2U_{ij}U_j(X_{\ell i}) \\
&= \mathrm{E}\{U^2(X_{\ell i}, X_{\ell j})|X_{\ell i}\} - 2\,\mathrm{E}\{U_i(X_{\ell j})U(X_{\ell i}, X_{\ell j})|X_{\ell i}\} \\
&\qquad\qquad + \mathrm{E}\{U_i^2(X_{\ell j})\} + 2U_{ij}U_j(X_{\ell i}) - U_j^2(X_{\ell i}) - U_{ij}^2 \\
&\le \mathrm{E}\{U^2(X_{\ell i}, X_{\ell j})|X_{\ell i}\} + \mathrm{E}\{U_i^2(X_{\ell j})\} + 2U_{ij}U_j(X_{\ell i}).
\end{aligned} \tag{2.C.32}$$

Next, we bound terms in (2.C.32).

Let $\varphi_j(x) = \mathrm{E}\{U^2(x, X_{\ell j})\}$, where U is defined in (2.C.15). Then, under (A2) and (A3), we have

$$\begin{aligned}
\varphi_j(x) &= \frac{1}{h^2} \int K^2\Big(\frac{x-y}{h}\Big) f(y, t_{\ell j}) \, \mathrm{d}y = \frac{1}{h} \int K^2(u) f(x - hu, t_{\ell j}) \, \mathrm{d}u \\
&\le \frac{M}{h} \int K^2(u) \, \mathrm{d}u = \frac{C_1}{h},
\end{aligned}$$

where $C_1 = M \int K^2(u) \, \mathrm{d}u$, and hence

$$\mathrm{E}\{U^2(X_{\ell i}, X_{\ell j})|X_{\ell i}\} = \varphi_j(X_{\ell i}) \le \frac{C_1}{h}. \tag{2.C.33}$$

Then, under (A2) and (A3), we have

$$U_i(x) = \frac{1}{h}\int K\Big(\frac{x-y}{h}\Big) f(y, t_{\ell i})\,\mathrm{d}y = \int K(u) f(x - hu, t_{\ell i})\,\mathrm{d}u$$
$$\leq M \int K(u)\,\mathrm{d}u = M, \tag{2.C.34}$$

and hence, using (2.C.34), we have

$$\mathrm{E}\{U_i^2(X_{\ell j})\} \leq M^2, \quad U_{ij}U_j(X_{\ell i}) = \mathrm{E}\{U_i(X_{\ell j})\}U_j(X_{\ell i}) \leq M^2 \tag{2.C.35}$$

Combining (2.C.33) and (2.C.35), we have, when Δt is small enough so that $h^{-1} \geq 1$,

$$\mathrm{E}(V_{ij}^2 | X_{\ell i}) \leq \frac{C_2}{h}, \tag{2.C.36}$$

where $C_2 = C_1 + 3M^2$.

Bounding $\mathrm{E}\,|V_{ij}|^p$**.** Applying (2.C.25) to V_{ij} defined in (2.C.15), we have $|V_{ij}| \leq C_3 h^{-1}$ for $C_3 = 4M$. Hence, we have

$$\mathrm{E}\,|V_{ij}|^p \leq \frac{C_3^p}{h^p}. \tag{2.C.37}$$

Conclusion. Now, plugging (2.C.36) and (2.C.37) into (2.C.31), we have

$$\mathrm{E}\,|Y_i|^p \leq C_p \left[C_2^{p/2} h^{-p/2} \left\{ \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j})^2 \right\}^{p/2} + C_3^p h^{-p} \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j})^p \right]$$
$$\leq C_p (2^{p/2} C_2^{p/2} h^{-p/2} b_\ell^{p/2} \gamma^p + 2^p C_3^p h^{-p} b_\ell \gamma^p)$$
$$\leq C_4^p C_p (b_\ell^{p/2} \gamma^p h^{-p/2} + b_\ell \gamma^p h^{-p}), \tag{2.C.38}$$

where, recalling the definition of $\tilde{w}_{\ell i}$ and $\tilde{w}_{\ell j}$ below (2.C.6) and the fact that $0 < \gamma < 1$, the second inequality follows from

$$\tilde{w}_{\ell i} + \tilde{w}_{\ell j} \leq 2\gamma, \tag{2.C.39}$$

and where $C_4 = (2C_2)^{1/2} \vee (2C_3)$.

Plugging (2.C.38) into (2.C.29), we have, under (B1),

$$\begin{aligned} \mathrm{E}\,|J_1|^p &\le 2b_\ell^{-p} C_p \Big[b_\ell \big\{ C_4^p C_p (b_\ell^{p/2} \gamma^p h^{-p/2} + b_\ell \gamma^p h^{-p}) \big\}^{2/p} \Big]^{p/2} \\ &= 2C_4^p C_p^2 (\gamma^p h^{-p/2} + b_\ell^{1-p/2} \gamma^p h^{-p}). \end{aligned} \tag{2.C.40}$$

To obtain a bound of b_ℓ, we use (2.1) to get

$$b_\ell \le \tau_\ell/\Delta t + 1 \le \delta^{-1} \Delta t^{a_3 - 1} + 1 = \delta^{-1} \Delta t^{-(5+\alpha)/7} + 1 \le \delta^{-2} \Delta t^{-(5+\alpha)/7},$$

where we used (A4) to get the the second inequality, (B3) to get the equality and, to get the last inequality, we used the fact that $\delta^{-2} > \delta^{-1} > 1$ and $\Delta t^{-(5+\alpha)/7} \to \infty$ as $\Delta t \to 0$. Using the above bound for b_ℓ and (B1), we have

$$\mathrm{E}\,|J_1|^p \le 2C_4^p C_p^2 \big\{ \delta^{-3p/2} \Delta t^{9p/14} + \delta^{-(p+2)} \Delta t^{(13+\alpha)p/14 - (5+\alpha)/7} \big\}, \tag{2.C.41}$$

According to (2.C.3), we can take $C_p = (C_0 p/\log p)^p$. Then, taking $p \ge 2(5+\alpha)/(4+\alpha) > 2$, (2.C.41) becomes

$$\begin{aligned} \mathrm{E}\,|J_1|^p &\le 2C_0^{2p} C_4^p (p/\log p)^{2p} \big\{ \delta^{-3p/2} \Delta t^{9p/14} + \delta^{-(p+2)} \Delta t^{(13+\alpha)p/14 - (5+\alpha)/7} \big\} \\ &\le 4\delta^{-2p} C_0^{2p} C_4^p (p/\log p)^{2p} \Delta t^{9p/14}, \end{aligned}$$

where the last inequality follows from the fact that, since $p > 2$ and $\delta^{-1} > 1$, both $\delta^{-3p/2}$ and $\delta^{-(p+2)}$ are bounded by δ^{-2p}. Now, by the Markov inequality, we have, for any $\eta > 0$,

$$\mathrm{P}\big(|J_1| > \eta \Delta t^{4/7}\big) \le \frac{\mathrm{E}\,|J_1|^p}{(\eta \Delta t^{4/7})^p} \le 4 \bigg(\frac{C_0^2 C_4}{\eta \delta^2 \log^2 p} \bigg)^p (p^2 \Delta t^{1/14})^p.$$

Taking $p = \Delta t^{-1/28}/\sqrt{e}$, we have, as $\Delta t \to 0$,

$$\mathrm{P}(|J_1| > \eta \Delta t^{4/7}) = o\big\{ \exp(-\Delta t^{-1/28}/\sqrt{e}) \big\},$$

which implies that, under (B1), for any $\eta > 0$,

$$\begin{aligned}
\mathrm{P}\left\{ \sup_{(\gamma,h)\in I^{\ell}_{\gamma,h}} |J_1| > \eta\Delta t^{4/7} \right\} &\leq \sum_{(\gamma,h)\in I^{\ell}_{\gamma,h}} \mathrm{P}(|J_1| > \eta\Delta t^{4/7}) \\
&\leq \#(I^{\ell}_{\gamma,h}) \sup_{(\gamma,h)\in I^{\ell}_{\gamma,h}} \mathrm{P}(|J_1| > \eta\Delta t^{4/7}) \\
&= o\big\{\Delta t^{-a} \exp(-\Delta t^{-1/28}/\sqrt{e})\big\} = o(1),
\end{aligned}$$

as $\Delta t \to 0$. Hence, we have

$$J_1 = o_p(\Delta t^{4/7}), \tag{2.C.42}$$

uniformly in $(\gamma, h) \in I^{\ell}_{\gamma,h}$.

2.C.4 Expansion of J_2

We first reorganise terms in J_2 at (2.C.20) to get

$$\begin{aligned}
J_2 &= \frac{1}{b_\ell} \sum_{i=2}^{b_\ell/2} \sum_{j=1}^{i-1} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j})\{U_i(X_{\ell j}) - U_{ij}\} + \frac{1}{b_\ell} \sum_{i=b_\ell/2+1}^{b_\ell} \sum_{j=1}^{b_\ell/2} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j})\{U_i(X_{\ell j}) - U_{ij}\} \\
&= \frac{1}{b_\ell} \sum_{j=1}^{b_\ell/2} \sum_{i=j+1}^{b_\ell/2} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j})\{U_i(X_{\ell j}) - U_{ij}\} + \frac{1}{b_\ell} \sum_{j=1}^{b_\ell/2} \sum_{i=b_\ell/2+1}^{b_\ell} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j})\{U_i(X_{\ell j}) - U_{ij}\} \\
&= \frac{1}{b_\ell} \sum_{j=1}^{b_\ell/2} \sum_{i=j+1}^{b_\ell} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j})\{U_i(X_{\ell j}) - U_{ij}\}.
\end{aligned} \tag{2.C.43}$$

Now, let $W_i(x) = U_i(x) - f(x, t_{\ell i})$. Then, using (2.C.15) and (2.C.43), we have $J_2 = J_{21} + J_{22}$, where

$$J_{21} = \frac{1}{b_\ell} \sum_{j=1}^{b_\ell/2} \sum_{i=j+1}^{b_\ell} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j})\big[W_i(X_{\ell j}) - \mathrm{E}\{W_i(X_{\ell j})\}\big], \tag{2.C.44}$$

$$J_{22} = \frac{1}{b_\ell} \sum_{j=1}^{b_\ell/2} \sum_{i=j+1}^{b_\ell} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j})\big[f(X_{\ell j}, t_{\ell i}) - \mathrm{E}\{f(X_{\ell j}, t_{\ell i})\}\big]. \tag{2.C.45}$$

Let $Z_{1j} = b_\ell^{-1} \sum_{i=j+1}^{b_\ell} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j})[W_i(X_{\ell j}) - \mathrm{E}\{W_i(X_{\ell j})\}]$. Then $J_{21} = \sum_{j=1}^{b_\ell/2} Z_{1j}$ is a sum of $b_\ell/2$ independent random variables since each summand Z_{1j} only depends on $X_{\ell j}$. To apply Theorem 2.C.2 on J_{21}, we first derive the bounds on Z_{1j} and $\sum_{j=1}^{b_\ell/2} \mathrm{E}(Z_{1j}^2)$.

Bounding Z_{1j}. Note from (2.C.15) that, under (A2) and (A3), we have

$$\begin{aligned}
|W_i(x)| &= \left| \int K_h(x-y) f(y, t_{\ell i}) \, \mathrm{d}y - f(x, t_{\ell i}) \right| \\
&= \left| \int K(u) f(x - hu, t_{\ell i}) \, \mathrm{d}u - f(x, t_{\ell i}) \right| \\
&= \left| \int K(u) \Big\{ f(x, t_{\ell i}) - huf_x(x, t_{\ell i}) \right. \\
&\qquad\qquad \left. + \frac{1}{2} h^2 u^2 K(u) f_{xx}(x - \theta_i hu, t_{\ell i}) \Big\} \, \mathrm{d}u - f(x, t_{\ell i}) \right| \\
&= \left| \frac{1}{2} h^2 \int u^2 K(u) f_{xx}(x - \theta_i hu, t_{\ell i}) \, \mathrm{d}u \right| \le C_1 h^2 \qquad \text{(2.C.46)}
\end{aligned}$$

uniformly in x, where $\theta_i \in [0, 1]$ and we used a first-order Taylor expansion of $f(x - hu, t_{\ell i})$ around x to get the third equation, and where $C_1 = M\mu_{K,2}/2$ with $\mu_{K,2}$ defined above Proposition 2.1.

Hence, using (B1) and (2.C.46), we have

$$\begin{aligned}
|Z_{1j}| &\le \frac{1}{b_\ell} \sum_{i=j+1}^{b_\ell} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j}) \{ |W_i(X_{\ell j})| + \mathrm{E}\,|W_i(X_{\ell j})| \} \\
&\le 2C_1 h^2 \frac{1}{b_\ell} \sum_{i=j+1}^{b_\ell} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j}) \qquad \text{(2.C.47)} \\
&\le 4C_1 h^2 \gamma \le C_2 \Delta t, \qquad \text{(2.C.48)}
\end{aligned}$$

where we used (2.C.39) to get the third inequality and where $C_2 = 4C_1/\delta^3$.

Bounding $\sum_{j=1}^{b_\ell/2} \mathrm{E}(Z_{1j}^2)$. Recalling the definition of $\tilde{w}_{\ell j}$ below (2.C.6) and using (B1), (B3) and (2.C.47), we have

$$\begin{aligned}
\sum_{j=1}^{b_\ell/2} \mathrm{E}(Z_{1j}^2) &\le \frac{4C_1^2h^4}{b_\ell^2} \sum_{j=1}^{b_\ell/2} \left\{ \sum_{i=j+1}^{b_\ell} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j}) \right\}^2 \\
&= \frac{4C_1^2h^4}{b_\ell^2} \sum_{j=1}^{b_\ell/2} \left\{ \gamma \sum_{i=j+1}^{b_\ell/2} (1-\gamma)^{b_\ell/2-i} + (b_\ell - j)\gamma(1-\gamma)^{b_\ell/2-j} \right\}^2 \\
&\le \frac{4C_1^2h^4}{b_\ell^2} \sum_{j=1}^{b_\ell/2} \big\{ 1 + (b_\ell - j)\gamma(1-\gamma)^{b_\ell/2-j} \big\}^2 \\
&= \frac{4C_1^2h^4}{b^2} \sum_{j=1}^{b_\ell/2} \big\{ 1 + (b_\ell - j)^2\gamma^2(1-\gamma)^{2(b_\ell/2-j)} + 2(b_\ell - j)\gamma(1-\gamma)^{b_\ell/2-j} \big\} \\
&= \frac{4C_1^2h^4}{b_\ell^2} \left\{ \frac{b_\ell}{2} + \sum_{i=b_\ell/2}^{b_\ell-1} (i\gamma)^2(1-\gamma)^{2(i-b_\ell/2)} + 2 \sum_{i=b_\ell/2}^{b_\ell-1} i\gamma(1-\gamma)^{i-b_\ell/2} \right\} \\
&\le \frac{4C_1^2h^4}{b_\ell^2} \left[\frac{b_\ell}{2} + \sum_{i=b_\ell/2}^{b_\ell-1} \big\{ (i\gamma)^2 + 2i\gamma \big\} (1-\gamma)^{i-b_\ell/2} \right] \\
&= \frac{4C_1^2h^4}{b_\ell^2} \left[\frac{b_\ell}{2} + \big\{ (b_\ell\gamma/2)^2 + b_\ell\gamma \big\} \sum_{i=b_\ell/2}^{b_\ell-1} (1-\gamma)^{i-b_\ell/2} \right. \\
&\qquad \left. + \sum_{j=b_\ell/2}^{b_\ell-2} \{(2j+1)\gamma + 2\} \gamma \sum_{i=j+1}^{b_\ell-1} (1-\gamma)^{i-b_\ell/2} \right] \\
&\le \frac{4C_1^2h^4}{b_\ell^2} \left[\frac{b_\ell}{2} + \gamma(b_\ell/2)^2 + b_\ell + \sum_{j=b_\ell/2}^{b_\ell-2} \{(2j+1)\gamma + 2\} \right] \\
&\le \frac{4C_1^2h^4}{b_\ell^2} \left\{ \frac{b_\ell}{2} + \gamma(b_\ell/2)^2 + b_\ell + b_\ell^2\gamma + b_\ell \right\} \\
&\le \frac{10C_1^2h^4}{b_\ell} (1 + b_\ell\gamma) \le 10C_1^2h^4 \{ (\tau_\ell/\Delta t - 1)^{-1} + \gamma \} \\
&\le \frac{20C_1^2}{\delta^4} \Delta t^{4/7} \big\{ \Delta t^{(5+\alpha)/7} + \delta^{-1}\Delta t^{5/7} \big\} \le C_3 \Delta t^{9/7}, \qquad (2.C.49)
\end{aligned}$$

where $C_3 = 40C_1^2/\delta^5$ and where we used summation by parts to get the seventh line and (2.1) for the second last line.

Conclusion. With (2.C.48) and (2.C.49), we apply Theorem 2.C.2 to J_{21} at (2.C.44) and get

$$\mathrm{P}\big(|J_{21}| > \eta\Delta t^{4/7}\big) \le \exp\bigg\{-\frac{\eta^2\Delta t^{8/7}}{2(C_3\Delta t^{9/7} + \eta C_2\Delta t^{11/7}/3)}\bigg\}.$$

Then, under (B1), we have

$$\begin{aligned}
\mathrm{P}\bigg\{\sup_{(\gamma,h)\in I^\ell_{\gamma,h}} |J_{21}| > \eta\Delta t^{4/7}\bigg\} &\le \sum_{(\gamma,h)\in I^\ell_{\gamma,h}} \mathrm{P}\big(|J_{21}| > \eta\Delta t^{4/7}\big) \\
&\le \#(I^\ell_{\gamma,h}) \sup_{(\gamma,h)\in I^\ell_{\gamma,h}} \mathrm{P}\big(|J_{21}| > \eta\Delta t^{4/7}\big) \\
&\le C\Delta t^{-a} \exp\bigg\{-\frac{\eta^2\Delta t^{8/7}}{2(C_3\Delta t^{9/7} + \eta C_2\Delta t^{11/7}/3)}\bigg\} = o(1),
\end{aligned}$$

which implies that $J_{21} = o_p(\Delta t^{4/7})$ uniformly in $(\gamma, h) \in I^\ell_{\gamma,h}$. Hence, from (2.C.44) and (2.C.45), we have

$$J_2 = J_{22} + o_p(\Delta t^{4/7}), \tag{2.C.50}$$

uniformly in $(\gamma, h) \in I^\ell_{\gamma,h}$.

2.C.5 Expansion of J_3

Using (2.C.21) and recalling the definition of $W_i(x)$ below (2.C.43), we have $J_3 = J_{31} + J_{32}$, where

$$J_{31} = \frac{1}{b_\ell}\sum_{i=2}^{b_\ell}\sum_{j=1}^{(i-1)\wedge(b_\ell/2)} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j})\big[W_j(X_{\ell i}) - \mathrm{E}\{W_j(X_{\ell i})\}\big], \tag{2.C.51}$$

$$J_{32} = \frac{1}{b_\ell}\sum_{i=2}^{b_\ell}\sum_{j=1}^{(i-1)\wedge(b_\ell/2)} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j})\big[f(X_{\ell i}, t_{\ell j}) - \mathrm{E}\{f(X_{\ell i}, t_{\ell j})\}\big]. \tag{2.C.52}$$

Next, we calculate J_{31} and postpone the calculation of J_{32} to §2.C.7.

Let $Z_{2i} = b_\ell^{-1}\sum_{j=1}^{(i-1)\wedge(b_\ell/2)}(\tilde{w}_{\ell i} + \tilde{w}_{\ell j})\big[W_j(X_{\ell i}) - \mathrm{E}\{W_j(X_{\ell i})\}\big]$. Then $J_{31} = \sum_{i=2}^{b_\ell} Z_{2i}$ is a sum of $b_\ell - 1$ independent random variables, since each summand Z_{2i} only depends on $X_{\ell i}$.

To apply Theorem 2.C.2 to J_{31}, we need to bound Z_{2i} and $\sum_{i=2}^{b_\ell} \mathrm{E}(Z_{2i}^2)$.

Bounding Z_{2i}. Using (B1) and (2.C.46), we have

$$
\begin{aligned}
|Z_{2i}| &\le \frac{1}{b_\ell} \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j})\{|W_j(X_{\ell i})| + \mathrm{E}\,|W_j(X_{\ell i})|\} \\
&\le 2C_1 h^2 \frac{1}{b_\ell} \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j}) \qquad (2.C.53) \\
&\le 4C_1 h^2 \gamma \le C_2 \Delta t, \qquad (2.C.54)
\end{aligned}
$$

where we used (2.C.39) to get the third inequality and C_2 is defined blow (2.C.48).

Bounding $\sum_{i=2}^{b_\ell} \mathrm{E}(Z_{2i}^2)$. Recall that $\tilde{w}_{\ell j}$ is defined below (2.C.6) and that below (2.C.17), we took the convention that $\tilde{w}_{\ell i} = 0$ for $i = b_\ell/2+1, \ldots, b_\ell$. Then, using (B1), (B3) and (2.C.53), we have

$$
\begin{aligned}
&\sum_{i=2}^{b_\ell} \mathrm{E}(Z_{2i}^2) \\
&\quad \le \frac{4C_1^2 h^4}{b_\ell^2} \sum_{i=2}^{b_\ell} \left\{ \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j}) \right\}^2 \\
&\quad = \frac{4C_1^2 h^4}{b_\ell^2} \left[\sum_{i=2}^{b_\ell/2} \left\{ \sum_{j=1}^{i-1} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j}) \right\}^2 + \sum_{i=b_\ell/2+1}^{b_\ell} \left(\sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} \right)^2 \right] \\
&\quad = \frac{4C_1^2 h^4}{b_\ell^2} \left[\sum_{i=2}^{b_\ell/2} \{ (i-1)\gamma(1-\gamma)^{b_\ell/2-i} \right. \\
&\qquad\qquad \left. + (1-\gamma)^{b_\ell/2-i+1} - (1-\gamma)^{b_\ell/2} \}^2 + \frac{b_\ell}{2} \{1 - (1-\gamma)^{b_\ell/2}\}^2 \right] \\
&\quad \le \frac{4C_1^2 h^4}{b_\ell^2} \left\{ \sum_{i=2}^{b_\ell/2} (i\gamma + 1)^2 (1-\gamma)^{2(b_\ell/2-i)} + \frac{b_\ell}{2} \right\} \\
&\quad \le \frac{4C_1^2 h^4}{b_\ell^2} \left\{ (b_\ell\gamma + 1)^2 \sum_{i=2}^{b_\ell/2} (1-\gamma)^{b_\ell/2-i} + \frac{b_\ell}{2} \right\}
\end{aligned}
$$

$$\leq \frac{4C_1^2h^4}{b_\ell^2}\left\{\frac{(b_\ell\gamma+1)^2}{\gamma}+\frac{b_\ell}{2}\right\} \leq \frac{10C_1^2h^4}{b_\ell^2}\big(b_\ell^2\gamma+b_\ell+\gamma^{-1}\big)$$
$$\leq \frac{10C_1^2}{\delta^7}\Delta t^{(14+2\alpha)/7}\big\{\Delta t^{-(5+2\alpha)/7}+\Delta t^{-(5+\alpha)/7}+\Delta t^{-5/7}\big\} \leq C_4\Delta t^{9/7}, \qquad (2.C.55)$$

where $C_4 = 30C_1^2/\delta^7$.

Conclusion. Using (2.C.54) and (2.C.55), we apply Theorem 2.C.2 to J_{31} and get

$$\mathrm{P}\big(|J_{31}| > \eta\Delta t^{4/7}\big) \leq \exp\left\{-\frac{\eta^2\Delta t^{8/7}}{2(C_4\Delta t^{9/7}+\eta C_2\Delta t^{11/7}/3)}\right\}.$$

Then, under (B1), we have

$$\begin{aligned}\mathrm{P}\left\{\sup_{(\gamma,h)\in I^\ell_{\gamma,h}}|J_{31}| > \eta\Delta t^{4/7}\right\} &\leq \sum_{(\gamma,h)\in I^\ell_{\gamma,h}}\mathrm{P}\big(|J_{31}| > \eta\Delta t^{4/7}\big)\\ &\leq \#(I^\ell_{\gamma,h})\sup_{(\gamma,h)\in I^\ell_{\gamma,h}}\mathrm{P}\big(|J_{31}| > \eta\Delta t^{4/7}\big)\\ &\leq C\Delta t^{-a}\exp\left\{-\frac{\eta^2\Delta t^{8/7}}{2(C_4\Delta t^{9/7}+\eta C_2\Delta t^{11/7}/3)}\right\} = o(1),\end{aligned}$$

which implies that $J_{31} = o_p(\Delta t^{4/7})$ uniformly in $(\gamma, h) \in I^\ell_{\gamma,h}$. Hence, from (2.C.51) and (2.C.52), we have

$$J_3 = J_{32} + o_p(\Delta t^{4/7}), \qquad (2.C.56)$$

uniformly in $(\gamma, h) \in I^\ell_{\gamma,h}$.

2.C.6 Calculating J_4

From (2.C.22), we have $J_4 = J_{41} + J_{42}$, where

$$J_{41} = \int\big[\tilde{f}(x, s_\ell) - \mathrm{E}\{\tilde{f}(x, s_\ell)\}\big]f(x, s_\ell)\,\mathrm{d}x, \qquad (2.C.57)$$

$$J_{42} = \int \mathrm{E}\{\tilde{f}(x, s_\ell)\} f(x, s_\ell)\, \mathrm{d}x - \frac{1}{b_\ell} \sum_{i=2}^{b_\ell} \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j}) U_{ij}. \tag{2.C.58}$$

For $j = 1, \ldots, b_\ell$, let $F(\cdot, t_{\ell j})$ be the distribution function of $X_{\ell j}$ and

$$F_{b_\ell/2}(y) = \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} \mathbb{1}\{X_{\ell j} \le y\}, \quad \bar{F}_{b_\ell/2}(y) = \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} F(y, t_{\ell j}), \tag{2.C.59}$$

where $\tilde{w}_{\ell j}$ is defined below (2.C.6).

Now, by (2.C.4), we have

$$\sup_{y \in \mathbb{R}} |F_{b_\ell/2}(y) - \bar{F}_{b_\ell/2}(y)| = O_p\left(\sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j}^2 \right)^{1/2} = O_p(\gamma^{1/2}), \tag{2.C.60}$$

where, recalling the definition of $\tilde{w}_{\ell j}$ below (2.C.6), we used the fact that

$$\sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j}^2 = \gamma^2 \sum_{j=1}^{b_\ell/2} (1-\gamma)^{2(b_\ell/2 - j)} \le \gamma^2 \sum_{j=1}^{b_\ell/2} (1-\gamma)^{b_\ell/2 - j} = \gamma^2 \sum_{i=0}^{b_\ell/2 - 1} (1-\gamma)^i \le \gamma,$$

since $0 < \gamma < 1$.

Next we calculate J_{41} and J_{42}.

2.C.6.1 Calculating J_{41}

Note that, from (2.C.6) and (2.C.59), we have

$$\tilde{f}(x, s_\ell) - \mathrm{E}\big\{\tilde{f}(x, s_\ell)\big\} = \int K_h(x - y)\, \mathrm{d}\big\{F_{b_\ell/2}(y) - \bar{F}_{b_\ell/2}(y)\big\}.$$

Hence, under (A2) and (A3), we have

$$\begin{aligned} J_{41} &= \iint K_h(x - y) f(x, s_\ell)\, \mathrm{d}x\, \mathrm{d}\big\{F_{b_\ell/2}(y) - \bar{F}_{b_\ell/2}(y)\big\} \\ &= \iint K(u) f(y + hu, s_\ell)\, \mathrm{d}u\, \mathrm{d}\big\{F_{b_\ell/2}(y) - \bar{F}_{b_\ell/2}(y)\big\} \end{aligned}$$

$$= \iint K(u)\Big\{ f(y, s_\ell) + huf_x(y, s_\ell) + \frac{1}{2}h^2u^2 f_{xx}(y + \theta hu, s_\ell)\Big\} \,\mathrm{d}u \,\mathrm{d}\big\{F_{b_\ell/2}(y) - \bar{F}_{b_\ell/2}(y)\big\}$$
$$= J_{411} + J_{412}, \tag{2.C.61}$$

where $\theta \in [0, 1]$ and we used a first-order Taylor expansion of $f(y-hu, s_\ell)$ around y, and where

$$J_{411} = \int f(y, s_\ell) \,\mathrm{d}\big\{F_{b_\ell/2}(y) - \bar{F}_{b_\ell/2}(y)\big\}, \tag{2.C.62}$$
$$J_{412} = \frac{h^2}{2}\int u^2K(u) \int f_{xx}(y + \theta hu, s_\ell) \,\mathrm{d}\big\{F_{b_\ell/2}(y) - \bar{F}_{b_\ell/2}(y)\big\} \,\mathrm{d}u. \tag{2.C.63}$$

Next we derive a bound for J_{412} and postpone the calculation of J_{411} to §2.C.7.

Bounding J_{412}. Note that, from (2.C.59), we have $F_{b_\ell/2}(-\infty) = \bar{F}_{b_\ell/2}(-\infty) = 0$ and $F_{b_\ell/2}(+\infty) = \bar{F}_{b_\ell/2}(+\infty) = 1$. Hence, using (A3), (B1), (B2), (2.C.60), (2.C.63), and by integration by parts, we have

$$\begin{aligned}
|J_{412}| &= \left|\frac{h^2}{2}\int u^2K(u) \int \big\{F_{b_\ell/2}(y) - \bar{F}_{b_\ell/2}(y)\big\} f_{xxx}(y + \theta hu, s_\ell) \,\mathrm{d}y \,\mathrm{d}u\right| \\
&\le \frac{h^2}{2}\int u^2K(u) \int |F_{b_\ell/2}(y) - \bar{F}_{b_\ell/2}(y)||f_{xxx}(y + \theta hu, s_\ell)| \,\mathrm{d}y \,\mathrm{d}u \\
&= \frac{h^2}{2}\int u^2K(u) \int |F_{b_\ell/2}(x - \theta hu) - \bar{F}_{b_\ell/2}(x - \theta hu)||f_{xxx}(x, s_\ell)| \,\mathrm{d}x \,\mathrm{d}u \\
&\le \frac{h^2}{2}\sup_{y\in\mathbb{R}} |F_{b_\ell/2}(y) - \bar{F}_{b_\ell/2}(y)| \int u^2K(u) \int |f_{xxx}(x, s_\ell)| \,\mathrm{d}x \,\mathrm{d}u \\
&= C_1h^2 \sup_{y\in\mathbb{R}} |F_{b_\ell/2}(y) - \bar{F}_{b_\ell/2}(y)| = O_p(h^2\gamma^{1/2}) = o_p(\Delta t^{4/7}),
\end{aligned} \tag{2.C.64}$$

uniformly in $(\gamma, h) \in I^\ell_{\gamma,h}$, where $C_1 = 2^{-1}\mu_{K,2}\int |f_{xxx}(x, s_\ell)| \,\mathrm{d}x$ with $\mu_{K,2}$ defined above Proposition 2.1.

Conclusion. Combining (2.C.61) and (2.C.64), we have

$$J_{41} = J_{411} + o_p(\Delta t^{4/7}), \tag{2.C.65}$$

uniformly in $(\gamma, h) \in I^{\ell}_{\gamma,h}$, where J_{411} is defined at (2.C.62).

2.C.6.2 Calculating J_{42}

First note that, from (2.C.6),

$$\int \mathrm{E}\big\{\tilde{f}(x, s_\ell)\big\} f(x, s_\ell)\,\mathrm{d}x = \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} \iint K_h(x-y) f(x, s_\ell) f(y, t_{\ell j})\,\mathrm{d}x\,\mathrm{d}y. \tag{2.C.66}$$

Then, note from (2.C.17) that we have

$$\begin{aligned}
&\frac{1}{b_\ell} \sum_{i=2}^{b_\ell} \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j}) U_{ij} \\
&\quad = \frac{1}{b_\ell} \sum_{i=2}^{b_\ell} \sum_{j=1, j\neq i}^{b_\ell/2} \tilde{w}_{\ell j} U_{ij} \\
&\quad = \frac{1}{b_\ell} \sum_{i=1}^{b_\ell} \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} U_{ij} - \frac{1}{b_\ell} \sum_{j=2}^{b_\ell/2} \tilde{w}_{\ell j} U_{jj} - \frac{1}{b_\ell} \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} U_{1j} \\
&\quad = \frac{1}{b_\ell} \sum_{i=1}^{b_\ell} \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} \iint K_h(x-y) f(x, t_{\ell i}) f(y, t_{\ell j})\,\mathrm{d}x\,\mathrm{d}y + o(\Delta t^{4/7}),
\end{aligned} \tag{2.C.67}$$

where we used the definition of U_{ij} in (2.C.15) and the fact that, by (2.C.34), $U_{ij} = \mathrm{E}\{U_j(X_{\ell j})\} \leq M$, so that, using (B3) and (2.1),

$$\begin{aligned}
\frac{1}{b_\ell} \sum_{j=2}^{b_\ell/2} \tilde{w}_{\ell j} U_{jj} + \frac{1}{b_\ell} \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} U_{1j} &\leq \frac{2M}{b_\ell} \gamma \sum_{j=1}^{b_\ell/2} (1-\gamma)^{b_\ell/2-j} \\
&= \frac{2M}{b_\ell} \gamma \sum_{i=0}^{b_\ell/2-1} (1-\gamma)^i \leq \frac{2M}{b_\ell} = o(\Delta t^{4/7}).
\end{aligned}$$

Hence, from (2.C.58), (2.C.66) and (2.C.67), and by the symmetry of K in (A3), we have

$$J_{42} = \iint K_h(x-y)g_1(x)g_2(y)\,\mathrm{d}x\,\mathrm{d}y + o(\Delta t^{4/7}), \tag{2.C.68}$$

where

$$g_1(x) = \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} f(x, t_{\ell j}), \tag{2.C.69}$$

$$g_2(y) = f(y, s_\ell) - \frac{1}{b_\ell}\sum_{i=1}^{b_\ell} f(y, t_{\ell i}). \tag{2.C.70}$$

Next, we calculate $g_1(x)$ and $g_2(y)$.

Calculating $g_1(x)$. Under (A2), by the mean value theorem, we have $f(x, t_{\ell j}) = f(x, s_\ell) + f_t(x, \theta_j)(t_{\ell j} - s_\ell)$ for some θ_j between $t_{\ell j}$ and s_ℓ, and hence, recalling the definition of $\tilde{w}_{\ell j}$ below (2.C.6),

$$\begin{aligned} g_1(x) &= f(x, s_\ell)\gamma \sum_{j=1}^{b_\ell/2} (1-\gamma)^{b_\ell/2-j} - r_1(x) \\ &= f(x, s_\ell)\{1 - (1-\gamma)^{b_\ell/2}\} - r_1(x), \end{aligned} \tag{2.C.71}$$

where

$$r_1(x) = \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j}(s_\ell - t_{\ell j}) f_t(x, \theta_j). \tag{2.C.72}$$

Now, using (A2), (B1) and (2.C.5), we have

$$\begin{aligned} |r_1(x)| &\le M\Delta t\gamma \sum_{j=1}^{b_\ell/2} (b_\ell/2 - j + 1)(1-\gamma)^{b_\ell/2-j} \\ &= M\Delta t\gamma \sum_{i=0}^{b_\ell/2-1} (i+1)(1-\gamma)^i \end{aligned} \tag{2.C.73}$$

$$
\begin{aligned}
&= M\Delta t\gamma \sum_{i=0}^{b_\ell/2-1} (1-\gamma)^i + M\Delta t\gamma \sum_{j=0}^{b_\ell/2-2} \sum_{i=j+1}^{b_\ell/2-1} (1-\gamma)^i \\
&\leq M\Delta t + M\Delta t \sum_{j=0}^{b_\ell/2-2} (1-\gamma)^{j+1} \\
&\leq M\Delta t + M\frac{\Delta t}{\gamma} \leq C_1 \Delta t^{2/7}, \qquad \text{(2.C.74)}
\end{aligned}
$$

where we used summation by parts to get the third line, and where $C_1 = 2M/\delta$.

Calculating $g_2(x)$. Using (A2) and (2.C.5), we have

$$
\begin{aligned}
f(y, t_{\ell i}) &= f(y, s_\ell) + f_t(y, s_\ell)(t_{\ell i} - s_\ell) + \frac{1}{2} f_{tt}(y, s_\ell)(t_{\ell i} - s_\ell)^2 + \frac{1}{6} f_{ttt}(y, \theta_i)(t_{\ell i} - s_\ell)^3 \\
&= f(y, s_\ell) - f_t(y, s_\ell)\Delta t(b_\ell/2 - i) + \frac{1}{2} f_{tt}(y, s_\ell)\Delta t^2 (b_\ell/2 - i)^2 \\
&\qquad - \frac{1}{6} f_{ttt}(y, \theta_i)\Delta t^3 (b_\ell/2 - i)^3 + O(\Delta t), \qquad \text{(2.C.75)}
\end{aligned}
$$

uniformly in y, where θ_i is between $t_{\ell i}$ and s_ℓ and we used a second-order Taylor expansion of $f(y, t_{\ell i})$ around s_ℓ. Hence, using (A2), (B3), (2.C.70) and (2.C.75), we have

$$
\begin{aligned}
g_2(y) &= f_t(y, s_\ell)\Delta t \left(\frac{1}{b_\ell} \sum_{j=-b_\ell/2}^{b_\ell/2-1} j \right) - \frac{\Delta t^2}{2} f_{tt}(y, s_\ell) \left(\frac{1}{b_\ell} \sum_{j=-b_\ell/2}^{b_\ell/2-1} j^2 \right) \\
&\qquad + \frac{1}{6} f_{ttt}(y, \theta_i)\Delta t^3 \left(\frac{1}{b_\ell} \sum_{j=-b_\ell/2}^{b_\ell/2-1} j^3 \right) + O(\Delta t) \\
&= -\frac{\Delta t}{2} f_t(y, s_\ell) - \left(\frac{b_\ell^2 \Delta t^2}{24} + \frac{\Delta t^2}{12} \right) f_{tt}(y, s_\ell) - \frac{1}{48} f_{ttt}(y, \theta_i)\Delta t^3 b_\ell^2 + O(\Delta t), \\
&= -\frac{b_\ell^2 \Delta t^2}{24} f_{tt}(y, s_\ell) + O(\Delta t), \qquad \text{(2.C.76)}
\end{aligned}
$$

uniformly in y, recalling that the remainder term in (2.C.75) is uniform in y, and where we used the fact that

$$\frac{1}{b_\ell}\sum_{j=-b_\ell/2}^{b_\ell/2-1} j^2 = \frac{2}{b_\ell}\sum_{j=1}^{b_\ell/2} j^2 - \frac{b_\ell}{4} = \frac{1}{6}(b_\ell/2+1)(b_\ell+1) - \frac{b_\ell}{4} = \frac{b_\ell^2}{12} + \frac{1}{6}.$$

Conclusion. Now, plugging (2.C.71) and (2.C.76) into (2.C.68) and recalling that the remainder terms in (2.C.71) and (2.C.76) are uniform in x and y, we have

$$\begin{aligned} J_{42} &= -\iint K(u)\big[f(x,s_\ell)\{1-(1-\gamma)^{b_\ell/2}\} - r_1(x)\big] \\ &\qquad\qquad \times \left\{\frac{b_\ell^2\Delta t^2}{24} f_{tt}(x-hu,s_\ell) + O(\Delta t)\right\} \mathrm{d}u\,\mathrm{d}x + o(\Delta t^{4/7}) \\ &= -\frac{b_\ell^2\Delta t^2}{24} J_{421}\{1-(1-\gamma)^{b_\ell/2}\} + \frac{b_\ell^2\Delta t^2}{24} J_{422} + O(\Delta t)(J_{423}-J_{424}) + o(\Delta t^{4/7}), \end{aligned} \tag{2.C.77}$$

since $0<\gamma<1$, and where

$$J_{421} = \iint K(u) f(x,s_\ell) f_{tt}(x-hu,s_\ell)\,\mathrm{d}u\,\mathrm{d}x, \tag{2.C.78}$$

$$J_{422} = \iint K(u) r_1(x) f_{tt}(x-hu,s_\ell)\,\mathrm{d}u\,\mathrm{d}x, \tag{2.C.79}$$

$$J_{423} = \iint K(u) f(x,s_\ell)\,\mathrm{d}u\,\mathrm{d}x, \tag{2.C.80}$$

$$J_{424} = \iint K(u) r_1(x)\,\mathrm{d}u\,\mathrm{d}x. \tag{2.C.81}$$

We calculate $J_{421},\ldots,J_{424}$ as follows. Under (A2), (A3) and (B2), we have

$$\begin{aligned} J_{421} &= \iint K(u) f(x,s_\ell)\big\{f_{tt}(x,s_\ell) - huf_{xtt}(x,s_\ell) + h^2u^2 f_{xxtt}(x-\theta hu,s_\ell)\big\}\,\mathrm{d}u\,\mathrm{d}x \\ &= \int f_{tt}(x,s_\ell) f(x,s_\ell)\,\mathrm{d}x + h^2 \iint u^2 K(u) f(x,s_\ell) f_{xxtt}(x-\theta hu,s_\ell)\,\mathrm{d}u\,\mathrm{d}x \\ &= \int f_{tt}(x,s_\ell) f(x,s_\ell)\,\mathrm{d}x + O\left\{h^2 \iint u^2 K(u) f(x,s_\ell)\,\mathrm{d}u\,\mathrm{d}x\right\} \end{aligned}$$

$$= \int f_{tt}(x, s_\ell) f(x, s_\ell) \, \mathrm{d}x + O(\Delta t^{2/7}), \tag{2.C.82}$$

where $\theta \in [0, 1]$ and we used a first-order Taylor expansion of $f_{tt}(x - hu, s_\ell)$ around x.

Using (A2), (A3), (B2) and (2.C.72), we have

$$\begin{aligned} |J_{422}| &\leq M \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j}(s_\ell - t_{\ell j}) \int |f_t(x, \theta_j)| \, \mathrm{d}x \\ &\leq M \max_j \left\{ \int |f_t(x, \theta_j)| \, \mathrm{d}x \right\} \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j}(s_\ell - t_{\ell j}) \\ &= M^2 \Delta t \gamma \sum_{j=1}^{b_\ell/2} (b_\ell/2 - j + 1)(1 - \gamma)^{b_\ell/2 - j} = O(\Delta t^{2/7}), \end{aligned} \tag{2.C.83}$$

where we used (2.C.73) and (2.C.74) to get the last line.

Finally, under (A3), we have

$$J_{423} = 1, \tag{2.C.84}$$

and, from (2.C.72),

$$\begin{aligned} |J_{424}| &\leq \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j}(s_\ell - t_{\ell j}) \int |f_t(x, \theta_j)| \, \mathrm{d}x \\ &\leq M \Delta t \gamma \sum_{j=1}^{b_\ell/2} (b_\ell/2 - j + 1)(1 - \gamma)^{b_\ell/2 - j} = O(\Delta t^{2/7}), \end{aligned} \tag{2.C.85}$$

where we used (2.C.73) and (2.C.74) to get the last line.

Note that, from (2.C.8),

$$(1 - \gamma)^{b_\ell/2} \leq (1 - \gamma_m)^{b_\ell/2} \leq \exp(-3^{-1} \delta^2 \Delta t^{-\alpha/7}) = o(\Delta t). \tag{2.C.86}$$

Hence, under (B3), plugging (2.C.82)–(2.C.86) into (2.C.77), we have

$$J_{42} = -\frac{b_\ell^2 \Delta t^2}{24} \int f_{tt}(x, s_\ell) f(x, s_\ell) \, \mathrm{d}x + o(\Delta t^{4/7}), \tag{2.C.87}$$

uniformly in $(\gamma, h) \in I_{\gamma,h}^{\ell}$.

2.C.6.3 Conclusion

Now, combining (2.C.57), (2.C.58), (2.C.61), (2.C.64) and (2.C.87), we have

$$J_4 = J_{411} - \frac{b_\ell^2 \Delta t^2}{24} \int f_{tt}(x, s_\ell) f(x, s_\ell) \, \mathrm{d}x + o_p(\Delta t^{4/7}) \tag{2.C.88}$$

uniformly in $(\gamma, h) \in I_{\gamma,h}^{\ell}$, where J_{411} is defined at (2.C.62).

2.C.7 Remaining terms

Combining (2.C.18), (2.C.42), (2.C.50), (2.C.56) and (2.C.88), we have

$$J = J_{411} - (J_{22} + J_{32}) - \frac{b_\ell^2 \Delta t^2}{24} \int f_{tt}(x, s_\ell) f(x, s_\ell) \, \mathrm{d}x + o_p(\Delta t^{4/7}), \tag{2.C.89}$$

uniformly in $(\gamma, h) \in I_{\gamma,h}^{\ell}$.

From (2.C.59) and (2.C.62), we have

$$J_{411} = \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} f(X_{\ell j}, s_\ell) - \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} \int f(x, s_\ell) f(x, t_{\ell j}) \, \mathrm{d}x. \tag{2.C.90}$$

From (2.C.45) and (2.C.52), we have

$$J_{22} + J_{32} = \frac{1}{b_\ell} \sum_{i=2}^{b_\ell} \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j}) \big[f(X_{\ell i}, t_{\ell j}) + f(X_{\ell j}, t_{\ell i}) - \mathrm{E}\{f(X_{\ell i}, t_{\ell j})\} - \mathrm{E}\{f(X_{\ell j}, t_{\ell i})\} \big] \tag{2.C.91}$$

Next, we reorganise terms in (2.C.91).

First note that $\mathrm{E}\{f(X_{\ell j}, t_{\ell i})\} = \int f(x, t_{\ell i}) f(x, t_{\ell j})\, \mathrm{d}x = \mathrm{E}\{f(X_{\ell i}, t_{\ell j})\}$. Then, similar to (2.C.17), we have

$$\begin{aligned}
&\frac{1}{b_\ell} \sum_{i=2}^{b_\ell} \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j}) \big[\mathrm{E}\{f(X_{\ell i}, t_{\ell j})\} + \mathrm{E}\{f(X_{\ell j}, t_{\ell i})\}\big] \\
&\qquad = \frac{2}{b_\ell} \sum_{i=1}^{b_\ell} \sum_{j=1, j\neq i}^{b_\ell/2} \tilde{w}_{\ell j}\, \mathrm{E}\{f(X_{\ell i}, t_{\ell j})\} \\
&\qquad = \frac{2}{b_\ell} \sum_{i=1}^{b_\ell} \sum_{j=1}^{\ell/2} \tilde{w}_{\ell j}\, \mathrm{E}\{f(X_{\ell i}, t_{\ell j})\} - \frac{2}{b_\ell} \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j}\, \mathrm{E}\{f(X_{\ell j}, t_{\ell j})\} \\
&\qquad = \frac{2}{b_\ell} \sum_{i=1}^{b_\ell} \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} \int f(x, t_{\ell i}) f(x, t_{\ell j})\, \mathrm{d}x + o(\Delta t^{4/7}), \qquad (2.C.92)
\end{aligned}$$

where the last equation follows from the fact that, under (A2) and (B3), we have

$$\frac{1}{b_\ell} \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j}\, \mathrm{E}\{f(X_{\ell j}, t_{\ell j})\} \leq M \frac{\gamma}{b_\ell} \sum_{j=1}^{b_\ell/2} (1-\gamma)^{b_\ell/2-j} = M \frac{\gamma}{b_\ell} \sum_{i=0}^{b_\ell/2-1} (1-\gamma)^i \leq \frac{M}{b_\ell} = o_{a.s.}(\Delta t^{4/7}).$$

Then, note that $f(X_{\ell i}, t_{\ell j}) + f(X_{\ell j}, t_{\ell i})$ is symmetric in i and j, so that, similar to (2.C.92), we have

$$\begin{aligned}
&\frac{1}{b_\ell} \sum_{i=2}^{b_\ell} \sum_{j=1}^{(i-1)\wedge(b_\ell/2)} (\tilde{w}_{\ell i} + \tilde{w}_{\ell j}) \big\{f(X_{\ell i}, t_{\ell j}) + f(X_{\ell j}, t_{\ell i})\big\} \\
&\qquad = \frac{1}{b_\ell} \sum_{i=1}^{b_\ell} \sum_{j=1, j\neq i}^{b_\ell/2} \tilde{w}_{\ell j} \big\{f(X_{\ell i}, t_{\ell j}) + f(X_{\ell j}, t_{\ell i})\big\} \\
&\qquad = \frac{1}{b_\ell} \sum_{i=1}^{b_\ell} \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} \big\{f(X_{\ell i}, t_{\ell j}) + f(X_{\ell j}, t_{\ell i})\big\} - \frac{2}{b_\ell} \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} f(X_{\ell j}, t_{\ell j}) \\
&\qquad = \frac{1}{b_\ell} \sum_{i=1}^{b_\ell} \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} \big\{f(X_{\ell i}, t_{\ell j}) + f(X_{\ell j}, t_{\ell i})\big\} + o_{a.s.}(\Delta t^{4/7}), \qquad (2.C.93)
\end{aligned}$$

where the last equation follows from the fact that, under (A2) and (B3), we have

$$\frac{1}{b_\ell}\sum_{j=1}^{b_\ell/2}\tilde{w}_{\ell j}f(X_{\ell j},t_{\ell j}) \le M\frac{\gamma}{b_\ell}\sum_{j=1}^{b_\ell/2}(1-\gamma)^{b_\ell/2-j} = M\frac{\gamma}{b_\ell}\sum_{i=0}^{b_\ell/2-1}(1-\gamma)^{i} \le \frac{M}{b_\ell} = o_{a.s.}(\Delta t^{4/7}).$$

Plugging (2.C.92) and (2.C.93) into (2.C.91), we have

$$\begin{aligned} J_{22}+J_{32} = \frac{2}{b_\ell}\sum_{i=1}^{b_\ell}\sum_{j=1}^{b_\ell/2}\tilde{w}_{\ell j}\int f(x,t_{\ell i})f(x,t_{\ell j})\,\mathrm{d}x \\ -\frac{1}{b_\ell}\sum_{i=1}^{b_\ell}\sum_{j=1}^{b_\ell/2}\tilde{w}_{\ell j}\big\{f(X_{\ell i},t_{\ell j})+f(X_{\ell j},t_{\ell i})\big\}+o_{a.s.}(\Delta t^{4/7}). \end{aligned} \tag{2.C.94}$$

Then, combining (2.C.89), (2.C.90) and (2.C.94), we have

$$J = S_1 - S_2 + 2S_3 - S_4 - S_5 - \frac{b_\ell^2\Delta t^2}{24}\int f_{tt}(x,s_\ell)f(x,s_\ell)\,\mathrm{d}x + o_p(\Delta t^{4/7}), \tag{2.C.95}$$

uniformly in $(\gamma,h)\in I^\ell_{\gamma,h}$, where

$$S_1 = \sum_{j=1}^{b_\ell/2}\tilde{w}_{\ell j}f(X_{\ell j},s_\ell), \tag{2.C.96}$$

$$S_2 = \sum_{j=1}^{b_\ell/2}\tilde{w}_{\ell j}\int f(x,s_\ell)f(x,t_{\ell j})\,\mathrm{d}x, \tag{2.C.97}$$

$$S_3 = \frac{1}{b_\ell}\sum_{i=1}^{b_\ell}\sum_{j=1}^{b_\ell/2}\tilde{w}_{\ell j}\int f(x,t_{\ell i})f(x,t_{\ell j})\,\mathrm{d}x, \tag{2.C.98}$$

$$S_4 = \frac{1}{b_\ell}\sum_{i=1}^{b_\ell}\sum_{j=1}^{b_\ell/2}\tilde{w}_{\ell j}f(X_{\ell i},t_{\ell j}), \tag{2.C.99}$$

$$S_5 = \frac{1}{b_\ell}\sum_{i=1}^{b_\ell}\sum_{j=1}^{b_\ell/2}\tilde{w}_{\ell j}f(X_{\ell j},t_{\ell i}). \tag{2.C.100}$$

Next, we calculate $S_1,\ldots,S_5$ as follows.

Calculating S_5. First note from (2.C.70) and (2.C.76) that

$$\frac{1}{b_\ell}\sum_{i=1}^{b_\ell} f(x, t_{\ell i}) = f(x, s_\ell) + \frac{b_\ell^2 \Delta t^2}{24} f_{tt}(x, s_\ell) + o(\Delta t^{4/7}), \tag{2.C.101}$$

uniformly in x. Plugging (2.C.101) into (2.C.100) and recalling the definition of S_1 at (2.C.96), we have

$$\begin{aligned} S_5 &= \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} \left\{ \frac{1}{b_\ell} \sum_{i=1}^{b_\ell} f(X_{\ell j}, t_{\ell i}) \right\} \\ &= \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} \left\{ f(X_{\ell j}, s_\ell) + \frac{b_\ell^2 \Delta t^2}{24} f_{tt}(X_{\ell j}, s_\ell) + o_p(\Delta t^{4/7}) \right\} \\ &= S_1 + \frac{b_\ell^2 \Delta t^2}{24} \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} f_{tt}(X_{\ell j}, s_\ell) + o_p(\Delta t^{4/7}). \end{aligned} \tag{2.C.102}$$

Then, using (B2) and (2.C.59), we have

$$\begin{aligned} &\left| \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} f_{tt}(X_{\ell j}, s_\ell) - \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} \int f_{tt}(x, s_\ell) f(x, t_{\ell j}) \, \mathrm{d}x \right| \\ &\quad = \left| \int f_{tt}(x, s_\ell) \, \mathrm{d}\big\{ F_{b_\ell/2}(x) - \bar{F}_{b_\ell/2}(x) \big\} \right| = \left| \int \big\{ F_{b_\ell/2}(x) - \bar{F}_{b_\ell/2}(x) \big\} f_{xtt}(x, s_\ell) \, \mathrm{d}x \right| \\ &\quad \le \sup_{x \in \mathbb{R}} \big| F_{b_\ell/2}(x) - \bar{F}_{b_\ell/2}(x) \big| \int |f_{xtt}(x, s_\ell)| \, \mathrm{d}x = O_p(\gamma^{1/2}), \end{aligned}$$

where we used (2.C.60) to get the last equation, and hence, under (B1) and (B3), we have

$$\frac{b_\ell^2 \Delta t^2}{24} \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} f_{tt}(X_{\ell j}, s_\ell) = \frac{b_\ell^2 \Delta t^2}{24} \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} \int f_{tt}(x, s_\ell) f(x, t_{\ell j}) \, \mathrm{d}x + o_p(\Delta t^{4/7}), \tag{2.C.103}$$

uniformly in $(\gamma, h) \in I^\ell_{\gamma,h}$.

Now, combining (2.C.102) and (2.C.103), we have

$$S_5 = S_1 + \frac{b_\ell^2 \Delta t^2}{24} \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} \int f_{tt}(x, s_\ell) f(x, t_{\ell j}) \, \mathrm{d}x + o_p(\Delta t^{4/7}), \tag{2.C.104}$$

uniformly in $(\gamma, h) \in I^\ell_{\gamma,h}$.

Calculating $S_3 - S_2$. Using (B1), (B3), (2.C.97) and (2.C.98), we have

$$\begin{aligned} S_3 - S_2 &= \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} \int \left\{ \frac{1}{b_\ell} \sum_{i=1}^{b_\ell} f(x, t_{\ell i}) - f(x, s_\ell) \right\} f(x, t_{\ell j}) \, \mathrm{d}x \\ &= \frac{b_\ell^2 \Delta t^2}{24} \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} \int f_{tt}(x, s_\ell) f(x, t_{\ell j}) \, \mathrm{d}x + o(\Delta t^{4/7}) \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} \int f(x, t_{\ell j}) \, \mathrm{d}x \\ &= \frac{b_\ell^2 \Delta t^2}{24} \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} \int f_{tt}(x, s_\ell) f(x, t_{\ell j}) \, \mathrm{d}x + o(\Delta t^{4/7}), \end{aligned} \tag{2.C.105}$$

where the second equation follows from (2.C.101) and the last equation from the fact that

$$\sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} \int f(x, t_{\ell j}) \, \mathrm{d}x = \gamma \sum_{j=1}^{b_\ell/2} (1-\gamma)^{b_\ell/2 - j} = \gamma \sum_{i=0}^{b_\ell/2 - 1} (1-\gamma)^i \leq 1,$$

since $0 < \gamma < 1$ and f is a density.

Calculating $S_3 - S_4$. Let

$$F_{b_\ell}(x) = \frac{1}{b_\ell} \sum_{i=1}^{b_\ell} \mathbb{1}\{X_{\ell i} \leq x\}, \quad \bar{F}_{b_\ell}(x) = \frac{1}{b_\ell} \sum_{i=1}^{b_\ell} F(x, t_{\ell i}),$$

where $F(\cdot, t_{\ell i})$ is defined above (2.C.59). Then, using (2.C.4), we have

$$\sup_{x \in \mathbb{R}} |F_{b_\ell}(x) - \bar{F}_{b_\ell}(x)| = O_p\left(\sum_{i=1}^{b_\ell} \frac{1}{b_\ell^2} \right)^{1/2} = O_p(b_\ell^{-1/2}). \tag{2.C.106}$$

Using (A2) and applying the mean value theorem to $f_x(x, t_{\ell j})$, we have $f_x(x, t_{\ell j}) = f_x(x, s_\ell) + f_{xt}(x, \theta_j)(t_{\ell j} - s_\ell)$ with θ_j between $t_{\ell j}$ and s_ℓ, so that

$$\sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} f_x(x, t_{\ell j}) = f_x(x, s_\ell) - r_3(x) - (1-\gamma)^{b_\ell/2} f_x(x, s_\ell), \tag{2.C.107}$$

uniformly in x, where $r_3(x) = \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} f_{xt}(x, \theta_j)(s_\ell - t_{\ell j})$. Combining (B1), (B2) and (2.C.5), we have

$$\begin{aligned}\int |r_3(x)|\,\mathrm{d}x &\le \Delta t\gamma \sum_{j=1}^{b_\ell/2} (b_\ell/2 - j + 1)(1-\gamma)^{b_\ell/2-j} \int |f_{xt}(x, \theta_j)|\,\mathrm{d}x \\ &= O(\Delta t)\gamma \sum_{j=1}^{b_\ell/2} (b_\ell/2 - j + 1)(1-\gamma)^{b_\ell/2-j} = O(\Delta t^{2/7}),\end{aligned} \tag{2.C.108}$$

where the last equation follows from (2.C.74).

Now, using (2.C.98), (2.C.99) and (2.C.107), we have

$$\begin{aligned}S_3 - S_4 &= -\sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} \int f(x, t_{\ell j})\,\mathrm{d}\big\{F_{b_\ell}(x) - \bar{F}_{b_\ell}(x)\big\} \\ &= \int \big\{F_{b_\ell}(x) - \bar{F}_{b_\ell}(x)\big\} \left\{\sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} f_x(x, t_{\ell j})\right\} \mathrm{d}x \\ &= \int \big\{F_{b_\ell}(x) - \bar{F}_{b_\ell}(x)\big\} f_x(x, s_\ell)\,\mathrm{d}x \\ &\qquad - \int \big\{F_{b_\ell}(x) - \bar{F}_{b_\ell}(x)\big\}\big\{r_3(x) + (1-\gamma)^{b_\ell/2} f_x(x, s_\ell)\big\}\,\mathrm{d}x.\end{aligned}$$

Using (B2), (2.C.86), (2.C.106) and (2.C.108), we have

$$\begin{aligned}&\left|\int \big\{F_{b_\ell}(x) - \bar{F}_{b_\ell}(x)\big\}\big\{r_3(x) + (1-\gamma)^{b_\ell/2} f_x(x, s_\ell)\big\}\,\mathrm{d}x\right| \\ &\le \sup_{x\in\mathbb{R}} |F_{b_\ell}(x) - \bar{F}_{b_\ell}(x)| \left\{\int |r_3(x)|\,\mathrm{d}x + (1-\gamma)^{b_\ell/2} \int |f_x(x, s_\ell)|\,\mathrm{d}x\right\} \\ &= O_p\big(\Delta t^{2/7} b_\ell^{-1/2}\big) = o_p(\Delta t^{4/7}),\end{aligned}$$

where, to get the last equality, we used the fact that, from (B3) and (2.1), we have $b_\ell \asymp \tau_\ell/\Delta t \asymp \Delta t^{-(5+\alpha)/7}$, so that $b_\ell^{-1/2} \asymp \Delta t^{(5+\alpha)/14} = o_p(\Delta t^{2/7})$. Hence

$$\begin{aligned}
S_3 - S_4 &= \frac{1}{b_\ell}\sum_{i=1}^{b_\ell}\int f(x,s_\ell)f(x,t_{\ell i})\,\mathrm{d}x - \frac{1}{b_\ell}\sum_{i=1}^{b_\ell} f(X_{\ell i},s_\ell) + o_p(\Delta t^{4/7}) \\
&= \int f(x,s_\ell)\left\{\frac{1}{b_\ell}\sum_{i=1}^{b_\ell} f(x,t_{\ell i})\right\}\mathrm{d}x - \frac{1}{b_\ell}\sum_{i=1}^{b_\ell} f(X_{\ell i},s_\ell) + o_p(\Delta t^{4/7}) \\
&= \int f(x,s_\ell)\left\{f(x,s_\ell) + \frac{b_\ell^2\Delta t^2}{24} f_{tt}(x,s_\ell) + o(\Delta t^{4/7})\right\}\mathrm{d}x \\
&\qquad\qquad - \frac{1}{b_\ell}\sum_{i=1}^{b_\ell} f(X_{\ell i},s_\ell) + o_p(\Delta t^{4/7}) \\
&= \int f^2(x,s_\ell)\,\mathrm{d}x + \frac{b_\ell^2\Delta t^2}{24}\int f(x,s_\ell)f_{tt}(x,s_\ell)\,\mathrm{d}x \\
&\qquad\qquad - \frac{1}{b_\ell}\sum_{i=1}^{b_\ell} f(X_{\ell i},s_\ell) + o_p(\Delta t^{4/7}), \quad \text{(2.C.109)}
\end{aligned}$$

uniformly in $(\gamma,h) \in I^\ell_{\gamma,h}$, where we used (2.C.101) to get third equation.

Conclusion. Combining (2.C.95), (2.C.104), (2.C.105) and (2.C.109), we have

$$J = \int f^2(x,s_\ell)\,\mathrm{d}x - \frac{1}{b_\ell}\sum_{i=1}^{b_\ell} f(X_{\ell i},s_\ell) + o_p(\Delta t^{4/7}), \quad \text{(2.C.110)}$$

uniformly in $(\gamma,h) \in I^\ell_{\gamma,h}$. Combining (2.C.14) and (2.C.110), we have

$$\mathrm{SCV}_\ell(\gamma,h) - \mathrm{ISE}_\ell(\gamma,h) = \int f^2(x,s_\ell)\,\mathrm{d}x - \frac{2}{b_\ell}\sum_{i=1}^{b_\ell} f(X_{\ell i},s_\ell) + o_p(\Delta t^{4/7}), \quad \text{(2.C.111)}$$

uniformly in $(\gamma,h) \in I^\ell_{\gamma,h}$, which proves Theorem 2.1.

2.D Proof of Proposition 2.2

The proof of Proposition 2.2 is very similar to the proof of Theorem 2.1. First note from (2.22) that

$$\begin{aligned}\mathrm{ISE}_\ell(\gamma,h)-\mathrm{MISE}_\ell(\gamma,h) &= \int \big[\check{f}^2(x,s_\ell)-\mathrm{E}\{\check{f}^2(x,s_\ell)\}\big]\,\mathrm{d}x \\ &\qquad -2\int \big[\check{f}(x,s_\ell)-\mathrm{E}\{\check{f}(x,s_\ell)\}\big]f(x,s_\ell)\,\mathrm{d}x \\ &= I_0-2J_{41}-2\int \big[\check{f}(x,s_\ell)-\mathrm{E}\{\check{f}(x,s_\ell)\} \\ &\qquad -\tilde{f}(x,s_\ell)+\mathrm{E}\{\tilde{f}(x,s_\ell)\}\big]f(x,s_\ell)\,\mathrm{d}x \\ &= I_0-2J_{41}+o_p(\Delta t), \end{aligned} \tag{2.D.1}$$

where J_{41} is defined at (2.C.57) and $I_0=\int \big[\check{f}^2(x,s_\ell)-\mathrm{E}\{\check{f}^2(x,s_\ell)\}\big]\,\mathrm{d}x$, and where we used (2.C.7) to get the last equation.

Plugging (2.C.64) and (2.C.90) into (2.C.61), we have

$$J_{41}=\sum_{i=1}^{b_\ell/2}\tilde{w}_{\ell i}f(X_{\ell i},s_\ell)-\sum_{i=1}^{b_\ell/2}\tilde{w}_{\ell i}\int f(x,s_\ell)f(x,t_{\ell i})\,\mathrm{d}x+o_p(\Delta t^{4/7}), \tag{2.D.2}$$

uniformly in $(\gamma,h)\in I^\ell_{\gamma,h}$, where $\tilde{w}_{\ell j}$ is defined under (2.C.6). Next, we show that

$$I_0=I+o(\Delta t), \tag{2.D.3}$$

where

$$I=\int \big[\tilde{f}^2(x,s_\ell)-\mathrm{E}\big\{\tilde{f}^2(x,s_\ell)\big\}\big]\,\mathrm{d}x, \tag{2.D.4}$$

so that

$$\mathrm{ISE}_\ell(\gamma,h)-\mathrm{MISE}_\ell(\gamma,h)=I-2J_{41}+o_p(\Delta t^{4/7}), \tag{2.D.5}$$

uniformly in $(\gamma,h)\in I^\ell_{\gamma,h}$.

To show (2.D.3), recalling the definition of γ_j at (2.18) and the definitions of γ_m and γ_M in

(A1) and using (A3), (B1), (2.C.6) and (2.C.7), we have

$$\begin{aligned}\int \{\check{f}(x,s_\ell)+\tilde{f}(x,s_\ell)\}\,\mathrm{d}x &= 2\int \tilde{f}(x,s_\ell)\,\mathrm{d}x + \int \{\check{f}(x,s_\ell)-\tilde{f}(x,s_\ell)\}\,\mathrm{d}x \\ &= 2\gamma\sum_{j=1}^{b_\ell/2}(1-\gamma)^{b_\ell/2-j} \\ &\qquad + \prod_{j=2}^{n_{s_\ell}}(1-\gamma_j) + \sum_{i=2}^{n_{s_\ell}-b_\ell/2}\gamma_i\prod_{j=i+1}^{n_{s_\ell}}(1-\gamma_j) \\ &\le 2 + \prod_{j=n_{s_\ell}-b_\ell/2+1}^{n_{s_\ell}}(1-\gamma_j) + \sum_{i=2}^{n_{s_\ell}-b_\ell/2}\gamma_i\prod_{j=n_{s_\ell}-b_\ell/2+1}^{n_{s_\ell}}(1-\gamma_j) \\ &\le 2 + (n_{s_\ell}\gamma_M+1)(1-\gamma_m)^{b_\ell/2} \\ &\le 2 + \Big(\frac{s_\ell}{\delta}\Delta t^{-2/7}+1\Big)\exp(-3^{-1}\delta^2\Delta t^{-\alpha/7}), \qquad (2.D.6)\end{aligned}$$

where we used the fact that $\gamma_j \in (0,1)$ to get the third line, and where, to get the last inequality, we used (1.2) and (2.C.8).

Hence, using (B1), (B3), (2.C.7) and (2.D.6), we have

$$\begin{aligned}\int |\check{f}^2(x,s_\ell)-\tilde{f}^2(x,s_\ell)|\,\mathrm{d}x &= \int |\check{f}(x,s_\ell)+\tilde{f}(x,s_\ell)||\check{f}(x,s_\ell)-\tilde{f}(x,s_\ell)|\,\mathrm{d}x \\ &\le 2\exp(-3^{-1}\delta^2\Delta t^{-\alpha/7})\int |\check{f}(x,s_\ell)+\tilde{f}(x,s_\ell)|\,\mathrm{d}x = o_p(\Delta t),\end{aligned}$$

and, similarly, $\int |\,\mathrm{E}\{\check{f}^2(x,s_\ell)\}-\mathrm{E}\{\tilde{f}^2(x,s_\ell)\}|\,\mathrm{d}x = o(\Delta t)$, which proves (2.D.3) and hence also (2.D.5). Next, we calculate I.

2.D.1 Calculating I

First note from (2.C.6) that

$$\int \tilde{f}^2(x,s_\ell)\,\mathrm{d}x = \sum_{j=1}^{b_\ell/2}\tilde{w}_{\ell j}^2\int K_h^2(x-X_{\ell j})\,\mathrm{d}x$$

$$+ 2\sum_{i=2}^{b_\ell/2}\sum_{j=1}^{i-1}\tilde{w}_{\ell i}\tilde{w}_{\ell j}\int K_h(x-X_{\ell i})K_h(x-X_{\ell j})\,\mathrm{d}x. \quad (2.D.7)$$

Then, for $i, j \in \{1, \ldots, b_\ell/2\}$, let

$$\begin{gathered} U(x,y) = \int K_h(z-x)K_h(z-y)\,\mathrm{d}z,\ U_i(x) = \mathrm{E}\{U(x, X_{\ell i})\}, \\ U_{ij} = \mathrm{E}\{U_i(X_{\ell j})\},\ V_{ij} = U(X_{\ell i}, X_{\ell j}) - U_i(X_{\ell j}) - U_j(X_{\ell i}) + U_{ij}. \end{gathered} \quad (2.D.8)$$

Note that, with the new definitions in (2.D.8), (2.C.16) still holds.

Now, combining (2.D.4), (2.D.7) and (2.D.8), we have $I = I_1 + 2I_2$, where

$$I_1 = \sum_{i=1}^{b_\ell/2}\tilde{w}_{\ell i}^2\{U(X_{\ell i}, X_{\ell i}) - U_{ii}\}, \quad (2.D.9)$$

$$I_2 = \sum_{i=2}^{b_\ell/2}\sum_{j=1}^{i-1}\tilde{w}_{\ell i}\tilde{w}_{\ell j}\{U(X_{\ell i}, X_{\ell j}) - U_{ij}\}. \quad (2.D.10)$$

Note that $I_2 = I_{21} + I_{22} + I_{23}$, where

$$I_{21} = \sum_{i=2}^{b_\ell/2}\sum_{j=1}^{i-1}\tilde{w}_{\ell i}\tilde{w}_{\ell j}V_{ij}, \quad (2.D.11)$$

$$I_{22} = \sum_{i=2}^{b_\ell/2}\sum_{j=1}^{i-1}\tilde{w}_{\ell i}\tilde{w}_{\ell j}\{U_i(X_{\ell j}) - U_{ij}\}, \quad (2.D.12)$$

$$I_{23} = \sum_{i=2}^{b_\ell/2}\sum_{j=1}^{i-1}\tilde{w}_{\ell i}\tilde{w}_{\ell j}\{U_j(X_{\ell i}) - U_{ij}\}. \quad (2.D.13)$$

Next, we calculate I_1, I_{21}, I_{22} and I_{23}.

2.D.1.1 Calculating I_1

Note from (2.D.9) that I_1 is sum of independent centred random variables. To apply Theorem 2.C.2 to I_1, we need bounds for $\tilde{w}_{\ell i}^2 U(X_{\ell i}, X_{\ell i})$ and $\sum_{i=1}^{b_\ell/2}\tilde{w}_{\ell i}^4\,\mathrm{E}\{U^2(X_{\ell i}, X_{\ell i})\}$. Recalling the

definition of $\tilde{w}_{\ell i}$ below (2.C.6) and using (A3) and (B1), we have

$$\begin{aligned}\tilde{w}_{\ell i}^2 U(X_{\ell i}, X_{\ell i}) &\le \frac{\gamma^2}{h^2}\int K^2\Big(\frac{x - X_{\ell i}}{h}\Big)\,\mathrm{d}x \\ &\le M\frac{\gamma^2}{h^2}\int K\Big(\frac{x - X_{\ell i}}{h}\Big)\,\mathrm{d}x \\ &= M\frac{\gamma^2}{h} \le C_1\Delta t^{9/7}. \end{aligned} \tag{2.D.14}$$

where $C_1 = M/\delta^3$.

Then, under (A3) and (B1), we have

$$\begin{aligned}\sum_{i=1}^{b_\ell/2} \tilde{w}_{\ell i}^4\,\mathrm{E}\{U^2(X_{\ell i}, X_{\ell i})\} &= \frac{\gamma^4}{h^4}\sum_{i=1}^{b_\ell/2}(1-\gamma)^{4(b_\ell/2-i)}\,\mathrm{E}\left\{\int K^2\Big(\frac{x - X_{\ell i}}{h}\Big)\,\mathrm{d}x\right\}^2 \\ &\le M^2\frac{\gamma^4}{h^2}\sum_{i=1}^{b_\ell/2}(1-\gamma)^{b_\ell/2-i}\,\mathrm{E}\left\{\int \frac{1}{h}K\Big(\frac{x - X_{\ell i}}{h}\Big)\,\mathrm{d}x\right\}^2 \\ &= M^2\frac{\gamma^4}{h^2}\sum_{j=0}^{b_\ell/2-1}(1-\gamma)^j \le M^2\frac{\gamma^3}{h^2} \le C_2\Delta t^{13/7}, \end{aligned} \tag{2.D.15}$$

where $C_2 = M^2/\delta^5$ and we used the fact that $0 < \gamma < 1$.

Now, under (B1), applying Theorem 2.C.2 and using (2.D.14) and (2.D.15), we have, for any $\eta > 0$,

$$\begin{aligned}\mathrm{P}\left\{\sup_{(\gamma,h)\in I^\ell_{\gamma,h}} |I_1| \ge \eta\Delta t^{4/7}\right\} &\le \#(I^\ell_{\gamma,h}) \sup_{(\gamma,h)\in I^\ell_{\gamma,h}} P\big(|I_1| \ge \eta\Delta t^{4/7}\big) \\ &\le C\Delta t^{-a}\exp\left\{-\frac{\eta^2\Delta t^{8/7}}{2(C_2\Delta t^{13/7} + C_1\eta\Delta t^{13/7}/3)}\right\} = o(1),\end{aligned}$$

which implies that

$$I_1 = o_p(\Delta t^{4/7}), \tag{2.D.16}$$

uniformly in $(\gamma, h) \in I^\ell_{\gamma,h}$.

2.D.1.2 Calculating I_{21}

Let $Y_i = \sum_{j=1}^{i-1} \tilde{w}_{\ell i}\tilde{w}_{\ell j}V_{ij}$ for $i = 2, \ldots, b_\ell/2$. Then, from (2.D.11), we have $I_{21} = \sum_{i=2}^{b_\ell/2} Y_i$. Next, we show that $\{\sum_{i=2}^{k} Y_i\}_{k=2,\ldots,b_\ell/2}$ is a martingale. Similar to §2.C.3, we only need to show that (2.C.24) is true.

Firstly, by (2.D.8), Y_i is a function of $X_{\ell 1}, \ldots, X_{\ell i}$, and hence measurable with respect to $\mathcal{F}_i$ defined above (2.C.24). Secondly, since, under (A3), we have

$$U(x,y) = \frac{1}{h^2}\int K\Big(\frac{z-x}{h}\Big)K\Big(\frac{z-y}{h}\Big)\,\mathrm{d}z \le \frac{M}{h}\int K(u)\,\mathrm{d}u = \frac{M}{h}, \tag{2.D.17}$$

so that $|V_{ij}| \le 4M/h$. Finally, similar to (2.C.26), we have $\mathrm{E}(Y_i|\mathcal{F}_{i-1}) = 0$.

Now, we can apply Theorem 2.C.1 to I_{21}. Similar to (2.C.29), we have, for any $p \ge 2$,

$$\mathrm{E}\,|I_{21}|^p \le 2C_p\left\{\sum_{i=2}^{b_\ell/2}(\mathrm{E}\,|Y_i|^p)^{2/p}\right\}^{p/2}. \tag{2.D.18}$$

Note that, conditioning on $X_{\ell i}$, Y_i is a sum of zero-mean independent random variables. Hence, similar to (2.C.31), we use (2.C.2) to get

$$\mathrm{E}\,|Y_i|^p \le C_p\left[\mathrm{E}\left\{\sum_{j=1}^{i-1}\tilde{w}_{\ell i}^2\tilde{w}_{\ell j}^2\,\mathrm{E}(V_{ij}^2|X_{\ell i})\right\}^{p/2} + \sum_{j=1}^{i-1}\tilde{w}_{\ell i}^p\tilde{w}_{\ell j}^p\,\mathrm{E}\,|V_{ij}|^p\right]. \tag{2.D.19}$$

Now, we need bounds for $\mathrm{E}(V_{ij}^2|X_{\ell i})$ and $\mathrm{E}\,|V_{ij}|^p$.

Bounding $\mathrm{E}(V_{ij}^2|X_{\ell i})$. Note that (2.C.32) still holds for the new definition of U at (2.D.8). Then, note that, under (A2) and (A3), we have, uniformly in x,

$$\begin{aligned} U_i(x) &= \frac{1}{h^2}\iint K\Big(\frac{z-x}{h}\Big)K\Big(\frac{z-y}{h}\Big)f(y,t_{\ell i})\,\mathrm{d}z\,\mathrm{d}y \\ &= \frac{1}{h}\iint K\Big(\frac{z-x}{h}\Big)K(u)f(z-hu,t_{\ell i})\,\mathrm{d}u\,\mathrm{d}z \\ &= \iint K(v)K(u)f(x+hv-hu,t_{\ell i})\,\mathrm{d}u\,\mathrm{d}v \end{aligned}$$

$$\leq M \iint K(v)K(u)\,\mathrm{d}u\,\mathrm{d}v = M. \tag{2.D.20}$$

Let $\varphi_j(x) = \mathrm{E}\{U^2(x, X_{\ell j})\}$. Then, under (A2) and (A3), we have, uniformly in x,

$$\varphi_j(x) = \int U^2(x,y)f(y,t_{\ell j})\,\mathrm{d}y \leq \frac{M}{h}\int U(x,y)f(y,t_{\ell j})\,\mathrm{d}y = \frac{M}{h}U_j(x) \leq \frac{M^2}{h},$$

where we used (2.D.17) to get the first inequality, the definition of $U_i(x)$ in (2.D.8) to get the second equation, and (2.D.20) to get the second inequality. Hence, we have

$$\mathrm{E}\{U^2(X_{\ell i}, X_{\ell j})|X_{\ell i}\} = \varphi_j(X_{\ell i}) \leq \frac{M^2}{h}. \tag{2.D.21}$$

Now, using (2.D.20), (2.C.35) still holds. Combining the above results with (2.C.32), taking there U as in (2.D.8), we have, when Δt is small enough such that $h^{-1} \geq 1$,

$$\mathrm{E}(V_{ij}^2|X_{\ell i}) \leq \frac{C_1}{h}, \tag{2.D.22}$$

where $C_1 = 4M^2$.

Bounding $\mathrm{E}\,|V_{ij}|^p$. Applying (2.D.17) to V_{ij} defined in (2.D.8), we have $|V_{ij}| \leq C_2 h^{-1}$ with $C_2 = 4M$. Hence, we have

$$\mathrm{E}\,|V_{ij}|^p \leq \frac{C_2^p}{h^p}. \tag{2.D.23}$$

Conclusion. Recall the definition of $\tilde{w}_{\ell j}$ below (2.C.6). Now, using (B1) and plugging (2.D.22) and (2.D.23) into (2.D.19), we have, for $i = 2, \ldots, b_\ell/2$ and $p \geq 2.5$,

$$\begin{aligned}
\mathrm{E}\,|Y_i|^p &\leq C_p\left\{C_1^{p/2}h^{-p/2}\left(\sum_{j=1}^{i-1}\tilde{w}_{\ell i}^2\tilde{w}_{\ell j}^2\right)^{p/2} + C_2^p h^{-p}\sum_{j=1}^{i-1}\tilde{w}_{\ell i}^p\tilde{w}_{\ell j}^p\right\} \\
&\leq C_p\left[C_1^{p/2}h^{-p/2}\left\{\gamma^4\sum_{j=1}^{i-1}(1-\gamma)^{b_\ell/2-j}\right\}^{p/2} + C_2^p h^{-p}\gamma^{2p}\sum_{j=1}^{i-1}(1-\gamma)^{b_\ell/2-j}\right] \\
&\leq C_p(C_1^{p/2}h^{-p/2}\gamma^{3p/2} + C_2^p h^{-p}\gamma^{2p-1}) \leq C_p\{C_1^{p/2}\delta^{-2p}\Delta t^p + C_2^p\delta^{1-3p}\Delta t^{(9p-5)/7}
\end{aligned}$$

$$\leq C_p(C_1^{p/2}\delta^{-2p} + C_2^p\delta^{1-3p})\Delta t^p, \tag{2.D.24}$$

where we used the fact that $0 < \gamma < 1$.

Plugging (2.D.24) into (2.D.18), we have

$$\mathrm{E}\,|I_{21}|^p \leq 2C_p^2(b_\ell/2)^{p/2}(C_1^{p/2}\delta^{-2p} + C_2^p\delta^{1-3p})\Delta t^p \leq C_3^p C_p^2 b_\ell^{p/2}\Delta t^p, \tag{2.D.25}$$

where $C_3 = 2^{1/2}(C_1^{1/2}\delta^{-2} + C_2\delta^{-3})$ is a constant. By (2.C.3), we have $\mathrm{E}\,|I_{21}|^p \leq \{(C_0^2 C_3 p^2/\log^2 p)b_\ell^{1/2}\Delta t\}^p$. Now, by Chebyshev's inequality, we have, for all $\eta > 0$,

$$\mathrm{P}(|I_{21}| > \eta\Delta t^{4/7}) \leq \frac{\mathrm{E}\,|I_{21}|^p}{(\eta\Delta t^{4/7})^p} \leq \left(\frac{C_0^2 C_3}{\eta\log^2 p}\right)^p (p^2 b_\ell^{1/2}\Delta t^{3/7})^p.$$

Under (B3), taking $p = b_\ell^{-1/4}\Delta t^{-3/14}/\sqrt{e}$, and then, recalling that $b_\ell^{-1} \asymp \Delta t^{(5+\alpha)/7}$ for some $0 < \alpha < 2/3$ using (B3) and (2.27), we have $p \to \infty$ as $\Delta t \to 0$ and

$$\mathrm{P}(|I_{21}| > \eta\Delta t^{4/7}) = o\big\{\exp(-b_\ell^{-1/4}\Delta t^{-3/14}/\sqrt{e})\big\}. \tag{2.D.26}$$

Finally, using (B1), (B3) and (2.D.26), we have

$$\begin{aligned}\mathrm{P}\left\{\sup_{(\gamma,h)\in I_{\gamma,h}^\ell} |I_{21}| > \eta\Delta t^{4/7}\right\} &\leq \#(I_{\gamma,h}^\ell)\sup_{(\gamma,h)\in I_{\gamma,h}^\ell}\mathrm{P}(|I_{21}| > \eta\Delta t^{4/7}) \\ &= o\big\{\Delta t^{-a}\exp(-b_\ell^{-1/4}\Delta t^{-3/14}/\sqrt{e})\big\} = o(1),\end{aligned}$$

which implies that

$$I_{21} = o_p(\Delta t^{4/7}), \tag{2.D.27}$$

uniformly in $(\gamma, h) \in I_{\gamma,h}^\ell$.

2.D.1.3 Calculating I_{22} and I_{23}

The calculations of I_{22} and I_{23} are similar and we first calculate I_{22} as follows. Let $W_i(x) = U_i(x) - f(x, t_{\ell i})$ for $i = 2, \ldots, b_\ell/2$, where $U_i(x)$ is defined in (2.D.8). Then we have $I_{22} = I_{221} + I_{222}$, where

$$I_{221} = \sum_{i=2}^{b_\ell/2} \sum_{j=1}^{i-1} \tilde{w}_{\ell i} \tilde{w}_{\ell j} \big[W_i(X_{\ell j}) - \mathrm{E}\{W_i(X_{\ell j})\} \big], \tag{2.D.28}$$

$$I_{222} = \sum_{i=2}^{b_\ell/2} \sum_{j=1}^{i-1} \tilde{w}_{\ell i} \tilde{w}_{\ell j} \big[f(X_{\ell j}, t_{\ell i}) - \mathrm{E}\{f(X_{\ell j}, t_{\ell i})\} \big]. \tag{2.D.29}$$

Next we show that $I_{221} = o_p(\Delta t^{4/7})$ uniformly in $(\gamma, h) \in I^\ell_{\gamma,h}$, which implies that

$$I_{22} = I_{222} + o_p(\Delta t^{4/7}), \tag{2.D.30}$$

uniformly in $(\gamma, h) \in I^\ell_{\gamma,h}$.

Calculating I_{221}. First note from (2.D.28) that

$$I_{221} = \sum_{j=1}^{b_\ell/2-1} \tilde{w}_{\ell j} \sum_{i=j+1}^{b_\ell/2} \tilde{w}_{\ell i} \big[W_i(X_{\ell j}) - \mathrm{E}\{W_i(X_{\ell j})\} \big].$$

Let $Z_j = \tilde{w}_{\ell j} \sum_{i=j+1}^{b_\ell/2} \tilde{w}_{\ell i} \big[W_i(X_{\ell j}) - \mathrm{E}\{W_i(X_{\ell j})\} \big]$ for $j = 1, \ldots, b_\ell/2 - 1$, so that $I_{221} = \sum_{j=1}^{b_\ell/2-1} Z_j$ is a sum of independent random variables since each Z_j only depends on one random variable $X_{\ell j}$. To apply Theorem 2.C.2 to I_{221}, we first derive bounds for $|Z_j|$ and $\sum_{j=1}^{b_\ell/2-1} \mathrm{E}(Z_j^2)$.

Using (A2), (A3) and (2.D.8), we have

$$\begin{aligned} U_i(x) &= \frac{1}{h^2} \iint K\Big(\frac{z-x}{h}\Big) K\Big(\frac{z-y}{h}\Big) f(y, t_{\ell i}) \, \mathrm{d}y \, \mathrm{d}z \\ &= \frac{1}{h} \int K\Big(\frac{z-x}{h}\Big) \int K(u) f(z - hu, t_{\ell i}) \, \mathrm{d}u \, \mathrm{d}z \\ &= \frac{1}{h} \int K\Big(\frac{z-x}{h}\Big) \int K(u) \Big\{ f(z, t_{\ell i}) \end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad - huf_x(z, t_{\ell i}) + \frac{1}{2}h^2u^2 f_{xx}(z - \theta_i hu, t_{\ell i})\Big\} \,\mathrm{d}u \,\mathrm{d}z \\
&= \frac{1}{h}\int K\Big(\frac{z-x}{h}\Big) f(z, t_{\ell i}) \,\mathrm{d}z + r_{1i}(x) \\
&= \int K(u) f(x + hu, t_{\ell i}) \,\mathrm{d}u + r_{1i}(x) \\
&= \int K(u)\Big\{ f(x, t_{\ell i}) + huf_x(x, t_{\ell i}) + \frac{1}{2}h^2u^2 f_{xx}(x + \xi_i hu, t_{\ell i})\Big\} \,\mathrm{d}u + r_{1i}(x) \\
&= f(x, t_{\ell i}) + r_{1i}(x) + r_{2i}(x), \qquad (2.D.31)
\end{aligned}
$$

where $\theta_i, \xi_i \in [0, 1]$, we used first-order Taylor expansions of $f(z - hu, t_{\ell i})$ and $f(x + hu, t_{\ell i})$ around z and x, and where

$$
\begin{aligned}
r_{1i}(x) &= \frac{1}{2}h\int u^2K(u)\int K\Big(\frac{z-x}{h}\Big) f_{xx}(z - \theta_i hu, t_{\ell i}) \,\mathrm{d}u \,\mathrm{d}z, \\
r_{2i}(x) &= \frac{1}{2}h^2\int u^2K(u) f_{xx}(x + \xi_i hu, t_{\ell i}) \,\mathrm{d}u.
\end{aligned}
$$

Note that, under (A2) and (A3), we have

$$
|r_{1i}(x)| \le \frac{M}{2}h\int u^2K(u) \,\mathrm{d}u \int K\Big(\frac{z-x}{h}\Big) \,\mathrm{d}z = \frac{M}{2}\mu_{K,2}h^2, \qquad (2.D.32)
$$

where $\mu_{K,2}$ is defined above Proposition 2.1, and we also have

$$
|r_{2i}(x)| \le \frac{M}{2}\mu_{K,2}h^2. \qquad (2.D.33)
$$

Now, plugging (2.D.31), (2.D.32) and (2.D.33) into the definition of $W_i(x)$ above (2.D.28), we have, under (B1),

$$
|W_i(x)| \le M\mu_{K,2}h^2 \le \frac{M\mu_{K,2}}{\delta^2}\Delta t^{2/7},
$$

and hence, recalling the definition of $\tilde{w}_{\ell j}$ below (2.C.6), we have

$$
|Z_j| \le 2\frac{M\mu_{K,2}}{\delta^2}\Delta t^{2/7}\tilde{w}_{\ell j}\sum_{i=j+1}^{b_\ell/2}\tilde{w}_{\ell i}
$$

$$= 2\frac{M\mu_{K,2}}{\delta^2}\Delta t^{2/7}\gamma^2(1-\gamma)^{b_\ell/2-j}\sum_{i=j+1}^{b_\ell/2}(1-\gamma)^{b_\ell/2-i}$$

$$= 2\frac{M\mu_{K,2}}{\delta^2}\Delta t^{2/7}\gamma^2(1-\gamma)^{b_\ell/2-j}\sum_{i=0}^{b_\ell/2-j-1}(1-\gamma)^{i} \tag{2.D.34}$$

$$\leq 2\frac{M\mu_{K,2}}{\delta^2}\Delta t^{2/7}\gamma \leq C_1\Delta t, \tag{2.D.35}$$

where $C_1 = 2M\mu_{K,2}/\delta^3$, since $0 < \gamma < 1$, and where we used (B1).

To bound $\sum_{j=1}^{b_\ell/2-1}\mathrm{E}(Z_j^2)$, using (2.D.34) and (B1), we have

$$\begin{aligned}
\sum_{j=1}^{b_\ell/2-1}\mathrm{E}(Z_j^2) &\leq 4\frac{M^2\mu_{K,2}^2}{\delta^4}\Delta t^{4/7}\gamma^4\sum_{j=1}^{b_\ell/2-1}(1-\gamma)^{2(b_\ell/2-j)}\left\{\sum_{i=0}^{b_\ell/2-j-1}(1-\gamma)^i\right\}^2 \\
&\leq 4\frac{M^2\mu_{K,2}^2}{\delta^4}\Delta t^{4/7}\gamma^3\sum_{j=1}^{b_\ell/2-1}(1-\gamma)^{b_\ell/2-j}\sum_{i=0}^{b_\ell/2-j-1}(1-\gamma)^i \\
&= 4\frac{M^2\mu_{K,2}^2}{\delta^4}\Delta t^{4/7}\gamma^2\sum_{j=1}^{b_\ell/2-1}(1-\gamma)^{b_\ell/2-j}\{1-(1-\gamma)^{b_\ell/2-j}\} \\
&\leq 4\frac{M^2\mu_{K,2}^2}{\delta^4}\Delta t^{4/7}\gamma^2\sum_{j=1}^{b_\ell/2-1}(1-\gamma)^{b_\ell/2-j} \\
&\leq 4\frac{M^2\mu_{K,2}^2}{\delta^4}\Delta t^{4/7}\gamma \leq C_2\Delta t^{9/7},
\end{aligned} \tag{2.D.36}$$

where $C_2 = 4M^2\mu_{K,2}^2/\delta^5$.

Applying Theorem 2.C.2 and using (2.D.35) and (2.D.36), we have, for any $\eta > 0$,

$$\mathrm{P}\big(|I_{221}| > \eta\Delta t^{4/7}\big) \leq \exp\left\{-\frac{\eta^2\Delta t^{8/7}}{2(C_2\Delta t^{9/7} + \eta C_1\Delta t^{11/7}/3)}\right\}.$$

Then, under (B1), we have

$$\begin{aligned}
\mathrm{P}\left\{\sup_{(\gamma,h)\in I_{\gamma,h}^\ell}|I_{221}| > \eta\Delta t^{4/7}\right\} &\leq \#(I_{\gamma,h}^\ell)\sup_{(\gamma,h)\in I_{\gamma,h}^\ell}\mathrm{P}\big(|I_{221}| > \eta\Delta t^{4/7}\big) \\
&\leq C\Delta t^{-a}\exp\left\{-\frac{\eta^2\Delta t^{8/7}}{2(C_1\Delta t^{9/7} + \eta C_2\Delta t^{11/7}/3)}\right\} = o(1),
\end{aligned}$$

which implies that

$$I_{221} = o_p(\Delta t^{4/7}), \tag{2.D.37}$$

uniformly in $(\gamma, h) \in I^{\ell}_{\gamma,h}$. Hence (2.D.30) is proved.

Calculating I_{23}. Note from (2.D.12) and (2.D.13) that I_{23} is the same as I_{22} except that $U_i(X_{\ell j})$ is replaced by $U_j(X_{\ell i})$. Therefore, similar to (2.D.30), we have

$$I_{23} = \sum_{i=2}^{b_\ell/2} \sum_{j=1}^{i-1} \tilde{w}_{\ell i} \tilde{w}_{\ell j} \big[f(X_{\ell i}, t_{\ell j}) - \mathrm{E}\{f(X_{\ell i}, t_{\ell j})\} \big] + o_p(\Delta t^{4/7}), \tag{2.D.38}$$

uniformly in $(\gamma, h) \in I^{\ell}_{\gamma,h}$.

2.D.1.4 Conclusion

Plugging (2.D.27), (2.D.30) and (2.D.38) into the formula for I_2 above (2.D.11), we have

$$\begin{aligned} I_2 = &\sum_{i=2}^{b_\ell/2} \sum_{j=1}^{i-1} \tilde{w}_{\ell i} \tilde{w}_{\ell j} \big[f(X_{\ell j}, t_{\ell i}) - \mathrm{E}\{f(X_{\ell j}, t_{\ell i})\} \big] \\ &+ \sum_{i=2}^{b_\ell/2} \sum_{j=1}^{i-1} \tilde{w}_{\ell i} \tilde{w}_{\ell j} \big[f(X_{\ell i}, t_{\ell j}) - \mathrm{E}\{f(X_{\ell i}, t_{\ell j})\} \big] + o_p(\Delta t^{4/7}). \end{aligned} \tag{2.D.39}$$

Then, plugging (2.D.16) and (2.D.39) into the formula for I above (2.D.9), we have

$$\begin{aligned} I &= 2 \left(\sum_{i=1}^{b_\ell/2-1} \sum_{j=i+1}^{b_\ell/2} + \sum_{i=2}^{b_\ell/2} \sum_{j=1}^{i-1} \right) \tilde{w}_{\ell i} \tilde{w}_{\ell j} \big[f(X_{\ell i}, t_{\ell j}) - \mathrm{E}\{f(X_{\ell i}, t_{\ell j})\} \big] + o_p(\Delta t^{4/7}) \\ &= 2I' - 2I_3 + o_p(\Delta t^{4/7}), \end{aligned} \tag{2.D.40}$$

where

$$I' = \sum_{i=1}^{b_\ell/2} \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell i} \tilde{w}_{\ell j} \big[f(X_{\ell i}, t_{\ell j}) - \mathrm{E}\{f(X_{\ell i}, t_{\ell j})\} \big], \tag{2.D.41}$$

$$I_3 = \sum_{i=1}^{b_\ell/2} \tilde{w}_{\ell i}^2 \big[f(X_{\ell i}, t_{\ell i}) - \mathrm{E}\{f(X_{\ell i}, t_{\ell i})\} \big]. \tag{2.D.42}$$

Finally, we show that

$$I_3 = o_p(\Delta t^{4/7}), \tag{2.D.43}$$

uniformly in $(\gamma, h) \in I_{\gamma,h}^\ell$, which implies that, using (2.D.40), we have

$$I = 2I' + o_p(\Delta t^{4/7}), \tag{2.D.44}$$

uniformly in $(\gamma, h) \in I_{\gamma,h}^\ell$.

Note that I_3 is a sum of independent random variables. Recalling the definition of $\tilde{w}_{\ell i}$ below (2.C.6) and using (A2) and (B1), we have $\tilde{w}_{\ell i}^2 f(X_{\ell i}, t_{\ell i}) \le M\gamma^2 \le C_1 \Delta t^{10/7}$, where $C_1 = M/\delta^2$, and, since $0 < \gamma < 1$, we have

$$\begin{aligned}\sum_{i=1}^{b_\ell/2} \tilde{w}_{\ell i}^4 \, \mathrm{E}\{f(X_{\ell i}, t_{\ell i})\}^2 &\le M^2\gamma^4 \sum_{i=1}^{b_\ell/2} (1-\gamma)^{4(b_\ell/2-i)} \le M^2\gamma^4 \sum_{i=1}^{b_\ell/2} (1-\gamma)^{b_\ell/2-i} \\ &= M^2\gamma^4 \sum_{j=0}^{b_\ell/2-1} (1-\gamma)^j \le M^2\gamma^3 \le C_2 \Delta t^{15/7},\end{aligned}$$

Using the above results, by Theorem 2.C.2, we have, for any $\eta > 0$,

$$\mathrm{P}\big(|I_3| > \eta \Delta t^{4/7}\big) \le \exp\left\{ -\frac{\eta^2 \Delta t^{8/7}}{2(C_2 \Delta t^{15/7} + \eta C_1 \Delta t^2/3)} \right\},$$

and hence, under (B1), we have

$$\begin{aligned}\mathrm{P}\left\{ \sup_{(\gamma,h)\in I_{\gamma,h}^\ell} |I_3| > \eta \Delta t^{4/7} \right\} &\le \#(I_{\gamma,h}^\ell) \sup_{(\gamma,h)\in I_{\gamma,h}^\ell} \mathrm{P}\big(|I_3| > \eta \Delta t^{4/7}\big) \\ &\le C \Delta t^{-a} \exp\left\{ -\frac{\eta^2 \Delta t^{8/7}}{2(C_2 \Delta t^{15/7} + \eta C_1 \Delta t^2/3)} \right\} = o(1),\end{aligned}$$

which proves (2.D.43) and hence also (2.D.44).

2.D.2 Remaining terms

Combining (2.D.1), (2.D.2), (2.D.3), (2.D.41) and (2.D.44), we have

$$\begin{aligned}\frac{1}{2}\big\{\mathrm{ISE}_\ell(\gamma,h)-\mathrm{MISE}_\ell(\gamma,h)\big\} &= I'-J_{41}+o_p(\Delta t^{4/7})\\ &= S_1-S_2-S_3+S_4+o_p(\Delta t^{4/7}),\end{aligned}\tag{2.D.45}$$

uniformly in $(\gamma,h)\in I^\ell_{\gamma,h}$, where, using (2.C.90),

$$S_1=\sum_{i=1}^{b_\ell/2}\sum_{j=1}^{b_\ell/2}\tilde{w}_{\ell i}\tilde{w}_{\ell j}f(X_{\ell i},t_{\ell j}),\tag{2.D.46}$$

$$S_2=\sum_{i=1}^{b_\ell/2}\sum_{j=1}^{b_\ell/2}\tilde{w}_{\ell i}\tilde{w}_{\ell j}\int f(x,t_{\ell i})f(x,t_{\ell j})\,\mathrm{d}x,\tag{2.D.47}$$

$$S_3=\sum_{i=1}^{b_\ell/2}\tilde{w}_{\ell i}f(X_{\ell i},s_\ell),\tag{2.D.48}$$

$$S_4=\sum_{i=1}^{b_\ell/2}\tilde{w}_{\ell i}\int f(x,s_\ell)f(x,t_{\ell i})\,\mathrm{d}x.\tag{2.D.49}$$

Next, we derive some technical results which will be used to calculate $S_1,\ldots,S_4$.

2.D.2.1 Technical results

Recalling the definition of $\tilde{w}_{\ell j}$ below (2.C.6) and using (B1), (B3) and (2.C.5), we have

$$\begin{aligned}\sum_{j=1}^{b_\ell/2}\tilde{w}_{\ell j}f(x,t_{\ell j}) &= \gamma\sum_{j=1}^{b_\ell/2}(1-\gamma)^{b_\ell/2-j}\Big\{f(x,s_\ell)-f_t(x,s_\ell)(s_\ell-t_{\ell j})\\ &\qquad+\frac{1}{2}f_{tt}(x,s_\ell)(s_\ell-t_{\ell j})^2-\frac{1}{6}(s_\ell-t_{\ell j})^3f_{ttt}(x,\theta_j)\Big\}\\ &= f(x,s_\ell)\gamma\sum_{i=0}^{b_\ell/2-1}(1-\gamma)^i\end{aligned}$$

$$
\begin{aligned}
&- f_t(x,s_\ell)\gamma\sum_{j=1}^{b_\ell/2}(1-\gamma)^{b_\ell/2-j}\{(b_\ell/2-j)\Delta t+O(\Delta t)\}\\
&\quad+\frac{1}{2}f_{tt}(x,s_\ell)\gamma\sum_{j=1}^{b_\ell/2}(1-\gamma)^{b_\ell/2-j}\{(b_\ell/2-j)\Delta t+O(\Delta t)\}^2\\
&\quad-\frac{1}{6}\gamma\sum_{j=1}^{b_\ell/2}(1-\gamma)^{b_\ell/2-j}\{(b_\ell/2-j)\Delta t+O(\Delta t)\}^3 f_{ttt}(x,\theta_j)\\
&= f(x,s_\ell)-\Delta t f_t(x,s_\ell)\gamma\sum_{i=0}^{b_\ell/2-1} i(1-\gamma)^i\\
&\quad+\Delta t^2\frac{1}{2}f_{tt}(x,s_\ell)\gamma\sum_{i=0}^{b_\ell/2-1} i^2(1-\gamma)^i\\
&\quad-\Delta t^3\frac{1}{6}\gamma\sum_{i=0}^{b_\ell/2-1} i^3(1-\gamma)^i f_{ttt}(x,\theta_{b_\ell/2-i})+r_1(x), \qquad (2.D.50)
\end{aligned}
$$

where θ_j is between $t_{\ell j}$ and s_ℓ and we used a second order Taylor expansion of $f(x,t_{\ell j})$ around time s_ℓ, and where

$$
r_1(x)=-f(x,s_\ell)(1-\gamma)^{b_\ell/2}+O(\Delta t)\gamma\sum_{i=0}^{b_\ell/2-1}(1-\gamma)^i=O(\Delta t), \qquad (2.D.51)
$$

using (2.C.86) and the fact that $0<\gamma<1$.

Then, note that, using summation by parts,

$$
\begin{aligned}
\gamma\sum_{i=0}^{b_\ell/2-1} i(1-\gamma)^i &= \gamma\sum_{j=0}^{b_\ell/2-2}\sum_{i=j+1}^{b_\ell/2-1}(1-\gamma)^i=\sum_{j=0}^{b_\ell/2-2}(1-\gamma)^{j+1}-(b_\ell/2-1)(1-\gamma)^{b_\ell/2}\\
&=\frac{1}{\gamma}\{1-\gamma-(1-\gamma)^{b_\ell/2-1}\}+o(\Delta t)=\frac{1}{\gamma}+O(1), \qquad (2.D.52)
\end{aligned}
$$

where we used (2.C.86).

Moreover, using (2.C.86), we have

$$
\gamma\sum_{i=0}^{b_\ell/2-1} i^2(1-\gamma)^i=\gamma\sum_{j=0}^{b_\ell/2-2}(2j+1)\sum_{i=j+1}^{b_\ell/2-1}(1-\gamma)^i
$$

$$
\begin{aligned}
&= \sum_{j=0}^{b_\ell/2-2} (2j+1)\{(1-\gamma)^{j+1} - (1-\gamma)^{b_\ell/2}\} \\
&= \sum_{j=0}^{b_\ell/2-2} (1-\gamma)^{j+1} \\
&\qquad + 2\sum_{i=0}^{b_\ell/2-3}\sum_{j=i+1}^{b_\ell/2-2} (1-\gamma)^{j+1} - \frac{1}{2}(1-\gamma)^{b_\ell/2}(b_\ell/2-1)^2 \\
&= \frac{2}{\gamma}\sum_{i=0}^{b_\ell/2-3}\{(1-\gamma)^{i+1} - (1-\gamma)^{b_\ell/2-1}\} + O(\gamma^{-1}) \\
&= \frac{2}{\gamma^2}\{1-\gamma-(1-\gamma)^{b_\ell/2-1}\} + O(\gamma^{-1}) = \frac{2}{\gamma^2} + O(\gamma^{-1}). \qquad \text{(2.D.53)}
\end{aligned}
$$

Finally, using (A2), (B1), (B3) and (2.D.53), we have

$$
\begin{aligned}
\left|\frac{\Delta t^3}{6}\gamma\sum_{i=0}^{b_\ell/2-1} i^3(1-\gamma)^i f_{ttt}(x,\theta_{b_\ell/2-i})\right| &\le \frac{M\Delta t^3}{6}\gamma\sum_{i=0}^{b_\ell/2-1} i^3(1-\gamma)^i \\
&\le \frac{Mb_\ell\Delta t^3}{6}\gamma\sum_{i=0}^{b_\ell/2-1} i^2(1-\gamma)^i \\
&= o(\Delta t^2/\gamma^2) = o(\Delta t^{4/7}), \qquad \text{(2.D.54)}
\end{aligned}
$$

uniformly in x.

Now, plugging (2.D.51), (2.D.52), (2.D.53) and (2.D.54) into (2.D.50), we have

$$
\sum_{j=1}^{b_\ell/2}\tilde{w}_{\ell j} f(x,t_{\ell j}) = f(x,s_\ell) - \frac{\Delta t}{\gamma} f_t(x,s_\ell) + \frac{\Delta t^2}{\gamma^2} f_{tt}(x,s_\ell) + o(\Delta t^{4/7}), \qquad \text{(2.D.55)}
$$

uniformly in x.

2.D.2.2 Conclusion

First, note from (2.D.46) and (2.D.48) that

$$\begin{aligned} S_1 - S_3 &= \sum_{i=1}^{b_\ell/2} \tilde{w}_{\ell i} \left\{ \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} f(X_{\ell i}, t_{\ell j}) - f(X_{\ell i}, s_\ell) \right\} \\ &= -\sum_{i=1}^{b_\ell/2} \tilde{w}_{\ell i} \left\{ \frac{\Delta t}{\gamma} f_t(X_{\ell i}, s_\ell) - \frac{\Delta t^2}{\gamma^2} f_{tt}(X_{\ell i}, s_\ell) \right\} + o_p(\Delta t^{4/7}), \end{aligned} \tag{2.D.56}$$

where we used (2.D.55).

Then, note from (2.D.47) and (2.D.49) that

$$\begin{aligned} -S_2 + S_4 &= \sum_{i=1}^{b_\ell/2} \tilde{w}_{\ell i} \int \left\{ f(x, s_\ell) - \sum_{j=1}^{b_\ell/2} \tilde{w}_{\ell j} f(x, t_{\ell j}) \right\} f(x, t_{\ell i}) \, \mathrm{d}x \\ &= \sum_{i=1}^{b_\ell/2} \tilde{w}_{\ell i} \int \left\{ \frac{\Delta t}{\gamma} f_t(x, s_\ell) - \frac{\Delta t^2}{\gamma^2} f_{tt}(x, s_\ell) \right\} f(x, t_{\ell i}) \, \mathrm{d}x + o(\Delta t^{4/7}), \end{aligned} \tag{2.D.57}$$

where we used (2.D.55).

Now, recalling the definitions of $F_{b_\ell/2}(x)$ and $\bar{F}_{b_\ell/2}(x)$ in (2.C.59) and using (2.D.56) and (2.D.57), we have

$$\begin{aligned} &|S_1 - S_2 - S_3 + S_4| \\ &\quad = \left| \int \left\{ \frac{\Delta t}{\gamma} f_t(x, s_\ell) - \frac{\Delta t^2}{\gamma^2} f_{tt}(x, s_\ell) \right\} \mathrm{d}x \left\{ F_{b_\ell/2}(x) - \bar{F}_{b_\ell/2}(x) \right\} \right| \\ &\quad = \left| \int \left\{ F_{b_\ell/2}(x) - \bar{F}_{b_\ell/2}(x) \right\} \left\{ \frac{\Delta t}{\gamma} f_{xt}(x, s_\ell) - \frac{\Delta t^2}{\gamma^2} f_{xtt}(x, s_\ell) \right\} \mathrm{d}x \right| \\ &\quad \leq \sup_{x \in \mathbb{R}} |F_{b_\ell/2}(x) - \bar{F}_{b_\ell/2}(x)| \left\{ \frac{\Delta t}{\gamma} \int |f_{xt}(x, s_\ell)| \, \mathrm{d}x + \frac{\Delta t^2}{\gamma^2} \int |f_{xtt}(x, s_\ell)| \, \mathrm{d}x \right\} \\ &\quad = O_p(\Delta t / \gamma^{1/2}) = o_p(\Delta t^{4/7}), \end{aligned} \tag{2.D.58}$$

uniformly in $(\gamma, h) \in I^\ell_{\gamma,h}$, where we used (2.C.60), (B1) and (B2) in the last line. Plugging (2.D.58) into (2.D.45), Proposition 2.2 is proved.

Chapter 3

Nonparametric regression for streaming data

In this chapter we propose a streaming regression algorithm (SRA) to estimate the time-varying regression model (1.3):

$$Y_i = m(X_i, t_i) + \epsilon_i\,, \quad i = 1, 2, \ldots,$$

where $m(x, t_i) = \mathrm{E}(Y_i|X_i = x)$ is a time-varying regression function and $\{\epsilon_i\}_{i=1,2,\ldots}$ is an error sequence.

The SRA is inspired by ensemble learning, an approach for nonstationarity adaptation commonly used by machine learning researchers in modelling streaming data (see §1.3.1.2). Specifically, the SRA computes a mild-sized ensemble of semi-recursive NW estimators, defined at (1.9), and adaptively selects the estimator that best fits the current data points. The ensemble is updated once in a while with promising new estimators added in and outdated ones discarded. Both the ensemble updating and the model selection are based on a computationally efficient recursive cross-validation (RCV) procedure.

Instead of considering i.n.i.d. streaming data as in Chapter 2, in this chapter we relax the independence assumption and consider a d.n.i.d. data stream $\{(X_i, Y_i)\}_{i=1,2,\ldots}$, with arrival times $\{t_i\}_{i=1,2,\ldots}$ defined at (1.1). We assume that the data are sequentially available from the time-varying regression model (1.3), where $\{X_i\}_{i=1,2,\ldots}$ is a weakly dependent sequence satisfying some mixing conditions (see §3.3). We also assume that the error sequence $\{\epsilon_i\}_{i=1,2,\ldots}$ is i.i.d. and independent of $\{X_i\}_{i=1,2,\ldots}$, satisfying $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{var}(\epsilon_i) = \sigma^2$ for all i. We make the

i.i.d. assumption about the errors to simplify theoretical derivations. It will be relaxed in future work, so that $\{\epsilon_i\}_{i=1,2,\ldots}$ may be d.n.i.d. as well. In that case, the regression methodology described in this chapter can still be applied without substantial modification. See §5.2.3 for an explanation.

To better illustrate the basic idea of the SRA, we first discuss a prototype algorithm, which is a naive application of ensemble learning to the streaming regression problem. The prototype algorithm is easier to interpret but difficult to implement due to its high computational cost. Then we present the SRA as a computationally efficient approximation to the prototype algorithm. Theoretical results in §3.3 will show that these two algorithms have some very similar asymptotic properties.

3.1 Prototype algorithm

In this section, we introduce a prototype algorithm for estimating the time-varying regression model (1.3) and then discuss how to make it more computationally efficient. Recalling the definition of the t_i's at (1.1), the goal here is to estimate the regression function $m(\cdot, t)$ at $t = t_1, t_2, \ldots,$ whenever a new data point is observed. This requires us to update the estimator at each time step. For that purpose, the prototype algorithm will compute the semi-recursive NW estimator, defined at (1.9), with smoothing parameters (γ, h) on a grid. A cross-validation criterion is used to select, for each t_i, the most appropriate (γ, h), and thence the most appropriate estimator. This algorithm can be viewed as a straightforward application of the ensemble learning idea with the semi-recursive NW estimator as the base learner.

3.1.1 Definition of base learner

Consider estimating the time-varying regression function $m(x, t)$ at (1.3) by

$$\tilde{m}(x, t|\gamma, h) = \hat{m}(x, t|\tilde{\boldsymbol{\gamma}}_t, \tilde{\boldsymbol{h}}_t)\,, \tag{3.1}$$

where $\hat{m}$ is defined at (1.9) and

$$\tilde{\boldsymbol{\gamma}}_t = \{\tilde{\gamma}_i\}_{i=1,\ldots,n_t} \text{ and } \tilde{\boldsymbol{h}}_t = \{\tilde{h}_i\}_{i=1,\ldots,n_t} \text{ with } (\tilde{\gamma}_i, \tilde{h}_i) \equiv (\gamma, h)\,, \tag{3.2}$$

for some $(\gamma, h) \in (0,1) \times \mathbb{R}_+$. That is, $\tilde{m}$ is a semi-recursive NW estimator using fixed stepsize γ and bandwidth h.

By (1.8)–(1.10), $\tilde{m}(x,t|\gamma,h) = \tilde{r}(x,t|\gamma,h)/\tilde{f}(x,t|\gamma,h)$ can be semi-recursively computed by

$$\begin{aligned}
\tilde{r}(x,t|\gamma,h) &= \begin{cases} K_h(x - X_1)Y_1\,, & \text{if } n_t = 1\,, \\ (1-\gamma)\tilde{r}(x,t_{n_t-1}|\gamma,h) + \gamma K_h(x - X_{n_t})Y_{n_t}\,, & \text{if } n_t \geq 2\,, \end{cases} \\
\tilde{f}(x,t|\gamma,h) &= \begin{cases} K_h(x - X_1)\,, & \text{if } n_t = 1\,, \\ (1-\gamma)\tilde{f}(x,t_{n_t-1}|\gamma,h) + \gamma K_h(x - X_{n_t})\,, & \text{if } n_t \geq 2\,. \end{cases}
\end{aligned} \tag{3.3}$$

The way we recursively compute $\tilde{r}$ and $\tilde{f}$ in (3.3) can be viewed as exponential smoothing (see §1.3.1.2) applied to kernel functions. For a fixed $\gamma \in (0,1)$, the weights for past observations decrease exponentially fast, so that effectively we are using only a number of recent observations to estimate $m(x,t)$. The stepsize γ plays a key role here since it controls how fast we 'forget' the past.

3.1.2 Definition of prototype algorithm

Considering that the time domain $\mathbb{R}_+$ of data stream $\{(X_i, Y_i)\}_{i=1,2,\ldots}$ is never-ending, using $\tilde{m}(\cdot,t|\gamma,h)$ defined at (3.1) with fixed (γ,h) for all $t \in \mathbb{R}_+$ is likely to be unsatisfactory, since the appropriate γ and h values may be time-varying. For example, suppose the true regression function $m(x,t) = \sin(x - a_t t)$, where $(x,t) \in \mathbb{R} \times \mathbb{R}_+$ and $a_t > 0$, is a sine curve with (non-time-independent) period 2π but time-varying locations. Here the time-variability of the regression function m is controlled by the parameter a_t. Suppose for two different time points $s_1, s_2 > 0$ we have $a_{s_1} > a_{s_2}$, i.e. m varies faster around time s_1 than around time s_2. Then

it is more appropriate to take two different stepsizes $\gamma_1 > \gamma_2$ and use $\tilde{m}(\cdot, s_1|\gamma_1, h)$ to estimate $m(\cdot, s_1)$ and $\tilde{m}(\cdot, s_2|\gamma_2, h)$ to estimate $m(\cdot, s_2)$. We use the same h at times s_1 and s_2 since $m(\cdot, s_1)$ and $m(\cdot, s_2)$ have the same curvature, but we use larger stepsize γ_1 at time s_1 so that, recalling (3.3), the resulting estimator $\tilde{m}(\cdot, s_1|\gamma_1, h)$ weighs down past observations more quickly.

The above considerations motivate us to compute an ensemble of $\tilde{m}$'s using (γ, h) on a grid $I_{\gamma,h}$. At each t_k, we compute an ensemble by updating the estimators of $m(\cdot, t_{k-1})$ computed at the previous time t_{k-1}, and then dynamically select the best estimator from the ensemble by minimising a cross-validation criterion. This is the key idea of our prototype algorithm. Next we iteratively define the algorithm.

At $t = t_1$, we observe the first data point (X_1, Y_1). We initialise the ensemble $\mathscr{E}_t^0 = \mathscr{E}_{t_1}^0 = \{\tilde{m}(\cdot, t_1|\gamma, h); \tilde{r}(\cdot, t_1|\gamma, h), \tilde{f}(\cdot, t_1|\gamma, h) : (\gamma, h) \in I_{\gamma,h}\}$, where each $\tilde{m} = \tilde{r}/\tilde{f}$ with $\tilde{r}$ and $\tilde{f}$ computed by (3.3). Here the regression estimator $\tilde{m}$ is the main quantity of interest in the ensemble, but we also store $\tilde{r}$ and $\tilde{f}$ in the ensemble. In the prototype algorithm, the idea of an ensemble $\mathscr{E}_t^0$ does not play a very important role, as the grid $I_{\gamma,h}$ will be non-time-varying. However, we will see in §3.2.1 that in our main algorithm we will use a time-varying ensemble.

At every consecutive time point t_k, for $k = 2, 3, \ldots$, we observe (X_k, Y_k). We use this new observation to update the ensemble $\mathscr{E}_{t_{k-1}}^0$ to the new ensemble $\mathscr{E}_{t_k}^0$ by updating $\tilde{r}(\cdot, t_{k-1}|\gamma, h)$, $\tilde{f}(\cdot, t_{k-1}|\gamma, h)$ and $\tilde{m}(\cdot, t_{k-1}|\gamma, h)$ to $\tilde{r}(\cdot, t_k|\gamma, h)$, $\tilde{f}(\cdot, t_k|\gamma, h)$ and $\tilde{m}(\cdot, t_k|\gamma, h)$ using (3.3). We see from (3.3) that this only involves computing $K_h(\cdot - X_k)$ and $K_h(\cdot - X_k)Y_k$ and adding them to the old estimators $\tilde{r}(\cdot, t_{k-1}|\gamma, h)$ and $\tilde{f}(\cdot, t_{k-1}|\gamma, h)$ using (3.3).

At time t_k, we select the best estimator of $m(\cdot, t_k)$ from the ensemble $\mathscr{E}_t^0$ by minimising with respect to $(\gamma, h) \in I_{\gamma,h}$ the following cross-validation criterion:

$$\mathrm{CV}_{t_k}^0(\gamma, h) = \sum_{i=k-b_{t_k}}^{k-1} \{\tilde{m}(X_{i+1}, t_i|\gamma, h) - Y_{i+1}\}^2, \tag{3.4}$$

where $b_{t_k} \in \{1, \ldots, k-1\}$. That is, let

$$(\gamma_k^{CV^0}, h_k^{CV^0}) = \underset{(\gamma,h)\in I_{\gamma,h}}{\arg\min}\ \mathrm{CV}_{t_k}^0(\gamma, h).$$

and we use $\tilde{m}(\cdot, t_k|\gamma_k^{CV^0}, h_k^{CV^0})$ to estimate $m(\cdot, t_k)$. This implies that each time t_k gives rise to a potentially different choice $(\gamma_k^{CV^0}, h_k^{CV^0})$ of the (γ, h) value, and hence we are not using the same (γ, h) value throughout.

Next we explain the motivation for the $\mathrm{CV}_{t_k}^0$ criterion at (3.4). Since we want to select (γ, h) to use at time t_k, so that $\check{m}(\cdot, t_k|\gamma, h)$ is a good estimator of $m(\cdot, t_k)$, ideally at time t_k we would use a sample of i.i.d. data from the distribution at time t_k to compute the cross-validation. However, the only data point from that distribution we have is (X_k, Y_k), so we have to use data from the recent past $\{(X_i, Y_i)\}_{i=k-b_{t_k}+1,\ldots,k}$. This is reasonable since, when the data distribution is smoothly time-varying and b_{t_k} is appropriately small, these data should be approximately distributed as (X_k, Y_k). In addition, note from (1.3) that, since the ϵ_i's are independent and they are independent of the X_i's, the Y_i's are conditionally independent from each other given the X_i's. This implies that, if b_{t_k} is appropriately chosen, we can view the data $\{(X_i, Y_i)\}_{i=k-b_{t_k}+1,\ldots,k}$ used to compute $\mathrm{CV}_{t_k}^0$ as nearly i.i.d., given $\{X_i\}_{i=k-b_{t_k}+1,\ldots,k}$.

The summand $\{\tilde{m}(X_{i+1}, t_i|\gamma, h) - Y_{i+1}\}^2$ in the cross-validation criterion (3.4) may be termed as the one-step-ahead error. Similar cross-validation criteria are sometimes used in the time series literature to evaluate the performance of prediction methods for time series. See, for example, Gijbels et al. (1999), where the authors considered a univariate time series $\{X_i\}_{i=1,\ldots,n}$ observed at times $t_1, \ldots, t_n$ and coming from the model $X_i = g(t_i) + \epsilon_i$, where g is a smooth function representing the trend of $\{X_i\}_{i=1,\ldots,n}$ and $\{\epsilon_i\}_{i=1,\ldots,n}$ is an mean-zero error sequence. To evaluate the performance of the one-step-ahead predictor $\hat{g}$, where $\hat{g}(t_i)$ computed from $X_1, \ldots, X_i$ is the predictor of X_{i+1}, Gijbels et al. (1999) considered the criterion $\sum_{i=1}^{n-1}\{\hat{m}(t_i) - X_{i+1}\}^2$.

3.1.3 Issues with prototype algorithm

We claimed above §3.1 that the prototype algorithm is not directly applicable to streaming data due to its high computational cost. To see this, note that the candidate set $I_{\gamma,h}$ in (3.4) should contain a large number of candidate (γ, h)'s on $(0,1) \times \mathbb{R}_+$, since appropriate γ and h values may be very different for different times. This means that, to compute the ensemble $\mathscr{E}^0_{t_k}$ for $k = 2, 3, \ldots$, we have to update the $\tilde{m}$'s from time t_{k-1} to t_k for all $(\gamma, h) \in I_{\gamma,h}$, where $I_{\gamma,h}$ is a very large set. Recall from (3.3) that this would involve computing a large number of K_h's and storing all the associated $\tilde{r}$'s and $\tilde{f}$'s. In addition, we have to compute a large number (one for each $(\gamma, h) \in I_{\gamma,h}$) of $\mathrm{CV}^0_{t_k}$'s, in order to select the appropriate (γ, h).

To solve this problem, ideally we would dynamically adjust the grid $I_{\gamma,h}$ so that it never needs to be very large, which means that at each time t_k we would only need to compute a much smaller number of estimates and cross-validation scores. However, this would make the situation even worse. This is because, to compute a new estimator $\tilde{m}(\cdot, t_k|\gamma_0, h_0)$ for (γ_0, h_0) for which $\tilde{m}(\cdot, t_{k-1}|\gamma_0, h_0)$ is not in the previous ensemble $\mathscr{E}^0_{t_{k-1}}$, we need to use all data observed up to time t_k and compute the $\tilde{r}$'s and $\tilde{f}$'s for the new values (γ_0, h_0). However, storing all past data is generally impossible in our streaming data setting and is exactly what we try to avoid by introducing the semi-recursive estimator at (3.3). We will solve this problem in §3.2.1, where a new estimator $\check{m}$ is introduced to approximate $\tilde{m}$. There, to compute $\check{m}(\cdot, t_k|\gamma_0, h_0)$ for some (γ_0, h_0) for which $\check{m}(\cdot, t_{k-1}|\gamma_0, h_0)$ is not in $\mathscr{E}_{t_{k-1}}$, we do not need to use all data up to t_k. As we will see, this approximation is based on the fact that, through the parameter γ, past data receive exponentially decreasing weights and $\tilde{m}(\cdot, t|\gamma_0, h_0)$ is computed using essentially a small number of recent data points.

Besides the computational issue, an additional challenge for the prototype algorithm is the selection of b_{t_k} at (3.4), i.e. the number of observations used to compute the cross-validation score $\mathrm{CV}^0_{t_k}$. There is a trade-off involved in selecting b_{t_k}. On the one hand, b_{t_k} should not be too small as this would make the cross-validation numerically unstable. On the other hand, if b_{t_k} is too large, then the data used to compute the cross-validation scores may have very different distributions than at time t_k. In this case, by minimising $\mathrm{CV}^0_{t_k}$, we would select a value (γ_k, h_k)

for which the estimator of $m(\cdot, t_k)$ may not be a very good one. Ideally, b_{t_k} should be adaptive to the time-variability of $m(\cdot, t_k)$, i.e. the slower $m(\cdot, t_k)$ changes in t_k, the more data we should use to compute $\text{CV}^0_{t_k}$, and vice versa. This trade-off between numerical stability and time sensitivity is sometimes referred to as the 'stability–plasticity dilemma' in the machine learning community and is identified as one of the major difficulties in streaming data analysis (Ditzler et al., 2015).

Despite these computational issues and those of selecting b_{t_k} discussed above, the prototype algorithm enjoys good interpretability. In §3.2, we will describe the SRA, which mimics the prototype algorithm, and hence shares its interpretability, but is free from the aforementioned problems.

3.2 Streaming regression algorithm

Inspired by the concept-adapting very fast decision tree algorithm of Hulten et al. (2001) (see §1.3.1.2), we modify the prototype algorithm in §3.1 into a more useful algorithm, the SRA. Recall that the classification algorithm of Hulten et al. (2001) constructs alternative sub-trees that may be used to update the tree classifier. For this purpose, whether a sub-tree is used is determined by a data-driven criterion. Here we further exploit this idea to solve the computational issue of the prototype algorithm.

The SRA computes a small (compared to the prototype algorithm in §3.1) ensemble of regression estimators while periodically checking if it is necessary to construct an alternative ensemble according to a recursive cross-validation (RCV) criterion (so-called because it can be recursively computed). When constructed, the alternative ensemble is not immediately used to replace the original ensemble, but will continue to be updated with some observations. Then, if some estimator from the alternative ensemble admits lower RCV score, we substitute the current ensemble with the alternative one. We refer to this substitution as an ensemble substitution. However, if using estimators in the alternative ensemble does not reduce the RCV score, we delete the alternative ensemble and check again later if an alternative ensemble is needed.

As discussed in §3.1.3, due to exponential decay (through γ) of the influence of past data, we

can approximate $\tilde{m}(\cdot, t|\gamma, h)$ by some estimator $\check{m}(\cdot, t|\gamma, h)$ (this argument will be made rigorous in Proposition 3.1 in §3.3). Therefore, the modified algorithm can be seen as an imitation of the prototype algorithm without the problems discussed in §3.1.3.

3.2.1 Definition of SRA

In this section we iteratively define the SRA, which sequentially processes the data $\{(X_i, Y_i)\}_{i=1,2,\ldots}$ as they arrive at times $\{t_i\}_{i=1,2,\ldots}$ defined at (1.1). The goal here is to estimate the regression function $m(\cdot, t)$ at $t = t_{N_0+1}, t_{N_0+2}, \ldots$, where N_0, a positive integer, is the number of data points we use to initialise our algorithm. During the initialisation stage, we will compute estimates and cross-validation scores for a range of (γ, h) values, but we do not choose the best (γ, h) among them. Indeed, N_0 is the minimum number of points needed to really start our algorithm (before that, we have too few observations to be able to select a meaningful value of (γ, h)).

3.2.1.1 Initialisation

We first choose an integer ν by hand, which determines how often we will check if it is necessary to construct an alternative ensemble. Ideally, it should be the largest number such that we do not expect the data stream to change significantly during a time period of $\nu\Delta t$ with Δt defined at (1.1). Hence the appropriate value of ν depends on the specific data set at hand. For example, for some meteorological data sets, it is often reasonable to view observations arriving in a 20-minute time interval as nearly stationary (Hall and Patil, 1994). We also choose the first candidate set $I^1_{\gamma,h} = I^1_\gamma \times I^1_h$ as an equidistant grid on some rectangular range

$$\square_1 = [\gamma_0/L, L\gamma_0] \times [h_0/L, Lh_0], \tag{3.5}$$

where $L > 1$ is some constant and γ_0 and h_0 are some initial γ and h values, all chosen by hand. Since the candidate set will be dynamically adjusted in the future, different from the prototype algorithm in §3.1, it needs only contain a relatively small number of (γ, h) values.

Next we use the first $N_0 = 2\nu$ observations to create our initial ensemble and the initial cross-validation scores. We do not check for alternative ensembles nor attempt to minimise the cross-validation score yet, as too few data have been observed at this initialisation stage.

For the first ν data, i.e. for $k = 1, \ldots, \nu$, we compute the ensemble $\mathscr{E}_{t_k} = \{\check{m}(\cdot, t_k|\gamma, h); \check{r}(\cdot, t_k|\gamma, h), \check{f}(\cdot, t_k|\gamma, h) : (\gamma, h) \in I^1_{\gamma,h}\}$ iteratively as follows. Let $\mathscr{E}_{t_1} = \{\check{m}(\cdot, t_1|\gamma, h); \check{r}(\cdot, t_1|\gamma, h), \check{f}(\cdot, t_1|\gamma, h) : (\gamma, h) \in I^1_{\gamma,h}\}$, where

$$\check{m}(x, t_1|\gamma, h) = \frac{\check{r}(x, t_1|\gamma, h)}{\check{f}(x, t_1|\gamma, h)} \tag{3.6}$$

with

$$\check{r}(x, t_1|\gamma, h) = K_h(x - X_1)Y_1 \quad \text{and} \quad \check{f}(x, t_1|\gamma, h) = K_h(x - X_1)\,.$$

Then, for $k = 2, \ldots, \nu$, let $\mathscr{E}_{t_k} = \{\check{m}(\cdot, t_k|\gamma, h); \check{r}(\cdot, t_k|\gamma, h), \check{f}(\cdot, t_k|\gamma, h) : (\gamma, h) \in I^1_{\gamma,h}\}$, where $\check{m}(x, t_k|\gamma, h) = \check{r}(x, t_k|\gamma, h)/\check{f}(x, t_k|\gamma, h)$ with

$$\begin{aligned} \check{r}(x, t_k|\gamma, h) &= (1 - \gamma)\check{r}(x, t_{k-1}|\gamma, h) + \gamma K_h(x - X_k)Y_k\,, \\ \check{f}(x, t|\gamma, h) &= (1 - \gamma)\check{f}(x, t_{k-1}|\gamma, h) + \gamma K_h(x - X_k)\,. \end{aligned} \tag{3.7}$$

Next, for $k = \nu + 1, \ldots, N_0$, we continue computing $\mathscr{E}_{t_k}$ by updating $\mathscr{E}_{t_{k-1}}$ as at (3.7). We also start computing the RCV scores iteratively as follows. At time $t_{\nu+1}$, let

$$\mathrm{RCV}_{t_{\nu+1}}(\gamma, h) = \left\{\check{m}(X_{\nu+1}, t_\nu|\gamma, h) - Y_{\nu+1}\right\}^2. \tag{3.8}$$

For $k = \nu + 2, \ldots, N_0$, let

$$\mathrm{RCV}_{t_k}(\gamma, h) = \mathrm{RCV}_{t_{k-1}}(\gamma, h) + \left\{\check{m}(X_k, t_{k-1}|\gamma, h) - Y_k\right\}^2. \tag{3.9}$$

Note that we start computing RCV only after N_0 data have been observed, because $\check{m}(\cdot, t_k|\gamma, h)$ for $k < N_0$ is not expected to be a good estimator of $m(\cdot, t_k)$, as it is constructed from too few data. Therefore, we do not use such poor estimators to compute RCV, which might

otherwise perform too poorly.

3.2.1.2 Routine updates and checks

After the initialisation, for $k = N_0 + 1, N_0 + 2, \ldots$, we continue computing iteratively $\check{m}(\cdot, t_k|\gamma, h)$ in $\mathscr{E}_{t_k}$ and the RCV scores RCV_{t_k} and, for each k, dynamically select the best estimator by minimising the RCV criterion with respect to γ and h. In addition, we will check regularly (every ν observations) whether the current candidate set for (γ, h), say $I^{\ell}_{\gamma,h}$, is still appropriate. Hence the updates and checks will be the routine of our algorithm. If there is evidence showing that $I^{\ell}_{\gamma,h}$ might no longer be appropriate, we start computing an alternative ensemble, corresponding to an alternative set, and the associated RCV scores. If there is evidence confirming that the new ensemble and candidate set are better than the current ones, we replace the latter by the former. If that happens, we call it an ensemble substitution and use $I^{\ell+1}_{\gamma,h}$ to denote the new candidate set. Before the first ensemble substitution, the routine updates are essentially an application of the prototype algorithm in §3.1, except that here the candidate set $I^{\ell}_{\gamma,h}$ is significantly smaller.

Recall that, after the initialisation step, we have computed $\mathscr{E}_{t_k}$ and RCV_{t_k} for $k = 1, \ldots, N_0$. For $k = N_0+1, \ldots, N_0+\nu$, we initialise the routine updates and, at time $t_{N_0+\nu}$ we will check for the first time if it is necessary to construct an alternative ensemble. These procedures are defined as follows. The subsequent updates and checks will be discussed towards the end of §3.2.1.3.

Routine updates. For each $k = N_0 + 1, \ldots, N_0 + \nu$, we do the following updates.

(R1) For all $(\gamma, h) \in I^1_{\gamma,h}$, update $\mathrm{RCV}_{t_{k-1}}(\gamma, h)$ to $\mathrm{RCV}_{t_k}(\gamma, h)$ by (3.9).

(R2) Compute the ensemble $\mathscr{E}_{t_k} = \{\check{m}(\cdot, t_k|\gamma, h); \check{r}(\cdot, t_k|\gamma, h), \check{f}(\cdot, t_k|\gamma, h) : (\gamma, h) \in I^1_{\gamma,h}\}$, where $\check{m}(\cdot, t_k|\gamma, h) = \check{r}(\cdot, t_k|\gamma, h)/\check{f}(\cdot, t_k|\gamma, h)$ with $\check{r}(\cdot, t_k|\gamma, h)$ and $\check{f}(\cdot, t_k|\gamma, h)$ updated from time t_{k-1} using (3.7).

(R3) Compute

$$(\gamma^{CV}_k, h^{CV}_k) = \underset{(\gamma,h)\in I^1_{\gamma,h}}{\arg\min}\ \mathrm{RCV}_{t_k}(\gamma, h) \tag{3.10}$$

and then select $\check{m}(\cdot, t_k|\gamma_k^{CV}, h_k^{CV})$ as the estimator of $m(\cdot, t_k)$.

From (R1)–(R3), we can see that, for $(\gamma, h) \in I^1_{\gamma,h}$, we have $\check{m}(\cdot, t_k|\gamma, h) = \tilde{m}(\cdot, t_k|\gamma, h)$, where $\tilde{m}$ is defined at (3.1), and $\text{RCV}_{t_k}(\gamma, h)$ is equal to $\text{CV}^0_{t_k}(\gamma, h)$ taking there $b_{t_k} = k - \nu$, where $\text{CV}^0_{t_k}(\gamma, h)$ is defined at (3.4). Indeed, as we mentioned above, before the first ensemble substitution the routine updates are the same as the prototype algorithm. After that, these equivalences are no longer valid, but they will still hold in an asymptotic sense (see Proposition 3.1).

Routine checks. At time $t_{N_0+\nu}$, we check if

$$(\gamma^{CV}_{N_0+\nu}, h^{CV}_{N_0+\nu}) \in \partial I^1_{\gamma,h}, \tag{3.11}$$

where

$$\partial I^1_{\gamma,h} = \big\{(\gamma, h) \in I^1_{\gamma,h} : \gamma = \min(I^1_\gamma),\ \text{or}\ \gamma = \max(I^1_\gamma),\ \text{or}\ h = \min(I^1_h),\ \text{or}\ h = \max(I^1_h)\big\} \tag{3.12}$$

denotes the boundary of $I^1_{\gamma,h}$.

If (3.11) holds, then this indicates that $I^1_{\gamma,h}$ may no longer include the right range of (γ, h) values. Therefore we start constructing an alternative grid $I^{\text{alt}}_{\gamma,h}$ corresponding to a different range. However, since the data evolve smoothly in time, we do not expect the appropriate range for (γ, h) to change drastically. Hence it is often enough to only slightly shift the range $\square_1$ of $I^1_{\gamma,h}$, defined at (3.5), and take the shifted range as the range of $I^{\text{alt}}_{\gamma,h}$.

Instead of criterion (3.11), another option is that we start to construct an alternative ensemble once $(\gamma^{CV}_{N_0+\nu}, h^{CV}_{N_0+\nu})$ is close to the boundary $\partial I^1_{\gamma,h}$, rather than waiting until it hits the boundary. This is because, when (3.11) holds, $I^1_{\gamma,h}$ may have already become inappropriate. We shall investigate the latter approach in future work.

However, criterion (3.11) is often not a strong enough evidence showing that $I^1_{\gamma,h}$ has become inappropriate, since it is only determined by $(\gamma^{CV}_{N_0+\nu}, h^{CV}_{N_0+\nu})$, which is chosen at a single time point $t_{N_0+\nu}$. Hence we should not discard $I^1_{\gamma,h}$ yet immediately after (3.11) is satisfied. Instead, if

(3.11) holds, all we do is to start constructing an alternative candidate set $I^{\text{alt}}_{\gamma,h}$ and start computing a corresponding alternative ensemble $\mathscr{E}^{\text{alt}}_{t_{N_0+\nu}}$ using data arriving after $t_{N_0+\nu}$, but we do not do anything with it just yet. Analogous to the classification tree algorithm of Hulten et al. (2001) (see page 124), we only replace $I^1_{\gamma,h}$ by $I^{\text{alt}}_{\gamma,h}$ later, when there is stronger evidence showing that the latter is better. See §3.2.1.3 for details.

If (3.11) does not hold, then, for $k = N_0 + \nu + 1, N_0 + \nu + 2, \ldots$, we continue the routine updates (R1)–(R3) and check after every ν data points if (3.11) holds for those data, i.e. for $i = 2, 3, \ldots$ if $(\gamma^{CV}_{N_0+i\nu}, h^{CV}_{N_0+i\nu}) \in \partial I^1_{\gamma,h}$ holds, until it does hold, say at time $t_{N_0+\lambda\nu}$ for some $\lambda \in \{2, 3, \ldots\}$. Then we start constructing an alternative candidate set $I^{\text{alt}}_{\gamma,h}$ and an alternative ensemble, as described above.

The motivation for checking if we should create an alternative ensemble for very ν data, instead of for every data point, is that we want to reduce the number of false alarms, since $I^1_{\gamma,h}$ may still be appropriate even when criterion (3.11) is satisfied. A similar idea is used in some ensemble learning works on streaming data classification. For example, in Kolter and Maloof (2007), the authors used an ensemble of base learners, each computed from a different block of past data. To classify a new data point, they used a weighted vote of base learners. They checked once in a while if some of the base learners have become inappropriate. Inappropriate base learners were discarded whilst those trained from more recent data points were added into the ensemble.

3.2.1.3 Alternative ensemble and ensemble substitution

If, at time $t_{N_0+\lambda\nu}$ for some $\lambda = 1, 2, \ldots$, criterion (3.11) is satisfied with $N_0 + \nu$ there replaced by $N_0 + \lambda\nu$, then we construct an alternative set $I^{\text{alt}}_{\gamma,h}$ and an alternative ensemble $\mathscr{E}^{\text{alt}}_{t_{N_0+\lambda\nu}}$ as follows.

Initialising alternative candidate set and ensemble. At time $t_{N_0+\lambda\nu}$, we initialise the alternative set $I^{\text{alt}}_{\gamma,h} = I^{\text{alt}}_{\gamma} \times I^{\text{alt}}_{h}$ by shifting the range of γ and h around $\gamma^{CV}_{N_0+\lambda\nu}$ and $h^{CV}_{N_0+\lambda\nu}$, respectively.

Specifically we take the range

$$\square^{\text{alt}} = [\gamma^{CV}_{N_0+\lambda\nu}/L, L\gamma^{CV}_{N_0+\lambda\nu}] \times [h^{CV}_{N_0+\lambda\nu}/L, Lh^{CV}_{N_0+\lambda\nu}], \tag{3.13}$$

where $L > 1$ is some constant defined under (3.5). Then I^{alt}_{γ} and I^{alt}_{h} are taken as equidistant grids on $[\gamma^{CV}_{N_0+\lambda\nu}/L, L\gamma^{CV}_{N_0+\lambda\nu}]$ and $[h^{CV}_{N_0+\lambda\nu}/L, Lh^{CV}_{N_0+\lambda\nu}]$.

Remark 3.1. An alternative way of constructing $\square^{\text{alt}}$ is, instead of consulting $(\gamma^{CV}_{N_0+\lambda\nu}, h^{CV}_{N_0+\lambda\nu})$ chosen by the RCV at just one time point, to use more than one (γ^{CV}, h^{CV}) values to determine $\square^{\text{alt}}$. Indeed, this alternative approach is reasonable and worth experimenting in future work. However, we observed in the simulation studies in §4.1.2 that the way we construct $\square^{\text{alt}}$ at (3.13) already worked reasonably well. This might be due to the fact that, since the $(\gamma^{CV}_i, h^{CV}_i)$'s with similar i values are chosen by (3.10) using nearly the same data points, these (γ^{CV}, h^{CV}) are strongly correlated, so that even if we use just one particular $(\gamma^{CV}_{N_0+\lambda\nu}, h^{CV}_{N_0+\lambda\nu})$, it already contains information from the (γ^{CV}, h^{CV}) values obtained at some nearby time points.

Next we initialise an alternative ensemble. Since we want to mimic the prototype algorithm at §3.1 for the new grid $I^{\text{alt}}_{\gamma,h}$, ideally we would compute the alternative ensemble $\mathscr{E}^{\text{alt}}_{t_{N_0+\lambda\nu}}$ containing $\tilde{m}(\cdot, t_{N_0+\lambda\nu}|\gamma, h)$ for $(\gamma, h) \in I^{\text{alt}}_{\gamma,h}$, where $\tilde{m}$ is defined at (3.1). However, computing $\tilde{m}(\cdot, t_{N_0+\lambda\nu}|\gamma, h)$ would require storing all data up to time $t_{N_0+\lambda\nu}$. Indeed, $I^{\text{alt}}_{\gamma,h}$ does not contain the same (γ, h) as $I^{\ell}_{\gamma,h}$, so the computations need to start from (X_1, Y_1). However, we cannot store all past data in the streaming data setting (recall that the data stream is never-ending). Hence the alternative ensemble should be constructed in such a way that access to the past data is minimised.

With the above constraint in mind, we construct an alternative ensemble which does not require access to any past data point. Let $\mathscr{E}^{\text{alt}}_{t_{N_0+\lambda\nu}} = \{\check{m}_{\text{alt}}(\cdot, t_{N_0+\lambda\nu}|\gamma, h); \check{r}_{\text{alt}}(\cdot, t_{N_0+\lambda\nu}|\gamma, h), \check{f}_{\text{alt}}(\cdot, t_{N_0+\lambda\nu}|\gamma, h) : (\gamma, h) \in I^{\text{alt}}_{\gamma,h}\}$, where $\check{m}_{\text{alt}}(\cdot, t_{N_0+\lambda\nu}|\gamma, h) = \check{m}(\cdot, t_{N_0+\lambda\nu}|\gamma^{CV}_{N_0+\lambda\nu}, h^{CV}_{N_0+\lambda\nu})$ for all $(\gamma, h) \in I^{\text{alt}}_{\gamma,h}$. That is, the initial $\check{m}_{\text{alt}}(\cdot, t_{N_0+\lambda\nu}|\gamma, h)$'s are all equal to the estimator $\check{m}$ selected at time $t_{N_0+\lambda\nu}$ from the

ensemble $\mathscr{E}_{t_{N_0+\lambda\nu}}$. More formally, for $(\gamma, h) \in I^{\text{alt}}_{\gamma,h}$, let

$$\check{m}_{\text{alt}}(x, t_{N_0+\lambda\nu}|\gamma, h) = \frac{\check{r}_{\text{alt}}(x, t_{N_0+\lambda\nu}|\gamma, h)}{\check{f}_{\text{alt}}(x, t_{N_0+\lambda\nu}|\gamma, h)} \tag{3.14}$$

with

$$\begin{aligned} \check{r}_{\text{alt}}(x, t_{N_0+\lambda\nu}|\gamma, h) &= \check{r}(x, t_{N_0+\lambda\nu}|\gamma^{CV}_{N_0+\lambda\nu}, h^{CV}_{N_0+\lambda\nu}) \,, \\ \check{f}_{\text{alt}}(x, t_{N_0+\lambda\nu}|\gamma, h) &= \check{f}(x, t_{N_0+\lambda\nu}|\gamma^{CV}_{N_0+\lambda\nu}, h^{CV}_{N_0+\lambda\nu}) \,. \end{aligned} \tag{3.15}$$

Letting $\check{m}_{\text{alt}}(\cdot, t_{N_0+\lambda\nu}|\gamma, h) = \check{m}(\cdot, t_{N_0+\lambda\nu}|\gamma^{CV}_{N_0+\lambda\nu}, h^{CV}_{N_0+\lambda\nu})$ for all $(\gamma, h) \in I^{\text{alt}}_{\gamma,h}$ as in (3.14) and (3.15) facilitates the derivation of some simple formulas for both $\check{m}$ and $\check{m}_{\text{alt}}$, as we will see in §3.2.2, making their theoretical analysis easier.

After this initialisation step, for $k = N_0 + \lambda\nu + 1, N_0 + \lambda\nu + 2, \ldots$, $\check{m}_{\text{alt}}(\cdot, t_k|\gamma, h)$ will be updated using $(\gamma, h) \in I^{\text{alt}}_{\gamma,h}$ (see the next two paragraphs). Recall from the first paragraph under Remark 3.1 that, ideally, we would have computed an alternative ensemble containing the $\tilde{m}(\cdot, t_k|\gamma, h)$'s for $(\gamma, h) \in I^{\text{alt}}_{\gamma,h}$, if we had access to all past data. Now, by updating the $\check{m}_{\text{alt}}(\cdot, t_k|\gamma, h)$'s in the above way, after some time, $\check{m}_{\text{alt}}(\cdot, t_k|\gamma, h)$ will be sufficiently close to the $\tilde{m}(\cdot, t_k|\gamma, h)$ we would have computed, since 'old' data would have contributed in a minimal way to $\tilde{m}$.

Updating alternative estimators. After the initialisation of the alternative ensemble $\mathscr{E}^{\text{alt}}_{t_{N_0+\lambda\nu}}$, we experience a 'transition period' for another 2ν data, i.e. for $k = N_0+\lambda\nu+1, \ldots, N_0+(\lambda+2)\nu$. During this transition period we continue the routine updates (R1)–(R3) with the ensemble $\mathscr{E}_{t_k} = \{\check{m}(\cdot, t_k|\gamma, h); \check{r}(\cdot, t_k|\gamma, h), \check{f}(\cdot, t_k|\gamma, h) : (\gamma, h) \in I^1_{\gamma,h}\}$. That is, we still compute the ensemble $\mathscr{E}_{t_k}$ and the RCV scores RCV_{t_k} and adaptively select $\check{m}(\cdot, t_k|\gamma^{CV}_k, h^{CV}_k)$ from $\mathscr{E}_{t_k}$ at each time t_k. However, we do not do routine checks since we have already decided, at time $t_{N_0+\lambda\nu}$, to start constructing an alternative ensemble. In addition to the routine updates, we also update the alternative ensembles and the alternative RCV scores as follows.

For the first ν data that arise after the initialisation of $\mathscr{E}^{\text{alt}}_{t_{N_0+\lambda\nu}}$, i.e. for $k = N_0 + \lambda\nu + 1, \ldots, N_0 + (\lambda + 1)\nu$, we compute the alternative ensemble $\mathscr{E}^{\text{alt}}_{t_k} = \{\check{m}_{\text{alt}}(\cdot, t_k|\gamma, h); \check{r}_{\text{alt}}(\cdot, t_k|\gamma, h), \check{f}_{\text{alt}}(\cdot, t_k|\gamma, h) : (\gamma, h) \in I^{\text{alt}}_{\gamma,h}\}$, where $\check{m}_{\text{alt}}(\cdot, t_k|\gamma, h) =$

$\check{r}_{\text{alt}}(\cdot, t_k|\gamma, h)/\check{f}_{\text{alt}}(\cdot, t_k|\gamma, h)$ with

$$\begin{aligned}\check{r}_{\text{alt}}(x, t_k|\gamma, h) &= (1-\gamma)\check{r}_{\text{alt}}(x, t_{k-1}|\gamma, h) + \gamma K_h(x - X_k)Y_k\,,\\ \check{f}_{\text{alt}}(x, t_k|\gamma, h) &= (1-\gamma)\check{f}_{\text{alt}}(x, t_{k-1}|\gamma, h) + \gamma K_h(x - X_k)\,.\end{aligned} \tag{3.16}$$

As at the initialisation stage at §3.2.1.1, we wait for more data to start computing alternative RCV scores. Specifically, for the next ν data, i.e. for $k = N_0+(\lambda+1)\nu+1, \ldots, N_0+(\lambda+2)\nu$, we calculate, for all $(\gamma, h) \in I^{\text{alt}}_{\gamma,h}$, the alternative RCV scores for the estimators in $\mathscr{E}^{\text{alt}}_{t_k}$ using

$$\text{RCV}^{\text{alt}}_{t_k}(\gamma, h) = \begin{cases} \left\{\check{m}_{\text{alt}}(X_k, t_{k-1}|\gamma, h) - Y_k\right\}^2\,, \ k = N_0 + (\lambda+1)\nu + 1\,,\\ \text{RCV}^{\text{alt}}_{t_{k-1}}(\gamma, h) + \{\check{m}_{\text{alt}}(X_k, t_{k-1}|\gamma, h) - Y_k\}^2\,,\\ \qquad\qquad k = N_0 + (\lambda+1)\nu + 2, \ldots, N_0 + (\lambda+2)\nu\,. \end{cases} \tag{3.17}$$

Then we compute $\mathscr{E}^{\text{alt}}_{t_k}$ by updating $\check{r}_{\text{alt}}(\cdot, t_{k-1}|\gamma, h)$, $\check{f}_{\text{alt}}(\cdot, t_{k-1}|\gamma, h)$ and $\check{m}_{\text{alt}}(\cdot, t_{k-1}|\gamma, h)$ to $\check{r}_{\text{alt}}(\cdot, t_k|\gamma, h)$, $\check{f}_{\text{alt}}(\cdot, t_k|\gamma, h)$ and $\check{m}_{\text{alt}}(\cdot, t_k|\gamma, h)$ using (3.16). As for RCV_t in §3.2.1.1, we do not use $\text{RCV}^{\text{alt}}_{t_k}$ to select (γ, h) at this stage since it is computed using too few data.

Now we have updated the alternative estimators $\check{m}_{\text{alt}}$ using 2ν data and updated the alternative RCV scores $\text{RCV}^{\text{alt}}_{t_k}$ using ν data. This is when we decide if we replace $I^1_{\gamma,h}$ and ensemble $\mathscr{E}_{t_k}$ with $I^{\text{alt}}_{\gamma,h}$ and alternative ensemble $\mathscr{E}^{\text{alt}}_{t_k}$ for future t_k's, using the alternative RCV scores. This is done as follows.

Criterion for ensemble substitution. At $t_{N_0+(\lambda+2)\nu}$, let

$$(\gamma^{\text{alt}}, h^{\text{alt}}) = \underset{(\gamma,h)\in I^{\text{alt}}_{\gamma,h}}{\arg\min}\ \text{RCV}^{\text{alt}}_{t_{N_0+(\lambda+2)\nu}}(\gamma, h). \tag{3.18}$$

Then we check whether the newly selected $(\gamma^{\text{alt}}, h^{\text{alt}})$ still lies in $\square_1$ (recall that $I^1_{\gamma,h}$ is a grid on $\square_1$). If this is the case, then there is not enough evidence supporting ensemble substitution, since the RCV still selects some $(\gamma^{\text{alt}}, h^{\text{alt}}) \in \square^1$, so that continuing to use $I^1_{\gamma,h}$ may still be appropriate. Hence we delete $\mathscr{E}^{\text{alt}}_{t_{N_0+(\lambda+2)\nu}}$ and $I^{\text{alt}}_{\gamma,h}$ and continue the routine updates and checks

using $I^1_{\gamma,h}$, RCV_t and $\mathscr{E}_t$ as in §3.2.1.2.

On the other hand, if

$$(\gamma^{\text{alt}}, h^{\text{alt}}) \in \square^{\text{alt}} \backslash \square^k, \tag{3.19}$$

where $\square^{\text{alt}}$, defined at (3.13), denotes the range of $I^{\text{alt}}_{\gamma,h}$, then it implies that the RCV selects some $(\gamma^{\text{alt}}, h^{\text{alt}})$ falling outside $\square^1$, so that that using $I^1_{\gamma,h}$ may no longer be appropriate. Recall that the way we construct $\square^{\text{alt}}$ at (3.13) guarantees that $\square^1 \cap \square^{\text{alt}} \neq \varnothing$. When (3.19) holds, we do the ensemble substitution as follows.

Ensemble substitution.

(ES1) Let $N_1 = N_0 + (\lambda + 2)\nu$ denote the number of observations up to the first ensemble substitution.

(ES2) Use the new candidate set $I^2_{\gamma,h} = I^{\text{alt}}_{\gamma,h}$ instead of $I^1_{\gamma,h}$. Let $\check{m}(\cdot, t_{N_1}|\gamma, h) = \check{m}_{\text{alt}}(\cdot, t_{N_1}|\gamma, h)$ for all $(\gamma, h) \in I^2_{\gamma,h}$. Let $\mathscr{E}_{t_{N_1}} = \{\check{m}(\cdot, t_{N_1}|\gamma, h); \check{r}(\cdot, t_{N_1}|\gamma, h), \check{f}(\cdot, t_{N_1}|\gamma, h) : (\gamma, h) \in I^2_{\gamma,h}\}$.

(ES3) Let $\text{RCV}_{t_{N_1}}(\gamma, h) = \text{RCV}^{\text{alt}}_{t_{N_1}}(\gamma, h)$ for all $(\gamma, h) \in I^2_{\gamma,h}$.

(ES4) Continue routine updates as in (R1)–(R3) at page 127 and routine checks as at (3.11), except that N_0 and $I^1_{\gamma,h}$ there are replaced by N_1 and $I^2_{\gamma,h}$.

Remark 3.2. Instead of considering whether $(\gamma^{\text{alt}}, h^{\text{alt}})$ lies in $\square_1$ or not as at (3.19), where $(\gamma^{\text{alt}}, h^{\text{alt}})$ is chosen at (3.18) by minimising the alternative RCV scores at only one time point $t_{N_0+(\lambda+2)\nu}$, we may use more than one $(\gamma^{\text{alt}}, h^{\text{alt}})$ values to determine if we do an ensemble substitution as in (ES1)–(ES4). We shall investigate the latter approach in future work. However, since different $(\gamma^{\text{alt}}, h^{\text{alt}})$ values chosen at consecutive time points would be strongly correlated (the $\text{RCV}^{\text{alt}}_{t_k}$'s with similar k values are computed using nearly the same data points), the benefit of using more $(\gamma^{\text{alt}}, h^{\text{alt}})$ values may be limited.

With the above procedure we have defined the initialisation of the SRA and how to proceed till the first ensemble substitution. When there have been $\ell - 1$ ensemble substitutions for some

$\ell = 2, 3, \ldots$, we proceed analogously till the the ℓ-th ensemble substitution. That is, replacing N_0, N_1 and $I^1_{\gamma,h}$ with $N_{\ell-1}$, N_ℓ and $I^\ell_{\gamma,h}$, we do the following: routine updates as in (R1)–(R3) at page 127, routine checks as at (3.11), initialisation of the alternative ensemble as under (3.13), initialisation of the alternative estimators as at (3.14), updating the alternative estimators as at (3.16), updating the alternative RCV scores as at (3.17), ensemble substitution as in (ES1)–(ES4) at page 133 if (3.19) holds.

Hence an iterative procedure for the SRA has been defined. See Figure 3.1 at page 144 for a flowchart illustrating the SRA and Appendix 3.F for the pseudocode.

3.2.2 Summary of the SRA

In §3.2, the estimator $\check{m}$ and the RCV score were iteratively defined at $t_1, t_2, \ldots$ For the convenience of theoretical analysis, here we present their definitions for a generic time $t \in \mathbb{R}_+$ satisfying

$$n_t \geq N_0 + 1, \tag{3.20}$$

where n_t is defined at (1.2) and $N_0 = 2\nu$ is the number of observations used at the initialisation step (see §3.2.1.1). That is, we only consider $\check{m}(\cdot, t|\gamma, h)$ and RCV_t after the initialisation. This is because, recalling from §3.2.1.1, the initialisation stage is only a short warm-up period for the SRA, during which we do not use the RCV to select any (γ, h).

At time $t \in \mathbb{R}_+$, let

$$\ell_t \text{ denote the index of the candidate set in use at time } t, \tag{3.21}$$

so that the candidate set at time t is $I^{\ell_t}_{\gamma,h}$ and the ensemble at time is equal to $\mathscr{E}^\circ_t = \{\check{m}(\cdot, t|\gamma, h); \check{r}(\cdot, t|\gamma, h), \check{f}(\cdot, t|\gamma, h) : (\gamma, h) \in I^{\ell_t}_{\gamma,h}\}$. Using this notation, there have been $\ell_t - 1$ ensemble substitutions up to time t and the ℓ_t-th ensemble substitution has not taken place yet. Recall that $N_0 = 2\nu$ and N_j, for $j = 1, 2, \ldots$, denotes the number of data observed up to the j-th ensemble substitution.

3.2.2.1 Definition of $\check{m}(x,t|\gamma,h)$

Recalling the definition of $\hat{m}$ at (1.9), in this section we will show that

$$\check{m}(x,t|\gamma,h) = \hat{m}(x,t|\check{\boldsymbol{\gamma}}_t, \check{\boldsymbol{h}}_t)\,,\ (\gamma,h) \in I^{\ell_t}_{\gamma,h}\,, \tag{3.22}$$

where $\check{\boldsymbol{\gamma}}_t = \{\check{\gamma}_i\}_{i=1,\ldots,n_t}$ and $\check{\boldsymbol{h}}_t = \{\check{h}_i\}_{i=1,\ldots,n_t}$ are some specific choices of γ and h values. For this, we need to see what specific choices of γ and h $\check{m}(x,t|\gamma,h)$ has taken.

First we discuss the case $\ell_t = 1$, i.e. no ensemble substitution has happened up to time t. In this case, recall from §3.2.1.1 and (R2) that $\check{m}$ is initialised at time t_1 by (3.6) and then we use the same $(\gamma,h) \in I^1_{\gamma,h}$ to update $\check{m}(x,t_k|\gamma,h)$ by (3.7), for $k = 2,\ldots,n_t$. This implies that, in the case $\ell_t = 1$, we have

$$(\check{\gamma}_i, \check{h}_i) = (\gamma,h) \text{ for } i = 1,\ldots,n_t\,. \tag{3.23}$$

In the case $\ell_t \geq 2$, for $\ell = 1,\ldots,\ell_t - 1$, $\check{m}(x,t|\gamma,h)$ uses the same (γ,h) value $(\gamma^{CV}_{N_\ell - 2\nu}, h^{CV}_{N_\ell - 2\nu})$, selected by the RCV at time $t_{N_\ell - 2\nu}$, for data arriving between times $t_{N_{\ell-1}-2\nu}$ and $t_{N_\ell - 2\nu}$. For observations arriving after $t_{N_{\ell_t-1}-2\nu}$, $\check{m}(x,t|\gamma,h)$ uses (γ,h). That is,

$$(\check{\gamma}_i, \check{h}_i) = \begin{cases} (\gamma^{CV}_{N_1-2\nu}, h^{CV}_{N_1-2\nu})\,, & \text{for } i = 1,\ldots,N_1 - 2\nu\,, \\ (\gamma^{CV}_{N_2-2\nu}, h^{CV}_{N_2-2\nu})\,, & \text{for } i = N_1 - 2\nu + 1,\ldots,N_2 - 2\nu\,, \\ \qquad\vdots & \\ (\gamma,h)\,, & \text{for } i = N_{\ell_t-1} - 2\nu + 1,\ldots,n_t\,, \end{cases} \tag{3.24}$$

Hence by (3.23) and (3.24) we have obtained the definition of the $(\check{\gamma}_i, \check{h}_i)$'s for both the cases $\ell_t = 1$ and $\ell_t \geq 2$.

Thus by (3.22), (3.23) and (3.24) we have shown that $\check{m}$ can be written as a special case of $\hat{m}$ at (1.9), taking $\{(\check{\gamma}_i, \check{h}_i)\}_{i=1,\ldots,n_t}$ defined at (3.24). That is, for a given $t \in \mathbb{R}_+$, $\check{m}(x,t|\gamma,h)$ uses blockwise defined (γ,h) values for data arriving up to time t. In Proposition 3.1 we will

see that, as long as ν is not too small, this guarantees that $\check{m}(x, t|\gamma, h)$ is close to $\tilde{m}(x, t|\gamma, h)$ at (3.1). That is, although $\check{m}(x, t|\gamma, h)$ takes some different (γ, h) values for data arriving up to time $t_{N_{\ell_t-1}}$, those (γ, h) values have very little influence on the estimator at time t.

3.2.2.2 Definitions of RCV_t, $\mathrm{RCV}_t^{\mathrm{alt}}$ and $\check{m}_t^{\mathrm{alt}}$

To deduce the definition of RCV_t, first note that, in the case $\ell_t = 1$, we initialise the RCV scores using (3.8) and update them using (3.9). This implies that, in the case $\ell_t = 1$, we have $\mathrm{RCV}_t(\gamma, h) = \sum_{i=\nu}^{n_t-1}\{\check{m}(X_{i+1}, t_i|\gamma, h) - Y_{i+1}\}^2$.

In the case $\ell_t \geq 2$, recall from §3.2.1.3 that the last ensemble substitution happens at time $t_{N_{\ell_t-1}}$. For the last ν data arriving up to this ensemble substitution, we update the alternative RCV scores by (3.17). That is, for $k = N_{\ell_t-1} - \nu + 1, \ldots, N_{\ell_t-1}$, we have, for $(\gamma, h) \in I_{\gamma,h}^{\mathrm{alt}}$,

$$\mathrm{RCV}_{t_k}^{\mathrm{alt}}(\gamma, h) = \sum_{i=N_{\ell_t-1}-\nu}^{k-1} \{\check{m}_{\mathrm{alt}}(X_{i+1}, t_i|\gamma, h) - Y_{i+1}\}^2 ,$$

where

$$\check{m}_{\mathrm{alt}}(x, t_i|\gamma, h) = \hat{m}(x, t_i|\boldsymbol{\gamma}_{t_i}^{\mathrm{alt}}, \boldsymbol{h}_{t_i}^{\mathrm{alt}}) , \tag{3.25}$$

with $\hat{m}$ defined at (1.9) and $\boldsymbol{\gamma}_{t_i}^{\mathrm{alt}} = \{\gamma_j^{\mathrm{alt}}\}_{j=1,\ldots,i}$, $\boldsymbol{h}_{t_i}^{\mathrm{alt}} = \{h_j^{\mathrm{alt}}\}_{j=1,\ldots,i}$ denoting some specific choices of γ and h values (we will derive the definition of the $(\gamma_j^{\mathrm{alt}}, h_j^{\mathrm{alt}})$'s towards the end of this section). Then, at time $t_{N_{\ell_t-1}}$, these alternative scores are redefined as the RCV scores (see (ES3) at page 133). After time $t_{N_{\ell_t-1}}$, we continue updating the RCV scores using (3.9).

We summarise the above derivations into the following formula: for $(\gamma, h) \in I_{\gamma,h}^{\ell_t}$,

$$\mathrm{RCV}_t(\gamma, h) = \begin{cases} \sum_{i=\nu}^{n_t-1} \{\check{m}(X_{i+1}, t_i|\gamma, h) - Y_{i+1}\}^2, & \text{if } \ell_t = 1\,, \\ \sum_{i=N_{\ell_t-1}-\nu}^{N_{\ell_t-1}-1} \{\check{m}_{\text{alt}}(X_{i+1}, t_i|\gamma, h) - Y_{i+1}\}^2 & \\ \qquad + \sum_{i=N_{\ell_t-1}}^{n_t-1} \{\check{m}(X_{i+1}, t_i|\gamma, h) - Y_{i+1}\}^2, & \text{if } \ell_t \geq 2\,. \end{cases} \tag{3.26}$$

where $\check{m}_{\text{alt}}(\cdot, t_i|\gamma, h)$ is defined at (3.25).

To conclude this section, we still need to derive what specific choices of γ and h $\check{m}_{\text{alt}}(x, t_i|\gamma, h)$ at (3.25) has taken. For this, it suffices to see that, since these $\check{m}_{\text{alt}}$'s are going to be redefined as $\check{m}$'s in the ensemble substitution at time $t_{N_{\ell_t-1}}$ (see (ES2) at page 133), $\check{m}_{\text{alt}}(x, t_i|\gamma, h)$ takes the same (γ, h) values for data arriving up to time t_i as $\check{m}_{\text{alt}}(x, t_{N_{\ell_t-1}}|\gamma, h)$. That is, we have $(\gamma_j^{\text{alt}}, h_j^{\text{alt}}) = (\check{\gamma}_j, \check{h}_j)$ for $j = 1, \ldots, i$, where $(\check{\gamma}_j, \check{h}_j)$ is defined at (3.24). Specifically, we have obtained

$$(\gamma_j^{\text{alt}}, h_j^{\text{alt}}) = \begin{cases} (\gamma_{N_1-2\nu}^{CV}, h_{N_1-2\nu}^{CV})\,, & \text{for } j = 1, \ldots, N_1 - 2\nu\,, \\ (\gamma_{N_2-2\nu}^{CV}, h_{N_2-2\nu}^{CV})\,, & \text{for } j = N_1 - 2\nu + 1, \ldots, N_2 - 2\nu\,, \\ \qquad\vdots & \\ (\gamma, h)\,, & \text{for } j = N_{\ell_t-1} - 2\nu + 1, \ldots, i\,, \end{cases} \tag{3.27}$$

3.3 Theoretical analysis

In this section, we consider the asymptotic behaviours of $\check{m}$ at (3.22) and RCV_t at (3.26) in the infill asymptotics setting (see §1.3.4.3), i.e. as $\Delta t \to 0$. Recalling from (1.1) that this implies that the arrival times $\{t_i\}_{i=1,2,\ldots}$ of the data stream are denser and denser on $\mathbb{R}_+$ and, by (1.2), we also have $n_t \to \infty$ for any given time $t \in \mathbb{R}_+$. Although in practice Δt is usually fixed,

infill asymptotics allow us to investigate how the estimator will behave if we have more and more observations locally in time.

We first present a result establishing that, for a time $t \in \mathbb{R}_+$, the difference between $\check{m}(x,t|\gamma,h)$ at (3.22) and $\tilde{m}(x,t|\gamma,h)$ at (3.1) is small. Recall that $\tilde{m}(x,t|\gamma,h)$ uses the same (γ,h) for all past data and hence it is often simpler to analyse than $\check{m}(x,t|\gamma,h)$. Indeed, we shall make use of this result to establish the asymptotic normality of $\check{m}(x,t|\gamma,h)$ in Theorem 3.1.

Recall that we observe a data stream $\{(X_i,Y_i)\}_{i=1,2,\ldots}$ from the model (1.3) with arrival times $\{t_i\}_{i=1,2,\ldots}$ defined at (1.1) and each $X_i \sim f(\cdot,t_i)$. Also recall that ν defined at the beginning of §3.2.1.1 determines how often we check for alternative ensembles. Let $\square_\ell$ denote the range of the candidate set $I^\ell_{\gamma,h}$ as at (3.5). Let $\#(I^\ell_{\gamma,h})$ denote the cardinality of $I^\ell_{\gamma,h}$. For a given $\Delta t > 0$, let $\alpha(\Delta t,k)$ denote the k-th order α-mixing coefficient of $\{X_i\}_{i=1,2,\ldots}$. See Appendix 3.B.1 for the definition of $\alpha(\Delta t,k)$.

For Proposition 3.1, we assume:

(C1) There exist constants $\delta \in (0,1)$ and $a_1,a_2 \in (0,1)$ with $1/2 < a_1 - a_2 < 1$ such that for $\ell = 1,2,\ldots$, we have $\square^\ell \subset \square$, where $\square = [\gamma_m,\gamma_M] \times [h_m,h_M] \subset (0,1)\times(0,\infty)$ with $\gamma_m = \delta\Delta t^{a_1}$, $\gamma_M = \Delta t^{a_1}/\delta$, $h_m = \delta\Delta t^{a_2}$ and $h_M = \Delta t^{a_2}/\delta$.

(C2) There exists $b \in (a_1,1)$ such that $\nu \in [\delta\Delta t^{-b}, \Delta t^{-b}/\delta]$, where δ and a_1 are defined in Condition (C1).

(C3) $\#(I^\ell_{\gamma,h}) = O(\Delta t^{-c})$ uniformly in ℓ with

$$0 \le c < \max\left\{\frac{(a+1)(2+\varsigma)(a_1-a_2)}{a+2+\varsigma} - 2a_2 - 2, (a+1)(a_1-a_2) - 2a_2 - 2\right\}, \quad (3.28)$$

where a_1 and a_2 are defined in Condition (C1), ς is defined in Condition (C7) and a is defined in Condition (C8).

(C4) m and all its partial derivatives up to order 3 exist and are uniformly bounded.

(C5) f and all its partial derivatives up to order 3 exist and are uniformly bounded.

(C6) The kernel function K is positive, continuous, symmetric, vanishes outside of $[-1, 1]$ and satisfies $\int K(u)\,\mathrm{d}u = 1$.

(C7) $\{\epsilon_i\}_{i=1,2,\ldots}$ is i.i.d. and independent from $\{X_i\}_{i=1,2,\ldots}$ and satisfies $\mathrm{E}(\epsilon_i) = 0$, $\mathrm{var}(\epsilon_i) = \sigma^2 > 0$ and $\mathrm{P}(|\epsilon_1| > x) \leq x^{-(2+\varsigma)}$ for some $\varsigma > \max\{0, (4a_2 - 2a_1 + 2)/(a_1 - a_2)\}$, where a_1, a_2 are defined in Condition (C1).

(C8) For any integer $k \geq 1$, $\sup_{\Delta t>0} \alpha(\Delta t, k) \leq Ck^{-a}$ for some constants $C > 0$ and

$$a > \max\left\{\frac{3}{2}, \frac{2(a_2+1)}{a_1 - a_2} - 1, \frac{(2+\varsigma)(2 - a_1 + 3a_2)}{(2+\varsigma)a_1 - (4+\varsigma)a_2 - 2}\right\},$$

where a_1 and a_2 are defined in Condition (C1) and ς is defined in Condition (C7).

Condition (C1) implies that $(\gamma_k^{CV}, h_k^{CV}) \in \square$, since by (3.10), $(\gamma_k^{CV}, h_k^{CV})$ is selected from some $I_{\gamma,h}^{\ell}$. Hence (3.24) and (3.27) imply that

$$(\check{\gamma}_i, \check{h}_i) \in \square \text{ and } (\gamma_i^{\text{alt}}, h_i^{\text{alt}}) \in \square. \tag{3.29}$$

Note from (1.2) that Condition (C2) implies that, as $\Delta t \to 0$, ν increases to infinity faster than γ_M^{-1}. This guarantees that, when we do the substitution $\check{m}(\cdot, t_{N_\ell}|\gamma, h) = \check{m}_{\text{alt}}(\cdot, t_{N_\ell}|\gamma, h)$ at (ES2) for $\ell = 1, 2, \ldots$, the alternative estimator $\check{m}_{\text{alt}}(\cdot, t_{N_\ell}|\gamma, h)$ is asymptotically equivalent to $\tilde{m}(\cdot, t_{N_\ell}|\gamma, h)$, where $\tilde{m}$ defined at (3.1) uses the same (γ, h) value for all past data points. Indeed, recall from §3.2.1.3 that for $k = N_\ell - 2\nu + 1, \ldots, N_\ell$, $\check{m}_{\text{alt}}(\cdot, t_{N_\ell}|\gamma, h)$ is updated using (3.16). That is, up to time t_{N_ℓ}, $\check{m}_{\text{alt}}(\cdot, t_{N_\ell}|\gamma, h)$ has been updated using the same (γ, h) value for 2ν data. Since the influence of old data decreases exponentially fast, when ν is large enough, the difference between $\check{m}_{\text{alt}}(\cdot, t_{N_\ell}|\gamma, h)$ and $\tilde{m}(\cdot, t_{N_\ell}|\gamma, h)$ becomes small. However, we still require that $\nu\Delta t \to 0$ as $\Delta t \to 0$, which implies, recalling from (1.2), that ν is small compared to n_t. This is because ν also controls how often we check if we need to construct alternative ensembles,

so that ν cannot be too large, otherwise we would use an inappropriate ensemble for too long before we can finally substitute it with an alternative one.

We impose Condition (C3) on the cardinality of $I^{\ell}_{\gamma,h}$ since Proposition 3.1 proves that $\check{m}$ and $\tilde{m}$ are close uniformly in (γ, h). Therefore, although $\#(I^{\ell}_{\gamma,h})$ can either stay finite ($c = 0$) or increases polynomially fast as $\Delta t \to 0$ ($c > 0$) under Condition (C3), it cannot grow too fast.

Conditions (C4)–(C6) are some standard smoothness conditions, requiring that the regression function m, the density f and the kernel K are all smooth enough. See, for example, Vogt (2012) and Zhang and Wu (2015).

Conditions (C3), (C7) and (C8) involve some complicated upper and lower bounds for the constants c, ε and a. We use an example to show how these conditions can hold simultaneously. First, we take $a_1 = 5/7$ and $a_2 = 1/7$ satisfying Condition (C1) (we will show below Theorem 3.1 that these choices of a_1 and a_2 lead to the optimal convergence rate for $\check{m}$). Then Condition (C7) implies that $\varepsilon > 2$ and Condition (C8) requires $a > 3 + 12/(\varsigma - 2)$, where the latter is comparable to $a > 3 + 8/\varsigma$ needed in Vogt (2012). Finally, under these restrictions, the right hand side of (3.28) is strictly positive, so that Condition (C3) can also hold.

The next proposition shows that the difference between $\check{m}(x, t|\gamma, h)$ at (3.22) and $\tilde{m}(x, t|\gamma, h)$ at (3.1) is small uniformly in x and $(\gamma, h) \in I^{\ell_t}_{\gamma,h}$, where ℓ_t is defined at (3.21). See Appendix 3.D for its proof.

Proposition 3.1. *Assume that Conditions (C1)–(C8) hold. Suppose $t \in \mathbb{R}_+$ is a given time satisfying* (3.20). *For any compact set $\mathcal{K} \subset \mathbb{R}$ such that* $\inf_{x\in\mathcal{K}} f(x, t) > 0$*, there exists $\zeta > 0$ such that*

$$\sup_{x\in\mathcal{K}} \sup_{(\gamma,h)\in I^{\ell_t}_{\gamma,h}} \left|\check{m}(x, t|\gamma, h) - \tilde{m}(x, t|\gamma, h)\right| = o_p\big\{\exp(-\zeta\Delta t^{-1})\big\}, \text{ as } \Delta t \to 0. \tag{3.30}$$

Note that the convergence rate $o_p\big\{\exp(-\zeta\Delta t^{-1})\big\}$ in (3.30) is not the fastest possible and can be further improved. However, since this exponentially fast convergence rate already serves our purpose well (see discussions below), we leave the sharpening of the convergence rate to future work.

The above proposition implies that, at time t, the estimators $\check{m}(x, t|\gamma, h) \in \mathscr{E}_t$ used in the

SRA in §3.2 is asymptotically equivalent to the estimators $\tilde{m}(x,t|\gamma,h) \in \mathscr{E}_t^0$ used in the prototype algorithm in §3.1, taking $I_{\gamma,h} = I_{\gamma,h}^{\ell_t}$. Hence, selecting an estimator from $\mathscr{E}_t$ is asymptotically equivalent to selecting one from $\mathscr{E}_t^0$. This is because the SRA mimics the prototype algorithm, although the former is free from the computational issues mentioned in §3.1.3.

Proposition 3.1 is also useful for analysing asymptotic properties of $\check{m}(x,t|\gamma,h)$, since $\tilde{m}(x,t|\gamma,h)$ uses (γ,h) for all observations and is simpler to analyse. Indeed, we shall make use of Proposition 3.1 to establish the asymptotic normality of $\check{m}(x,t|\gamma,h)$. For that, we will need some additional conditions as follows.

For a given $\Delta t > 0$, let $\rho(\Delta t, k)$ denote the k-th order ρ-mixing coefficient of $\{X_i\}_{i=1,2,\ldots}$ (see Appendix 3.B.1 for the definition of $\rho(\Delta t, k)$). Let X and Y denote two random variables with marginal densities f_X and f_Y and joint density $f_{X,Y}$. Following equation (1.12) of Bosq (1998, p. 22), we define the local measure of dependence between two random variables X and Y as

$$g_{X,Y}(x,y) = f_{X,Y}(x,y) - f_X(x)f_Y(y). \tag{3.31}$$

Now, let $g_{ij} = g_{X_i,X_j}$ denote the local measure of dependence between two covariates X_i and X_j for some $i \neq j$. Then we have $g_{ij}(x,y) = f_{ij}(x,y) - f(x,t_i)f(y,t_j)$, where f_{ij} denotes the joint density of X_i and X_j.

In addition to Conditions (C1)–(C8), we make the following assumptions:

(C9) $\sup_{\Delta t>0} \rho(\Delta t, 1) < 1$.

(C10) Given some $i \neq j$, there exists a constant $L > 0$ such that

$$|g_{ij}(x',y') - g_{ij}(x,y)| \leq L(|x'-x|^2 + |y'-y|^2)^{1/2},$$

for any $x, x', y, y' \in \mathbb{R}$.

Condition (C9) implies that the correlation coefficient between X_i and X_{i+1} is bounded away from ± 1, which is sufficiently mild (Peligrad, 1996). Condition (C10), implying that the local

measure of dependence g_{ij} is Lipschitz continuous, is standard in the literature of nonparametric estimation for dependent data (Bosq, 1998, p. 23).

We introduce the following notations: for functions $g, g_1, g_2 : \mathbb{R} \times \mathbb{R}_+ \mapsto \mathbb{R}$, let $g_x = \partial g/\partial x$, $g_t = \partial g/\partial t$, $(g_1 \cdot g_2)(x,t) = g_1(x,t)g_2(x,t)$ and

$$\mu_{K,2}^2 = \int u^2 K(u)\,\mathrm{d}u,\ R_K = \int K^2(u)\,\mathrm{d}u. \tag{3.32}$$

Then the next theorem establishes the asymptotic normality of $\check{m}$, defined at (3.22). See Appendix 3.E for its proof.

Theorem 3.1. *Assume that Conditions (C1)–(C10) hold. Assume also that* $1/7 \le a_2 < a_1 \le 5/7$*, where* a_1 *and* a_2 *are defined in Condition (C1). Suppose* $(\gamma, h) \in I_{\gamma,h}^{\ell_t}$ *and* $(x,t) \in \mathbb{R} \times \mathbb{R}_+$ *is given and satisfies* $f(x,t) > 0$*. Then we have*

$$\frac{1}{\sqrt{V}}\big\{\check{m}(x,t|\gamma,h) - m(x,t) - B\big\} \Rightarrow N(0,1),\ \textit{as}\ \Delta t \to 0, \tag{3.33}$$

where $\Rightarrow$ *denotes convergence in distribution and where*

$$V = \frac{R_K}{2}\frac{\gamma}{h}\frac{\sigma^2}{f(x,t)} + o(\gamma/h) \tag{3.34}$$

and

$$\begin{aligned} B = \Big\{\frac{1}{2}m_{xx}(x,t) + \Big(\frac{m_x \cdot f_x}{f}\Big)(x,t)\Big\}\mu_{K,2}h^2 \\ + \Big\{\Big(\frac{m \cdot f_t}{f}\Big)(x,t) - m_x(x,t) - \Big(\frac{m \cdot f_x}{f}\Big)(x,t)\Big\}\frac{\Delta t}{\gamma}. \end{aligned} \tag{3.35}$$

Equations (3.33)–(3.35) imply that the fastest convergence rate of $\check{m}(x,t|\gamma,h)$ to $m(x,t)$ is attained when we take $\gamma \asymp \Delta t^{5/7}$ and $h \asymp \Delta t^{1/7}$ so that $B \asymp \sqrt{V} \asymp \Delta t^{2/7}$, which implies that the fastest convergence rate in distribution is $\Delta t^{2/7}$. Using (1.2), we have $\Delta t^{2/7} \asymp n_t^{-2/7}$, which is slightly slower than the convergence rate in distribution $n_t^{-1/3}$ of the regression estimators in Vogt (2012) and Zhang and Wu (2015). Recall from §2.2 that SKDE also has slower convergence

rate compared to the temporal KDE (1.13) in Hall et al. (2006). Discussions on the different convergence rates of density estimators in §2.3 are still applicable in the regression case.

For an illustration of the finite-sample behaviour of the SRA, see §4.1.2 and §4.2.2 for some simulation studies and real data examples.

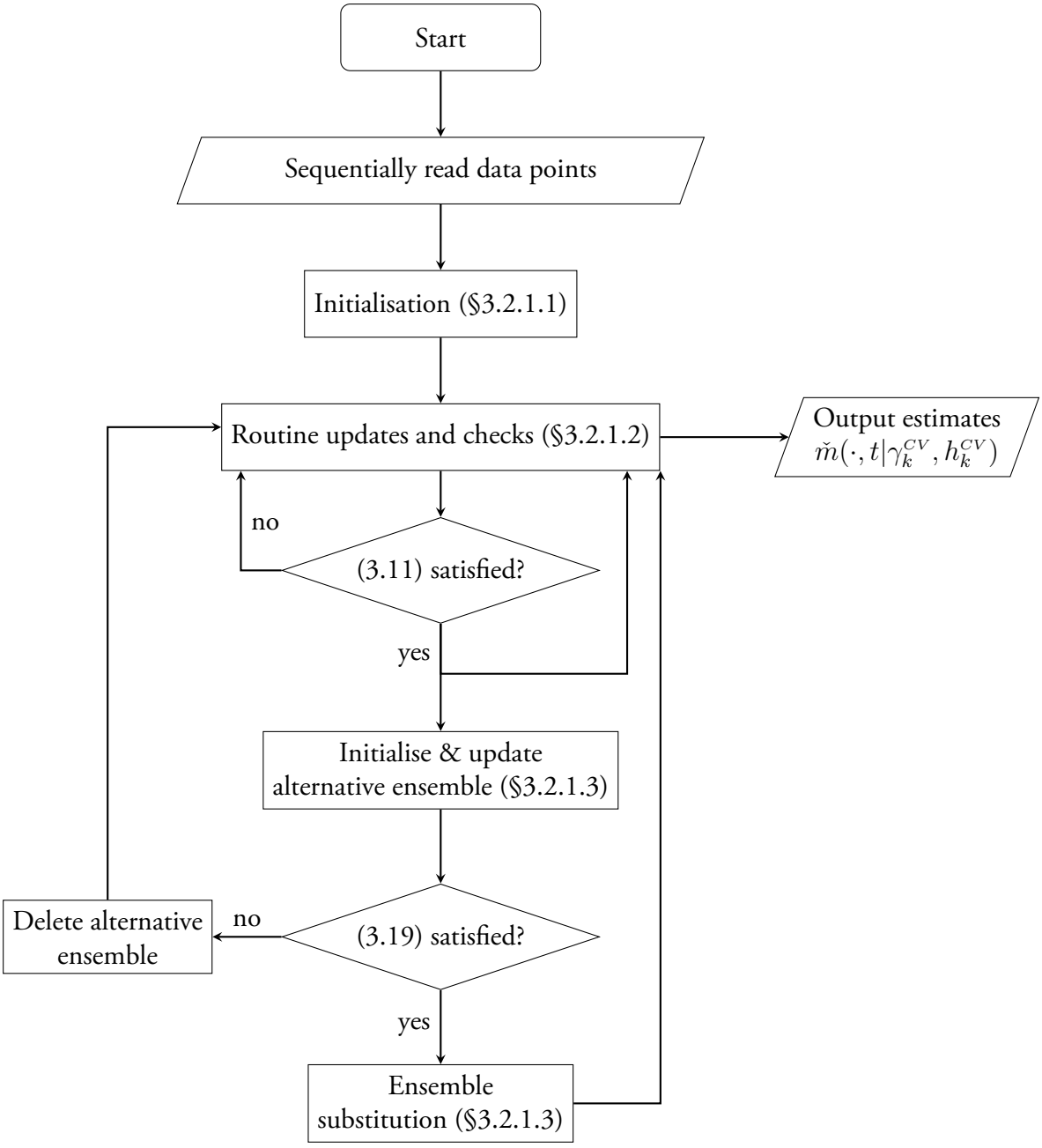

Figure 3.1: Flowchart of the SRA defined in §3.2. There is no 'end' node since the data stream might be never-ending.

Appendix

3.A Notation

We summarise the notations that will be used in the appendices.

Data. We observe a data stream $\{(X_i, Y_i)\}_{i=1,2,\dots}$ from the model (1.3) with arrival times $\{t_i\}_{i=1,2,\dots}$ defined at (1.1) and each $X_i \sim f(\cdot, t_i)$. Let n_t denote the number of observations arriving up to time t satisfying (1.2).

Streaming regression algorithm (SRA). Appendix 3.F gives the pseudocode for the SRA defined in §3.2. Note from there that ν is an integer determining how often the algorithm checks for alternative ensembles. Also note that $N_0 = 2\nu$ (line 2 in Algorithm F2) and N_ℓ denotes the number of observations arriving up to the ℓ-th ensemble substitution (line 7 in Algorithm F7). In Algorithm F1, we can see that λ increases by at least 1 after the while loop (lines 11–14) and it increase by 2 after the for loop (lines 16–23). Hence we have $\lambda \geq 3$ in line 7, which implies $N_\ell - N_{\ell-1} \geq 3\nu$, for $\ell = 1, 2, \dots$ Finally, recall that for a time $t > 0$, ℓ_t denote the integer satisfying (3.21).

Regression estimators. In view of (1.11) and (1.12), we obtain from (3.1), (3.22) and (3.25) that

$$\tilde{r}(x, t|\gamma, h) = \sum_{i=1}^{n_t} \tilde{w}_{n_t,i} K_h(x - X_i) Y_i, \quad \tilde{f}(x, t|\gamma, h) = \sum_{i=1}^{n_t} \tilde{w}_{n_t,i} K_h(x - X_i), \qquad \text{(3.A.1)}$$

$$\check{r}(x,t|\gamma,h)=\sum_{i=1}^{n_t}\check{w}_{n_t,i}K_{\check{h}_i}(x-X_i)Y_i,\quad \check{f}(x,t|\gamma,h)=\sum_{i=1}^{n_t}\check{w}_{n_t,i}K_{\check{h}_i}(x-X_i),\quad (3.A.2)$$

$$\check{r}_{\text{alt}}(x,t_i|\gamma,h)=\sum_{j=1}^{i}w_{i,j}^{\text{alt}}K_{h_j^{\text{alt}}}(x-X_j)Y_j,\ \check{f}_{\text{alt}}(x,t_i|\gamma,h)=\sum_{j=1}^{i}w_{i,j}^{\text{alt}}K_{h_j^{\text{alt}}}(x-X_j),\quad (3.A.3)$$

where

$$\tilde{w}_{n_t,i}=\begin{cases}(1-\gamma)^{n_t-1}, & \text{for } i=1,\\ \gamma(1-\gamma)^{n_t-i}, & \text{for } i=2,\dots,n_t,\end{cases} \quad (3.A.4)$$

$$\check{w}_{n_t,i}=\begin{cases}\prod_{j=2}^{n_t}(1-\check{\gamma}_j), & \text{for } i=1,\\ \check{\gamma}_i\prod_{j=i+1}^{n_t}(1-\check{\gamma}_j), & \text{for } i=2,\dots,n_t,\end{cases} \quad (3.A.5)$$

and

$$w_{i,j}^{\text{alt}}=\begin{cases}\prod_{k=2}^{i}(1-\gamma_k^{\text{alt}}), & \text{for } j=1,\\ \gamma_j^{\text{alt}}\prod_{k=j+1}^{i}(1-\gamma_k^{\text{alt}}), & \text{for } j=2,\dots,i,\end{cases} \quad (3.A.6)$$

with $(\check{\gamma}_i,\check{h}_i)$ defined by (3.23) and (3.24) and $(\gamma_j^{\text{alt}},h_j^{\text{alt}})$ defined by (3.27).

Essential supremum. Let $\|\cdot\|_\infty$ denote the essential supremum. That is, for a real-valued random variable, $\|X\|_\infty=\inf\{x>0:|X|\le x \text{ a.s.}\}$; under Lebesgue measure, for a real and measurable function $f:\mathbb{R}\mapsto\mathbb{R}$, $\|f\|_\infty=\inf\{C>0:|f(x)|\le C \text{ for almost every } x\}$.

Asymptotic orders. Let $a=a(\Delta t)$ and $b=b(\Delta t)$ be two functions of Δt and $b\neq 0$ for all $\Delta t>0$. We write $a\sim b$ if $\lim_{\Delta t\to 0}ab^{-1}=1$, $a=o(b)$ if $\lim_{\Delta t\to 0}ab^{-1}=0$, $a=O(b)$ if $\limsup_{\Delta t\to 0}|ab^{-1}|<\infty$ and $a\asymp b$ if both $a=O(b)$ and $b=O(a)$.

3.B Some known results

3.B.1 Mixing conditions

Mixing conditions are often used to characterise the dependence structure of time series (Bosq, 1998, Chapter 1; Bradley, 2005; Peligrad, 1996). Among various mixing conditions, α-mixing and ρ-mixing are two commonly used ones. Here we present the definitions of the α-mixing and the ρ-mixing coefficients.

Let $(\Sigma, \mathcal{F}, \mathrm{P})$ denote a probability space, the α-mixing coefficient, measuring the dependence between two σ-algebras $\mathcal{A}, \mathcal{B} \in \mathcal{F}$, is defined as

$$\alpha_0(\mathcal{A}, \mathcal{B}) = \sup_{A \in \mathcal{A}, B \in \mathcal{B}} |\,\mathrm{P}(A \cap B) - \mathrm{P}(A)\,\mathrm{P}(B)|. \tag{3.B.1}$$

The ρ-mixing coefficient is given by

$$\rho_0(\mathcal{A}, \mathcal{B}) = \sup\big\{|\,\mathrm{corr}(X, Y)| : X \in L^2(\mathcal{A}), Y \in L^2(\mathcal{B})\big\}, \tag{3.B.2}$$

where $\mathrm{corr}(\cdot, \cdot)$ denotes the correlation coefficient and $L^2(\mathcal{A})$ the space of all $\mathcal{A}$-measurable random variables X satisfying $\mathrm{E}(|X|^2) < \infty$. When $\mathcal{A}$ and $\mathcal{B}$ are independent, we have $\alpha_0(\mathcal{A}, \mathcal{B}) = \rho_0(\mathcal{A}, \mathcal{B}) = 0$.

Now consider the data stream $\{(X_i, Y_i)\}_{i=1,2,\ldots}$ with arrival times defined at (1.1). For a given $\Delta t > 0$, the k-th order α-mixing coefficient of the covariate sequence $\{X_i\}_{i=1,2,\ldots}$ is defined as

$$\alpha(\Delta t, k) = \sup_{i=1,2,\ldots} \alpha_0\big\{\sigma(X_j, j \le i), \sigma(X_j, j \ge i + k)\big\}, \tag{3.B.3}$$

where $\sigma(X_j, j \le i)$ and $\sigma(X_j, j \ge i + k)$ denote the σ-algebras generated by $\{X_j\}_{j=1,\ldots,i}$ and $\{X_j\}_{j=i+k,j+k+1,\ldots}$, respectively. Note that we have Δt in (3.B.3) since, by (1.1), the arrival time t_j of X_j depends on Δt, and hence its density $f(\cdot, t_j)$ also depends on Δt. Similarly, the k-th

order ρ-mixing coefficient of $\{X_i\}_{i=1,2,\ldots}$ is defined as

$$\rho(\Delta t, k) = \sup_{i=1,2,\ldots} \rho_0\big\{\sigma(X_j, j \leq i), \sigma(X_j, j \geq i+k)\big\}. \tag{3.B.4}$$

3.B.2 Technical results

The next lemma, known as summation by parts (Elaydi, 2005, pp. 62–63), will be frequently used in the proofs.

Lemma 3.B.1. *Let $\{a_i\}_{i=0,\ldots,n}$ and $\{b_i\}_{i=0,\ldots,n}$ denote two sequences of real numbers, then*

$$\sum_{i=0}^{n} a_i b_i = a_0 \sum_{i=0}^{n} b_i + \sum_{j=0}^{n-1} (a_{j+1} - a_j) \sum_{i=j+1}^{n} b_i.$$

The next lemma gives the asymptotic orders of the p-series $\sum_{i=1}^{n} i^{-p}$. See Goel and Rodriguez (1987) for its proof.

Lemma 3.B.2. *Let $p \in (0, \infty)$ denote a constant, then we have, as $n \to \infty$,*

$$\sum_{i=1}^{n} \frac{1}{i^p} = \begin{cases} O(1), & \text{if } p \in (1, \infty), \\ O(\log n), & \text{if } p = 1, \\ O(n^{1-p}), & \text{if } p \in (0, 1), \end{cases} \tag{3.B.5}$$

If $p > 1$, then we also have

$$\sum_{i=n}^{\infty} \frac{1}{i^p} = O(n^{1-p}), \text{ as } n \to \infty. \tag{3.B.6}$$

The next lemma can be found in Chung (2000, p. 52).

Lemma 3.B.3. *Let X be a positive random variable. Then*

$$\mathrm{E}(X) = \int_0^{\infty} \mathrm{P}(X > x)\, \mathrm{d}x.$$

The following two lemmas are Lemma 1.3 and formula (1.11) from Bosq (1998). The latter is also known as Billingsley's inequality. In both lemmas, let X and Y denote two real-valued univariate random variables, $\sigma(X)$ and $\sigma(Y)$ denote the σ-algebras generated by X and Y and let $\alpha_0\{\sigma(X), \sigma(Y)\}$ denote their α-mixing coefficient, defined at (3.B.1). Let $g_{(X,Y)}$ denote their local measure of dependence, defined at (3.31).

Lemma 3.B.4. *If there exists a constant $L > 0$ such that*

$$|g_{(X,Y)}(x', y') - g_{(X,Y)}(x, y)| \leq L\big\{|x' - x|^2 + |y' - y|^2\big\}^{1/2},$$

for any $x, x', y, y' \in \mathbb{R}$, then there exists a constant $C = C(L)$ such that

$$\|g_{(X,Y)}\|_\infty \leq C\alpha_0^{1/3}\{\sigma(X), \sigma(Y)\}.$$

Lemma 3.B.5. *Suppose $\|X\|_\infty < \infty$ and $\|Y\|_\infty < \infty$. Then*

$$|\mathrm{cov}(X, Y)| \leq 4\|X\|_\infty\|Y\|_\infty\alpha_0\{\sigma(X), \sigma(Y)\}.$$

The following result is Theorem 2.2 from Peligrad (1996). See also Theorem 1.1 in Bradley and Tone (2017) for a more general version. Let $\{\xi_{n,i} : 1 \leq i \leq n\}$ denote a triangular array of zero-mean random variables. For each n, define the k-th order α-mixing coefficient of $\{\xi_{n,i}\}$ by

$$\tilde{\alpha}(n, k) = \sup_{1 \leq j \leq n-k} \alpha_0\big\{\sigma(\xi_{n,i}, 1 \leq i \leq j), \sigma(\xi_{n,i}, j + k \leq i \leq n)\big\}, \tag{3.B.7}$$

for $k = 1, \ldots, n - 1$, where α_0 is defined at (3.B.1). Similarly, define the ρ-mixing coefficient of $\{\xi_{n,i}\}$ by

$$\tilde{\rho}(n, k) = \sup_{1 \leq j \leq n-k} \rho_0\big\{\sigma(\xi_{n,i}, 1 \leq i \leq j), \sigma(\xi_{n,i}, j + k \leq i \leq n)\big\}, \tag{3.B.8}$$

for $k = 1, \ldots, n-1$, where ρ_0 is defined at (3.B.2). Let

$$S_n = \sum_{i=1}^{n} \xi_{n,i} \text{ and } \sigma_n^2 = \text{var}(S_n). \tag{3.B.9}$$

Then the next theorem establishes a central limit theorem for $\{\xi_{n,i}\}$.

Theorem 3.B.1. *Assume that the following two mixing conditions hold:*

(C1) $\sup_{n\geq 1} \tilde{\alpha}(n,k) \to 0$ *as* $k \to \infty$.

(C2) $\sup_{n\geq 1} \tilde{\rho}(n,1) < 1$.

Assume also that the following Lyapunov condition holds:

(C3) There exists some constant $c > 0$ *such that*

$$\lim_{n\to\infty} \frac{1}{\sigma_n^{2+c}} \sum_{i=1}^{n} \mathrm{E}(\xi_{n,i}^{2+c}) = 0,$$

where S_n *and* σ_n *are defined in* (3.B.9)

Then we have

$$\frac{S_n}{\sigma_n} \Rightarrow N(0,1), \text{ as } n \to \infty,$$

where $\Rightarrow$ *denotes convergence in distribution.*

The following is the Fuk–Nagaev inequality for α-mixing sequences (Rio, 2013, p. 102–105).

Theorem 3.B.2. *Let* $\{\xi_{n,i}\}$ *denote a triangular array of zero-mean random variables with* α*-mixing coefficients* $\{\tilde{\alpha}(n,k)\}$ *defined at* (3.B.7). *Assume* $\tilde{\alpha}(n,k) \leq Ck^{-a}$*, for some constants* $C > 0$ *and* $a > 1$. *Assume also that there exists some constant* $C_0, C_1 > 0$ *and* $p > 2$ *such that*

$$\mathrm{P}(|\xi_{n,i}| > x) \leq C_1 x^{-p}, \text{ for any } x > C_0\,. \tag{3.B.10}$$

Let $\bar{\sigma}_n^2 = \sum_{i,j=1}^n |\operatorname{cov}(\xi_{n,i}, \xi_{n,j})|$. *Then for any* $\lambda \geq 1$, *there exists a constant* $\kappa = \kappa(\lambda, a, p)$ *that depends on* λ, a *and* p *such that for all* $\eta > 0$,

$$\mathrm{P}(|S_n| \geq \eta) \leq \kappa \left\{ \left(1 + \frac{\eta^2}{\lambda \bar{\sigma}_n^2}\right)^{-\lambda/2} + \frac{n}{\lambda}\left(\frac{\lambda}{\eta}\right)^{(a+1)p/(a+p)} \right\}. \tag{3.B.11}$$

Furthermore, if $\|\xi_{n,i}\|_\infty \leq 1$ *holds for all* i, *then for any* $\lambda \geq 1$, *there exists a constant* $\kappa = \kappa(\lambda, a)$ *that depends on* λ *and* a *such that for all* $\eta > 0$

$$\mathrm{P}(|S_n| \geq \eta) \leq \kappa \left\{ \left(1 + \frac{\eta^2}{\lambda \bar{\sigma}_n^2}\right)^{-\lambda/2} + \frac{n}{\lambda}\left(\frac{\lambda}{\eta}\right)^{a+1} \right\}. \tag{3.B.12}$$

3.C Some technical lemmas

In this section we prove some lemmas needed the proofs of Proposition 3.1 and Theorem 3.1.

Lemma 3.C.1. *Under Condition (C7), we have* $\mathrm{E}\,|\epsilon_1|^{2+c} < \infty$, *for any* $c \in (0, \varsigma)$.

Proof. For some $c \in (0, \varsigma)$, where ς is defined in Condition (C7), we have

$$\begin{aligned} \mathrm{E}\,|\epsilon_i|^{2+c} &= \int_0^\infty \mathrm{P}\big(|\epsilon_i|^{2+c} > x\big)\,\mathrm{d}x = \int_0^\infty \mathrm{P}\big\{|\epsilon_i| > x^{1/(2+c)}\big\}\,\mathrm{d}x \\ &\leq \int_0^\infty x^{-(2+\varsigma)/(2+c)}\,\mathrm{d}x < \infty, \end{aligned} \tag{3.C.1}$$

where we used Lemma 3.B.3 to get the first line and Condition (C7) and the fact that $(2 + \varsigma)/(2 + c) > 1$ to get the last inequality. □

The following lemma states some standard arithmetic results, useful for the proof of some other lemmas in this section.

Lemma 3.C.2. *For* $\gamma \in (0, 1)$, *we have*

$$\sum_{i=0}^{n} (1-\gamma)^i = \frac{1}{\gamma} - \frac{(1-\gamma)^{n+1}}{\gamma}, \tag{3.C.2}$$

$$\sum_{i=0}^{n}(1-\gamma)^{2i} = \frac{1}{\gamma(2-\gamma)} - \frac{(1-\gamma)^{2(n+1)}}{\gamma(2-\gamma)}, \tag{3.C.3}$$

$$\sum_{i=0}^{n} i(1-\gamma)^{i} = \frac{1}{\gamma^2} - \frac{1}{\gamma} - \Big(\frac{1}{\gamma^2} + \frac{n}{\gamma}\Big)(1-\gamma)^{n+1}\,, \tag{3.C.4}$$

$$\sum_{i=0}^{n} i^2(1-\gamma)^{i} = \frac{2}{\gamma^3} - \frac{3}{\gamma^2} + \frac{1}{\gamma} - \Big(\frac{2}{\gamma^3} + \frac{2n-1}{\gamma^2} + \frac{n^2}{\gamma}\Big)(1-\gamma)^{n+1}\,. \tag{3.C.5}$$

Proof. We only demonstrate (3.C.4) and (3.C.5) below since the proof of (3.C.2) and (3.C.3) are straightforward.

To show (3.C.4), note that in view of Lemma 3.B.1, we have

$$\begin{aligned}\sum_{i=0}^{n} i(1-\gamma)^{i} &= \sum_{i=0}^{n} i(1-\gamma)^{i} = \sum_{j=0}^{n-1}\sum_{i=j+1}^{n}(1-\gamma)^{i} \\ &= \frac{1}{\gamma}\sum_{j=0}^{n-1}\{(1-\gamma)^{j+1} - (1-\gamma)^{n+1}\} = \frac{1}{\gamma}\sum_{j=0}^{n-1}(1-\gamma)^{j+1} - \frac{n(1-\gamma)^{n+1}}{\gamma}\,.\end{aligned}$$

From there, as (3.C.2) implies that the first term of the right hand side of the last equality can be expressed as

$$\frac{1}{\gamma}\sum_{j=0}^{n-1}(1-\gamma)^{j+1} = \frac{(1-\gamma)}{\gamma}\sum_{j=0}^{n-1}(1-\gamma)^{j} = \frac{(1-\gamma)}{\gamma^2}\Big\{1-(1-\gamma)^{n}\Big\}\,,$$

it follows that

$$\begin{aligned}\sum_{i=0}^{n} i(1-\gamma)^{i} &= \frac{(1-\gamma)}{\gamma^2}\Big\{1-(1-\gamma)^{n}\Big\} - \frac{n(1-\gamma)^{n+1}}{\gamma} \\ &= \frac{1}{\gamma^2} - \frac{1}{\gamma} - \frac{(1-\gamma)^{n+1}}{\gamma^2} - \frac{n(1-\gamma)^{n+1}}{\gamma}\,,\end{aligned}$$

which proves (3.C.4).

Next we show (3.C.5). For this, using (3.C.4), we first compute that

$$\sum_{j=0}^{n-1}(2j+1)(1-\gamma)^{j+1} = (1-\gamma)\Big\{2\sum_{j=0}^{n-1} j(1-\gamma)^{j} + \sum_{j=0}^{n-1}(1-\gamma)^{j}\Big\}$$

$$= (1-\gamma)\Big\{\frac{2}{\gamma^2} - \frac{2}{\gamma} - \frac{2(1-\gamma)^n}{\gamma^2} - \frac{2(n-1)(1-\gamma)^n}{\gamma} + \frac{1-(1-\gamma)^n}{\gamma}\Big\}$$
$$= (1-\gamma)\Big\{\frac{2}{\gamma^2} - \frac{1}{\gamma} - \frac{2(1-\gamma)^n}{\gamma^2} - \frac{(2n-1)(1-\gamma)^n}{\gamma}\Big\}$$

Therefore,

$$\sum_{j=0}^{n-1}(2j+1)(1-\gamma)^{j+1} = 1 - \frac{3}{\gamma} + \frac{2}{\gamma^2} - \frac{2(1-\gamma)^{n+1}}{\gamma^2} - \frac{(2n-1)(1-\gamma)^{n+1}}{\gamma}\,.$$

Since

$$\sum_{j=0}^{n-1}(2j+1) = n^2\,,$$

we deduce from the above computations and combined with an application of Lemma 3.B.1 that

$$\begin{aligned}\sum_{i=0}^{n} i^2(1-\gamma)^i &= \sum_{j=0}^{n-1}(2j+1)\sum_{i=j+1}^{n}(1-\gamma)^i = \frac{1}{\gamma}\sum_{j=0}^{n-1}(2j+1)\{(1-\gamma)^{j+1} - (1-\gamma)^{n+1}\}\\ &= \frac{1}{\gamma}\sum_{j=0}^{n-1}(2j+1)(1-\gamma)^{j+1} - \frac{(1-\gamma)^{n+1}}{\gamma}\sum_{j=0}^{n-1}(2j+1)\\ &= \frac{2}{\gamma^3} - \frac{3}{\gamma^2} + \frac{1}{\gamma} - (1-\gamma)^{n+1}\Big(\frac{2}{\gamma^3} + \frac{2n-1}{\gamma^2} + \frac{n^2}{\gamma}\Big)\,.\end{aligned}$$

This concludes the proof of (3.C.5). □

Recall the definition of ν at the beginning of §3.2.1.1 and the definitions of γ_m and γ_M in Condition (C1).

Lemma 3.C.3. *Let $t > 0$. Under Conditions (C1) and (C2), there exists a constant $\kappa > 0$ such that for any $\gamma \in [\gamma_m, \gamma_M]$, it holds uniformly in $s \in [t_{N_0}, t]$ that as $\Delta t \to 0$:*

$$(1-\gamma)^{n_s - 1} \le (1-\gamma)^{\nu} \le \exp\big(-\Delta t^{-\kappa}\big)\,. \tag{3.C.6}$$

Proof. By a first-order Taylor expansion of e^x around 0, we have $e^x = 1 + x + e^{\theta}x^2/2$ for some θ between 0 and x, which implies $1 - x \le e^{-x}$. As under Condition (C1) we have

$\gamma_m = \delta\Delta t^{a_1} \leq \gamma$, and since Condition (C2) guarantees that $\nu \in [\delta\Delta t^{-b}, \Delta t^{-b}/\delta]$ for a given $\delta \in (0,1)$, we therefore have that

$$(1-\gamma)^\nu \leq (1-\gamma_m)^\nu \leq e^{-\gamma_m \nu} \leq \exp\bigl(-\delta^2 \Delta t^{a_1 - b}\bigr)\,.$$

As by Condition (C2) $\delta > 0$ and $a_1 - b < 0$, taking any $\kappa' \in (0, b - a_1)$ entails $\exp(-\delta^2\Delta t^{a_1-b}) \leq \exp(-\Delta t^{-\kappa'})$ as $\Delta t \to 0$. Hence, we have proved that

$$(1-\gamma)^\nu \leq \exp\bigl(-\Delta t^{-\kappa'}\bigr) \quad \text{as } \Delta t \to 0\,. \tag{3.C.7}$$

To conclude the proof of the announced result, note from (3.20) that $n_s \geq N_0$ for any $s \in [t_{N_0}, t]$. Also, from the definition of N_0 at the beginning of §3.2.1.1, we have $N_0 = 2\nu \geq \nu + 1$ since $\nu \geq 1$, and hence $(1-\gamma)^{n_s - 1} \leq (1-\gamma)^\nu$ as $\gamma \in (0,1)$. Combining the latter inequality to (3.C.7) yields (3.C.6). □

Recalling the definitions of $\tilde{w}_{n_t,i}$, t_i and n_t at (3.A.4), (1.1) and (1.2), respectively, we have the next lemma.

Lemma 3.C.4. *For any $\gamma \in (0,1)$ and $t > 0$ satisfying $n_t \geq 2$, we have $\sum_{i=1}^{n_t} \tilde{w}_{n_t,i} = 1$.*

Proof. From (3.A.4), we have

$$\begin{aligned}\sum_{i=1}^{n_t} \tilde{w}_{n_t,i} &= (1-\gamma)^{n_t-1} + \gamma \sum_{i=2}^{n_t} (1-\gamma)^{n_t - i} = (1-\gamma)^{n_t-1} + \gamma \sum_{j=0}^{n_t-2} (1-\gamma)^j \\ &= (1-\gamma)^{n_t-1} + 1 - (1-\gamma)^{n_t-1} = 1,\end{aligned}$$

where we used (3.C.2) with $n = n_t - 2$. □

Lemma 3.C.5. *Let $t > 0$ satisfy* (3.20). *Then, under Conditions (C1) and (C2), for any $\gamma \in [\gamma_m, \gamma_M]$, where γ_m, γ_M are defined in Condition (C1), we have uniformly in $s \in [t_{N_0+1}, t]$ and as*

$\Delta t \to 0$ *that*

$$\sum_{i=1}^{n_s} \tilde{w}_{n_s,i}^2 = \frac{\gamma}{2} + o(\Delta t)\,. \tag{3.C.8}$$

Proof. Using (3.A.4), we compute that

$$\begin{aligned}\sum_{i=1}^{n_s} \tilde{w}_{n_s,i}^2 &= (1-\gamma)^{2(n_s-1)} + \gamma^2 \sum_{i=2}^{n_s} (1-\gamma)^{2(n_s-i)} = \gamma^2 \sum_{j=0}^{n_s-2} (1-\gamma)^{2j} + (1-\gamma)^{2(n_s-1)} \\ &= \frac{\gamma}{2-\gamma} + \frac{(1-\gamma)^{2(n_s-1)}}{2-\gamma} + (1-\gamma)^{2(n_s-1)}\,,\end{aligned}$$

where the last line followed by an application of (3.C.3) with $n = n_s - 2$.

From there, as $\gamma \in (0,1)$, and since by Condition (C1) we have $\gamma \asymp \Delta t^{a_1}$ with $a_1 > 1/2$, it follows that $\gamma/(2-\gamma) = \gamma/2 + o(\Delta t)$ as $\Delta t \to 0$. Moreover, since $\gamma < 1$ implies $(2-\gamma)^{-1}(1-\gamma)^{2(n_s-1)} \leq (1-\gamma)^{2(n_s-1)} \leq (1-\gamma)^{n_s-1}$, we have in view of Lemma 3.C.3 that under Conditions (C1) and (C2) the second and third terms on the last line of the above display are $o(\Delta t)$ uniformly in $s \in [t_{N_0+1}, t]$. Plugging the latter results into the above equation concludes the proof of the lemma. □

Lemma 3.C.6. *Let $g : \mathbb{R}^+ \to \mathbb{R}$ denote a twice-differentiable function such that there exists a constant $M > 0$ satisfying $\max\{\|g\|_\infty, \|g'\|_\infty, \|g''\|_\infty\} \leq M$. Under Conditions (C1) and (C2), for any $\gamma \in [\gamma_m, \gamma_M]$ and any $t > 0$ satisfying* (3.20)*, we have uniformly in $s \in [t_{N_0+1}, t]$ that as $\Delta t \to 0$,*

$$\sum_{i=1}^{n_s} \tilde{w}_{n_s,i} g(t_i) = g(s) - \frac{\Delta t}{\gamma} g'(s) + o(\Delta t/\gamma) \tag{3.C.9}$$

and

$$\sum_{i=1}^{n_s} \tilde{w}_{n_s,i}^2 g(t_i) = \frac{\gamma}{2} g(s) + o(\Delta t)\,. \tag{3.C.10}$$

Proof. We first prove (3.C.9). In view of (3.A.4) and of Lemma 3.C.4, we obtain using first-order

Taylor expansions of the $g(t_i)$'s around s that there exist $\theta_1, \ldots, \theta_{n_s}$ such that

$$\begin{aligned}\sum_{i=1}^{n_s} \tilde{w}_{n_s,i} g(t_i) &= \sum_{i=1}^{n_s} \tilde{w}_{n_s,i} \Big\{ g(s) - (s-t_i) g'(s) + \frac{1}{2}(s-t_i)^2 g''(\theta_i) \Big\} \\ &= g(s) - g'(s) S_1(s) + \frac{1}{2} S_2(s), \end{aligned} \tag{3.C.11}$$

where $S_1(s) = \sum_{i=1}^{n_s} \tilde{w}_{n_s,i}(s-t_i)$ and $S_2(s) = \sum_{i=1}^{n_s} \tilde{w}_{n_s,i}(s-t_i)^2 g''(\theta_i)$. From there, to show (3.C.9), we next derive the representations of $S_1(s)$ and $S_2(s)$ as $\Delta t \to 0$.

We start with $S_1(s)$. As $s - t_i = (s - t_{n_s}) + (t_{n_s} - t_i)$, we obtain from Lemma 3.C.4 and (3.A.4) that

$$S_1(s) = (s - t_{n_s}) + \gamma \sum_{i=2}^{n_s} (1-\gamma)^{n_s - i}(t_{n_s} - t_i) + (1-\gamma)^{n_s - 1}(t_{n_s} - t_1)\,.$$

From there, in view of the fact that (1.2) entails $0 \leq s - t_{n_s} \leq \Delta t$ and of the inequalities $t_{n_s} - t_1 \leq t_{n_s} \leq s$, we deduce from Lemma 3.C.3 and (1.1) that, uniformly in $s \in [t_{N_0+1}, t]$,

$$S_1(s) = \Delta t \gamma \sum_{i=2}^{n_s} (n_s - i)(1-\gamma)^{n_s - i} + O(\Delta t) = \Delta t \gamma \sum_{j=0}^{n_s - 2} j (1-\gamma)^j + O(\Delta t)\,.$$

An application of (3.C.4) with $n = n_s - 2$ together with the inequality $n_s \leq s/\Delta t$ in (1.2) then yields that, uniformly in $s \in [t_{N_0+1}, t]$,

$$S_1(s) = \Delta t \gamma \Big\{ \frac{1}{\gamma^2} - \frac{1}{\gamma} - \Big(\frac{1}{\gamma^2} + \frac{n_s - 2}{\gamma} \Big)(1-\gamma)^{n_s - 1} \Big\} + O(\Delta t) = \frac{\Delta t}{\gamma} + O(\Delta t)\,, \tag{3.C.12}$$

where the last equality is a consequence of the fact that Conditions (C1) and (C2) ensure the result stated in Lemma 3.C.3.

We next turn our attention to $S_2(s)$ defined below (3.C.11). In view of (3.A.4) and of the assumption that $\|g''\|_\infty \leq M$ in the lemma's statement, we first compute that

$$|S_2(s)| \leq M \Big\{ \gamma \sum_{i=2}^{n_s} (1-\gamma)^{n_s - i}(s - t_i)^2 + (1-\gamma)^{n_s - 1}(s - t_1)^2 \Big\}$$

$$= M\gamma \sum_{i=2}^{n_s} (1-\gamma)^{n_s-i} \{s - t_{n_s}) + (t_{n_s} - t_i)\}^2 + o(\Delta t)\,,$$

uniformly in $s \in [t_{N_0+1}, t]$, where we use Lemma 3.C.3 together with the fact that $s - t_1 \leq t - t_1 = O(1)$.

From there, as (1.2) implies $s - n_s = O(\Delta t)$ uniformly in $s \in [t_{N_0+1}, t]$, it follows that, uniformly in $s \in [t_{N_0+1}, t]$,

$$\begin{aligned}
|S_2(s)| &\leq M\gamma \sum_{i=2}^{n_s} (1-\gamma)^{n_s-i} \{(t_{n_s} - t_i) + O(\Delta t)\}^2 + o(\Delta t) \\
&= M\gamma \sum_{i=2}^{n_s} (1-\gamma)^{n_s-i} (t_{n_s} - t_i)^2 + O(\Delta t) \times \gamma \sum_{i=0}^{n_s-2} (1-\gamma)^i \\
&= M\gamma \sum_{i=2}^{n_s} (1-\gamma)^{n_s-i} (t_{n_s} - t_i)^2 + O(\Delta t) \\
&= M\gamma \Delta t^2 \sum_{i=0}^{n_s-2} i^2 (1-\gamma)^i + O(\Delta t)\,,
\end{aligned}$$

where, to obtain the one-to-last line, we used (3.C.2) in Lemma 3.C.2, while the last equality followed by the fact that $t_{n_s} - t_i = \Delta t(n_s - i)$.

In view of Lemma 3.C.3, an application of (3.C.5) in Lemma 3.C.2 with $n = n_s - 2$ entails $\sum_{j=0}^{n_s-2} j^2(1-\gamma)^j = 2/\gamma^3 + O(\gamma^{-2})$ as $\Delta t \to 0$. Thus, as Condition (C1) guarantees that $\gamma^{-1} = o(\Delta t^{-1})$, we obtain from the above computations, that

$$|S_2(s)| = 2M\frac{\Delta t^2}{\gamma^2} + O(\Delta t^2/\gamma + \Delta t) = o(\Delta t/\gamma)\,, \text{ uniformly in } s \in [t_{N_0+1}, t]\,.$$

Plugging (3.C.12) and the above equation into (3.C.11) combined with the fact that $\|g'\|_\infty \leq M$ concludes the proof of (3.C.9).

Next we prove (3.C.10). For this, note that an application of the mean-value theorem to

$g(t_i)$ with θ_i between t_i and s yields

$$\sum_{i=1}^{n_s} \tilde{w}_{n_s,i}^2 g(t_i) = \sum_{i=1}^{n_s} \tilde{w}_{n_s,i}^2 \big\{ g(s) - (s - t_i) g'(\theta_i) \big\} = \frac{\gamma}{2} g(s) - S_3(s) + o(\Delta t), \quad (3.C.13)$$

uniformly in $s \in [t_{N_0+1}, t]$, where we used Lemma 3.C.5 and the fact that $\|g\|_\infty \leq M$, and where we let $S_3(s) = \sum_{i=1}^{n_s} \tilde{w}_{n_s,i}^2 (s - t_i) g'(\theta_i)$.

Recalling (3.A.4) and using the fact that $\|g'\|_\infty \leq M$ and $0 < \gamma < 1$, we have uniformly in $s \in [t_{N_0+1}, t]$ that

$$\begin{aligned}
S_3(s) &\leq M \bigg\{ (s - t_{n_s}) \sum_{i=1}^{n_s} \tilde{w}_{n_s,i}^2 + \sum_{i=1}^{n_s} \tilde{w}_{n_s,i}^2 (t_{n_s} - t_i) \bigg\} = M \Delta t \sum_{i=1}^{n_s} \tilde{w}_{n_s,i}^2 (n_s - i) + o(\Delta t) \\
&= M \bigg\{ \Delta t (n_s - 1)(1 - \gamma)^{2(n_s - 1)} + \Delta t \gamma^2 \sum_{i=2}^{n_s} (n_s - i)(1 - \gamma)^{2(n_s - i)} \bigg\} + o(\Delta t) \\
&= M \Delta t \gamma^2 \sum_{j=0}^{n_s - 2} j (1 - \gamma)^{2j} + o(\Delta t) \\
&\leq M \Delta t \gamma^2 \sum_{j=0}^{n_s - 2} j (1 - \gamma)^{j} + o(\Delta t) = O(\Delta t), \qquad (3.C.14)
\end{aligned}$$

where we used (1.1) and (1.2) to get the first inequality, Lemma 3.C.3 to get the third equality, and where the last equality followed from the fact that, under Conditions (C1) and (C2), Lemma 3.C.3 combined with (3.C.4) with $n = n_s - 2$ yields $\sum_{j=0}^{n_s - 2} j (1 - \gamma)^j = O(\gamma^{-2})$.

Now, plugging (3.C.14) into (3.C.13) concludes the proof of (3.C.10) and hence of Lemma 3.C.6. □

The next lemma gives the expectation and the variance of $\tilde{r}$ defined in (3.A.1).

Lemma 3.C.7. *Let $t > 0$ satisfy* (3.20). *Under Conditions (C1), (C2), (C4)–(C8) and (C10), if $(\gamma, h) \in \square$ with $\square$ defined in Condition (C1), then we have uniformly in $(x, s) \in \mathbb{R} \times [t_{N_0+1}, t]$ and as $\Delta t \to 0$ that*

$$\mathrm{E}\{\tilde{r}(x, s | \gamma, h)\} = (m \cdot f)(x, s) - \frac{\Delta t}{\gamma} \big\{ (m_x \cdot f)(x, s) + (m \cdot f_x)(x, s) \big\}$$

$$+ \frac{1}{2}h^2\mu_{K,2}\psi_1(x,s) + o(h^2 + \Delta t/\gamma) \qquad (3.C.15)$$

and

$$\text{var}\{\tilde{r}(x,s|\gamma,h)\} = \frac{R_K}{2}\frac{\gamma}{h}\psi_2(x,s) + o(\gamma/h)\,, \qquad (3.C.16)$$

where $\mu_{K,2}^2$ *and* R_K *are defined in* (3.32) *and*

$$\psi_1(x,s) = (m \cdot f_{xx})(x,s) + 2(m_x \cdot f_x)(x,s) + (m_{xx} \cdot f)(x,s), \qquad (3.C.17)$$

$$\psi_2(x,s) = (m^2 \cdot f)(x,s) + \sigma^2 f(x,s), \qquad (3.C.18)$$

with σ^2 *defined in Condition (C7).*

Proof. We first prove (3.C.15). Using the model assumption in (1.3) that $Y_i = m(X_i,t_i) + \epsilon_i$, with ϵ_i in dependent of X_i (see Condition (C7)), we deduce from the definition of $\tilde{r}$ at (3.A.1) that uniformly in $(x,s) \in \mathbb{R} \times [t_{N_0+1}, t]$,

$$\begin{aligned}\text{E}\{\tilde{r}(x,s|\gamma,h)\} &= \sum_{i=1}^{n_s} \tilde{w}_{n_s,i}\,\text{E}\{K_h(x - X_i)m(X_i,t_i)\} \\ &= \sum_{i=1}^{n_s} \tilde{w}_{n_s,i} \int K(u)m(x-hu,t_i)f(x-hu,t_i)\,\text{d}u. \end{aligned} \qquad (3.C.19)$$

To derive the asymptotic representation of the above expression under $\Delta t \to 0$, which implies $h \to 0$ under Condition (C1), first note that Conditions (C4) and (C5) allow us to deduce from a second-order Taylor expansions of $m(x-hu,t_i)$ and $f(x-hu,t_i)$ around x that there exist $\theta_{i1}, \theta_{i2} \in [0,1]$ such that

$$\begin{aligned} m(x-hu,t_i) &= m(x,t_i) - hum_x(x,t_i) \\ &\qquad + \frac{1}{2}h^2u^2m_{xx}(x,t_i) - \frac{1}{6}h^3u^3m_{xxx}(x-\theta_{i1}hu,t_i)\,, \\ f(x-hu,t_i) &= f(x,t_i) - huf_x(x,t_i) \end{aligned}$$

$$+ \frac{1}{2}h^2u^2 f_{xx}(x,t_i) - \frac{1}{6}h^3u^3 f_{xxx}(x-\theta_{i2}hu,t_i). \quad (3.C.20)$$

Plugging the above equations into (3.C.19), and using the fact that under Condition (C6) $\int uK(u)\,\mathrm{d}u = 0$ and that K vanishes outside of $[-1,1]$, we obtain that uniformly in x, and as $\Delta t \to 0$,

$$\mathrm{E}\{K_h(x-X_i)m(X_i,t_i)\} = (m\cdot f)(x,t_i) + \frac{1}{2}h^2\mu_{K,2}\psi_1(x,t_i) + o(h^2)\,, \quad (3.C.21)$$

where ψ_1 is defined at (3.C.17) and $\mu_{K,2}$ is defined in (3.32).

From there, as under Conditions (C4) and (C5) both $(m\cdot f)$ and ψ_1 as functions of t satisfy the conditions of Lemma 3.C.6, the proof of (3.C.15) follows from (3.C.19), (3.C.21), Lemma 3.C.6 and Lemma 3.C.4.

Next we prove (3.C.16). For this, first note that $\mathrm{var}\{\tilde{r}(x,s|\gamma,h)\} = V_1 + V_2$, where

$$V_1 = \sum_{i=1}^{n_s} \tilde{w}_{n_s,i}^2\,\mathrm{var}\{K_h(x-X_i)Y_i\}, \quad \text{and} \quad |V_2| \le \sum_{1\le|i-j|\le n_s-1} \tilde{w}_{n_s,i}\tilde{w}_{n_s,j}c_{ij}, \quad (3.C.22)$$

where, under (1.3) and Condition (C7),

$$c_{ij} = \left|\,\mathrm{cov}\{K_h(x-X_i)m(X_i,t_i), K_h(x-X_j)m(X_j,t_j)\}\right|. \quad (3.C.23)$$

In what follows, we show that, uniformly in $(x,s)\in\mathbb{R}\times[t_{N_0+1},t]$ and as $\Delta t\to 0$, $V_2 = o(\gamma/h)$ and V_1 is equivalent to the first term on the right hand side of the equality at (3.C.16).

We start by deriving the asymptotic order of V_2. For this, recalling the definition of g_{ij} at (3.31), we compute that under (C6), we have, on the one hand, that

$$\begin{aligned} c_{ij} &= \left|\iint K(u)K(v)m(x-hu,t_i)m(x-hv,t_j)g_{ij}(x-hu,x-hv)\,\mathrm{d}u\,\mathrm{d}v\right| \\ &\le \|m\|_\infty^2\|g_{ij}\|_\infty\,. \end{aligned}$$

In view of the above display, and since Condition (C4) implies $\|m\|_\infty^2 < \infty$, we obtain from

an application of Lemma 3.B.4 that under Condition (C8), there exists a constant $C_1 > 0$ such that

$$c_{ij} \leq C_1 \alpha_0^{1/3} \big\{ \sigma(X_i), \sigma(X_j) \big\} \leq C_1 \alpha^{1/3}(\Delta t, |i-j|), \tag{3.C.24}$$

where α is defined at (3.B.3). On the other hand, in view of Lemma 3.B.5, we obtain from (3.C.23) that under Conditions (C4) and (C6), there exists a constant $C_2 > 0$ such that

$$c_{ij} \leq \frac{C_2}{h^2} \alpha(\Delta t, |i-j|). \tag{3.C.25}$$

From the definition of V_2 at (3.C.22), and in view of the bounds at (3.C.24) and (3.C.25), we use the fact that there exist constants $C, a > 0$ such that that $\alpha(\Delta t, k) \leq C k^{-a}$ (see Condition (C8)) to derive that, uniformly in $(x, s) \in \mathbb{R} \times [t_{N_0+1}, t]$,

$$\begin{aligned} |V_2| &\leq \sum_{1 \leq |i-j| \leq n_s - 1} \tilde{w}_{n_s,i} \tilde{w}_{n_s,j} \min\big\{ C_1 \alpha^{1/3}(\Delta t, |i-j|), C_2 h^{-2} \alpha(\Delta t, |i-j|) \big\} \\ &\leq 2 \max(C_1, C_2) \sum_{k=1}^{n_s - 1} \sum_{j=k+1}^{n_s} \tilde{w}_{n_s,j-k} \tilde{w}_{n_s,j} \min\big\{ \alpha^{1/3}(\Delta t, k), h^{-2} \alpha(\Delta t, k) \big\} \\ &\leq 2C \max(C_1, C_2) \sum_{k=1}^{n_s - 1} \min\left(k^{-a/3}, h^{-2} k^{-a} \right) \sum_{j=k+1}^{n_s} \tilde{w}_{n_s,j-k} \tilde{w}_{n_s,j} \,. \end{aligned} \tag{3.C.26}$$

From there, to derive the asymptotic order of V_2, we use (3.A.4) to compute that for any k, and uniformly in $s \in [t_{N_0+1}, t]$,

$$\begin{aligned} \sum_{j=k+1}^{n_s} \tilde{w}_{n_s,j-k} \tilde{w}_{n_s,j} &= \gamma (1-\gamma)^{2n_s - k - 2} + \gamma^2 \sum_{j=k+2}^{n_s} (1-\gamma)^{2(n_s - j) + k} \\ &\leq \gamma (1-\gamma)^{n_s - 1} + \gamma^2 \sum_{j=0}^{n_s - 2} (1-\gamma)^{2j} = O(\gamma) \end{aligned} \tag{3.C.27}$$

where we used (3.C.3) and the fact that under Conditions (C1) and (C2), the result stated in Lemma 3.C.3 holds.

Now let $k_0 = \min\big(\lfloor h^{-3/a} \rfloor, n_s - 1\big)$ and observe that Conditions (C1) and (C8) guarantee that $k_0 \geq 1$ and that $k_0 \to \infty$ as $\Delta t \to 0$. As under Condition (C8) we have $a > 3/2$, an application of Lemma 3.B.2 entails, uniformly in $s \in [t_{N_0+1}, t]$:

$$\sum_{k=1}^{n_s-1} \min\big(k^{-a/3}, h^{-2}k^{-a}\big) \leq \sum_{k=1}^{k_0} k^{-a/3} + h^{-2} \sum_{k=k_0+1}^{n_s} k^{-a} = O(k_0^{1-a/3}) + O(h^{-2}k_0^{1-a})\,.$$

Since Condition (C8) guarantees $a > 3/2$, and as $h \to 0$ as $\Delta t \to 0$ (see Condition (C1)), we deduce that as $\Delta t \to 0$, and uniformly in $s \in [t_{N_0+1}, t]$,

$$\sum_{k=1}^{n_s-1} \min\big(k^{-a/3}, h^{-2}k^{-a}\big) = o(h^{-1})\,. \tag{3.C.28}$$

Plugging the latter result and into (3.C.26) proves that $|V_2| = o(\gamma h^{-1})$ uniformly in $(x, s) \in \mathbb{R} \times [t_{N_0+1}, t]$. Hence, to derive (3.C.16), it remains to prove that V_1 is equivalent to the first term on the right hand side of the equality at (3.C.16).

In order to do this, using the definition of V_1 at (3.C.22), we compute from (1.3) and Condition (C7) that

$$\begin{aligned}
\mathrm{E}\big[\{K_h(x - X_i)Y_i\}^2\big] &= \frac{1}{h^2}\,\mathrm{E}\left\{K^2\Big(\frac{x - X_i}{h}\Big)m^2(X_i, t_i)\right\} + \frac{\sigma^2}{h^2}\,\mathrm{E}\Big\{K^2\Big(\frac{x - X_i}{h}\Big)\Big\} \\
&= \frac{1}{h}\int K^2(u)m^2(x - hu, t_i)f(x - hu, t_i)\,\mathrm{d}u \\
&\qquad + \frac{\sigma^2}{h}\int K^2(u)f(x - hu, t_i)\,\mathrm{d}u\,. \qquad (3.C.29)
\end{aligned}$$

Applying the mean-value theorem to $m^2(x - hu, t_i)$ and $f(x - hu, t_i)$ in (3.C.29), there exists constants $\theta_{i1}, \theta_{i2} \in [0, 1]$ such that the following equality holds uniformly in $x \in \mathbb{R}$ and $i \in \{1, \ldots, n_s\}$:

$$\begin{aligned}
\mathrm{E}\big[\{K_h(x - X_i)Y_i\}^2\big] = \frac{1}{h}\int K^2(u)\big\{&m^2(x, t_i) - 2hu(m \cdot m_x)(x - \theta_{i1}hu, t_i)\big\} \\
&\times \big\{f(x, t_i) - huf_x(x - \theta_{i2}hu, t_i)\big\}\,\mathrm{d}u
\end{aligned}$$

$$+ \frac{\sigma^2}{h} \int K^2(u)\big\{f(x,t_i) - huf_x(x-\theta_{i2}hu,t_i)\big\}\,\mathrm{d}u$$
$$= \frac{R_K}{h}\psi_2(x,t_i) + O(1)\,, \tag{3.C.30}$$

where R_K is defined at (3.32), ψ_2 is as at (3.C.18) and where we used Conditions (C4) to (C6) to get the last line.

Since $h \to 0$ as $\Delta t \to 0$ (see Condition (C1)), we have under Conditions (C4) and (C5) that (3.C.21) implies $\mathrm{E}\{K_h(x-X_i)Y_i\} = \mathrm{E}\{K_h(x-X_i)m(X_i,t_i)\} = O(1)$ uniformly in $x \in \mathbb{R}$ and $i \in \{1,\dots,n_s\}$. Hence from (3.C.30) we have uniformly in $x \in \mathbb{R}$ and $i \in \{1,\dots,n_s\}$ that

$$\mathrm{var}\{K_h(x-X_i)Y_i\} = \frac{R_K}{h}\psi_2(x,t_i) + O(1)\,. \tag{3.C.31}$$

To conclude, note from (3.C.18) that under Conditions (C4) and (C5) $t \mapsto \psi_2(\cdot,t)$ satisfies the conditions of Lemma 3.C.6. Hence, plugging (3.C.31) in (3.C.22) and using Lemma 3.C.6, we have, uniformly in $(x,s) \in \mathbb{R} \times [t_{N_0+1},t]$ and as $\Delta t \to 0$ that

$$V_1 = \frac{R_K}{h}\sum_{i=1}^{n_s} \tilde{w}_{n_s,i}^2 \psi_2(x,t_i) = \frac{R_K}{2}\frac{\gamma}{h}\psi_2(x,s) + o(\gamma/h) \tag{3.C.32}$$

This concludes the proof of Lemma 3.C.7. □

The next lemma gives the expectation and variance of $\tilde{f}$ defined in (3.A.1).

Lemma 3.C.8. *Let $t > 0$ satisfy* (3.20). *Under Conditions (C1), (C2), (C5)–(C8) and (C10), for all $(\gamma,h) \in \square$ with $\square$ defined in Condition (C1), we have uniformly in $(x,s) \in \mathbb{R} \times [t_{N_0+1},t]$ and as $\Delta t \to 0$ that*

$$\mathrm{E}\{\tilde{f}(x,s|\gamma,h)\} = f(x,s) + \frac{1}{2}f_{xx}(x,s)\mu_{K,2}h^2 - f_t(x,s)\frac{\Delta t}{\gamma} + o(h^2 + \Delta t/\gamma), \tag{3.C.33}$$
$$\mathrm{var}\{\tilde{f}(x,s|\gamma,h)\} = \frac{R_K}{2}\frac{\gamma}{h}f(x,s) + o(\gamma/h), \tag{3.C.34}$$

where $\mu_{K,2}^2$ and R_K are defined in (3.32).

Proof. We start with the proof of (3.C.33). From (3.A.1), we have uniformly in $(x, s) \in \mathbb{R} \times [t_{N_0+1}, t]$ that

$$\mathrm{E}\{\tilde{f}(x, s|\gamma, h)\} = \sum_{i=1}^{n_s} \tilde{w}_{n_s,i} \, \mathrm{E}[K_h(x - X_i)] = \sum_{i=1}^{n_s} \tilde{w}_{n_s,i} \int K(u) f(x - hu, t_i) \, \mathrm{d}u \,. \quad (3.C.35)$$

Using that fact that under Condition (C5) the Taylor expansion of f at (3.C.20) holds, we compute that under Condition (C6), $\mathrm{E}\{K_h(x - X_i)\} = f(x, t_i) + 2^{-1}h^2\mu_{K,2}f_{xx}(x, t_i) + o(h^2)$ uniformly in $x \in \mathbb{R}$ and in $i \in \{1, \ldots, n_s\}$. Therefore, as under Condition (C5), both f and f_x as functions of t satisfy the conditions of Lemma 3.C.6, (3.C.33) follows from (3.C.35) and Lemma 3.C.6.

Next we prove (3.C.34). We first decompose $\mathrm{var}\{\tilde{f}(x, s|\gamma, h)\} = V_3 + V_4$, where

$$V_3 = \sum_{i=1}^{n_s} \tilde{w}_{n_s,i}^2 \, \mathrm{var}\{K_h(x - X_i)\} \quad \text{and} \quad |V_4| \leq \sum_{1 \leq |i-j| \leq n_s - 1} \tilde{w}_{n_s,i}\tilde{w}_{n_s,j} d_{ij}, \quad (3.C.36)$$

with $d_{ij} = |\operatorname{cov}\{K_h(x - X_i), K_h(x - X_j)\}|$. In what follows, to show that (3.C.34) holds uniformly in $(x, s) \in \mathbb{R} \times [t_{N_0+1}, t]$ as $\Delta t \to 0$, we will prove that, uniformly in $(x, s) \in \mathbb{R} \times [t_{N_0+1}, t]$, we have $V_4 = o(\gamma/h)$ and that V_3 is equivalent, up to a negligible term, to the first term on the right hand side of (3.C.34).

To show that $V_4 = o(\gamma/h)$, note that taking $m \equiv 1$ at (3.C.23) allows us to deduce from (3.C.24) and (3.C.25) that under our Conditions, there exists a constant $\kappa > 0$ such that

$$d_{ij} \leq \kappa \min\left\{\alpha^{1/3}(\Delta t, |i - j|), h^{-2}\alpha(\Delta t, |i - j|)\right\} \leq C\kappa \min\left\{|i - j|^{-a/3}, h^{-2}|i - j|^{-a}\right\}, \quad (3.C.37)$$

where the last line is a consequence of Condition (C8). Therefore, we have, uniformly in $(x, s) \in \mathbb{R} \times [t_{N_0+1}, t]$, that

$$\begin{aligned} |V_4| &= 2\sum_{k=1}^{n_s-1} \sum_{j=k+1}^{n_s} \tilde{w}_{n_s,i}\tilde{w}_{n_s,j} d_{ij} \leq 2\kappa \sum_{k=1}^{n_s-1} \min\left(k^{-a/3}, h^{-2}k^{-a}\right) \sum_{j=k+1}^{n_s} \tilde{w}_{n_s,j-k}\tilde{w}_{n_s,j} \\ &= o(\gamma/h)\,, \end{aligned} \quad (3.C.38)$$

where, to obtain the last line, we used (3.C.27) and (3.C.28).

We now turn our attention to V_3. Recalling the definition of R_K in (3.32), we apply the mean-value theorem to $f(x-hu,t_i)$ in (3.C.36) to deduce that there exists a constant $\theta_{i1} \in [0,1]$ such that the following equality holds uniformly in $x \in \mathbb{R}$ and $i \in \{1,\dots,n_s\}$:

$$\begin{aligned}
\mathrm{E}\big[\{K_h(x-X_i)\}^2\big] &= \frac{1}{h}\int K^2(u)\big\{f(x,t_i) - huf_x(x-\theta_i hu,t_i)\big\}\,\mathrm{d}u \\
&= \frac{R_K}{h}f(x,t_i) - \int uK^2(u)f_x(x-\theta_i hu,t_i)\,\mathrm{d}u \\
&= \frac{R_K}{h}f(x,t_i) + O(1)\,, \qquad (3.C.39)
\end{aligned}$$

where, to get the last line, we used Conditions (C5) and (C6).

Now from similar computations we have under Condition (C5) and (C6) that $\mathrm{E}\{K_h(x - X_i)\} = O(1)$ uniformly in x. Hence from (3.C.39) we have

$$\mathrm{var}\{K_h(x-X_i)\} = \frac{R_K}{h}f(x,t_i) + O(1), \text{uniformly in } x \text{ and in } i. \qquad (3.C.40)$$

Then, under Condition (C5), f as a function of t satisfies the conditions of Lemma 3.C.6. Hence, plugging (3.C.40) in the definition of V_3 at (3.C.36), and using Lemma 3.C.6, we deduce that it holds uniformly in $(x,s) \in \mathbb{R} \times [t_{N_0+1}, t]$ that

$$V_3 = \frac{R_K}{h}\sum_{i=1}^{n_s}\tilde{w}_{n_s,i}^2 f(x,t_i) = \frac{R_K}{2}\frac{\gamma}{h}f(x,s) + o(\gamma/h). \qquad (3.C.41)$$

Hence, (3.C.34) follows from (3.C.36), (3.C.38) and the above equation. This concludes the proof of Lemma 3.C.8. □

The next lemma derives the uniform consistency of $\tilde{f}$ defined in (3.A.1). Recall from (3.21) that ℓ_t defines the unique integer such that $t \in [t_{N_{\ell_t-1}}, t_{N_{\ell_t}})$, where N_k denotes the number of observations arriving up to the k-th ensemble substitution.

Lemma 3.C.9. *Assume that Conditions (C1) to (C3), (C5) to (C8) and (C10) hold. Suppose* $(\gamma,h) \in \square$ *with* $\square$ *defined in Condition (C1). For any fixed* $t > 0$ *and any compact interval*

$\mathcal{K} \subset \mathbb{R}$, *we have, as* $\Delta t \to 0$, *that*

$$\sup_{x \in \mathcal{K}} \sup_{(\gamma,h) \in I^{\ell_t}_{\gamma,h}} |\tilde{f}(x,t|\gamma,h) - f(x,t)| = o_p(1)\,.$$

Proof. Assume without loss of generality that $\mathcal{K} = [0,1]$. Set $P_{\Delta t} = \lfloor \log(\Delta t^{-2})/\Delta t^{2a_2} \rfloor$ and let $\mathcal{K}_0 = \{x_i : x_i = i/P_{\Delta t} \text{ for } i = 0, \ldots, P_{\Delta t}\}$ be a grid of equally spaced points in $[0,1]$. We decompose

$$\sup_{x \in \mathcal{K}} \sup_{(\gamma,h) \in I^{\ell_t}_{\gamma,h}} \left|\tilde{f}(x,t|\gamma,h) - f(x,t)\right| \leq J_1 + J_2, \tag{3.C.42}$$

where

$$J_1 = \sup_{x \in \mathcal{K}_0} \sup_{(\gamma,h) \in I^{\ell_t}_{\gamma,h}} \left|\tilde{f}(x,t|\gamma,h) - f(x,t)\right|, \tag{3.C.43}$$

$$J_2 = \sup_{|x-x'| \leq P^{-1}_{\Delta t}} \sup_{(\gamma,h) \in I^{\ell_t}_{\gamma,h}} \left|\tilde{f}(x,t|\gamma,h) - f(x,t) - \{\tilde{f}(x',t|\gamma,h) - f(x',t)\}\right|. \tag{3.C.44}$$

In what follows, to prove Lemma 3.C.9, we show that $J_1 = o_p(1)$ and $J_2 = o_p(1)$.

Dealing first with J_1, note that under Conditions (C1), (C2), (C5) to (C8) and (C10), for all $(\gamma, h) \in \square$ with $\square$ defined in Condition (C1), we have uniformly in x and as $\Delta t \to 0$ that $\mathrm{E}\{\tilde{f}(x,t|\gamma,h)\} = f(x,t) + O(h^2 + \Delta t/\gamma)$ (see Lemma 3.C.8). Therefore, since $h \to 0$ and $\Delta t/\gamma \to 0$ as $\Delta t \to 0$ (see Condition (C1)), we deduce that as $\Delta t \to 0$,

$$J_1 = \sup_{x \in \mathcal{K}_0} \sup_{(\gamma,h) \in I^{\ell_t}_{\gamma,h}} \left|\tilde{f}(x,t|\gamma,h) - \mathrm{E}\{\tilde{f}(x,t|\gamma,h)\}\right| + o_p(1)\,. \tag{3.C.45}$$

In view of (3.C.45), to prove that $J_1 = o_p(1)$, it suffices to show that for any $\eta > 0$,

$$\lim_{\Delta t \to 0} \mathrm{P}\left[\sup_{x \in \mathcal{K}_0} \sup_{(\gamma,h) \in I^{\ell_t}_{\gamma,h}} \left|\tilde{f}(x,t|\gamma,h) - \mathrm{E}\{\tilde{f}(x,t|\gamma,h)\}\right| > \eta\right] = 0\,, \tag{3.C.46}$$

i.e. the first term on the right hand side of (3.C.45) converges to 0 in probability as $\Delta t \to 0$.

To prove (3.C.46), our strategy is to apply (3.B.12) in Theorem 3.B.2. For this, recall that under Condition (C6) $\|K\|_\infty < \infty$, and set

$$\xi_{n_t,i} = \frac{\tilde{w}_{n_t,i}}{2\|K\|_\infty}\frac{h}{\gamma}\big[K_h(x - X_i) - \mathrm{E}\{K_h(x - X_i)\}\big]. \tag{3.C.47}$$

Then by (3.A.1), we have

$$\frac{1}{2\|K\|_\infty}\frac{h}{\gamma}\big[\tilde{f}(x,t|\gamma,h) - \mathrm{E}\{\tilde{f}(x,t|\gamma,h)\}\big] = \sum_{i=1}^{n_t}\xi_{n_t,i} := S_{n_t}\,. \tag{3.C.48}$$

From there, to apply (3.B.12) in Theorem 3.B.2 to S_{n_t}, and in view of Condition (C8), we need to verify that $|\xi_{n_t,i}| \leq 1$. For this, in view of (3.A.4) and of Lemma 3.C.3, we have under Condition (C1) and (C2) that $\tilde{w}_{n_t,i}/\gamma = (1-\gamma)^{n_t-i} \leq 1$, for $i = 2,\ldots,n_t$, and $\tilde{w}_{n_t,1}/\gamma = (1-\gamma)^{n_t-1}/\gamma = o(1)$ as $\Delta t \to 0$. Hence, for small enough Δt, $\tilde{w}_{n_t,i}/\gamma \leq 1$, and therefore form the definition of $\xi_{n_t,i}$ at (3.C.47), we have $|\xi_{n_t,i}| \leq 1$ for all i as $\Delta t \to 0$.

Since we have just shown that $|\xi_{n_t,i}| \leq 1$ for all i when Δt is small enough, (3.B.12) in Theorem 3.B.2 applies to S_{n_t}, i.e. for any (fixed) $\lambda > 0$, and letting $\bar{\sigma}^2_{n_t} = \sum_{i,j=1}^{n_t} |\operatorname{cov}(\xi_{n_t,i},\xi_{n_t,j})|$, there exists some constant $\Upsilon = \Upsilon(\lambda,a) > 0$ that depends on λ and on the α-mixing rate a defined in Condition (C8) such that for all $\eta > 0$,

$$\mathrm{P}(S_{n_t} \geq \eta) \leq \Upsilon\left\{\left(1 + \frac{\eta^2}{\lambda\bar{\sigma}^2_{n_t}}\right)^{-\lambda/2} + \frac{n_t}{\lambda}\left(\frac{\lambda}{\eta}\right)^{a+1}\right\}. \tag{3.C.49}$$

As under Condition (C5) f is bounded, we obtain from (3.C.38) and (3.C.41) in the proof of Lemma 3.C.8 that there exists a constant $\upsilon > 0$ such that $\bar{\sigma}^2_{n_t} \leq \upsilon h/\gamma$ when Δt is small enough. Therefore, we obtain from (3.C.49) that for any $\eta_0 > 0$, and for sufficiently small Δt,

$$\begin{aligned}\mathrm{P}\left\{\big|\tilde{f}(x,t|g,h) - \mathrm{E}\{\tilde{f}(x,t|g,h)\}\big| > \frac{\eta_0}{2\|K\|_\infty}\right\} &= \mathrm{P}\big(|S_{n_t}| > \eta_0 h/\gamma\big)\\ &\leq \Upsilon\left\{\left(1 + \frac{\eta_0^2 h/\gamma}{\lambda\upsilon}\right)^{-\lambda/2} + \frac{n_t\lambda^a}{\eta_0^{a+1}(h/\gamma)^{a+1}}\right\}.\end{aligned}$$

As Condition (C1) ensures that $\gamma/h \sim \Delta t^{a_1-a_2} \to 0$ as $\Delta t \to 0$, and since $n_t \leq t/\Delta t$

(see (1.2)), we deduce from the above equation that there exists a constant $\Upsilon'(\lambda, a, \eta_0) > 0$ that depends on λ, a and η_0, independent of $(\Delta t, x, t, \gamma, h)$, such that for sufficiently small Δt,

$$\mathrm{P}\Big[\big|\tilde{f}(x,t|g,h) - \mathrm{E}\{\tilde{f}(x,t|g,h)\}\big| > \frac{\eta_0}{2\|K\|_\infty}\Big] \leq \Upsilon'(\lambda, a, \eta_0)\big\{\Delta t^{\lambda(a_1-a_2)/2} + \Delta t^{(a+1)(a_1-a_2)-1}\big\}\,.$$

We are now ready to show that (3.C.46) holds. For this, fix $\eta > 0$ and set $\lambda \geq 2\{(a+1)(a_1-a_2)-1\}/(a_1-a_2)$. From the inequality above, we deduce that

$$\mathrm{P}\big\{\big|\tilde{f}(x,t|\gamma,h) - \mathrm{E}\{\tilde{f}(x,t|\gamma,h)\}\big| > \eta\big\} \leq 2\Upsilon'(\lambda, a, 2\|K\|_\infty\eta)\Delta t^{(a+1)(a_1-a_2)-1}\,. \quad (3.C.50)$$

Therefore, we obtain using the inequality above that

$$\begin{aligned}
&\mathrm{P}\Big[\sup_{x\in\mathcal{K}_0}\sup_{(\gamma,h)\in I^{\ell_t}_{\gamma,h}}\big|\tilde{f}(x,t|\gamma,h) - \mathrm{E}\{\tilde{f}(x,t|\gamma,h)\}\big| > \eta\Big] \\
&\qquad\leq \sum_{x\in\mathcal{K}_0}\sum_{(\gamma,h)\in I^{\ell_t}_{\gamma,h}}\mathrm{P}\big\{\big|\tilde{f}(x,t|\gamma,h) - \mathrm{E}\{\tilde{f}(x,t|\gamma,h)\}\big| > \eta\big\} \\
&\qquad\leq (P_{\Delta t}+1)\#(I^{\ell_t}_{\gamma,h}) \times 2\Upsilon'(\lambda, a, 2\|K\|_\infty\eta)\Delta t^{(a+1)(a_1-a_2)-1} \\
&\qquad= O\big\{P_{\Delta t}\Delta t^{-c}\Delta t^{(a+1)(a_1-a_2)-1}\big\}\,,
\end{aligned}$$

where we used Conditions (C1), (C3) and (C8) to obtain the last equality. In view of the definition of $P_{\Delta t}$ above (3.C.42), it follows that the last line of the display above is $o(1)$ as $\Delta t \to 0$. Hence we have proved (3.C.46), which also implies that $J_1 = o_p(1)$ (see (3.C.45)).

Next we prove that $J_2 = o_p(1)$. By Condition (C5), note that for any $x, x' \in \mathbb{R}$, an application of the mean-value Theorem entails $|f(x,t) - f(x',t)| \leq \|f'\|_\infty|x - x'|$. Hence, recalling the definition of $P_{\Delta t}$ above (3.C.42), we have

$$J_{21} := \sup_{|x-x'|\leq P^{-1}_{\Delta t}}\big|f(x,t) - f(x',t)\big| \leq \frac{\|f'\|_\infty}{P_{\Delta t}} \leq \frac{\|f'\|_\infty\Delta t^{2a_2}}{\log(\Delta t^{-2})}\,. \quad (3.C.51)$$

Also, note that under Condition (C6), for any i and $x, x' \in \mathbb{R}$, we have $|K_h(X_i - x) - K(X_i - x')| \leq \|K'\|_\infty h^{-1}|x - x'|$, where $\|K'\|_\infty < \infty$. Therefore, in view of (3.A.1), and using again the definition of $P_{\Delta t}$ above (3.C.42), we have

$$\begin{aligned} J_{22} &:= \sup_{|x-x'|\leq P_{\Delta t}^{-1}} \sup_{(\gamma,h)\in I_{\gamma,h}^{\ell_t}} \left|\tilde{f}(x,t|\gamma,h) - \tilde{f}(x',t|\gamma,h)\right| \\ &\leq \sup_{(\gamma,h)\in I_{\gamma,h}^{\ell_t}} \left\{ \frac{1}{h} \sum_{i=1}^{n_t} \tilde{w}_{n_t,i} \sup_{|x-x'|\leq P_{\Delta t}^{-1}} \left| K\left(\frac{x - X_i}{h}\right) - K\left(\frac{x' - X_i}{h}\right) \right| \right\} \\ &\leq \frac{\|K'\|_\infty}{h^2 P_{\Delta t}} \sup_{(\gamma,h)\in I_{\gamma,h}^{\ell_t}} \left(\sum_{i=1}^{n_t} \tilde{w}_{n_t,i} \right) = \|K'\|_\infty \delta^{-2} \{\log(\Delta t^{-2})\}^{-1}, \end{aligned} \tag{3.C.52}$$

where δ is as in Condition (C1) and where we used Lemma 3.C.4 to get the last equality.

In view of (3.C.44), since $J_2 \leq J_{21} + J_{22}$, we deduce from (3.C.51) and (3.C.52) that as $\Delta t \to 0$ we have $J_2 = o_p(1)$. The announced result follows. □

The next lemma proves the uniform consistency of $\tilde{r}$ defined in (3.A.1). Recall the definition of ℓ_t at (3.21).

Lemma 3.C.10. *Assume that Conditions (C1)–(C8) hold. Suppose $(\gamma, h) \in \square$ with $\square$ defined in Condition (C1). For a given time $t > 0$ and a compact set $\mathcal{K} \subset \mathbb{R}$, we have*

$$\sup_{x\in\mathcal{K}} \sup_{(\gamma,h)\in I_{\gamma,h}^{\ell_t}} |\tilde{r}(x,t|\gamma,h) - (m \cdot f)(x,t)| = o_p(1), \text{ as } \Delta t \to 0.$$

Proof. As in the proof of Lemma 3.C.9, assume without loss of generality that $\mathcal{K} = [0,1]$, set $P_{\Delta t} = \lfloor \log(\Delta t^{-2})/\Delta t^{2a_2} \rfloor$ and let $\mathcal{K}_0 = \{x_i : x_i = i/P_{\Delta t} \text{ for } i = 0, \ldots, P_{\Delta t}\}$. We have

$$\sup_{x\in\mathcal{K}} \sup_{(\gamma,h)\in I_{\gamma,h}^{\ell_t}} \left|\tilde{r}(x,t|\gamma,h) - (m \cdot f)(x,t)\right| \leq J_3 + J_4,$$

where

$$J_3 = \sup_{x\in\mathcal{K}_0} \sup_{(\gamma,h)\in I^{\ell_t}_{\gamma,h}} \big|\tilde{r}(x,t|\gamma,h) - (m\cdot f)(x,t)\big|, \tag{3.C.53}$$

$$J_4 = \sup_{|x-x'|\le P_{\Delta t}^{-1}} \sup_{(\gamma,h)\in I^{\ell_t}_{\gamma,h}} \big|\tilde{r}(x,t|\gamma,h) - (m\cdot f)(x,t) - \{\tilde{r}(x',t|\gamma,h) - (m\cdot f)(x',t)\}\big| . \tag{3.C.54}$$

In view of the above, to prove Lemma 3.C.9, it suffices to show that $J_3 = o_p(1)$ and that $J_4 = o_p(1)$ as $\Delta t \to 0$.

Starting with the asymptotic order of J_3, note that Lemma 3.C.7 ensures that under Conditions (C1), (C2), (C4)–(C8), $\mathrm{E}\{\tilde{r}(x,t|\gamma,h)\} = (m\cdot f)(x,t) + o_p(1)$ as $\Delta t \to 0$ uniformly in $(h,\gamma) \in \square$ and $x \in \mathbb{R}$. Hence to prove that $J_3 = o_p(1)$, it suffices to demonstrate that for all $\eta > 0$,

$$P\Big[\sup_{x\in\mathcal{K}_0} \sup_{(\gamma,h)\in I^{\ell_t}_{\gamma,h}} \big|\tilde{r}(x,t|\gamma,h) - \mathrm{E}\{\tilde{r}(x,t|\gamma,h)\}\big| > \eta\Big] \to 0\,, \text{ as } \Delta t \to 0. \tag{3.C.55}$$

To prove that (3.C.55) holds, our goal is to apply (3.B.11) of Theorem 3.B.2. For this, recalling the definition of $\tilde{w}_{n_t,i}$ at (3.A.4), let $\xi_{n_t,i} = \tilde{w}_{n_t,i} h\gamma^{-1}[K_h(x-X_i)Y_i - \mathrm{E}\{K_h(x-X_i)Y_i\}]$. Then, in view of (3.A.1), we have

$$\frac{h}{\gamma}\big[\tilde{r}(x,t|\gamma,h) - \mathrm{E}\{\tilde{r}(x,t|\gamma,h)\}\big] = \sum_{i=1}^{n_t} \xi_{n_t,i} := S_{n_t}\,. \tag{3.C.56}$$

From there, to apply (3.B.11) in Theorem 3.B.2 to S_{n_t}, and in view of Condition (C8), we need to verify that there exists $p > 2$ and constants $C_0, C_1 > 0$ such that $\mathrm{P}(|\xi_{n,i}| > x) \le C_1 x^{-p},$ for any $x > C_0$.

To show that it is the case, first recall from below (3.C.48) that we have $\tilde{w}_{n_t,i}\gamma^{-1} \le 1$ for all $i \in \{1,\dots,n_t\}$ as $\Delta t \to 0$. As $Y_i = m(X_i,t_i) + \epsilon_i$ (see (1.3)), it therefore follows that $\tilde{w}_{n_t,i} h\gamma^{-1}|K_h(X_i-x)Y_i| \le \|K\|_\infty(\|m\|_\infty+1)(|\epsilon_i|+1)$ for all $i \in \{1,\dots,n_t\}$. As under Condition (C7) $\mathrm{E}\{K_h(X_i-x)Y_i\} = \mathrm{E}\{K_h(X_i-x)m(X_i,t)\} = m\cdot f(x,t_i) + o(h)$ (see (3.C.21)),

it follows that under Condition (C1), $\tilde{w}_{n_t,i}h\gamma^{-1}|\,\mathrm{E}\{K_h(X_i-x)Y_i\}| \le h\|m\|_\infty\|f\|_\infty = o(1)$ as $\Delta t \to 0$.

In view of the above arguments, we deduce that when Δt is sufficiently small, $|\xi_{n_t,i}| \le 2\|K\|_\infty(\|m\|_\infty+1)(|\epsilon_i|+1)$. Therefore, for any $x > 2\|K\|_\infty(\|m\|_\infty+1) := C_0$, we obtain that under Condition (C7) and for sufficiently small Δt,

$$\mathrm{P}(|\xi_{n_t,i}| > x) \le \mathrm{P}(|\epsilon_i| > C_0^{-1}x - 1) \le C_0^{-2-\varsigma}(x-C_0)^{-2-\varsigma} \le C_1 x^{-(2+\varsigma)}\,,$$

for some large enough constant C_1.

We have just shown that the $\xi_{n_t,i}$'s fulfil the requirements of Theorem 3.B.2 with $p = 2+\varsigma$ when Δt is small enough. Therefore, letting $\bar{\sigma}_n^2 = \sum_{i,j=1}^n |\,\mathrm{cov}(\xi_{n,i},\xi_{n,j})|$, we deduce from Theorem 3.B.2 that for any (fixed) $\lambda > 0$, there exists a constant $\Upsilon = \Upsilon(\lambda,a,p) > 0$ that depends on λ, p and on the mixing rate a defined in Condition (C8) such that for all $\eta > 0$,

$$\mathrm{P}(S_{n_t} \ge \eta) \le \Upsilon\left\{\left(1+\frac{\eta^2}{\lambda\bar{\sigma}_n^2}\right)^{-\lambda/2} + \frac{n_t}{\lambda}\left(\frac{\lambda}{\eta}\right)^{(a+1)(2+\varsigma)/(a+2+\varsigma)}\right\}. \tag{3.C.57}$$

Since under our conditions (3.C.26), (3.C.28) and (3.C.32) in Lemma 3.C.8 ensures that as $\Delta t \to 0$ we have $\sum_{i=1}^n\sum_{j=1}^n |\mathrm{Cov}\{K_h(x-X_i)Y_i, K_h(x-X_j)Y_j\}|\tilde{w}_{n_t,i}\tilde{w}_{n_t,j} = \gamma h^{-1}R_K\psi_2(x,t) + o(\gamma h^{-1})$, where ψ_2 is defined at (3.C.18), we deduce from the definition of $\xi_{n_t,i}$ above (3.C.56) and Conditions (C4)–(C5) that there exist a constant $\vartheta > 0$ such that as $\Delta t \to 0$,

$$\bar{\sigma}_{n_t} = h^2\gamma^{-2}\sum_{i=1}^n\sum_{j=1}^n |\mathrm{Cov}\{K_h(x-X_i)Y_i, K_h(x-X_j)Y_j\}|\tilde{w}_{n_t,i}\tilde{w}_{n_t,j} \le \vartheta h\gamma^{-1}\,.$$

We are now ready to show (3.C.55). For this, fix $\eta > 0$ and $\lambda \ge 2\{(a+1)(2+\varsigma)(a_1-a_2)/(a+2+\varsigma)-1\}/(a_1-a_2)$. Using (3.C.57) and the above equation, we have

$$\mathrm{P}\big\{\big|\tilde{r}(x,t|g,h) - \mathrm{E}\{\tilde{r}(x,t|g,h)\}\big| > \eta\big\} = \mathrm{P}\big(|S_{n_t}| > \eta h/\gamma\big)$$

$$\leq \Upsilon\left\{\left(1+\frac{\eta^2}{\lambda\vartheta}\frac{h}{\gamma}\right)^{-\lambda/2}+\frac{n_t}{\lambda}\left(\frac{\lambda}{\eta_0}\frac{\gamma}{h}\right)^{(a+1)(2+\varsigma)/(a+2+\varsigma)}\right\}$$

From there, since Condition (C1) ensures that $\gamma/h \sim \Delta t^{a_1-a_2} \to 0$ as $\Delta t \to 0$, and since from (1.2) we have $n_t \leq t/\Delta t$, we deduce from the above equation that there exists a constant $\Upsilon' > 0$, independent of Δt (and of (x,t,g,h)), but that may depend on λ, a and η, such that for sufficiently small Δt,

$$\begin{aligned}\mathrm{P}\big\{\big|\tilde{r}(x,t|g,h)-\mathrm{E}\{\tilde{r}(x,t|g,h)\}\big|>\eta\big\} &\leq \Upsilon'(\Delta t^{\lambda(a_1-a_2)/2}+\Delta t^{(a+1)(2+\varsigma)(a_1-a_2)/(a+2+\varsigma)-1}) \\ &\leq 2\Upsilon'\Delta t^{(a+1)(2+\varsigma)(a_1-a_2)/(a+2+\varsigma)-1}\,, \qquad (3.C.58)\end{aligned}$$

where the last line followed from the choice of λ.

Therefore, as $\Delta t \to 0$, we have

$$\begin{aligned}\mathrm{P}(J_{31}>\eta) &\leq \sum_{x\in\mathcal{K}_0}\sum_{(\gamma,h)\in I^{\ell_t}_{\gamma,h}}\mathrm{P}\big\{\big|\tilde{r}(x,t|g,h)-\mathrm{E}\{\tilde{r}(x,t|g,h)\}\big|>\eta\big\} \\ &\leq 2\Upsilon'(P_{\Delta t}+1)\#(I^{\ell_t}_{\gamma,h})\Delta t^{(a+1)(2+\varsigma)(a_1-a_2)/(a+2+\varsigma)-1}=o(1)\,,\end{aligned}$$

where we used Conditions (C3), (C7) and (C8) and the definition of $P_{\Delta t}$ above (3.C.42) to get the last equality. Hence we have proved (3.C.55) and $J_3 = o_p(1)$ follows.

We now turn our attention to J_4. By Conditions (C4)–(C5) and the definition of $P_{\Delta t}$ above (3.C.42), we obtain using the mean-value theorem that

$$J_{41} := \sup_{|x-x'|\leq P^{-1}_{\Delta t}}\big|(m\cdot f)(x,t)-(m\cdot f)(x',t)\big| \leq \frac{\|(m\cdot f)'_x\|_\infty}{P_{\Delta t}} \leq \frac{\|(m\cdot f)'_x\|_\infty \Delta t^{2a_2}}{\log(\Delta t^{-2})}, \qquad (3.C.59)$$

where $\|(m\cdot f)'_x\|_\infty \leq \|m'\|_\infty\|f\|_\infty + \|m\|_\infty\|f'\|_\infty < \infty$ (see Conditions (C4)–(C5)).

Next, from (3.A.1), and using the inequality $|K_h(X_i-x)-K_h(X_i-x') \leq h^{-2}\|K\|_\infty|x-$

$x'|$, we obtain that

$$\begin{aligned} J_{42} &:= \sup_{|x-x'|\le P_{\Delta t}^{-1}} \sup_{(\gamma,h)\in I_{\gamma,h}^{\ell_t}} \big|\tilde{r}(x,t|\gamma,h) - \tilde{r}(x',t|\gamma,h)\big| \\ &\le \sup_{(\gamma,h)\in I_{\gamma,h}^{\ell_t}} \left(\sum_{i=1}^{n_t} \tilde{w}_{n_t,i}|\epsilon_i|\right) \sup_{|x-x'|\le P_{\Delta t}^{-1}} |K_h(X_i-x) - K_h(X_i-x')| \\ &\le \frac{\|K'\|_\infty}{h^2 P_{\Delta t}} \sup_{(\gamma,h)\in I_{\gamma,h}^{\ell_t}} \left(\sum_{i=1}^{n_t} \tilde{w}_{n_t,i}|\epsilon_i|\right) \le \frac{\|K'\|_\infty}{\delta^2 \log(\Delta t^{-2})} \sup_{(\gamma,h)\in I_{\gamma,h}^{\ell_t}} \left(\sum_{i=1}^{n_t} \tilde{w}_{n_t,i}|\epsilon_i|\right), \quad (3.C.60) \end{aligned}$$

where we used Conditions (C1) to get the last line.

Now, if we can show that

$$\sup_{(\gamma,h)\in I_{\gamma,h}^{\ell_t}} \left(\sum_{i=1}^{n_t} \tilde{w}_{n_t,i}|\epsilon_i|\right) = o_p\big\{\log(\Delta t^{-2})\big\}, \quad (3.C.61)$$

then $J_4 = o_p(1)$ will follow from (3.C.60). To show (3.C.61), first let $n_0 = \lceil \log(\Delta t^{-1})/\gamma_m \rceil$, where γ_m is defined in Condition (C1). Then we have $\sum_{i=1}^{n_t} \tilde{w}_{n_t,i}|\epsilon_i| = T_1 + T_2$, where

$$T_1 = \sum_{i=1}^{n_t-n_0} \tilde{w}_{n_t,i}|\epsilon_i|, \quad T_2 = \sum_{i=n_t-n_0+1}^{n_t} \tilde{w}_{n_t,i}|\epsilon_i|, \quad (3.C.62)$$

noting from Condition (C1) and the definition of n_0 below (3.C.61) that we have $n_t > n_0$ when Δt is small enough. Note that T_1 and T_2 do not depend on h but only depend on γ though $\tilde{w}_{n_t,i}$ (see (3.A.4)). We next bound T_1 and T_2.

Starting with T_1, using (3.A.4) and the fact as under Condition (C1) $\gamma \to 0$ as $\Delta t \to 0$, we have $1-\gamma \ge 1/2$ for sufficiently small Δt, it follows that $\tilde{w}_{n_t,i} \le 2\gamma(1-\gamma)^{n_0} \le \gamma e^{-\gamma n_0} \le \delta^{-1}\Delta t^{a_1+\gamma/\gamma_m}$ for all $i \in \{2,\dots,n_t-n_0\}$, where we also used the inequality $1-x \le e^{-x}$ (see e.g. below (3.C.6)), Condition (C1) and the definition of n_0. In view of Lemma 3.C.3, we have also have $\tilde{w}_{n_t,i} \le \delta^{-1}\Delta t^{a_1+\gamma/\gamma_m}$ for sufficiently small Δt. Since $\gamma > \gamma_m$, we deduce that $\tilde{w}_{n_t,i} \le \delta^{-1}\Delta t^{a_1+1}$ for all $i \in \{1,\dots,n_0\}$ when Δt is small enough, and therefore, unformly

in $(\gamma, h) \in I^{\ell_t}_{\gamma,h}$, we have

$$T_1 \leq \delta^{-1}\Delta t^{a_1+1}(n_t - n_0)\Big\{(n_t - n_0)^{-1}\sum_{i=1}^{n_t-n_0}|\epsilon_i|\Big\} \leq t\Delta t^{a_1}\Big\{(n_t - n_0)^{-1}\sum_{i=1}^{n_t-n_0}|\epsilon_i|\Big\}, \tag{3.C.63}$$

where we used (1.2) to get the last equality.

Since $\mathrm{E}(|\epsilon_i|^2) < \infty$ (see Condition (C7)), the weak law of large numbers (Chung, 2000, p. 114) ensures that as $\Delta t \to 0$,

$$(n_t - n_0)^{-1}\sum_{i=1}^{n_t-n_0}|\epsilon_i| = O_p(1)\,. \tag{3.C.64}$$

Combining (3.C.63) and (3.C.64) and noting that (3.C.64) does not depend on γ nor h, we conclude that $\sup_{(\gamma,h)\in I^{\ell_t}_{\gamma,h}} T_1 = O_p(\Delta t)$.

Next we bound T_2. From (3.A.4) and the fact that $0 < \gamma < 1$ we have $\tilde{w}_{n_t,i} \leq \gamma$ for all $i > 1$. Hence, we compute that as $\Delta t \to 0$ we have

$$T_2 \leq \gamma \sum_{i=n_t-n_0+1}^{n_t}|\epsilon_i| \leq 2\gamma\log(\Delta t^{-1})/\gamma_m \times \Big(n_0^{-1}\sum_{i=n_t-n_0+1}^{n_t}|\epsilon_i|\Big)\,,$$

From there, as the week law of large numbers implies that $n_0^{-1}\sum_{i=n_t-n_0+1}^{n_t}|\epsilon_i| = O_P(1)$, we deduce from the fact that $\gamma/\gamma_m \leq 1$ (see Condition (C1)) that $\sup_{(\gamma,h)\in I^{\ell_t}_{\gamma,h}} T_2 = O_p(\log(\Delta t^{-1}))$. This shows that (3.C.61) holds as $\Delta t \to 0$, and hence proves that $J_4 = o_p(1)$.

Since we have also shown $J_3 = o_p(1)$, Lemma 3.C.10 follows. □

3.D Proof of Proposition 3.1

Throughout this section, for any $A \subset R^p$, let $\|\cdot\|_A = \sup_{(a_1,\ldots,a_p)\in A}|\cdot|$.

To save space, let $A_{\ell_t} = \mathcal{K} \times I^{\ell_t}_{\gamma,h}$. Recall that when $\ell_t = 1$ with ℓ_t defined at (3.21), we have $\check{m}(x,t|\gamma,h) = \tilde{m}(x,t|\gamma,h)$ for all $(\gamma,h) \in I^{\ell_t}_{\gamma,h}$, which implies that (3.30) holds trivially when $\ell_t = 1$. Hence, to prove Proposition 3.1 it suffices to show (3.30) in the case $\ell_t \geq 2$.

For this, observe that as $\check{m} - \tilde{m} = (\tilde{f}\check{r} - \check{f}\tilde{r})/(\check{f}\tilde{f})$, and since $\tilde{f} \geq f - \|\tilde{f} - f\|_{A_{\ell_t}}$, combining Lemma 3.C.9 and the fact that $\inf_{x\in\mathcal{K}} f(x,t) > \upsilon$ for some $\upsilon > 0$ guarantee that for any t satisfying $\ell_t \geq 2$, and as $\Delta t \to 0$, we have

$$\begin{aligned}
&\|\check{m} - \tilde{m}\|_{A_{\ell_t}} \leq \|\tilde{f}\check{r} - \check{f}\tilde{r}\|_{A_{\ell_t}} \|\check{f}^{-1}\|_{A_{\ell_t}} \big\{\upsilon - o_p(1)\big\}^{-1} \\
&\leq \big(\|\tilde{f} - \check{f}\|_{A_{\ell_t}} \|\tilde{r}\|_{A_{\ell_t}} + \|\check{r} - \tilde{r}\|_{A_{\ell_t}} \|\tilde{f}\|_{A_{\ell_t}}\big) \|\check{f}^{-1}\|_{A_{\ell_t}} \big\{\upsilon - o_p(1)\big\}^{-1} \\
&\leq \big(\|\tilde{f} - \check{f}\|_{A_{\ell_t}} \|\tilde{r}\|_{A_{\ell_t}} + \|\check{r} - \tilde{r}\|_{A_{\ell_t}} \|\tilde{f}\|_{A_{\ell_t}}\big) \big\{\upsilon - o_p(1) - \|\check{f} - \tilde{f}\|_{A_{\ell_t}}\big\}^{-1} \big\{\upsilon - o_p(1)\big\}^{-1}.
\end{aligned}$$

Further, as $\|\tilde{r}\|_{A_{\ell_t}} \leq \|m \cdot f\|_\infty + \|\tilde{r} - m \cdot f\|_{A_{\ell_t}}$ and $\|\tilde{f}\|_{A_{\ell_t}} \leq \|f\|_\infty + \|\tilde{f} - f\|_{A_{\ell_t}}$, and since the continuous mapping theorem entails $\{\upsilon - o_p(1)\}^{-1} = O_p(1)$, we deduce from Conditions (C4)–(C5), Lemma 3.C.9 and Lemma 3.C.10 that as $\Delta t \to 0$,

$$\|\check{m} - \tilde{m}\|_{A_{\ell_t}} \leq \big(\|\tilde{f} - \check{f}\|_{A_{\ell_t}} + \|\check{r} - \tilde{r}\|_{A_{\ell_t}}\big)\big\{\upsilon - o_p(1) - \|\check{f} - \tilde{f}\|_{A_{\ell_t}}\big\}^{-1} \times O_p(1)\,. \quad \text{(3.D.1)}$$

In view of the above inequality, to prove to prove (3.30) in the case $\ell_t \geq 2$ it suffices to show there exists $\zeta > 0$ such that as $\Delta t \to 0$

$$\sup_{(x,\gamma,h)\in A_{\ell_t}} |\check{r}(x,t|\gamma,h) - \tilde{r}(x,t|\gamma,h)| = o_p\big\{\exp(-\zeta\Delta t^{-1})\big\}\,, \quad \text{(3.D.2)}$$

$$\sup_{(x,\gamma,h)\in A_{\ell_t}} |\check{f}(x,t|\gamma,h) - \tilde{f}(x,t|\gamma,h)| = o_p\big\{\exp(-\zeta\Delta t^{-1})\big\}\,. \quad \text{(3.D.3)}$$

We start by proving (3.D.2). Recall from (3.A.1) and (3.A.2) that

$$\tilde{r}(x,t|\gamma,h) = \sum_{i=1}^{n_t} \tilde{w}_{n_t,i} K_h(x - X_i) Y_i, \quad \text{and} \quad \check{r}(x,t|\gamma,h) = \sum_{i=1}^{n_t} \check{w}_{n_t,i} K_{\check{h}_i}(x - X_i) Y_i\,,$$

where, from (3.A.4) and (3.A.5), we have $\tilde{w}_{n_t,i} = \check{w}_{n_t,i}$ and $\check{h}_i = h$ for all $i \in (N_{\ell_t - 1} - 2\nu, n_t]$ (see (3.24)). Therefore, we compute that

$$|\check{r}(x,t|\gamma,h) - \tilde{r}(x,t|\gamma,h)| = \left| \sum_{i=1}^{N_{\ell_t-1}-2\nu} (\tilde{w}_{n_t,i} K_h(x - X_i) - \check{w}_{n_t,i} K_{\check{h}_i}(x - X_i)) Y_i \right|$$

$$\leq \|K\|_\infty \Big(\max_{1\leq i\leq n_t} |Y_i| \Big) \times \sum_{i=1}^{N_{\ell_t-1}-2\nu} (h^{-1}\tilde{w}_{n_t,i} + \check{h}_i^{-1}\check{w}_{n_t,i}) \,. \quad \text{(3.D.4)}$$

where we also used the inequality $N_{\ell_t-1} - 2\nu \leq n_t$.

As $|Y_i| \leq \|m\|_\infty + |\epsilon_i|$, where under Condition (C4) $\|m\|_\infty < \infty$, and since Condition (C7) ensures that for any $\kappa > 0$,

$$\begin{aligned} P\Big\{ \max_{1\leq j\leq n_t} \frac{|\epsilon_j|}{(n_t \log n_t)^{1/(2+\varepsilon)}} > \kappa \Big\} &\leq n_t P\big\{ |\epsilon_j| > (n_t \log n_t)^{1/(2+\varepsilon)} \kappa \big\} \\ &\leq \frac{1}{(\log n_t)\kappa^{2+\varsigma}} \to 0 \quad \text{as } \Delta t \to 0 \,, \end{aligned}$$

it follows that $\max_{1\leq j\leq n_t} |Y_i| = o_p(n_t \log n_t) = o_p\{\Delta t^{-1} \log(\Delta t^{-1})\}$. Plugging this result into (3.D.4) entails that as $\Delta t \to 0$ we have

$$|\check{r}(x,t|\gamma,h) - \tilde{r}(x,t|\gamma,h)| = \sum_{i=1}^{N_{\ell_t-1}-2\nu} (h^{-1}\tilde{w}_{n_t,i} + \check{h}_i^{-1}\check{w}_{n_t,i}) \times o_p\{\Delta t^{-1} \log(\Delta t^{-1})\} \,. \quad \text{(3.D.5)}$$

In view of the latter display, to prove (3.D.2), it remains to derive a bound for the first term of the product on the right hand side of the above equation.

To do this, observe that since under Conditions (C1) and from (3.29) we have $(\check{\gamma}_j, \check{h}_j) \in [\gamma_m, \gamma_M] \times [h_m, h_M]$, it follows that $\min(\check{h}_i, h) \geq h_m^{-1}$, that $\max(\tilde{w}_{n_t,1}, \check{w}_{n_t,1}) \leq (1-\gamma_m)^{n_t-1}$ and that $\max(\tilde{w}_{n_t,j}, \check{w}_{n_t,j}) \leq \gamma_M(1-\gamma_m)^{n_t-j} \leq (1-\gamma_m)^{n_t-N_{\ell_t-1}+2\nu}$ for all $j \in \{2, \ldots, i_{\ell_t-1}\}$ (see (3.A.4) and (3.A.5)). Therefore, uniformly in $(x, \gamma, h) \in A_{\ell_t}$, we have

$$\sum_{i=1}^{N_{\ell_t-1}-2\nu} (h^{-1}\tilde{w}_{n_t,i} + \check{h}_i^{-1}\check{w}_{n_t,i}) \leq 2h_m^{-1}\big\{ (1-\gamma_m)^{n_t-1} + (N_{\ell_t-1} - 2\nu)(1-\gamma_m)^{n_t-N_{\ell_t-1}+2\nu} \big\} \,.$$

By the definition of ℓ_t at (3.21), we have $N_{\ell_t-1} \leq n_t$, and hence $n_t - N_{\ell_t-1} + 2\nu \geq 2\nu > \nu$. We therefore deduce from Lemma (3.C.3) and the above equation that there exists a constant

$\widetilde{\zeta} > 0$ such that as $\Delta t \to 0$,

$$\sum_{i=1}^{N_{\ell_{t-1}}-2\nu} (h^{-1}\tilde{w}_{n_t,i} + \check{h}_i^{-1}\check{w}_{n_t,i}) \leq 2n_t h_m^{-1} \exp(-\widetilde{\zeta}\Delta t^{-1}) = o\{\exp(-\widetilde{\zeta}\Delta t^{-1}/2)\}\,. \quad (3.D.6)$$

To obtain the last equality, we used Condition (C1) and the fact that $n_t \leq t\Delta t^{-1}$.

Plugging (3.D.6) into (3.D.5) proves that (3.D.2) holds with $\zeta = \widetilde{\zeta}/4$. Therefore, to conclude the proof of Proposition 3.1, it only remains to demonstrate (3.D.3).

For this, we proceed as before to deduce that

$$\begin{aligned} |\check{f}(x,t|\gamma,h) - \tilde{f}(x,t|\gamma,h)| &= \left| \sum_{i=1}^{N_{\ell_{t-1}}-2\nu} \{\tilde{w}_{n_t,i} K_h(x - X_i) - \check{w}_{n_t,i} K_{\check{h}_i}(x - X_i)\} \right| \\ &\leq \|K\|_\infty \times \sum_{i=1}^{N_{\ell_{t-1}}-2\nu} (h^{-1}\tilde{w}_{n_t,i} + \check{h}_i^{-1}\check{w}_{n_t,i}) = o\{\exp(-\widetilde{\zeta}\Delta t^{-1}/4)\}\,, \quad (3.D.7) \end{aligned}$$

where we used the fact that Condition (C6) ensures $\|K\|_\infty < \infty$ and (3.D.6). This shows that (3.D.3) holds with $\zeta = \widetilde{\zeta}/4$ as above and concludes the proof of the announced result.

3.E Proof of Theorem 3.1

Since $(\gamma, h) \in I_{\gamma,h}^{\ell_t} \subseteq \square^k \subset \square$, we have under Condition (C1) that $h/\gamma \leq h_M/\gamma_m \leq 2\delta^{-1}\Delta t^{a_2-a_1}$. Hence, from the definition of V at (3.34), and since (C5) and $(C7)$ respectively imply $\|f\|_\infty < \infty$ and $\sigma^2 > 0$, it follows that $V^{-1/2} = O(\Delta t^{(a_2-a_1)/2})$. Therefore, we obtain from Proposition 3.1 that as $\Delta t \to 0$,

$$\frac{1}{\sqrt{V}}\big\{\check{m}(x,t|\gamma,h) - m(x,t) - B\big\} = \frac{1}{\sqrt{V}}\big\{\tilde{m}(x,t|\gamma,h) - m(x,t) - B\big\} + o_p(1)\,. \quad (3.E.1)$$

Hence, to prove (3.33) it suffices to show that the first term on the right hand side of the above equation converges in distribution to a normal random variable as $\Delta t \to 0$. To do so,

recall that $\tilde{m} = \tilde{r}/\tilde{f}$, and observe that we can express $\tilde{m} - m$ above as

$$\tilde{m}(x,t|\gamma,h) - m(x,t) = \frac{1}{\tilde{f}(x,t \mid \gamma,h)} \sum_{i=1}^{n_t} \tilde{w}_{n_t,i} K_h(x - X_i)\{Y_i - m(x,t)\}\,.$$

Introducing the random variables $\zeta_{n_t,i} = \tilde{w}_{n_t,i} K_h(x - X_i)\{Y_i - m(x,t)\}$ and $\mathcal{J}_t = \sum_{i=1}^{n_t} \zeta_{n_t,i}$ and $\mathbb{J}_t = V^{-1/2}\{\mathcal{J}_t - \mathrm{E}(\mathcal{J}_t)\}$, we can express the right hand side of (3.E.1) as

$$\frac{\tilde{m}(x,t|\gamma,h) - m(x,t) - B}{\sqrt{V}} = \frac{\mathbb{J}_t + V^{-1/2}[\mathrm{E}(\mathcal{J}_t) - B\{f(x,t) + o_p(1)\}]}{f(x,t) + o_p(1)}\,, \tag{3.E.2}$$

where we used to fact that $f(x,t) > 0$ and that Lemma 3.C.9 implies $\tilde{f}(x,t \mid \gamma,h) = f(x,t) + o_p(1)$ and $\tilde{f}^{-1}(x,t \mid \gamma,h) = f^{-1}(x,t) + o_p(1)$.

In view of (3.E.2), to prove Theorem 3.1, it suffices to show that (i) $\mathbb{J}_t$ converges in distribution to a normal random variable, (ii) that $V^{-1/2}\{\mathrm{E}(\mathcal{J}_t) - Bf(x,t)\} = o(1)$ and (iii) that $V^{-1/2}B = O(1)$. Below we show (i), (ii) and (iii).

Starting with (ii), observe that as $\mathcal{J}_t = \tilde{r}(x,t \mid \gamma,h) - m(x,t)\tilde{f}(x,t \mid \gamma,h)$, it follows from Lemma 3.C.7 and Lemma 3.C.8 that $\mathrm{E}(\mathcal{J}_t) - Bf(x,t) = o(h^2 + \Delta t/\gamma)$. As $V^{-1/2} = O(\Delta t^{(a_2-a_1)/2})$ (see above (3.E.1)), and since under Condition (C1) we have $h^2 + \Delta t/\gamma = O(\Delta t^{2a_2} + \Delta t^{1-a_1})$, it follows from the assumption that $1/7 < a_2 < a_1 < 5/7$ that $V^{-1/2}(h^2 + \Delta t/\gamma) = O(1)$. This concludes the proof of (ii).

Next we show (iii). For this, observe that under Condition (C5) and (C4), and since we have assumed that $f(x,t) > 0$, we have $|B| = O(h^2 + \Delta t/\gamma) = O(\Delta t^{2a_2} + \Delta t^{1-a_1})$. Hence, (iii) follows from the fact that $V^{-1/2}(h^2 + \Delta t/\gamma) = O(1)$ (see the previous paragraph).

As we have just shown that (ii) and (iii) hold, to prove the announced result in Theorem 3.1, it only remains to show that (i) is in force, i.e. that $\mathbb{J}_t$ is asymptotically normally distributed.

To prove that it is the case, our strategy is to apply Theorem 3.B.1 to $\mathbb{J}'_t := V^{1/2}\mathbb{J}_t$. This will allow us to prove that $\mathbb{J}'_t/\sqrt{\mathrm{var}(\mathbb{J}'_t)}$ is asymptotically distributed as a standard normal random variable. From there, we will show that $Vf^2(x,t) = \mathrm{var}(\mathbb{J}'_t) + o(1)$, which will conclude the proof of the announced result.

Let $\xi_{n_t,i} = \zeta_{n_t,i} - \mathrm{E}(\zeta_{n_t,i})$, so that $\mathbb{J}'_t = \sum_{i=1}^{n_t} \xi_{n_t,i}$. In this notation, to apply Theo-

rem 3.B.1, we need to show that (C1) the α-mixing coefficients of the sequence $\{\xi_{n_t,i}\}$ are such that $\sup_n \tilde{\alpha}(n,k) \to 0$ as $k \to \infty$, where $\tilde{\alpha}$ is as at (3.B.7); (C2) the $\rho-$mixing coefficients of the sequence $\{\xi_{n_t,i}\}$ are such that $\sup_n \tilde{\rho}(n,1) < 1$, where $\tilde{\rho}$ is as at (3.B.8); and (C3) that there exist a constant $c > 0$ such that

$$\lim_{\Delta t \to 0} \frac{1}{\{\text{var}(\mathbb{J}_t')\}^{1+c/2}} \sum_{i=1}^{n} \text{E}(|\xi_{n_t,i}|^{2+c}) = 0\,. \tag{3.E.3}$$

Since in view of Condition (C8) and (C9), (C1) and (C2) are trivially satisfied, to prove that $\mathbb{J}_t$ is asymptotically normally distributed, we need to show that the $\xi_{n_t,i}$'s satisfy (3.E.3). This is what we do below.

Fix $c \in (0,\varsigma)$, where ς is defined in Condition (C7). As $|\zeta_{n_t,i}| = \tilde{w}_{n_t,i} K_h(X_i - x)|m(X_i,t_i) - m(x,t) + \epsilon_i| \leq \tilde{w}_{n_t,i} K_h(X_i - x)(2\|m\|_\infty + |\epsilon_i|)$, where we used (1.3), it follows that

$$\begin{aligned}
\text{E}(|\zeta_{n_t,i}|^{2+c}) &\leq \tilde{w}_{n_t,i}^{2+c}\, \text{E}\{K_h^{2+c}(X_i - x)\}\, \text{E}\{(2\|m\|_\infty + |\epsilon_i|)^{2+c}\} \\
&\leq \tilde{w}_{n_t,i}^{2+c} h^{-1-c} \|K\|_\infty^{1+c}\, \text{E}\{K_h(X_i - x)\}\, \text{E}\{(2\|m\|_\infty + |\epsilon_i|)^{2+c}\} \\
&= \tilde{w}_{n_t,i}^{2+c} h^{-1-c} \times O(1)\,,
\end{aligned} \tag{3.E.4}$$

where, to obtain the last equality, we used the fact that from Conditions (C6) and (C4) we have $\max(\|K\|_\infty, \|m\|_\infty) < \infty$ and Condition (C7) which ensures that $\text{E}(|\epsilon_i|^{2+c}) < \infty$ since $c < \varsigma$.

Using similar derivations, we also compute that

$$|\,\text{E}(\zeta_{n_t,i})| \leq \text{E}(|\zeta_{n_t,i}|) \leq \tilde{w}_{n_t,i}\, \text{E}\{K_h(X_i - x)\}\, \text{E}(2\|m\|_\infty + |\epsilon_i|) = \tilde{w}_{n_t,i} \times O(1)\,. \tag{3.E.5}$$

From the latter inequality, the result at (3.E.4) and the inequality $|a+b|^{2+c} \leq 2^{1+c}(|a|^{2+c} + |b|^{1+c})$, we deduce that as $\Delta t \to 0$,

$$\text{E}(|\xi_{n_t,i}|^{2+c}) \leq 2^{1+c}\{\text{E}(|\zeta_{n_t,i}|^{2+c}) + |\,\text{E}(\zeta_{n_t,i})|^{2+c}\} = \tilde{w}_{n_t,i}^{2+c}(1 + h^{-1-c}) \times O(1)\,.$$

Therefore, as the $\tilde{w}_{n_t,i}$'s are bounded by 1, and since Lemma 3.C.5 guarantees that as $\Delta t \to 0$ we have $\sum_{i=1}^{n_t} \tilde{w}_{n_t,i}^2 = O(\gamma)$, we obtain that as $\Delta t \to 0$,

$$\sum_{i=1}^{n_t} \mathrm{E}(|\xi_{n_t,i}|^{2+c}) = O(h^{-1-c}\gamma)\,.$$

In view of the above equation, (3.E.3) will be proved if we show that as $\Delta t \to 0$,

$$\mathrm{var}(\mathbb{J}_t') = \frac{R_K \gamma}{2h}\{\sigma^2 f(x,t) + o(1)\}\,. \tag{3.E.6}$$

To demonstrate (3.E.6), we decompose $\mathrm{var}(\mathbb{J}_t') = \mathcal{V}_1 + \mathcal{V}_2$, where

$$\mathcal{V}_1 = \sum_{i=1}^{n_t} \mathrm{var}(\zeta_{n_t,i}) \qquad \mathcal{V}_2 = \sum_{1\le|i-j|\le n_t-1} \tilde{w}_{n_t,i}\tilde{w}_{n_t,j} e_{ij}\,, \tag{3.E.7}$$

with $e_{ij} = \mathrm{cov}[K_h(x-X_i)\{Y_i - m(x,t)\}, K_h(x-X_j)\{Y_j - m(x,t)\}]$.

Mimicking the arguments presented below (3.C.23) in the proof of Lemma 3.C.7 allows to conclude that there exists a constant $\kappa > 0$ such that $|e_{ij}| \le \kappa \min\{\alpha^{1/3}(\Delta t, k), h^{-2}\alpha(\Delta t, |i-j|)\}$, so that

$$\begin{aligned}
|\mathcal{V}_2| &\le \kappa \sum_{1\le|i-j|\le n_t-1} \tilde{w}_{n_t,i}\tilde{w}_{n_t,j} \min\big\{\alpha^{1/3}(\Delta t, |i-j|), h^{-2}\alpha(\Delta t, |i-j|)\big\} \\
&\le 2\kappa \sum_{k=1}^{n_t-1} \min\left(k^{-a/3}, h^{-2}k^{-a}\right) \sum_{j=k+1}^{n_t} \tilde{w}_{n_t,j-k}\tilde{w}_{n_t,j}\,.
\end{aligned}$$

From the above display, and in view of (3.C.27) and (3.C.28), we conclude that as $\Delta t \to 0$ we have $|\mathcal{V}_2| = o(\gamma/h)$. Thus, to prove (3.E.6), it remains to show that $\mathcal{V}_1$ is asymptotically equivalent to the term on the right hand side of (3.E.6).

To show that it is the case, we first compute that

$$\begin{aligned}
&\mathrm{E}[K_h^2(x-X_i)\{Y_i - m(x,t)\}^2\} \\
&= \mathrm{E}\{K_h^2(x-X_i)Y_i^2\} - 2m(x,t)\,\mathrm{E}\{K_h^2(x-X_i)Y_i\} + m^2(x,t)\,\mathrm{E}\{K_h^2(x-X_i)\}
\end{aligned}$$

$$= \frac{R_K}{h}\{2(m^2 \cdot f)(x,t_i) + \sigma^2 f(x,t_i)\} + O(1) - 2m(x,t)\,\mathrm{E}\{K_h^2(x-X_i)m(X_i,t_i)\}\,, \tag{3.E.8}$$

where we used (3.C.30) and (3.C.39), and with R_K as at (3.32).

Then, using the mean-value theorem, we obtain that for all $i \in \{1,\ldots,n_t\}$, there exists $\theta_{i1}, \theta_{i2} \in [0,1]$ such that

$$\begin{aligned}\mathrm{E}\big\{K_h^2(x-X_i)m(X_i,t_i)\big\} &= \frac{1}{h}\int K^2(u)\big\{m(x,t_i) - hum_x(x-\theta_{i1}hu,t_i)\big\} \\ &\qquad \times \big\{f(x,t_i) - huf_x(x-\theta_{i2}hu,t_i)\big\}\,\mathrm{d}u \\ &= \frac{R_K}{h}(m\cdot f)(x,t_i) + O(1)\,,\end{aligned}$$

where, to obtain the last line, we used Conditions (C4)–(C6).

Plugging the latter result into (3.E.8) entails, as $\Delta t \to 0$, that

$$\mathrm{E}[K_h^2(x-X_i)\{Y_i - m(x,t)\}^2\} = \frac{R_K}{h}\sigma^2 f(x,t_i) + O(1)\,. \tag{3.E.9}$$

In view of the latter result and the one at (3.E.5), we compute that as $\Delta t \to 0$,

$$\mathcal{V}_1 = \frac{R_K}{h}\sigma^2 \sum_{i=1}^{n_t} \tilde{w}_{n_t,i}^2 f(x,t_i) + O(\sum_{i=1}^{n_t} \tilde{w}_{n_t,i}^2) = \frac{R_K}{h}\sigma^2 \sum_{i=1}^{n_t} \tilde{w}_{n_t,i}^2 f(x,t_i) + O(\gamma)\,,$$

where we used (3.C.8).

To conclude the proof of that $\mathcal{V}_1$ is asymptotically equivalent to the term on the right hand side of (3.E.6), note that by Conditions (C4) and (C5), $(m\cdot f)$ as a function of t satisfies the conditions of Lemma 3.C.6, and therefore an application of that lemma yields, as $\Delta t \to 0$, that

$$\mathcal{V}_1 = \frac{R_K\gamma}{2h}\sigma^2 f(x,t) + O(\Delta t/h) + O(\gamma) = \frac{R_K\gamma}{2h}\sigma^2\{f(x,t) + o(1)\}\,,$$

where, to obtain last equality, we used Condition (C1). Therefore, (3.E.3) is in force, and the announced result follows from the fact that $Vf^2(x,t) = \mathcal{V}_1 + o(1)$.

3.F Pseudocode for SRA

In this section we present the pseudocode for the SRA explained in §3.2.1. Algorithm F1 is the framework of the SRA, consisting of several modules, such as INITIALISE for the initialisation (§3.2.1.1). These modules are defined by Algorithms F2–F7. The function `Boolean`(x) returns `True` (`False`) if the statement x is `True` (`False`), otherwise it returns `False`.

Algorithm F1 Streaming regression algorithm.

Inputs: ν, an integer determining how often we check for alternative ensembles;
    $S = \{(X_i, Y_i)\}_{i=1,2,\ldots}$, sequentially observed;
    $I^1_{\gamma,h}$, initial candidate set;
    $L > 0$, a constant used in (3.13);
    g, cardinality of I^1_γ and I^1_h;
procedure SRA($\nu, S, I^1_{\gamma,h}, L$)
    INITIALISE($\nu, B_0, I^1_{\gamma,h}$); ▷ Algorithm F2
    $k \leftarrow 1$ and $\lambda \leftarrow 0$;
    `initialise.alt` $\leftarrow$ `False`;
    repeat ▷ repeat the following, could be never-ending
        while `initialise.alt` $==$ `False` **do** ▷ while (3.11) not satisfied
            ROUTINE($\nu, k, I^k_{\gamma,h}, N_{k-1}, \lambda, B_{k,\lambda}, L$); ▷ Algorithm F3
            $\lambda \leftarrow \lambda + 1$;
        end while
        INITIALISEALT($(\gamma^{CV}_{n_t}, h^{CV}_{n_t}), L, g$); ▷ Algorithm F4
        for $i = 1, 2$ **do**
            if $i == 2$ **then**
                UPDATEALTCV($B_{k,\lambda}, \mathscr{E}^{\text{alt}}_t, I^{\text{alt}}_{\gamma,h}$); ▷ Algorithm F6
            end if
            UPDATEALT($B_{k,\lambda}, \mathscr{E}^{\text{alt}}_t, I^{\text{alt}}_{\gamma,h}$); ▷ Algorithm F5
            ROUTINE($\nu, k, I^k_{\gamma,h}, N_{k-1}, \lambda, B_{k,\lambda}, L$);

22: $\lambda \leftarrow \lambda + 1$;
23: **end for**
24: Let `ES` $\leftarrow$ `Boolean`(if (3.19) holds);
25: **if** `ES` $==$ `True` **then**
26: EnsSubs($k, \lambda, \mathscr{E}_t^{\text{alt}}, I_{\gamma,h}^{\text{alt}}, \text{RCV}_t^{\text{alt}}$); ▷ Algorithm F7
27: **end if**
28: **until** data stream stops;
29: **end procedure**

The following algorithm defines the initialisation procedure described in §3.2.1.1.

Algorithm F2 Initilisation.

1: **Inputs**: ν, an integer determining how often we check for alternative ensembles;
2: $B_0 = \{(X_i, Y_i)\}_{i=1,\ldots,N_0}$, sequentially observed, where $N_0 = 2\nu$;
3: $I_{\gamma,h}^1$, initial candidate set;
4: **procedure** Initialise($\nu, B_0, I_{\gamma,h}^1$)
5: **for** $t = t_1, \ldots, t_{N_0}$ **do**
6: **if** $t \geq t_{\nu+1}$ **then**
7: Compute $\text{RCV}_t(\gamma, h)$ defined by (3.8);
8: **end if**
9: Compute $\check{m}(\cdot, t|\gamma, h) \in \mathscr{E}_t$ by (3.6);
10: **end for**
11: **return** $\mathscr{E}_{t_{N_0}}$, $\text{RCV}_{t_{N_0}}(\gamma, h)$ for $(\gamma, h) \in I_{\gamma,h}^1$;
12: **end procedure**

The following algorithms defines the routine updates (R1)–(R3), described in §3.2.1.2.

Algorithm F3 Routine updates.

1: **Inputs**: ν, an integer determining how often we check for alternative ensembles;
2: k, number of already-happened ensemble substitutions plus one;

3: $I^k_{\gamma,h}$, the current candidate set;

4: N_{k-1}, number of observations up to $(k-t)$-th ensemble substitution;

5: λ, number of checks since $t_{N_{k-1}}$;

6: $B_{k,\lambda} = \{(X_i, Y_i)\}_{i=N_{k-1}+\lambda\nu+1,\ldots,N_{k-1}+(\lambda+1)\nu}$, sequentially observed;

7: $L > 0$, a constant used in (3.13);

8: **procedure** ROUTINE($\nu, k, I^k_{\gamma,h}, N_{k-1}, \lambda, B_{k,\lambda}, L$)

9: **for** $t = t_{N_{k-1}+\lambda\nu+1}, \ldots, t_{N_{k-1}+(\lambda+1)\nu}$ **do**

10: (R1) Update RCV_t by (3.9);

11: (R2) Update $\check{m}(\cdot, t|\gamma, h) \in \mathscr{E}_t$ by (3.7);

12: (R3) Compute $(\gamma^{CV}_{n_t}, h^{CV}_{n_t})$ by (3.10);

13: **return** $\check{m}(\cdot, t|\gamma^{CV}_{n_t}, h^{CV}_{n_t})$, estimate of $m(\cdot, t)$;

14: **end for**

15: `initialise.alt` $\leftarrow$ `Boolean`(if (3.11) holds);

16: **return** `initialise.alt`;

17: **end procedure**

The next algorithm defines the initialisation of alternative ensemble as described in §3.2.1.3.

Algorithm F4 Initialisation of alternative ensemble.

1: **Inputs:** $\gamma^{CV}_{n_t}, h^{CV}_{n_t}$, γ and h values selected by RCV;

2: $L > 0$, a constant used in (3.13);

3: g, cardinality of I^1_γ and I^1_γ;

4: **procedure** INITIALISEALT($\gamma^{CV}_{n_t}, h^{CV}_{n_t}, L, g$)

5: Let $I^{\text{alt}}_{\gamma,h} = I^{\text{alt}}_\gamma \times I^{\text{alt}}_h$, where I^{alt}_γ and I^{alt}_h are equidistant grids of size g on $[\gamma^{CV}_{n_t}/L, L\gamma^{CV}_{n_t}]$ and $[h^{CV}_{n_t}/L, Lh^{CV}_{n_t}]$;

6: Let $\mathscr{E}^{\text{alt}}_t = \{\check{m}_{\text{alt}}(\cdot, t|\gamma, h) : (\gamma, h) \in I^{\text{alt}}_{\gamma,h}\}$, where $\check{m}_{\text{alt}}(\cdot, t|\gamma, h)$ is initialised by (3.14);

7: **return** $\mathscr{E}^{\text{alt}}_t$ and $I^{\text{alt}}_{\gamma,h}$;

8: **end procedure**

The next algorithm defines the updating of alternative ensemble as described in §3.2.1.3.

Algorithm F5 Updating alternative ensemble.

Inputs: $B_{k,\lambda} = \{(X_i, Y_i)\}_{i=N_{k-1}+\lambda\nu+1,\ldots,N_{k-1}+(\lambda+1)\nu}$, sequentially observed;
    $\mathscr{E}_t^{\text{alt}}$, initial alternative ensemble;
    $I_{\gamma,h}^{\text{alt}}$, alternative candidate set;
procedure UPDATEALT($B_{k,\lambda}, \mathscr{E}_t^{\text{alt}}, I_{\gamma,h}^{\text{alt}}$)
    for $t = t_{N_{k-1}+\lambda\nu+1}, \ldots, t_{N_{k-1}+(\lambda+1)\nu}$ **do**
        Update $\check{m}_{\text{alt}}(\cdot, t|\gamma, h) \in \mathscr{E}_t^{\text{alt}}$ by (3.16);
    end for
    return $\mathscr{E}_t^{\text{alt}}$;
end procedure

The next algorithm defines the updating of alternative cross-validation scores as described in §3.2.1.3.

Algorithm F6 Updating alternative cross-validation scores.

Inputs: $B_{k,\lambda} = \{(X_i, Y_i)\}_{i=N_{k-1}+\lambda\nu+1,\ldots,N_{k-1}+(\lambda+1)\nu}$, sequentially observed;
    $\mathscr{E}_t^{\text{alt}}$, updated alternative ensemble;
    $I_{\gamma,h}^{\text{alt}}$, alternative candidate set;
procedure UPDATEALTCV($B_{k,\lambda}, \mathscr{E}_t^{\text{alt}}, I_{\gamma,h}^{\text{alt}}$)
    for $t = t_{N_{k-1}+\lambda\nu+1}, \ldots, t_{N_{k-1}+(\lambda+1)\nu}$ **do**
        Update $\text{RCV}_t^{\text{alt}}(\gamma, h)$ for $(\gamma, h) \in I_{\gamma,h}^{\text{alt}}$ by (3.17);
    end for
    return $\text{RCV}_t^{\text{alt}}(\gamma, h)$ for $(\gamma, h) \in I_{\gamma,h}^{\text{alt}}$;
end procedure

The next algorithm defines the ensemble substitution as described in §3.2.1.3.

Algorithm F7 Ensemble substitution.

Inputs: k, number of already-happened ensemble substitutions plus one;

λ, number of checks since $t_{N_{k-1}}$;

$\mathscr{E}_t^{\text{alt}}$, alternative ensemble;

$I_{\gamma,h}^{\text{alt}}$, alternative candidate set;

$\text{RCV}_t^{\text{alt}}(\gamma, h)$, $(\gamma, h) \in I_{\gamma,h}^{\text{alt}}$, alternative cross-validation scores;

procedure EnsSubs($k, \lambda, \mathscr{E}_t^{\text{alt}}, I_{\gamma,h}^{\text{alt}}, \text{RCV}_t^{\text{alt}}$)

(ES1) $N_k = N_{k-1} + \lambda\nu$;

(ES2) $I_{\gamma,h}^{k+1} \leftarrow I_{\gamma,h}^{\text{alt}}$, $\check{m}(\cdot, t|\gamma, h) \leftarrow \check{m}_{\text{alt}}(\cdot, t|\gamma, h)$ for all $(\gamma, h) \in I_{\gamma,h}^{k+1}$, delete $\mathscr{E}_t^{\text{alt}}$ and $I_{\gamma,h}^{\text{alt}}$;

(ES3) $\text{RCV}_t(\gamma, h) \leftarrow \text{RCV}_t^{\text{alt}}(\gamma, h)$ for all $(\gamma, h) \in I_{\gamma,h}^{k+1}$, delete all $\text{RCV}_t^{\text{alt}}(\gamma, h)$;

(ES4) $k \leftarrow k + 1$, $\lambda \leftarrow 0$;

`initialise.alt` $\leftarrow$ `False`;

return k, λ, $I_{\gamma,h}^{k}$, N_{k-1}, $\mathscr{E}_t$, $\text{RCV}_t(\gamma, h)$ for $(\gamma, h) \in I_{\gamma,h}^{\text{alt}}$ and `initialise.alt`;

end procedure

Chapter 4

Numerical illustrations

In this chapter we present some simulation and real data examples to illustrate the performances of the streaming density and regression estimation methods described in Chapters 2 and 3.

4.1 Simulations

4.1.1 Density estimation

We ran simulations to illustrate the performance of the SKDE defined at (2.2), with smoothing parameters chosen by the SCV procedure described in §2.4. We simulated data streams $\{X_i\}_{i=1,\ldots,n_1}$ arriving on a finite time interval $[0,1]$ for 3 different sample sizes $n_1 = 5 \times 10^3$, 10^4 and 1.5×10^4, where n_1 is defined at (1.2), taking $t = 1$. These sample sizes are larger than the sample size used in the simulation studies of Hall et al. (2006). This is to reflect the fact that streaming data often have higher frequency than the conventional offline time series data, and hence it is easier to collect a large number of streaming data within a short period of time. Indeed, our sample sizes are comparable to those in the simulation studies of some works on density estimation for streaming data (e.g. Zhou et al., 2003; Cao et al., 2012; García-Treviño and Barria, 2012).

Suppose each X_i arrives at $t_i = i/n_1$ and has a density $f(x, t_i)$. Let $\phi_{\mu,\sigma}$ denote the density

function of a $N(\mu, \sigma^2)$. We consider the following four data models:

(i) $f(x,t) = \phi_{\mu_t,\sigma_t}(x)$, where $\mu_t = 100t$ and $\sigma_t = 1 + 10t$;

(ii) $f(x,t) = \phi_{\mu_t,\sigma_t}(x)$, where $\mu_t = 5\sin(10t\pi) + 5$ and $\sigma_t = 0.4\sin(10t\pi) + 2$;

(iii) $f(x,t) = 2^{-1}\sum_{k=1}^{2} \phi_{\mu_{t,k},\sigma_{t,k}}(x)$, where $\mu_{t,1} = 5t$, $\mu_{t,2} = 10 - 5t$ and $\sigma_{t,1} = \sigma_{t,2} = 1$;

(iv) $f(x,t) = 4^{-1}\sum_{k=1}^{4} \phi_{\mu_{t,k},\sigma_{t,k}}(x)$, where $\mu_{t,1} = 100t$, $\mu_{t,2} = 1 + 100t$, $\mu_{t,3} = 2 + 100t$, $\mu_{t,4} = 3 + 100t$ and $\sigma_{t,1} = \sigma_{t,2} = \sigma_{t,3} = \sigma_{t,4} = 0.2$.

In models (i) and (ii) the true densities are both Gaussian. In model (i) both the mean and variance are increasing; in model (ii) the mean and variance are both periodically varying. In models (iii) and (iv) the true densities are mixtures of 2 and 4 normal densities with fixed variance, respectively. In model (iii) the two normal densities approach each other and finally collide to form one mode; in model (iv) the location of the true density with 4 modes changes monotonically. See Figures 4.1–4.4 for an visual illustration of models (i)–(iv), respectively.

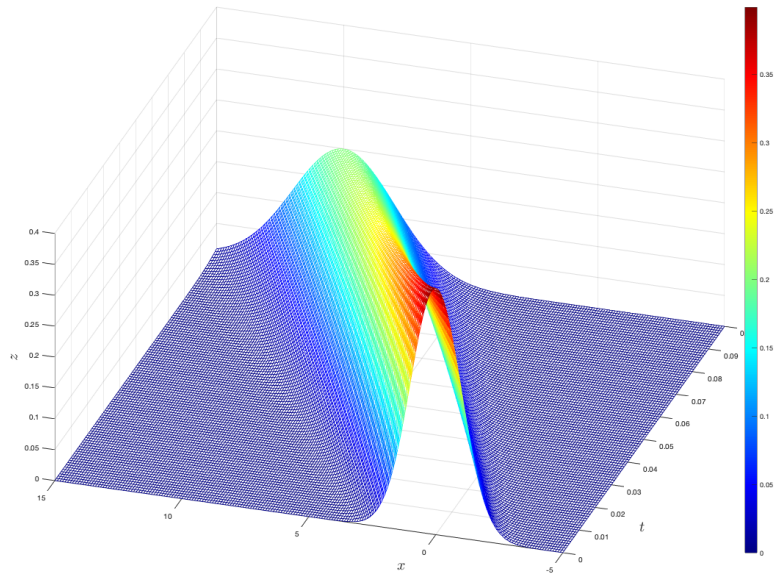

Figure 4.1: Illustration of density model (i), where x and t axes represent the spatial and temporal domains of $f(x,t)$ and $z = f(x,t)$ represents the density values. For better visual effects, only $f(x,t)$, $t \in (0, 0.1)$ are shown, with $f(x,0)$ at the front and $f(x,0.1)$ at the back.

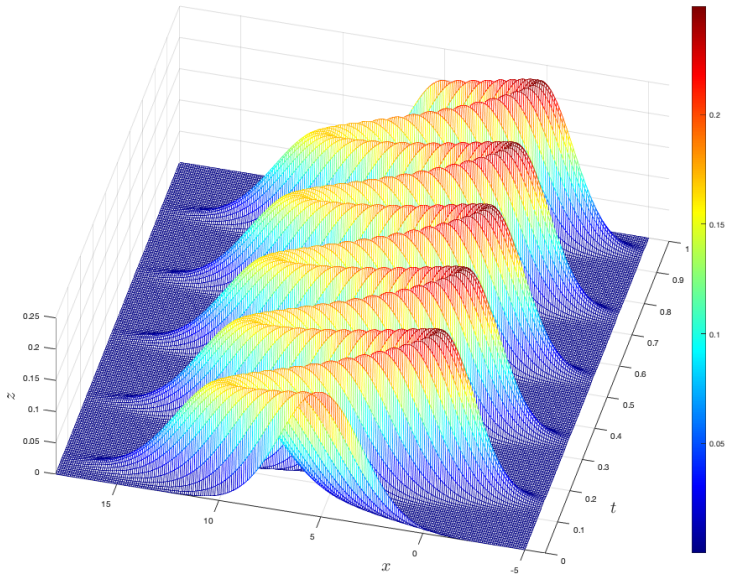

Figure 4.2: Illustration of density model (ii), where x and t axes represent the spatial and temporal domains of $f(x,t)$ and $z = f(x,t)$ represents the density values, with $f(x,0)$ at the front and $f(x,1)$ at the back.

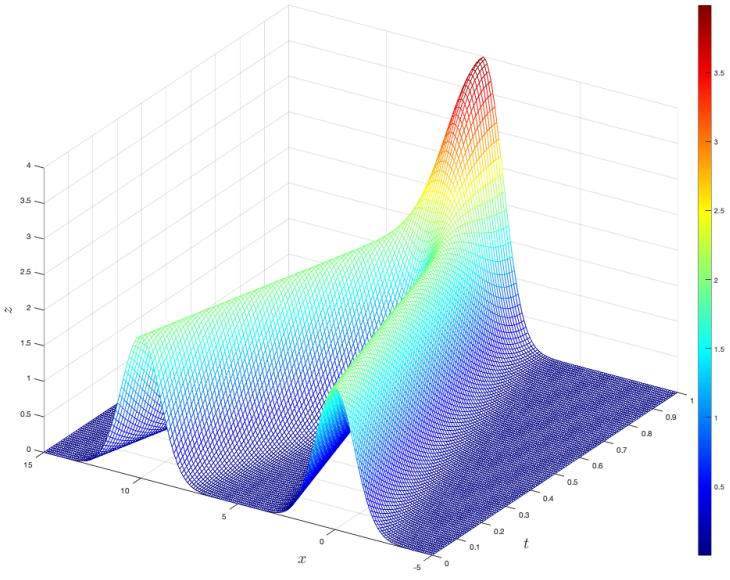

Figure 4.3: Illustration of density model (iii), where x and t axes represent the spatial and temporal domains of $f(x,t)$ and $z = f(x,t)$ represents the density values, with $f(x,0)$ at the front and $f(x,1)$ at the back.

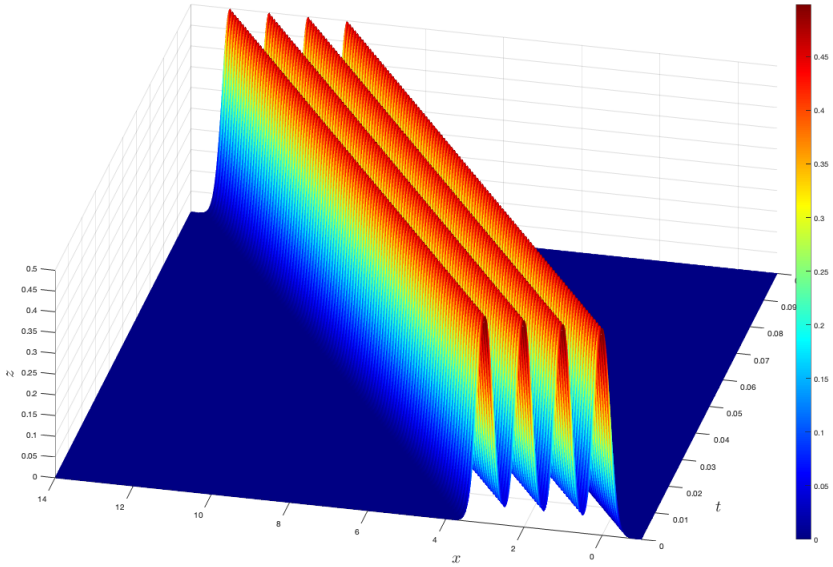

Figure 4.4: Illustration of density model (iv), where x and t axes represent the spatial and temporal domains of $f(x,t)$ and $z = f(x,t)$ represents the density values. For better visual effects, only $f(x,t)$, $t \in (0, 0.1)$ are shown, with $f(x,0)$ at the front and $f(x,0.1)$ at the back.

Models (i) and (ii) are similar to those used in Hall et al. (2006), where data are simulated from bivariate normal distributions with time-varying means and variances. However, Hall et al. (2006) did not consider time-varying mixture normal densities such as models (iii) and (iv). Many works on nonparametric density estimation for streaming data used simulated samples from time-varying mixture normal densities (Caudle and Wegman, 2009; Cao et al., 2012; García-Treviño and Barria, 2012; Qahtan et al., 2017).

We first applied the SKDE $\check{f}$ equipped with the SCV procedure to data streams simulated from models (i)–(iv). To initialise the SCV procedure, we need to select the first block size b_1, which depends on the specific data set at hand. Ideally, it should be the largest integer such that the first b_1 data points are nearly i.i.d. In the simulation studies we took $b_1 = 50$. To initialise the candidate set I_h^1 for the h values, we applied the normal reference rule (Scott and Sain, 2004) to the first b_1 observations to obtain an initial bandwidth $h_0 = 1.06\hat{\sigma}_0 b_1^{-1/5}$, where $\hat{\sigma}_0$ is the sample standard deviation computed from the first b_1 observations. Then I_h^1 was chosen as an equidistant grid of size $g = 10$ on the interval $[0.2h_0, 2h_0]$, where the lower and upper limits are chosen in a way to take into account the fact that the normal reference rule often produces

overly large bandwidths. We selected the set I_γ^1 as an equidistant grid of size $g = 10$ on the interval $[b_1^{-1}, 0.1]$, where b_1^{-1} is a value for γ we would take if the first b_1 data were i.i.d. and 0.1 is the γ value we would take if the appropriate sample size used to calculate an up-to-date density estimate is around 10. Hence we have initialised the set $I_{\gamma,h}^1 = I_\gamma^1 \times I_h^1$ of cardinality $\#(I_{\gamma,h}^1) = g^2 = 100$. For the ACV and the SCV criteria at (2.20) and (2.26), we imposed the restriction that $b_\ell \geq 10$, so that these cross-validation criteria have to be computed from more than 10 data points. This is to prevent the ACV and the SCV from being numerically unstable.

To illustrate the superiority of our algorithm over the sliding window KDE at (1.6) and the temporal KDE of Hall et al. (2006) defined at (1.13), we also applied the latter two methods to the simulated data streams. We computed the sliding window KDE $\hat{f}_w$ with four choices of window sizes $w = 25, 50, 100, 200$. We tried a range of w values because, for streaming data with different types of time variabilities, we did not expect that any particular choice of w would always result in a good estimator $\hat{f}_w$. These particular w values were chosen to make sure that, when estimating models (i)–(iv) with different sample sizes, there is at least one $\hat{f}_w$ performing reasonably well. At time t, we used the direct plug-in method proposed by Sheather and Jones (1991) and the leave-one-out least squares cross-validation (LSCV, see e.g. Wand and Jones, 1995, pp. 63–65) to select the bandwidth h_{n_t}, using data $\{X_i\}_{i=n_t-w+1,\ldots,n_t}$ in the sliding window. Let $\hat{f}_w^{\mathrm{PI}}$ and $\hat{f}_w^{\mathrm{CV}}$ denote the sliding window KDEs equipped with the plug-in and the LSCV bandwidth selectors, respectively. We used the `kde`, `hpi` and `hlscv` functions from the R package `ks` (Duong et al., 2019) to compute the KDE, the plug-in bandwidths and the LSCV bandwidths.

We computed the temporal KDE of Hall et al. (2006), denoted by $\hat{f}_{\mathrm{HMW}}$, with temporal kernel $K_T(u) = 4 - 6u$, for $u \in [0, 1]$, as in Hall et al. (2006). The temporal bandwidth λ was selected by the time dynamic least squares cross-validation procedure described in Hall et al. (2006). However, since their cross-validation method is not designed for streaming data, we made some modifications to it as follows, both to improve its performance and to reduce its computational cost. First, for each n_t, the temporal bandwidth λ_{n_t} is selected from an equidistant grid of size 20 on the interval $[20\Delta t, 500\Delta t]$, instead of on $[0, 1]$. That is, the number of data

falling in the time interval $[t-\lambda_{n_t}, t]$ is between 20 and 500. This is to prevent λ_{n_t} from being either too small, causing numerical instability for the estimator and the cross-validation procedure; or being too large, making the computation of $\hat{f}_{\mathrm{HMW}}$ too time-consuming. We observed that the performance of $\hat{f}_{\mathrm{HMW}}$ with this restriction was slightly better than without, while the average computation time was significantly reduced compared to the original cross-validation method in Hall et al. (2006). For selecting the spatial bandwidth h_{n_t}, we used the plug-in method instead of the normal reference rule as in Hall et al. (2006), since the latter does not work well when the true density is very different from the normal density (e.g. in model (iii) or (iv)).

To compare the performances of $\check{f}$, $\hat{f}_w^{\mathrm{PI}}$, $\hat{f}_w^{\mathrm{CV}}$, for $w = 25, 50, 100, 200$, and $\hat{f}_{\mathrm{HMW}}$, we generated 100 data streams for each of models (i)–(iv) and the 3 different sample sizes ($4 \times 3 \times 100$ data streams in total). For each data stream and each estimator, we calculated the integrated square error $\mathrm{ISE}(\check{f}) = \int_{t_{200}}^{1} \int_{C_t}^{D_t} \{\bar{f}(x,t) - f(x,t)\}^2 \, \mathrm{d}x \, \mathrm{d}t$, where $\bar{f}$ denotes one of $\check{f}$, $\hat{f}_w^{\mathrm{PI}}$, $\hat{f}_w^{\mathrm{CV}}$ and $\hat{f}_{\mathrm{HMW}}$ and $[C_t, D_t]$ denotes an interval large enough such that $\int_{C_t}^{D_t} f(x,t) \, \mathrm{d}x \approx 1$. We took $C_t = \min(B_{\ell_t}) - 3\sigma(B_{\ell_t})$ and $D_t = \max(B_{\ell_t}) + 3\sigma(B_{\ell_t})$, where, recalling from above (2.6), B_{ℓ_t} denotes the block of data arriving on a time interval containing t and where $\sigma(B_{\ell_t})$ denotes the sample standard deviation computed from B_{ℓ_t}. Note that $\mathrm{ISE}(\check{f})$ is defined for time interval $[t_{200}, 1]$ instead of $[0, 1]$, since $\hat{f}_{200}^{\mathrm{PI}}$ and $\hat{f}_{200}^{\mathrm{CV}}$ can only be computed when there are 200 or more data. In our setting, t_{200} was close to time 0 anyway, and we cannot compute any of the estimators well without at least some data to compute it.

Table 4.1 reports the median and the inter-quartile range (IQR) of the 100 ISEs for estimating models (i)–(iv) with $\check{f}$, $\hat{f}_w^{\mathrm{PI}}$ and $\hat{f}_w^{\mathrm{CV}}$, for $w = 25, 50, 100, 200$. These results show that $\hat{f}_w^{\mathrm{PI}}$ significantly outperformed $\hat{f}_w^{\mathrm{CV}}$, for $w = 25, 50, 100, 200$ and all data-generating models, in terms of the median ISE. They also show that $\check{f}$ significantly outperformed $\hat{f}_w^{\mathrm{CV}}$, for $w = 25, 50, 100, 200$. Furthermore, in all except two cases (models (iii) and (iv) when $n_1 = 5 \times 10^3$), $\check{f}$ also significantly outperformed $\hat{f}_w^{\mathrm{PI}}$, for $w = 25, 50, 100, 200$. In the case of estimating model (iii) when $n_1 = 5 \times 10^3$, the median ISE of the SKDE $\check{f}$ is slightly larger than, but still comparable to, the median ISE of $\hat{f}_{200}^{\mathrm{PI}}$. For estimating model (iv) when $n_1 = 5 \times 10^3$, the median ISE of $\check{f}$ is the same as that of $\hat{f}_{25}^{\mathrm{PI}}$.

In addition to the superiority in performance, another strength of the SKDE $\check{f}$ is that, by selecting the γ values using the SCV, it automatically selects an appropriate number of recent observations to compute the density estimate (recall that, with different choices of the γ values the effective number of data used to compute $\check{f}$ differs). In contrast, the sliding window KDEs $\hat{f}_w^{\mathrm{PI}}$ and $\hat{f}_w^{\mathrm{CV}}$ only performs well when the window size w is carefully chosen by the user. For example, from Table 4.1 we can see that models (ii) and (iv) require smaller w to obtain a small median ISE compared to models (i) and (iii). This is because the time variabilities of the density f in the latter two models are relatively mild compared to the former. Hence, the adaptivity to different kinds of time variabilities makes our method more preferable in practice.

Table 4.1: Simulation results for estimating models (i)–(iv) with $\check{f}$, $\hat{f}_w^{\mathrm{PI}}$ and $\hat{f}_w^{\mathrm{CV}}$, for $w = 25, 50, 100, 200$, when $n_1 = 5 \times 10^3$, 10^4 or 1.5×10^4. The numbers show $10^3 \times$ ISE$(\bar{f})$[1st quartile, 3rd quartile] calculated from 100 data streams, where $\bar{f}$ denotes one of $\check{f}$, $\hat{f}_w^{\mathrm{PI}}$ and $\hat{f}_w^{\mathrm{CV}}$.

$n_1 = 5 \times 10^3$	$\check{f}$	$\hat{f}_{25}^{\mathrm{PI}}$	$\hat{f}_{50}^{\mathrm{PI}}$	$\hat{f}_{100}^{\mathrm{PI}}$	$\hat{f}_{200}^{\mathrm{PI}}$
Model (i)	2.70[2.52, 2.94]	4.56[4.38, 4.72]	3.09[2.94, 3.29]	4.53[4.26, 4.78]	11.9[11.6, 12.3]
Model (ii)	13.0[12.6, 13.6]	16.1[15.5, 16.7]	18.1[17.5, 19.0]	39.2[38.3, 40.1]	74.9[74.0, 75.9]
Model (iii)	4.72[4.37, 5.06]	18.2[17.8, 18.6]	10.7[10.4, 10.9]	6.37[6.12, 6.60]	4.30[4.04, 4.57]
Model (iv)	135[134, 136]	134[134, 135]	145[144, 146]	179[178, 180]	234[232, 234]
$n_1 = 10^4$	$\check{f}$	$\hat{f}_{25}^{\mathrm{PI}}$	$\hat{f}_{50}^{\mathrm{PI}}$	$\hat{f}_{100}^{\mathrm{PI}}$	$\hat{f}_{200}^{\mathrm{PI}}$
Model (i)	2.08[1.92, 2.26]	4.68[4.55, 4.88]	2.74[2.64, 2.88]	2.46[2.34, 2.66]	5.21[4.95, 5.47]
Model (ii)	9.04[8.88, 9.35]	14.0[13.6, 14.3]	10.1[9.78, 10.4]	15.3[14.9, 15.8]	39.0[38.5, 39.7]
Model (iii)	3.30[3.11, 3.49]	18.3[18.0, 18.7]	10.7[10.4, 11.0]	6.26[6.03, 6.41]	3.79[3.61, 3.98]
Model (iv)	111[110, 113]	128[127, 128]	127[126, 127]	144[143, 144]	178[177, 179]
$n_1 = 1.5 \times 10^4$	$\check{f}$	$\hat{f}_{25}^{\mathrm{PI}}$	$\hat{f}_{50}^{\mathrm{PI}}$	$\hat{f}_{100}^{\mathrm{PI}}$	$\hat{f}_{200}^{\mathrm{PI}}$
Model (i)	1.71[1.62, 1.82]	4.75[4.63, 4.92]	2.68[2.60, 2.78]	1.97[1.86, 2.06]	3.04[2.89, 3.21]
Model (ii)	7.44[7.18, 7.70]	13.6[13.3, 13.9]	8.59[8.34, 8.86]	9.43[9.17, 9.72]	21.9[21.3, 22.2]
Model (iii)	2.65[2.53, 2.78]	18.3[18.1, 18.6]	10.7[10.5, 10.8]	6.24[6.10, 6.37]	3.70[3.59, 3.80]
Model (iv)	93.2[92.4, 94.0]	126[126, 127]	119[119, 120]	131[131, 132]	154[154, 155]
$n_1 = 5 \times 10^3$	$\check{f}$	$\hat{f}_{25}^{\mathrm{CV}}$	$\hat{f}_{50}^{\mathrm{CV}}$	$\hat{f}_{100}^{\mathrm{CV}}$	$\hat{f}_{200}^{\mathrm{CV}}$
Model (i)	2.70[2.52, 2.94]	5.48[5.22, 5.71]	3.66[3.46, 3.87]	4.80[4.54, 5.08]	12.30[11.80, 12.70]
Model (ii)	13.0[12.6, 13.6]	19.0[18.2, 19.7]	20.0[19.2, 20.6]	40.7[39.7, 41.8]	83.2[82.0, 84.2]
Model (iii)	4.72[4.37, 5.06]	21.0[20.5, 21.6]	12.1[11.7, 12.6]	7.23[6.87, 7.55]	4.85[4.49, 5.26]
Model (iv)	135[134, 136]	154[153, 156]	150[149, 151]	185[183, 185]	264[262, 265]
$n_1 = 10^4$	$\check{f}$	$\hat{f}_{25}^{\mathrm{CV}}$	$\hat{f}_{50}^{\mathrm{CV}}$	$\hat{f}_{100}^{\mathrm{CV}}$	$\hat{f}_{200}^{\mathrm{CV}}$
Model (i)	2.08[1.92, 2.26]	5.70[5.51, 5.95]	3.43[3.24, 3.57]	2.82[2.69, 3.01]	5.44[5.14, 5.83]
Model (ii)	9.04[8.88, 9.35]	16.9[16.4, 17.5]	12.0[11.5, 12.4]	16.2[15.7, 16.7]	39.1[38.7, 40.5]
Model (iii)	3.30[3.11, 3.49]	21.0[20.3, 21.5]	12.2[11.7, 12.6]	7.04[6.78, 7.35]	4.32[4.02, 4.67]
Model (iv)	111[110, 113]	132[131, 133]	148[147, 150]	147[146, 148]	185[184, 186]
$n_1 = 1.5 \times 10^4$	$\check{f}$	$\hat{f}_{25}^{\mathrm{CV}}$	$\hat{f}_{50}^{\mathrm{CV}}$	$\hat{f}_{100}^{\mathrm{CV}}$	$\hat{f}_{200}^{\mathrm{CV}}$
Model (i)	1.71[1.62, 1.82]	5.81[5.65, 6.01]	3.38[3.27, 3.50]	2.34[2.25, 2.48]	3.37[3.16, 3.66]
Model (ii)	7.44[7.18, 7.70]	16.6[16.3, 17.1]	10.6[10.3, 10.8]	10.5[10.2, 10.7]	22.3[21.9, 23.0]
Model (iii)	2.65[2.53, 2.78]	20.90[20.40, 21.20]	12.00[11.80, 12.30]	6.99[6.76, 7.24]	4.24[4.00, 4.99]
Model (iv)	93.2[92.4, 94.0]	122[122, 123]	118[117, 118]	160[159, 161]	155[155, 156]

Table 4.2 reports the median and the inter-quartile range (IQR) of the 100 ISEs for estimating

models (i)–(iv) with $\hat{f}_{\text{HMW}}$, when $n_1 = 5\times 10^3$. It also reports the average computation times (in seconds) for computing $\hat{f}_{\text{HMW}}$ and $\check{f}$ from one sample. The computation time, averaged over 100 simulations, includes the time for selecting smoothing parameters. Since the time for computing $\hat{f}_{\text{HMW}}$ is already prohibitively long for even the smallest sample size $n_1 = 5 \times 10^3$ ($\hat{f}_{\text{HMW}}$ takes more than 11 minutes to process one data stream, nearly 30 times more time-consuming than $\check{f}$), we did not compute it for larger sample sizes. Comparing Table 4.2 to Table 4.1, we can see that the performance of $\hat{f}_{\text{HMW}}$ is significantly worse than $\check{f}$ in terms of the ISE, although its computation time is dozens of times longer.

Table 4.2: Simulation results for estimating models (i)–(iv) with $\hat{f}_{\text{HMW}}$ and $\check{f}$ when $n_1 = 5\times 10^3$. Numbers in first two rows show $10^3 \times \text{ISE}(\bar{f})$[1st quartile, 3rd quartile], where $\bar{f}$ denotes one of $\hat{f}_{\text{HMW}}$ and $\check{f}$, calculated from 100 data streams. Numbers in the last two rows show average computation time for one sample (in seconds).

$n_1 = 5 \times 10^3$	Model (i)	Model (ii)	Model (iii)	Model (iv)
$\hat{f}_{\text{HMW}}$	$14.2[13.5, 14.8]$	$47.4[45.7, 49.1]$	$33.2[31.9, 34.6]$	$173[171, 174]$
$\check{f}$	$2.70[2.52, 2.94]$	$13.0[12.6, 13.6]$	$4.72[4.37, 5.06]$	$135[134, 136]$
Time of $\hat{f}_{\text{HMW}}$ (sec)	717	708	732	675
Time of $\check{f}$ (sec)	24.6	25.3	24.2	25.7

To visually illustrate the superiority of $\check{f}$ over $\hat{f}_w^{\text{PI}}$ for $w = 25, 50, 100, 200$ and $\hat{f}_{\text{HMW}}$, we computed these estimators from the first of the 100 samples simulated from models (i)–(iv) and plotted them in Figures 4.8–4.16. Recall that $\hat{f}_w^{\text{PI}}$ outperformed $\hat{f}_w^{\text{CV}}$ in all cases, so we did not plot the density estimates generated by $\hat{f}_w^{\text{CV}}$. Since the computation time of $\hat{f}_{\text{HMW}}$ for one sample is already very long for the smallest sample size $n_1 = 5 \times 10^3$, we did not compute it for the larger sample sizes $n_1 = 10^4, 1.5 \times 10^4$. In each of Figures 4.8–4.16, the 9 subplots correspond to 9 equidistant time points on the time interval $[t_{200}, 1]$. The true density curves are plotted in solid red lines, the estimated density curves produced by $\check{f}$ in solid blue lines, the $\hat{f}_{\text{HMW}}$ curves in dashed blue lines, the $\hat{f}_{25}^{\text{PI}}$ curves in dashed black lines, the $\hat{f}_{50}^{\text{PI}}$ curves in dotted black lines, the $\hat{f}_{100}^{\text{PI}}$ curves in dotted dashed black lines and the $\hat{f}_{200}^{\text{PI}}$ curves in long dashed black lines.

From Figures 4.5–4.7, we can see that, expect for estimating the density at time t_{200} (the first subplot), most of the estimated density curves are reasonably close to the true density curves. At

time t_{200}, $\hat{f}_{50}$ often produced the best results, whilst $\hat{f}_{200}$ often produced the worst. The $\check{f}$ curves at time t_{200} in these plots capture the location of the true density quite accurately, but they are somewhat oversmoothed.

Due to its complicated time variabilities, model (ii) is difficult to estimate, especially when the sample size $n_1 = 5 \times 10^3$ is relatively small. From Figures 4.8–4.8, we can see that the estimated density curves produced by the SKDE $\check{f}$ (solid blue line) are often closer to the true density curves. In particular, in Figure 4.8, we observe that the $\hat{f}_{\text{HMW}}$ estimates are sometimes too wiggly, since the cross-validation procedure in Hall et al. (2006) sometimes chooses an overly small temporal bandwidth λ_{n_t}. Also observe that the $\hat{f}_{\text{HMW}}$ estimates are sometimes negative. This is because $\hat{f}_{\text{HMW}}$ uses a one-sided second order temporal kernel $K_T(u) = 4 - 6u$, for $u \in [0, 1]$, giving some data points negative weights.

For estimating model (iii), we can see from Figures 4.11–4.13 that, again, $\hat{f}_{\text{HMW}}$ often produced very wiggly density estimates and negative density values. The estimator $\hat{f}_{200}$ sometimes produced oversmoothed density estimates, whilst density curves produced by $\hat{f}_{25}$, $\hat{f}_{50}$ and $\hat{f}_{100}$ are often fairly close to the truth. Overall, however, $\check{f}$ produced the best density estimates for model (iii).

Estimating model (iv) is very challenging since the true density has large curvature and its location varies very fast in time. Indeed, from Figure 4.14 we can see that, when $n_1 = 5 \times 10^3$, all density estimates are oversmoothed. This might indicate that this particular sample size was too small for estimating the fast-varying density well. We also observe that the density estimates produced by $\hat{f}^{\text{PI}}_{200}$ are often very bad at capturing the location of the true density, since they were computed using too many data points. As the sample size n_1 increases (see Figures 4.15 and 4.16), we observe that the density curves produced by $\check{f}$ become much closer to the true curves. In contrast, the other density estimates often fail to estimate the true number of modes (i.e. 4) due to serious oversmoothing.

From the above simulation results we conclude that the SKDE $\check{f}$ with smoothing parameters chosen by the SCV procedure has shown superior adaptivity to different types of time variabilities of the underlying density function, compared to classic KDE $\hat{f}_w$ equipped with sliding window

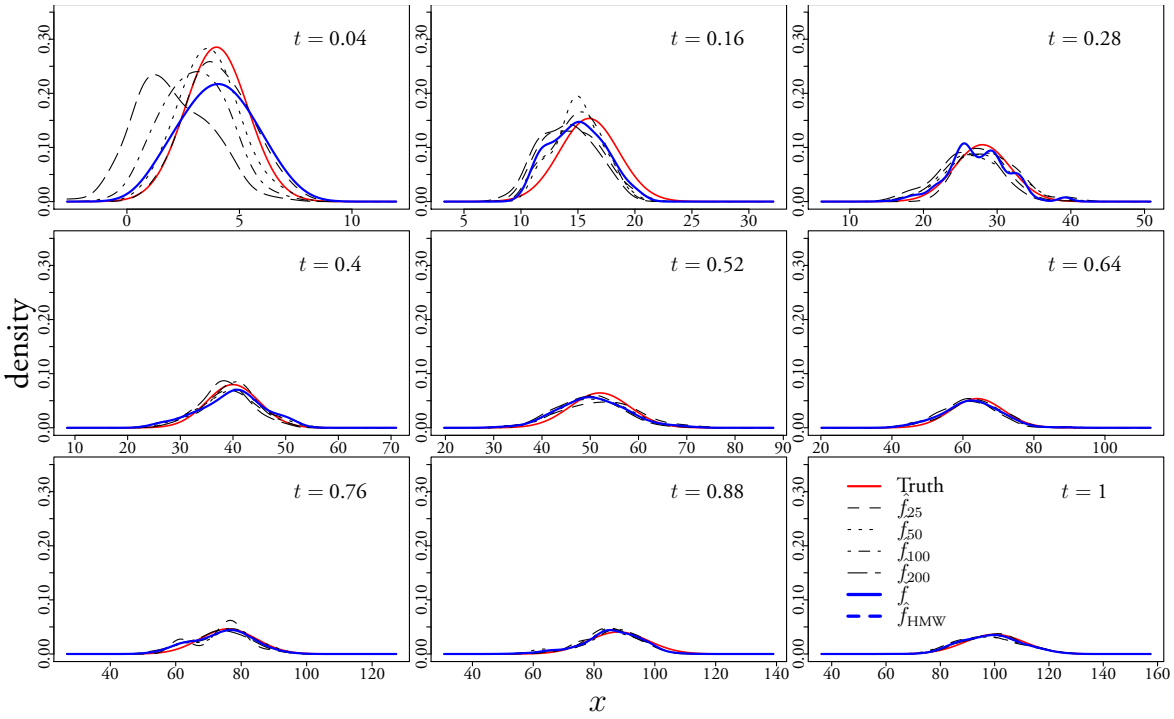

Figure 4.5: Density estimates for model (i) when $n_1 = 5 \times 10^3$. The 1-st to the 9-th subplots, corresponding to 9 equidistant time points $t = 0.04, 0.16, 0.28, 0.4, 0.52, 0.64, 0.76, 0.88, 1$ on time interval $[t_{200}, 1]$, show the true density curves (red line) and density estimates produced by $\hat{f}_{25}$ (dashed black line), $\hat{f}_{50}$ (dotted black line), $\hat{f}_{100}$ (dotted dashed black line), $\hat{f}_{200}$ (long dashed black line), $\check{f}$ (solid blue line) and $\hat{f}_{\mathrm{HMW}}$ (dashed blue line).

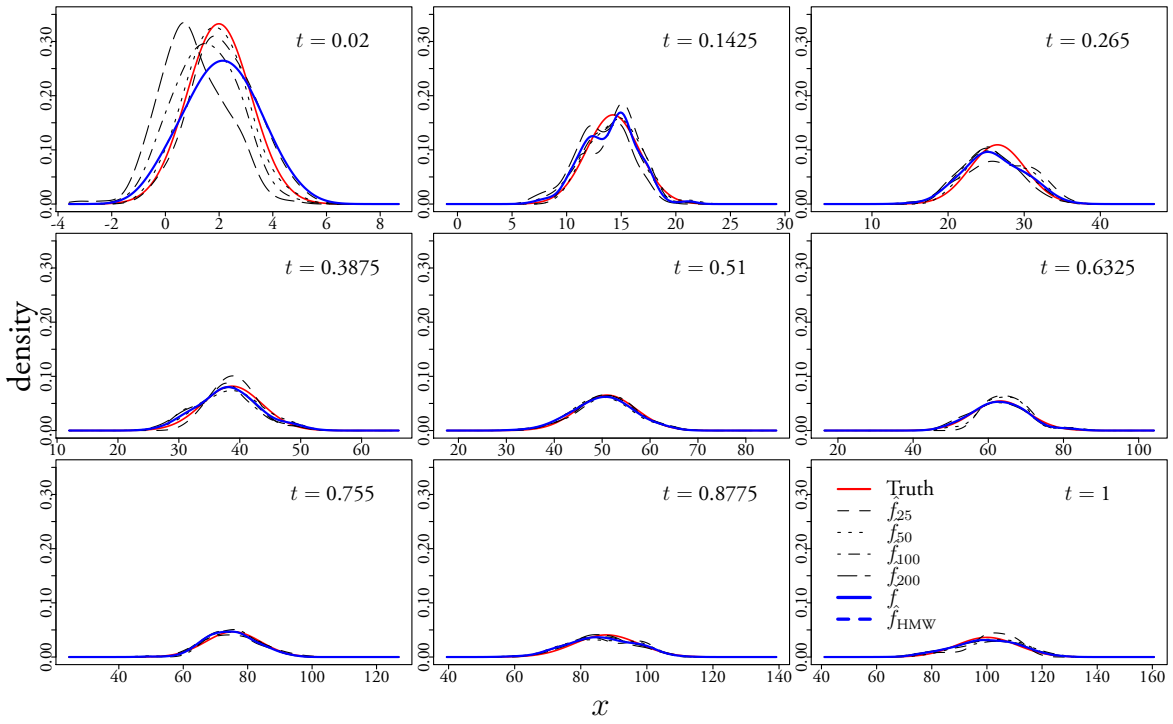

Figure 4.6: Density estimates for model (i) when $n_1 = 10^4$. The 1-st to the 9-th subplots, corresponding to 9 equidistant time points $t = 0.02, 0.1425, 0.265, 0.3875, 0.51, 0.6325, 0.755, 0.8775, 1$ on time interval $[t_{200}, 1]$, show the true density curves (red line) and density estimates produced by $\hat{f}_{25}$ (dashed black line), $\hat{f}_{50}$ (dotted black line), $\hat{f}_{100}$ (dotted dashed black line), $\hat{f}_{200}$ (long dashed black line) and $\check{f}$ (solid blue line).

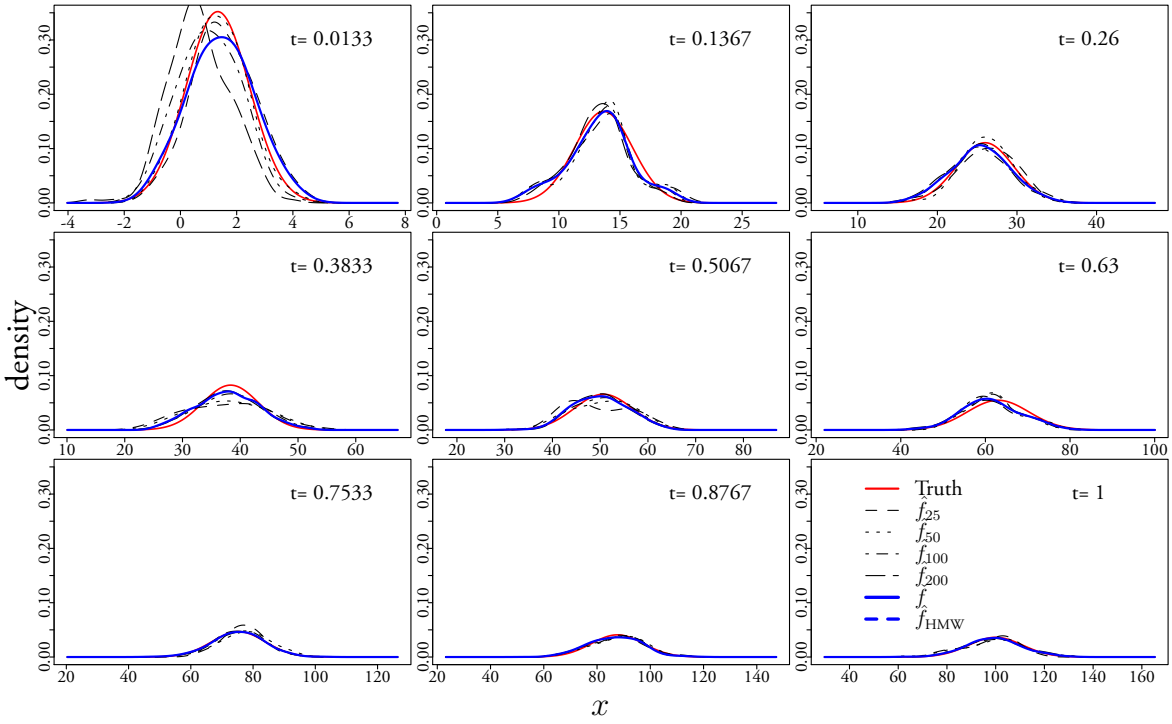

Figure 4.7: Density estimates for model (i) when $n_1 = 1.5 \times 10^4$. The 1-st to the 9-th subplots, corresponding to 9 equidistant time points $t = 0.0133, 0.1367, 0.26, 0.3833, 0.5067, 0.63, 0.7533, 0.8776, 1$ on time interval $[t_{200}, 1]$, show the true density curves (red line) and density estimates produced by $\hat{f}_{25}$ (dashed black line), $\hat{f}_{50}$ (dotted black line), $\hat{f}_{100}$ (dotted dashed black line), $\hat{f}_{200}$ (long dashed black line) and $\check{f}$ (solid blue line).

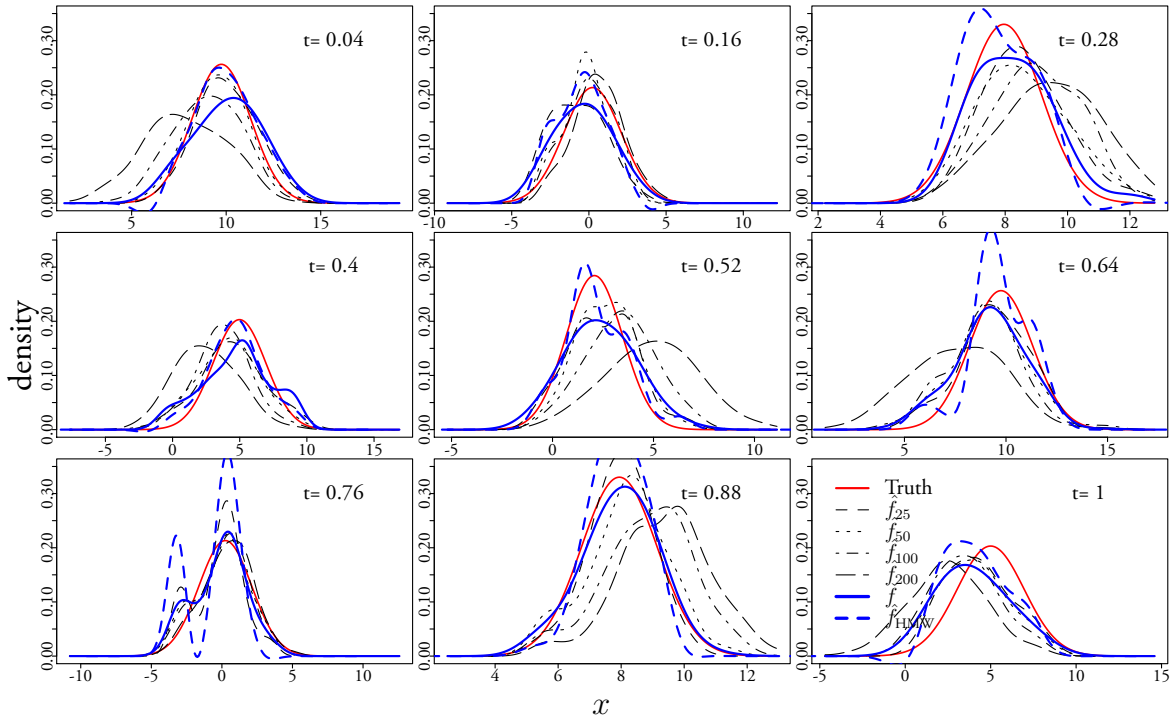

Figure 4.8: Density estimates for model (ii) when $n_1 = 5 \times 10^3$. The 1-st to the 9-th subplots, corresponding to 9 equidistant time points $t = 0.04, 0.16, 0.28, 0.4, 0.52, 0.64, 0.76, 0.88, 1$ on time interval $[t_{200}, 1]$, show the true density curves (red line) and density estimates produced by $\hat{f}_{25}$ (dashed black line), $\hat{f}_{50}$ (dotted black line), $\hat{f}_{100}$ (dotted dashed black line), $\hat{f}_{200}$ (long dashed black line), $\check{f}$ (solid blue line) and $\hat{f}_{\mathrm{HMW}}$ (dashed blue line).

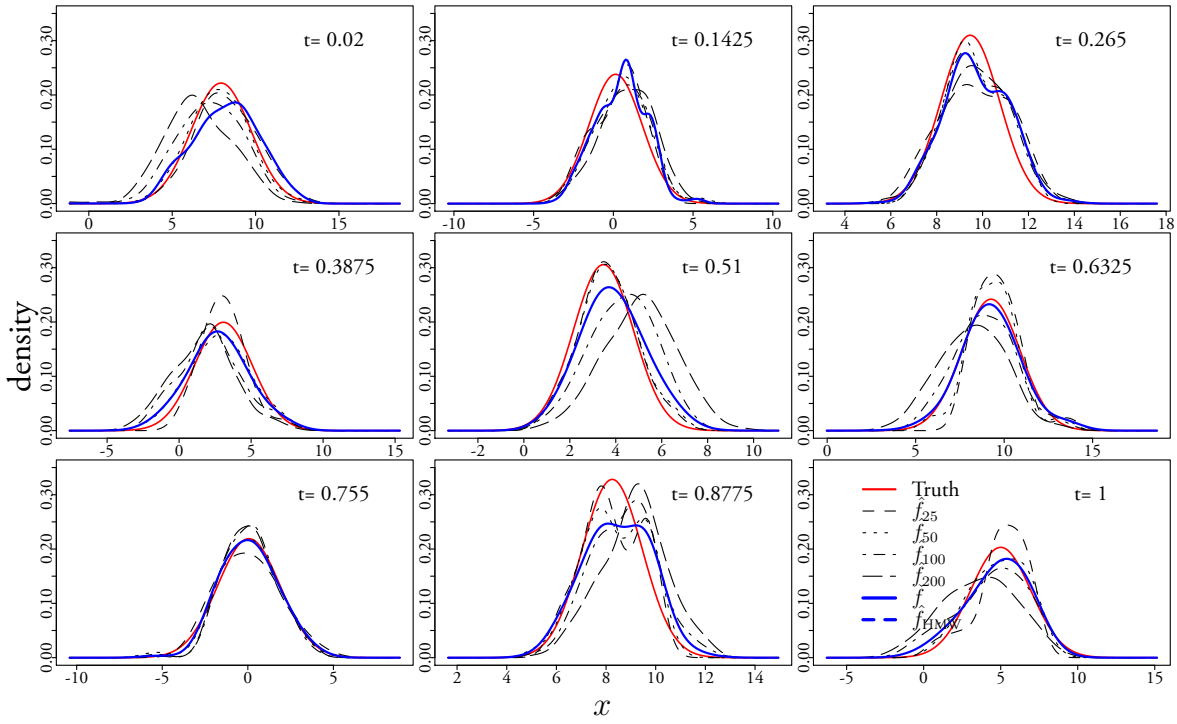

Figure 4.9: Density estimates for model (ii) when $n_1 = 10^4$. The 1-st to the 9-th subplots, corresponding to 9 equidistant time points $t = 0.02, 0.1425, 0.265, 0.3875, 0.51, 0.6325, 0.755, 0.8775, 1$ on time interval $[t_{200}, 1]$, show the true density curves (red line) and density estimates produced by $\hat{f}_{25}$ (dashed black line), $\hat{f}_{50}$ (dotted black line), $\hat{f}_{100}$ (dotted dashed black line), $\hat{f}_{200}$ (long dashed black line) and $\check{f}$ (solid blue line).

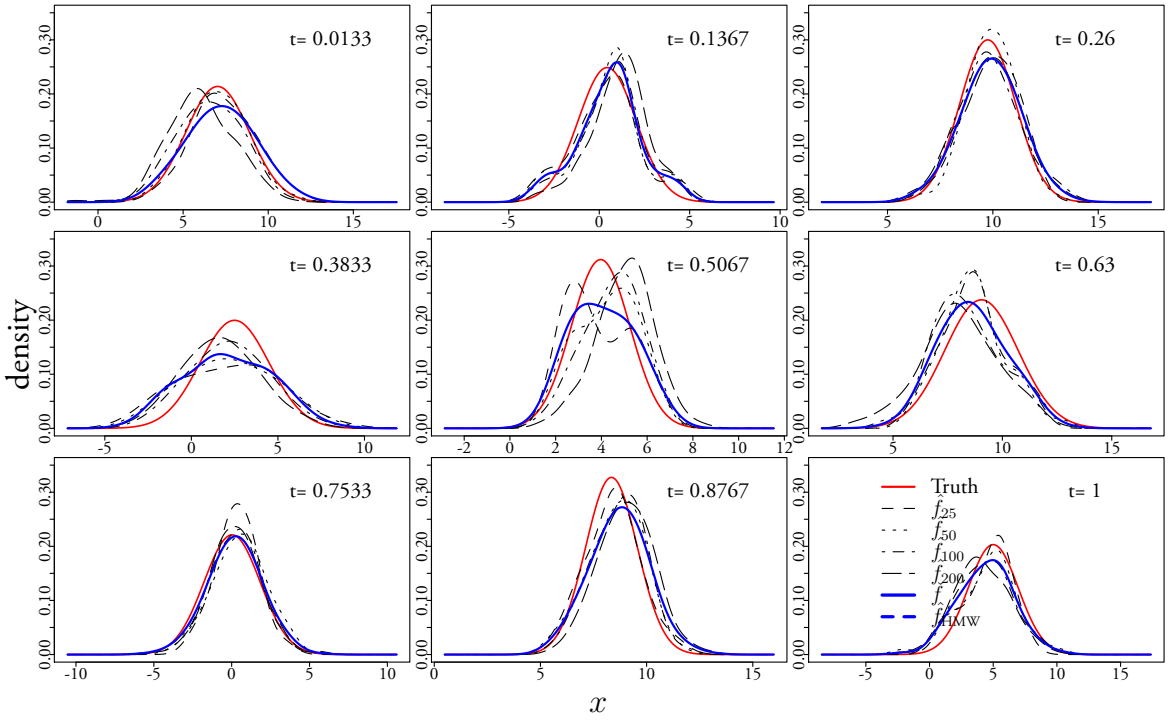

Figure 4.10: Density estimates for model (ii) when $n_1 = 1.5 \times 10^4$. The 1-st to the 9-th subplots, corresponding to 9 equidistant time points $t = 0.0133, 0.1367, 0.26, 0.3833, 0.5067, 0.63, 0.7533, 0.8776, 1$ on time interval $[t_{200}, 1]$, show the true density curves (red line) and density estimates produced by $\hat{f}_{25}$ (dashed black line), $\hat{f}_{50}$ (dotted black line), $\hat{f}_{100}$ (dotted dashed black line), $\hat{f}_{200}$ (long dashed black line) and $\check{f}$ (solid blue line).

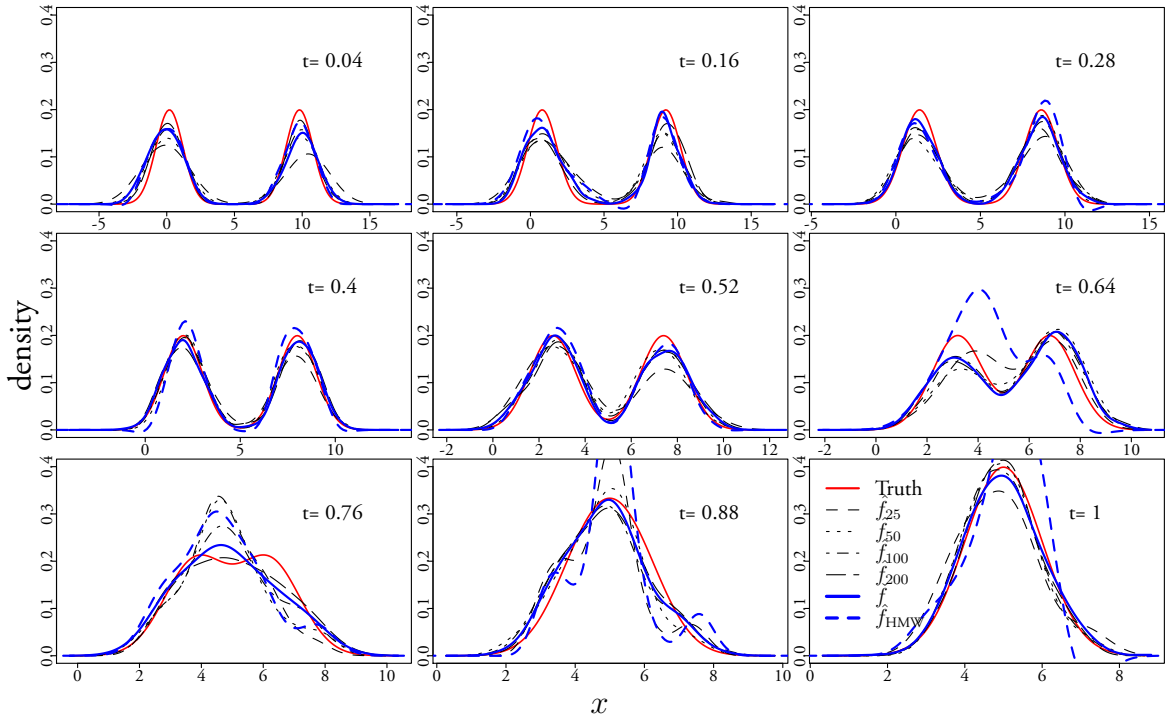

Figure 4.11: Density estimates for model (iii) when $n_1 = 5 \times 10^3$. The 1-st to the 9-th subplots, corresponding to 9 equidistant time points $t = 0.04, 0.16, 0.28, 0.4, 0.52, 0.64, 0.76, 0.88, 1$ on time interval $[t_{200}, 1]$, show the true density curves (red line) and density estimates produced by $\hat{f}_{25}$ (dashed black line), $\hat{f}_{50}$ (dotted black line), $\hat{f}_{100}$ (dotted dashed black line), $\hat{f}_{200}$ (long dashed black line), $\check{f}$ (solid blue line) and $\hat{f}_{\mathrm{HMW}}$ (dashed blue line).

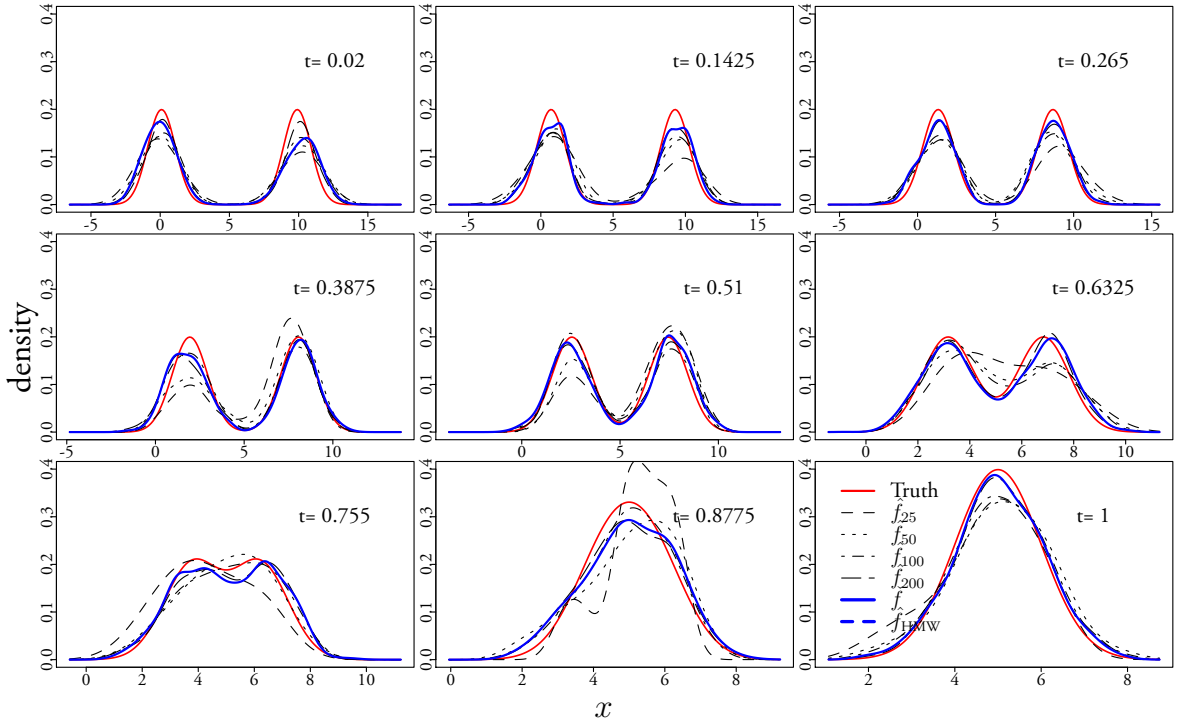

Figure 4.12: Density estimates for model (iii) when $n_1 = 10^4$. The 1-st to the 9-th subplots, corresponding to 9 equidistant time points $t = 0.02, 0.1425, 0.265, 0.3875, 0.51, 0.6325, 0.755, 0.8775, 1$ on time interval $[t_{200}, 1]$, show the true density curves (red line) and density estimates produced by $\hat{f}_{25}$ (dashed black line), $\hat{f}_{50}$ (dotted black line), $\hat{f}_{100}$ (dotted dashed black line), $\hat{f}_{200}$ (long dashed black line) and $\check{f}$ (solid blue line).

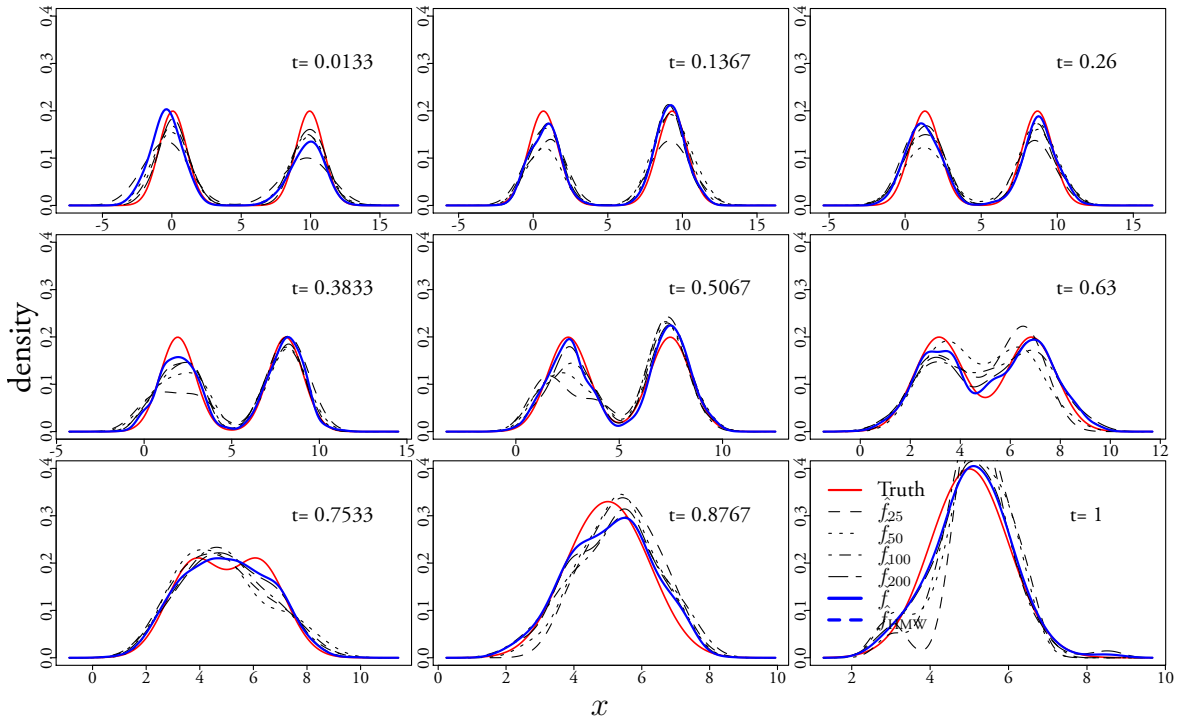

Figure 4.13: Density estimates for model (iii) when $n_1 = 1.5 \times 10^4$. The 1-st to the 9-th subplots, corresponding to 9 equidistant time points $t = 0.0133, 0.1367, 0.26, 0.3833, 0.5067, 0.63, 0.7533, 0.8776, 1$ on time interval $[t_{200}, 1]$, show the true density curves (red line) and density estimates produced by $\hat{f}_{25}$ (dashed black line), $\hat{f}_{50}$ (dotted black line), $\hat{f}_{100}$ (dotted dashed black line), $\hat{f}_{200}$ (long dashed black line) and $\check{f}$ (solid blue line).

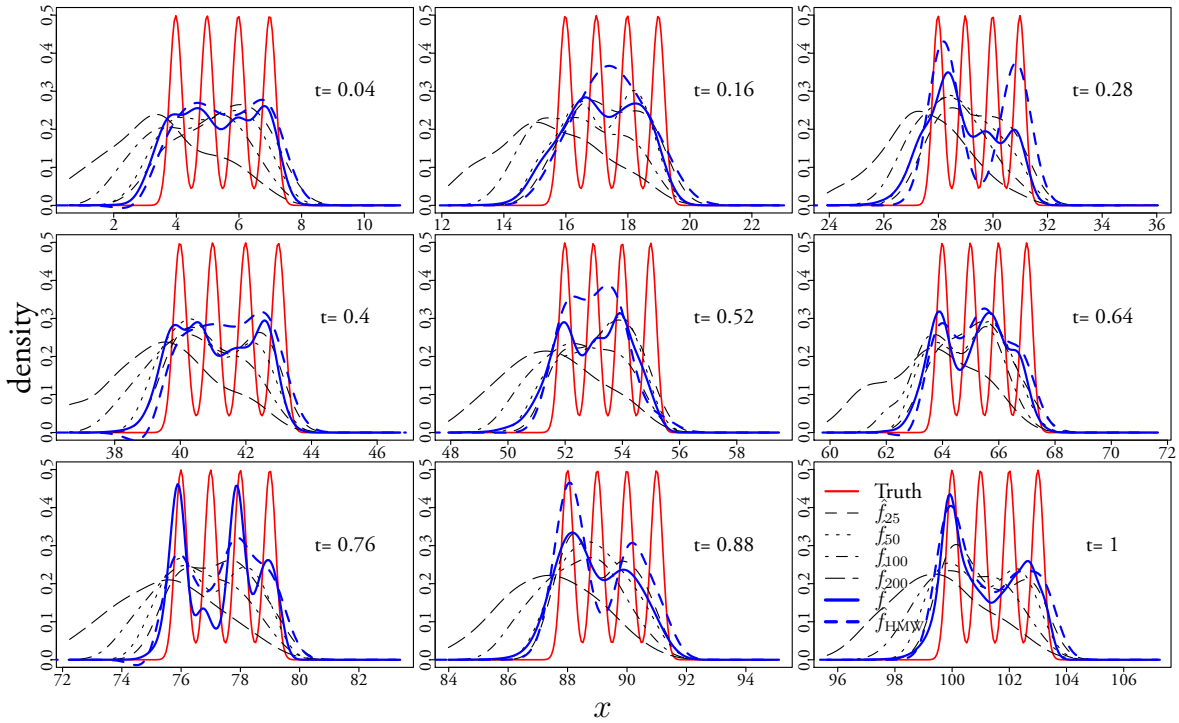

Figure 4.14: Density estimates for model (iv) when $n_1 = 5 \times 10^3$. The 1-st to the 9-th subplots, corresponding to 9 equidistant time points $t = 0.04, 0.16, 0.28, 0.4, 0.52, 0.64, 0.76, 0.88, 1$ on time interval $[t_{200}, 1]$, show the true density curves (red line) and density estimates produced by $\hat{f}_{25}$ (dashed black line), $\hat{f}_{50}$ (dotted black line), $\hat{f}_{100}$ (dotted dashed black line), $\hat{f}_{200}$ (long dashed black line), $\check{f}$ (solid blue line) and $\hat{f}_{\mathrm{HMW}}$ (dashed blue line).

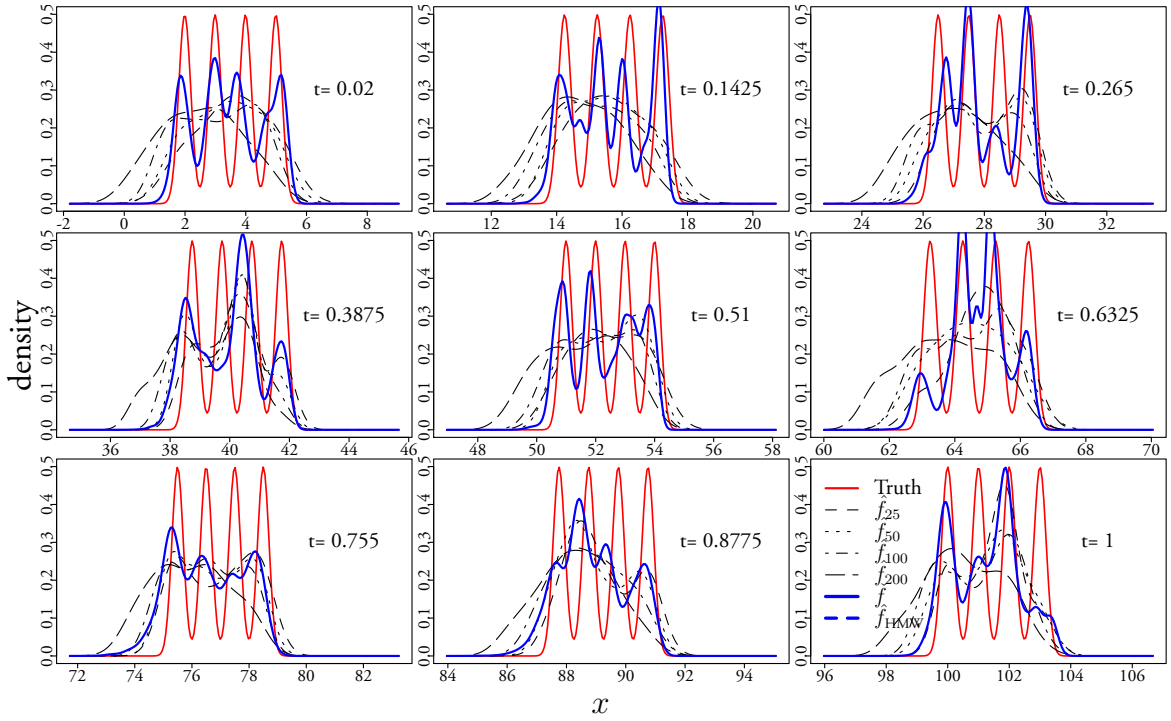

Figure 4.15: Density estimates for model (iv) when $n_1 = 10^4$. The 1-st to the 9-th subplots, corresponding to 9 equidistant time points $t = 0.02, 0.1425, 0.265, 0.3875, 0.51, 0.6325, 0.755, 0.8775, 1$ on time interval $[t_{200}, 1]$, show the true density curves (red line) and density estimates produced by $\hat{f}_{25}$ (dashed black line), $\hat{f}_{50}$ (dotted black line), $\hat{f}_{100}$ (dotted dashed black line), $\hat{f}_{200}$ (long dashed black line) and $\check{f}$ (solid blue line).

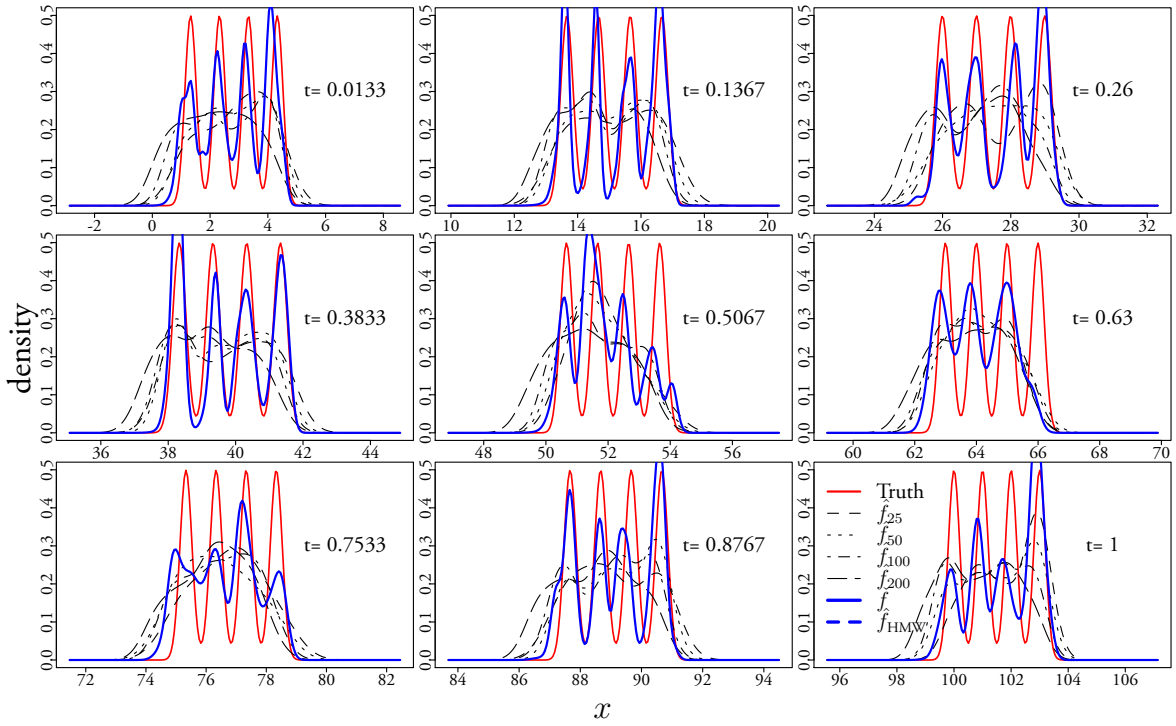

Figure 4.16: Density estimates for model (iv) when $n_1 = 1.5 \times 10^4$. The 1-st to the 9-th subplots, corresponding to 9 equidistant time points $t = 0.0133, 0.1367, 0.26, 0.3833, 0.5067, 0.63, 0.7533, 0.8776, 1$ on time interval $[t_{200}, 1]$, show the true density curves (red line) and density estimates produced by $\hat{f}_{25}$ (dashed black line), $\hat{f}_{50}$ (dotted black line), $\hat{f}_{100}$ (dotted dashed black line), $\hat{f}_{200}$ (long dashed black line) and $\check{f}$ (solid blue line).

and the temporal KDE $\hat{f}_{\text{HMW}}$ of Hall et al. (2006).

4.1.2 Regression

We ran simulations to illustrate the performance of the SRA defined in §3.2.1 when the underlying regression function has different types of time-variabilities. We simulated data streams $\{(X_i, Y_i)\}_{i=1,\ldots,n_1}$ arriving on a fixed time interval $[0, 1]$ for 3 different sample sizes $n_1 = 4\times10^3$, 8×10^3 and 1.2×10^4, where n_1 is defined at (1.2), taking $t = 1$. We simulated $\{X_i\}_{i=1,\ldots,n_1}$ from an ARMA$(2, 2)$ model $X_i = \varepsilon_i + \sum_{j=1}^{2} \varphi_j X_{i-j} + \sum_{j=1}^{2} \theta_j \varepsilon_{i-j}$, where $\{\varepsilon_j\}_{i=1,2,\ldots}$ is an i.i.d. $N(0, 0.1796)$ sequence and where $\varphi_1 = 0.8879$, $\varphi_2 = -0.4858$, $\theta_1 = -0.2279$ and $\theta_2 = 0.2488$. These particular parameters were selected to make sure that the serial dependence of $\{X_i\}_{i=1,\ldots,n_1}$ is not particularly strong. See Figure 4.17 for one realisation of $\{X_i\}_{i=1,\ldots,n_1}$ with $n_1 = 4 \times 10^3$.

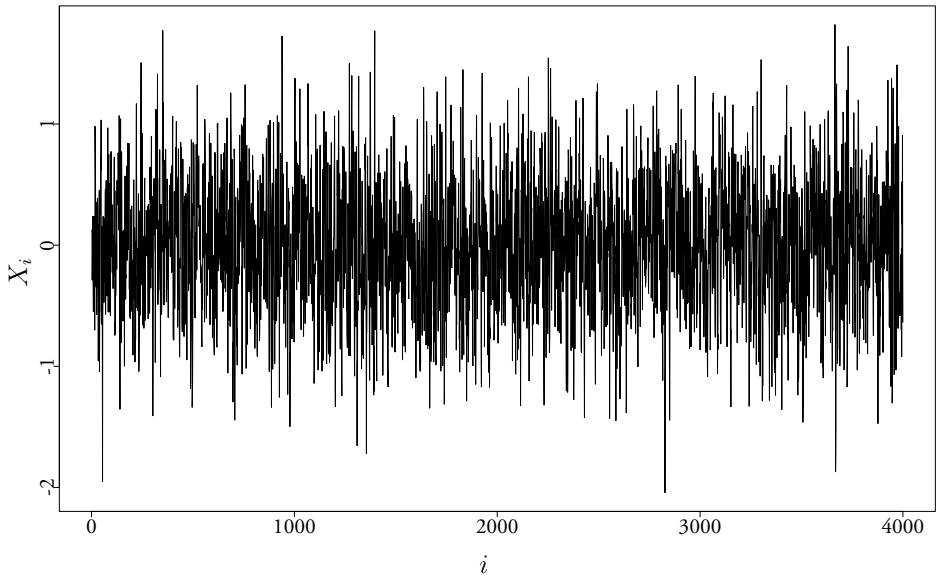

Figure 4.17: Time series simulated from an ARMA$(2, 2)$ model $X_i = \varepsilon_i + \sum_{j=1}^{2} \varphi_j X_{i-j} + \sum_{j=1}^{2} \theta_j \varepsilon_{i-j}$, where $\{\varepsilon_j\}_{i=1,2,\ldots}$ is an i.i.d. $N(0, 0.1796)$ sequence and where $\varphi_1 = 0.8879$, $\varphi_2 = -0.4858$, $\theta_1 = -0.2279$ and $\theta_2 = 0.2488$.

Furthermore, we simulated $\{Y_i\}_{i=1,\ldots,n_1}$ from the regression model (1.3) with $\sigma^2 = \text{var}(\epsilon_i) = 0.04$ and $m(x, t) = \sin\{A_t(x - B_t)\}$ with different values of (A_t, B_t). Similar time-

varying regression models have also been investigated in the simulation studies of e.g.Zhang and Wu (2015) and Bedi et al. (2019). We considered the following 4 models:

(i) $A_t = 3$, $B_t = 6t$, for $t \in [0, 1]$;

(ii) $A_t = 1 + 3t$, $B_t = 0$, for $t \in [0, 1]$;

(iii) $A_t = 1 + 3\sin(\pi t)$, $A_t = 6\sin(\pi t)$, for $t \in [0, 1]$;

(iv) $A_t = 3$, $B_t = 0$, for $t \in [0, 1]$.

In each model above, the true m is a sine function with time-varying or non-time-varying period A_t and location B_t. In model (i) and (ii), one of A_t and B_t is fixed and the other parameter is time-varying with a constant speed. In model (iii) both A_t and B_t are time-varying and in model (iv) both A_t and B_t are fixed.

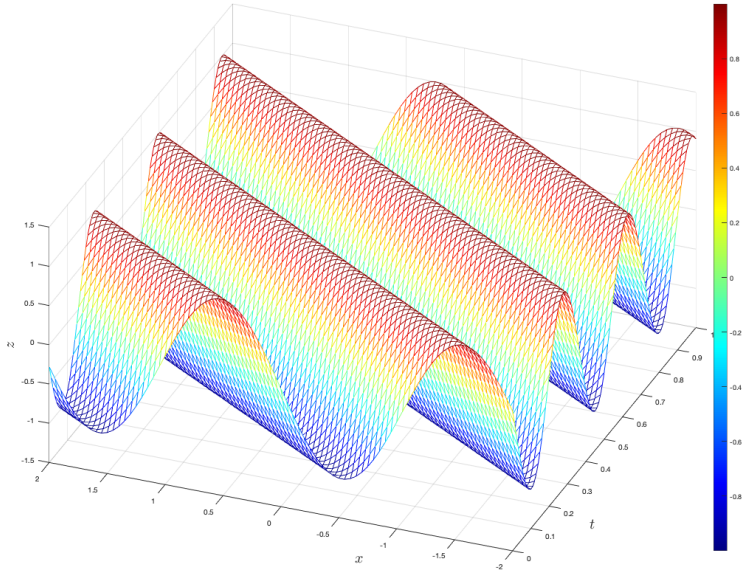

Figure 4.18: Illustration of density model (i), where x and t axes represent the spatial and temporal domains of $m(x, t)$ and $z = m(x, t)$, with $m(x, 0)$ at the front and $m(x, 1)$ at the back.

We first applied the SRA defined in §3.2.1 to data streams simulated from models (i)–(iv) when $n_1 = 4 \times 10^3$, 8×10^3 or 1.2×10^4. To initialise the algorithm, we took $\nu = 50$, 100, 150 when $n_1 = 4 \times 10^3$, 8×10^3 or 1.2×10^4, respectively (recall from §3.2.1.1 that the modified

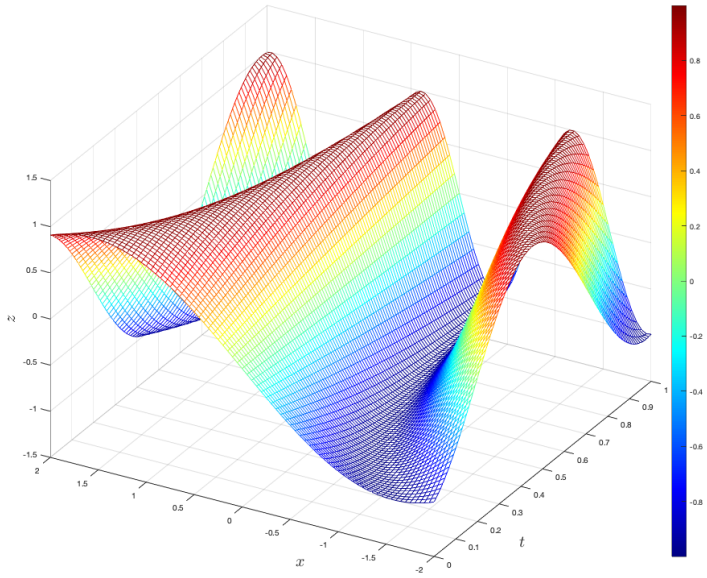

Figure 4.19: Illustration of density model (ii), where x and t axes represent the spatial and temporal domains of $m(x,t)$ and $z = m(x,t)$, with $m(x,0)$ at the front and $m(x,1)$ at the back.

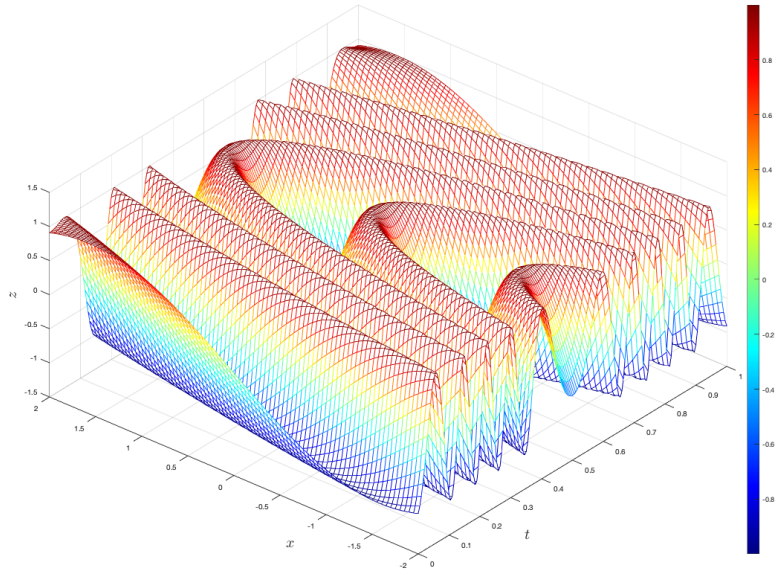

Figure 4.20: Illustration of density model (iii), where x and t axes represent the spatial and temporal domains of $m(x,t)$ and $z = m(x,t)$, with $m(x,0)$ at the front and $m(x,1)$ at the back.

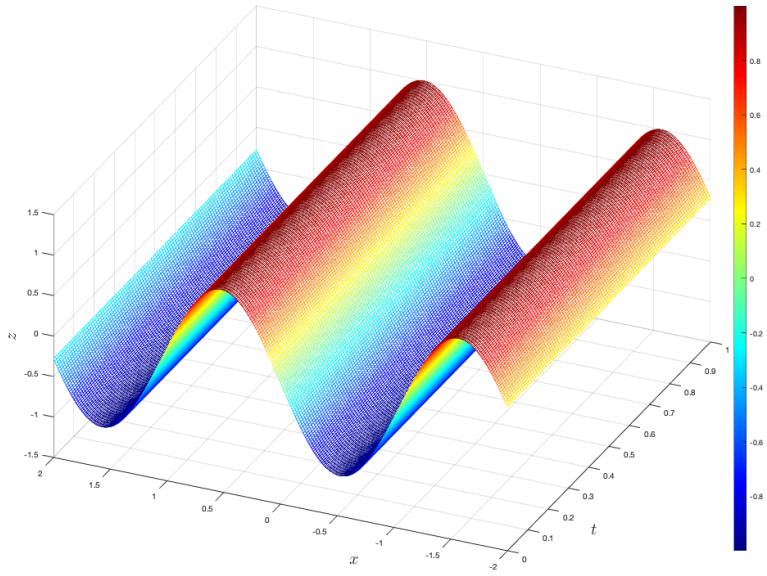

Figure 4.21: Illustration of density model (iv), where x and t axes represent the spatial and temporal domains of $m(x,t)$ and $z = m(x,t)$, with $m(x,0)$ at the front and $m(x,1)$ at the back.

algorithm uses the first $N_0 = 2\nu$ observations from the data stream for initialisation). Recall from §3.2.1.1 that, ideally, ν should be the largest number such that every ν data can be viewed as approximately stationary. We will see later in Table 4.4 that the SRA does not appear to be very sensitive to the choice of ν.

To initialise $I^1_{\gamma,h}$, similar to §4.1.1, we took I^1_γ as an equidistant grid of size $g = 10$ on intervals $[N_0^{-1}, 0.1]$, where N_0^{-1} is the γ value we would take if the first N_0 data were stationary. To initialise I^1_h, we used the plug-in rule for the local linear regression estimator described in Ruppert et al. (1995) to compute an initial bandwidth h_0 from the first N_0 data. Then I^1_h was chosen as an equidistant grid of size g on the interval $[0.2h_0, 5h_0]$. Hence $I^1_{\gamma,h} = I^1_\gamma \times I^1_h$ had cardinality $\#(I^1_{\gamma,h}) = g^2 = 100$. For numerical stability, we imposed the constraint $I^\ell_\gamma \in (0, 0.1)$ for all ℓ to prevent $\check{m}$ from selecting overly large γ values (recall from (3.7) that taking γ overly large implies that $\check{m}$ is computed from very few data). Lastly, we took $L = 1.2$ at (3.13).

To illustrate the superiority of our algorithm over the conventional kernel smoothing methodology, we also computed the local linear regression estimators (Fan and Gijbels, 1996) for the simulated data. Considering the time variability, we computed the local linear estimator $\hat{m}_w$

using data from a sliding window of size w. That is, for time $t \in [t_w, 1]$, $\hat{m}_w$ estimates $m(\cdot, t)$ using data $\{(X_i, Y_i)\}_{i=n_t-w+1,\ldots,n_t}$ arriving on a sliding window $(t - w\Delta t, t]$, as if they were identically distributed. Since it is well known that the local linear regression estimator enjoys some theoretical advantage over the NW estimator (see e.g. Wand and Jones, 1995, Chapter 5 and Fan and Gijbels, 1996, Chapters 2 and 3), we expected that the sliding window local linear estimator would perform better than the sliding window NW estimator at (1.7). Hence we computed the sliding window local linear estimates and compared them to estimates produced by our SRA.

To see the influence of the sliding window size w on the performance of $\hat{m}_w$, we computed two local linear estimators $\hat{m}_\nu$ and $\hat{m}_{2\nu}$, taking $w = \nu$ and $w = 2\nu$, respectively. The bandwidths for $\hat{m}_\nu$ and $\hat{m}_{2\nu}$ were selected by the leave-one-out LSCV. See Fan and Gijbels (1996, p. 44–45) and Härdle (1991, pp. 152–159) for details. We used the R package `locpol` (Cabrera, 2018) to implement the local linear regression methodology. In particular, the LSCV was implemented by the `regCVBwSelC` function, which is programmed in C.

For each of models (i)–(iv) and each of the 3 different sample sizes $n_1 = 4 \times 10^3$, 8×10^3 and 1.2×10^4, we generated 100 data streams ($4 \times 3 \times 100$ data streams in total). For each sample and each estimator, we calculated the integrated squared error $\mathrm{ISE}(\bar{m}) = \int_{t_{N_0}}^{1} \int_{C}^{D} \{\bar{m}(x,t) - m(x,t)\}^2 \, \mathrm{d}x \, \mathrm{d}t$, where $\bar{m}$ denotes one of $\check{m}$, $\hat{m}_\nu$ and $\hat{m}_{2\nu}$ and C, D denote the 0.05 and 0.95 sample quantiles of the sequence $\{X_i\}_{i=1,\ldots,n_1}$.

Table 4.3 reports the median and the inter-quartile range (IQR) of the 100 ISEs for estimating models (i)–(iv) with $\check{m}$, $\hat{m}_\nu$ and $\hat{m}_{2\nu}$. These results show that, for estimating model (i), the median ISE of $\check{m}$ is comparable to that of $\hat{m}_\nu$ and significantly smaller than that of $\hat{m}_{2\nu}$. For estimating model (ii), the median ISE of $\check{m}$ is comparable to that of $\hat{m}_{2\nu}$ and significantly smaller than that of $\hat{m}_\nu$. In all the remaining cases, $\check{m}$ significantly outperforms both $\hat{m}_\nu$ and $\hat{m}_{2\nu}$ in terms of the median ISE.

Compared to $\hat{m}_w$, for $w = \nu$ and 2ν, our $\check{m}$ shows adaptivity to different kinds of time variabilities. The estimator $\hat{m}_w$ behaves reasonably well when w is appropriate but its performance may dramatically worsen with an inappropriate w value. For example, note from Table

Table 4.3: Simulation results of $\check{m}$, $\hat{m}_\nu$ and $\hat{m}_{2\nu}$ for estimating models (i)–(iv), when $n_1 = 4 \times 10^3$, $n_1 = 8 \times 10^3$ or $n_1 = 1.2 \times 10^4$. The numbers show $10^3 \times \text{ISE}(\bar{m})$[1st quartile, 3rd quartile] calculated from 100 samples and the average computational time for computing each estimator on a grid of 200 design points at all time points from one sample of size n_1.

$n_1 = 4 \times 10^3$, $\nu = 50$	$\check{m}$	$\hat{m}_\nu$	$\hat{m}_{2\nu}$
Model (i)	21.4[20.6, 22.3]	33.0[28.0, 42.6]	31.5[30.7, 32.7]
Model (ii)	5.26[4.90, 5.61]	19.3[16.8, 24.9]	6.36[5.80, 7.12]
Model (iii)	44.4[43.2, 45.9]	80.4[76.0, 87.3]	186[183, 190]
Model (iv)	3.14[2.46, 3.75]	23.5[20.1, 32.3]	7.42[6.79, 8.21]
Average time (min)	0.154	0.101	0.250
$n_1 = 8 \times 10^3$, $\nu = 100$	$\check{m}$	$\hat{m}_\nu$	$\hat{m}_{2\nu}$
Model (i)	13.4[12.9, 13.9]	13.6[12.8, 14.3]	27.8[27.3, 28.4]
Model (ii)	3.53[3.28, 3.75]	6.39[6.00, 7.12]	3.33[3.15, 3.51]
Model (iii)	26.2[25.4, 26.9]	58.1[56.5, 59.2]	180[179, 182]
Model (iv)	2.21[1.85, 2.70]	7.67[7.29, 8.27]	3.83[3.56, 4.06]
Average time (min)	0.311	0.497	1.483
$n_1 = 1.2 \times 10^4$, $\nu = 150$	$\check{m}$	$\hat{m}_\nu$	$\hat{m}_{2\nu}$
Model (i)	10.0[9.81, 10.2]	10.8[10.6, 11.1]	26.6[26.4, 26.7]
Model (ii)	2.69[2.54, 2.81]	4.08[3.94, 4.34]	2.36[2.24, 2.57]
Model (iii)	19.6[19.1, 20.2]	54.7[54.0, 55.1]	180[179, 182]
Model (iv)	1.83[1.61, 2.09]	4.96[4.75, 5.23]	2.67[2.54, 2.87]
Average time (min)	0.545	1.28	4.68

4.3 that $\hat{m}_{2\nu}$ estimated model (ii) reasonably well but it always yielded the largest median ISE or estimating model (iii), which does not improve even when n_1 doubled or tripled. In contrast, $\check{m}$ often yielded the smallest median ISE and its median ISEs for all models significantly reduces when n_1 increases.

Table 4.3 also shows how the average computation times of $\check{m}$, $\hat{m}_\nu$ and $\hat{m}_{2\nu}$ at all n_1 time points for one data stream varies when n_1 increases. Except for the case $n_1 = 4 \times 10^3$, where $\check{m}$ was slower than $\hat{m}_\nu$, $\check{m}$ on average took significantly less time to process one data stream compared to the other two estimators. Furthermore, the computation time of $\check{m}$ increased roughly linearly as n_1 increased, whereas the computation times of $\hat{m}_\nu$ and $\hat{m}_{2\nu}$ increased much faster. Considering the fact that the bandwidth selection algorithm in the `locpol` package is programmed in C (Cabrera, 2018) and our modified algorithm is programmed purely in R, the difference in computational efficiency between $\check{m}$ and $\hat{m}_w$ may have been underestimated.

Furthermore, we argue that the performance of $\check{m}$ did not appear to be very sensitive to the selection of ν. Table 4.4 reports the median and the IQR of the 100 ISEs for estimating model (i)–(iv) with $\check{m}$, when $n_1 = 4 \times 10^3$, 8×10^3 or 1.2×10^4, and ν took different values compared to Table 4.3. These results show that, compared to Table 4.3, even when we halved or doubled the ν values, the changes in the median ISEs were insignificant.

Finally, for a visual illustration of the superiority of $\check{m}$ over $\hat{m}_\nu$ and $\hat{m}_{2\nu}$, we computed these estimators from the first of the 100 samples simulated from models (i)–(iv) for all 3 sample sizes and plotted them in Figures 4.22–4.33. In each of Figures 4.22–4.33, the 9 subplots correspond to 9 equidistant time points on the time interval $[t_{2\nu}, 1]$. The true regression curves were plotted in solid red lines. The estimated regression curves produced by $\check{m}$, $\hat{m}_\nu$ and $\hat{m}_{2\nu}$ were plotted in solid blue lines, dotted black lines and dashed black lines, respectively.

For estimating model (i), (ii) and (iv), we can see from Figures 4.22–4.27 and 4.33–4.33 that all 3 estimators produced some reasonable estimated regression curves. The $\check{f}$ curves are often closer to the truth compared to the $\hat{m}_\nu$ and the $\hat{m}_{2\nu}$ curves, although they are sometimes a bit wiggly. Moreover, $\hat{m}_\nu$ and $\hat{m}_{2\nu}$ occasionally produced oversmoothed estimates (e.g. the central subplot corresponding to $t = 0.5606$ in Figure 4.23).

Table 4.4: The role of ν. The numbers show $10^3 \times \mathrm{ISE}(\check{m})$[1st quartile, 3rd quartile] calculated from 100 samples.

	$n_1 = 4 \times 10^3$			$n_1 = 8 \times 10^3$			$n_1 = 1.2 \times 10^4$		
	$\nu = 25$	$\nu = 50$	$\nu = 100$	$\nu = 50$	$\nu = 100$	$\nu = 200$	$\nu = 75$	$\nu = 150$	$\nu = 300$
Model (i)	21.7[20.9, 23.1]	21.4[20.6, 22.3]	20.4[19.7, 21.1]	13.5[13.2, 13.9]	13.4[12.9, 13.9]	12.6[12.3, 13.0]	10.3[10.1, 10.7]	10.0[9.81, 10.2]	9.78[9.56, 10.1]
Model (ii)	5.60[5.16, 5.97]	5.26[4.90, 5.61]	4.92[4.56, 5.20]	3.60[3.37, 3.75]	3.53[3.28, 3.75]	3.36[3.18, 3.53]	2.75[2.61, 3.00]	2.69[2.54, 2.81]	2.69[2.57, 2.77]
Model (iii)	44.1[42.8, 45.5]	44.4[43.2, 45.9]	44.6[42.9, 46.6]	26.1[25.6, 26.9]	26.2[25.4, 26.9]	26.8[25.5, 27.5]	19.6[18.9, 19.9]	19.6[19.1, 20.2]	19.7[19.2, 20.1]
Model (iv)	3.53[2.79, 4.42]	3.14[2.46, 3.75]	3.30[2.61, 3.71]	2.29[2.04, 2.76]	2.21[1.85, 2.70]	2.02[1.86, 2.35]	1.72[1.51, 1.97]	1.83[1.61, 2.09]	1.96[1.85, 2.14]

Model (iii) is particularly difficult to estimate compared to the other 3 models, since, except for a short time period around $t = 0.5$, both A_t and B_t in model (iii) change very fast. From Figures 4.28–4.30, we can see that the $\check{m}$ curves are often fairly close to the truth, although sometimes the curves can be somewhat wiggly. In contrast, the $\hat{m}_{2\nu}$ curves are often smoother but more distant from the truth. Generally speaking, the $\hat{m}_{\nu}$ curves are closer to the truth compared to the $\hat{m}_{2\nu}$ curves but further from the truth compared to the $\check{m}$ curves.

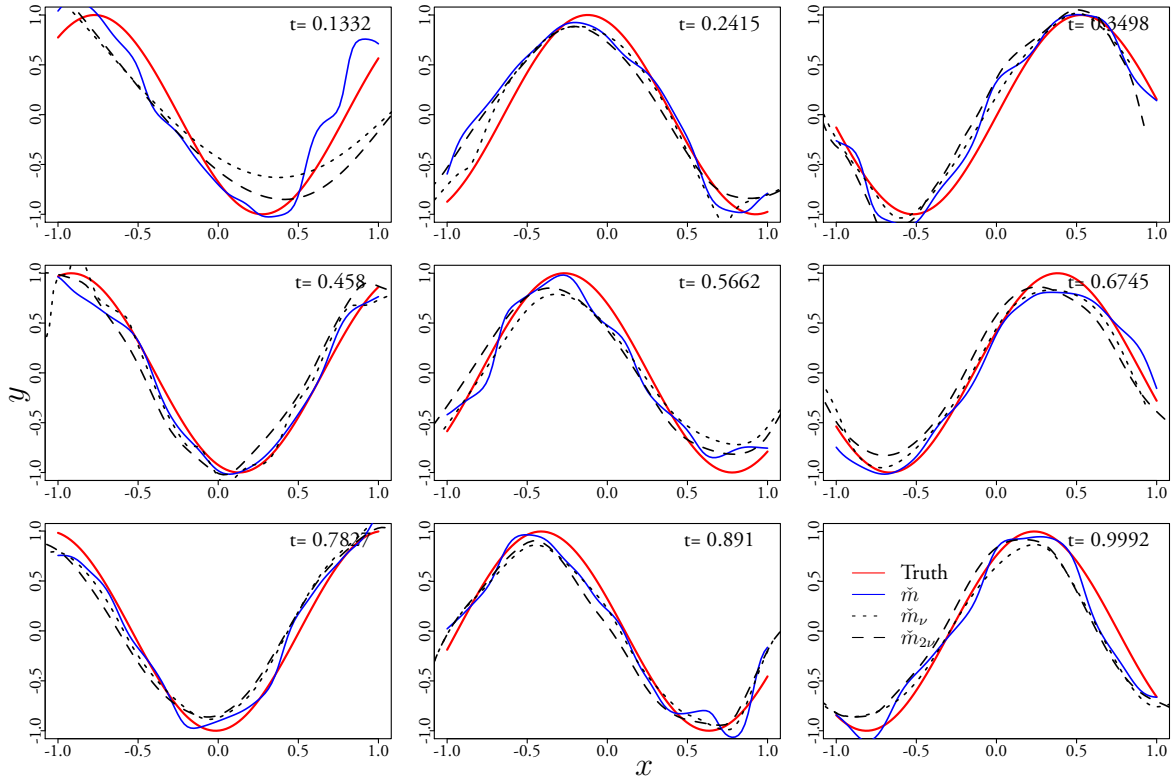

Figure 4.22: Comparison of $\check{m}$, $\hat{m}_{\nu}$ and $\hat{m}_{2\nu}$ for estimating model (i) when $n_1 = 4 \times 10^3$. The 9 subplots, corresponding to 9 equidistant time points on $[t_{N_0}, 1]$, show curves representing the true regression curve (solid line) and the estimates corresponding to $\check{m}$ (dotted dashed line), $\hat{m}_{\nu}$ (dotted line) and $\hat{m}_{2\nu}$ (dashed line).

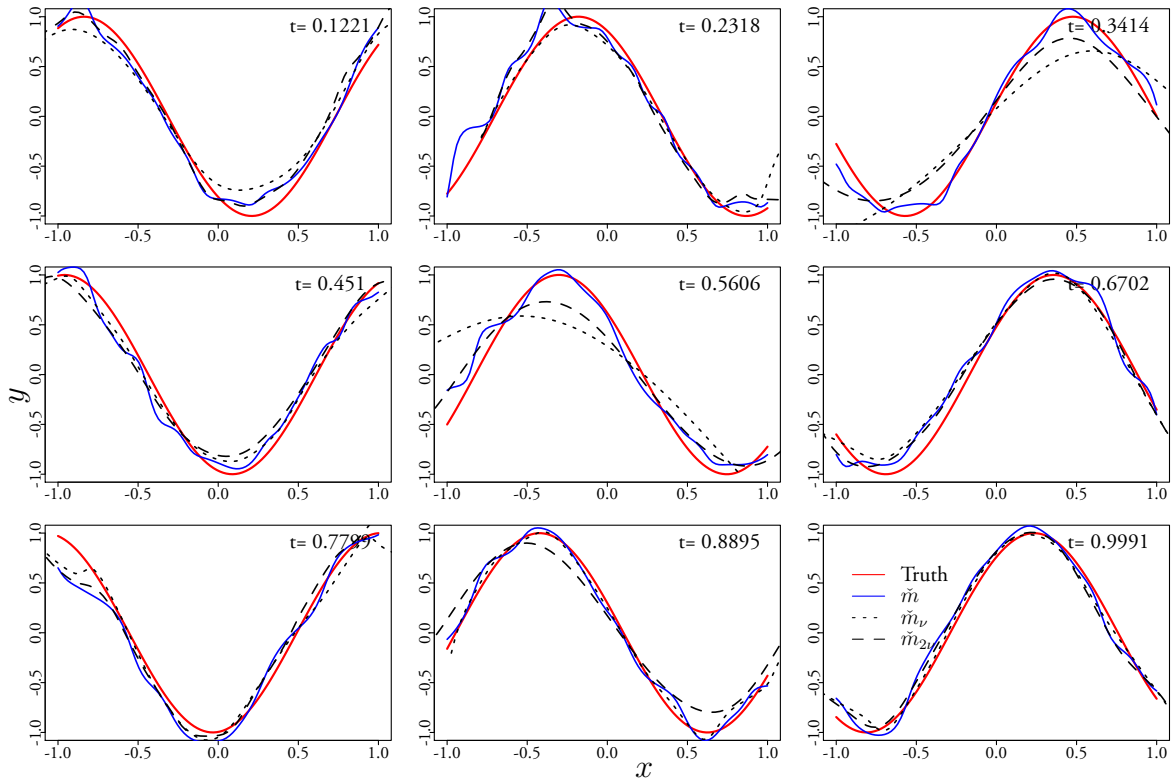

Figure 4.23: Comparison of $\check{m}$, $\hat{m}_\nu$ and $\hat{m}_{2\nu}$ for estimating model (i) when $n_1 = 8 \times 10^3$. The 9 subplots, corresponding to 9 equidistant time points on $[t_{N_0}, 1]$, show curves representing the true regression curve (solid line) and the estimates corresponding to $\check{m}$ (dotted dashed line), $\hat{m}_\nu$ (dotted line) and $\hat{m}_{2\nu}$ (dashed line).

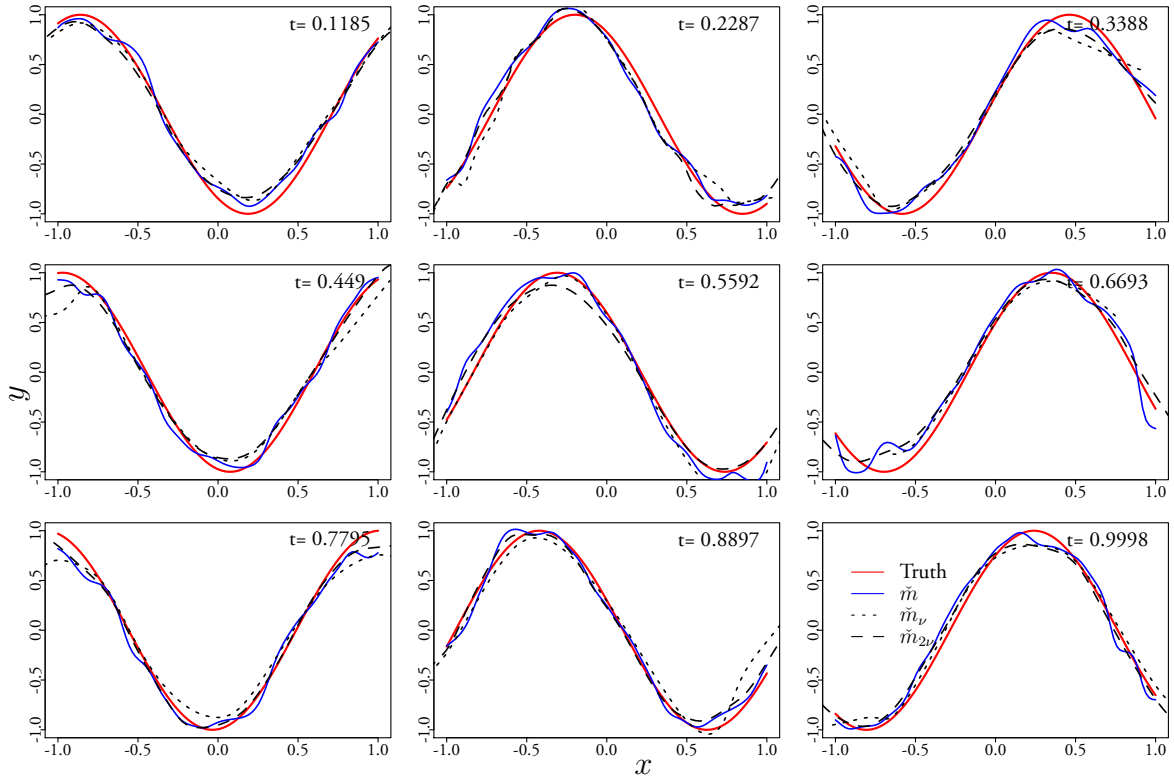

Figure 4.24: Comparison of $\check{m}$, $\hat{m}_\nu$ and $\hat{m}_{2\nu}$ for estimating model (i) when $n_1 = 1.2 \times 10^4$. The 9 subplots, corresponding to 9 equidistant time points on $[t_{N_0}, 1]$, show curves representing the true regression curve (solid line) and the estimates corresponding to $\check{m}$ (dotted dashed line), $\hat{m}_\nu$ (dotted line) and $\hat{m}_{2\nu}$ (dashed line).

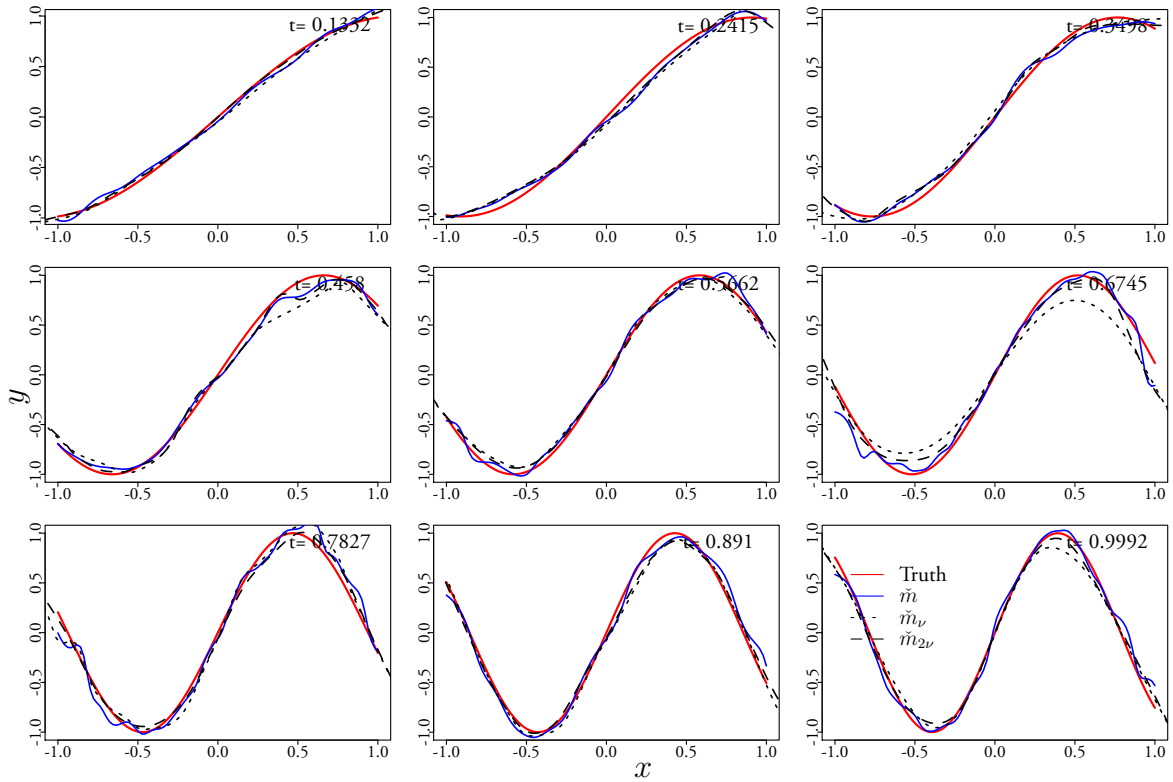

Figure 4.25: Comparison of $\check{m}$, $\hat{m}_\nu$ and $\hat{m}_{2\nu}$ for estimating model (ii) when $n_1 = 4 \times 10^3$. The 9 subplots, corresponding to 9 equidistant time points on $[t_{N_0}, 1]$, show curves representing the true regression curve (solid line) and the estimates corresponding to $\check{m}$ (dotted dashed line), $\hat{m}_\nu$ (dotted line) and $\hat{m}_{2\nu}$ (dashed line).

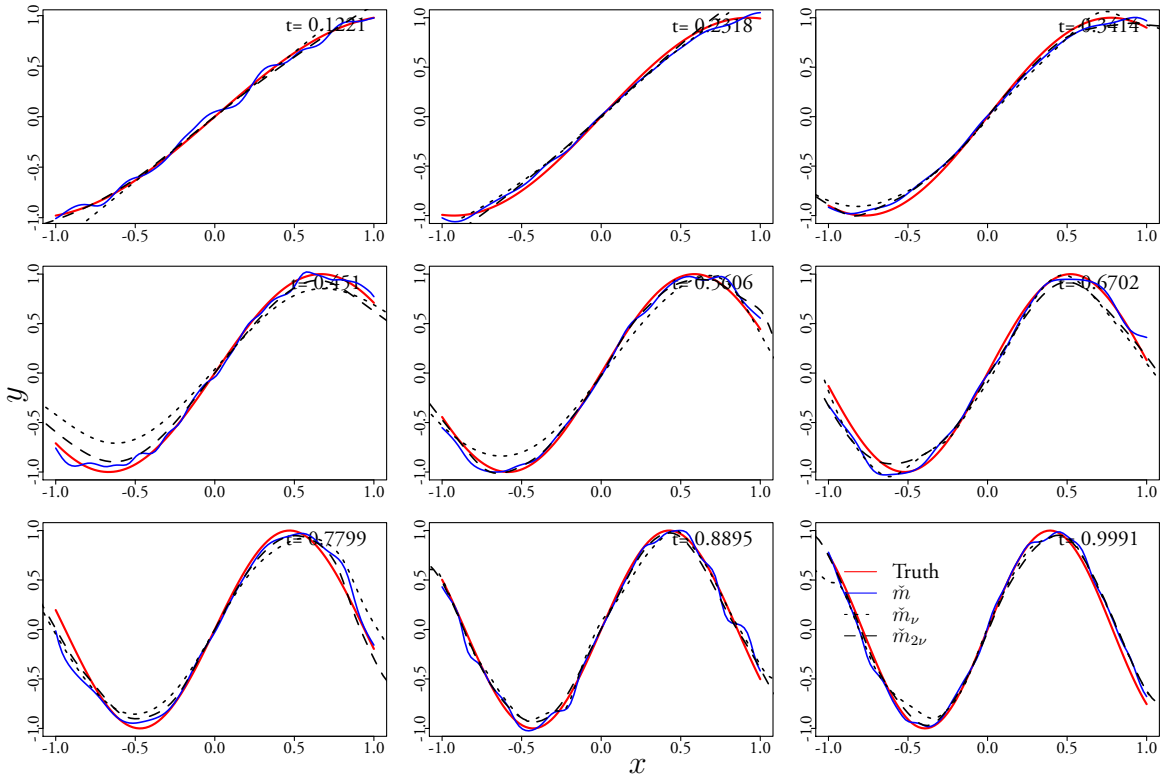

Figure 4.26: Comparison of $\check{m}$, $\hat{m}_\nu$ and $\hat{m}_{2\nu}$ for estimating model (ii) when $n_1 = 8 \times 10^3$. The 9 subplots, corresponding to 9 equidistant time points on $[t_{N_0}, 1]$, show curves representing the true regression curve (solid line) and the estimates corresponding to $\check{m}$ (dotted dashed line), $\hat{m}_\nu$ (dotted line) and $\hat{m}_{2\nu}$ (dashed line).

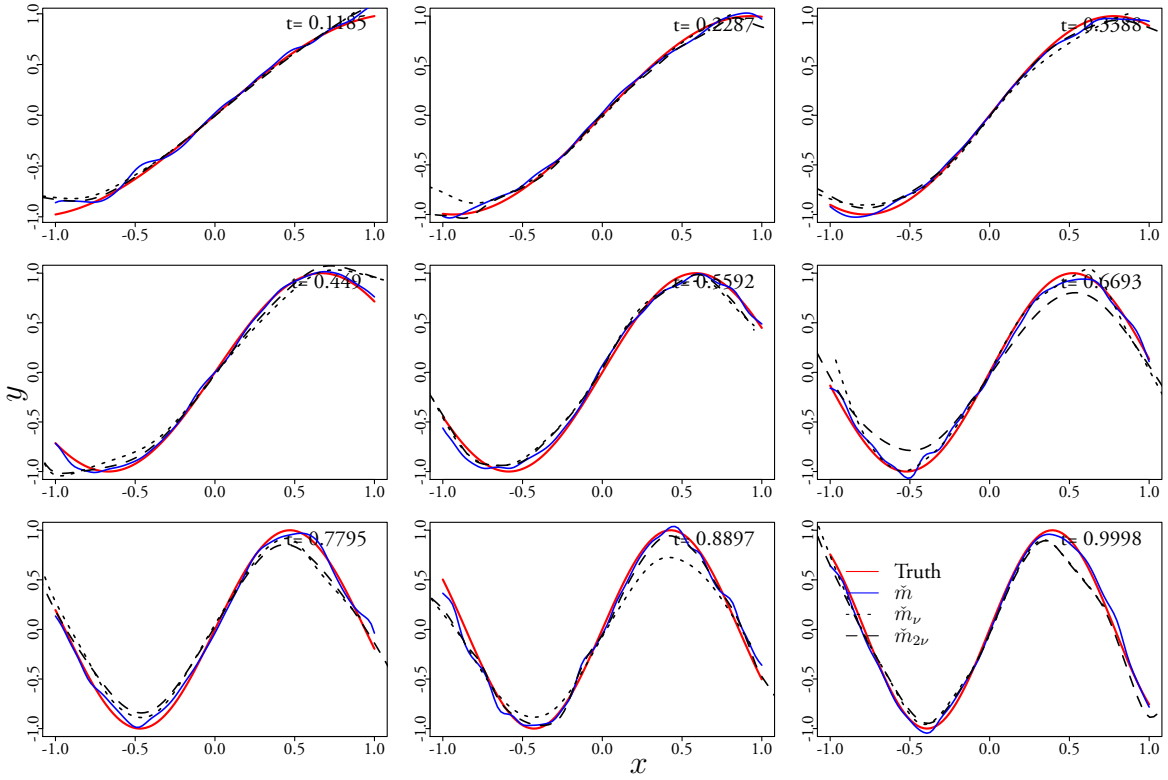

Figure 4.27: Comparison of $\check{m}$, $\hat{m}_\nu$ and $\hat{m}_{2\nu}$ for estimating model (ii) when $n_1 = 1.2 \times 10^4$. The 9 subplots, corresponding to 9 equidistant time points on $[t_{N_0}, 1]$, show curves representing the true regression curve (solid line) and the estimates corresponding to $\check{m}$ (dotted dashed line), $\hat{m}_\nu$ (dotted line) and $\hat{m}_{2\nu}$ (dashed line).

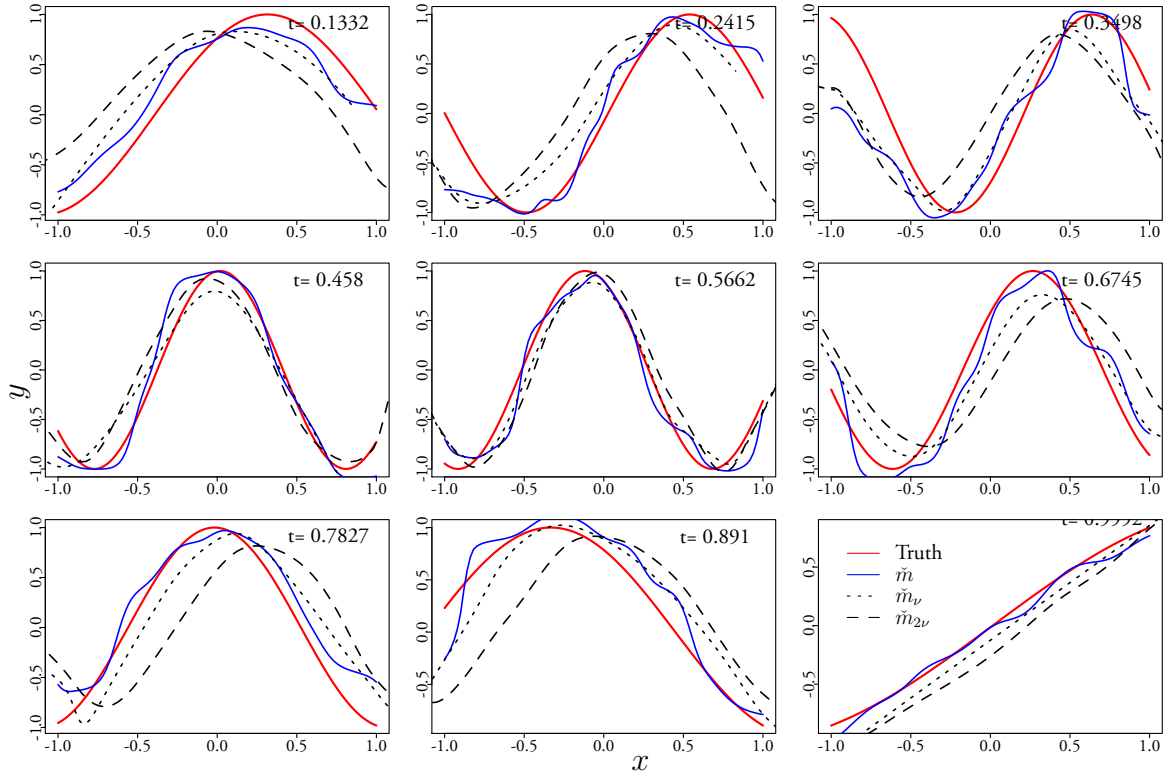

Figure 4.28: Comparison of $\check{m}$, $\hat{m}_\nu$ and $\hat{m}_{2\nu}$ for estimating model (ii) when $n_1 = 4 \times 10^3$. The 9 subplots, corresponding to 9 equidistant time points on $[t_{N_0}, 1]$, show curves representing the true regression curve (solid line) and the estimates corresponding to $\check{m}$ (dotted dashed line), $\hat{m}_\nu$ (dotted line) and $\hat{m}_{2\nu}$ (dashed line).

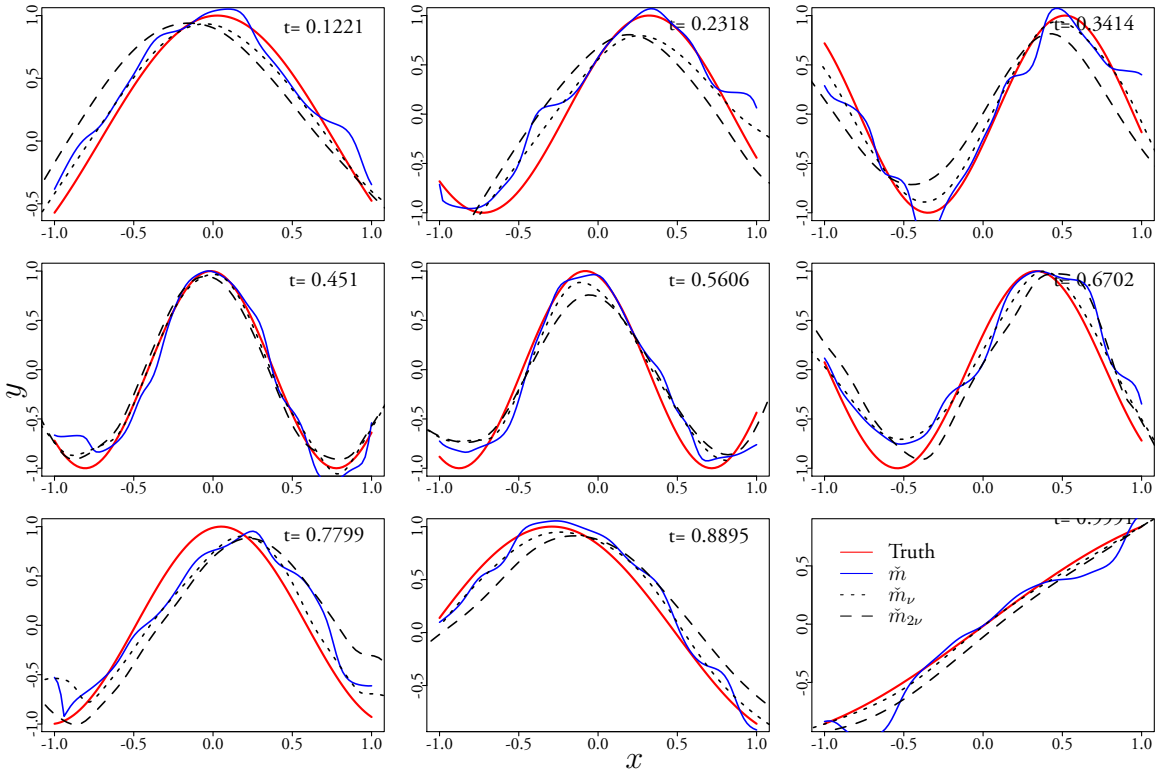

Figure 4.29: Comparison of $\check{m}$, $\hat{m}_\nu$ and $\hat{m}_{2\nu}$ for estimating model (ii) when $n_1 = 8 \times 10^3$. The 9 subplots, corresponding to 9 equidistant time points on $[t_{N_0}, 1]$, show curves representing the true regression curve (solid line) and the estimates corresponding to $\check{m}$ (dotted dashed line), $\hat{m}_\nu$ (dotted line) and $\hat{m}_{2\nu}$ (dashed line).

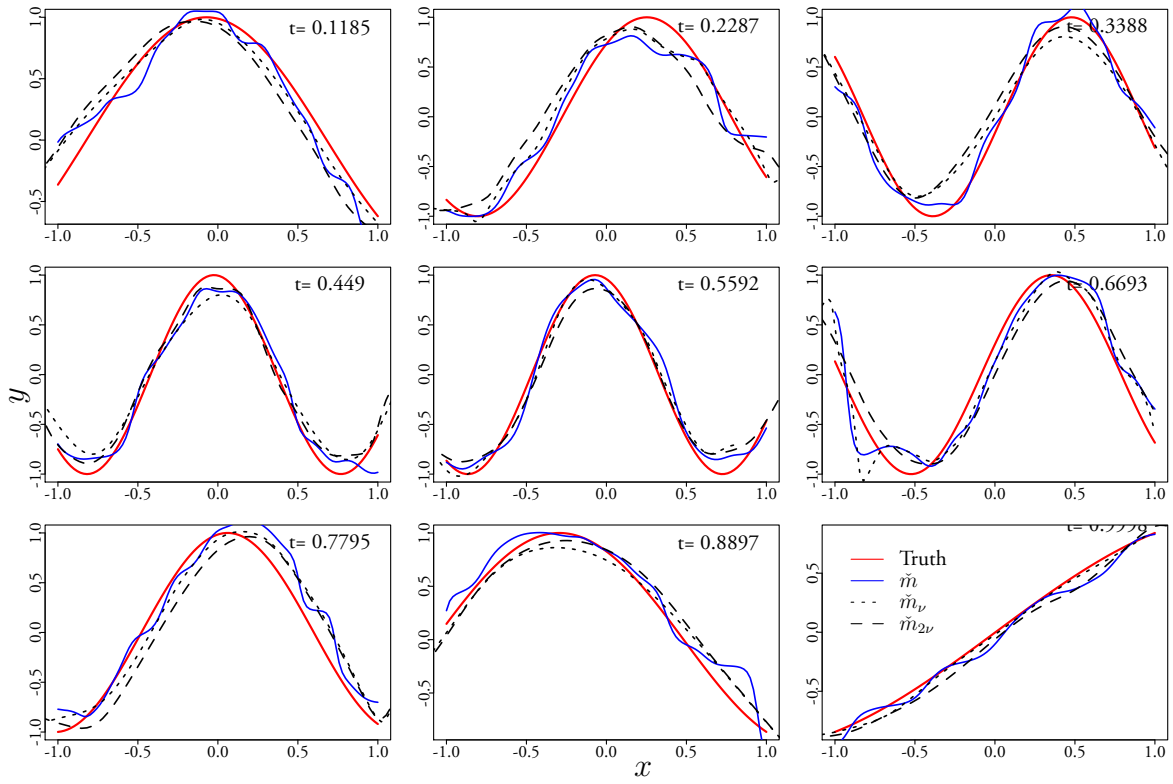

Figure 4.30: Comparison of $\check{m}$, $\hat{m}_\nu$ and $\hat{m}_{2\nu}$ for estimating model (ii) when $n_1 = 1.2 \times 10^4$. The 9 subplots, corresponding to 9 equidistant time points on $[t_{N_0}, 1]$, show curves representing the true regression curve (solid line) and the estimates corresponding to $\check{m}$ (dotted dashed line), $\hat{m}_\nu$ (dotted line) and $\hat{m}_{2\nu}$ (dashed line).

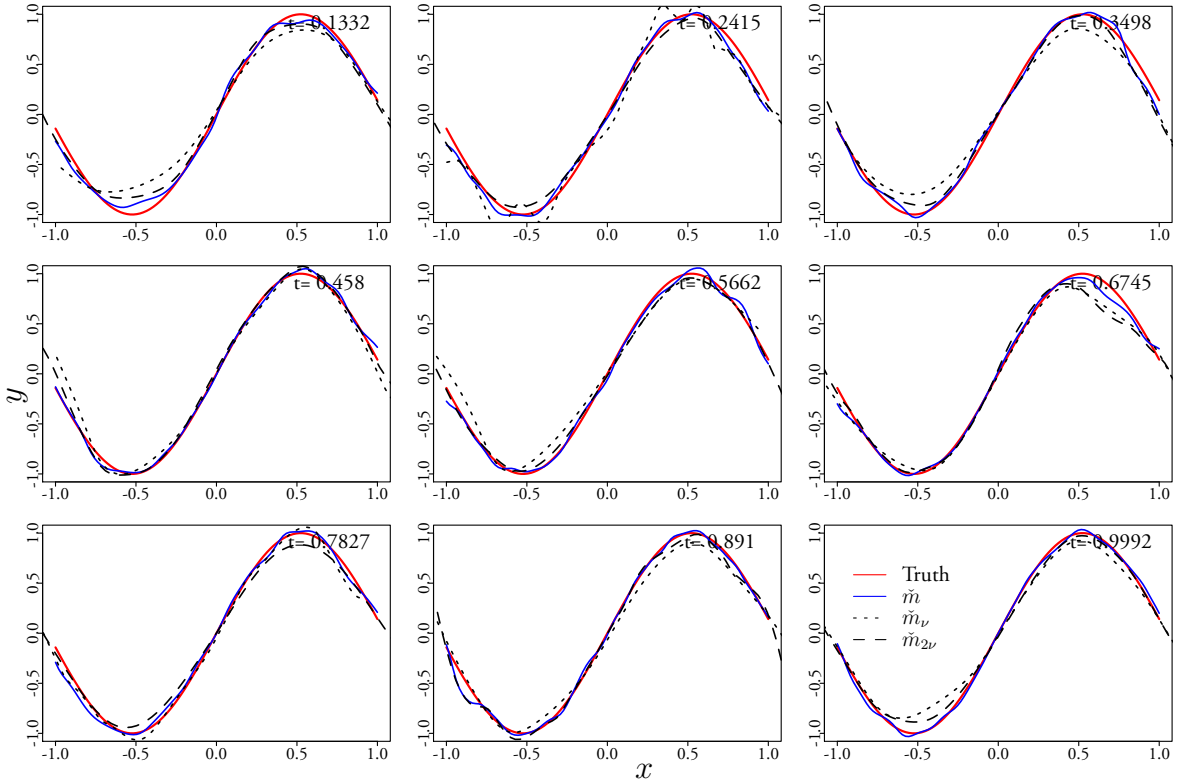

Figure 4.31: Comparison of $\check{m}$, $\hat{m}_\nu$ and $\hat{m}_{2\nu}$ for estimating model (ii) when $n_1 = 4 \times 10^3$. The 9 subplots, corresponding to 9 equidistant time points on $[t_{N_0}, 1]$, show curves representing the true regression curve (solid line) and the estimates corresponding to $\check{m}$ (dotted dashed line), $\hat{m}_\nu$ (dotted line) and $\hat{m}_{2\nu}$ (dashed line).

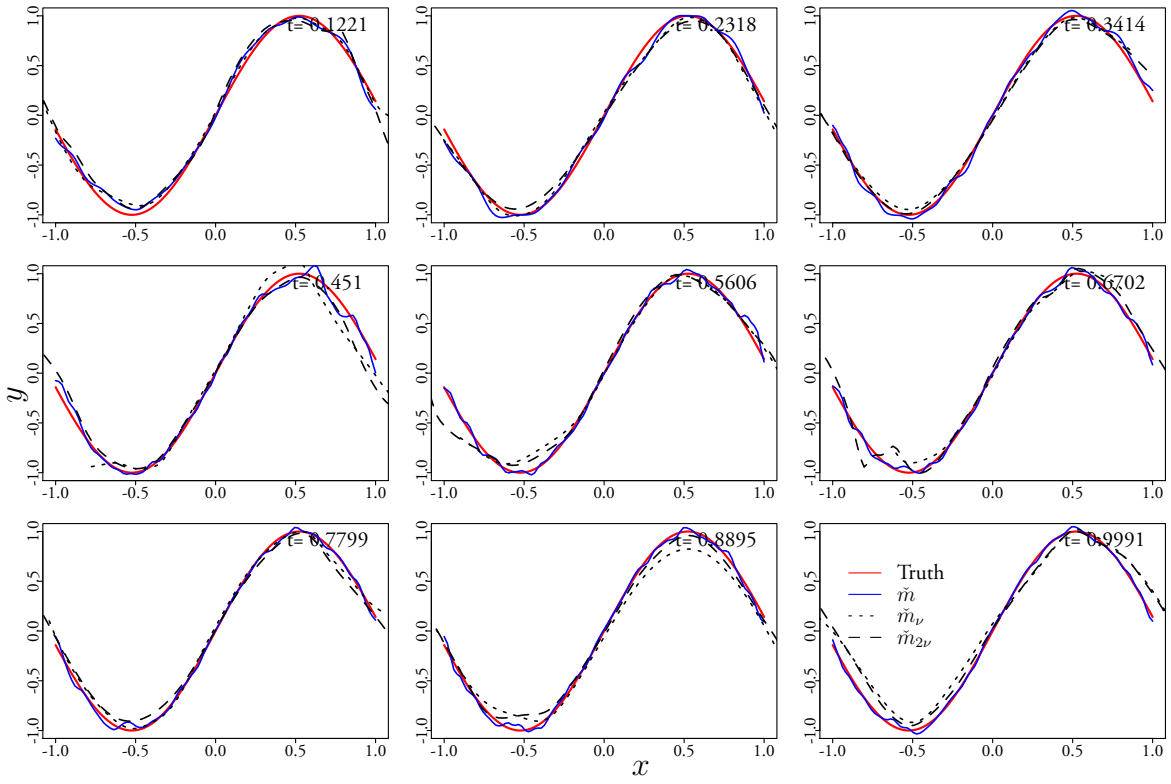

Figure 4.32: Comparison of $\check{m}$, $\hat{m}_\nu$ and $\hat{m}_{2\nu}$ for estimating model (ii) when $n_1 = 8 \times 10^3$. The 9 subplots, corresponding to 9 equidistant time points on $[t_{N_0}, 1]$, show curves representing the true regression curve (solid line) and the estimates corresponding to $\check{m}$ (dotted dashed line), $\hat{m}_\nu$ (dotted line) and $\hat{m}_{2\nu}$ (dashed line).

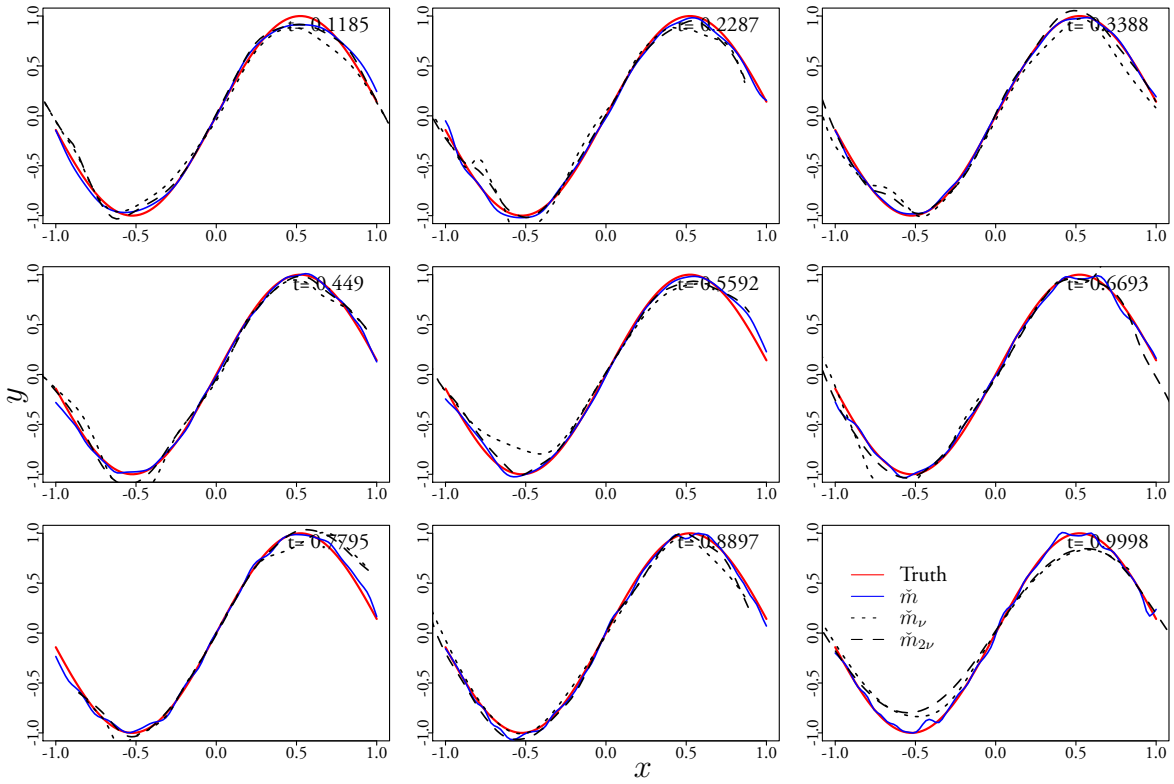

Figure 4.33: Comparison of $\check{m}$, $\hat{m}_\nu$ and $\hat{m}_{2\nu}$ for estimating model (ii) when $n_1 = 1.2 \times 10^4$. The 9 subplots, corresponding to 9 equidistant time points on $[t_{N_0}, 1]$, show curves representing the true regression curve (solid line) and the estimates corresponding to $\check{m}$ (dotted dashed line), $\hat{m}_\nu$ (dotted line) and $\hat{m}_{2\nu}$ (dashed line).

4.2 Real data examples

4.2.1 Density estimation

The Kepler mission of NASA (Thompson et al., 2016) is designed to survey a region of the Milky Way to detect Earth-size planets. This is done by observing changes in the brightness of stars in the same region of the Milky Way for 4 years between May 2009 and May 2013. One of the data types produced by the telescope is in the form of light curves, i.e. time series of the level of electromagnetic waves, such as the X-ray. Real time density estimation for light curves are potentially useful for tasks such as anomaly detection and analysis of the nature of certain astronomical events (Bloom et al., 2012; Bhatt and Bhattacharyya, 2012).

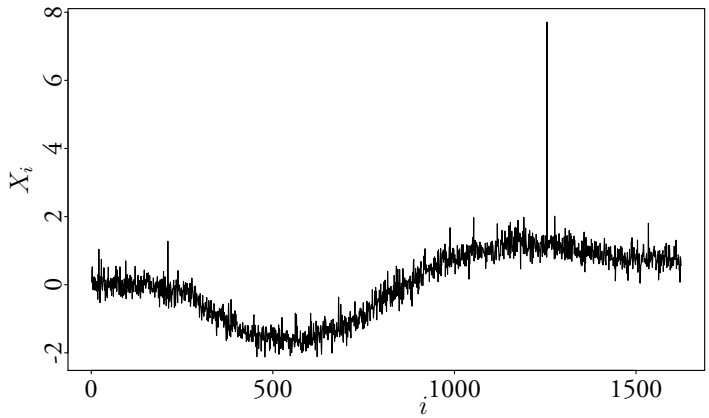

Figure 4.34: Kepler light curve `kplr001026992-2009166043257`.

Here we apply the SKDE at (2.2) with smoothing parameters selected by the SCV procedure described in §2.4 to a light curve with Kepler dataset name `kplr001026992-2009166043257`, observed from 13 May 2009 to 15 June 2009 (retrieved from `https://archive.stsci.edu/kepler/data_search/search.php`). The original dataset contains 15 missing values and 1624 measurements of the flux rate (units of electrons per second). When there is no missing value, the time gap Δt between two consecutive measurements is 30 minutes. However, since the missing values are relatively scarce, we ignore the possible effects of them on the density estimation. Figure 4.34 shows the raw data minus the sample mean 4590.44 and divided by the

sample standard deviation 14.29.

To test if the light curve data after detrending are serially correlated, we plotted the estimated auto-correlation functions of the detrended data in Figure 4.35. We detrended the data by subtracting from them the fitted values produced by a local linear regression, computed using the `locpoly` function from the R package `KernSmooth` (Wand and Jones, 2019) with bandwidth selected by the `dpill` function from the same package. Figure 4.35 shows the estimated lag 0–30 auto-correlation functions, where the blue dashed lines correspond to auto-correlation function values ± 0.05. From the plot we can see that the detrended data are not likely to be serially correlated.

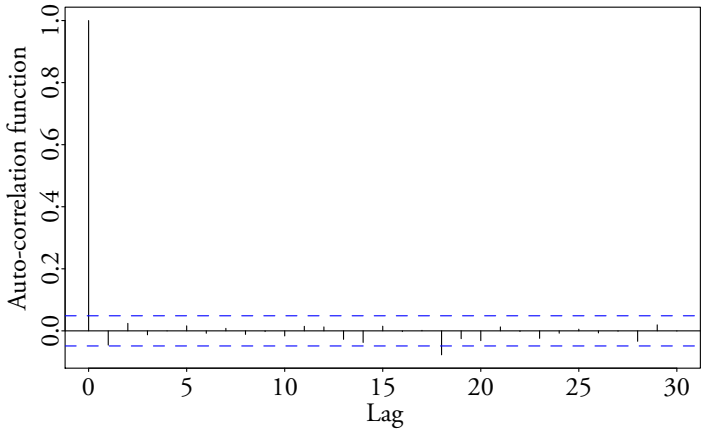

Figure 4.35: Auto-correlation function estimation for detrended Kepler light curve data.

Figure 4.36 shows the density estimates produced by $\check{f}$ and $\hat{f}_w$, for $w = 25, 50, 100, 200$, at 20 equidistant time points. The estimated density curves are sometimes fairly close to each other, indicating that the time variability at those time points may be mild (recall that estimators $\hat{f}_w$ is computed from a sliding window of size w). For other time points the estimated density curves show significant differences, indicating that the true density on those occasions may be varying fast. Overall the density estimates produced by $\check{f}$ are fairly smooth and close to those produced by $\hat{f}_{50}$.

Figure 4.37 is a mesh plot showing the density estimates produced by $\check{f}$ at 1600 time points, where the x-axis represents time points, the y-axis represents the rescaled flux rate and the z-axis represents the density value. Compared to Figure 4.34, we can see that these density estimates

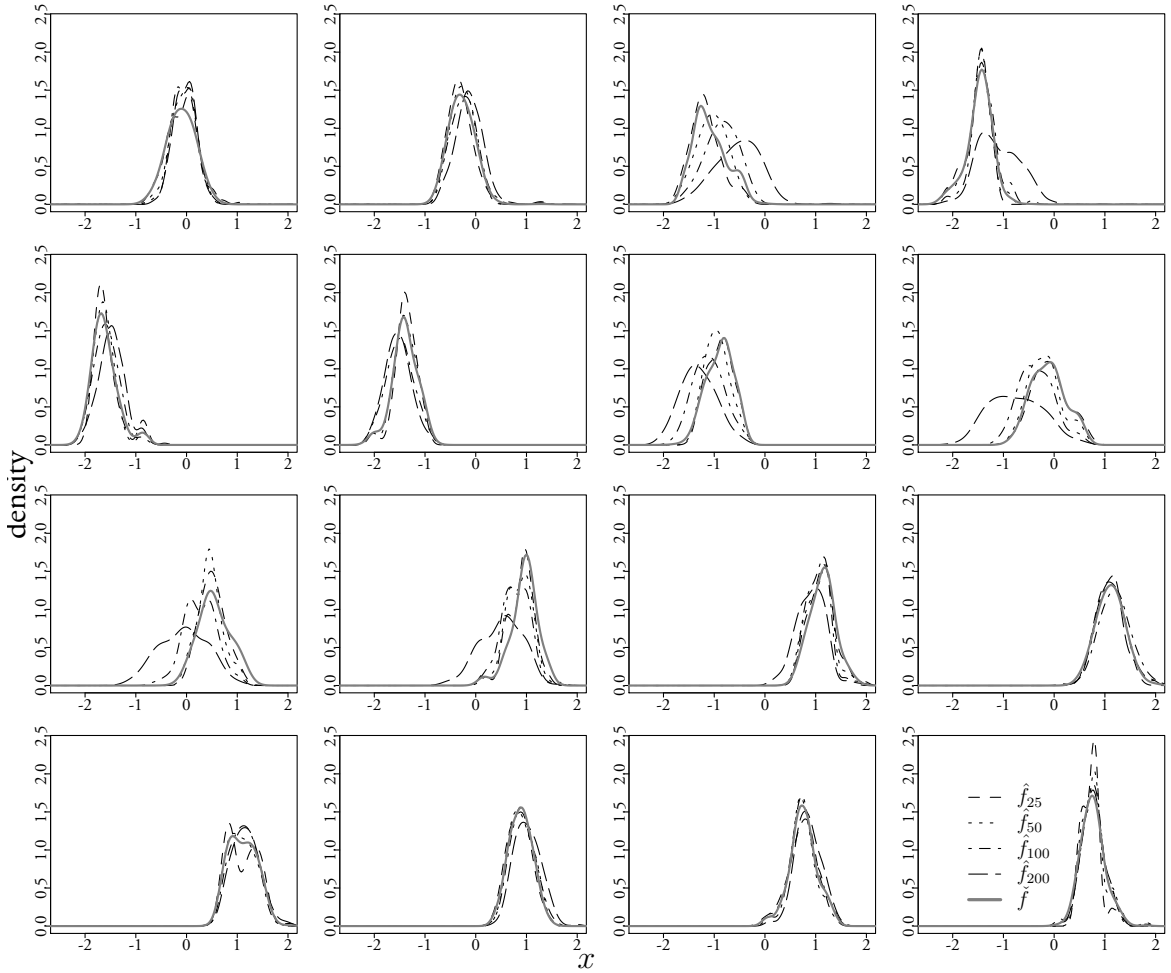

Figure 4.36: Density estimates produced by $\hat{f}_{25}$ (dashed line), $\hat{f}_{50}$ (dotted line), $\hat{f}_{100}$ (dotted dashed line), $\hat{f}_{200}$ (long dashed line) and $\check{f}$ (grey solid line), for the Kepler light curve data at 20 equidistant time points.

have captured the changes in the location of the data. In addition, changes in the scale of these density estimates may have reflected the changes in the variance of the underlying data distribution, which is difficult to observe directly from the raw data plot (Figure 4.34).

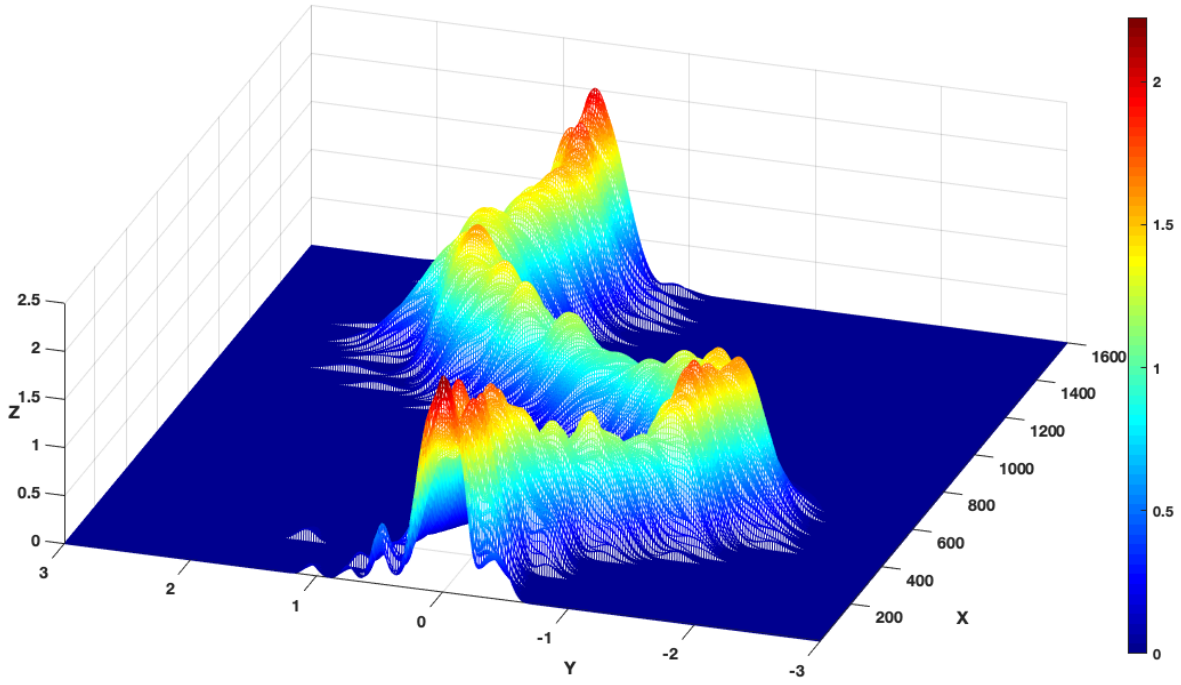

Figure 4.37: Density estimates produced by the SKDE $\check{f}$ for the Kepler light curve data as a 3D mesh plot. X-axis represents the index of data points, y-axis represents the domain of the density estimates and z-axis represents the density values.

4.2.2 Regression

We modelled the time-varying relationship between the hourly stock returns of Microsoft and Intel, using data running from 1 January 2018 to 31 August 2019 (7 data points per day). The data can be downloaded from `https://www.dukascopy.com/trading-tools/widgets/quotes/historical_data_feed`. Excluding weekends and non-working hours, the data set comprises of $2,919$ hourly observations. Let open_i and close_i denote the i-th open and close prizes of a stock, then the i-th stock return is defined as $100 \times (\text{close}_i - \text{open}_i)/\text{open}_i$, where we multiply by 100 for convenience.

Luts et al. (2014) demonstrated their real-time semiparametric regression model using stock prizes of Microsoft and Intel, without considering the possible time variability of the underlying regression function. We modelled stock returns instead by applying $\check{m}$, $\hat{m}_\nu$ and $\hat{m}_{2\nu}$ to the hourly data. To initialise the algorithm, we took $\nu = 28$ (4 days of data), so that $N_0 = 2\nu = 56$, and defined $I^1_{\gamma,h}$ and L as in §4.1.2.

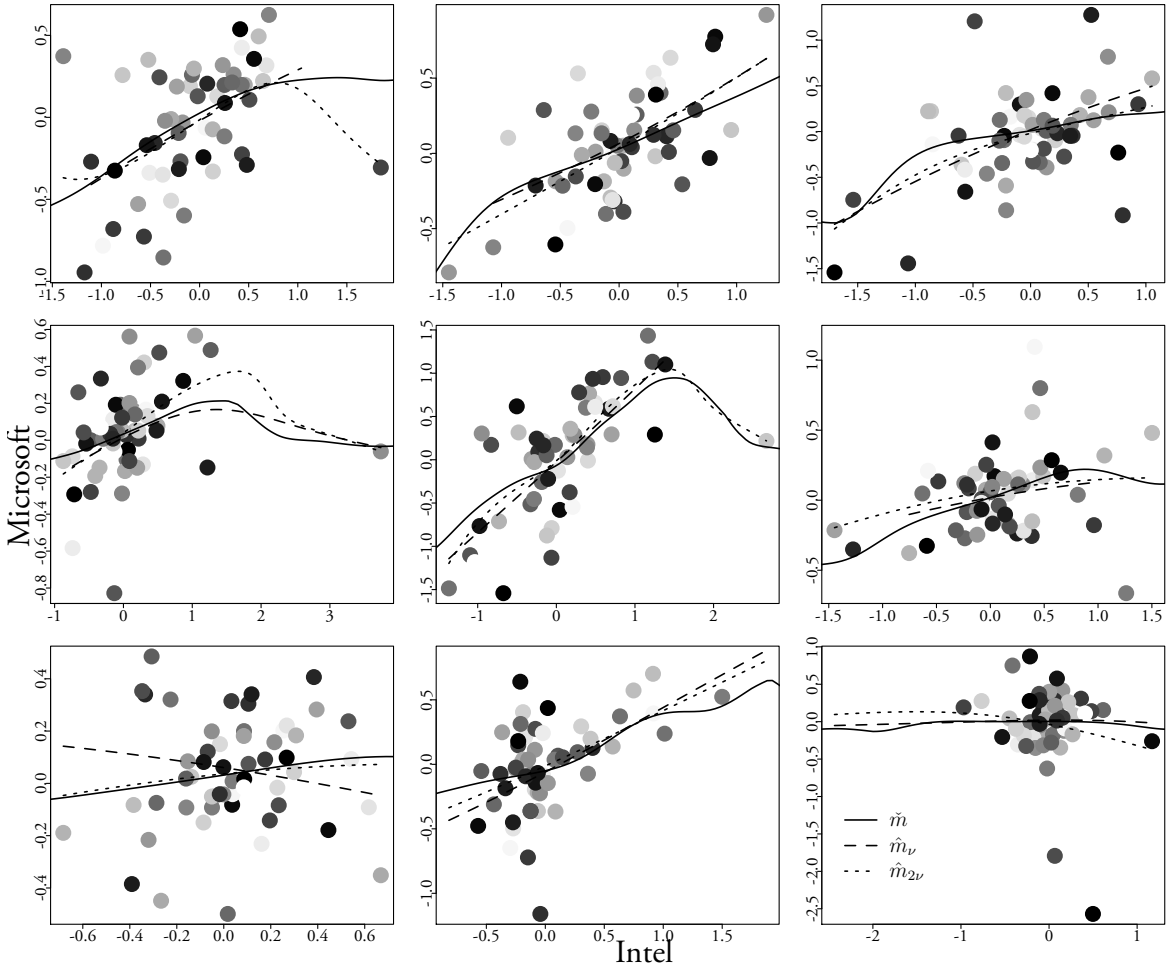

Figure 4.38: Regression curves produced by $\check{m}$ (solid line), $\hat{m}_\nu$ (dashed line) and $\hat{m}_{2\nu}$ (dotted line) for the Microsoft vs Intel stock returns data. The subplots, from left to right and top to bottom, correspond to 9 equidistant time points. The grayscale points represent the most recent 2ν observations, the darker the more recent.

Figure 4.38 shows the regression curves estimated by $\check{m}$ (solid line), $\hat{m}_\nu$ (dashed line) and $\hat{m}_{2\nu}$ (dotted line) for 9 equidistant time points. The points in each subplot show the most recent

2ν data points and their colours represent their arrival times, the darker the more recent. The figure shows that the curves produced by $\check{m}$, $\hat{m}_\nu$ and $\hat{m}_{2\nu}$ are often close together, implying that the time variability might be mild. However, the $\check{m}$ curves are often more bent towards the darker (more recent) data points. For example, in the subplot at the top right corner in Figure 4.38, the $\check{m}$ curve significantly differs from the $\hat{m}_{2\nu}$ curve towards the right boundary. This is because $\check{m}$ is less affected by the lighter (further in the past) point close to the right boundary.

Chapter 5

Conclusion and future works

5.1 Conclusion

In the era of big and complex data, statistical modelling faces new challenges arising from data sets collected in unconventional ways. Streaming data are a type of big and complex data: big, since the data arrive in high frequency and huge volume of data accumulate in relatively short period of time; complex, due to the time variabilities of the data distribution on a potentially infinite time horizon. Hence the major challenge of modelling streaming data is that any method has to be computationally efficient while being adaptive to the nonstationarity. Considering the lack of theoretically founded nonparametric estimation methods for streaming data, in this thesis we have proposed some computationally efficient and theoretically justified algorithms for nonparametric density and regression estimations for streaming data.

In Chapter 2 we proposed the SKDE for estimating nonparametrically the time-varying density of an i.n.i.d. data stream, with smoothing parameters selected by the SCV procedure. In §2.2 we proved that the SKDE is consistent when the smoothing parameters were selected appropriately. In §2.4.3 we showed the asymptotic optimality of the SCV procedure. Simulation examples in §4.1.1 illustrated that the proposed method was adaptive to various types of nonstationarity of the simulated data and was superior, in terms of estimation accuracy, to conventional offline KDE equipped with sliding windows and the plug-in bandwidth selector. In §4.2.1 we

applied the proposed method to an astronomical data stream.

In Chapter 3 we proposed the SRA for adaptively estimating the time-varying regression function of a d.n.i.d. data stream. The SRA consists of a semi-recursive NW estimator and a RCV procedure for selecting smoothing parameters. We established the asymptotic normality of the semi-recursive NW estimator in the streaming data settings. Simulation examples in §4.1.2 showed that the proposed method was adaptive to various types of nonstationarity of the simulated data and was superior to the conventional offline NW estimator equipped with the least squares cross-validation bandwidth selector, in terms of computational time and estimation accuracy. In §4.2.2 we applied the proposed method to financial and air quality data streams.

The major methodological innovation of this thesis is that it extends the conventional kernel smoothing techniques, such as the KDE, the NW regression estimator and the least squares cross-validation, which were originally proposed for offline i.i.d. samples of mild sample size, to i.n.i.d. or d.n.i.d. data streams. Other techniques have been proposed before, e.g. in the machine learning literature, but they were somewhat ad hoc and not theoretically founded. In addition, the theoretical framework we use in this thesis, featuring the infill asymptotics and some mild smoothness conditions, is flexible enough for the discussion of other modelling tasks. Considering the popularity of kernel smoothing in the machine learning community as a convenient tool for visualisation and data reduction (see e.g. Gramacki, 2018), machine learning researchers may also find these works interesting.

5.2 Future work

5.2.1 Improving convergence rates of density and regression estimators

As discussed in §2.3 and under Theorem 3.1, the SKDE $\check{f}$ and the semi-recursive NW estimator $\check{m}$ discussed in this thesis have slightly slower convergence rate compared to some existing time-varying nonparametric estimators, such as those in Hall et al. (2006), Vogt (2012) and Zhang and Wu (2015). As analysed in §2.3, this is because the way $\check{f}$ (and analogously $\check{m}$) does smoothing

over time is similar to using a one-sided positive (first-order) kernel on the temporal domain, which leads to higher bias of $\check{f}$. In §2.3, we also discussed a possible way to reduce the bias of $\check{f}$, using double-exponential smoothing (see §2.3.2). However, the resulting density estimator $\breve{f}$ has an unsatisfactory feature, namely, it cannot guarantee the positivity of the density estimate (the temporal KDE of Hall et al. (2006) also suffers from this problem, a side effect similar to that of using a higher-order kernel in the conventional KDE).

Therefore, an interesting open problem for future research is whether we can modify $\check{f}$ to reduce its bias while maintaining its positivity. This problem is somewhat similar to the boundary correction problem in the literature of nonparametric density estimation for offline i.i.d. data (see e.g. Jones, 1993b). It is well known that, for a density with bounded support, the conventional KDE using two-sided kernels suffers from higher bias close to the boundary points than in the interior of its domain. Reduction of bias (boundary correction) can be achieved by using one-sided kernels close to the boundary points. However, since a one-sided second-order kernel may take negative values, the resulting KDE with boundary correction may take negative values near the boundary points. A simple approach for boundary correction maintaining the positivity of the density estimator is discussed in Jones and Foster (1996), which is potentially useful for developing a bias-corrected version of our SKDE $\check{f}$.

While positivity of the density estimator is desirable for density estimation, it is far less of a concern in the regression problem. We can use double-exponential smoothing, without positivity correction, to construct estimators $\breve{r}(x,t)$ and $\breve{f}(x,t)$ of $r(x,t) = m(x,t)f(x,t)$ and $f(x,t)$, respectively, and then define the regression estimator as $\breve{m}(x,t) = \breve{r}(x,t)/\breve{f}(x,t)$. Motivated by the discussion in §2.3, we expect that the new estimator $\breve{m}$ will have smaller bias compared to $\check{m}$ at (3.22). This will be one of the main focuses of our future research. A related task for future research is to investigate the minimax optimal convergence rates for the density and regression estimation for streaming data.

5.2.2 Density derivative estimation for streaming data

Density derivative estimation is an important topic in the nonparametric density estimation literature. For example, it is an essential component of different versions of plug-in bandwidth selectors for the conventional KDE (see e.g. Section 3.5 of Wand and Jones, 1995). In Chapter 2 we proposed a cross-validation procedure to select smoothing parameters (γ, h) for the SKDE. Since, from the research on the conventional KDE for offline data (e.g. Park and Marron, 1990), we know that cross-validation is often less robust compared to the plug-in approach in bandwidth selection, therefore an intriguing open question is to develop plug-in smoothing parameter selectors for streaming data. From Proposition 2.1 we know that this will involve the estimation of density derivatives f_{xx} and f_t. The main challenge here is to iteratively compute these (time-varying) quantities.

Density derivative estimation also finds its application in various machine learning problems, such as mode-based clustering, mean-shift tracking and filament detection; see e.g. Chacón (2015) and Chapter 6 of Chacón and Duong (2018). Considering that clustering is one of the most important problems in streaming data analysis (see Aggarwal, 2018, for a review), it is also intriguing to investigate the connection between density derivative estimation and clustering in the streaming data setting.

5.2.3 More general assumption about error sequence in regression model

As mentioned on page 118, in future works we will relax the assumption that the error sequence $\{\epsilon_i\}_{i=1,2,\ldots}$ in the regression model (1.3) is i.i.d. and independent from the covariate sequence $\{X_i\}_{i=1,2,\ldots}$. Indeed, in a more general setting, ϵ_i may depend on X_i, as in the literature of semiparametric and nonparametric nonlinear time series analysis (Fan and Yao, 2003, Chapter 8; Gao, 2007, Chapter 1). There, since the sequence $\{X_i\}_{i=1,2,\ldots}$ is d.n.i.d., the error sequence $\{\epsilon_i\}_{i=1,2,\ldots}$ is no longer i.i.d.

One way to characterise the dependence of ϵ_i on X_i is to let

$$\epsilon_i = \varsigma(X_i, t_i)\eta_i \,, \tag{5.1}$$

where $\{\eta_i\}_{i=1,2,\ldots}$ is a sequence of i.i.d. random variables independent from $\{X_i\}_{i=1,2,\ldots}$ satisfying $\mathrm{E}(\eta_i) \equiv 0$ and $\mathrm{var}(\eta_i) \equiv 1$ and where $\varsigma(\cdot,\cdot)$ is a smooth function. This assumption is used in Vogt (2012) and Zhang and Wu (2015), where the goal is the offline estimation of model (1.3). The i.i.d. assumption of error sequence used in this thesis can be viewed as a special case of (5.1), taking $\varsigma(x,t) \equiv \sigma$. Next we discuss how dependent and heteroscedastic errors may affect the SRA described in Chapter 3.

5.2.3.1 Error dependence

It is well documented in the literature on nonparametric regression that dependent errors may have a negative impact on bandwidth selection procedures such as the leave-one-out cross-validation (Zhou et al., 2003; Altman, 1990; Chu and Marron, 1991; Hart, 1991, 1994; Hall, Lahiri and Polzehl, 1995; Opsomer et al., 2001; Hall and Van Keilegom, 2003; De Brabanter et al., 2018). However, as pointed out by Yao and Tong (1998) and Fan and Yao (2003, p. 273), this negative impact is more prominent for fixed design regression models where the design points are time points. For random design models, such as (1.3), where the errors are serially dependent, i.e. the dependence between two error terms ϵ_i and ϵ_j decreases as their observational times t_i and t_j are further apart, the bandwidth selection is very similar to that for independent data, when the dependence is weak enough[1].

To understand why serially dependent errors have a much smaller effect on random design regression models, first consider the following fixed design model:

$$Y_i = m(t_i) + \epsilon_i \,,\ i = 1, \ldots, n \,, \tag{5.2}$$

[1]Here we only consider serial dependence of errors. Opsomer et al. (2001) and De Brabanter et al. (2018) considered spatially dependent errors, where the dependence between ϵ_i and ϵ_j decreases as the covariates X_i and X_j are further apart. In the latter case, the bandwidth selection for random design regression models will also suffer from the dependence of errors, as that for the fixed design models.

where $t_i = i/n$ is the observational time of Y_i and $\{\epsilon_i\}_{i=1,\ldots,n}$ is a stationary error sequence satisfying $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{var}(\epsilon_i) = \sigma^2$ for some $\sigma \in \mathbb{R}_+$. That is, the time series $\{Y_i\}_{i=1,\ldots,n}$ is observed on a equidistant time grid on $[0, 1]$ and it is composed by a deterministic trend $m(\cdot)$ plus errors. Suppose the error sequence admits the following serial dependence structure:

$$\mathrm{corr}(\epsilon_i, \epsilon_j) = \rho(|t_i - t_j|)\,, \tag{5.3}$$

where $\mathrm{corr}(\cdot, \cdot)$ denotes correlation and $\rho : \mathbb{R}_+ \to [0, 1]$ is a monotone decreasing function. That is, the errors are positively correlated (positively correlated errors are much more common in practice, see, e.g., Hart, 1991) and the correlation of two error terms ϵ_i, ϵ_j decreases as their observational times t_i, t_j are further apart.

If we conduct leave-one-out cross-validation for model (5.2), then, due to due to (5.3), the leave-out data point (t_i, Y_i) will be positively correlated with its neighbours (t_{i-1}, Y_{i-1}) and (t_{i+1}, Y_{i+1}), since $|t_i - t_{i-1}|$ and $|t_i - t_{i+1}|$ are small. As a result, the cross-validation tend to select an overly small bandwidth, causing undersmoothing. However, this problem will be less significant for a random design model

$$Y_i = m(X_i) + \epsilon_i\,,\ i = 1, \ldots, n\,, \tag{5.4}$$

where $\{(X_i, Y_i)\}_{i=1,\ldots,n}$ is a weakly dependent random process and $\{\epsilon_i\}_{i=1,\ldots,n}$ satisfies (5.3). This is because, in model (5.4), the nearest neighbour (X_{j_i}, Y_{j_i}) of the leave-out data point (X_i, Y_i), may be only weakly correlated with (X_i, Y_i), since $|t_{j_i} - t_i|$ may be large.

To summarise, random design regression models suffer much less from serially dependent errors, since the nearest neighbours in the state space (space of the regressor) may not be close in the time space. Hence Yao and Tong (1998) suggests that leave-one-out cross-validation may still be applicable for these models. Analogously, the RCV procedure at (3.26) may still be applicable when the errors are dependent, since the one-step-ahead data point (X_{i+1}, Y_{i+1}), although close to (X_i, Y_i) in the time space, may be far away from it in the state space.

5.2.3.2 Heteroscedasticity

Jones (1993a) argues that nonparametric regression estimators often suffer much less from heteroscedasticity compared to linear regression model, so that we can afford to ignore heteroscedasticity when the goal is simply to estimate the regression function nonparametrically. However, Müller and Stadtmüller (1987) show that taking heteroscedasticity into account can improve the performance of the kernel regression estimator, in terms of the MISE. They propose to estimate the local error variance and to use variable bandwidth. That is, we use larger bandwidth (hence more smoothing) where the error variance is larger, since higher error variance implies noisier data. However, in the streaming data setting, such a procedure, involving estimation of the error variance and selection of local bandwidth, may be computationally heavy.

Hence, to apply the SRA under the more general condition (5.1) on the error sequence, we may safely ignore the heteroscedasticity when it is mild. Otherwise, we may estimate the function $\sigma(\cdot, \cdot)$ in (5.1) and use the variance-stabilised data $\{(X_i, Y_i/\hat{\sigma}(X_i, t_i))\}_{i=1,2,\dots}$ instead. Estimating time-varying local variance is a challenging task by itself and we shall leave it for future work.

5.2.4 Higher dimensions

For the convenience of theoretical derivations, in this thesis we have focused on univariate streaming data. In this section we discuss possible ways of extending our density and regression estimation methods to higher (than 1) dimensions. We consider two types of higher dimensionality, namely, the low-dimensional case (e.g. 2 or 3) and the high-dimensional case (e.g. 100).

5.2.4.1 Low-dimensional case

Low-dimensional kernel smoothing is useful for visualisation. For example, a bivariate KDE can be used to model spatial phenomena (Hall et al., 2006) or for object tracking (Qahtan et al., 2017). Let $\{X_i\}_{i=1,2,\dots}$ denote a data stream with arrival times $\{t_i\}_{i=1,2,\dots}$ defined at (1.1), where each $X_i = (X_{i1}, \dots, X_{id})$ is a d-dimensional random vector with density $f(\cdot, t_i)$, for some small

integer $d > 1$ (e.g. $d = 2$). To estimate the time-varying density f, we can still use the recursive KDE at (1.8) with a d-dimensional kernel $K : \mathbb{R}^d \to \mathbb{R}$ satisfying $\int_{\mathbb{R}^d} K = 1$ and $K_h(\cdot) = h^{-d}K(\cdot/h)$. Then we can use a cross-validation procedure to select γ and h, similar to the SCV in §2.4. However, the relation (2.28) needs to be modified based on a multivariate version of Theorem 2.1. Similarly, by using a multivariate kernel K, we can modify the semi-recursive NW estimator at (1.9) to estimate a time-varying regression function m with multivariate covariate.

5.2.4.2 High-dimensional case

Regression It is well-known that the kernel smoothing estimators suffer greatly from the curse of dimensionality. Specifically, for i.i.d. data, the order of their convergence rate with respect to, e.g. the MISE, decreases fast as dimension increases. See, for example, Chapter 4 and Section 5.9 of Wand and Jones (1995). A possible remedy for the curse of dimensionality for kernel regression is the generalised additive models (GAM).

Let $\{(X_i, Y_i)\}_{i=1,2,\ldots}$ denote a data stream with arrival times $\{t_i\}_{i=1,2,\ldots}$ defined at (1.1), where each $X_i = (X_{i1}, \ldots, X_{id})$ is a d-dimensional random vector, for some large integer $d > 1$ (e.g. $d = 100$). Analogous to the GAM for offline i.i.d. data (see, e.g., Chapter 7 of Fan and Gijbels, 1996), we may assume the following time-varying GAM

$$Y_i = m_0(t_i) + \sum_{j=1}^{d} m_j(X_{ij}, t_i) + \epsilon_i, \tag{5.5}$$

where m_0 is the time-varying intercept term, m_j is a time-varying function for the j-th dimension and ϵ_i is the error term. Model (5.5) has been investigated by Vogt (2012) for offline d.n.i.d. data, where each m_j is estimated by an estimator defined at (1.14) with a symmetric temporal kernel K_T. The model fitting is done by an adaptation of the smooth backfitting technique proposed by Mammen et al. (1999). For streaming data, we may apply the semi-recursive NW estimator defined at (1.9). However, developing a computationally-efficient model-fitting algorithm for this 'streaming GAM' remains an open problem.

Density estimation The nonparametric estimation of high-dimensional densities for i.i.d. data has received attention of some recent statistical works. For example, Nagler and Czado (2016) assume a simplified vine copula model to evade the curse of dimensionality. That is, they assume that the multivariate density to be estimated has a special dependence structure among the marginal densities, so that to estimate the original high-dimensional density we only need to estimate all the marginal densities and all their pairwise joint densities. Then the convergence rate of their density estimator has the same order as the estimator of the pairwise (bivariate) joint densities. That is, the convergence rate is independent of the dimensionality. However, a drawback of their estimator is the computation – for a d-dimensional density their estimator requires estimating d marginal densities and $d(d-1)/2$ bivariate joint densities. This may prevent the use of their estimator for high-dimensional streaming data.

Xu and Samworth (2019) propose a nonparametric method which evades the curse of dimensionality via symmetry and shape constraints. They assume that the density to be estimated belongs to a class of log-concave densities on $\mathbb{R}^d$ satisfying some symmetry conditions. The main symmetry condition is that the super-level sets of the density f (sets of the form $\{x \in \mathbb{R}^d : f(x) \geq c\}$ with some $c > 0$) are scalar products of some convex set $\mathcal{K}$. In particular, when $\mathcal{K}$ is known, their estimator computed from n i.i.d. data has a worst-case risk bound with respect to, e.g., square Hellinger loss, of $O(n^{-4/5})$, independent of d. The drawback of their method is twofold. Firstly, in practice $\mathcal{K}$ is often unknown and hence needs to be estimated from data. In this case the risk bound of their density estimator is no longer independent from d, so that the curse of dimensionality will still be present. Secondly, the shape constraints that their method relies on may be somewhat stringent in some cases, since they require, e.g., that the data-generating density has to be unimodal and its super-level sets have to have the same convex shape.

Due to their heavy computation or rather stringent shape constraints, the above-mentioned two methods for high-dimensional density estimation cannot be directly applied to streaming data. Furthermore, the potential nonstationarity of high-dimensional streaming data poses some additional challenge to density estimation besides the curse of dimensionality. However, before

discussing possible approaches for countering this double challenge, we first ask the question: 'what is the purpose of high-dimensional density estimation?' This is an important question since it will directly affect the appropriate loss function to use for the estimation problem. We shall discuss this question and its implications for our future research in more details in §5.2.5.

5.2.5 Density-based classification

Unlike the low-dimensional (e.g. $d = 2$) case, high-dimensional density estimators are not directly useful for, say, visualisation. A more straightforward application of those estimators is density-based classification. For example, Nagler and Czado (2016) applied their copula-based high-dimensional density estimator to the classification of some astrophysical imaging data. In this section we briefly review the density-based approach for the binary classification of offline i.i.d. data. Then we shall discuss their implications for constructing classifiers for high-dimensional streaming data.

Let $\mathcal{Z} = \{(X_i, Y_i)\}_{i=1,\dots,n}$ denote a sample with n i.i.d. observations, where $X_i = (X_{i1}, \dots, X_{id})$ denotes a random vector called the feature vector and where $Y_i \in \{0, 1\}$ denotes a Bernoulli random variable called the (binary) class label. Let (X, Y) denote a new random vector that is i.i.d. as the data in $\mathcal{Z}$. Then the goal of binary classification is to compute from $\mathcal{Z}$ a classifier $C_{\mathcal{Z}} : \mathbb{R}^d \to \{0, 1\}$ which accurately predicts the class label $\widehat{Y} = C_{\mathcal{Z}}(X)$ given the feature vector X.

Mathematically, we seek to minimise the following misclassification risk given $X = x$:

$$\begin{aligned}\mathcal{R}(x) &= \mathrm{P}\{C_{\mathcal{Z}}(x) \neq y\} \\ &= \mathrm{P}\{C_{\mathcal{Z}}(x) = 0 | X = x\}\mathbb{1}(Y = 1) + \mathrm{P}\{C_{\mathcal{Z}}(x) = 1 | X = x\}\mathbb{1}(Y = 0), \end{aligned} \tag{5.6}$$

where y is the class label for feature vector $x \in \mathbb{R}^d$. This risk is minimised by the so-called Bayes rule

$$C(x) = \mathbb{1}\{\mathrm{P}(Y = 1 | X = x) \geq 1/2\}. \tag{5.7}$$

See, e.g., Friedman (1997) for more detailed discussions of binary classification and the misclassification risk.

The Bayes rule $C(x)$ is an ideal classifier but it requires us knowing $\mathrm{P}(Y = 1|X = x)$. In practice, of course, we can only estimate this probability from the sample $\mathcal{Z}$. The density-based classification seeks to do this in a straightforward way. Let $f_0 = f_{X|Y=0}$ and $f_1 = f_{X|Y=1}$ denote the conditional densities of feature vector X given its class label Y. By the Bayes theorem, we have

$$\mathrm{P}(Y = 1|X = x) = \frac{\pi_1 f_1(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)}, \tag{5.8}$$

where $\pi_0 = \mathrm{P}(Y = 0)$ and $\pi_1 = \mathrm{P}(Y = 1)$ are the unconditional probabilities of each class. From (5.7) and (5.8), a density-based classifier has the form

$$C_{\mathcal{Z}}(x) = \hat{C}(x) = \mathbb{1}\left\{\frac{\pi_1 \hat{f}_1(x)}{\pi_0 \hat{f}_0(x) + \pi_1 \hat{f}_1(x)} \geq \frac{1}{2}\right\}, \tag{5.9}$$

where $\hat{f}_0$ and $\hat{f}_1$ are some density estimates. Comparing (5.9), (5.7) and (5.8), $C_{\mathcal{Z}}(x)$ at (5.9) can be viewed as a plug-in estimator of $C(x)$.

Depending on how we estimate f_0 and f_1, (5.9) induces various classifiers. For example, suppose f_0 and f_1 are multivariate normal densities of $N(\mu_0, \Sigma_0)$ and $N(\mu_1, \Sigma_1)$, where μ_0, μ_1 are mean vectors and Σ_0, Σ_1 covariance matrices. If we assume $\Sigma_0 = \Sigma_1$, then (5.9) leads to the linear discriminant analysis (LDA); if we allow $\Sigma_0 \neq \Sigma_1$, then (5.9) leads to the quadratic discriminant analysis (QDA)[2]. See, e.g., Section 4.3 of Hastie et al. (2009) for details.

A prominent feature of LDA (QDA) is that the decision boundary it generates is linear (quadratic). In contrast, naive Bayes, another density-based classification method, generates more general decision boundaries. It is called 'naive' since it makes the following simplistic

[2]R. A. Fisher, who proposed LDA in Fisher (1936), did not derive LDA from the normality assumptions, but from an analysis of variance (ANOVA) viewpoint. This partially explains why LDA and QDA still work reasonably well even when the data are far from normally distributed. See (Hastie et al., 2009, p. 111) for a brief discussion on the popularity of these methods and the role of the normality assumptions.

assumptions: for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$,

$$f_i(x) = \prod_{k=1}^{d} f_{ik}(x_k), \text{ for } i = 1, 2, \tag{5.10}$$

where f_{ik} denotes the density of the k-th margin of the feature vector X. That is, naive Bayes assumes that the different margins of X are conditionally independent given Y, which is often untrue in practice. Then, the marginal density f_{ik} can be modelled either parametrically or nonparametrically.

Despite its sheer 'naivety', naive Bayes proves to be one of the most efficient and effective classification methods in practice (Zhang, 2004). Its competitive performance even extends to the case when the dimension d of feature vector X is high (Hastie et al., 2009, pp. 210–211). Bickel and Levina (2004) compare the asymptotic behaviours of LDA and parametric naive Bayes (marginal densities assumed to be normal) in binary classification. They assume that the two conditional densities f_0 and f_1 are normal and the dimension d grows faster than sample size n as $n \to \infty$. They find that naive Bayes classifier greatly outperforms LDA (even when the normality assumptions for LDA are satisfied). One of the intuitions to the success of naive Bayes is that the bias induced by its simplistic assumptions often does not harm the estimation of (5.8), especially for x values close to the decision boundary (Hastie et al., 2009, pp. 210–211). Hence a bad density estimator (in terms of pointwise behaviour measured by, e.g., the MISE) may still lead to a good classifier (in terms of the classification risk $\mathcal{R}$ at (5.6)).

In addition to its competitive performance both in theory and in practice, another attractive feature of naive Bayes is its computational efficiency. Recall that the copula-based high-dimensional density estimator proposed by Nagler and Czado (2016) is computationally heavy since it requires modelling the complex dependence structure among the marginal densities of the feature vector X. Naive Bayes is free from this burden by assuming the conditional independence of the margins. This is especially attractive for the classification of streaming data.

In our future research, we aim at developing some computationally efficient and theoretically justified algorithm, using the naive Bayes approach, for the classification of streaming data. For

the estimation of the time-varying conditional densities, we may apply the SKDE described in Chapter 2. However, due to the change in the goal from density estimation to classification, the SCV procedure proposed in §2.4 may no longer be appropriate for the selection of smoothing parameters. Hence a key component of our future research is to develop an algorithm for the selection of smoothing parameters for our streaming naive Bayes classifier.

References

Adams, N. M., Tasoulis, D. K., Anagnostopoulos, C. and Hand, D. J. (2010). Temporally-adaptive linear classification for handling population drift in credit scoring. In *Proceedings of COMPSTAT 2010*, Ed. by Y. Lechevallier and G. Saporta, 97–106.

Aggarwal, C. C. (2018). A survey of stream clustering algorithms. In C. C. Aggarwal and C. K. Reddy (Ed.) *Data Clustering (pp. 231–258)*. Chapman and Hall/CRC.

Ahmad, I. A. and Lin, P. E. (1976). Nonparametric sequential estimation of a multiple regression function. *Bull. Math. Statist.*, **17**, 63–75.

Altman, N. S. (1990). Kernel smoothing of data with correlated errors. *J. Amer. Statist. Assoc.*, **85**, 749–759.

Amiri, A. (2009). Sur une famille paramétrique d'estimateurs séquentiels de la densité pour un processus fortement mélangeant. *C. R. Acad. Sci. Paris, Sér. I*, **347**, 309–314.

Amiri, A. (2012). Recursive regression estimators with application to nonparametric prediction. *J. Nonparametr. Stat.*, **24**, 169–186.

Anagnostopoulos, C. (2010). A statistical framework for streaming data analysis. *PhD thesis*. Cambridge University

Anagnostopoulos, C., Tasoulis, D. K., Adams, N. M. and Hand, D. J. (2009). Temporally adaptive estimation of logistic classifiers on data streams. *Adv. Data Anal. Classif.*, **3**, 243–261.

Anagnostopoulos, C., Tasoulis, D. K., Adams, N. M., Pavlidis, N. G. and Hand, D. J. (2012). Online linear and quadratic discriminant analysis with adaptive forgetting for streaming classification. *Stat. Anal. Data Min.*, **5**, 139–166.

Bedi, A. S., Koppel, A., Rajawat, K. and Sadler, B. M. (2019). Nonstationary nonparametric online learning: Balancing dynamic regret and model parsimony. *arXiv:1909.05442*.

Bertini Jr., J. R. and Nicoletti, M. d. C. (2019). An iterative boosting-based ensemble for streaming data classification. *Inf. Fusion*, **45**, 66–78.

Bhatt, N. and Bhattacharyya, S. (2012). Time evolution of the probability density function of a gamma-ray burst: A possible indication of the turbulent origin of gamma-ray bursts. *Mon. Not. R. Astron. Soc.*, **420**, 1706–1713.

Bickel, P. J. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, **10**, 989–1010.

Bloom, J., Richards, J., Nugent, P., Quimby, R., Kasliwal, M., Starr, D., Poznanski, D., Ofek, E., Cenko, S., Butler, N., Kulkarni, S., Gal-Yam, A. and Law, N. (2012). Automating discovery and classification of transients and variable stars in the synoptic survey era. *Publ. Astron. Soc. Pac.*, **124**, 1175–1196.

Bodenham, D. A. and Adams, N. M. (2017). Continuous monitoring for changepoints in data streams using adaptive estimation. *Stat. Comput.*, **27**, 1257–1270.

Boedihardjo, A. P., Lu, C. and Chen, F. (2008). A framework for estimating complex probability density structures in data streams. In *Proc. 17th ACM Conf. Inf. Knowl. Manage.*, **27**, 619–628.

Bosq, D. (1998). *Nonparametric statistics for stochastic processes: Estimation and prediction*. Springer.

Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Bradley, R. C. (2005). Basic properties of strong mixing conditions: A survey and some open questions. *Probab. Surv.*, **2**, 107–144.

Bradley, R. C. and Tone, C. (2017). A central limit theorem for non-stationary strongly mixing random fields. *J. Theor. Probab.*, **30**, 655–674.

Breiman, L. (2001). Random forests. *Mach. Learn.*, **45**, 5–32.

Cabrera, J. L. O. (2018). locpol: Kernel Local Polynomial Regression. *R package version 0.7-0*.

Cao, Y., He, H., and Man, H. (2012). SOMKE: Kernel density estimation over data streams by sequences of self-organizing maps. *IEEE T. Neural Netw. Learn. Syst.*, **23**, 1254–1268.

Caudle, K. A. and Wegman, E. (2009). Nonparametric density estimation of streaming data using orthogonal series. *Comput. Stat. Data Anal.*, **53**, 3980–3986.

Chacón, J. E. (2015). A population background for nonparametric density-based clustering. *Stat. Sci.*, **30**, 518–532.

Chacón, J. E. and Duong, T. (2018). *Multivariate kernel smoothing and its applications*. Chapman & Hall, Boca Raton, USA.

Chan, H. P. (2017). Optimal sequential detection in multi-stream data. *Ann. Statist.*, **45**, 2736–2763.

Chen, H. (2019). Sequential change-point detection based on nearest neighbors. *Ann. Statist.*, **47**, 1381–1407.

Chu, C.-K. and Marron, J. S. (1991). Comparison of two bandwidth selectors with dependent errors. *Ann. Statist.*, **19**, 1906–1918.

Chung, K. L. (2000). *A course in probability*. 3rd Ed., Academic Press.

Davies, H. L. (1973). Strong consistency of a sequential estimator of a probability density function. *Bull. Math. Statist.*, **15**, 49–54.

De Brabanter, K., Cao, F., Gijbels, I. and Opsomer, J. (2004). Local polynomial regression with correlated errors in random design and unknown correlation structure. *Biometrika*, **105**, 681–690.

Delaigle, A. and Gijbels, I. (2004). Bootstrap bandwidth selection in kernel density estimation from a contaminated sample. *Ann. Inst. Stat. Math.*, **56**, 19–47.

Devroye, L. (1979). On the pointwise and integral convergence of recursive kernel estimates of probability densities. *Utilitas Math.*, **15**, 113–128.

Devroye, L. and Wagner, T. J. (1980). On the L_1 convergence of kernel estimators of regression functions with applications in discrimination. *Z. Wahrsch. Verw. Gebiete.*, **51**, 15–21..

Ditzler, G., Roveri, M., Alippí, C., and Polikar, R. (2015). Learning in nonstationary environments: A survey. *IEEE Comput. Intell. Mag.*, **10**, 12–25.

Domingos, P. and Hulten, G. (2000). Mining high-speed data streams. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 71–80.

Doukhan, P. (1994). *Mixing: Properties and examples*. Springer.

Duda, P., Jaworski, M. and Rutkowski, L. (2018). Knowledge discovery in data streams with the orthogonal series-based generalized regression neural networks. *Inf. Sci.*, **460–461**, 497–518.

Duda, P., Rutkowski, L., Jaworski, M. and Rutkowska, D. (2018). On the Parzen kernel-based probability density function learning procedures over time-varying streaming data with applications to pattern classification. *IEEE Trans. Cybern.*

Duong, T., Wand, M. P., Chacon, J. and Gramacki, A. (2019). ks: Kernel Smoothing. *R package version 1.11.6*.

Elaydi, S. (2005). *An introduction to difference equations*. 3rd Ed., Springer.

Eliseyev, A., Auboiroux, V., Costecalde, T., Langar, L., Charvet, G., Mestais, C., Aksenova, T. and Benabid, A.-L. (2017). Recursive exponentially weighted n-way partial least squares regression with recursive-validation of hyper-parameters in brain-computer interface applications. *Scientific Reports*, 7, 1–15.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman and Hall.

Fan, J., Hall, P., Martin, M. A., and Patil, P. (1996). On local smoothing of nonparametric curve estimators. *J. Amer. Statist. Assoc.*, **91**, 258–266.

Fan, J. and Yao, Q. (2003) *Nonlinear time series: Nonparametric and parametric methods*. Springer.

Faraway, J. (1990). Bootstrap selection of bandwidth and confidence bands for nonparametric regression. *J. Statist. Comput. Simul.*, **37**, 37–44.

Faraway, J. and Jhun, M. (1990). Bootstrap choice of bandwidth for density estimation. *J. Amer. Statist. Assoc.*, **85**, 1119–1122.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, 7, 179–188.

Friedman, J. H. (1997). On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data. Min. Knowl. Discov.*, **1**, 55–77.

Gama, J. (2010). *Knowledge discovery from data streams*. Chapman and Hall.

Gama, J. and Gaber, M. M. (Eds.). (2007). *Learning from data streams: Processing techniques in sensor networks*. Springer.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M. and Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Comput. Surv.*, **46**, 1–37.

Gao, J. (2007) *Nonlinear time series: Semiparametric and nonparametric methods*. Chapman & Hall/CRC.

García-Treviño, E. and Barria, J. (2012). Online wavelet-based density estimation for non-stationary streaming data. *Comp. Statist. Data Anal.*, **56**, 327–344.

Gehrels, N., Chincarini, G., Giommi, P., Mason, K. O., Nousek, J. A., Wells, A. A., White, N. E., Barthelmy, S. D., Burrows, D. N., Cominsky, L. R., Hurley, K. C., Marshall, F. E., Mészáros, P., Roming, P. W. A., Angelini, L., Barbier, L. M., Belloni, T., Campana, S., Caraveo, P. A., … Zhang, W. W. (2004). The swift gamma-ray burst mission. *Astrophys. J.*, **611**, 1005–1020.

Gijbels, I., Pope, A., and Wand, M. P. (1999). Understanding exponential smoothing via kernel regression. *J. R. Stat. Soc. Ser. B*, **61**, 39–50.

Goel, S. K. and Rodriguez, D. M. (1987). A note on evaluating limits using Riemann sums. *Math. Mag.*, **60**, 225–228.

Gomes, H. M., Barddal, J. P., Enembreck, F. and Bifet, A. (1987). A survey on ensemble learning for data stream classification. *ACM Comput. Surv.*, **23**, 23–36.

Gramacki, A. (2018). *Nonparametric Kernel Density Estimation and Its Computational Aspects*. Springer.

Gray, A. and Moore, A. (2003). Nonparametric density estimation: Toward computational tractability. In *Proc. SIAM Int. Conf. Data Mining*, 203–211.

Greblicki, W. and Pawlak, M. (1987). Necessary and sufficient pointwise consistency conditions for recursive kernel regression estimate. *J. Multivar. Anal.*, **23**, 67–76.

Greblicki, W. and Rutkowska, D. and Rutkowski, L. (1983). An orthogonal series estimate of time-varying regression. *Ann. Inst. Statist. Math.*, **35**, 215–228.

Grillenzoni, C. (2000). Nonparametric regression for nonstationary processes. *J. Nonparametric Stat.*, **12**, 265–282.

Györfi, L., Kohler, M., Krzyżak, A. and Walker, H. (2002). *A distribution-free theory of nonparametric regression*. Springer.

Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.*, **11**, 1156–1174.

Hall, P. (1987). On Kullback–Leibler loss and density estimation. *Ann. Statist.*, **15**, 1491–1519.

Hall, P. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *J. Multivar. Anal.*, **32**, 177–203.

Hall, P. and Heyde, C. C. (1980). *Martingale limit theory and its application*. Academic Press.

Hall, P., Lahiri, S. N. and Polzehl, J. (1995). On bandwidth choice in nonparametric regression with both short- and long-range dependent errors. *Ann. Statist.*, **23**, 1921–1936.

Hall, P., Lahiri, S. N. and Truong, Y. K. (1995). On bandwidth choice for density estimation with dependent data. *Ann. Statist.*, **23**, 2241–2263.

Hall, P. and Marron, J. S. (1991). Local minima in cross-validation functions. *J. R. Stat. Soc. Ser. B*, **53**, 245–252.

Hall, P., Müller, H.-G. and Wu, P.-S. (2006). Real-time density and mode estimation with application to time-dynamic mode tracking. *J. Comput. Graph. Stat.*, **15**, 82–100.

Hall, P. and Patil, P. (1994). On the efficiency of on-line density estimators. *IEEE Trans. Inf. Theory*, **40**, 1504–1512.

Hall, P. and Van Keilegom, I. (2003). Using difference-based methods for inference in nonparametric regression with time series errors. *J. R. Stat. Soc. Ser. B*, **65**, 443–456.

Härdle, W. (1991). *Smoothing techniques: With implementation in S*. Springer.

Härdle, W. and Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.*, **13**, 1465-1481.

Hart, J. D. (1991). Kernel regression estimation with time series errors. *J. Roy. Statist. Soc. Ser. B* , **53**, 173–187.

Hart, J. D. (1994). Automated kernel smoothing of dependent data by using time series cross-

validation. *J. Roy. Statist. Soc. Ser. B* , **56**, 529–542.

Harvey, A. C. (1990). *Forecasting, structural time series models and the Kalman filter*. Cambridge University press.

Harvey, A. and Oryshchenko, V. (2012). Kernel density estimation for time series data. *Int. J. Forecast.*, **28**, 3–14.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. 2nd Ed., Springer.

Heinz, C. and Seeger, B. (2008). Cluster kernels: Resource-aware kernel density estimators over streaming data. *IEEE Trans. Knowl. Data Eng.*, **20**, 880–893.

Hitczenko, P. (1990). Best constants in martingale version of Rosenthal's inequality. *Ann. Probab.*, **18**, 1656–1668.

Hofmeyr, D. P., Pavlidis, N. G. and Eckley, I. A. (2016). Divisive clustering of high dimensional data streams. *Stat. Comput.*, **26**, 1101–1120.

Holt, C. C. (1957). Forecasting seasonals and trends by exponentially weighted moving averages. *ONR Research Memorandum 52*, Carnegie Institute of Technology. Pittsburgh, Pennsylvania.

Huang, Y., Chen, X. and Wu, W. B. (2014). Recursive nonparametric estimation for time series. *IEEE Trans. Inf. Theory*, **60**, 1301–1312.

Hulten, G., Spencer, L. and Domingos, P. (2001). Mining time-changing data streams. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 97–106.

Inoue, A., Jin, L. and Rossi, B. (2017). Rolling window selection for out-of-sample forecasting with time-varying parameters. *J. Econom.*, **196**, 55–67.

Jones, M. C. (1993a). Do not weight for heteroscedasticity in nonparametric regression. *Austral. J. Statist.*, **35**, 89–92.

Jones, M. C. (1993b). Simple boundary correction for kernel density estimation. *Stat. Comput.*, **3**, 135–146.

Jones, M. C. and Foster, P. J. (1996). A simple nonnegative boundary correction method for kernel density estimation. *Statistica Sinica*, **6**, 1005–1013.

Kolter, J. Z. and Maloof, M. A. (2007). Dynamic weighted majority: An ensemble method for drifting concepts. *J. Mach. Learn. Res.*, **8**, 2755–2790.

Krzyżak, A. and Pawlak, M. (1984). Almost everywhere convergence of recursive regression function estimate and classification. *IEEE Trans. Inform. Theory*, **30**, 91–93.

Krzyżak, A. (1992). Global convergence of the recursive kernel regression estimates with applications in classification and nonlinear system estimation. *IEEE Trans. Inform. Theory*, **38**, 1323–1338.

Lam, H. T. and Bouillet, E. (2015). Flexible sliding windows for kernel regression based bus arrival time prediction. In *Machine Learning and Knowledge Discovery in Databases*, Edited by A. Bifet, M. May, B. Zadrozny, R. Gavalda, D. Pedreschi, F. Bonchi, J. Cardoso and M. Spiliopoulou, 68–84.

Lin, Z. and Bai, Z. (2010). *Probability inequalities*. Science Press.

Luo, L. and Song, P. X.-K. (2020). Renewable estimation and incremental inference in generalized linear models with streaming data sets. *J. R. Stat. Soc. Ser. B*, **82**, 69–97.

Luts, J., Broderick, T. and Wand, M. P. (2014). Real-time semiparametric regression. *J. Comput. Graph. Stat.*, **23**, 589–615.

Mammen, E., Linton, O. and Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.*, **27**, 1443–1490.

Marron, J. S. and Härdle, W. (1986). Random approximations to some measures of accuracy in nonparametric curve estimation. *J. Multivar. Anal.*, **18**, 1656–1668.

McLeish, D. L. (2005). *Monte Carlo Simulation and Finance*. Wiley.

Minku, L. L. and Yao, X. (2012). DDD: A new ensemble approach for dealing with concept drift. *IEEE Trans. Knowl. Data Eng.*, **24**, 619–633.

Mokkadem, A., Pelletier, M., and Slaoui, Y. (2009a). Revisiting Révész's stochastic approximation method for the estimation of a regression function. *ALEA. Lat. Am. J. Probab. Math. Stat.*, **6**, 63–114.

Mokkadem, A., Pelletier, M., and Slaoui, Y. (2009b). The stochastic approximation method for the estimation of a multivariate probability density. *J. Stat. Plan. Inference*, **139**, 2459–2478.

Müller, H.-G. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.*, **15**, 610–625.

Nagler, T. and Czado, C. (2016). Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *J. Multivar. Anal.*, **151**, 69–89.

Noble, J. and Adams, N. M. (2018). Real-time dynamic network anomaly detection. *IEEE Intell. Syst.*, **33**, 5–18.

Opsomer, J., Wang, Y. and Yang, Y. (2001). Nonparametric regression with correlated errors. *Stat. Sci.*, **16**, 134–153.

Padilla, O. H. M., Athey, A., Reinhart, A. and Scott J. G. (2019). Sequential nonparametric tests for a change in distribution: An application to detecting radiological anomalies. *J. Amer. Statist. Assoc.*, **114**, 514–528.

Park, B. U. and Marron, J. S. (1990). Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc.*, **85**, 66–72.

Parzen, E. (1962). On the estimation of a probability density function and the mode. *Ann. Math. Statist.*, **33**, 1065-1076.

Pavlidis, N. G., Tasoulis, D. K., Adams, N. M. and Hand, D. J. (2011). λ-Perceptron: An adaptive classifier for data streams. *Pattern Recognit.*, **44**, 78–96.

Peligrad, M. (1996). On the asymptotic normality of sequences of weak dependent random variables. *J. Theor. Probab.*, **9**, 703–715.

Qahtan, A., Wang, S., and Zhang, X. (2017). KDE-Track: An efficient dynamic density estimator for data streams. *IEEE Trans. Knowl. Data Eng.*, **29**, 642–655.

Qin, S. J. (1998). Recursive PLS algorithms for adaptive data modelling. *Comput. Chem. Eng.*, **22**, 503–514.

Rey, D. and Neuhäuser, M. (2011). Wilcoxon-signed-rank test. In *International Encyclopedia of Statistical Science*, Ed. by M. Lovric, 1658–1659. Wiley.

Révész, P. (1973). Robbins-Monro procedure in a Hilbert space and its application in the theory of learning processes I, II. *Studia Sci. Math. Hungar.*, **8**, 391–398, 469–472.

Rio, E. (2013). *Inequalities and limit theorems for weakly dependent sequences*. Technical report,

Université de Versailles.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, **27**, 832-837.

Ross, G. J., Adams, N. M., Tasoulis, D. K. and Hand, D. J. (2012). Exponentially weighted moving average charts for detecting concept drift. *Pattern Recognit. Lett.*, **33**, 191–198.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.*, **90**, 1257–1270..

Rutkowski, L. (1982a). On Bayes risk consistent pattern recognition procedures in a quasi-stationary environment. *IEEE Trans. Pattern Anal. Mach. Intell.*, **4**, 84–87.

Rutkowski, L. (1982b). On-line identi

cation of time-varying systems by nonparametric technique. *IEEE Trans. Autom. Control*, **27**, 228–230.

Rutkowski, L. (1984). On nonparametric identification with prediction of time-varying systems. *IEEE Trans. Autom. Control*, **29**, 58–60.

Rutkowski, L. (1985). Nonparametric identification of quasi-stationary systems. *Syst. Control Lett.*, **6**, 33–35.

Rutkowski, L. (1989a). Application of multiple Fourier series to identification of multivariate non-stationary systems. *International J. of Systems Science*, **20**, 1993–2002.

Rutkowski, L. (1989b). Non-parametric learning algorithms in time-varying environments. *Signal Process.*, **18**, 129–137.

Rutkowski, L. (2004a). Adaptive probabilistic neural networks for pattern classification in time-varying environment. *IEEE Trans. Neural Netw.*, **15**, 811–827.

Rutkowski, L. (2004b). Density-free Bayes risk consistency of nonparametric pattern recognition procedures. *IEEE Trans. Neural Netw.*, **15**, 576–596.

Sayed-Mouchaweh, M. (Ed.). (2019). *Learning from data streams in evolving environments: Methods and applications*. Springer.

Scott, D. W. and Sain, S. R. (2004). Multidimensional density estimation. In *Data mining and data visualization*, Ed. by C. R. Rao and E. J. Wegman, 229–261. Elsevier.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Stat. Soc. Ser. B*, **53**, 683–690.

Shen, Y., Chen, T. and Giannakis, G. B. (2019). Random feature-based online multi-kernel learning in environments with unknown dynamics. *J. Mach. Learn. Res.*, **20**, 773–808.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical processes with applications to statistics*. Wiley.

Slaoui Y. (2016). Optimal bandwidth selection for semi-recursive kernel regression estimators. *Stat. Interface*, **9**, 375–388.

Street, W. N. and Kim, Y. (2001). A streaming ensemble algorithm (SEA) for large-scale classification. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 377–382. ACM.

Subramaniam, S., Palpanas, T., Papadopoulos, D., Kalogeraki, V. and Gunopulos, D. (2006). Online outlier detection in sensor data using non-parametric models. In *Proceedings of the VLDB Conference 2006*, **6**, 187–198.

Talagala, P. D., Hyndman, R. J., Smith-Miles, K., Kandanaarachchi, S. and Muñoz, M. A. (2019). Anomaly detection in streaming nonstationary temporal data. *J. Comput. Graph. Stat.*, **0**, 1–21.

Tasoulis, D. K., Adams, N. M. and Hand, D. J. (2006). Unsupervised clustering in streaming data. In *Sixth IEEE International Conference on Data Mining – Workshops (ICDMW'06)*, 638–642.

Taylor, C. (1989). Bootstrap Choice of Smoothing Parameter in Kernel Density Estimation. *Biometrika*, **76**, 705–712.

Thompson, S. E., Fraquelli, D., van Cleve, J. E., and Caldwell, D. A. (2016). *Kepler: A search for terrestrial planets*. Retrieved from: `http://archive.stsci.edu/kepler/documents.html`.

Tveten, M., and Glad, I. K. (2019). Online detection of sparse changes in high-dimensional data streams using tailored projections. *arXiv:1908.02029v1*.

Vogt, M. (2012). Nonparametric regression for locally stationary time series. *Ann. Statist.*, **40**, 2601–2633.

Walk, H. (2001). Strong universal pointwise consistency of recursive regression estimates. *Ann. Inst. Statist. Math.*, **53**, 691–707.

Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing*. CRC Press.

Wand, M. P. and Jones, M. C. (2019). KernSmooth: Functions for kernel smoothing supporting Wand & Jones (1995). *R package version 2.23-16*.

Wegman, E. J. and Davies, H. I. (1979). Remarks on some recursive estimators of a probability density. *Ann. Statist.*, 7, 316–327.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Manag. Sci.*, **6**, 324–342.

Wolverton, C. T. and Wagner, T. J. (1969). Asymptotically optimal discriminant functions for pattern recognition. *IEEE Trans. Inform. Theory*, **15**, 258–265.

Xu, M. and Samworth, R. J. (2019). High-dimensional nonparametric density estimation via symmetry and shape constraints. *arXiv: 1903.06092*.

Wolverton, C. T. and Wagner, T. J. (1969). Asymptotically optimal discriminant functions for pattern recognition. *IEEE Trans. Inform. Theory*, **15**, 258–265.

Yao, Q. and Tong, H. (1998). Cross-validatory bandwidth selections for regression estimation based on dependent data. *J. Stat. Plan. Inference*, **68**, 387–415.

Zhang, H. (2004). The optimality of naive Bayes. In *FLAIRS2004 Conference*. AAAI Press.

Zhang, R., Mei, Y. and Shi, J. (2019). Robust real-time monitoring of high-dimensional data streams. *arXiv: 1906.02265*.

Zhang, T. and Wu, W. B. (2015). Time-varying nonlinear regression models: Nonparametric estimation and model selection. *Ann. Statist.*, **43**, 741–768.

Zheng, Y., Jestes, J., Phillips, J. and Li, F. (2013). Quality and efficiency in kernel density estimates for large data. In *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 433–-444.

Zhou, D., Guo, J., Zhang, Y., Chai, J., Liu, H., Liu, Y., Huang, C., Gui, X. and Liu, Y. (2016). Distributed data analytics platform for wide-area synchrophasor measurement systems. *IEEE Trans. Smart Grid*, 7, 2397–2405.

Zhou, A., Cai, Z., Wei, L. and Qian, W. (2003). Towards density estimation over data streams.

In *Proc. 8th Int. Conf. Database Syst. Adv. Appl.*, 285–292.