DEPARTMENT OF ECONOMICS
AND BUSINESS ECONOMICS
AARHUS UNIVERSITY

CREATES
Center for Research in Econometric Analysis of Time Series

Uniform Convergence Rates of Kernel-Based Nonparametric Estimators for Continuous Time Diffusion Processes: A Damping Function Approach

Shin Kanaya

CREATES Research Paper 2015-50

Department of Economics and Business Economics
Aarhus University
Fuglesangs Allé 4
DK-8210 Aarhus V
Denmark

Email: oekonomi@au.dk
Tel: +45 8716 5515

Uniform Convergence Rates of Kernel-Based Nonparametric Estimators for Continuous Time Diffusion Processes: A Damping Function Approach*

Shin Kanaya†
Aarhus University and CREATES

Abstract

In this paper, we derive uniform convergence rates of nonparametric estimators for continuous time diffusion processes. In particular, we consider kernel-based estimators of the Nadaraya-Watson type with introducing a new technical device called a *damping function*. This device allows us to derive sharp uniform rates over an infinite interval with minimal requirements on the processes: The existence of the moment of any order is not required and the boundedness of relevant functions can be significantly relaxed. Restrictions on kernel functions are also minimal: We allow for kernels with discontinuity, unbounded support and slowly decaying tails. Our proofs proceed by using the covering-number technique from empirical process theory and exploiting the mixing and martingale properties of the processes. We also present new results on the path-continuity property of Brownian motions and diffusion processes over an infinite time horizon. These path-continuity results, which should also have an independent interest, are used to control discretization biases of the nonparametric estimators. The obtained convergence results are useful for non/semiparametric estimation and testing problems of diffusion processes.

This version: November 2015.
Keywords: Diffusion process; uniform convergence; kernel estimation; nonparametric.
JEL codes: C14; C32; C58.

*This paper was part of my Ph.D. dissertation at the University of Wisconsin-Madison. I am indebted to my supervisor Bruce E. Hansen for his guidance and support. I am very grateful to Dennis Kristensen and Jack R. Porter for their valuable advice and encouragements. I also thank Martin Browning, Xiaohong Chen, Valentina Corradi, Bonsoo Koo, Bent Nielsen, Andrew Patton, Olivier Scaillet, Neil Shephard and seminar participants at University of Wisconsin-Madison and University of Oxford, for helpful comments and suggestions. I gratefully acknowledge support from CREATES, Center for Research in Econometric Analysis of Time Series, funded by the Danish National Research Foundation (DNRF78). I would like to thank Co-Editor, Oliver B. Linton, and two anonymous referees for their constructive and valuable comments, which have greatly improved the original version of this paper.

†Address: Department of Economics and Business Economics, Aarhus University, Fuglesangs Alle 4, Aarhus V, 8210, Denmark. E-mail: skanaya@econ.au.dk. Phone: +45 87165984.

1 Introduction

In this paper, we derive the uniform convergence rates of kernel-based nonparametric estimators for continuous time processes. We specifically consider diffusion processes described by the following type of stochastic differential equation (SDE):

$$dX_s = \mu(X_s)\,ds + \sigma(X_s)\,dW_s, \tag{1}$$

where $\{W_s\}$ is the standard Brownian motion. $\mu(\cdot)$ and $\sigma^2(\cdot)$ are called the drift and diffusion (volatility) functions, respectively, which are of our interest in the estimation. To estimate $\mu(\cdot)$ and $\sigma^2(\cdot)$ nonparametrically, it is standard to use the Nadaraya-Watson (NW) type estimators:

$$\tilde{\mu}(x) := \frac{(1/T)\sum_{j=1}^{n-1} K_h(X_{j\Delta} - x)\left[X_{(j+1)\Delta} - X_{j\Delta}\right]}{(\Delta/T)\sum_{j=1}^{n} K_h(X_{j\Delta} - x)}; \tag{2}$$

$$\tilde{\sigma}^2(x) := \frac{(1/T)\sum_{j=1}^{n-1} K_h(X_{j\Delta} - x)\left[X_{(j+1)\Delta} - X_{j\Delta}\right]^2}{(\Delta/T)\sum_{j=1}^{n} K_h(X_{j\Delta} - x)}, \tag{3}$$

where $K_h(z) := K(z/h)/h$; K is a kernel function; and h is a bandwidth. Florens-Zmirou (1993) first considered this type of estimator, and several authors, such as Jiang and Knight (1997), Bandi and Phillips (2003), Nicolau (2003) and Aït-Sahalia and Park (2013), have further developed these estimators. While they have established asymptotic properties of the estimators, such as the consistency and the asymptotic (mixed) normality, they have focused on the pointwise convergence. The uniform convergence has not yet been fully considered in the literature. This study presents the uniform convergence rates of the NW type estimators for continuous time diffusion processes of the type (1).

In the discrete time setting, several authors, including Bierens (1983), Andrews (1995), Liebscher (1996), Masry (1996), Bosq (1998), Fan and Yao (2003), Ango Nze and Doukhan (2004), Hansen (2008), Kristensen (2009), Kong, Linton and Xia (2010), and Gao, Kanaya, Li and Tjøstheim (2015), have studied the uniform convergence of kernel-based estimators. Their results may not be directly imported to our continuous time setting; however, some of their techniques may be used in our context to some extent. There are several difficulties in dealing with diffusion processes: First, the results in the discrete time setting require the existence of the higher order moment of the process and the (uniform) boundedness of some relevant functions. These requirements may be too strong for some class of diffusion processes, which, in fact, are not satisfied by several parametric models used in the econometrics literature.

Second, the estimators (2) and (3) are based on the discrete time sample $\{X_{j\Delta}\}_{j=1}^{n}$ and incur biases due to the discretization, which do not appear in the discrete time setting. Note that our estimation targets $\mu(\cdot)$ and $\sigma^2(\cdot)$ are the *instantaneous* conditional mean and variance functions, respectively. In estimating such objects without assuming the availability of a continuously recorded path of the process, we generally experience the discretization biases, unless some parametric restriction or the so-called cross restriction are exploited (see arguments in Hansen and Scheinkman, 1995; Aït-Sahalia, 1996a; Bandi and Phillips, 2002; Kristensen, 2010a). To control the discretization biases, we consider the *infill* assumption, which means that the time distance between adjacent observations, Δ, shrinks to zero as the sample size n tends to infinity. Given the infill, we have the effects due to the discretization asymptotically negligible (and obtain the consistency). Our derivation of the rate of discretization bias is based on the (sample) path continuity of the process. However, the existing results on the path continuity are not sufficient for our purpose. We prove new results on the uniform/global modulus of continuity of Brownian motions and diffusion processes, which should also have an independent interest.

Third, the aforementioned studies that consider the discrete time setting exploit the asymptotic independence property of the process for the uniform results, which is (typically) implied by the mixing condition. While our proofs also use this property (through the large deviation type inequality), we need to consider a different treatment since we work with the infill. The dependence between consecutive observations becomes stronger as $n \to \infty$ under the infill ($\Delta \to 0$), which leads to slower convergence (than in the standard discrete time setting). Note that we also work with the longspan assumption where the time horizon of the observations, $T(= n\Delta) \to \infty$. This is necessary to exploit the asymptotic independence.

To circumvent the difficulties in the continuous time setting, we propose to introduce a technical device $B(\cdot)$, which we call a *damping* function, and consider the following modified estimators for the drift and diffusion functions, instead of (2) and (3):

$$\hat{\mu}(x) := \hat{\Psi}_{\mu}(x) / \hat{\Pi}(x); \tag{4}$$

$$\hat{\sigma}^2(x) := \hat{\Psi}_{\sigma^2}(x) / \hat{\Pi}(x), \tag{5}$$

where

$$\hat{\Psi}_{\mu}(x) := (1/T) \sum_{j=1}^{n-1} K_h(X_{j\Delta} - x) B(X_{j\Delta}) \left[X_{(j+1)\Delta} - X_{j\Delta}\right]; \tag{6}$$

$$\hat{\Psi}_{\sigma^2}(x) := (1/T) \sum_{j=1}^{n-1} K_h(X_{j\Delta} - x) B(X_{j\Delta}) \left[X_{(j+1)\Delta} - X_{j\Delta}\right]^2; \tag{7}$$

$$\hat{\Pi}(x) := (\Delta/T) \sum_{j=1}^{n} K_h(X_{j\Delta} - x) B(X_{j\Delta}). \tag{8}$$

Unlike the standard estimators in (2) and (3), each component of the new estimators includes $B(X_{j\Delta})$. We call (4) and (5) the *damped* versions of the NW estimators. We subsequently derive the uniform convergence rates of $\hat{\mu}(x)$ and $\hat{\sigma}^2(x)$. The function $B(\cdot)$ should take strictly positive values over the entire support of the process. The econometrician may choose this function arbitrarily so that the products of $B(\cdot)$ and relevant functions, e.g., $B(x)\mu(x)$ and $B(x)\sigma^2(x)$, are uniformly bounded. Note that $\mu(x)$ and/or $\sigma^2(x)$ are not bounded in many parametric models found in the literature. However, we can let $B(x)\mu(x)$ and $B(x)\sigma^2(x)$ uniformly bounded by choosing $B(x)$ with exponential decay rate (to zero as $|x| \to \infty$), if $\mu(x)$ and $\sigma^2(x)$ are at most of polynomial order (we provide conditions on $B(\cdot)$ and an example of $B(\cdot)$ in Section 3). We also note that by using the damping function, we can also work with a process whose invariant density is unbounded, allowing for highly skewed distributions.[1] This point is discussed in the Supplementary Material.

The introduction of the damping function does not affect the consistency property of the estimators. This is because $\hat{\Psi}_{\mu}(x)$ and $\hat{\Psi}_{\sigma^2}(x)$ converge to $B(x)\mu(x)\pi(x)$ and $B(x)\sigma^2(x)\pi(x)$ respectively, where $\pi(x)$ denotes the invariant density of the process, and $\hat{\Pi}(x)$ converges to $B(x)\pi(x)$. Thus, our new estimators (4) and (5) are respectively consistent for $\mu(x)$ and $\sigma^2(x)$, since $B(x)$ (as well as $\pi(x)$) is cancelled out. We also note that the limit normal distributions are the same for the new and standard estimators while asymptotic biases of the estimators are affected by the damping function (see discussions in Section 5 and the Supplementary Material). The uniform rate we have derived for the diffusion function estimator is optimal in the sense of Stone (1982). We also conjecture that the rate for the drift function estimator is also optimal since our uniform rate is $\sqrt{(\log T)/Th}$ and the pointwise rate (derived in the previous studies) is $\sqrt{1/Th}$, while further studies are need to confirm the optimality.

[1] For example, we allow for the gamma distribution with the shape parameter less than 1, whose density is unbounded (∞) at the left endpoint 0.

There are several advantages of introducing the damping function device. It enables us to employ a technique from empirical process theory, i.e., the method by the covering number. As discussed in Andrews (1994), van der Vaart (1998) and others, empirical process theory provides very useful techniques for establishing asymptotic theory in econometric and statistical problems. However, it often requires relatively strong conditions, such as the uniform boundedness of relevant functions, and thus may have limited applications. As stated above, even when some function $f(x)$ is not bounded, we can choose $B(x)$ so that $B(x) f(x)$ is (uniformly) bounded. Given this *induced* boundedness, we can more easily employ the technique from empirical process theory. The use of the covering number allows us to proceed without the so-called truncation technique, as used in Bosq (1998) and Hansen (2008), to prove the uniform convergence over an unbounded support, which in turns allows us to proceed without assuming the existence of the moment of the process (of any order). We can comfortably accommodate the infinite mean/variance case, for example. This is in contrast to the results in Andrews (1995), Bosq (1998) and Hansen (2008), which require the existence of the higher order moments to derive uniform convergence rates. We note that the covering number technique can work with almost all forms of the kernels functions, e.g., ones with discontinuity, unbounded support, and/or slowly decreasing tails. We also note that $B(x)$ plays an important role in controlling the discretization biases. As mentioned previously, for this purpose, we use the path continuity property of the diffusion processes, the so-called modulus of continuity. This property, unfortunately, may not generally hold under the longspan (with $T \to \infty$) due to the potential unboundedness of the drift and diffusion functions. However, we prove the modified version of the modulus of continuity, i.e., the continuity with a weighted sup-norm (setting the damping function as weight), which allows us to proceed.

Our uniform convergence results are useful in various econometric/statistical problems for diffusion processes. In fact, the author applies the results of the paper to nonparametric specification testing of Markov processes (Kanaya, 2014). They may be used for deriving asymptotic distributions results in specification testing problems of volatility components as found in Li (2007) and Corradi and Distaso (2010), as well as in derivative-security pricing problems as in Aït-Sahalia (1996a) and Kristensen (2008).

Some ideas found in econometrics and statistics may seem as similar to those of the damping function. For example, the so-called trimming device is often used to eliminate aberrant behaviors of nonparametric estimators (see Sec. 6 of Ichimura and Todd, 2007 for an overview) in two-step semiparametric estimation problems (e.g., as in Robinson, 1988; Ai, 1997; Cosslett, 2004). This is similar in its concept to our damping device. However, trimming completely discards some part of estimated values, typically very large ones (or small ones if the estimator is in the denominator). In contrast, the damping function puts less weight on some observations and does not discard any part of observations or estimated values. Chen and Fan (2006) consider a weighting function to verify asymptotic results of copula-based semiparametric models. Roughly, they show the convergence of the (rescaled) empirical distribution function based on a weighed norm, $\sup_x |f(x) - g(x)| \tilde{w}(x)$, instead of the usual sup-norm $\sup_x |f(x) - g(x)|$ (Sec. 4 in Chen and Fan, 2006). The function $\tilde{w}(x)$ is introduced to suppress large values in their semiparametric score function (and its derivative) near the boundaries. Our damping function plays the same role as Chen and Fan's weighting function, in that it is used to suppress large values in the nonparametric estimators. However, we note that Chen and Fan's weighting function is used only to modify the definition of the metric. In contrast, our damping function is used to modify the estimators themselves.

In related studies, Fan and Zhang (2003) and Xu (2009, 2010) also consider kernel-based nonparametric estimators for diffusion processes, working with the local polynomial and/or re-weighted estimators. It is known that these estimators in general possess better bias properties compared to the simple NW

type as (2) and (3). While their results are pointwise, our techniques of the damping function may also be used to establish the uniform results of their estimators. Krinsenten (2008) consider convergence rates of nonparametric estimators with respect to the L_2 integral norm (the proof of Theorem 5 in p. 405), which is required for his derivations of asymptotic results for estimators of derivative-security pricing. Our uniform convergence theorems can be used to derive such L_2-convergence results, and complement his results. Koo and Linton (2012) consider a kind of time varying semiparametric diffusion model and present the uniform convergence results for their semi/nonparametric estimators. Since they exploit the drift function's parametric restriction as well as the (so-called) cross restriction of stationary diffusions, the infill assumption is not required. As a result, their estimators are free of the discretization biases, where the existing results for the discrete time processes may be applied directly. Finally, Kutoyants (1999) and van Zanten (2000) consider the (invariant) density estimation for diffusion processes and present the uniform convergence of the kernel-based density estimator. However, their results are distinct from ours since they assume the availability of a continuous path of the process.

The rest of the paper is organized as follows: In Section 2, we set up our framework. In Section 3, we derive new results on the path continuity of Brownian motions and diffusion processes. Section 4 presents general convergence theorems for functionals of diffusion processes. Section 5 presents uniform convergence theorems for our new estimators (4) and (5). Proofs are found in an Appendix, and some additional proofs, results and discussions are provided in Supplementary Material to this article.

For definitional equations, we write $A := B$ and $C =: D$ throughout the text. The former means that A is defined by B, and the latter means that D is defined by C. We also write $\partial f(x)$ and $\partial^k f(x)$ to denote the first and k-th derivatives of a function $f(x)$, respectively.

2 Framework

This section formally describes our framework. Let $\{X_s\}_{s\geq 0}$ be a time-homogeneous Markov diffusion process described by the stochastic differential equation (1) with $\{W_s\}_{s\geq 0}$ a standard Brownian motion. The processes are defined on a filtered probability space $(\Omega, \mathfrak{F}, \{\mathfrak{F}_s\}_{s\geq 0}, \Pr)$, which satisfies the usual conditions. The functional forms of $\mu(\cdot)$ and $\sigma(\cdot)$ are assumed to be unknown and are of our interest in the estimation. The domain of X_s is denoted by I, which is a real interval whose left and right boundaries are given by l and r respectively ($-\infty \leq l < r \leq \infty$). While we mainly consider $I = \mathbb{R} := (-\infty, \infty)$, all the subsequent discussions and results hold true for some other choices of I upon suitable/slight modifications. In the Supplementary Material, we provide some discussions/results in particular for $I = (0, \infty)$ and $[0, \infty)$.

We require the following conditions for establishing the uniform convergence results of the estimators over $x \in (l, r)$:

A1. **(i)** $\mu(\cdot)$ ($: (l,r) \to \mathbb{R}$) and $\sigma(\cdot)$ ($: (l,r) \to (0,\infty)$) are twice continuously differentiable on (l, r)**.** **(ii)** The process $\{X_s\}$, as a solution to (1), is recurrent.

A2. **(i)** It holds that $\int_l^r m(x)\,dx < \infty$, where $m(\cdot)$ is the speed density:

$$m(x) := \sigma^{-2}(x) \exp\left\{2\int_c^x \left[\mu(u)/\sigma^2(u)\right] du\right\} \quad \text{for } x \in (l, r). \tag{9}$$

(ii) $\{X_s\}$ is strictly stationary with the invariant probability density $\pi(\cdot)$ which is bounded at any interior point of I (i.e., $\pi(x) < \infty$ for each $x \in (l, r)$), and is α-mixing (strongly mixing) with

mixing coefficients $\alpha(s)$ satisfying

$$\alpha(s) \leq As^{-\beta} \text{ for some } \beta > 0 \text{ and } A > 0. \tag{10}$$

These conditions are standard. (A1.i) is sufficient for the existence of a unique strong solution to (1) up to an explosion time (and up to at least the first hitting time on $l = 0$ if $I = [0, \infty)$) for any initial distribution of X_0.[2] Given (A1.ii), the solution to (1) should be non-explosive and should not be killed at any point in I. Under (A1.i), a simple sufficient condition for (A1.ii) is that

$$S(x) \to -\infty \text{ as } x \to l; \text{ and } S(x) \to \infty \text{ as } x \to r, \tag{11}$$

where $S(x)$ is the scale function:

$$S(x) := \int_c^x \exp\left\{-2\int_c^y \left[\mu(u)/\sigma^2(u)\right] du\right\} dy \text{ for } x \in (l, r), \tag{12}$$

with c representing a generic element in (l, r) (see Proposition of 5.22(a) in p. 345 of Karatzas and Shreve, 1991; henceforth, KS91). The condition in (11) means that neither the left nor the right boundary is attracting (at the same time, neither is attainable). Therefore, the process is non-explosive. If either of the boundaries is bounded, we allow X_s to attain boundaries in some finite time with a positive probability. For example, if $l = 0$ and $I = [0, \infty)$, $\lim_{x \to l} S(x)$ may be finite (see Sec. 6 in Ch. 15 of Karlin and Taylor, 1981; KT81, henceforth). In this case, the SDE (1) by itself may not be able to fully describe the behavior of the process through infinite time horizon, in particular after the hit on $l = 0$, while we need some additional specification.[3]

(A2.i) is sufficient (and necessary) for the process to be positively recurrent and to have the invariant probability density $\pi(\cdot)$ under (A1). In particular, $\pi(\cdot)$ is given as $\pi(x) = m(x) / \int_l^r m(z)\, dz$ (see, e.g., Ch. 15 of KT81). Then, given (A1.i), $\pi(x)$ is twice continuously differentiable at $x \in (l, r)$. We also note that (A1.i) guarantees the existence of the transition density $p_s(x, y)\, dy = \Pr[X_{t+s} \in dy | X_t = x]$ (for any $s, t \geq 0$; see, e.g., McKean 1956, Sec. 5), which, together with the existence of the invariant density, in turn implies the existence of joint density of (X_t, X_{t+s}), $\pi_{t,t+s}(x, y)\, dxdy = \Pr[X_t \in dx, X_{t+s} \in dy]$. This fact is effectively used to derive sharp convergence rates (see the proof of Lemma 4).

[2] This statement can be illustrated as follows: The continuity of $\mu(\cdot)$ and $\sigma^2(\cdot)$ in (A1.i) is sufficient for the existence of a unique weak solution up to an explosion time (and up to the first hitting time on $l = 0$ if $I = [0, \infty)$), given any initial distribution (Theorem 5.15 in p. 341 of Karatzas and Shreve, 1991; KS91, henceforth). The differentiability in (A1.i) implies the local Lipschitz continuity of $\mu(\cdot)$ and $\sigma^2(\cdot)$, and thus the pathwise uniqueness of the solution to (1) (Theorem 2.5 in p. 287 and Remark 3.3 in p. 301 of KS91). Then, by applying Yamada and Watanabe's theorem (Corollary 3.23 in p. 310 of KS91), which states that weak existence and pathwise uniqueness imply strong existence, we obtain the desired result.

[3] We can impose such a specification as follows. Let l be an instantaneously reflecting boundary (that is, the process, having attained at $l = 0$, returns to the interior immediately (with the Lebesgue measure of time spent at l equal to zero). In particular, we consider that $\{X_s\}$ is a diffusion in Feller's sense determined by the scale function $S(\cdot)$ given by (12) and the speed measure $M(\cdot)$, where $M(\cdot)$ is a Borel measure with $M(\{0\}) = 0$; $M((a; b]) = \int_a^b m(x)dx$ for $0 < a < b < \infty$; and $m(\cdot)$ given by (9). We note that $M(\{0\}) = 0$ means that the boundary $l = 0$ is instantaneously reflecting. For the definition of diffusion processes in terms of the scale function and the speed measure, see e.g., Ch. I-III of Mandl (1968), or Sec. 8-9 of Kent (1978). The process $\{X_s\}$ constructed in this way, by $S(\cdot)$ and $M(\cdot)$, is a weak solution to the SDE (1), which can be verified by using the same arguments as in V.48 of Rogers and Williams (2000). By regarding $\{X_s\}$ as a weak solution to (1), we can proceed in the same way as in the case where (1) fully determines the behavior of $\{X_s\}$.

We also note that given that $\lim_{x \to l} S(x)$ is finite and $S(x) \to \infty$ as $x \to r$, $\{X_s\}$ is recurrent if l is made an instantaneously reflecting boundary, which can be proved by arguments similar to those for natural-scale diffusions, as Theorem 20.15 of Ch. 20 of Kallenberg (2002).

Under (A2.i) and boundedness condition on $\pi(\cdot)$ in (A2.ii), it is not restrictive to further assume that $\pi(\cdot)$ is uniformly bounded if $I = \mathbb{R}$ (as we do so in Section 5). While we might be able to construct some pathological example where π is not uniformly bounded on $\mathbb{R}$, many parametric models found in the literature satisfy this uniform boundedness. On the other hand, if either of the boundaries is finite, e.g., $l = 0$, we allow $\pi(x)$ to be unbounded around 0, i.e., it may hold that $\pi(x) \to \infty$ as $x \to 0$. $\{X_s\}$ typically has this kind of unbounded density when $l = 0$ is attainable, as we exemplify in the Supplementary Material. Even in this unbounded case, we can derive the uniform convergence results by suitably choosing the form of a damping function B, as we discuss in the Supplementary Material.

The strict stationarity condition in (A2.ii) is imposed for simplicity. We can work with some heterogeneous processes and remove this condition by using arguments similar to those in Kristensen (2009) (as long as the other conditions are fulfilled). Almost all (parametric) models found in the econometrics literature can satisfy the mixing condition in (A2.ii). We may use various results to check the condition (10) (see, e.g., Doukhan, 1994; Hansen and Scheinkman, 1995; Hansen, Scheinkman and Touzi, 1998; Veretennikov, 1987, 1997, 1999, Kusuoka and Yoshida, 2000; Chen, Hansen and Carrasco, 2010).[4] Some previous studies which consider the uniform convergence rates of the nonparametric estimators, such as Andrews (1995) and Hansen (2008), assume that β is sufficiently large (e.g., $\beta > 2$ at least). In contrast, our uniform convergence results are applicable to any $\beta(> 0)$, while the corresponding rate may be slower for smaller β (see Theorem 7 in the Supplementary Material). Chen *et al.* (2010) present a class of diffusion processes with very slowly decaying mixing coefficients (small β), and argue that such processes may exhibit a property resembling long memory. These processes are in the scope of our convergence theorems. Note that our results on the path continuity (presented in Section 3) do not require the conditions in (A2) and they are applicable to both stationary and nonstationary processes if (A1) is satisfied.

Asymptotic Scheme: Before concluding this section, we describe the asymptotic scheme we consider throughout the paper. We assume that the continuous-time process $\{X_s\}$ is observed at discrete time points, $s = \Delta, 2\Delta, \ldots n\Delta$ over the time interval $(0, T]$, where T is some positive number; n is the number of observations; and $\Delta = T/n$ is the time distance between adjacent observations. We work under the infill and longspan asymptotic scheme, i.e., $\Delta \to 0$ and $T \to \infty$ (as $n \to \infty$). The availability of the equi-spaced data $\{X_{j\Delta}\}_{j=1}^n$ is assumed only for (notational) simplicity. All of our convergence theorems may be applied to non equi-spaced data $\{X_{t_j}\}_{j=1}^n$ if the shrinking rates of $\bar{\Delta}$ and $\underline{\Delta}$ are the same as that of Δ given in each theorem, where $\bar{\Delta}$ and $\underline{\Delta}$ are defined as: $\bar{\Delta} := \max_{1\leq j\leq n-1}(t_{j+1} - t_j)$ and $\underline{\Delta} := \min_{1\leq j\leq n-1}(t_{j+1} - t_j)$.

We note that the long-span is generally required to identify the drift term nonparametrically without relying on the cross-restriction (see arguments in Bandi and Phillips, 2003; Kristensen, 2010b). In contrast, it is not necessary to obtain the pointwise consistency in the diffusion function estimation (see, e.g., Florens-Zmirou, 1993; Bandi and Phillips, 2003). However, our proofs for the uniform convergence exploit the asymptotic independence between distant observations as $T \to \infty$, which is implied by the mixing condition of the process, and thus we need to work with the longspan assumption.

[4] If $\{X_s\}$ satisfies (A1) and (A2.i) and is strictly stationary, it is necessarily α-mixing (with an unknown rate). This is because a strictly stationary Markov process is β-mixing if and only if it is (Harris) recurrent and aperiodic (see, p. 157 of Chen *et al.*, 2010), and (A1) and (A2.i) imply the recurrency and aperiodicity of $\{X_s\}$.

3 Path Continuity of Diffusion Processes

In this section, we present some new results on the path continuity of diffusion processes. Since we suppose only the availability of discretely sampled data from $\{X_s\}$ (instead of its full and continuously-recorded trajectory), our estimators for continuous time processes incur biases due to discretization. To control the discretization biases and obtain the consistency of the estimators, we rely on the infill assumption $\Delta \to 0$, under which the effects due to the discretization are expected to be asymptotically negligible. Several approaches can verify its negligibility: for example, 1) computing the moments of the discretization biases; 2) using the (almost sure) path continuity based on the Kolmogorov-Čentsov criterion (see, e.g., Theorem 2.8 in p. 53 of KS91); and 3) using the Brownian/diffusion modulus of continuity (see the arguments subsequently). The first approach allows us to derive the sharpest rate, and it has been used in Florens-Zmirou (1989), Yoshida (1992), Kessler (1997), Nicolau (2003), Jacod (2006) and Phillips and Yu (2009), for example. However, it generally requires the existence of the higher-order moment. The second approach also requires the existence of the higher order moment, particularly for obtaining a sharper rate. It is applicable not only to diffusion processes and Brownian semimartingales but also to a wider class of general processes including ones based on the fractional Brownian motion (see Embrechts and Maejima, 2002). This approach has been used in Fan, Fan and Jiang (2007), and Kanaya and Kristensen (2015) for example.

In this study, we adopt the third approach, since it works without assuming the existence of the moment and it always gives us a sharper rate than the second approach. While the convergence rate obtained by our approach is inferior to that obtained by the first, its loss is only minor.[5] The Brownian/diffusion modulus of continuity states that the increments of the process, $|X_{s+\Delta} - X_s|$, are $O_{a.s.}(\sqrt{\Delta \log(1/\Delta)})$ uniformly over $s \in [0, T]$ (as $\Delta \to \infty$). This result is *local* in that it should be applicable to the case where $T = \bar{T} < \infty$. As previously mentioned, we work with the longspan assumption of $T \to \infty$, and this local result is not sufficient for our purpose. The modulus of continuity might still hold under the longspan ($T \to \infty$) or even globally over $s \in [0, \infty)$. However, such result has not been fully investigated. We subsequently discuss this point, present some new results on the global modulus of continuity, and clarify required conditions. Indeed, we have not been able to show that the modulus of continuity holds under the longspan or globally for general diffusion processes (with potentially unbounded $\mu(\cdot)$ and $\sigma^2(\cdot)$), and thus, we pursue an alternative approach. We show that *weighted* increments, $B(X_s)|X_{s+\Delta} - X_s|$, are $O_{a.s.}(\sqrt{\Delta \log(1/\Delta)})$. To discuss this point, we start by reviewing a classical result.

McKean's Classical Result: The following property of a diffusion process (as a solution to (1)) is well-known (McKean, 1969, pp. 46-47 and 96-97): If the process $\{X_s\}_{s \geq 0}$ is stopped at some fixed time $T = \bar{T} < \infty$, then there exists some random variable $C_T = O_{a.s.}(1)$ such that

$$\Pr[\limsup_{\Delta \searrow 0} \sup_{s,t \in [0,T];\ |t-s| \in (0,\Delta]} |X_t - X_s| / \sqrt{\Delta \log(1/\Delta)} \leq C_T] = 1, \tag{13}$$

where we write $C_T = O_{a.s.}(1)$ if and only if $C_T < \infty$ for each $\omega \in \Omega^*$ with some Ω^* such that $\Pr[\Omega^*] = 1$ (i.e., almost surely all $\omega \in \Omega$; we often use this notation in the sequel). In the estimation of diffusion/volatility functions, we can generally obtain the consistency under $T < \infty$ (without letting $T \to \infty$). Then, this local path-continuity property (13) ensures that the discretization biases of the diffusion estimators are negligible with the almost sure rate $\sqrt{\Delta \log(1/\Delta)}$ as $\Delta \to 0$ (see, e.g., Florens-Zmirou, 1993;

[5] In general, we obtain the discretization bias of order $\sqrt{\Delta}$ by using the first approach, but we have its order $\sqrt{\Delta \log(1/\Delta)}$ in the third approach. We also note that Jeong and Park (2014) consider an approach based on extremal processes.

Jiang and Knight, 1997; Bandi and Phillips, 2003; Xu, 2009). However, this may not hold under the longspan asymptotic scheme where $T \to \infty$, which is necessary for our uniform results. To the author's knowledge, the longspan/global counterpart of (13) has not been available in the literature.[6] The local result (13) depends up on the following two facts: (i) each path of the Brownian motion $\{W_s\}$ is uniformly continuous over $s \in [0,T]$ with the degree of continuity $\sqrt{\Delta \log(1/\Delta)}$ (in the almost sure sense); and (ii) $\sup_{s\in[0,T]} |\mu(X_s)|$ and $\sup_{s\in[0,T]} \sigma^2(X_s)$ are $O_{a.s.}(1)$. We note that (ii) is trivially satisfied if T is finite (as long as $\mu(\cdot)$ and $\sigma^2(\cdot)$ are continuous and $\{X_s\}$ is nonexplosive) but may not be so for various processes if $T \to \infty$, where we can generally consider the bound C_T in (13) as follows:

$$C_T := \sup_{s\in[0,T]} |\mu(X_s)| + \max\left\{1, \sup_{s\in[0,T]} \sigma^2(X_s)\right\}. \tag{14}$$

We refer to the proof of Theorem 1 to see how this bound can be derived (or McKean, 1969, pp. 96-97).

New Results on the Global Modulus of Continuity: Here, we show a version of (13) with allowing for $T \to \infty$. For obtaining such a new result, we need to tackle the two points (i) and (ii) mentioned in the previous paragraph. As for (i), we prove that the modulus of continuity of the Brownian motion $\{W_s\}$ actually holds globally over the infinite interval $[0,\infty)$:

$$\Pr[\limsup_{\Delta\searrow 0} \sup_{s,t\in[0,\infty);\ |t-s|\in(0,\Delta]} |W_t - W_s| / \sqrt{2\Delta\log(1/\Delta)} = 1] = 1, \tag{15}$$

which is stated formally with its proof in the Appendix. While this result does not seem to have been available in the literature, it can be obtained by a slight modification of the proof for the finite-interval case, where we use factorial rationals (instead of dyadic rationals) to construct partitions of a certain time interval.

An immediate consequence of (15) is the modulus of continuity of a Brownian martingale, that is, for each $\omega \in \Omega^*$ where Ω^* is an event satisfying $\Pr[\Omega^*] = 1$, there exists some $\tilde{\Delta} > 0$ such for any $\Delta \in [0, \tilde{\Delta}]$,

$$\sup_{s,t\in[0,\infty);\ |t-s|\in[0,\Delta]} \left|\int_s^t \rho_u dW_u\right| \le \sqrt{2\Delta\log(1/\Delta)\max\{1, \sup_{s\in[0,\infty)} \rho_s^2\}}, \tag{16}$$

where $\{\rho_s\}_{s\ge 0}$ is a uniformly bounded process (over $s \in [0,\infty)$) with which a (local) martingale process $\{M_s\}_{s\ge 0}$ through a stochastic integral $M_s := \int_0^s \rho_u dW_u$ is well-defined (e.g., if $\{\rho_s\}$ is also adapted and predictable, then such $\{M_s\}$ is well-defined); and $\sup_{s\in[0,\infty)} \rho_s^2 > 0$. This result (16) seems to have an independent interest, which can be a theoretical basis for jump thresholding (as in Mancini, 2009) under the longspan asymptotics (see Kanaya and Kristensen, 2015). The proof of this statement, which is provided in the Appendix, is based on (15) and the so-called time-change argument.

While the result (16) is often used in our subsequent proofs, it is not necessarily sufficient for our purpose to uniformly control discretization biases of the nonparametric estimators. To see this, note that if both the drift and diffusion functions, $\mu(x)$ and $\sigma^2(x)$, were bounded uniformly over $x \in I$, then (16) would imply the almost sure uniform continuity of $\{X_s\}$ with the degree of $\sqrt{\Delta\log(1/\Delta)}$. However, such uniform boundedness excludes many (parametric) models commonly used in the economics/finance literature. In particular, provided that I is unbounded as $\mathbb{R}$ or $(0,\infty)$, the boundedness of $\mu(x)$ and $\sigma^2(x)$ does not generally guarantee the stationarity/ergodicity of the process.[7] To handle the unboundedness of

[6] Note that the longspan is often necessary in identifying the drift function nonparametrically. It seems that some previous studies concerning the drift estimation simply assume that the modulus of continuity of diffusion processes holds globally over $s \in [0,\infty)$.

[7] It is known that any diffusion process (on $\mathbb{R}$) whose drift function is compactly supported and whose diffusion function is (uniformly) bounded is null recurrent (see, e.g., Has'minskiĭ, 1980, Ch. IV). See also discussions in Nicolau (2005) and Chen *et al.* (2010) on volatility-induced stationarity.

$\mu(x)$ and $\sigma^2(x)$, we introduce a technical devise of a *damping* function $B(x)(>0)$, and verify a modified version of the modulus of continuity with the damping function as a weight:

$$B(X_s)\,|\psi(X_{s+\Delta})-\psi(X_s)| = O_{a.s.}(\sqrt{\Delta\log(1/\Delta)}) \text{ uniformly over } s\in[0,T] \text{ as } T\to\infty, \tag{17}$$

by imposing some additional conditions on $B(\cdot)$, where $\psi(\cdot)$ is some function, such as $f(\cdot)$, $\mu(\cdot)$ or $\sigma^2(\cdot)$ ($f(x)=x$). Quantities of the form on the left-hand side (LHS) of (17) frequently appear in analyzing our nonparametric estimators. To obtain the result as in (17), we restrict a class of functions of $\psi(\cdot)$. Restrictions imposed on $\psi(\cdot)$ depend upon the damping function $B(\cdot)$, which the econometrician can choose arbitrarily, as well as the property of the underlying process $\{X_s\}$. To clarify such restrictions, we introduce the following conditions:

A3. **(i)** There exists some constant $p\geq 0$ such that $|x|\to\infty$, $|\mu(x)|=O(|x|^{p+1})$ and $\sigma^2(x)=O(|x|^{p+2})$. **(ii)** Let $\{\xi_T\}_{T\geq 0}$ be a sequence of positive real numbers, satisfying $\max_{s\in[0,T]}|X_s| = O_{a.s.}(\xi_T)$ as $T\to\infty$. Then, $\xi_T^p\sqrt{\Delta\log(1/\Delta)}=O(1)$ as $T\to\infty$ and $\Delta\to 0$.

B1. $B(\cdot)(:I\to(0,\infty))$ is twice differentiable; $\sup_{x\in I}B(x)<\bar{B}$ for some $\bar{B}\in(0,\infty)$; and

$$B(x)=O\left(\exp\{-c_1(\log|x|)^{1+c_2}\}\right) \text{ as } |x|\to\infty, \text{ for some } c_1,c_2>0. \tag{18}$$

The polynomial growth condition (A3.i) on $\mu(\cdot)$ and $\sigma^2(\cdot)$ is quite mild. While $p=0$ corresponds to the the classical linear condition for the existence of SDE solutions (e.g., Sec. 5.2 of Karatzas and Shreve, 1991), (A3.i) is much milder, allowing for any $p\geq 0$. For example, it allows for hyperbolic diffusion models (see Bibby and Sørensen, 1997, 2003), and models with volatility-induced stationarity (see Conley, Hansen, Luttmer and Scheinkman, 1997; Nicolau, 2005), where diffusion functions of these models are generally unbounded. Indeed the author does not know of a parametric diffusion model with its state space $I=\mathbb{R}$ (used in economics and finance) that would violate (A3.i).

(A3.ii) restricts the growing rate of the extremal/maximal process $\max_{s\in[0,T]}|X_s|$ through the shrinking rate of Δ, while no restriction on ξ_T is required for $p=0$. We can find similar conditions on the extremal processes in Aït-Sahalia and Park (2013), Jeong and Park (2014), and Kanaya and Kristensen (2015). While it is generally not an easy task to find the exact rates of extremal processes, they have been investigated in the literature (e.g., Davis, 1982; Borkovec and Klüpperlberg, 1998; Jeong and Park, 2014). Their results mainly imply O_p rates of extremal processes, but we conjecture that such results can be strengthened to a.s. results with some extra efforts. As an instructive example, we can check that the Brownian motion $\{W_s\}$ satisfies $\sup_{s\in[0,T]}|W_s| = o_{a.s.}(T^{1/2}\log T)$ as $T\to\infty$ (the proof of this result is provided in Kanaya and Kristensen, 2015).

The condition (B1) presents requirements for the damping function. Its tail needs to decay sufficiently fast: The decaying rate should be faster than any polynomial functions (but need not to be of an exponential order). In a numerical example in Section 11 in the Supplementary Material, we set $B(x)=\exp\{-cx^2\}$ (a scaled version of the standard-normal density) with some $c>0$. We also provide some discussions on a choice of $B(x)$ in Sections 5 and 11. We note that (A3.ii) may be removed at the price of having more rapidly decaying B. That is, we can verify all the subsequent theorems and related results without (A3.ii) if (B1) is replaced by

B1'. $B(x)$ satisfies the same conditions as those in (B1) but the tail-decay condition (18) is replaced by the following one: $B(x)=O(\exp\{-c_1|x|^{p+c_2}\})$ as $|x|\to\infty$, for some $c_1,c_2>0$, where $p\geq 0$ is the constant given in (A3.i).[8]

[8] The proof of Theorem 1 under this alternative condition is provided in the Appendix.

Given these conditions, we formally state the result (17) as the following theorem:

Theorem 1. *Suppose that $\{X_s\}$ is a solution to (1) with (A1.i) and (A2) satisfied, and (B1) holds. Let $\psi : I \to \mathbb{R}$ be a continuously differentiable function satisfying $|\psi'(x)| = O(|x|^q)$ as $|x| \to \infty$ for some $q \geq 0$ (i.e., the first derivative of ψ grows at most with a polynomial order). Then, there exists some positive-valued random variable $\check{C} = O_{a.s.}(1)$ satisfying*

$$\Pr\left[\limsup_{\Delta \searrow 0;\, T\to\infty} \sup_{s,t\in[0,T];\ |t-s|\in(0,\Delta]} B(X_s)\,|\psi(X_t) - \psi(X_s)|/\sqrt{\Delta \log(1/\Delta)} \leq \check{C}\right] = 1. \tag{19}$$

While the formal proof of this theorem is provided in the Appendix, we see its basic idea here. To this end, write the LHS of (17) as

$$B(X_s)\,|\psi(X_{s+\Delta}) - \psi(X_s)| = B\left(\gamma^{-1}(Y_s)\right)\,|\psi\left(\gamma^{-1}(Y_{s+\Delta})\right) - \psi\left(\gamma^{-1}(Y_s)\right)|,$$

where γ is some function which transform X_s into $Y_s = \gamma(X_s)$ ($\in I_\gamma = \left(\bar{l}, \bar{r}\right)$). E.g., we can set the range I_γ of $\{Y_s\}$ as a bounded set $(-1, 1)$, so that $\{Y_s\}$ is well behaved in that it is globally uniformly continuous with the degree of continuity $\sqrt{\Delta \log(1/\Delta)}$ (we use the result (16)). However, as the price for obtaining the well-behaved process $\{Y_s\}$, the tail behavior of $\psi(\gamma^{-1}(\cdot))$ may be aberrant since the slope of $\gamma^{-1}(y)$ is very steep as $y \to \bar{r}$ or $\bar{l}$ (this is the case if I_γ is bounded). To suppress/damp such tail behavior, we put $B(X_s)$, or equivalently $B(\gamma^{-1}(Y_s))$, as a weight.

We use the result of Theorem 1 to control the discretization bias of our nonparametric estimators. Its important implication is that a component such as $B(X_{j\Delta})\psi(X_s)$ is bounded almost surely as $n \to \infty$ and $\Delta \to 0$ uniformly, that is,

$$B(X_{j\Delta})\psi(X_s) = O_{a.s.}(1) \quad \text{uniformly over } j \in \{1, \ldots, n-1\} \text{ and } s \in [j\Delta, (j+1)\Delta]. \tag{20}$$

This holds because $B(X_{j\Delta})\psi(X_s)$ can be decomposed into a uniformly bounded part and the other negligible part:

$$\begin{aligned} B(X_{j\Delta})\,\psi(X_s) \;=\; & \left\{B(X_{j\Delta})\psi(X_j) + B(X_{j\Delta})\left[\psi(X_s) - \psi(X_{j\Delta})\right]\mathbf{1}_{\{B(X_{j\Delta})|\psi(X_s)-\psi(X_{j\Delta})|\leq\sqrt{\Delta}\log(1/\Delta)\}}\right\} \\ & + B(X_{j\Delta})\left[\psi(X_s) - \psi(X_{j\Delta})\right]\mathbf{1}_{\{B(X_{j\Delta})|\psi(X_s)-\psi(X_{j\Delta})|>\sqrt{\Delta}\log(1/\Delta)\}}. \end{aligned}$$

The first term on the right-hand side (RHS) is uniformly bounded, given the uniform boundedness of $B(x)\psi(x)$ (which is imposed in Definition 1 below). By Theorem 1, for each $\omega \in \Omega^*$ with $\Pr[\Omega^*] = 1$, we can find some $\tilde{\Delta}(= \tilde{\Delta}(\omega)) > 0$ such that for any $\Delta \leq \tilde{\Delta}$, $B(X_{j\Delta})|\psi(X_s) - \psi(X_{j\Delta})| \leq \check{C}\sqrt{\Delta \log(1/\Delta)} \leq \sqrt{\Delta}\log(1/\Delta)$. This means that the second term on the LHS of (20) almost surely converges to zero with an arbitrary fast rate (since it is exactly zero for any small Δ). From these, we can conclude (20).

4 General Convergence Results

We here present general convergence theorems used for deriving the convergence of the nonparametric estimators. To set out additional conditions on the damping function, we introduce the following class:

Definition 1. $\mathcal{D}(B, \pi)$ is a class of functions defined for each pair of the damping function $B(\cdot)$ (which satisfies (B1)) and the invariant density $\pi(\cdot)$ of the process. A function $\psi : I \to \mathbb{R}$ is said to belong to $\mathcal{D}(B, \pi)$ if it is continuously twice differentiable on I and satisfies the following conditions: **(i)** There exists some constant $\bar{B}_1 > 0$ such that $\sup_{x\in I} |B(x)\psi(x)| \leq \bar{B}_1$; **(ii)** There exists some constant $\bar{B}_2 > 0$ such that $\sup_{x\in I} \left|\left(d^k/dx^k\right) H(x)\right| \leq \bar{B}_2$ for $k = 0, 2$, where $H(x) := B(x)\psi(x)\pi(x)$.

This class $\mathcal{D}(B,\pi)$ is introduced to derive convergence results for the so-called smoothing-bias and variance terms of the nonparametric estimators. $\psi(\cdot)$ is typically set as the drift or diffusion function $\mu(\cdot)$, $\sigma^2(\cdot)$. In the sequel, we suppose that some relevant functions belong to $\mathcal{D}(B,\pi)$ (see the subsequent theorems' conditions). The condition (ii) in Definition 1 is a weakening of the stationary version of Andrews' (1995) condition in the discrete time setting. He assumes that $m(x)\pi(x)$ and its derivatives are uniformly bounded over $x \in \mathbb{R}$ (see Assumption NP3 in Andrews, 1995), where $m(x)$ is the conditional expectation function $E[Y_t|X_t = x]$ in a discrete time process $\{Y_t, X_t\}$ and $\pi(x)$ is the invariant density of $\{X_t\}$. Note that $m(x)$ needs not to be bounded if the tail decay of $\pi(x)$ is sufficiently fast. For example, his condition is satisfied if $\pi(x)$ decays geometrically and $m(x)$ grows at most polynomially (as $|x| \to \infty$). However, we may not be able to expect $\pi(x)$ to decay in that way with $m(x)\pi(x)$ uniformly bounded. In fact, the heavy-tailed distribution is often a key feature in a financial time series, where $\pi(x)$ approaches zero very slowly (only polynomially) as $|x| \to \infty$. Moreover, we can easily find several examples of parametric diffusion models (used in economics and finance) which violate a continuous-time-process analogue of Andrews' condition. For example, in a model with $\mu(x) = x/(1+x^2)$ and $\sigma^2(x) = 1 + x^2$ (Nicolau, 2005), neither of $\mu(x)\pi(x)$ nor $\sigma^2(x)\pi(x)$ is bounded. We can find many examples violate the analogue of Andrew's condition in a class of processes with volatility-induced stationarity. By having $B(\cdot)$, we allow for the case where the boundedness of $\mu(x)\pi(x)$ and/or $\sigma^2(x)\pi(x)$ is not satisfied.

Before presenting our convergence theorems, we provide a set of conditions for the kernel function:

B2. $K(\cdot)(:\mathbb{R}\to\mathbb{R})$ is of bounded variation and satisfies the following conditions: **(i)** $\int_{-\infty}^{\infty} K(x)\,dx = 1$ and $\int_{-\infty}^{\infty} xK(x)\,dx = 0$; **(ii)** There exists some $\bar{K} \in (0,\infty)$ such that $\sup_{x\in\mathbb{R}} |K(x)| \leq \bar{K}$ and $\int_{-\infty}^{\infty} x^2 |K(x)|\,dx \leq \bar{K}$.

(B2) allows for most of symmetric kernels, including the normal kernel and the polynomial kernels while excluding the Dirichlet kernel $K(x) = \sin(x)/(\pi x)$. We do not require the continuity and may choose the uniform kernel, which is excluded in Hansen (2008). It also allows for so-called higher-order (bias-reducing) kernels as considered first in Bartlett (1963). For simplicity, we only consider the second-order kernels (and, as a result, the smoothing biases of the nonparametric estimators are of order h^2). However, when imposing appropriate conditions on the differentiability of relevant functions, the use of the higher-order kernel also leads to the faster convergence rates in our case as in Andrews (1995) and Hansen (2008). Bosq (1998) and Hansen (2008) assume the tail decay of the kernel function is sufficiently fast (e.g., the conditions of Corollary 2.2 in Bosq, 1998; Assumption 3 and the conditions of Theorem 4 in Hansen, 2008). They use the tail decay assumption, together with the condition on the existence of the higher order moment and the Markov inequality, to show that the outside of the expanding compact set is asymptotically negligible (or to use the so-called truncation argument). In contrast, (B2) does not impose any condition on the tail of the kernel except for the one implied by the integrability. The flexibility in the choice of the kernel is a benefit by using the covering-number technique. Note that we subsequently impose some continuity and compact-support conditions for obtaining a sharp convergence rate of the diffusion function estimator (Theorem 5), while such conditions are not required for a less sharp rate (Theorem 10 in the Supplementary Material)

Convergence Theorems: We here present two convergence theorems for components which constitute the functional estimators for diffusion processes. First, we investigate the uniform convergence rate of an object of the following form:

$$G_{n,T}(x) := (1/T)\sum_{j=1}^{n-1} K_h(X_{j\Delta} - x)\, B(X_{j\Delta}) \int_{j\Delta}^{(j+1)\Delta} \psi(X_s)\,ds, \tag{21}$$

to the expectation of its discretized version $\bar{G}_{n,T}(x) := E[K_h(X_\Delta - x) B(X_\Delta) \psi(X_\Delta)]$, where we note that $G_{n,T}(x)$ may be regarded as an estimator of $H(x) = B(x)\psi(x)\pi(x)$:

Theorem 2. *Suppose that $\{X_s\}$ is a solution to (1) with (A1)-(A3) and $\sup_{x\in I}\pi(x) < \infty$. Let $B(\cdot)$ be the damping function satisfying (B1), and $\psi(\cdot) \in \mathcal{D}(B,\pi)$ with its first derivative satisfying $|\psi'(x)| = O(|x|^q)$ as $|x| \to \infty$ for some $q \geq 0$. Suppose also that the kernel function $K(\cdot)$ satisfies (B2), and there exists some $\theta \in (0,1)$ such that as $T \to \infty$ and $h \to 0$,*

$$(\log T)/T^\theta h \to 0; \tag{22}$$

$$\beta \geq 5(1+\theta)/(1-\theta). \tag{23}$$

Then, it holds that as $n, T \to \infty$ and $\Delta, h \to 0$,

$$\sup_{x\in I}\left|G_{n,T}(x) - \bar{G}_{n,T}(x)\right| = O_p(\sqrt{\Delta\log(1/\Delta)}) + O_p(\sqrt{(\log T)/Th}). \tag{24}$$

The two terms on the RHS of (24) correspond to the discretization bias and the variance-effect component. Note that Bandi and Phillips (2003) assume that $\sqrt{\Delta\log(1/\Delta)}/h \to 0$ to control the discretization bias, whose order is given by $\sqrt{\Delta\log(1/\Delta)}/h$. In contrast, we only require $\sqrt{\Delta\log(1/\Delta)} \to 0$ for the discretization bias to be negligible, where h is not involved and we can work with weaker conditions on h and Δ. This relaxation of the conditions is made possible by considering a *fully* discretized process as $G_{n,T,2}(x) := (1/n)\sum_{j=1}^{n-1} K_h(X_{j\Delta} - x) B(X_{j\Delta})\psi(X_{j\Delta})$ for deriving the rate of the variance-effect component. In this case, the discretization bias corresponds to $G_{n,T,1}(x) := (1/T)\sum_{j=1}^{n-1} K_h(X_{j\Delta} - x) B(X_{j\Delta}) \times \int_{j\Delta}^{(j+1)\Delta}[\psi(X_s) - \psi(X_{j\Delta})]ds$, where we note that $G_{n,T}(x) = G_{n,T,1}(x) + G_{n,T,2}(x)$. We can also think of an alternative decomposition of $G_{n,T}(x)$, based on a fully continuous-time process, which is convenient to consider a wider class of nonstationary diffusion processes but in general requires some stronger conditions on the rates of h and Δ. For details, see Section 3.2 of Kanaya (2015).

The condition (23) requires that the exponent of the mixing coefficients β is necessarily larger than 5 (since $\theta \in (0,1)$). We can still derive the convergence result even when $\beta \in (0,5]$, while its rate is slower than that in Theorem 2 (i.e., for smaller β, the second term on the RHS of (24) is replaced by $O_p(\sqrt{(\log T)/T^\theta h})$). Previous studies have only considered the case where β is sufficiently large (see the previous arguments concerning the condition (A2.ii)), and uniform results for small β seem to be new and are presented as Theorem 7 in the Supplementary Material.

The rate $\sqrt{(\log T)/Th}$ in (24) is an infill counterpart of that presented in Hansen (2008, Theorem 2) for discrete time processes, where he derived $\sqrt{(\log n)/nh}$. Our rate is necessarily slower than Hansen's rate due to the infill assumption $\Delta(= T/n) \to 0$. Under the infill asymptotics, the dependence between the consecutive observations becomes stronger as $n \to \infty$, and therefore, we achieve a slower rate. Hansen's rate of $\sqrt{(\log n)/nh}$ is optimal for discrete time processes, which is the same as Stone's (1982) optimal rate for independent and identically distributed (i.i.d.) data. While it seems that no study has investigated the uniform optimal rate for the continuous time case with the infill and longspan assumptions, we conjecture that the rate in (24) is optimal for an object of the form (21) in such cases (note that if Δ were fixed, then our rate should coincide with Hansen's rate since $T = \Delta n = O(n)$). For the diffusion function estimator $\hat{\sigma}^2(x)$, we can derive the optimal rate of $\sqrt{(\log n)/nh}$ under the infill and longspan (in Theorem 5), but this is because $\hat{\sigma}^2(x)$ consists of some components whose variance property is different from that of $G_{n,T}(x)$ (see decomposition into five terms in the proof of Theorem 5). We also note that the condition imposed on β is different from that in Hansen (2008). This is because we use the covering-number-based

technique in Lemma 2, while he uses arguments based on the truncation of variables, expanding compact sets and the Markov inequality (see the proof of Theorem 2 in Hansen, 2008). Some other remarks on Theorem 2 are in order:

Remark 1. (i) If the mixing coefficients have geometric decay, i.e., $\alpha(s) \leq \tilde{A}\exp\{-\tilde{\beta}s\}$ for some $\tilde{\beta}$, $\tilde{A} > 0$ (corresponding to $\beta = \infty$ in (A2.ii)), then the result of Theorem 2 holds under $(\log T)/Th \to 0$ instead of $(\log T)/T^{\theta}h \to 0$ (this result can be proven analogously, whose proof is omitted for brevity). The same/similar remark also applies to Theorems 4, 5 and 10.
(ii) If $\psi = 1$, $G_{n,T}(x) = \hat{\Pi}(x)$, where $\hat{\Pi}(x)$ is the estimator of the *damped* invariant density $\Pi(x) = B(x)\pi(x)$, defined in (8). In this case, we do not need to consider the discretization bias. Accordingly, we can drop $O_p(\sqrt{\Delta\log(1/\Delta)})$ on the RHS of (24).

As seen in (1), the increments of diffusion processes consist of two components, the conditional mean part and the martingale (difference) one. The previous theorem concerns an object involving the former component. As a general form involving the latter component, we consider the following quantity:

$$M_{n,T}(x) := (1/T)\sum\nolimits_{j=1}^{n-1} K_h(X_{j\Delta} - x)\int_{j\Delta}^{(j+1)\Delta} \rho_s dW_s,$$

where $\{\rho_s\}_{s\geq 0}$ is an adapted and predictable process uniformly bounded over $s\,(\geq 0)$ with which a stochastic integral $\int_0^t \rho_s dW_s$ is well-defined for any $t > 0$. The next theorem derives the convergence property of $M_{n,T}(x)$:

Theorem 3. *Suppose that $\{X_s\}$ is a solution to (1) with (A1)-(A2) and $\sup_{x\in I}\pi(x) < \infty$, and the kernel function K satisfies (B2). Then, as $n, T \to \infty$ and $\Delta, h \to 0$, it holds that for each $a(> 0)$ large enough, and for each $x \in I$,*

$$\Pr[|M_{n,T}(x)| \geq a\sqrt{(\log T)/Th}] \leq 2T^{-a^2/2} + T^{-aC_M} + 4AT^{-\beta}h^{-(\beta+1)}(\log T)^{1-\beta}, \tag{25}$$

where $C_M\,(> 0)$ is some constant independent of x, n, T and a.

For generality, we write this theorem without explicitly specifying the form of ρ_s, which is set as $B(X_s)\sigma(X_s)$ for proving the convergence of the drift estimator, for example. (25) is based on the exponential inequality for continuous martingales and that for mixing processes, where we note that $M_{n,T}(x)$ can be represented as a continuous martingale. The application of the former inequality requires some boundedness condition on the quadratic variation process of $M_{n,T}(x)$. The quadratic variation in fact grows as $T \to \infty$ and is not bounded, but we can control its growing rate by the latter inequality.

Note also that Theorems 2-3 require the uniform boundedness of the invariant density π. While this requirement does not seem restrictive for the case $I = \mathbb{R}$, we can find some processes with an unbounded π in particular when the either of the end points is bounded, say $I = (0, \infty)$ or $[0, \infty)$. Even in such unbounded cases, we can still verify the convergence results as in Theorems 2-3, given slight modifications of the condition and a suitable choice of B (see discussions in the Supplementary Material).

5 Uniform Convergence Rates of Nadaraya-Watson Type Estimators

In this section, we present two theorems on the uniform convergence rates for the drift and diffusion estimators (4) and (5), respectively. We require the following conditions:

Assumption 1. *$\{X_s\}$ is a solution to (1) satisfying (A1)-(A3) with the state space $I = \mathbb{R}$. Let $B(\cdot)$ be the damping function satisfying (B1) and $K(\cdot)$ be the kernel function satisfying (B2).*

Theorem 4 (Drift Function Estimation)**.** *Suppose that Assumption 1 holds; $\sup_{x\in\mathbb{R}} |\pi(x)| < \infty$; the observation interval Δ and the bandwidth h satisfy*

$$\Delta^{-1} = O(T^{\kappa}) \quad \text{and} \quad (\log T)/T^{\theta}h \to 0, \tag{26}$$

as $T \to \infty$ and $\Delta, h \to 0$, for some constants $\kappa\,(>0)$ and $\theta \in (0,1)$;

$$\mu(\cdot) \in \mathcal{D}(B,\pi)\text{; and } |\partial\mu(x)| + |\partial\sigma(x)| = O(|x|^{\bar{q}_1}) \text{ as } |x| \to \infty \text{ for some } \bar{q}_1 \geq 0. \tag{27}$$

Let

$$a^{\star}_{n,T} := h^2 + \sqrt{\Delta \log(1/\Delta)} + \sqrt{(\log T)/Th}. \tag{28}$$

Then, if $\beta \geq \max\{5(1+\theta)/(1-\theta), (3\theta+2+2\kappa)/(1-\theta)\}$, it holds that as $n, T \to \infty$ and $\Delta, h \to 0$,

$$\sup_{x\in\mathbb{R}} |\hat{\Psi}_{\mu}(x) - B(x)\mu(x)\pi(x)| = O_p(a^{\star}_{n,T}); \tag{29}$$

and further if $a^{\star}_{n,T}/\delta_{n,T} \to 0$,

$$\sup_{|x|\leq c_T} |\hat{\mu}(x) - \mu(x)| = O_p(a^{\star}_{n,T}/\delta_{n,T}), \tag{30}$$

where $c_{n,T}$ is any sequence tending to infinity (as $n, T \to \infty$), and $\delta_{n,T} := \inf_{|x|\leq c_{n,T}} B(x)\pi(x) > 0$.

Theorem 5 (Diffusion Function Estimation)**.** *Suppose that Assumption 1 holds; $\max_{k=0,1,2}\sup_{x\in\mathbb{R}} |\partial^k \pi(x)| < \infty$; the observation interval Δ and the bandwidth h satisfy*

$$\Delta^{-1} = O(n^{\varkappa}), \;\; (\log n)/n^{\vartheta}h = O(1) \quad \text{and} \quad nh^5/(\log n) = O(1), \tag{31}$$

as $n \to \infty$ and $\Delta, h \to 0$, for some constants $\varkappa \in (0, 1/2)$ and $\vartheta \in (0,1)$ with $(1-\vartheta-\varkappa) > 0$; the kernel function K is Lipschitz continuous (i.e., $|K(u) - K(v)| \leq C_K|u-v|$ for some $C_K > 0$), whose support is included in $[-c_K, c_K]$ with some $c_K > 0$; $\sigma^2(\cdot)$ is three-times continuously differentiable;

$$\begin{gathered}\sigma^2(\cdot) \in \mathcal{D}(B,\pi)\text{; } |\partial\mu(x)| + \textstyle\sum_{k=0}^{3} |\partial^k\sigma(x)| = O(|x|^{\bar{q}_2}) \text{ as } |x| \to \infty \text{ for some } \bar{q}_2 \geq 0; \\ \max_{k=1,2} |\partial^k B(x)| < C_B \times B(x) \quad \text{for some } C_B > 0.\end{gathered} \tag{32}$$

Let $a^{\blacklozenge}_{n,T} := h^2 + \Delta\log(1/\Delta) + \sqrt{(\log n)/nh}$. Then, if $\beta > \max\{(2+3\vartheta-2\varkappa)/(1-\vartheta-\varkappa), 2/(1-2\varkappa)\}$, it holds that as $n, T \to \infty$ and $\Delta, h \to 0$,

$$\sup_{x\in\mathbb{R}} |\hat{\Psi}_{\sigma^2}(x) - \hat{\Pi}(x)\sigma^2(x)| = O_p(a^{\blacklozenge}_{n,T}); \tag{33}$$

and further if $a^{\blacklozenge}_{n,T}/\delta_{n,T} \to 0$,

$$\sup_{|x|\leq c_T} |\hat{\sigma}^2(x) - \sigma^2(x)| = O_p(a^{\blacklozenge}_{n,T}/\delta_{n,T}),$$

where $c_{n,T}$ and $\delta_{n,T}$ are sequences defined in Theorem 4.

The form we consider in (29) for the drift estimation is different from the one in (33) for the diffusion estimation. Instead of the expression in (33), we can also consider the uniform convergence of $|\hat{\Psi}_{\sigma^2}(x) - B(x)\sigma^2(x)\pi(x)|$ as in (33) (see Theorem 10 in the Supplementary Material), which can lead to the uniform rate of $\left|\hat{\sigma}^2(x) - \sigma^2(x)\right|$ (with the penalty $\delta_{n,T}^{-1}$) as well. The rate we can derive for $|\hat{\Psi}_{\sigma^2}(x) - \hat{\Pi}(x)\sigma^2(x)|$ is faster than that for $|\hat{\Psi}_{\sigma^2}(x) - B(x)\sigma^2(x)\pi(x)|$ (compare the form of $\mathcal{V}_5$ with that of U_2 in the proofs of Theorems 5 and 4, respectively). However, we need to work with some stronger conditions for the former, where we impose the continuity and the compact-support conditions of the kernel K as well as the boundedness conditions in the derivatives of the density, which are not required for Theorem 10. This is because the potential unboundedness of $\sigma^2(x)$ can be directly suppressed by the multiplication of $B(x)$ in $|\hat{\Psi}_{\sigma^2}(x) - B(x)\sigma^2(x)\pi(x)|$ while it is not so in $|\hat{\Psi}_{\sigma^2}(x) - \hat{\Pi}(x)\sigma^2(x)|$, where we note that the boundedness of the latter component can be induced to some extent by the combination of the damping function and the compactly-supported kernel.

The growing condition of Δ^{-1} in (26) or (31), controlled through the restriction on κ or $\varkappa$, is required mainly to control the rate of the martingale components of the estimators. This can be significantly relaxed for the diffusion estimation case: We only require Δ^{-1} to grow at most with *some* polynomial rate of T (Theorem 10 in the Supplementary Material), which essentially impose no restriction on the growing rate of Δ^{-1} while it results in a slower convergence rate.

The sequence $a_{n,T}^{\blacklozenge}$ which determines the rate for the diffusion estimation involves $\sqrt{(\log n)/nh}$. If we let $h = O([(\log n)/n]^{1/5})$ (i.e., $\vartheta = 1/5$) and $\varkappa \in (2/5, 1/2)$, then we can obtain $a_{n,T}^{\blacklozenge} = [(\log n)/n]^{2/5}$ for β large enough, which is known to be Stone's optimal rate for the i.i.d. case with estimation objects being twice differentiable. The pointwise convergence rate of the diffusion function estimator is $1/\sqrt{nh}$, and this is obtained only under the infill without the longspan $T \to \infty$ (see arguments/results in Bandi and Phillips, 2003). However, Theorem 5 requires the longspan since $\Delta^{-1} = O(n^{\varkappa})$ for $\varkappa < 1$ (note that we must have $\varkappa = 1$ if the fixed span $T = \bar{T} < \infty$ were assumed). The pointwise asymptotic normality/distribution results for $\hat{\mu}(x)$ and $\sigma^2(x)$ rely on the central limit theorems for martingales (cf. the second term on the RHS of the SDE (1)). While our uniform results also exploit the martingale property of the process (as in Theorem 3), we need to show uniform convergence of some terms which are not martingales. To show uniform convergence of such terms, we apply the Bernstein exponential inequality for mixing processes, exploiting the asymptotic independence implied by the mixing, and this is the reason why we need the longspan "$T \to \infty$" even for the diffusion estimation case (see also discussions after Theorem 2). It is uncertain if we can derive a sharp uniform convergence rate of the diffusion estimator as $a_{n,T}^{\blacklozenge}$ under the fixed span case.

Penalized Uniform Rates: The uniform rates of $\hat{\mu}(x)$ and $\hat{\sigma}^2(x)$ are penalized by $\delta_{n,T}^{-1}$, which are of the ratio types. The rate of $\delta_{n,T}$ depends upon the tail thickness/thinness of $B(x)\pi(x)$. By the nature of the damping, the tail of $B(x)\pi(x)$ is necessarily thinner than that of $\pi(x)$. For some sequence of $c_{n,T}$, the growing rate of $\delta_{n,T}^{-1}$ may be very fast. However, given the uniform boundedness of $\pi(\cdot)$, we may be able to use the following modified estimators in order to avoid such the fast growing $\delta_{n,T}^{-1}$:

$$\check{\mu}(x) := \hat{\Psi}_{\mu}(x)/B(x)\hat{\pi}(x); \quad \text{and} \quad \check{\sigma}^2(x) := \hat{\Psi}_{\sigma^2}(x)/B(x)\hat{\pi}(x),$$

where $\hat{\pi}(x)$ is the estimator of the invariant density: $\hat{\pi}(x) := (\Delta/T)\sum_{j=1}^{n} K_h(X_{j\Delta} - x)$. Let $\bar{\delta}_{n,T} := \inf_{|x|\leq \bar{c}_T} \pi(x) (> 0)$. Then, corresponding to Theorems 4 and 5,

$$\sup_{|x|\leq \bar{c}_T} B(x)|\check{\mu}(x) - \mu(x)| = O_p\left(a_{n,T}^{\star}/\bar{\delta}_T\right); \quad \text{and} \quad \sup_{|x|\leq \bar{c}_T} B(x)|\check{\sigma}^2(x) - \sigma^2(x)| = O_p\left(a_{n,T}^{\blacklozenge}/\bar{\delta}_T\right).$$

These convergence results with the weighted sup-norm may be similar to that found in Chen and Fan (2006), who consider the uniform convergence of the empirical distribution function.

In applying our uniform results to semiparametric estimation, we may not necessarily take into account the penalty due to $\delta_{n,T}$ or $\bar{\delta}_{n,T}$. For example, Ai (1997) and Coslette (2004) consider the trimming device in their semiparametric problems to establish distribution theory. They use the ratio-type nonparametric estimators to construct the objective function for the estimation of a finite dimensional parameter, where very small values of the nonparametric estimator in the denominator are trimmed. Then, a sequence used for trimming may be chosen as irrelevant to the tail decay of the object in the denominator (in our case, it may be chosen independently from the tail thinness of $B(x)\pi(x)$). This argument seems to apply to various semiparametric problems, where the uniform convergence results for the estimators in the numerator and the denominator may be sufficient.

Effects of the Damping Function: Here, we briefly consider the choice of the damping function and its effects on the estimators. One way is to select the form of $B(\cdot)$ so that it is only effective in the region where only few observations exist. In this case, there is practically no effect due to damping. This may reflect the view that the damping function is *only* a *technical* device for establishing theoretical properties.

Alternatively, we may be able to select a form of the damping function through minimizing mean-squared errors (MSE; or some other objective) of the estimators with respect to $B(\cdot)$, as the Epanechnikov kernel is obtained as the MSE-optimal one (Epanecnikov, 1969). However, it seems generally difficult to select such an optimal B. To see this point, observe that an approximation (leading component) of the pointwise bias of the estimator $\hat{f}(x) = \hat{\mu}(x)$ or $\hat{\sigma}^2(x)$ can be obtained as

$$\mathcal{B}_f(x) = h^2 \left\{ (d/dx)[B(x) f(x)] \times \frac{\pi'(x)}{B(x)\pi(x)} + \frac{(d^2/dx^2)[B(x) f(x)]}{2B(x)} \right\} \int z^2 K(z)\, dz,$$

where $f = \mu$ or σ^2; and $\mathcal{B}_\mu(x)$ is an approximated bias of $\hat{\mu}(x)$ and $\mathcal{B}_{\sigma^2}(x)$ is that of $\hat{\sigma}^2(x)$. While the choice of $B(\cdot)$ affect the bias expressions, it has no effect on the (asymptotic) variances (since we have some cancellation between the numerator and denominator; see the Supplementary Material for details). Therefore the MSE minimization problem is reduced to minimizing $\mathcal{B}_\mu(x)$ or $\mathcal{B}_{\sigma^2}(x)$, given *unknown* π, and μ or σ^2, with respect to B (probably with some regularization restriction such as $\int B(x)\, dx = 1$). While one can select the optimal kernel independent of the underlying process' structure, the optimal B (if it could be found) depends up on unknown components. We may be able to develop a two-step procedure to estimate some *optimal* damping function based on preliminary estimators of $\mu(\cdot)$ and $\sigma^2(\cdot)$, but we leave it to future research.

Finally, we provide some graphical illustration of the finite-sample effects of the damping function in the Supplementary Material, which compares the standard NW estimator and our damped one. Our graphical result seems to suggest that the damping function B does not have significant effects, as its effects are cancelled out between numerator and denominator parts.

Extensions: We here briefly discuss two possible extensions of our results. First, while our current results are on univariate diffusion processes, we can derive uniform convergence rates of nonparametric estimators of multivariate diffusion processes without significant changes in proof arguments. We note that the arguments on the degree of path continuity in Section 3 hold true irrelevant of the univariate or multivariate settings, upon a suitable choice of a multivariate version of a damping function, and the exponential inequalities for mixing processes and continuous martingales can still hold for the multivariate

case.[9] We expect the orders of the smoothing and discretization biases in the multivariate case are the same as those in the univariate case, i.e., the first two terms in $a^{\star}_{n,T}$ and $a^{\blacklozenge}_{n,T}$. In contrast, the orders of variance effect terms should be changed to $(\log T)/\sqrt{Th^d}$ and $(\log n)/\sqrt{nh^d}$ for drift and diffusion estimators, respectively, where d is the dimension of the process. These slower convergence rates for high-dimension cases are analogous to those in the standard discrete-time setting as in Hansen (2008).

Second, we can also think of an extension to general non-ergodic/non-stationary cases. A key is the use of Bernstein-type exponential inequalities, where we have used those for mixing processes and martingales. The mixing property does not necessarily hold for non-ergodic diffusion processes, and accordingly we cannot use the exponential inequality as in Lemma 3, while the martingale-based inequality can be still used. Instead, we may exploit implications of the Markov property of general diffusion processes. That is, under the assumption of recurrence, we can split a process $\{X_s\}_{s\in[0,T]}$ into blocks which possess some independent property, and remaining negligible parts, by using a continuous-time counterpart of Nummelin's (1984) Markov regeneration/splitting method as in Fukasawa (2008), Löcherbacha and Loukianova (2008), and Loukianova and Loukianov (2008). Then, we can apply the exponential inequality for independent/mixing processes to the constructed blocks. Nummelin's method has been used in Gao, Kanaya, Li and Tjøstheim (2015) for deriving sharp uniform convergence rates of kernel-based estimators in a discrete-time setting. We expect that its continuous-time version can be effectively used to derive uniform convergence rates of drift and diffusion function estimators, while the application of the exponential inequality requires establishing some uniform moment bounds of the continuous-time processes (e.g., ones corresponding to Lemmas B.1-B.3 of Gao *et al.*, 2015), which may not be trivial. We note that the order of discretization biases which we can derive through these techniques should be $\sqrt{\Delta \log(1/\Delta)}/h$ (see discussions after Theorem 2), and variance effects' rates should be slower than those in the ergodic/mixing case, as found in Gao *et al.* (2015).

6 Concluding Remark

We have proposed to use the damping function device for establishing the uniform convergence results of the NW type estimators of the diffusion processes. Using the damping function and the covering-number technique allows us to work with quite mild conditions on the underlying processes and the kernel functions. Our results should be useful in various estimation and testing problems for diffusion processes. Note that the same idea/method may also be applied to the discrete time setting, where we can significantly relax various restrictions imposed in previous studies to obtain uniform convergence results of the nonparametric kernel estimators. This idea will be pursued in future studies.

References

[1] Ai, C. (1997) A semiparametric maximum likelihood estimator, *Econometrica* 65, 933-963.

[2] Aït-Sahalia, Y. (1996a) Nonparametric pricing of interest rate derivative securities, *Econometrica* 64, 527-560.

[3] Aït-Sahalia, Y. (1996b) Testing continuous-time models of the spot interest rate, *Review of Financial Studies* 9, 385-426.

[9] The upper bound of the covering number of a multivariate scaled kernel can be also easily calculated if it takes a product form $\Pi_{i=1}^{d} K(z_i)$ (see the proof of Theorem 9.15 of Kosorok, 2008).

[4] Aït-sahalia, Y. (1999) Transition densities for interest rate and other nonlinear diffusions, *Journal of Finance* 54, 1361-1395.

[5] Aït-Sahalia, Y. & J.Y. Park (2013) Bandwidth selection and asymptotic properties of local nonparametric estimators in possibly nonstationary continuous-time models. Working paper, Indiana University.

[6] Andrews, D.W.K. (1994a) Empirical process methods in econometrics, *Handbook of Econometrics*, Vol. 4, Ch. 37, 2247-2294, Ed. by R.F. Engle and D. McFadden, North-Holland.

[7] Andrews, D.W.K. (1994b) Asymptotics for semiparametric econometric models via stochastic equicontinuity, *Econometrica* 62, 43-72.

[8] Andrews, D.W.K. (1995) Nonparametric kernel estimation for semiparametric models, *Econometric Theory* 11, 560-596.

[9] Ango Nze, P. & P. Doukhan (2004) Weak dependence: Models and applications to econometrics, *Econometric Theory* 20, 995-1045.

[10] Bandi, F.M. & P.C.B. Phillips (2003) Fully nonparametric estimation of scholar diffusion models, *Econometrica*, 71, 241-283.

[11] Bartlett, M.S. (1963) Statistical estimation of density functions, *Sankhya A* 25, 245-254.

[12] Bibby, B.M. & M. Sørensen (1997) A hyperbolic diffusion model for stock prices, *Finance Stochastics* 1, 25-41.

[13] Bibby, B.M. & M. Sørensen (2003) Hyperbolic processes in finance, *Handbook of Heavy Tailed Distributions in Finance*, S. Rachev (ed.), Ch.6, 211 - 248, Elsevier Science.

[14] Bierens, H. J. (1983) Uniform consistency of kernel estimators of a regression function under generalized conditions, *Journal of the American Statistical Association* 78, 699-707.

[15] Bosq, D. (1998) *Nonparametric Statistics for Stochastic Processes*, 2nd edition, Springer.

[16] Borkovec, N.H. & C. Klüpperlberg (1998) Extremal behaviour of diffusion models in finance, Extremes 1, 47-80.

[17] Bouezmarni, T. & O. Scaillet (2005) Consistency of asymmetric kernel density estimators and smoothed histograms with application to income data, *Econometric Theory* 21, 390-412.

[18] Chapman, D.A. & Pearson, N.D. (2003) Is the short rate drift actually nonlinear? *Journal of Finance* 55, 355-388.

[19] Chen, X. & Y. Fan (2006) Estimation of copula-based semiparametric time series models, *Journal of Econometrics* 130, 307-335.

[20] Chen, X., L.P. Hansen & M. Carrasco (2010) Nonlinearity and temporal dependence, *Journal of Econometrics* 155, 155-169.

[21] Conley, T., L.P. Hansen, E. Luttmer & J. Scheinkman (1997) Short-term interest rates as subordinated diffusions, *The Review of Financial Studies* 10, 525-577.

[22] Corradi, V. & W. Distaso (2010) Diagnostic tests for volatility models, Working paper, University of Warwick.

[23] Cosslett, S.R. (2004) Efficient estimation of censored and truncated regressions via a smoothed self-consistency equation, *Econometrica* 72, 1277-1293.

[24] Davis, R.A. (1982) Maximum and minimum of one-dimensional diffusions, *Stochastic Processes and their Applications* 13, 1-9.

[25] Doukhan, P. (1994) *Mixing: Properties and Example*, Lecture Notes in Statistics 85, Springer-Verlag, New York.

[26] Embrechts, P & M. Maejima (2002) *Selfsimilar Processes*, Princeton University Press.

[27] Epanecnikov, V.A. (1969) Non-parametric estimation of a multivariate probability density, *Theory of Probability and its Applications* 14, 153-158.

[28] Escanciano, J.C., D.T. Jacho-Chávez, A. Lewbel (2014) Uniform convergence of weighted sums of non- and semi-parametric residuals for estimation and testing, *Journal of Econometrics* 178, 426-443.

[29] Fan, J., Y. Fan & J. Jiang (2007) Dynamic integration of time- and state domain methods for volatility estimation, *Journal of the American Statistical Association* 102, 618-631.

[30] Fan, J. & Q. Yao (2003) *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer-Verlag.

[31] Fan, J. & C. Zhang (2003) A reexamination of diffusion estimators with applications to financial Model validation, *Journal of the American Statistical Association* 98, 118-134.

[32] Florens-Zmirou, D. (1989) Approximate discrete-time schemes for statistics of diffusion processes, *Statistics* 4, 547-557.

[33] Florens-Zmirou, D. (1993) On estimating the diffusion coefficient from discrete observations, *Journal of Applied Probability* 30, 790-804.

[34] Forman, J.Y. & M. Sørensen (2008) The Pearson diffusions: A class of statistically tractable diffusion processes, *Scandinavian Journal of Statistics* 35, 438-465.

[35] Fukasawa, M. (2008) Edgeworth expansion for ergodic diffusions, Probability Theory and Related Fields 142, 1-20.

[36] Gao, J., S. Kanaya, D. Li & D. Tjøstheim (2015) Uniform consistency for nonparametric estimators in null recurrent time series, *Econometric Theory 31, 911-952.*

[37] Gospodinov, N. & M. Hirukawa (2012) Nonparametric estimation of scalar diffusion processes of interest rates using asymmetric kernels, *Journal of of Empirical Finance 19, 595-609.*

[38] Working Paper 08-011, Department of Economics, Concordia University.

[39] Jacod, J. (2006) Parametric inference for discretely observed non-ergodic diffusions, *Bernoulli* 12, 383-401.

[40] Jeong, M. & J. Park (2010) Asymptotic theory of maximum likelihood estimator for diffusion model, Working Paper, Indiana University.

[41] Jiang, G.J. & J.L. Knight (1997) A Nonparametric approach to the estimation of diffusion processes, with an application to a short-term interest rate model, *Econometric Theory* 13, 615-645.

[42] Has'minskiĭ, R.Z. (1980) *Stochastic Stability of Differential Equations,* Stijthoff & Noordoff International Publishers B.V., Alphen aan denRijn.

[43] Hansen, B. E. (2008) Uniform convergence rates for kernel estimation with dependent data, *Econometric Theory* 24, 726-748.

[44] Hansen, L.P. & J.A. Scheinkman (1995) Back to the future: generating moment implication for continuous-time Markov processes, *Econometrica* 63, 767-804.

[45] Hansen, L.P., J.A. Scheinkman & N. Touzi (1998) Spectral methods for identifying scalar diffusions, *Journal of Econometrics* 86, 1-32.

[46] Ichimura, H. & P.E. Todd (2007) Implementing nonparametric and semiparametric estimators, *Handbook of Econometrics* Vol. 6-2, Ch. 74, 5369-5468, Ed. by J.J. Heckman and E.E. Leamer, North-Holland.

[47] Ikeda, N. & S. Watanabe (1981) *Stochastic Differential Equations and Diffusion Processes,* Kodansha with North-Holland.

[48] Kallenberg, O. (2002) *Foundations of Modern Probability*, 2nd Ed., Springer.

[49] Kanaya, S. (2014) Non-parametric specification testing for continuous-time markov processes: Do the processes follow diffusions? Working paper, University of Aarhus.

[50] Kanaya, S. (2015) Convergence rates of sums of α-mixing triangular arrays: with an application to non-parametric drift function estimation of continuous-time processes, Working paper, University of Aarhus.

[51] Kanaya, S. & D. Kristensen (2014) Optimal sampling and bandwidth selection for kernel estimators of diffusion processes, Working paper, University of Aarhus and University College London.

[52] Kanaya, S. & D. Kristensen (2015) Estimation of stochastic volatility models by nonparametric filtering, forthcoming in *Econometric Theory*.

[53] Karatzas, I. & S.E. Shereve (1991) *Brownian Motion and Stochastic Calculus*, Springer.

[54] Karlin, S. & H.M. Taylor (1981) *A Second Course in Stochastic Processes*, Academic Process, New York.

[55] Kent, J. (1978) Time-reversible diffusions, *Advances in Applied Probability* 10, 819-835.

[56] Kessler, M. (1997) Estimation of diffusion processes from discrete observations, *Scandinavian Journal of Statistics* 24, 211-229.

[57] Kong, E., O. Linton & Y. Xia (2010) Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model, *Econometric Theory* 26, 1529-1564.

[58] Koo, B. & O. Linton (2012) Semiparametric estimation of locally stationary diffusion models, *Journal of Econometrics* 170, 210-23.

[59] Kosorok, M. (2008) *Introduction to Empirical Processes and Semiparametric Inference*, Springer.

[60] Kristensen, D. (2008) Estimation of partial differential equation with applications in finance, *Journal of Econometrics* 144, 392-408.

[61] Kristensen, D. (2009) Uniform convergence rates of kernel estimators with heterogeneous dependent data, *Econometric Theory* 25, 1433-1445.

[62] Kristensen, D. (2010a) Pseudo-maximum likelihood estimation in two classes of semiparametric diffusion models, *Journal of Econometrics* 156, 239-259.

[63] Kristensen, D. (2010b) Nonparametric filtering of the realized spot volatility: A kernel-based approach, *Econometric Theory* 26, 60-93.

[64] Kusuoka, S. & N. Yoshida (2000) Malliavian calculus, geometric mixing, and expansion of diffusion functionals, *Probability Theory and Related Fields* 116, 457-484.

[65] Kutoyants, Y.A. (1999) Efficient density estimation for ergodic diffusion processes, *Statistical Inference for Stochastic Processes* 1, 131-155.

[66] Li, F. (2007) Testing the parametric specification of the diffusion function in a diffusion process, *Econometric Theory* 23, 221-250.

[67] Liebscher, E. (1996) Strong convergence of sums of α-mixing random variables with application to density estimation, *Stochastic Processes and their Applications* 65, 69-80.

[68] Löcherbacha, E. & D. Loukianova (2008) On Nummelin splitting for continuous time Harris recurrent Markov processes and application to kernel estimation for multi-dimensional diffusions, *Stochastic Processes and their Applications* 118, 1301-1321.

[69] Loukianova, D. & O. Loukianov (2008) Deterministic equivalents of additive functionals of recurrent diffusions and drift estimation, *Statistical Inference for Stochastic Processes* 11, 107-121.

[70] Mancini, C. (2009) Non-parametric Threshold Estimation for Models with Stochastic Diffusion Coefficient and Jumps, *Scandinavian Journal of Statistics* 36, 270-296.

[71] Masry, E. (1996) Multivariate local polynomial regression for time series: uniform strong consistency and rates, *Journal of Time Series Analysis* 17, 571-599.

[72] Mandl, P (1968) *Analytical Treatment of One-dimensional Markov Processes*, Springer-Verlag.

[73] McKean, H.P. (1969) *Stochastic Integrals*, Academic Press.

[74] Nicolau, J. (2003) Bias reduction in nonparametric diffusion coefficient estimation, *Econometric Theory* 19, 754-777.

[75] Nicolau, J. (2005) Processes with volatility-induced stationarity: an application for interest rates, *Statistica Neerlandica* 59, 376-396.

[76] Nolan, D. & D. Pollard (1987) U-processes: rates of convergence," *The Annals of Statistics*, 15, 780-799.

[77] Nummelin, E. (1984) *General Irreducible Markov Chains and Non-Negative Operators*, Cambridge University Press.

[78] Pagan, A. & A. Ullah (1999) *Nonparametric Econometrics*, Cambridge University Press.

[79] Pakes, A. & D. Pollard (1989) Simulation and the asymptotics of optimization estimators, *Econometrica* 57, 1027-1057.

[80] Parzen, E. (1962) On estimation of a probability density function and mode, *Annals of Mathematical Statistics* 33, 1065-1076.

[81] Phillips, P.C.B. & J. Yu (2009) A two-stage realized volatility approach to estimation of diffusion processes with discrete data, *Journal of Econometrics* 150, 139-150.

[82] Pollard, D. (1984) *Convergence of Stochastic Processes*, Springer-Verlag.

[83] Revuz, D. & M. Yor (1999) *Continuous Martingales and Brownian Motion*, 3rd Ed. (corrected 3rd printing, 2005), Springer.

[84] Rio, E. (1995) The functional law of the iterated logarithm for stationary strongly mixing sequences, *Annals of Probability* 23, 1188-1203.

[85] Robinson, P.M. (1988) Root-N-consistent semiparametric regression, *Econometrica* 56, 931-954.

[86] Rogers, L.C.G. & D. Williams (2000) *Diffusions, Markov Processes and Martingales*, Vol. 2, Cambridge University Press.

[87] Stanton, R. (1997) A nonparametric model of term structure dynamics and the market price of interest rate risk, *Journal of Finance 52*, 1973-2002.

[88] Stone, C.J. (1982) Optimal global rates of convergence for nonparametric regression, *Annals of Statistics* 10, 1040-1053.

[89] Van der Vaart, A.W. (1998) *Asymptotic Statistics*, Cambridge University Press.

[90] Van der Vaart, A.W. & J. A. Wellner (1996) *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer.

[91] Van Zanten, J.H. (2000) On the uniform convergence of the empirical density of an ergodic diffusion, *Statistical Inference for Stochastic Processes* 3, 251-262.

[92] Veretennikov, A.Y. (1987) Bounds for the mixing rate in the theory of stochastic equations, *Theory of Probability and its Applications* 32, 273-281.

[93] Veretennikov, A.Y. (1997) On polynomial mixing bounds for stochastic differential equations, *Stochastic Processes and their Applications* 70, 115-127.

[94] Veretennikov, A.Y. (1999) On polynomial mixing and convergence rate for stochastic difference and differential equations, *Theory of Probability and its Applications* 44, 361-374.

[95] Xu, K.-L. (2009) Empirical likelihood based inference for recurrent nonparametric diffusions, *Journal of Econometrics* 153, 65-82.

[96] Xu, K.-L. (2010) Re-weighted functional estimation of diffusion models, *Econometric Theory* 26, 541-563.

[97] Yoshida, N. (1992) Estimation for diffusion processes from discrete observation, *Journal of Multivariate Analysis* 41, 220-242.

7 Appendix

7.1 Proofs of Path Continuity Results

In this section, we provide proofs of results on the path continuity of a diffusion process, as well as several auxiliary results. We start with the global modulus of continuity of a Brownian motion:

Theorem 6. *Let $\{W_s\}_{s\geq 0}$ be a standard Brownian motion on a filtered probability space $(\Omega, \mathfrak{F}, \{\mathfrak{F}_s\}_{s\geq 0}, \Pr)$ which satisfies the usual conditions. Then, the global Brownian modulus of continuity (15) holds.*

Proof of Theorem 6. Our proof resembles that of Theorem 2.9.25 in Karatzas and Shreve (1991) where s, t are supposed to take values in some finite interval $[0, T]$ ($T \leq \bar{T}$ fixed), but extends it to the case with an infinite interval $[0, \infty)$. For notational simplicity, define $g(\delta) := \sqrt{2\delta \log(1/\delta)}$ for $\delta > 0$.

First, we prove $\underline{\limsup \geq 1}$. Look at

$$\Pr[\max_{1\leq j\leq m!} \left|W_{jm/m!} - W_{(j-1)m/m!}\right| \leq (1-\theta)^{1/2} g(m/m!)] = (1-\zeta)^{m!} \leq \exp\{-\zeta m!\}, \tag{34}$$

where $\zeta := 2\Pr[(m/m!)^{-1/2} W_{m/m!} \leq x]$ and $x := \sqrt{(1-\theta)2\log(m!/m)}$. The inequality in (34) holds since $(1-\zeta)^{m!} \leq \exp\{-\zeta m!\}$. From the inequality: $\int_x^\infty e^{-u^2/2} du \geq x e^{-x^2/2}/(1+x^2)$, we have

$$\zeta \geq 2x[\sqrt{2\pi}(1+x^2)]^{-1} \exp\{-x^2/2\} \geq C(1-\theta)^{1/2} [1 + 2\log(m!/m)]^{-1} (m/m!)^{1-\theta}$$

with some appropriate constant $C > 0$, which implies that

$$\Pr\left[\max_{1\leq j\leq m!} \left|W_{jm/m!} - W_{(j-1)m/m!}\right| \leq (1-\theta)^{1/2} g(m/m!)\right] \leq \exp\left\{\frac{-C(1-\theta)^{1/2}}{1+2\log(m!/m)} m^{1-\theta} (m!)^{\theta}\right\}.$$

By the Borel-Cantelli lemma, there exists an event Ω_θ with $\Pr(\Omega_\theta) = 1$ such that for any $\omega \in \Omega_\theta$, $\exists M_\theta$, $\forall m \geq M_\theta$,

$$[1/g(m/m!)] \max_{1\leq j\leq m!} \left|W_{jm/m!} - W_{(j-1)m/m!}\right| > (1-\theta)^{1/2}.$$

Letting $\Delta = m/m!$, we have for any $\omega \in \Omega_\theta$, $\exists M_\theta$, $\forall m \geq M_\theta$,

$$[1/g(\Delta)] \max_{s,t\in[0,\infty);\ |t-s|\leq\Delta} |W_t - W_s| \geq [1/g(m/m!)] \max_{1\leq j\leq m!} \left|W_{jm/m!} - W_{(j-1)m/m!}\right|$$

and thus, for any $\omega \in \Omega_\theta$,

$$\limsup_{\Delta\searrow 0} [1/g(\Delta)] \max_{s,t\in[0,\infty);\ |t-s|\leq\Delta} |W_t - W_s| \geq (1-\theta)^{1/2}.$$

We can conclude $\underline{\limsup \geq 1}$ by letting $\theta \searrow 0$.

For the proof of $\underline{\limsup \leq 1}$, let $\theta \in (0,1)$ and $\varepsilon > (1+\theta)(1-\theta)^{-1} - 1$. Observe the following inequalities:

$$\begin{aligned}
&\Pr\left[\max_{0\leq i<j\leq m!;\ k=j-i\leq(m!/m)^\theta} \tfrac{1}{g(km/m!)} \left|W_{jm/m!} - W_{im/m!}\right| > 1+\varepsilon\right] \\
&\leq \sum\nolimits_{k=1}^{\lceil (m!/m)^\theta \rceil} \Pr\left[\max_{0\leq i<i+k\leq m!} \left|W_{(k+i)m/m!} - W_{im/m!}\right| > (1+\varepsilon) g\left(\tfrac{km}{m!}\right)\right] \\
&\leq m! \sum\nolimits_{k=1}^{\lceil (m!/m)^\theta \rceil} \Pr\left[\frac{|W_{km/m!}|}{\sqrt{km/m!}} > (1+\varepsilon)\sqrt{2\log\left(\tfrac{m!}{km}\right)}\right] \\
&\leq \frac{m!}{\sqrt{2\pi}} \sum\nolimits_{k=1}^{\lceil (m!/m)^\theta \rceil} k^{(1+\varepsilon)^2} \left(\frac{m}{m!}\right)^{(1+\varepsilon)^2} \\
&\leq (1/\sqrt{2\pi}) m! \left(\frac{m}{m!}\right)^{(1+\varepsilon)^2} \frac{\left[(m!/m)^\theta + 1\right]^{1+(1+\varepsilon)^2}}{1+(1+\varepsilon)^2} \leq \text{const.} \times \frac{m^{(1-\theta)(1+\varepsilon)^2-\theta}}{(m!)^\rho},
\end{aligned}$$

where $\lceil x \rceil$ denotes the largest integer that is less than or equal to x; $\rho = (1-\theta)(1+\varepsilon)^2 - (1+\theta) > 0$; and the third and fourth inequalities hold since $\int_x^\infty e^{-u^2/2}du \leq \frac{1}{x}e^{-x^2/2}$ and $\sum_{k=1}^{\lceil (m!/m)^\theta \rceil} k^{(1+\varepsilon)^2} \leq \int_0^{(m!/m)^\theta+1} x^{(1+\varepsilon)^2} = \frac{\left[(m!/m)^\theta+1\right]^{1+(1+\varepsilon)^2}}{1+(1+\varepsilon)^2}$, respectively. By the Borel-Cantelli lemma, $\exists M_\theta$ such that

$$(M_\theta/M_\theta!)^{1-\theta} \leq 1/e; \tag{35}$$

$$\forall m \geq M_\theta, \quad \max_{0\leq i<j\leq m!;\ k=j-i\leq (m!/m)^\theta} \frac{1}{g(km/m!)} \left|W_{jm/m!} - W_{im/m!}\right| \leq 1+\varepsilon. \tag{36}$$

Now, fix $\omega \in \Omega_\theta$ and $m \geq M_\theta$, and define $E_l^m := \{km/l! \mid k = 0, 1, \ldots, l!\}$. We prove the following lemma (whose proof is given below):

Lemma 1.

$$|W_t - W_s| \leq (1+\varepsilon)\left[2\sum_{j=m+1}^{l} J^2 g\left(\frac{J+1}{(J+1)!}\right) + g\left(\max\left\{|t-s|, m/(l-1)!\right\}\right)\right] \tag{37}$$

is valid for every pair (s,t) *in* $\forall t, s \in E_l^m = \{km/l! \mid k = 0, 1, \ldots, l!\}$ *with* $0 < |t-s| < (m!/m)^\theta \times m/m! = (m/m!)^{1-\theta}$.

Given this lemma, we define $E^m := \bigcup_{l=1}^\infty E_l^m$ and let $l \to \infty$ in (37). Then,

$$\begin{array}{l} \forall t, s \in E^m \text{ with } 0 < |t-s| < \lceil \theta m \rceil! \times m/m!; \\ |W_t - W_s| \leq (1+\varepsilon)\left[2\sum_{j=m+1}^{\infty} J^2 g\left(\frac{J+1}{(J+1)!}\right) + g\left(|t-s|\right)\right]. \end{array} \tag{38}$$

For any s, t satisfying $s, t \in E_m$ for some m and $0 < |t-s| < (M_\theta/M_\theta!)^{1-\theta}$, we can select $\tilde{m}\,(\geq M_\theta)$ that satisfies $s, t \leq \tilde{m}$ and

$$\left\{(\tilde{m}+1)/(\tilde{m}+1)!\right\}^{1-\theta} \leq \Delta\,(:= t-s) < (\tilde{m}/\tilde{m}!)^{1-\theta}. \tag{39}$$

Since g is increasing on $(0, 1/e]$, we have

$$\sum_{j=\tilde{m}+1}^\infty J^2 g\left(\frac{J+1}{(J+1)!}\right) \leq \sum_{j=\tilde{m}}^\infty J^2 g\left(\frac{J+1}{(J+1)!}\right) \leq C\tilde{m}^2 g\left(\frac{\tilde{m}+1}{(\tilde{m}+1)!}\right) \leq \frac{C\tilde{m}^2}{\sqrt{1-\theta}}\left(\frac{\tilde{m}+1}{(\tilde{m}+1)!}\right)^{\theta/2} g(\Delta),$$

for some constant $C > 0$. We conclude from (38), (39) and the continuity of $W_\cdot(\omega)$ that for every $\omega \in \Omega_\theta$ and $\tilde{m} \geq M_\theta$,

$$\frac{1}{g(\Delta)} \sup_{0\leq s<t\leq m;\ t-s=\Delta} |W_t - W_s| \leq (1+\varepsilon)\left[\frac{2C\tilde{m}^2}{\sqrt{1-\theta}}\left(\frac{\tilde{m}+1}{(\tilde{m}+1)!}\right)^{\theta/2} + 1\right] \tag{40}$$

holds for any $\Delta \in [\{(\tilde{m}+1)/(\tilde{m}+1)!\}^{1-\theta}, (\tilde{m}/\tilde{m}!)^{1-\theta})$. Since the strict increasingness of g on $(0, 1/e]$ and the continuity of the Brownian path, we may replace the condition $t - s = \Delta$ by $t - s \leq \Delta$. Since (40) holds with any m, we have

$$\sup_{s,t\in[0,\infty);\ |t-s|\leq\Delta} \frac{1}{g(\Delta)} |W_t - W_s| \leq (1+\varepsilon)\left[\frac{2C\tilde{m}^2}{\sqrt{1-\theta}}\left(\frac{\tilde{m}+1}{(\tilde{m}+1)!}\right)^{\theta/2} + 1\right]$$

for all $\Delta \in [\{(\tilde{m}+1)/(\tilde{m}+1)!\}^{1-\theta}, (\tilde{m}/\tilde{m}!)^{1-\theta})$. Letting $\tilde{m} \to \infty$, we obtain

$$\limsup_{\Delta\searrow 0} \sup_{s,t\in[0,\infty);\ |t-s|\leq\Delta} |W_t - W_s|/g(\Delta) \leq (1+\varepsilon).$$

Finally, by letting $\theta \searrow 0$ and hence simultaneously $\varepsilon \searrow 0$ along the rationals, we establish that

$$\limsup_{\Delta\searrow 0} \sup_{s,t\in[0,\infty);\ |t-s|\leq\Delta} |W_t - W_s|/g(\Delta) \leq 1,$$

as desired. □

Proof of Lemma 1. Note that g is strictly increasing on $(0, 1/e]$. To show (37), we use the inductive method. First, for $l = m$, (37) follows from (36), since $|W_t - W_s| \leq (1+\varepsilon)\, g(|t-s|)$ for $s, t \in E_m^m$ with $|t-s| < (m/m!)^{1-\theta}$. Second, suppose that (37) is valid for $l = m+1, \ldots, L-1$. For $s, t \in E_L^m$ with $s < t$ and $0 < t - s < (m/m!)^{1-\theta}$, consider the numbers $s_1 := \min\left\{u \in E_{L-1}^m : u \geq s\right\}$ and $t_1 := \max\left\{u \in E_{L-1}^m : u \leq t\right\}$, and notice the relationships $s_1, t_1 \in E_{L-1}^m \subset E_{L+1}^{L+1}$; $s, t \in E_L^m \subset E_{L+1}^{L+1}$; $s_1 - s \leq m/(L-1)!$; and $t - t^1 \leq m/(L-1)!$. Then, by the inequality (36),

$$|W_{s_1}(\omega) - W_s(\omega)| \leq mL(1+\varepsilon)\, g((L+1)/(L+1)!)\,;$$
$$|W_t(\omega) - W_{t_1}(\omega)| \leq mL(1+\varepsilon)\, g((L+1)/(L+1)!)\,.$$

There are two possible relationships among s, t, s_1 and t_1: (i) If $|t-s| \geq m/(L-1)!$, it holds that $s \leq s_1 \leq t_1 \leq t$ (with at least one inequality strict); (ii) If $|t-s| < m/(L-1)!$, either of $|t-s| < |s_1 - t_1| = m/(L-1)!$ or $|s_1 - t_1| = 0$. Noting that $\theta \in (0,1)$, we have

$$|t_1 - s_1| \leq \max\{|t-s|, m/(L-1)!\} \leq \max\{|t-s|, (m/m!)^{1-\theta}\} \leq (m/m!)^{1-\theta}$$

with at least one inequality strict. Thus,

$$|W_{t_1}(\omega) - W_{s_1}(\omega)| \leq (1+\varepsilon)\left[2\textstyle\sum_{J=m+1}^{L-1} J^2 g\left(\frac{J+1}{(J+1)!}\right) + g(\max\{|t-s|, m/(L-1)!\})\right],$$

by the induction assumption with $l = L-1$. By the triangle inequalities,

$$\begin{aligned}
&|W_t(\omega) - W_s(\omega)| \leq 2mL(1+\varepsilon)\, g((L+1)/(L+1)!) \\
&+ (1+\varepsilon)\left[2\textstyle\sum_{J=m+1}^{L-1} J^2 g\left(\frac{J+1}{(J+1)!}\right) + g(\max\{|t-s|, m/(L-1)!\})\right] \\
&\leq (1+\varepsilon)\left[2\textstyle\sum_{J=m+1}^{L} J^2 g\left(\frac{J+1}{(J+1)!}\right) + g(\max\{|t-s|, m/(L-1)!\})\right].
\end{aligned}$$

Now, we have shown (37) for any $l\,(> m)$, as desired. □

Proof of the Statement (16). Define a process $\{M_s\}_{s\geq 0}$ by $M_s := \int_0^s \rho_u dW_u$. Since $\{M_s\}$ is a local martingale, its quadratic variation process $\{\mathbf{t}(s)\}$ is given by $\mathbf{t}(s) := \int_0^s \rho_u^2 du$, which satisfies $\int_0^s \rho_u^2 du < \infty$ almost surely for each $s \in [0,\infty)$ since $\{\rho_s\}$ is uniformly bounded. By the representation theorem for continuous local martingales (see Sec. 7 of Ikeda and Watanabe, 1981), we have a time-changed Brownian motion $\{\tilde{W}_s\}$ such that $\tilde{W}_{\mathbf{t}(t)} = M_t$ almost surely, where we consider an enlarged probability space if necessary (i.e., if $\mathbf{t}(\infty) < \infty$; see Theorem 7.2' of Ikeda and Watanabe, 1981). Now, by Theorem 6, it almost surely holds that

$$1 = \limsup_{\Delta \searrow 0} \sup_{\substack{\mathbf{t}(s), \mathbf{t}(t) \in [0,\infty); \\ |\mathbf{t}(s) - \mathbf{t}(t)| \in (0, \Delta]}} \frac{|\tilde{W}_{\mathbf{t}(t)} - \tilde{W}_{\mathbf{t}(s)}|}{\sqrt{2\Delta \log(1/\Delta)}} \geq \limsup_{\Delta \searrow 0} \sup_{\substack{s,t \in [0,\infty); \\ |t-s| \max\{1, \sup_{u\in[0,\infty)} \rho_u^2\} \leq \Delta}} \frac{|\tilde{W}_{\mathbf{t}(t)} - \tilde{W}_{\mathbf{t}(s)}|}{\sqrt{2\Delta \log(1/\Delta)}}, \tag{41}$$

where the inequality follows from the fact that if $|t-s| \max\{1, \sup_{u\in[0,\infty)} \rho_u^2\} \leq \Delta$, then $|\mathbf{t}(t) - \mathbf{t}(s)| = \int_s^t \rho_u^2 du \leq \Delta$; and that $u \in [0,\infty)$ implies that $\mathbf{t}(u) \in [0,\infty)$ by the non-explosiveness of $\{\mathbf{t}(s)\}$. Then, it almost surely holds that

$$\begin{aligned}
&\sqrt{\max\{1, \sup_{u\in[0,\infty)} \rho_u^2\}} \\
&\geq \limsup_{\Delta \searrow 0} \sup_{s,t \in [0,\infty);\, |t-s| \leq \tilde{\Delta}} \frac{|\tilde{W}_{\mathbf{t}(t)} - \tilde{W}_{\mathbf{t}(s)}|}{\sqrt{2[\Delta / \max\{1, \sup_{u\in[0,\infty)} \rho_u^2\}] \log(\max\{1, \sup_{u\in[0,\infty)} \rho_u^2\}/\Delta)}} \\
&= \limsup_{\tilde{\Delta} \searrow 0} \sup_{s,t \in [0,\infty);\, |t-s| \leq \tilde{\Delta}} \frac{|\tilde{W}_{\mathbf{t}(t)} - \tilde{W}_{\mathbf{t}(s)}|}{\sqrt{2\tilde{\Delta} \log(1/\tilde{\Delta})}} = \limsup_{\tilde{\Delta} \searrow 0} \sup_{s,t \in [0,\infty);\, |t-s| \leq \tilde{\Delta}} \frac{|M_t - M_s|}{\sqrt{2\tilde{\Delta} \log(1/\tilde{\Delta})}},
\end{aligned} \tag{42}$$

where $\tilde{\Delta} = \Delta / \max\{1, \sup_{u\in[0,\infty)} \rho_u^2\}$; the first inequality follows from (41) and the fact that $\max\{1, \sup_{u\in[0,\infty)} \rho_u^2\} \geq 1$; and the equalities hold since "$\Delta \to 0$" is equivalent to "$\tilde{\Delta} \to 0$" for each $\omega \in \Omega^*$ with $\Pr[\Omega^*] = 1$ (as $\sup_{u\in[0,\infty)} \rho_u^2 < \infty$) and $\tilde{W}_{\mathbf{t}(t)} = M_t$ almost surely. Now, letting Δ denote the time interval (instead of $\tilde{\Delta}$), (42) means that

$$\Pr[\limsup_{\Delta \searrow 0} \sup_{s,t\in[0,\infty);\ |t-s|\leq\Delta} |\textstyle\int_s^t \rho_u dW_u| / \sqrt{2\Delta \log(1/\Delta)} \leq \sqrt{\max\{1, \sup_{u\in[0,\infty)} \rho_u^2\}}] = 1,$$

leading to the desired result. □

Given the result (16), we are ready to prove our main result on the weighted version of the global modulus of continuity:

Proof of Theorem 1. We first consider the case where $p > 0$. Let $\gamma : I \to (-1, 1)$ be a strictly increasing function:

$$\gamma(x) = -1 + |x|^{-p} \text{ if } x < -1; \quad = \gamma_m(x) \text{ if } x \in [-1, 1]; \ = 1 - x^{-p} \text{ if } x > 1, \tag{43}$$

where $\gamma_m(\cdot)$ is a bridging function on $[-1, 1]$ which is picked so that $\gamma(\cdot)$ is strictly increasing and twice continuously differentiable (such $\gamma_m(\cdot)$ can be constructed by some polynomial function). Then, by the Ito lemma,

$$\gamma(X_t) - \gamma(X_s) = \textstyle\int_s^t \left[\gamma'(X_u)\mu(X_u) + \gamma''(X_u)\sigma^2(X_u)/2\right] du + \int_s^t \gamma'(X_u)\sigma(X_u) dW_u.$$

By the construction of $\gamma(\cdot)$ and the growth condition of μ and σ in (A3), $[\gamma'(x)\mu(x) + \gamma''(x)\sigma^2(x)/2]$ and $[\gamma'(x)\sigma(x)]$ are uniformly bounded. Thus, by the result (16), there exists some random variable $\tilde{C}_\gamma$ such that for each $\omega \in \Omega^*$ with $\Pr(\Omega^*) = 1$,

$$\sup_{s,t\in[0,\infty);\ |t-s|\in(0,\Delta]} |\gamma(X_t) - \gamma(X_s)| \leq \tilde{C}_\gamma \sqrt{\Delta \log(1/\Delta)}, \tag{44}$$

for any Δ small enough, where we note that $\tilde{C}_\gamma$ is bounded for each $\omega(\in \Omega^*)$ and is independent of Δ. Now, look at

$$\begin{aligned}
&\sup_{s,t\in[0,T];\ |t-s|\in(0,\Delta]} B(X_s) |\psi(X_t) - \psi(X_s)| \\
&= \sup_{s,t\in[0,T];\ |t-s|\in(0,\Delta]} B\left(\gamma^{-1}(\gamma(X_s))\right) \times \left|\psi\left(\gamma^{-1}(\gamma(X_t))\right) - \psi\left(\gamma^{-1}(\gamma(X_s))\right)\right| \\
&= \sup_{s,t\in[0,T];\ |t-s|\in(0,\Delta]} B\left(\gamma^{-1}(\gamma(X_s))\right) |\eta([1-\lambda_{s,t}]\gamma(X_s) + \lambda_{s,t}\gamma(X_t))| \times \sup_{s,t\in[0,\infty);\ |t-s|\in(0,\Delta]} |\gamma(X_t) - \gamma(X_s)|
\end{aligned} \tag{45}$$

where $\eta(y) = (d/dy)\psi(\gamma^{-1}(y)) = \psi'(\gamma^{-1}(y))/\gamma'(\gamma^{-1}(y))$; and $\lambda_{s,t}$ is some random variable whose values are in $[0, 1]$ (which depends on s and t). Since the second term on the RHS of (45) is $O_{a.s.}(\sqrt{\Delta \log(1/\Delta)})$, we can obtain the desired result if we show that the first term is $O_{a.s.}(1)$.

To show the first term's $O_{a.s.}(1)$-boundedness, note that given the a.s. growth rate of the extremal process specified in (A3), for each $\omega \in \Omega^*$ with $\Pr(\Omega^*) = 1$, we can find some $c = c(\omega)$ such that $|X_s|, |X_t| \leq c\xi_T$ (for $s, t \leq T$) and thus, $|\gamma(X_s)|, |\gamma(X_t)| \leq 1 - (c\xi_T)^{-p}$. Given (B.2), we can find some continuous function $\tilde{B}(x)$ such that $B(x) \leq \tilde{B}(x)$ for any $x \in \mathbb{R}$, and $\tilde{B}(x) = |x|^{-(p+q+1)}$ for $|x| \geq c_B$

with c_B sufficiently large (independent of $\omega \in \Omega^*$). Then, the first term on the RHS of (45) can be bounded as

$$\begin{aligned}
&\sup_{s,t\in[0,T];\ |t-s|\in(0,\Delta]} B\left(\gamma^{-1}\left(\gamma\left(X_s\right)\right)\right)\left|\eta\left(\left[1-\lambda_{s,t}\right]\gamma\left(X_s\right)+\lambda_{s,t}\gamma\left(X_t\right)\right)\right| \\
\leq &\sup_{|y|,|w|\in[0,(1-(c\xi_T)^{-p}];\ |y-w|\leq\tilde{C}_\gamma\sqrt{\Delta\log(1/\Delta)}} \tilde{B}\left(\gamma^{-1}\left(y\right)\right)\left|\eta\left(w\right)\right| \\
\leq &\sup_{|y|,|w|\in[0,(1-(c\xi_T)^{-p}];\ |y-w|\leq\tilde{C}_\gamma\sqrt{\Delta\log(1/\Delta)}} \frac{\tilde{B}\left(\gamma^{-1}\left(y\right)\right)}{\tilde{B}\left(\gamma^{-1}\left(w\right)\right)} \times \sup_{w\in(-1,1)} \tilde{B}\left(\gamma^{-1}\left(w\right)\right)\left|\eta\left(w\right)\right| =: \mathbf{B}_1 \times \mathbf{B}_2. \quad (46)
\end{aligned}$$

To bound the term $\mathbf{B}_1$, we consider some random variable $\bar{c} = \bar{c}(\omega) > 0$ that is close enough to 1 (but with with $\bar{c} < (1-(c\xi_T)^{-p})$ so that $\gamma^{-1}(z) = (1-z)^{-1/p}$ for $z \in [\bar{c}, (1-(c\xi_T)^{-p}]$, where we can find such $\bar{c}$ independent of T when T is large enough. Then,

$$\begin{aligned}
\mathbf{B}_1 \leq &\sup_{y,w\in[-\bar{c},\bar{c}];\ |y-w|\leq\tilde{C}_\gamma\sqrt{\Delta\log(1/\Delta)}} \frac{\tilde{B}\left(\gamma^{-1}\left(y\right)\right)}{\tilde{B}\left(\gamma^{-1}\left(w\right)\right)} + \sup_{y,w\in[\bar{c},(1-(c\xi_T)^{-p}];\ |y-w|\leq\tilde{C}_\gamma\sqrt{\Delta\log(1/\Delta)}} \frac{\tilde{B}\left(\gamma^{-1}\left(y\right)\right)}{\tilde{B}\left(\gamma^{-1}\left(w\right)\right)} \\
&+ \sup_{y,w\in[-1+(c\xi_T)^{-p},-\bar{c}];\ |y-w|\leq\tilde{C}_\gamma\sqrt{\Delta\log(1/\Delta)}} \frac{\tilde{B}\left(\gamma^{-1}\left(y\right)\right)}{\tilde{B}\left(\gamma^{-1}\left(w\right)\right)} =: \mathbf{B}_{11} + \mathbf{B}_{12} + \mathbf{B}_{13}.
\end{aligned}$$

By the continuity of $\tilde{B}$ and $\gamma^{-1}(\cdot)$, as well as by the compactness of the domain of y and w for each $\omega \in \Omega^*$, we can check $\mathbf{B}_{11} = O_{a.s.}(1)$. To bound $\mathbf{B}_{12}$, look at

$$\begin{aligned}
&\mathbf{B}_{12} = \sup_{|y|,|w|\in[0,(1-(c\xi_T)^{-p}];\ |y-w|\leq\tilde{C}_\gamma\sqrt{\Delta\log(1/\Delta)}} \left\{\left(1-w\right)/\left(1-y\right)\right\}^{1+(q+1)/p} \\
\leq\ &\sup_{|y|,|w|\in[0,(1-(c\xi_T)^{-p}];\ |y-w|\leq\tilde{C}_\gamma\sqrt{\Delta\log(1/\Delta)}} \left\{1+\left|y-w\right|/\left(1-y\right)\right\}^{1+(q+1)/p} \\
\leq\ &\left\{1+\tilde{C}_\gamma\sqrt{\Delta\log\left(1/\Delta\right)}\times\left(c\xi_T\right)^p\right\}^{1+(q+1)/p} = O_{a.s.}\left(1\right),
\end{aligned}$$

where the last equality follows from the condition (A.3). By an analogous argument, we can also show $\mathbf{B}_{13} = O_{a.s.}(1)$. As for the term $\mathbf{B}_2$, we note that $1/|\gamma'(x)|$ and $|\psi'(x)|$ are bounded by a $(p+1)$-th polynomial function and a q-th polynomial one respectively, and therefore, for some $\bar{C} > 0$,

$$\mathbf{B}_2 \leq \sup_{|x|\leq c_B} \tilde{B}\left(x\right)\left|\psi'\left(x\right)/\gamma'\left(x\right)\right| + \sup_{|x|\geq c_B}\left|x\right|^{-(p+q+1)}\bar{C}\left[1+\left|x\right|^{p+q+1}\right] < \infty,$$

where we have used the continuity of $\tilde{B}(x)|\psi'(x)/\gamma'(x)|$. Now, we have shown that the first term on the RHS of (45) is bounded by $\mathbf{B}_1 \times \mathbf{B}_2 = O_{a.s.}(1)$.

For the case $p = 0$, we outline only main points. Instead of (43), we set

$$\gamma\left(x\right) = -1-\left(\log\left|x\right|\right)\ \text{if}\ x<-1;\quad = \tilde{\gamma}_m\left(x\right)\ \text{if}\ x\in[-1,1];\ = 1+\left(\log x\right)\ \text{if}\ x>1,$$

with some suitable bridging function $\tilde{\gamma}_m(x)$. By the linear growth condition and this specification of $\gamma(\cdot)$, we can obtain the result (44). We also consider $\tilde{B}(x)$ such that $\tilde{B}(x) = |x|^{-(p+q+1)}$ for $|x|$ large enough (instead of $\tilde{B}(x) = \exp\{-(\log|x|)^2\}$ as above). Then we can check the term corresponding to $\mathbf{B}_2$ is bounded. Since $\gamma^{-1}(y) = \exp\{y-1\}$ for sufficiently large y, the term corresponding to $\mathbf{B}_{12}$ is bounded by $\exp\{(p+q+1)|y-w|\} \leq \exp\{(p+q+1)\tilde{C}_\gamma\sqrt{\Delta\log(1/\Delta)}\}$. The rest of the proof is quite analogous and we omit details for brevity, completing the proof. □

Proof of Theorem 1 without (A3.ii). We note that (A3.ii) imposes no restriction if $p = 0$, and we here let $p > 0$. The proof proceeds quite analogously to that for the case with (A3.ii) and $p > 0$, and we only outline main points. We select the same $\tilde{\gamma}$ as in (43) but a different $\tilde{B}$. That is, by (B1'), we can find some $\tilde{B}(x) = \bar{B}\exp\{-|x|^p\}$ satisfying $B(x) \leq \tilde{B}(x)$ for $x \in \mathbb{R}$ (with some constant $\bar{B} > 0$ large enough). Then, the LHS of (46) is bounded by

$$\sup_{|y-w|\leq \tilde{C}_\gamma \sqrt{\Delta \log(1/\Delta)}} \exp\{-|y-w|\} \times \sup_{w\in(-1,1)} \tilde{B}\left(\gamma^{-1}(w)\right)|\eta(w)|,$$

where the first term is $O_{a.s.}(1)$, and the boundedness of the second term also follows by arguments similar to previous ones. □

7.2 Proofs of Convergence Results: Theorems 2-4 and Related Results

Here, we provide proofs of two general convergence results (Theorems 2 - 3) in Section 4, and that of the drift estimator's convergence (Theorem 4) in Section 5. The proof of the diffusion estimator's convergence can be found in the Supplementary Material. Here, we present two useful results on the covering number and the Bernstein exponential inequality, which we below use repeatedly.

Covering Number for Kernel Transformations: We start with some discussions on conditions for the kernel function K. As stated in the Introduction, we establish the uniform convergence results of the kernel-based estimators over an unbounded support. The technical difficulty arises due to this unboundedness. Two methods are commonly used to circumvent this difficulty. The classical method uses a Fourier transformation. This was probably first considered by Parzen (1962) and has been used in several studies, including Bierens (1983) and Andrews (1995), which exploit the multiplicative expression of the kernel function by the inversion formula and the boundedness of $\exp\{ix\}$ (i is the imaginary unit). Some studies, including Bosq (1998) and Hansen (2008), employ another method. They verify the uniform convergence over a compact set which expands as the sample size increases, and show that the outside of the compact set is asymptotically negligible. Our method differs from these, which is based on a covering-number technique from empirical process theory. The advantage of our method is that it does not require the process' finite moments. The classical method requires the existence of the moment in order to bound a component involving the Fourier inversion (see Sec. 2 of Bierens, 1983; or Lemma A-1 of Andrews, 1995). The aforementioned second method also requires the existence of the moment to show the negligibility of the outside of the expanding compact set (see the proof of Corollary 2.2 of Bosq, 1998) or to use the so-called truncation argument (see the proofs of Theorems 2-4 of Hansen, 2008). The use of the covering-number technique also allows for a very flexible form of the kernel function, which imposes only minimal restrictions. The classical method requires the Fourier invertibility of the kernel function. In addition, some studies assume that the kernel function is continuous and/or has truncated support (or sufficiently fast tail decay). We do not require any of these (see the conditions in (B2) and its discussions in Section 4).

Here, we formally introduce the covering number. To this end, let $\mathcal{L}_r(Q)$ denote the set of functions $g : \mathbb{R} \to \mathbb{R}$ such that $\|g\|_{Q,r} := [\int |g|^r dQ]^{1/r} < \infty$, where $r \geq 1$ and Q is a probability measure on $\mathbb{R}$. The covering number $N(\epsilon, \mathcal{G}, \mathcal{L}_r(Q))$ is the minimum number of ϵ-balls in $\mathcal{L}_r(Q)$ needed to cover $\mathcal{G}$, where an ϵ-ball in $\mathcal{L}_r(Q)$ around a function $g \in \mathcal{L}_r(Q)$ is the set $\{f \in \mathcal{L}_r(Q) \mid \|f - g\|_{Q,r} < \epsilon\}$. For a collection of balls to cover $\mathcal{G}$, all elements of $\mathcal{G}$ must be included in at least one of the balls, but the centers of the balls need not to belong to $\mathcal{G}$. An envelope function for $\mathcal{G}$ is any function G such that $|g| \leq G$ for all $g \in \mathcal{G}$. Given these notions, we have the following lemma:

Lemma 2. *Suppose that a function $K(\cdot)(:\mathbb{R}\to\mathbb{R})$ is of bounded variation and that there exists some constant $\bar{K}\in(0,\infty)$ such that $\sup_{x\in\mathbb{R}}|K(x)|\leq\bar{K}$. Let $\mathcal{K}$ be a set of all rescaled translates of K, i.e.,*

$$\mathcal{K}:=\left\{K\left(\tfrac{p-x}{h}\right)\;\middle|\;x\in\mathbb{R}\text{ and }h>0\right\}. \tag{47}$$

Then, the covering numbers of $\mathcal{K}$ satisfy

$$\sup_Q N\left(\epsilon 8\bar{K},\mathcal{K},\mathcal{L}_r(Q)\right)\leq\Lambda\epsilon^{-4r}\quad\text{for }\epsilon\in(0,1), \tag{48}$$

where the supremum is over all probability measures on $\mathbb{R}$; and $\Lambda>0$ is some constant independent of Q.

A set of functions is called *Euclidean* if its covering number increases at the rate of ϵ^{-V} with some constant $V>0$. This lemma says that $\mathcal{K}$ is Euclidean with $V\leq 4r$. Several studies, including Pollard (1984), Nolan and Pollard (1987), and Pakes and Pollard (1989), have argued that the set $\mathcal{K}$ defined in (47) is Euclidean, but they have not provided a further result on the bound of V (a similar result can be also found in Lemma B.3 of Escanciano, Jacho-Chávez and Lewbel, 2014). The proof of this lemma, which is based on several basic results in empirical process theory, is provided in the Supplementary Material.

The Bernstein-Type Exponential Inequality for Strong Mixing Arrays: We next present a useful inequality, which is presented in Liebscher (1996) (see Theorem 2.1 and the arguments in Section 3, p. 73) and is derived from Theorem 5 of Rio (1995):

Lemma 3. *Let $\{Z_{n,j}\}_{j=1}^n$ be a stationary zero-mean real-valued triangular array such that $|Z_{n,j}|\leq C_Z$, with strong mixing coefficients $\alpha(s)$. Then, for each positive integer $m\leq n$ and each real number η such that $m\leq\eta/4C_Z$,*

$$\Pr(|\textstyle\sum_{j=1}^n Z_{n,j}|\geq\eta)\leq 4\exp\{-\eta^2\left[64n\left(\Sigma_m^2/m\right)+(8/3)C_Z\eta m\right]^{-1}\}+(4n/m)\alpha(m\Delta), \tag{49}$$

where $\Sigma_m^2:=E[(\sum_{j=1}^m Z_{n,j})^2]$.

Given two lemmas, we are ready to prove Theorem 2. For notational simplicity, we often write the product of functions as $f\cdot g(x):=f(x)g(x)$ in the subsequent proofs.

Proof of Theorem 2. By the triangle inequalities, the LHS of (24) is bounded by the sum of three terms:

$$\sup_{x\in I}\left|G_{n,T}(x)-\bar{G}_{n,T}(x)\right|\leq R_1+R_2+(1/n)\sup_{x\in I}\left|\bar{G}_{n,T}(x)\right|, \tag{50}$$

where

$$R_1:=\sup_{x\in I}\left|(1/Th)\textstyle\sum_{j=1}^{n-1}K\left(\frac{X_{j\Delta}-x}{h}\right)B(X_{j\Delta})\int_{j\Delta}^{(j+1)\Delta}[\psi(X_s)-\psi(X_{j\Delta})]\,ds\right|;$$
$$R_2:=h^{-1}\sup_{x\in I}\left|(\Delta/T)\textstyle\sum_{j=1}^{n-1}\left[K\left(\frac{X_{j\Delta}-x}{h}\right)B\cdot\psi(X_{j\Delta})-\int K\left(\frac{p-x}{h}\right)B\cdot\psi\cdot\pi(p)\,dp\right]\right|.$$

We subsequently show that

$$R_1=O_p(\sqrt{\Delta\log(1/\Delta)}); \tag{51}$$
$$R_2=O_p(\sqrt{(\log T)/Th}), \tag{52}$$

where θ is given in the statement of the theorem. Since we can easily check that $\sup_{x\in I}|\bar{G}_{n,T}(x)| = O(1)$ (as in (55, provided that $\sup_{x\in I}\pi(x) < \infty$ and $\psi \in \mathcal{D}(B,\pi)$), (51)-(52) imply the desired result.

Proof of (51). We also look at

$$\begin{aligned} R_1 &\leq \sup_{x\in I}(\Delta/Th)\sum_{j=1}^{n-1}\left|K\left(\tfrac{X_{j\Delta}-x}{h}\right)\right| \times \max_{1\leq j\leq n-1}\sup_{s\in[j\Delta,(j+1)\Delta]}B(X_{j\Delta})\left|\psi(X_s)-\psi(X_{j\Delta})\right| \\ &= : R_{11}\times R_{12}, \end{aligned} \tag{53}$$

To find the bound of R_{11}, observe that

$$R_{11} \leq \sup_{x\in I}\left|(1/nh)\sum_{j=1}^{n-1}\left|K\left(\tfrac{X_{j\Delta}-x}{h}\right)\right| - E[\left|K\left(\tfrac{X_{j\Delta}-x}{h}\right)\right|]\right| + \sup_{x\in I}\int\tfrac{1}{h}\left|K\left(\tfrac{p-x}{h}\right)\right|\pi(p)\,dp, \tag{54}$$

where the stationarity of the process is used; the first term on the RHS is $o_p(1)$, which can be verified by the same arguments as those for R_2 (below) using the boundedness of $\pi(\cdot)$ in (A2); and the second term is bounded since

$$\sup_{x\in I}\int K(q)\,\pi(qh+x)\,dq \leq \int|K(q)|\,dq \times \sup_{x\in I}\pi(x) < \infty. \tag{55}$$

Then, we have $R_{11} = O_p(1)$. Since ψ satisfies the conditions of Theorem 1, we have $R_{12} = O_{a.s.}(\sqrt{\Delta\log(1/\Delta)})$. From these, we can obtain (51) as desired.

Proof of (52). Define two measures on $\mathbb{R}$, Q_n and Q_0, as follows: for every Borel set E on $\mathbb{R}$,

$$Q_n(E) := n^{-1}\sum_{j=1}^{n-1}\mathbf{1}_{\{X_{j\Delta}\in E\}}B\cdot\psi(X_{j\Delta}); \quad Q_0(E) := \int_E B\cdot\psi\cdot\pi(p)\,dp. \tag{56}$$

Q_n and Q_0 can be considered to be weighted versions of an empirical measure and a probability measure, respectively, with a weight $B\cdot\psi$. Recalling the conditions in Definition 1, we can check that both are finite measures, i.e., $Q_n(I) \leq \sup_{p\in I}|B\cdot\psi(p)| < \bar{B}_1$ for any n, and $Q_0(I) = \int B\cdot\psi\cdot\pi(p)\,dp \leq \sup_{p\in I}|B\cdot\psi(p)| < \bar{B}_2$. Since we do not assume ψ is positive-valued, $Q_n(E)$ and $Q_0(E)$ may take negative values for some E. Thus, they are finite signed measures (instead of usual finite measures). It is known that any finite signed measure v can be expressed as a difference of two (positive) finite measures v_1 and v, that is, $v = v_1 - v_2$, which can be checked by letting $v_1 = (|v|+v)/2 \geq 0$ and $v_2 = (|v|-v)/2 \geq 0$. While our empirical-process technique based on Lemma 2 is designed for probability measures, we can still apply it to general finite signed measures by considering the decomposition as $v = v_1 - v_2$ and re-normalization of Q_n and Q_0. For brevity, we regard Q_n and Q as probability measures throughout this proof.

Now, consider the following set of functions (indexed by x) for each $h\,(>0)$: $\mathcal{K}(h) := \{K((p-x)/h) \mid x\in\mathbb{R}\}$. For any h, $\mathcal{K}(h)\subset\mathcal{K}$ where $\mathcal{K}$ is defined in (47) of Lemma 2. (B2) implies all the conditions of Lemma 2. Thus, by Lemma 2, the covering numbers satisfy

$$\sup_Q N\left(8\varepsilon\bar{K},\mathcal{K}(h),\mathcal{L}_r(Q)\right) \leq \sup_Q N\left(8\varepsilon\bar{K},\mathcal{K},\mathcal{L}_r(Q)\right) \leq \Lambda\varepsilon^{-4r} \quad \text{for } \varepsilon\in(0,1), \tag{57}$$

uniformly over any $h>0$. Here, we consider the case with $r=1$. Then, for each h, we can construct a partition of $\mathcal{K}(h)$, $\{\mathcal{K}_k(h)\}_{k=1}^{\nu(h)}$, which satisfies

$$\begin{aligned} &\mathcal{K}(h)\subseteq\textstyle\bigcup_{k=1}^{\nu(h)}\mathcal{K}_k(h)\text{; each }\mathcal{K}_k(h)\text{ has the center }g_k(\cdot) := K\left(\tfrac{\cdot-x_k}{h}\right)\text{, such that} \\ &\forall\varepsilon\in(0,1),\ \forall g\in\mathcal{K}_k(h),\ \int|g-g_k|\,dQ \leq 8\varepsilon\bar{K}\text{ for any probability measure }Q. \end{aligned} \tag{58}$$

To find the bound of R_2, we write it in terms of two measures Q_n and Q_0 (defined in (56)) and the partition of $\mathcal{K}(h)$:

$$R_2 = (1/h) \sup_{x \in I} \left| \int g dQ_n - \int g dQ_0 \right| \leq R_{21} + R_{22}, \tag{59}$$

where

$$R_{21} := (1/h) \max_{k \in \{1,...,\nu(h)\}} \sup_{g \in \mathcal{K}_k(h)} \left| \int (g - g_k) \, dQ_n - \int (g - g_k) \, dQ_0 \right|; \tag{60}$$

$$R_{22} := (1/h) \max_{k \in \{1,...,\nu(h)\}} \left| \int g_k dQ_n - \int g_k dQ_0 \right|, \tag{61}$$

where we subsequently derive the bounds these two terms. Given any ε in (58), we can bound R_{21} as

$$R_{21} \leq (1/h) \left\{ 8\varepsilon \bar{K} + 8\varepsilon \bar{K} \right\} = O(\varepsilon/h), \tag{62}$$

By setting

$$\varepsilon = h \left[(\log T) / Th \right]^{1/2}, \tag{63}$$

which tends to zero, we have $R_{21} = O([(\log T)/Th]^{1/2})$.

To find the probability bound of R_{22}, we write $R_{22} = (1/nh) \max_{k \in \{1,...,\nu(h)\}} | \sum_{j=1}^{n-1} Y_{n,j}(k,h) |$, where $\{Y_{n,j}(k,h)\}_{j=1}^{n}$ is a triangular array defined as

$$Y_{n,j}(k,h) := K\left(\tfrac{X_{j\Delta} - x_k}{h} \right) B \cdot \psi(X_{j\Delta}) - \int K\left(\tfrac{p - x_k}{h} \right) B \cdot \psi \cdot \pi(p) \, dp,$$

for each (k,h). We apply the Bernstein inequality in Lemma 3 to $\sum_{j=1}^{n-1} Y_{n,j}(k,h)$ (with $Z_{n,j} = Y_{n,j}(k,h)$ and $\Sigma_m^2 = \Sigma_m^2(k,h) := E[| \sum_{j=1}^{m} Y_{n,j}(k,h) |^2]$ for each (k,h)). To this end, we note that $\sup_{j,k} |Y_{n,j}(k,h)| \leq 2\bar{K}\bar{B}_2 =: C_Y$ (by the boundedness of K and $B \cdot \psi$), and the following bound:

Lemma 4. *If the mixing exponent $\beta > 1$, there exists some constant $\tilde{\varpi}$ such that for $m \leq (n-1)$ with n sufficiently large*

$$\Sigma_m^2(k,h) = E[| \sum_{j=1}^{m} Y_{n,j}(k,h) |^2] \leq \tilde{\varpi} mh \left(1 + h^{1-2/\beta} \Delta^{-1} \right), \text{ uniformly over } k \text{ and } h. \tag{64}$$

Given this, we look at

$$\begin{aligned}
&\Pr(R_{22} \geq a \left[(\log T)/Th \right]^{1/2}) \leq \textstyle\sum_{k=1}^{\nu(h)} \Pr(| \sum_{j=1}^{n-1} Y_{n,j}(k,h) | \geq a \sqrt{(\log T)/Thnh}) \\
&\leq 4\Lambda \varepsilon^{-4} \left\{ 4 \exp\left\{ -\frac{a^2 (\log T)}{64\varpi (\Delta + h^{1-2/\beta}) + (8/3) C_Y a} \right\} + 4AT^{1/2-\beta/2} (\log T)^{\beta/2-1/2} h^{-1/2-\beta/2} \right\} \\
&\leq 4\Lambda T^2 (\log T)^{-2} h^{-2} \left\{ \exp\left\{ -\frac{a^2 (\log T)}{64\varpi + (8/3) C_Y a} \right\} + AT^{1/2-\beta/2} (\log T)^{\beta/2-1/2} h^{-1/2-\beta/2} \right\} \\
&\leq 4\Lambda (\log T)^{-4} \left\{ T^{2+2\theta - a^2/[64\varpi + (8/3) C_Y a]} + A (\log T)^{-1} T^{5/2-\beta/2+(5/2+\beta/2)\theta} \right\},
\end{aligned}$$

where the second inequality holds by (57), (49) and (64) with setting $\eta = a[(\log T)/Th]^{1/2} nh$ and $m = [Th/(\log T)]^{1/2} \Delta^{-1}$ ($m \leq (n-1)$ and $m < \eta/4C_Y$ are satisfied for large T and a); the third holds by (63) and "$\Delta + h^{1-2/\beta} \leq 1$" (as $h, \Delta \to 0$ for $\beta \geq 2$), and the last inequality holds since $h^{-1} \leq T^{\theta} (\log T)^{-1}$ (which follows from $(\log T)/T^{\theta} h \to 0$). Therefore, if $5/2 - \beta/2 + (5/2 + \beta/2)\theta \leq 0 \Leftrightarrow \beta(1-\theta) \geq 5(\theta+1)$, we have $\Pr(R_{22} \geq a[(\log T)/Th]^{1/2}) \to 0$ for large $a (> 0)$, which means that $R_{22} = o_p([(\log T)/Th]^{1/2})$. This, together with the rate of R_{21}, leads to the desired result: $R_2 = O_p([(\log T)/Th]^{1/2})$. □

Proof of Lemma 4. We consider bounding the covariance in two ways. First, for any $j \geq 1$,

$$\begin{aligned}
&E\left[Y_{n,1}\left(k,h\right)Y_{n,j+1}\left(k,h\right)\right] \\
&= h^2 \int\int K\left(r\right)K\left(s\right)B\cdot\psi\left(r+x_k h\right)B\cdot\psi\left(s+x_k h\right)\pi_{\Delta,(j+1)\Delta}\left(r+x_k h, s+x_k h\right)drds \\
&\leq h^2 \int \left|B\cdot\psi\left(r+x_k h\right)\right|^2 \pi\left(r+x_k h\right)K^2\left(r\right)dr \leq h^2 \bar{B}_1\bar{B}_2\bar{K}\int\left|K\left(r\right)\right|dr = \varpi h^2, \qquad (65)
\end{aligned}$$

where $\pi_{\Delta,(j+1)\Delta}$ is the joint density of X_Δ and $X_{\Delta(j+1)}$, whose existence can be checked by (A1.i); the first equality holds by changing variables; the inequalities follows from the Schwartz inequality and (i)-(ii) of Definition 1; and the last equality holds with $\varpi := \bar{B}_1\bar{B}_2\bar{K}\int\left|K\left(q\right)\right|dq\,(<\infty)$. Second, by the Billingsley inequality (Bosq, 1998, Corollary 1.1), for $j \geq 1$,

$$E\left[Y_{n,1}\left(k,h\right)Y_{n,j+1}\left(k,h\right)\right] \leq 4\alpha\left(j\Delta\right)\times\left\|Y_{1,n}(k,h)\right\|_\infty^2 \leq 4A\left(j\Delta\right)^{-\beta}C_Y^2. \qquad (66)$$

We also have the moment bound:

$$E\left[Y_{n,j}^2\left(k,h\right)\right] = h\int K^2\left(q\right)B^2\cdot\psi^2\cdot\pi\left(qh+x_k\right)dq \leq \varpi h, \qquad (67)$$

for any j, k and h, which follows from the stationarity of the process, changing variables and (i)-(ii) of Definition 1. Then,

$$\begin{aligned}
&E[|\textstyle\sum_{j=1}^m Y_{n,j}\left(k,h\right)|^2] \leq 2m\sum_{2\leq j+1\leq\Delta^{-1}h^{-2/\beta}}E\left[Y_{n,1}\left(k,h\right)Y_{n,j+1}\left(k,h\right)\right] \\
&+ 2m\sum_{j+1>\Delta^{-1}h^{-2/\beta}}E\left[Y_{1,n}\left(k,h\right)Y_{j+1,n}\left(k,h\right)\right] + \sum_{1\leq j\leq m}E\left[Y_{n,j}^2\left(k,h\right)\right] \\
&\leq 2m\Delta^{-1}h^{-2/\beta}\varpi h^2 + 2m\sum_{j+1>\Delta^{-1}h^{-2/\beta}}4A\left(j\Delta\right)^{-\beta}C_Y^2 + m\varpi h \\
&\leq 2mh^{2-2/\beta}\Delta^{-1}\left[\varpi + 4AC_Y^2/\left(\beta-1\right)\right] + \varpi mh,
\end{aligned}$$

where the first inequality holds by the stationarity; the second by (65)-(67); and the last inequality uses the following fact:

$$\textstyle\sum_{j+1>\Delta^{-1}h^{-2/\beta}}j^{-\beta} \leq \int_{\Delta^{-1}h^{-2/\beta}}^\infty x^{-\beta}dx \leq \left[\Delta^{-1}h^{-2/\beta}\right]^{1-\beta}/\left(\beta-1\right) = \Delta^{\beta-1}h^{2-2/\beta}/\left(\beta-1\right), \qquad (68)$$

(note that $\beta > 1$ and $\Delta^{-1}h^{-2/\beta} \geq 1$ for small Δ and h). We now have shown that (64) holds with $\tilde{\varpi} = 3\varpi + 8AC_Y^2/\left(\beta-1\right)$. □

Proof of Theorem 3. For each (x,h,n,T), let $\left\{N_r\left(x,h;n,T\right)\right\}_{r\in[0,1]}$ be a process on $[0,1]$ defined as

$$N_r\left(x,h;n,T\right) := \int_0^r K\left(\tfrac{X_{\lceil sn\rceil\Delta}-x}{h}\right)B\left(X_{\lceil sn\rceil\Delta}\right)\rho_{sn\Delta}dW_{sn\Delta},$$

where $\lceil z\rceil$ is the largest integer which is less than or equal to z. For notational simplicity, we write $N_r\left(x,h\right) = N_r\left(x,h;n,T\right)$ in the sequel. Then, we can write

$$M_{n,T}\left(x\right) = \left(1/Th\right)N_1\left(x,h\right). \qquad (69)$$

Note that for each (x,h), $\left\{N_r\left(x,h\right)\right\}$ is a continuous martingale with respect to the filtration $\left\{\mathfrak{G}_r\right\}_{r\in[0,1]}$ where $\mathfrak{G}_r := \mathfrak{F}_{rn\Delta}$. This martingale vanishes at $r=0$, and its quadratic variation process is given by $\left\langle N\left(x,h\right)\right\rangle_r = \int_0^r K^2\left(\frac{X_{\lceil sn\rceil\Delta}-x}{h}\right)\rho_{sn\Delta}^2 d\left(sn\Delta\right)$. We can write $\left\langle N\left(x,h\right)\right\rangle_1 = \sum_{j=1}^{n-1}K^2\left(\frac{X_{j\Delta}-x}{h}\right)\int_{j\Delta}^{(j+1)\Delta}\rho_u^2 du$.

To find the probability bound of $N_1(x,h)$, we apply the Bernstein-type inequality for continuous martingales (see, e.g., Ex. 3.16 in Ch. IV of Revuz and Yor, 1999): for any $a>0$ and $y>0$,

$$\Pr\left(|N_1(x,h)|\geq\eta,\ \ \langle N(x,h)\rangle_1\leq y\right)\leq 2\exp\left\{-\eta^2/2y\right\}. \tag{70}$$

Then, let $a\,(>0)$ be any constant.

$$\begin{aligned}
&\Pr(|M_{n,T}(x)|\geq a\sqrt{(\log T)/Th})\\
&\leq \Pr(|N_1(x,h)|\geq aTh\sqrt{(\log T)/Th},\ \ \langle N(x,h)\rangle_1\leq y)+\Pr\left(\langle N(x,h)\rangle_1>y\right)\\
&\leq 2\exp\left\{-a^2(\log T)/2\right\}+\Pr\left(\langle N(x,h)\rangle_1>aTh\right)=2T^{-a^2/2}+\Pr\left(\langle N(x,h)\rangle_1>aTh\right),
\end{aligned} \tag{71}$$

where the second inequality holds by (70) with $\eta=aTh\sqrt{(\log T)/Th}$ and $y=aTh$.

We next derive the probability bound of $\langle N(x,h)\rangle_1$. To this end, define an array $\{\tilde{Z}_{n,j}(x,h)\}$ by

$$\tilde{Z}_{n,j}(x,h):=K^2\left(\tfrac{X_{j\Delta}-x}{h}\right)\textstyle\int_{j\Delta}^{(j+1)\Delta}\rho_u^2du-E[K^2\left(\tfrac{X_{j\Delta}-x}{h}\right)\int_{j\Delta}^{(j+1)\Delta}\rho_u^2du],$$

Observe that $|\tilde{Z}_{n,j}(x)|\leq\tilde{C}\Delta$ for some constant $\tilde{C}>0$ (uniformly over x), and

$$\langle N(x,h)\rangle_1=\sum_{j=1}^{n-1}\tilde{Z}_{n,j}(x,h)+\sum_{j=1}^{n-1}E[K^2\left(\tfrac{X_{j\Delta}-x}{h}\right)\textstyle\int_{j\Delta}^{(j+1)\Delta}\rho_u^2du]\leq\sum_{j=1}^{n-1}\tilde{Z}_{n,j}(x,h)+C_\rho Th,$$

where the last inequality holds with some $C_\rho\in(0,\infty)$ by the change-of-variable argument. Given this, we have

$$\Pr\left(\langle N(x,h)\rangle_1>aTh\right)\leq\Pr(|\textstyle\sum_{j=1}^{n-1}\tilde{Z}_{n,j}(x,h)|\geq(a/2)Th)+\Pr\left(C_\rho Th\geq(a/2)Th\right), \tag{72}$$

where the second term on the RHS is zero for any a large enough (and thus is negligible). To find the bound of the first term, we use the Bernstein-type inequality in Lemma 3 with the following bound of $\tilde{\Sigma}_m(x,h)$:

$$\tilde{\Sigma}_m(x,h):=E[|\textstyle\sum_{j=1}^{m}\tilde{Z}_{n,j}(x,h)|^2]\leq\tilde{C}_\rho\Delta^2hm^2\ \ \text{for } m\leq(n-1), \tag{73}$$

which holds by the change-of-variable argument. Then,

$$\begin{aligned}
&\Pr(|\textstyle\sum_{j=1}^{n-1}\tilde{Z}_{n,j}(x,h)|\geq(a/2)Th)\leq 4\exp\left\{-\frac{(a^2/4)T^2h^2}{64n\left(\tilde{C}_\rho\Delta^2hm\right)+(8/3)\tilde{C}\Delta\left[(a/2)Th\right]m}\right\}+4Anm^{-1-\beta}\Delta^{-\beta}\\
&\leq 4\exp\left\{-\frac{(a^2/4)(\log T)}{64\tilde{C}_\rho+(4/3)\tilde{C}a}\right\}+4AT^{-\beta}h^{-(\beta+1)}(\log T)^{-1-\beta}\leq T^{-aC_M}+4AT^{-\beta}h^{-(\beta+1)}(\log T)^{-1-\beta},
\end{aligned} \tag{74}$$

where the first two inequalities use (49) with (73), $m=nh/(\log T)$ and $\eta=(a/2)Th$, which satisfy $m\leq\min\{(n-1),\eta/4(\tilde{C}\Delta)\}$; and the last inequality holds with $C_M=1/4[1+(4/3)\tilde{C}]$ for a large enough with $64\tilde{C}_\rho/a\leq 1$. Now, (71), (72) and (74) imply the desired result. □

Proof of Theorem 4. We consider the following decomposition: $\sup_{x\in\mathbb{R}}|\hat{\Psi}_\mu(x)-B(x)\mu(x)\pi(x)|\leq\sum_{i=1}^{3}U_i$, where

$$\begin{aligned}
U_1 &:= \sup_{x\in\mathbb{R}}\left|(1/Th)\textstyle\sum_{j=1}^{n-1}\left[K\left(\tfrac{X_{j\Delta}-x}{h}\right)B(X_{j\Delta})\int_{j\Delta}^{(j+1)\Delta}\mu(X_s)\,ds-E[K\left(\tfrac{X_{j\Delta}-x}{h}\right)B\cdot\mu(X_{j\Delta})]\right]\right|;\\
U_2 &:= \sup_{x\in\mathbb{R}}\left|(1/h)E[K\left(\tfrac{X_{j\Delta}-x}{h}\right)B\cdot\mu(X_{j\Delta})]-B\cdot\mu\cdot\pi(x)\right|;\\
U_3 &:= \sup_{x\in\mathbb{R}}\left|(1/Th)\textstyle\sum_{j=1}^{n-1}K\left(\tfrac{X_{j\Delta}-x}{h}\right)B(X_{j\Delta})\int_{j\Delta}^{(j+1)\Delta}\sigma(X_s)\,dW_s\right|.
\end{aligned}$$

Then, applying Theorem 2 to U_1 (with $\psi = \mu$), we can immediately obtain $U_1 = O_p(a_T^\star)$. We can also easily derive $U_2 = O(h^2)$, which is the smoothing bias term, by standard change-of-variable and Taylor-expansion arguments for kernel estimators with the condition that $\psi \in \mathcal{D}(B, \pi)$. Therefore, if we have shown that $U_3 = O_p(\sqrt{(\log T)/Th})$, we obtain the desired result (29). Given the rate result (29) we can verify the convergence result for $\hat{\mu}(x)$ in the same way as in the proof of Theorem 8 of Hansen (2008) (we omit details for brevity).

It remain to derive the rate of U_3**.** To this end, let $B(X_{j\Delta}) \int_{j\Delta}^{(j+1)\Delta} \sigma(X_s)\, dW_s = \int_{j\Delta}^{(j+1)\Delta} \phi_s dW_s + \int_{j\Delta}^{(j+1)\Delta} \tilde{\phi}_s dW_s$ for each j, where $\{\phi_s\}$ and $\{\tilde{\phi}_s\}$ are defined as

$$\phi_s := B \cdot \sigma(X_{\lceil s/\Delta \rceil \Delta}) + B(X_{\lceil s/\Delta \rceil \Delta})[\sigma(X_s) - \sigma(X_{\lceil s/\Delta \rceil \Delta})] \mathbf{1}_{\left\{B(X_{\lceil s/\Delta \rceil \Delta}) \left| \sigma(X_s) - \sigma(X_{\lceil s/\Delta \rceil \Delta}) \right| \leq \sqrt{\Delta} \log \Delta^{-1} \right\}}; \quad (75)$$

$$\tilde{\phi}_s := B(X_{\lceil s/\Delta \rceil \Delta}) \left[\sigma(X_s) - \sigma(X_{\lceil s/\Delta \rceil \Delta})\right] \mathbf{1}_{\left\{B(X_{\lceil s/\Delta \rceil \Delta}) \left| \sigma(X_s) - \sigma(X_{\lceil s/\Delta \rceil \Delta}) \right| > \sqrt{\Delta} \log \Delta^{-1} \right\}}. \quad (76)$$

Recalling the conditions in (27) and the result (20), we can see that ϕ_s is uniformly bounded over $s \geq 0$ and $\tilde{\phi}_s$ is zero a.s. for any sufficiently small Δ. Therefore, the convergence rate of U_3 is determined by that of $\bar{U}_3 := \sup_{x \in \mathbb{R}} |(1/Th) \sum_{j=1}^{n-1} K(\frac{X_{j\Delta} - x}{h}) \int_{j\Delta}^{(j+1)\Delta} \phi_s dW_s|$. To derive the rate of $\bar{U}_3$, define a class of functions as $\mathcal{K}(h) := \left\{ K\left(\frac{p-x}{h}\right) \middle| \, x \in \mathbb{R} \right\}$ for each $h\,(>0)$. By Lemma 2 (with $r = 1$), we can construct a finite covering $\{\mathcal{K}_k(h)\}_{k=1}^{\nu(h)}$ of $\mathcal{K}(h)$ satisfying the following conditions: $\mathcal{K}(h) \subseteq \bigcup_{k=1}^{\nu(h)} \mathcal{K}_k(h)$; each $\mathcal{K}_k(h)$ has the center $g_k(\cdot) := K\left(\frac{\cdot - x_k}{h}\right)$ such that for any probability measure Q,

$$\forall \varepsilon \in (0,1), \ \forall g \in \mathcal{K}_k(h), \ \int |g - g_k| \, dQ \leq \varepsilon 8 \bar{K}; \ \text{ and } \ \nu(h) \leq \Lambda \varepsilon^{-4} \ \text{ for each } \varepsilon \in (0,1), \quad (77)$$

for some constant $\Lambda\,(>0)$ (independent of h). Using this covering, we can obtain

$$\begin{aligned} \bar{U}_3 &\leq (1/Th) \max_{k \in \{1,\ldots,\nu(h)\}} \sup_{g \in \mathcal{K}_k(h)} \left| \sum_{j=1}^{n-1} [g_k(X_{j\Delta}) - g(X_{j\Delta})] \int_{j\Delta}^{(j+1)\Delta} \phi_s dW_s \right| \\ &\quad + \max_{k \in \{1,\ldots,\nu(h)\}} \left| (1/Th) \sum_{j=1}^{n-1} K\left(\frac{X_{j\Delta} - x_k}{h}\right) \int_{j\Delta}^{(j+1)\Delta} \phi_s dW_s \right| =: \bar{U}_{31} + \bar{U}_{32}. \end{aligned} \quad (78)$$

We subsequently derive the bounds of $\bar{U}_{31}$ and $\bar{U}_{32}$. Now, consider the bound of $\bar{U}_{31}$:

$$\begin{aligned} \bar{U}_{31} &\leq \tfrac{(n-1)}{Th} \max_{k \in \{1,\ldots,\nu(h)\}} \sup_{g \in \mathcal{K}_k(h)} \tfrac{1}{n-1} \sum_{j=1}^{n-1} |g_k(X_{j\Delta}) - g(X_{j\Delta})| \times \max_{1 \leq j \leq n-1} \left| \int_{j\Delta}^{(j+1)\Delta} \rho_s dW_s \right| \\ &\leq \tfrac{(n-1)}{Th} \varepsilon 8 \bar{K} \times O_{a.s.}(\sqrt{\Delta \log(1/\Delta)}) = O_{a.s.}(\varepsilon \sqrt{\Delta \log(1/\Delta)} \times (1/\Delta h)), \end{aligned} \quad (79)$$

where the second inequality holds by the following two results: (i) for $Q_n(E) := \frac{1}{n-1} \sum_{j=1}^{n-1} \mathbf{1}_{\{X_{j\Delta} \in E\}}$ (the empirical probability measure),

$$\max_{k \in \{1,\ldots,\nu(h)\}} \sup_{g \in \mathcal{K}_k(h)} \tfrac{1}{n-1} \sum_{j=1}^{n-1} |g_k(X_{j\Delta}) - g(X_{j\Delta})| = \max_{k \in \{1,\ldots,\nu(h)\}} \sup_{g \in \mathcal{K}_k(h)} \int |g_k - g| \, dQ_n \leq \varepsilon 8 \bar{K},$$

which follows from (77); (ii) $\max_{1 \leq j \leq n-1} \left| \int_{j\Delta}^{(j+1)\Delta} \phi_s dW_s \right| = O_{a.s.}(\sqrt{\Delta \log(1/\Delta)})$, which follows from (16) since $\{\phi_s\}$ is uniformly bounded. Next, we consider the probability bound of $\bar{U}_{32}$. By Theorem 3 (with $\rho_s = \phi_s$) and the bound of $\nu(h)$ in (77), we can find some constant $C_M\,(>0)$ satisfying

$$\Pr(\bar{U}_{32} \geq a\sqrt{(\log T)/Th}) \leq \Lambda \varepsilon^{-4} [2T^{-a^2/2} + T^{-aC_M} + 4AT^{-\beta} h^{-(\beta+1)} (\log T)^{1-\beta}]. \quad (80)$$

Now, by letting $\varepsilon = \sqrt{\Delta h/T}$, we obtain $\bar{U}_{31} = O_{a.s.}(\sqrt{(\log T)/Th})$ since $\log(1/\Delta) = \log T$ by the first condition in (26). Given this ε and (26), we have $\varepsilon^{-4} = T^2 \Delta^{-2} h^{-2} \leq T^{2+2\theta+2\kappa} (\log T)^{-2}$ as well as $T^{-\beta} h^{-(\beta+1)} \leq T^{-\beta+\theta(\beta+1)} (\log T)^{-(\beta+1)}$. Therefore,

$$\Pr(\bar{U}_{32} \geq a\sqrt{(\log T)/Th}) = O(T^{2+2\theta+2\kappa} (\log T)^{-2} [T^{-a^2/2} + T^{-aC_M}] + T^{-\beta+2+2\theta+2\kappa+\theta(\beta+1)} (\log T)^{-(2+\beta)}).$$

This implies that $\bar{U}_{32} = O_p(\sqrt{(\log T)/Th})$ if $-\beta + 2 + 2\theta + 2\kappa + \theta(\beta+1) \leq 0 \Longleftrightarrow 3\theta + 2 + 2\kappa \leq \beta(1-\theta)$. From these arguments, we have now shown that $U_3 = O_p(\sqrt{(\log T)/Th})$, completing the proof. □

Supplementary Material to "Uniform Convergence Rates of Kernel-Based Nonparametric Estimators for Continuous Time Diffusion Processes: A Damping Function Approach"

Shin Kanaya

University of Aarhus and CREATES

8 Proof of Lemma 2 (Covering Number of Kernel Transformations)

Terminologies in this proof follow those of van der Vaart and Wellner (1996; hereafter referred to as VW96). First, observe that K can be decomposed into

$$K(z) = K_1(z) - K_2(z), \tag{S.2}$$

since it is of bounded variation, where K_1 and K_2 are bounded monotone functions. By (S.2) and re-parametrization, we have

$$\mathcal{K} \subset \{K_1 - K_2 \mid K_1 \in \mathcal{K}_1;\ K_2 \in \mathcal{K}_2\}, \tag{S.3}$$

where $\mathcal{K}_l := \{K_l(ap+b) \mid a \in (0,\infty);\ b \in \mathbb{R}\}$ for $l = 1, 2$. Since K is bounded by $\bar{K}$, there exists an envelope function M_l for $\mathcal{K}_l$ $(l = 1, 2)$.

Note that the set of subgraphs of functions $\{ap + b \mid a \in (0,\infty);\ b \in \mathbb{R}\}$ is a VC-class with the VC index 3, which is shown by a similar argument to the proof of Lemma 2.6.16 of VW96 (the set of subgraphs of functions $\{ap + b \mid a \in (0,\infty)$ and $b \in \mathbb{R}\}$ cannot shatter any three-point set while it can shatter some two-point set in $\mathbb{R}^2$). Then, since K_l is monotone, $\mathcal{K}_l$ is also a VC-class with the VC-index at most 3, which follows from Lemma 9.9 (viii) of Kosorok (2008). We now have shown that each $\mathcal{K}_l$ is a VC-class with the VC-index of the set of subgraphs of functions in $\mathcal{K}_l$ is 3.

By Theorem 2.6.7 of VW96, we can find the uniform covering number of $\mathcal{K}_l$, i.e., the exists some constant $\Lambda_l\,(>0)$ that is independent of Q satisfying

$$\sup_Q N(\epsilon \|M_l\|_{Q,r}, \mathcal{K}_l, \mathcal{L}_r(Q)) \le \Lambda_l \epsilon^{-2r} \quad \text{for } \epsilon \in (0,1). \tag{S.4}$$

Then, it holds that for each Q,

$$\begin{aligned} N(\epsilon 4 \|M_1 + M_2\|_{Q,r}, \mathcal{K}, \mathcal{L}_r(Q)) &\le N(\epsilon 4 \|M_1 + M_2\|_{Q,r}, \mathcal{K}_1 - \mathcal{K}_2, \mathcal{L}_r(Q)) \\ &\le N(\epsilon \|M_1\|_{Q,r}, \mathcal{K}_1, \mathcal{L}_r(Q)) N(\epsilon \|M_2\|_{Q,r}, -\mathcal{K}_2, \mathcal{L}_r(Q)) \\ &= \textstyle\prod_{l=1,2} N(\epsilon \|M_l\|_{Q,r}, \mathcal{K}_l, \mathcal{L}_r(Q)), \end{aligned} \tag{S.5}$$

where the first inequality holds by (S.3); the second holds by Lemma 16 of Nolan and Pollard (1987); and the last equality by the fact $-\mathcal{K}_2$ and $\mathcal{K}_2$ have the same covering number. From (S.4) and (S.5), it holds that

$$\begin{aligned} \sup_Q N(\epsilon 4 \|M_1 + M_2\|_{Q,r}, \mathcal{K}, \mathcal{L}_r(Q)) &\le \textstyle\prod_{l=1,2} \sup_Q N(\epsilon \|M_l\|_{Q,r}, \mathcal{K}_l, \mathcal{L}_r(Q)) \\ &\le \Lambda_1 \Lambda_2 \epsilon^{-4r}. \end{aligned}$$

Since $4\|M_1 + M_2\|_{Q,r} \le 8\bar{K}$, we have shown that (48) holds with some constant $\Lambda = \Lambda_1\Lambda_2$, completing the proof. □

9 Proof of Theorem 5 (Uniform Convergence of the Diffusion Estimator)

Using the Ito lemma:

$$\left[X_{(j+1)\Delta}-X_{j\Delta}\right]^2-\Delta\sigma^2(x)=2\int_{j\Delta}^{(j+1)\Delta}\left[X_s-X_{j\Delta}\right]\mu(X_s)\,ds$$
$$+2\int_{j\Delta}^{(j+1)\Delta}\left[X_s-X_{j\Delta}\right]\sigma(X_s)\,dW_s+\int_{j\Delta}^{(j+1)\Delta}\left[\sigma^2(X_s)-\sigma^2(x)\right]ds, \tag{S.6}$$

we have the following decomposition:

$$\sup_{x\in\mathbb{R}}|\hat{\Psi}_{\sigma^2}(x)-\hat{\Pi}(x)\,\sigma^2(x)|\leq\sum_{i=1}^5\mathcal{V}_i,$$

where

$$\mathcal{V}_1:=2\sup_{x\in\mathbb{R}}\left|(1/Th)\sum_{j=1}^{n-1}K\left(\tfrac{X_{j\Delta}-x}{h}\right)B(X_{j\Delta})\int_{j\Delta}^{(j+1)\Delta}[X_s-X_{j\Delta}]\mu(X_s)\,ds\right|;$$
$$\mathcal{V}_2:=2\sup_{x\in\mathbb{R}}\left|(1/Th)\sum_{j=1}^{n-1}K\left(\tfrac{X_{j\Delta}-x}{h}\right)B(X_{j\Delta})\int_{j\Delta}^{(j+1)\Delta}[X_s-X_{j\Delta}]\sigma(X_s)\,dW_s\right|;$$
$$\mathcal{V}_3:=\sup_{x\in\mathbb{R}}\left|(1/Th)\sum_{j=1}^{n-1}K\left(\tfrac{X_{j\Delta}-x}{h}\right)B(X_{j\Delta})\int_{j\Delta}^{(j+1)\Delta}[\sigma^2(X_s)-\sigma^2(X_{j\Delta})]ds\right|;$$
$$\mathcal{V}_4:=\sup_{x\in\mathbb{R}}\left|\tfrac{n-1}{n}\int K\left(\tfrac{p-x}{h}\right)B(p)\,\pi(p)\,[\sigma^2(p)-\sigma^2(x)]dp\right|;$$
$$\mathcal{V}_5:=\sup_{x\in\mathbb{R}}(1/nh)\left|\sum_{j=1}^{n-1}\{\Gamma_{j\Delta}(x)-E[\Gamma_{j\Delta}(x)]\}\right|;$$

and

$$\Gamma_{j\Delta}(x):=K\left(\tfrac{X_{j\Delta}-x}{h}\right)B(X_{j\Delta})\,[\sigma^2(X_{j\Delta})-\sigma^2(x)].$$

We below verify the following results:

$$\mathcal{V}_1 = O_p(\Delta\log(1/\Delta)+\sqrt{(\log n)/nh}); \tag{S.7}$$
$$\mathcal{V}_2 = O_p(\sqrt{(\log n)/nh}); \tag{S.8}$$
$$\mathcal{V}_3 = O_p(\Delta+\sqrt{(\log n)/nh}); \tag{S.9}$$
$$\mathcal{V}_4 = O(h^2); \tag{S.10}$$
$$\mathcal{V}_5 = O_p(\sqrt{(\log n)/nh}), \tag{S.11}$$

under the stated conditions, which imply the first part of the theorem. The convergence of $\hat{\mu}(x)$ can be verified in the same way as in the proof of Theorem 8 of Hansen (2008) and its proof is omitted.

Proof of (S.7). Plugging the expression $X_s-X_{j\Delta}=\int_{j\Delta}^s\mu(X_u)\,du+\int_{j\Delta}^s\sigma(X_u)\,dW_u$, we have

$$\begin{aligned}B(X_{j\Delta})\int_{j\Delta}^{(j+1)\Delta}\left[X_s-X_{j\Delta}\right]\mu(X_s)\,ds &= \int_{j\Delta}^{(j+1)\Delta}B^{1/2}(X_{j\Delta})\left[X_s-X_{j\Delta}\right]B^{1/2}(X_{j\Delta})[\mu(X_s)-\mu(X_{j\Delta})]ds\\ &\quad+B^{1/2}(X_{j\Delta})\mu(X_{j\Delta})\int_{j\Delta}^{(j+1)\Delta}\int_{j\Delta}^sB^{1/2}(X_{j\Delta})\,\mu(X_u)\,duds\\ &\quad+B(X_{j\Delta})\,\mu(X_{j\Delta})\int_{j\Delta}^{(j+1)\Delta}\int_{j\Delta}^s\sigma(X_u)\,dW_uds.\end{aligned}$$

The first and second terms on the RHS are $O_{a.s.}(\Delta^2\log(1/\Delta))$ and $O_{a.s.}(\Delta^2)$ uniformly over j, respectively, which holds by Theorem 1 and (20), where we note that $\mu(\cdot)$ is at most of polynomial growth (as supposed

in (A3.i)). Therefore, we can write

$$\begin{aligned}\mathcal{V}_1 &\leq 2\sup_{x\in\mathbb{R}}(1/nh)\textstyle\sum_{j=1}^{n-1}\left|K\left(\frac{X_{j\Delta}-x}{h}\right)\right|\times[O_{a.s.}(\Delta\log(1/\Delta))+O_{a.s.}(\Delta)]\\ &\quad +2\sup_{x\in\mathbb{R}}(1/nh)\textstyle\sum_{j=1}^{n-1}K\left(\frac{X_{j\Delta}-x}{h}\right)B(X_{j\Delta})\mu(X_{j\Delta})\int_{j\Delta}^{(j+1)\Delta}\int_{j\Delta}^{s}\sigma(X_u)\,dW_u ds\\ &= :2\mathcal{V}_{11}+2\mathcal{V}_{12}.\end{aligned}$$

By the uniform boundedness of $\pi(\cdot)$, we have $\sup_{x\in\mathbb{R}}(1/nh)\sum_{j=1}^{n-1}\left|K(\frac{X_{j\Delta}-x}{h})\right|=O_p(1)$, which is derived in the proof of Theorem 2 (see the term R_{11}), and therefore, $\mathcal{V}_{11}=O_P(\Delta\log(1/\Delta))$. For deriving the rate of $\mathcal{V}_{12}$, observe that

$$\begin{aligned}\int_{j\Delta}^{(j+1)\Delta}\int_{j\Delta}^{s}\sigma(X_u)\,dW_u ds &= \int_{j\Delta}^{(j+1)\Delta}\int_{u}^{(j+1)\Delta}ds\sigma(X_u)\,dW_u\\ &= \int_{j\Delta}^{(j+1)\Delta}[(j+1)\Delta-u]\,\sigma(X_u)\,dW_u,\end{aligned}$$

which holds by changing the order of (stochastic) integrals. Therefore, $\mathcal{V}_{12}$ can be written as the sum of martingale differences. It can be represented by a continuous martingale with index $r\in[0,1]$ (in the same way as the term U_3 in the proof of Theorem 4), and we can show that $\mathcal{V}_{12}=O_p(\sqrt{(\log n)/nh})$ in the same way as for $\mathcal{V}_2$ below (and we omit details for brevity).

Proof of (S.8). Let

$$\begin{aligned}\varrho_{s,j\Delta} &:= B(X_{j\Delta})\left[\textstyle\int_{j\Delta}^{s}\mu(X_u)\,du+\int_{j\Delta}^{s}\sigma(X_u)\,dW_u\right]\sigma(X_s);\\ \mathbf{e}_\Delta(s,j) &:= \mathbf{1}_{\left\{\sup_{v\in[j\Delta,s]}B^{1/2}(X_{j\Delta})\left[\left|\mu(X_v)-\mu(X_{j\Delta})\right|+\left|\sigma(X_v)-\sigma(X_{j\Delta})\right|\right]\leq 1\right\}},\end{aligned}$$

where $\mathbf{e}_\Delta$ is an indicator function defined for each (s,j) with $s\in[j\Delta,(j+1)\Delta]$. Using this $\mathbf{e}_\Delta$, we also define

$$\begin{aligned}\bar{\varrho}_{s,j\Delta} &:= \left\{\textstyle\int_{j\Delta}^{s}\{B^{1/2}(X_{j\Delta})\mu(X_{j\Delta})+B^{1/2}(X_{j\Delta})[\mu(X_u)-\mu(X_{j\Delta})]\mathbf{e}_\Delta(u,j)\}du\right.\\ &\quad \left.+\textstyle\int_{j\Delta}^{s}\{B^{1/2}(X_{j\Delta})\sigma(X_{j\Delta})+B^{1/2}(X_{j\Delta})[\sigma(X_u)-\sigma(X_{j\Delta})]\}\mathbf{e}_\Delta(u,j)\,dW_u\right\}\\ &\quad \times\left\{B(X_{j\Delta})\sigma(X_{j\Delta})+B(X_{j\Delta})[\sigma(X_s)-\sigma(X_{j\Delta})]\,\mathbf{e}_\Delta(s,j)\right\}. \qquad \text{(S.12)}\end{aligned}$$

Then, we can write

$$\varrho_{s,j\Delta}=\bar{\varrho}_{s,j\Delta}+\tilde{\varrho}_{s,j\Delta},$$

where $\tilde{\varrho}_{s,j\Delta}$ is defined through the same form as $\bar{\varrho}_{s,j\Delta}$ with $\mathbf{e}_\Delta(u,j)$ and $\mathbf{e}_\Delta(s,j)$ replaced by $[1-\mathbf{e}_\Delta(u,j)]$ and $[1-\mathbf{e}_\Delta(s,j)]$, respectively. Therefore, we can also write

$$\begin{aligned}\mathcal{V}_2 &= 2\sup_{x\in\mathbb{R}}\left|(1/Th)\textstyle\sum_{j=1}^{n-1}K\left(\frac{X_{j\Delta}-x}{h}\right)\int_{j\Delta}^{(j+1)\Delta}\bar{\varrho}_{s,j\Delta}dW_s\right|\\ &\quad +2\sup_{x\in\mathbb{R}}\left|(1/Th)\textstyle\sum_{j=1}^{n-1}K\left(\frac{X_{j\Delta}-x}{h}\right)\int_{j\Delta}^{(j+1)\Delta}\tilde{\varrho}_{s,j\Delta}dW_s\right|\\ &= :2\mathcal{V}_{21}+2\mathcal{V}_{22}.\end{aligned}$$

By Theorem 1, there exists some $\tilde{\Delta}>0$ such that for any $\Delta\leq\tilde{\Delta}$, $\max_{1\leq j\leq n-1}\sup_{s\in[j\Delta,(j+1)\Delta]}|1-\mathbf{e}_\Delta(s,j)|=0$ almost surely, implying that $\mathcal{V}_{22}=0$ almost surely for sufficiently small Δ. Therefore, the convergence rate of $\mathcal{V}_2$ is determined by that of $\mathcal{V}_{21}$.

To derive the rate of $\mathcal{V}_{21}$, we note that

$$\bar{\varrho}_{s,j\Delta} \leq C_0 \left\{ \Delta + \int_{j\Delta}^{s} \{ B^{1/2}(X_{j\Delta})\, \sigma(X_{j\Delta}) + B^{1/2}(X_{j\Delta}) \left[\sigma(X_u) - \sigma(X_{j\Delta}) \right] \} \mathbf{e}_{\Delta}(u,j)\, dW_u \right\} \tag{S.13}$$

for some constant $C_0 > 0$, which follows from the definition of $\bar{\varrho}_{s,j\Delta}$. Since the integrand of the stochastic integral on the RHS of (S.13) is uniformly bounded over j and $u \in [j\Delta, s]$, we can apply the same argument as those for (16) and (20) again. That is, we let

$$\bar{\varrho}_{s,j\Delta} = \bar{\varrho}_{s,j\Delta} \mathbf{1}_{\left\{ \bar{\varrho}_{s,j\Delta} \leq \sqrt{\Delta} \log \Delta^{-1} \right\}} + \bar{\varrho}_{s,j\Delta} \mathbf{1}_{\left\{ \bar{\varrho}_{s,j\Delta} > \sqrt{\Delta} \log \Delta^{-1} \right\}},$$

where the second term is exactly zero for sufficiently small Δ (uniformly over j and $s \in [j\Delta, (j+1)\Delta]$), implying that the rate of $\mathcal{V}_{21}$ is determined by that of

$$\bar{\mathcal{V}}_{21} := \sup_{x \in \mathbb{R}} \left| (1/Th) \sum_{j=1}^{n-1} K\left(\frac{X_{j\Delta} - x}{h} \right) \int_{j\Delta}^{(j+1)\Delta} \bar{\varrho}_{s,j\Delta} \mathbf{1}_{\left\{ \bar{\varrho}_{s,j\Delta} \leq \sqrt{\Delta} \log \Delta^{-1} \right\}} dW_s \right|.$$

To derive the rate of $\bar{\mathcal{V}}_{21}$, we consider a finite covering $\{\mathcal{K}_k(h)\}_{k=1}^{\nu(h)}$ of a set of functions $\mathcal{K}(h) := \{ K\left(\frac{p-x}{h} \right) \mid x \in \mathbb{R} \}$ (for each h). By Lemma 2, we can find $\{\mathcal{K}_k(h)\}_{k=1}^{\nu(h)}$ such that each $\mathcal{K}_k(h)$ has the center $g_k(\cdot) := K\left(\frac{\cdot - x_k}{h} \right)$; for any probability measure Q,

$$\forall \varepsilon \in (0,1),\ \forall g \in \mathcal{K}_k(h),\ \left\{ \int |g - g_k|^{\bar{r}}\, dQ \right\}^{1/\bar{r}} \leq \varepsilon 8 \bar{K}; \quad \text{and} \quad \nu(h) \leq \Lambda \varepsilon^{-4\bar{r}} \ \text{ for each } \varepsilon \in (0,1), \tag{S.14}$$

for some constant $\Lambda\,(>0)$ (independent of h) and for any $\bar{r} \geq 1$. Then, we have

$$\begin{aligned}
\bar{\mathcal{V}}_{21} &\leq \frac{n-1}{Th} \max_{k \in \{1,\ldots,v(h)\}} \sup_{g \in \mathcal{K}_k(h)} \frac{1}{n-1} \sum_{j=1}^{n-1} \left| g_k(X_{j\Delta}) - g(X_{j\Delta}) \right| \left| \int_{j\Delta}^{(j+1)\Delta} \bar{\varrho}_{s,j\Delta} \mathbf{1}_{\left\{ \bar{\varrho}_{s,j\Delta} \leq \sqrt{\Delta} \log \Delta^{-1} \right\}} dW_s \right| \\
&\quad + (1/Th) \max_{k \in \{1,\ldots,v(h)\}} \left| \sum_{j=1}^{n-1} g_k(X_{j\Delta}) \int_{j\Delta}^{(j+1)\Delta} \bar{\varrho}_{s,j\Delta} \mathbf{1}_{\left\{ \bar{\varrho}_{s,j\Delta} \leq \sqrt{\Delta} \log \Delta^{-1} \right\}} dW_s \right| \\
&= : \bar{\mathcal{V}}_{211} + \bar{\mathcal{V}}_{212}.
\end{aligned}$$

By the Hölder and Burkholder-Davis-Gundy (BDG) inequaliites, we have for any $\bar{r} > 1$,

$$\begin{aligned}
\bar{\mathcal{V}}_{211} &\leq \frac{n-1}{Th} \Big\{ \max_{k \in \{1,\ldots,v(h)\}} \sup_{g \in \mathcal{K}_k(h)} \frac{1}{n-1} \sum_{j=1}^{n-1} \left| g_k(X_{j\Delta}) - g(X_{j\Delta}) \right|^{\bar{r}} \Big\}^{1/\bar{r}} \\
&\quad \times \Big\{ \frac{1}{n-1} \sum_{j=1}^{n-1} \left| \int_{j\Delta}^{(j+1)\Delta} \bar{\varrho}_{s,j\Delta} \mathbf{1}_{\left\{ \bar{\varrho}_{s,j\Delta} \leq \sqrt{\Delta} \log \Delta^{-1} \right\}} dW_s \right|^{\bar{r}/(\bar{r}-1)} \Big\}^{(\bar{r}-1)/\bar{r}} \\
&\leq \frac{n-1}{Th} \varepsilon 8 \bar{K} \times \left\{ O_p(\Delta^{\bar{r}/(\bar{r}-1)} (\log \Delta^{-1})^{\bar{r}/(\bar{r}-1)}) \right\}^{(\bar{r}-1)/\bar{r}} \\
&= O_p(h^{-1} \varepsilon \log \Delta^{-1}) = O_p(h^{-1} \varepsilon \log n) = O_p(\sqrt{(\log n)/nh}),
\end{aligned} \tag{S.15}$$

where the last two equalities have used the condition $\Delta^{-1} \leq n^{\varkappa}$, implying that $\log \Delta^{-1} = O(\log n)$, as well as

$$\varepsilon = \sqrt{h/n(\log n)}. \tag{S.16}$$

To find the probability bound of the second term $\bar{\mathcal{V}}_{212}$, we note that $\sum_{j=1}^{n-1} g_k(X_{j\Delta}) \int_{j\Delta}^{(j+1)\Delta} \bar{\varrho}_{s,j\Delta} dW_s$ can be written as a continuous martingale indexed by $r \in [0,1]$ in the same was as in the proof Theorem 3 (see the expression for the term U_3), whose quadratic variation at $r = 1$ is given by

$$J_{n,T}(k) = \sum_{j=1}^{n-1} K^2\left(\frac{X_{j\Delta} - x_k}{h} \right) \int_{j\Delta}^{(j+1)\Delta} \bar{\varrho}_{s,j\Delta}^2 \mathbf{1}_{\left\{ \bar{\varrho}_{s,j\Delta} \leq \sqrt{\Delta} \log \Delta^{-1} \right\}} ds$$

with x_k being a point in $\mathbb{R}$ satisfying $g_k^2(\cdot) = K^2\left(\frac{\cdot - x_k}{h}\right)$. This $J_{n,T}(k)$ satisfies

$$J_{n,T}(k) \leq \bar{K}^2 T\Delta[\log(1/\Delta)]^2 \text{ uniformly over } k,$$

as well as

$$\begin{aligned} E[J_{n,T}(k)] &\leq n \max_{1\leq j\leq n-1} E[K^2\left(\tfrac{X_{j\Delta}-x_k}{h}\right)\bar{\varrho}_{s,j\Delta}^2] \\ &\leq C_J Th\Delta[\log(1/\Delta)]^2 \text{ uniformly over } k, \end{aligned}$$

for some constant $C_J > 0$, which follows from the BDG inequality (with the upper bound (S.13)) and the change-of-variable argument (with the uniform boundedness of π). Applying the exponential inequality for continuous martingales (Ex. 3.16 in Ch. IV of Revuz and Yor, 1999), for each $a > 0$ and each y, we have

$$\begin{aligned} &\Pr(\bar{\mathcal{V}}_{212} \geq a\sqrt{(\log n)/nh}) \\ \leq\ & v(h)\left\{\Pr\left((1/Th)\left|\textstyle\sum_{j=1}^{n-1} g_k(X_{j\Delta})\int_{j\Delta}^{(j+1)\Delta}\bar{\varrho}_{s,j\Delta}\mathbf{1}_{\left\{\bar{\varrho}_{s,j\Delta}\leq\sqrt{\Delta}\log\Delta^{-1}\right\}}dW_s\right| \geq a\sqrt{(\log n)/nh},\ J_{n,T}(k)\leq y\right)\right. \\ &\left. + \Pr\left(|J_{n,T}(k) - E[J_{n,T}(k)]| \geq y/2\right) + \Pr\left(E[J_{n,T}(k)] \geq y/2\right)\right\} \\ \leq\ & O(n^{2\bar{r}}h^{-2\bar{r}}(\log n)^{2\bar{r}}) \times \left\{2\exp\left\{-\frac{a^2[(\log n)/nh]T^2h^2}{aTh\Delta}\right\} + \Pr\left(|J_{n,T}(k) - E[J_{n,T}(k)]| \geq aTh\Delta\right)\right\} \\ =\ & O(n^{2\bar{r}}h^{-2\bar{r}}(\log n)^{2\bar{r}} \times n^{-a}) + O(n^{2\bar{r}}h^{-2\bar{r}}(\log n)^{2\bar{r}}) \times \Pr\left(|J_{n,T}(k) - E[J_{n,T}(k)]| \geq aTh\Delta\right), \end{aligned} \tag{S.17}$$

where the second inequality holds with $y = 2aTh\Delta$ (for sufficiently large a) since $v(h) \leq \Lambda\varepsilon^{-4\bar{r}}$, $\varepsilon^{-4\bar{r}} = O\left(n^{2\bar{r}}h^{-2\bar{r}}(\log n)^{2\bar{r}}\right)$ (recall $\varepsilon = \sqrt{h/n(\log n)}$ in (S.16)) and

$$\Pr\left(E\left[J_{n,T}(k)\right] \geq y/2\right) \leq \Pr\left(C_J Th\Delta \geq aTh\Delta\right) = 0.$$

The first term on the RHS of (S.17) tends to zero as $n \to \infty$ (for n large enough). To find the bound of the second term on the RHS of (S.17), we apply the Bernstein inequality in Lemma 3. To this end, write

$$J_{n,T}(k) - E\left[J_{n,T}(k)\right] = \textstyle\sum_{j=1}^{n-1}\left\{\mathcal{Z}_{j,n} - E\left[\mathcal{Z}_{j,n}\right]\right\},$$

where

$$\mathcal{Z}_{j,n} := K^2\left(\tfrac{X_{j\Delta}-x_k}{h}\right)\int_{j\Delta}^{(j+1)\Delta}\bar{\varrho}_{s,j\Delta}^2\mathbf{1}_{\{\bar{\varrho}_{s,j\Delta}\leq\sqrt{\Delta}\log\Delta^{-1}\}}ds.$$

By the boundedness of K and the definition of $\bar{\varrho}_{s,j\Delta}$ in (S.12), we can find some constant $C_{\mathcal{Z}} > 0$ such that $\max_{1\leq j\leq n-1}|\mathcal{Z}_{j,n}| \leq \bar{K}^2\Delta^2\left(\log\Delta^{-1}\right)^2$. By the change-of-variable argument and the BDG inequality, we can also find some constant $\varpi_{\mathcal{Z}}$ satisfying $E[|\sum_{j=1}^m \mathcal{Z}_{j,n}|^2] \leq \varpi_{\mathcal{Z}} m^2 h\Delta^4$. Given these, we have for $m \leq \min\{aTh\Delta/4\bar{K}^2\Delta^2\left(\log\Delta^{-1}\right)^2, n-1\}$, and for each $a > 0$,

$$\begin{aligned} &\Pr\left(|J_{n,T}(k) - E\left[J_{n,T}(k)\right]| \geq aTh\Delta\right) = \Pr(\textstyle\sum_{j=1}^{n-1}\mathcal{Z}_{j,n} \geq aTh\Delta) \\ \leq\ & 4\exp\left\{-\frac{-a^2T^2h^2\Delta^2}{64n\left(\varpi_{\mathcal{Z}}m\Delta^4h\right) + (4/3)\left[C_{\mathcal{Z}}\bar{K}^2\Delta^2\left(\log\Delta^{-1}\right)^2\right](aTh\Delta)\,m}\right\} + 4Anm^{-1-\beta}\Delta^{-\beta} \\ \leq\ & 4\exp\left\{-\frac{-a^2(\log n)}{64\varpi_{\mathcal{Z}}/\left(\log\Delta^{-1}\right)^2 + (4/3)\,C_{\mathcal{Z}}\bar{K}^2 a}\right\} + O(h^{-1-\beta}n^{-\beta}(\log n)^{3(1+\beta)}\Delta^{-\beta}) \\ =\ & O(n^{-a} + h^{-1-\beta}n^{-\beta}(\log n)^{3(1+\beta)}\Delta^{-\beta}), \end{aligned}$$

where we have set $m = nh/\left(\log n\right)\left(\log \Delta^{-1}\right)^2$ and used $\log \Delta^{-1} = O\left(\log n\right)$ to derive the last two lines. Therefore, using $h^{-1} = O(n^{\vartheta}\left(\log n\right)^{-1})$ and $\Delta^{-1} = O(n^{\varkappa})$, we can write the second term on the RHS of (S.17) as

$$\begin{aligned} & O(n^{2\bar{r}}h^{-2\bar{r}}(\log n)^{2\bar{r}})[n^{-a} + h^{-1-\beta}n^{-\beta}\left(\log n\right)^{3(1+\beta)}\Delta^{-\beta}] \\ = \; & O(n^{-a+2\bar{r}+4\vartheta} + n^{-\beta(1-\vartheta-\varkappa)+2\bar{r}+\vartheta(2\bar{r}+1)}(\log n)^{4\bar{r}+4(1+\beta)}), \end{aligned}$$

which tends to zero as $n \to \infty$ for a large enough if

$$\frac{2\bar{r} + \vartheta(2\bar{r}+1)}{1-\vartheta-\kappa} < \beta.$$

Since we may pick any $\bar{r} > 1$ in (S.15), we have this inequality satisfied as long as

$$\frac{2+3\vartheta}{1-\vartheta-\kappa} < \beta.$$

Therefore, given this condition on β, we have shown that $\bar{\mathcal{V}}_{212} = O_p(\sqrt{(\log n)/nh})$, completing the proof of (S.8).

Proof of (S.9). The convergence rate of $\mathcal{V}_3$ can be derived in the same way as those of $\mathcal{V}_1$ and $\mathcal{V}_2$, and we here outline only main points. Since $\sigma^2\left(\cdot\right)$ is twice continuously differentiable, we can apply the Ito lemma to $\sigma^2\left(X_s\right) - \sigma^2\left(X_{j\Delta}\right)$ to obtain

$$\begin{aligned} \mathcal{V}_3 \leq & \sup_{x\in\mathbb{R}}\left|(1/Th)\textstyle\sum_{j=1}^{n-1} K\left(\frac{X_{j\Delta}-x}{h}\right) B\left(X_{j\Delta}\right)\int_{j\Delta}^{(j+1)\Delta}\int_{j\Delta}^{s} m_3\left(X_u\right) du ds\right| \\ & + \sup_{x\in\mathbb{R}}\left|(1/Th)\textstyle\sum_{j=1}^{n-1} K\left(\frac{X_{j\Delta}-x}{h}\right) B\left(X_{j\Delta}\right)\int_{j\Delta}^{(j+1)\Delta}\int_{j\Delta}^{s} \partial\sigma^2\left(X_u\right)\sigma\left(X_u\right) dW_u ds\right|, \end{aligned}$$

where $m_3\left(x\right) := \partial\sigma^2\left(x\right)\mu\left(x\right) + \partial^2\sigma^2\left(x\right)\sigma^2\left(x\right)/2$. Since this $\psi\left(x\right) = m_3\left(x\right)$ satisfies the conditions of Theorem 1, we have $B\left(X_{j\Delta}\right) m_3\left(X_u\right) = O_{a.s.}\left(1\right)$ uniformly as discussed in (20), and we can show that the first term on the RHS is $O_p\left(\Delta\right)$. The second term is $O_p(\sqrt{\left(\log n\right)/nh})$, which follows from the same arguments as those for $\mathcal{V}_{12}$ and $\mathcal{V}_2$ (we omit details for brevity).

Proof of (S.10). We look at

$$\begin{aligned} & \mathcal{V}_4 = \textstyle\int K\left(q\right)\left[H\left(qh+x\right) - H\left(x\right)\right] dq + \int K\left(q\right)\left[l\left(qh+x\right) - l\left(x\right)\right]\sigma^2\left(x\right) dp \\ \leq \; & \left(h^2/2\right)\textstyle\int q^2 K\left(q\right) dq \times \sup_{x\in\mathbb{R}}\left|H''\left(x\right)\right| + \left(h^2/2\right)\int q^2 K\left(q\right)\left|l''\left(\lambda qh + x\right)\right|\sigma^2\left(x\right) dq, \end{aligned}$$

where we have set $H\left(x\right) = B\left(x\right)\pi\left(x\right)\sigma^2\left(x\right)$ and $l\left(x\right) = B\left(x\right)\pi\left(x\right)$; the inequality follows from the usual Talyor-expansion argument with $\lambda \in [0,1]$ (which depends on q, h and x). Then, we can check that $\mathcal{V}_4 = O\left(h^2\right)$, since we have $\sup_{x\in\mathbb{R}}\left|H''\left(x\right)\right| < \infty$ (by the condition that $\sigma^2 \in \mathcal{D}(B,\pi)$), as well as the compactness of the support of K and the following bound:

$$\begin{aligned} & \left|l''\left(\lambda qh + x\right)\right|\sigma^2\left(x\right)\mathbf{1}_{\{|q|\leq c_K\}} \leq 4C_B\left[\max_{k=0,1,2}\sup_{x\in\mathbb{R}}\left|\partial^k\pi\left(x\right)\right|\right] \times \sup_{x,y\in\mathbb{R}} B\left(y\right)\sigma^2\left(x\right)\mathbf{1}_{\{|x-y|\leq c_K h\}} \\ \leq \; & 4C_B\left[\max_{k=0,1,2}\sup_{x\in\mathbb{R}}\left|\partial^k\pi\left(x\right)\right|\right] \times \sup_{x,y\in\mathbb{R}} B\left(y\right)\bar{C}_0[1+|x|^{\bar{q}_2}]\mathbf{1}_{\{|x-y|\leq c_K h\}} \\ \leq \; & \bar{C}_1 + \bar{C}_1 B\left(y\right)\left[|y| + |x-y|\right]^{\bar{q}_2+1}\mathbf{1}_{\{|x-y|\leq c_K h\}} \\ \leq \; & \bar{C}_1 + \bar{C}_1 B\left(y\right) 2^{\bar{q}_2}\left[|y|^{\bar{q}_2+1} + |x-y|^{\bar{q}_2+1}\right]\mathbf{1}_{\{|x-y|\leq c_K h\}} \\ \leq \; & \bar{C}_1 + \bar{C}_1 2^{\bar{q}_2}\left[\sup_{y\in\mathbb{R}} B\left(y\right)|y|^{\bar{q}_2+1} + \sup_{y\in\mathbb{R}} B\left(y\right)\left(c_K h\right)^{\bar{q}_2+1}\right] < \infty, \end{aligned} \tag{S.18}$$

where the first inequality holds since $\max_{k=0,1,2}\sup_{x\in\mathbb{R}}\left|\partial^k\pi\left(x\right)\right| < \infty$ and $\max_{k=1,2}\left|\partial^k B\left(x\right)\right| \leq C_B \times B\left(x\right)$; the second holds since there exists some constant $\bar{C}_0 > 0$ such that $\sigma^2\left(x\right) \leq \bar{C}_0\left[1+\left|x\right|^{\bar{q}_2+1}\right]$ (by the condition that $\sigma^2\left(x\right) = O(|x|^{\bar{q}_2})$ as $|x| \to \infty$ for $\bar{q}_2 \geq 0$); the third holds with a constant $\bar{C}_1 = 4C_B\left[\max_{k=0,1,2}\sup_{x\in\mathbb{R}}\left|\partial^k\pi\left(x\right)\right|\right]\bar{C}_0 \in (0,\infty)$; and the fourth follows from the Jensen inequality.

Proof of (S.11). To bound the term $\mathcal{V}_5$, we consider a compact set $[-\mathcal{T}_n, \mathcal{T}_n] \subset \mathbb{R}$ with $\mathcal{T}_n \to \infty$ whose growing rate is specified below, and its finite covering $\{\mathcal{I}_k\}_{k=1}^{\nu(n)}$ such that $[-\mathcal{T}_n, \mathcal{T}_n] \subset \cup_{k=1}^{\nu(n)}\mathcal{I}_k$, each $\mathcal{I}_k$ is a closed ball in $\mathbb{R}$ with its center x_k and radius r_n, and $\nu\left(n\right) = \mathcal{T}_n/r_n$. Then, we can write

$$\begin{aligned}
\mathcal{V}_5\left(x\right) &\leq \sup_{|x|>\mathcal{T}_n}\left(1/nh\right)\textstyle\sum_{j=1}^{n-1}\left|\Gamma_{j\Delta}\left(x\right) - E\left[\Gamma_{j\Delta}\left(x\right)\right]\right| \\
&\quad + \max_{k\in\{1,\ldots,\nu(h)\}}\sup_{x\in\mathcal{I}_k}\left(1/nh\right)\textstyle\sum_{j=1}^{n-1}\{\left|\Gamma_{j\Delta}\left(x\right) - \Gamma_{j\Delta}(x_k)\right| + \left|E[\Gamma_{j\Delta}\left(x\right)] - E[\Gamma_{j\Delta}(x_k)]\right|\} \\
&\quad + \max_{k\in\{1,\ldots,\nu(h)\}}\left|\left(1/nh\right)\textstyle\sum_{j=1}^{n-1}\{\Gamma_{j\Delta}(x_k) - E[\Gamma_{j\Delta}(x_k)]\}\right| \\
&= : \mathcal{V}_{51} + \mathcal{V}_{52} + \mathcal{V}_{53},
\end{aligned}$$

where we below consider the bounds of these three terms.

To find the bound of $\mathcal{V}_{51}$, let

$$\bar{\Gamma}_{j\Delta}\left(x\right) := \Gamma_{j\Delta}\left(x\right)\mathbf{1}_{\{|X_{j\Delta}|\leq\mathcal{T}_n/2\}}; \quad \text{and} \quad \tilde{\Gamma}_{j\Delta}\left(x\right) := \Gamma_{j\Delta}\left(x\right)\mathbf{1}_{\{|X_{j\Delta}|>\mathcal{T}_n/2\}},$$

and observe that

$$\mathcal{V}_{51} \leq \sup_{|x|>\mathcal{T}_n}\left(1/nh\right)\textstyle\sum_{j=1}^{n-1}\left|\bar{\Gamma}_{j\Delta}\left(x\right) - E[\bar{\Gamma}_{j\Delta}\left(x\right)]\right| + \sup_{|x|>\mathcal{T}_n}\left(1/nh\right)\textstyle\sum_{j=1}^{n-1}\left|\tilde{\Gamma}_{j\Delta}\left(x\right) - E[\tilde{\Gamma}_{j\Delta}\left(x\right)]\right|, \tag{S.19}$$

For $|x| > \mathcal{T}_n$ and $|X_{j\Delta}| \leq \mathcal{T}_n/2$, it holds that $\left(X_{j\Delta} - x\right)/h \geq \mathcal{T}_n/2h$. Therefore, for sufficiently large n with $\mathcal{T}_n/2h \geq c_K$, $K\left(\left(X_{j\Delta} - x\right)/h\right) = 0$ and the first term on the RHS of (S.19) is zero. To find the bound of the second term, we observe that

$$\begin{aligned}
&\sup_{|x-y|\leq c_K h} B^{1/2}\left(y\right)\left|\sigma^2\left(y\right) - \sigma^2\left(x\right)\right| = \sup_{|x-y|\leq c_K h} B^{1/2}\left(y\right)\left|\partial\sigma^2\left(y + \lambda(x-y)\right)\right|\left|x-y\right| \\
\leq\ &\sup_{|x-y|\leq c_K h} B^{1/2}\left(y\right)\bar{C}_2\left[1 + \left|y + \lambda(x-y)\right|^{\bar{q}_2+1}\right]c_K h \\
\leq\ &O\left(h\right) + \bar{C}_2\sup_{|x-y|\leq c_K h} B^{1/2}\left(y\right)2^{\bar{q}_2}\left[\left|y\right|^{\bar{q}_2+1} + \left|\lambda(x-y)\right|^{\bar{q}_2+1}\right]c_K h = O(h),
\end{aligned} \tag{S.20}$$

for some constant $\bar{C}_2 > 0$, where the first equality follows from the mean-value theorem with some $\lambda \in [0,1]$ (which depends on x and y); the two inequalities use the polynomial growth condition of $\partial\sigma^2$ and the Jensen inequality. Then, we have

$$\begin{aligned}
&\sup_{|x|>\mathcal{T}_n}\left(1/nh\right)\textstyle\sum_{j=1}^{n-1}\left|\tilde{\Gamma}_{j\Delta}\left(x\right)\right| \\
\leq\ &\left(1/nh\right)\bar{K}\textstyle\sum_{j=1}^{n-1}\mathbf{1}_{\{|X_{j\Delta}-x|\leq c_K h\}}B^{1/2}(X_{j\Delta})\left|\sigma^2(X_{j\Delta}) - \sigma^2\left(x\right)\right| \times B^{1/2}(X_{j\Delta})\mathbf{1}_{\{|X_{j\Delta}|>\mathcal{T}_n/2\}} \\
\leq\ &\left(1/nh\right)\bar{K}\textstyle\sum_{j=1}^{n-1}\sup_{|x-y|\leq c_K h} B^{1/2}\left(y\right)\left|\sigma^2\left(y\right) - \sigma^2\left(x\right)\right| \times B^{1/2}\left(X_{j\Delta}\right)\left|X_{j\Delta}\right|^{\bar{d}}/\left(\mathcal{T}_n/2\right)^{\bar{d}} \\
=\ &O(1/\mathcal{T}_n^{\bar{d}}),
\end{aligned}$$

for any $\bar{d} > 0$, where the last equality holds since $B^{1/2}\left(x\right)\left|x\right|^d$ is uniformly bounded. We can also show that $\sup_{|x|>\mathcal{T}_n}\left(1/nh\right)\sum_{j=1}^{n-1}\left|\tilde{\Gamma}_{j\Delta}\left(x\right)\right| = O(1/\mathcal{T}_n^{\bar{d}})$ exactly in the same way, and therefore, can conclude that $\mathcal{V}_{51} = O(1/\mathcal{T}_n^{\bar{d}})$.

To bound the term $\mathcal{V}_{52}$, note that for any $x \in \mathcal{I}_k$, $|x - x_k| \leq r_n$. This implies that for an event

$$E_{n,j\Delta}(x, x_k) := \{\max\{|X_{j\Delta} - x|, |X_{j\Delta} - x_k|\} \leq c_K h + r_n\},$$

which is defined in Ω for each $(n, j\Delta, x, x_k)$, we have

$$\begin{aligned} \{|X_{j\Delta} - x| &\leq c_K h\} = \{|X_{j\Delta} - x| \leq c_K h \ \& \ |X_{j\Delta} - x_k| \leq c_K h + r_n\} \subset E_{n,j\Delta}(x, x_k); \\ \{|X_{j\Delta} - x_k| &\leq c_K h\} \subset E_{n,j\Delta}(x, x_k). \end{aligned}$$

Therefore, for any $x \in \mathcal{I}_k$,

$$\begin{aligned} & |\Gamma_{j\Delta}(x) - \Gamma_{j\Delta}(x_k)| \\ \leq\ & B(X_{j\Delta})\sigma^2(X_{j\Delta})\left|K\left(\tfrac{X_{j\Delta}-x}{h}\right) - K\left(\tfrac{X_{j\Delta}-x_k}{h}\right)\right| + K\left(\tfrac{X_{j\Delta}-x}{h}\right) B(X_{j\Delta}) \mathbf{1}_{E_{n,j}(x,x_k)} \left|\sigma^2(x) - \sigma^2(x_k)\right| \\ & + B(X_{j\Delta})\sigma^2(x_k)\mathbf{1}_{\left\{|X_{j\Delta}-x| \leq c_K h \text{ or } |X_{j\Delta}-x| \leq c_K h\right\}} \left|K\left(\tfrac{X_{j\Delta}-x}{h}\right) - K\left(\tfrac{X_{j\Delta}-x_k}{h}\right)\right| \\ \leq\ & \left\{\sup_{x\in\mathbb{R}} B(x)\sigma^2(x)\right\}\bar{K}|x_k - x|/h + \bar{K}B(X_{j\Delta})\mathbf{1}_{E_{n,j}(x,x_k)}\left|\sigma^2(x) - \sigma^2(x_k)\right| \\ & + B(X_{j\Delta})\sigma^2(x_k)\mathbf{1}_{E_{n,j}(x,x_k)}\bar{K}|x_k - x|/h \\ \leq\ & O(r_n/h) + O(r_n) + O(r_n/h) = O(r_n/h), \end{aligned}$$

uniformly over x, x_k and j, where the first inequality uses the triangular inequalities as well as the fact that if $|X_{j\Delta} - x| > c_K h$ and $|X_{j\Delta} - x| > c_K h$, then $K(\frac{X_{j\Delta}-x}{h}) = K(\frac{X_{j\Delta}-x_k}{h}) = 0$ and $\mathbf{1}_{\{|X_{j\Delta}-x| \leq c_K h \text{ or } |X_{j\Delta}-x| \leq c_K h\}} \leq \mathbf{1}_{E_{n,j}(x,x_k)}$; the second inequality uses the Lipschitz continuity and uniform boundedness of K; and the third inequality holds since we have

$$\begin{aligned} & B(X_{j\Delta})\mathbf{1}_{E_{n,j}(x,x_k)}\left|\sigma^2(x) - \sigma^2(x_k)\right| \\ \leq\ & B(X_{j\Delta})\mathbf{1}_{E_{n,j}(x,x_k)}\left|\partial\sigma^2(\lambda x + (1-\lambda)x_k)\right| \times |x - x_k| \\ \leq\ & \left\{B(X_{j\Delta})\left|\partial\sigma^2(X_{j\Delta})\right| + B(X_{j\Delta})\left|\partial\sigma^2(X_{j\Delta}) - \partial\sigma^2(\lambda x + (1-\lambda)x_k)\right|\right\}\mathbf{1}_{E_{n,j}(x,x_k)} \times |x - x_k| \\ \leq\ & \left\{\sup_{x\in\mathbb{R}} B(x)\left|\partial\sigma^2(x)\right| + \sup_{|x-y|\leq c_K h + r_n} B(x)\left|\partial\sigma^2(x) - \partial\sigma^2(y)\right|\right\} \times r_n \\ =\ & O(r_n); \text{ and} \end{aligned}$$

$$B(X_{j\Delta})\sigma^2(x_k)\mathbf{1}_{E_{n,j}(x,x_k)} = O(1),$$

uniformly over x, x_k and j, which follow from the same arguments as used to derive (S.18) and (S.20). From these we can conclude that $\mathcal{V}_{52} = O(r_n/h^2)$. Now, by setting $\mathcal{T}_n = [nh/(\log n)]^{1/2\bar{d}}$ and $r_n = \sqrt{h^3(\log n)/n}$, we can let both $\mathcal{V}_{51}(= (1/\mathcal{T}_n^{\bar{d}}))$ and $\mathcal{V}_{52}$ be $O_P(\sqrt{(\log n)/nh})$.

Finally, to investigate the rate of $\mathcal{V}_{53}$, we derive the bound $\Sigma^2_{\Gamma,m} := \sum_{j=1}^m \{\Gamma_{j\Delta}(x_k) - E[\Gamma_{j\Delta}(x_k)]\}$ for $m \leq (n-1)$ and apply the exponential inequality. For this purpose, we observe that

$$\begin{aligned} & E[|\{\Gamma_{j\Delta}(x_k) - E[\Gamma_{j\Delta}(x_k)]\}|^2] \\ \leq\ & 2h\int K^2(q)B^2 \cdot \pi(qh + x_k)[\sigma^2(qh + x_k) - \sigma^2(x_k)]^2 dq = O(h^3), \end{aligned}$$

uniformly over k and j, where the last equality follows from (S.20). Then, we can find some constant $\tilde{\omega} \in (0, \infty)$ such that $\Sigma^2_{\Gamma,m} \leq \tilde{\omega}m^2h^3$, and also some constant $C_\Gamma > 0$ satisfying

$$\begin{aligned} \Gamma_{j\Delta}(x) &= K\left(\tfrac{X_{j\Delta}-x}{h}\right)B(X_{j\Delta})\left|\partial\sigma^2(\lambda X_{j\Delta} - (1-\lambda)x)\right||X_{j\Delta} - x|\mathbf{1}_{\left\{|X_{j\Delta}-x| \leq c_K h\right\}} \\ &\leq \bar{K}\sup_{x,y\in\mathbb{R};\ |\epsilon|\leq\bar{\epsilon}} B(x)\left|\partial\sigma^2(y)\right| c_K h \leq C_\Gamma h, \end{aligned}$$

which follows from the compactness of the support of K. Then, applying Lemma 3, we have for each $a > 0$,

$$\begin{aligned}
& \Pr(\mathcal{V}_{53} \geq a\sqrt{(\log n)/nh}) \\
\leq\ & \textstyle\sum_{k=1}^{\nu(n)} \Pr\left(\left|\sum_{j=1}^{n-1}\{\Gamma_{j\Delta}(x_k) - E[\Gamma_{j\Delta}(x_k)]\}\right| \geq a\sqrt{(\log n)/nh}nh\right) \\
\leq\ & \nu(n)\left\{4\exp\left\{-\frac{a^2(\log n)}{64\tilde{\omega}mh^2 + (8/3)(2C_\Gamma h)\, a\sqrt{(\log n)/nh}m}\right\} + \frac{4n}{m}A(m\Delta)^{-\beta}\right\} \\
\leq\ & (\mathcal{T}_n/r_n)\left\{4\exp\left\{-\frac{a^2(\log n)}{64\tilde{\omega}\sqrt{nh^5/(\log n)} + 16C_\Gamma a/3}\right\} + 4Anm^{-1-\beta}\Delta^{-\beta}\right\} \\
=\ & (\mathcal{T}_n/r_n) \times 4n^{-a^2/\left[64\tilde{\omega}\tilde{C}+16C_\Gamma a/3\right]} + 4A(\mathcal{T}_n/r_n)\, nm^{-1-\beta}\Delta^{-\beta}, \qquad \text{(S.21)}
\end{aligned}$$

where the third inequality holds by setting $m = \sqrt{n/h(\log n)}$ ($\leq \min\{(a\sqrt{(\log n)/nh}nh)/C_\Gamma h, n-1\}$ for a large enough), and the last equality holds since $nh^5/(\log n) \leq \bar{C}_5$ for some constant $C_5 > 0$ (31). Now, by the definitions of $\mathcal{T}_n(= [nh/(\log n)]^{1/2\bar{d}})$ and $r_n(= \sqrt{h^3(\log n)/n})$, we have

$$\nu(n) = \mathcal{T}_n/r_n = (\log n)^{-\left(1+1/\bar{d}\right)/2} h^{-\left(3-1/\bar{d}\right)/2} n^{\left(1+1/\bar{d}\right)/2},$$

which is a polynomial order of n, and the first term on the RHS of (S.21) tends to zero as $n \to \infty$ for a large enough. As for the second term, recalling the definition of m and the conditions $\Delta^{-1} \leq n^\kappa$, we have

$$(\mathcal{T}_n/r_n) \times nm^{-1-\beta}\Delta^{-\beta} \leq (\log n)^{-\left(\beta+2+1/\bar{d}\right)/2} h^{\left(\beta-2+1/\bar{d}\right)/2} n^{\left(-\beta+2\beta\kappa+2+1/\bar{d}\right)/2},$$

which tends to zero (as $n \to \infty$) if

$$-\beta + 2\beta\kappa + 2 < 0 \Longleftrightarrow 2/(1-2\kappa) < \beta,$$

where we note that $\bar{d}$ can be any arbitrarily large integer. Now, the proof of Theorem 5 is completed. □

10 Convergence Results When Mixing Coefficients Decay Slowly

In this section, we present some results which complement convergence results in Theorems 2-5, focusing on the case when the decaying rate of the mixing coefficients in (10) is slow.

General Convergence Results with Possibly Small β: Theorem 2 requires at least $\beta > 5$ as in the condition (23), but the following theorem allows for any $\beta > 0$. At the price of possibly small β, we must have a slower convergence rate of $\sqrt{(\log T)/T^\theta h}$ (than that of $\sqrt{(\log T)/Th}$ in Theorem 2). We also note that the smaller β requires the smaller θ, implying the slower convergence rate of the bandwidth h through (22).

Theorem 7. *Suppose that the same conditions as in Theorem 2 hold with the condition (23) replaced by $\beta \geq 5\theta/(1-\theta)$. Then, it holds that as $n, T \to \infty$ and $\Delta, h \to 0$,*

$$\sup_{x\in I} |G_{n,T}(x) - E[G_{n,T}(x)]| = O_p(\sqrt{\Delta \log(1/\Delta)} + \sqrt{(\log T)/T^\theta h}).$$

Proof of Theorem 7. The proof proceeds in the same way as in that of Theorem 2. Since only the rate of the term R_2 differs, we omit details and outline only main points for $R_2 \leq R_{21} + R_{22}$. To find the rate of R_{21}, we set

$$\varepsilon = h\sqrt{(\log T)/T^{\theta}h}, \tag{S.22}$$

instead of (63). This means that $R_{21} = O(\sqrt{(\log T)/T^{\theta}h})$. For deriving the rate of R_{22}, we use the Bernstein inequality. To this end, observe the following moment bound:

$$E[|\sum_{j=1}^{m} Y_{n,j}(k,h)|^2] \leq m\sum_{j=1}^{m} E\left[Y_{n,j}^2(k,h)\right] \leq \varpi m^2 h,$$

uniformly over k and h, where the first inequality holds by the Jensen inequality for $m \leq (n-1)$, and the second by the moment bound derived in (67).

Now, we apply (49) to $\sum_{j=1}^{n-1} Y_{n,j}(k,h)$ with $Z_{n,j} = Y_{n,j}(k,h)$ and $\Sigma_m^2 = \Sigma_m^2(k,h) := E[(\sum_{j=1}^{m} Y_{n,j}(k,h))^2]$ for each (k,h). Let $\eta = a\left[(\log T)/T^{\theta}h\right]^{1/2} nh$ and $m = T^{(1-\theta)}/\Delta$ in (49), where $m \leq (n-1)$ and $m < \eta/4C_Y$ are satisfied for large T (since $\theta \in (0,1)$ and $(\log T)/T^{\theta}h \to 0$). Then, it holds that for any $a > 0$,

$$\begin{aligned}
&\Pr(R_{22} \geq a\sqrt{(\log T)/T^{\theta}h}) \\
&\leq \textstyle\sum_{k=1}^{\nu(h)} \Pr(|\sum_{j=1}^{n-1} Y_{n,j}(k,h)| \geq a\sqrt{(\log T)/T^{\theta}h}nh) \\
&\leq \nu(h)\left\{4\exp\left\{-\frac{a^2\left[(\log T)/T^{\theta}h\right]n^2h^2}{64n\varpi mh + (8/3)C_Y a\sqrt{(\log T)/T^{\theta}h}nhm}\right\} + \frac{4n}{m}\alpha(m\Delta)\right\} \\
&\leq 4\Lambda\varepsilon^{-4}\left\{\exp\left\{-\frac{a^2\log T}{64\varpi + (8/3)C_Y a\sqrt{(\log T)/T^{\theta}h}}\right\} + AT^{\theta-\beta(1-\theta)}\right\} \\
&\leq 4\Lambda(\log T)^{-4}\left\{T^{4\theta-a^2/[64\varpi+(8/3)C_Y a]} + AT^{5\theta-\beta(1-\theta)}\right\},
\end{aligned} \tag{S.23}$$

where the second inequality holds by (49) and (67); the third inequality uses (10) in (A2), "$\nu(h) \leq \Lambda\varepsilon^{-4}$" and "$m = T^{(1-\theta)}/\Delta$"; and the last inequality holds for large T since $\varepsilon = h\sqrt{(\log T)/T^{\theta}h}$ (which is set in (S.22)), $h^{-2} \leq T^{2\theta}(\log T)^{-2}$ and $\sqrt{(\log T)/T^{\theta}h} \leq 1$ (for large T). Therefore, for $a > 0$ large enough,

$$\Pr(R_{22} \geq a\sqrt{(\log T)/T^{\theta}h}) \to 0 \ \text{ as } T \to \infty, \tag{S.24}$$

if $5\theta - \beta(1-\theta) \leq 0$ ($\Leftrightarrow \theta \leq \beta/(5+\beta)$). This, together with the rate of R_{21}, we have $R_{22} = O_p(\sqrt{(\log T)/T^{\theta}h})$ as desired. □

The next theorem also concerns the small-β case. While Theorem 3 allows for any $\beta > 0$ (unlike Theorem 2), its probability bound is associated with the convergence rate of $\sqrt{(\log T)/Th}$. If we have a slower rate of $\sqrt{(\log T)/T^{\theta}h}$ (as in Theorem 7), we can derive a sharper inequality for the probability bound of $M_{n,T}(x)$:

Theorem 8. *Suppose that the same conditions as in Theorem 3 hold. Then, as $n, T \to \infty$ and $\Delta, h \to 0$, it holds that for each $a(>0)$ and each $x \in I$,*

$$\Pr(M_{n,T}(x) \geq a\sqrt{(\log T)/T^{\theta}h}) \leq 2T^{-a^2/2} + 4\exp\{-aC_M(Th)^{1-\theta}\} + 4AT^{-\beta}h^{-\theta(\beta+1)}, \tag{S.25}$$

for each $\theta \in (0,1)$, where $C_M(>0)$ is some constant independent of x.

Proof of Theorem 8. Since we use the same arguments as those for Theorem 3, we only outline main points. Given the same notation as in the proof of Theorem 3 we have for any $\theta \in (0,1)$,

$$\begin{aligned}
&\Pr(M_{n,T}(x) \geq a\sqrt{(\log T)/T^{\theta}h}) \\
&\leq \Pr(|N_1(x,h)| \geq aTh\sqrt{(\log T)/T^{\theta}h}) \\
&\leq \Pr(|N_1(x,h)| \geq aTh\sqrt{(\log T)/T^{\theta}h},\ \langle N(x,h)\rangle_1 \leq y) + \Pr(\langle N(x,h)\rangle_1 > y) \\
&\leq 2\exp\left\{-a^2(\log T)/2\right\} + \Pr\left(\langle N(x,h)\rangle_1 > aT^{2-\theta}h\right) \\
&= 2T^{-a^2/2} + \Pr\left(\langle N(x,h)\rangle_1 > aT^{2-\theta}h\right), \qquad \text{(S.26)}
\end{aligned}$$

where the third inequality holds by (70) with $\eta = aTh\sqrt{(\log T)/T^{\theta}h}$ and $y = aT^{2-\theta}h$. By applying the Bernstein inequality for mixing arrays in Lemma 3 to $\langle N(x,h)\rangle_1$, and using arguments quite analogous to those for (72)-(74), we can also derive

$$\Pr(\langle N(x,h)\rangle_1 > aT^{2-\theta}h) \leq 4\exp\{-aC_M(Th)^{1-\theta}\} + 4AT^{-\beta}h^{-\theta(\beta+1)},$$

which, together with (S.26), implies the desired result. □

Uniform Convergence Rates of Nadaraya-Watson Type Estimators with Possibly Small β: The next two theorems are small-β counterparts of Theorems 4-5 in Section 54. While Theorem 10 on the convergence of the diffusion function estimator provides rates in terms of T (slower than $\sqrt{(\log n)/nh}$), it imposes only minimal conditions, allowing for discontinuous kernels with unbounded support and relaxing conditions on derivatives of π and σ:

Theorem 9 (Drift Function Estimation with Possibly Small β)**.** *Suppose that the same conditions as in Theorem 4 hold, but replace the condition on the exponent of the mixing coefficient β by*

$$\beta \geq \max\left\{5\theta/(1-\theta), \left(4\theta+\theta^2+2\kappa\right)/\left(1-\theta^2\right)\right\}.$$

Then the convergence results in (29)-(30) hold with $a^{\star}_{n,T}$ replaced by

$$a^{*}_{n,T} := h^2 + \sqrt{\Delta\log(1/\Delta)} + \sqrt{(\log T)/T^{\theta}h}.$$

Proof of Theorem 9. Proof arguments proceed in the same way as those for the proof of Theorem 4, while we employ convergence results of Theorems 7-8, instead of those of Theorems 2-3. We omit details for brevity. □

Theorem 10 (Diffusion Function Estimation with Possibly Small β)**.** *Suppose that Assumption 1 holds;* $\sup_{x\in\mathbb{R}}\pi(x) < \infty$*; the observation interval Δ and the bandwidth h satisfy*

$$\Delta^{-1} = O(T^{\kappa}) \quad \text{and} \quad (\log T)/T^{\theta}h \to 0,$$

as $T \to \infty$ and $\Delta, h \to 0$, for some constants $\kappa > 0$ and $\theta \in (0,1)$;

$$\sigma^2(\cdot) \in \mathcal{D}(B,\pi); \quad [|\partial\mu(x)| + |\partial\sigma(x)|] = O(|x|^{\tilde{q}_2}) \text{ as } |x| \to \infty \text{ for some } \tilde{q}_2 \geq 0.$$

Let $c_{n,T}$, $\delta_{n,T}$, $a^{\star}_{n,T}$ and $a^{*}_{n,T}$ be sequences defined in Theorems 4 and 9. Then, the following results hold (as $n, T \to \infty$ and $\Delta, h \to 0$):

(i-a) *If*

$$\beta \geq \max\left\{5\theta/(1-\theta), \left(4\theta+\theta^2\right)/\left(1-\theta^2\right)\right\},$$

then,

$$\sup_{x\in\mathbb{R}} |\hat{\Psi}_{\sigma^2}(x) - B(x)\sigma^2(x)\pi(x)| = O_p\left(a^*_{n,T}\right). \tag{S.27}$$

(i-b) *Further if* $a^*_{n,T}/\delta_{n,T} \to 0$,

$$\sup_{|x|\le c_T} |\hat{\sigma}^2(x) - \sigma^2(x)| = O_p\left(a^*_{n,T}/\delta_{n,T}\right). \tag{S.28}$$

(ii) *If*

$$\beta \ge \max\left\{5(1+\theta)/(1-\theta), (2+3\theta)/(1-\theta)\right\},$$

then, the convergence results in (i-a) and (i-b) hold with $a^*_{n,T}$ *replaced by* $a^\star_{n,T}$.

Proof of Theorem 10. Using (S.6), we split the LHS of (S.27) into three terms:

$$\sup_{x\in\mathbb{R}} |\hat{\Psi}_{\sigma^2}(x) - B(x)\sigma^2(x)\pi(x)| \le \textstyle\sum_{i=1}^4 V_i,$$

where

$$\begin{aligned}
V_1 &:= \sup_{x\in\mathbb{R}} \left|(2/Th)\textstyle\sum_{j=1}^{n-1} K\left(\frac{X_{j\Delta}-x}{h}\right) B(X_{j\Delta}) \int_{j\Delta}^{(j+1)\Delta} [X_s - X_{j\Delta}]\,\mu(X_s)\,ds\right|;\\
V_2 &:= \sup_{x\in\mathbb{R}} \left|(2/Th)\textstyle\sum_{j=1}^{n-1} K\left(\frac{X_{j\Delta}-x}{h}\right) B(X_{j\Delta}) \int_{j\Delta}^{(j+1)\Delta} [X_s - X_{j\Delta}]\,\sigma(X_s)\,dW_s\right|;\\
V_3 &:= \sup_{x\in\mathbb{R}} \Big\{(1/Th)\textstyle\sum_{j=1}^{n-1} K\left(\frac{X_{j\Delta}-x}{h}\right) B(X_{j\Delta}) \int_{j\Delta}^{(j+1)\Delta} \sigma^2(X_s)\,ds\\
&\quad - (1/h)\,E[K\left(\frac{X_{j\Delta}-x}{h}\right) B(X_{j\Delta})\sigma^2(X_{j\Delta})]\Big\};\\
V_4 &:= \sup_{x\in\mathbb{R}} \left|(1/h)\textstyle\int K\left(\frac{p-x}{h}\right) B(p)\sigma^2(p)\pi(p)\,dp - B(x)\sigma^2(x)\pi(x)\right|.
\end{aligned}$$

We below investigate these four terms. First, by Theorem 1 and (20), we have

$$\begin{aligned}
B(X_{j\Delta})[X_s - X_{j\Delta}]\mu(X_s) &= B^{1/2}(X_{j\Delta})[X_s - X_{j\Delta}] \times B^{1/2}(X_{j\Delta})\mu(X_s)\\
&= O_{a.s.}(\sqrt{\Delta\log(1/\Delta)}),
\end{aligned}$$

uniformly. This implies that $V_1 = O_p(\sqrt{\Delta\log(1/\Delta)})$, since $\sup_{x\in\mathbb{R}}(1/nh)\sum_{j=1}^{n-1}\left|K(\frac{X_{j\Delta}-x}{h})\right| = O_P(1)$, which is derived in the proof of Theorem 2. Next, applying Theorem 7 (Theorem 2) to V_3 with $\psi(\cdot) = \sigma^2(\cdot)$, we can immediately obtain $V_3 = O_p(a^*_T)$ (resp. $O_p(a^\star_T)$) under the condition on β in part (i) (resp. part (ii)). We can also show that $V_4 = O(h^2)$ in the same arguments as those for the term U_2 in the proof of Theorem 4. We subsequently show that

$$V_2 = O_p(\sqrt{(\log T)/T^\theta h}) \text{ for part (i); and } O_p(\sqrt{(\log T)/Th}) \text{ for part (ii)}, \tag{S.29}$$

under the stated conditions. Given these, we have obtained the desired convergence results for $\sup_{x\in\mathbb{R}}|\hat{\Psi}_{\sigma^2}(x) - B(x)\sigma^2(x)\pi(x)|$, which in turn allow us to drive the desired convergence results for $|\hat{\sigma}^2(x) - \sigma^2(x)|$ in the same way as in the proof of Theorem 8 in Hansen (2008) (we omit the details for brevity).

Proof of (S.29). We derive the convergence rates of V_2 by using arguments analogous to those for U_2 in Theorem 4. Thus, we only outline main points. Look at

$$\begin{aligned}
&B(X_{j\Delta})[X_s - X_{j\Delta}]\sigma(X_s)\\
=\ &B^{1/2}(X_{j\Delta})\left[\textstyle\int_{j\Delta}^s \mu(X_u)\,du + \int_{j\Delta}^s \sigma(X_u)\,dW_u\right] \times B^{1/2}(X_{j\Delta})\sigma(X_{j\Delta})\\
&+B^{1/2}(X_{j\Delta})\left[\textstyle\int_{j\Delta}^s \mu(X_u)\,du + \int_{j\Delta}^s \sigma(X_u)\,dW_u\right] \times B^{1/2}(X_{j\Delta})[\sigma(X_s) - \sigma(X_{j\Delta})],
\end{aligned}$$

and define

$$\mathbf{f}_{\Delta}(s,j) := \mathbf{1}_{\left\{B^{1/2}(X_{j\Delta})[|\int_{j\Delta}^{s}\mu(X_u)du|+|\int_{j\Delta}^{s}\sigma(X_u)dW_u|+\left|\sigma(X_s)-\sigma(X_{j\Delta})\right|]\leq\Delta^{1/2}\log(1/\Delta)\right\}}.$$

Then,

$$\begin{aligned} V_2 &= (2/Th)\sup_{x\in\mathbb{R}}|\textstyle\sum_{j=1}^{n-1}K\left(\frac{X_{j\Delta}-x}{h}\right)\int_{j\Delta}^{(j+1)\Delta}B(X_{j\Delta})[X_s-X_{j\Delta}]\sigma(X_s)\mathbf{f}_{\Delta}(s,j)\,dW_s| \\ &\quad + (2/Th)\sup_{x\in\mathbb{R}}|\textstyle\sum_{j=1}^{n-1}K\left(\frac{X_{j\Delta}-x}{h}\right)\int_{j\Delta}^{(j+1)\Delta}B(X_{j\Delta})[X_s-X_{j\Delta}]\sigma(X_s)[1-\mathbf{f}_{\Delta}(s,j)]dW_s| \\ &= : \bar{V}_2+\tilde{V}_2. \end{aligned}$$

By the same arguments as in deriving the result (S.8) in the proof of Theorem 5, we have $\tilde{V}_2=0$ almost surely for sufficiently small Δ. Therefore, the convergence rate of V_2 is determined by that of $\bar{V}_2$. Letting

$$q(s,j\Delta) := B(X_{j\Delta})[X_s-X_{j\Delta}]\sigma(X_s)\mathbf{f}_{\Delta}(s,j),$$

we have

$$\begin{aligned} \bar{V}_2 &\leq (1/Th)\max_{k\in\{1,...,\nu(h)\}}\sup_{g\in\mathcal{K}_k(h)}\textstyle\sum_{j=1}^{n-1}|g_k(X_{j\Delta})-g(X_{j\Delta})|\left|\int_{j\Delta}^{(j+1)\Delta}q(s,j\Delta)\,dW_s\right| \\ &\quad + \max_{k\in\{1,...,\nu(h)\}}\left|(1/Th)\textstyle\sum_{j=1}^{n-1}K\left(\frac{X_{j\Delta}-x}{h}\right)\int_{j\Delta}^{(j+1)\Delta}q(s,j\Delta)\,dW_s\right| \\ &= : \bar{V}_{21}+\bar{V}_{22}, \end{aligned}$$

where $\{\mathcal{K}_k(h)\}_{k=1}^{\nu(h)}$ is the finite covering of $\mathcal{K}(h)$ as defined in the proof of Theorem 5 satisfying (S.14) with $\nu(h)\leq\Lambda\varepsilon^{-4\bar{r}}$ (for some constant $\Lambda>0$ and any $\bar{r}>1$). In the same way as in (S.15), we can show that

$$\bar{V}_{21} = O_p(h^{-1}\varepsilon\log\Delta^{-1}). \tag{S.30}$$

By Theorems 8 and 3, for any $a>0$,

$$\Pr(\bar{V}_{22}\geq a\sqrt{(\log T)/T^{\theta}h})\leq\Lambda\varepsilon^{-4\bar{r}}[2T^{-a^2/2}+4\exp\{-aC_M(Th)^{1-\theta}\}+4AT^{-\beta}h^{-\theta(\beta+1)}]; \tag{S.31}$$

$$\Pr(\bar{V}_{22}\geq a\sqrt{(\log T)/Th})\leq\Lambda\varepsilon^{-4\bar{r}}[2T^{-a^2/2}+T^{-aC_M}+4AT^{-\beta}h^{-(\beta+1)}(\log T)^{1-\beta}]. \tag{S.32}$$

Now, we can derive the convergence result of $\bar{V}_2$ under the condition on β of part (i). We let $\varepsilon=\sqrt{h/T^{\theta}(\log T)}$ and obtain $\bar{V}_{21}=O_p(\sqrt{(\log T)/T^{\theta}h})$ since $\Delta^{-1}=O(T^{\kappa})$ and $\log\Delta^{-1}=O(\log T)$. Then, by using (S.31) and the condition that $h^{-1}=O(T^{\theta}/(\log T))$, we can show that as $T\to\infty$,

$$\Pr(\bar{V}_{22}\geq a\sqrt{(\log T)/T^{\theta}h})\to 0,$$

for any a large enough if

$$\varepsilon^{-4\bar{r}}\times T^{-\beta}h^{-\theta(\beta+1)} = O((\log T)^{2\bar{r}-\theta(\beta+1)}\times T^{-\beta+4\bar{r}\theta+\theta^2(\beta+1)})$$

tends to zero, which occurs as long as

$$-\beta+4\bar{r}\theta+\theta^2(\beta+1)<0 \Longleftrightarrow 4\bar{r}\theta+\theta^2<\beta\left(1-\theta^2\right).$$

Recalling that any $\bar{r} > 1$ can be selected, we can see that the last inequality is satisfied if $4\theta + \theta^2 < \beta\left(1-\theta^2\right)$. We now have $\bar{V}_2 = O_p(\sqrt{(\log T)/T^{\theta}h})$ as desired for the part (i) case.

Finally, suppose that the condition on β of part (ii) holds. In this case, plugging $\varepsilon = \sqrt{h/T(\log T)}$ into (S.30) and (S.32), we have $\bar{V}_{21} = O_p(\sqrt{(\log T)/Th})$, and $\Pr(\bar{V}_{22} \geq a\sqrt{(\log T)/Th}) \to 0$ as $T \to \infty$ for any a large enough if

$$\varepsilon^{-4\bar{r}} \times T^{-\beta} h^{-(\beta+1)} (\log T)^{1-\beta} = O((\log T)^{2\bar{r}-\theta(\beta+1)} \times T^{-\beta(1-\theta)+2\bar{r}+(2\bar{r}+1)\theta})$$

tends to zero, which occurs as long as

$$2\bar{r} + (2\bar{r}+1)\theta < \beta(1-\theta).$$

We can obtain this inequality if $2 + 3\theta < \beta(1-\theta)$ since any $\bar{r} > 1$ can be picked. The proof is now completed. □

11 Effects of the Damping Function

In this section, we briefly investigate effects of the damping function by presenting bias, variance and mean-squared-error (MSE) expressions of the estimators (4) and (5), as well as by providing some results in finite samples.

Bias and Variance Expressions: The exact expressions are hard to analyze, and we derive their approximations: for each $x \in (l, r)$,

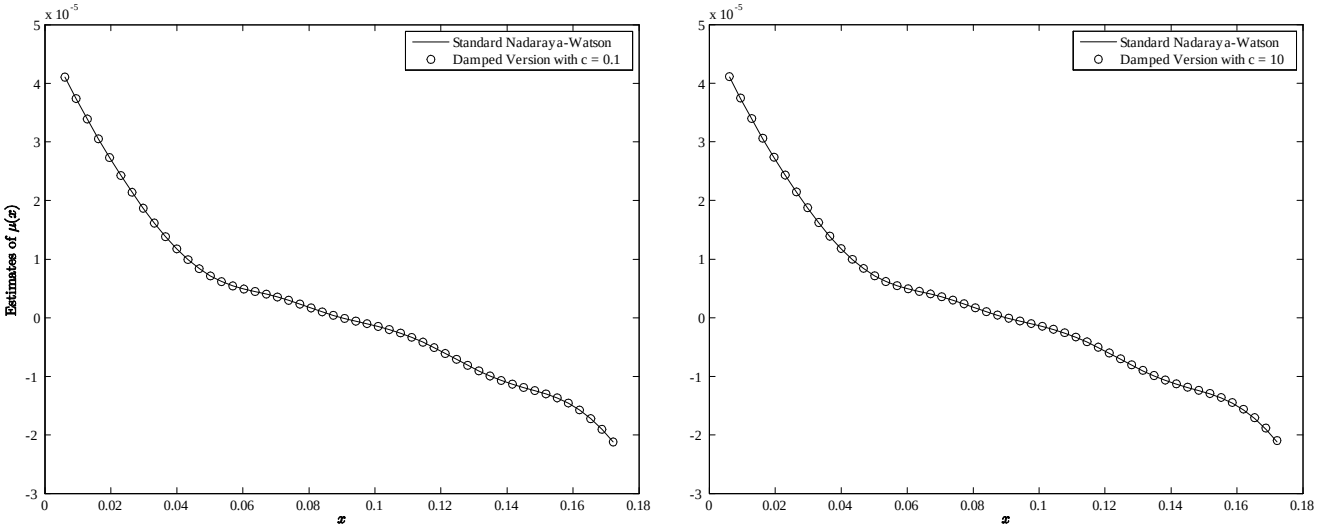
$$E\left[\left(\hat{\mu}(x) - \mu(x)\right)^2\right] \simeq \mathcal{B}_{\mu}^2(x) + \mathcal{V}_{\mu}(x); \; E\left[\left(\hat{\sigma}^2(x) - \sigma^2(x)\right)^2\right] \simeq \mathcal{B}_{\sigma^2}^2(x) + \mathcal{V}_{\sigma^2}(x);$$

where

$$\begin{aligned}
\mathcal{B}_{\mu}(x) &:= h^2 \left\{ (d/dx)\left[B(x)\mu(x)\right] \times \frac{\pi'(x)}{B(x)\pi(x)} + \frac{(d^2/dx^2)\left[B(x)\mu(x)\right]}{2B(x)} \right\} \int z^2 K(z)\, dz; \\
\mathcal{V}_{\mu}(x) &:= (1/Th)\left[\sigma^2(x)/\pi(x)\right] \int K^2(z)\, dz; \\
\mathcal{B}_{\sigma^2}(x) &:= h^2 \left\{ (d/dx)\left[B(x)\sigma^2(x)\right] \times \frac{\pi'(x)}{B(x)\pi(x)} + \frac{(d^2/dx^2)\left[B(x)\sigma^2(x)\right]}{2B(x)} \right\} \int z^2 K(z)\, dz; \\
\mathcal{V}_{\sigma^2}(x) &:= (1/nh)\, 2\left[\sigma^4(x)/\pi(x)\right] \int K^2(z)\, dz.
\end{aligned}$$

We can derive these approximations from using the standard method as in Pagan and Ullah (1999). For their validation, we require some conditions on the existence of moments, the decay rate of the mixing coefficients and the shrinking rate of Δ. For brevity, we omit the detailed conditions and derivations for the approximations, which are obtained analogously to ones provided in Kanaya and Kristensen (2014), who also provide the precise meaning of "$\simeq$." $\mathcal{B}_{\mu}(x)$ and $\mathcal{B}_{\sigma^2}(x)$ correspond to the biases of the estimators. $\mathcal{V}_{\mu}(x)$ and $\mathcal{V}_{\sigma^2}(x)$ correspond to their variances, which are the same as the variances of the asymptotic normal distributions. Obviously, the damping function affects only the bias properties, and the variance components are of the same form as those of the standard NW estimators.

Graphical Illustration of Effects of the Damping Function: To see the effects of $B(\cdot)$ in finite samples, we compare the standard NW estimator $\tilde{\mu}(x)$ and its damped version $\hat{\mu}(x)$ with $B(x) = \exp\{-cx^2\}$ with $c = 0.1$ and 10. The following figures are based on the same simulated path of the Ornstein-Uhlenbeck process $dX_s = \lambda(m - X_s)\,dt + \sigma dW_s$, where $(\lambda, m, \sigma^2) = (0.85837, 0.089102, 0.0021854)$, which is Aït-Sahalia's, (1996a) estimate for short-term interest rates; $(T, \Delta, n) = (25, 1/52, 1300)$; $h = 4\hat{s}n^{-1/5}$ (this bandwidth has been used in Stanton, 1997; see p. 360 of Chapman and Pearson, 2000); $\tilde{\mu}(x)$ and $\hat{\mu}(x)$ are evaluated over equally-spaced 50 grid points between 1 and 99 percentiles of the invariant distribution of the process (0.0061 and 0.1721, respectively). As we can see in the two figures, the NW estimates and its damped version perfectly coincide both for $c = 0.1$ and 10, even with the one-hundred time difference in the scale parameter c. While these are only based on one sample, it has been quite difficult to obtain some other samples/examples in which $\tilde{\mu}(x)$ and $\hat{\mu}(x)$ look significantly different for some other choices of date-generating-process, sample-size and bandwidth settings. We have found a similar result for the diffusion function estimation. From these, we conclude that the effects of the damping function B are not significant, where we again note that its effects are cancelled out between the numerator and denominator parts.



12 Estimation of Non-Negative Valued Processes

In this section, we provide some discussions and results for estimating processes with $I = (0, \infty)$ or $[0, \infty)$, where we note that many parametric models for short-term interest rates have such state space I. While we here focus on the case where the left-end point of I is 0, we can also think of some other cases (e.g., the left-end point is finite and non-zero and/or the right one is also bounded), to which results for $I = (0, \infty)$ or $[0, \infty)$ carry over with suitable modifications.

When I has a finite end point, the choice of $B(\cdot) = \exp\{-cx^2\}$ as considered in Sections 3 and (11) may not be sufficient. For example, Aït-Sahalia (1996b, 1999) considers a parametric diffusion model with $I = (0, \infty)$ and the drift function $\bar{\mu}(x) := \alpha_0 + \alpha_1 x + \alpha_2^2 + \alpha_3/x$, which diverges as $x \to 0$. To accommodate this kind of models, we can think of a damping function such as

$$B(x) = x^b \exp\left\{-x^2\right\}, \tag{S.33}$$

with some $b > 2$. This choice of B allow us to establish Theorem 1, even with a drift function such as $\bar{\mu}(x)$.

When the left end point of I is 0, we can also think of a case where the invariant density π may not be bounded around $x = 0$. Processes with this feature can be easily found. Among others, we can think of the CIR process:

$$dX_s = \kappa(\theta - X_s)ds + \sigma\sqrt{X_s}dW_s, \tag{S.34}$$

with $\kappa, \theta, \sigma > 0$. If $2\kappa\theta/\sigma^2 > 0$, the stationary solution to (S.34) can exist whose invariant density $\pi(\cdot)$ is given by the Gamma distribution with $2\kappa\theta/\sigma^2$ and $\sigma^2/2\kappa$ representing the shape and scale parameters, respectively. Given that $2\kappa\theta/\sigma^2 \in (0,1)$, the left boundary $l = 0$ is attainable. In this case, the process can have the Gamma distribution as its invariant distribution by making l instantaneously reflecting (for construction of this kind of process, see discussions in Section 2 on the behavior of the process running over infinite time horizon, after the hit on $l = 0$; see also discussions in p. 441 of Forman and Sørensen, 2008). In this case, $\pi(x) \to \infty$ as $x \to 0$. This process with $2\kappa\theta/\sigma^2 \in (0,1)$ satisfies the mixing condition in (A2.ii) with a geometric decaying rate (i.e. $\beta = \infty$).[10] This sort of process with unbounded $\pi(x)$ at the end point can be also handled through the choice of B as in (S.33), by which we can ensure the uniform boundedness of $B(x)\pi(x)$, where we note that the integrability of the density implies that $\pi(x) \sim x^{-q}$ (with some $q \in (0,1)$) in the neighborhood of 0 and thus the divergence rates of $\pi'(x)$ and $\pi''(x)$ around zero are also at most of the polynomial order.

For the case $\pi(x) \to \infty$ as $x \to 0$, we can still verify the uniform convergence of $\hat{\mu}(x)$ and $\hat{\sigma}^2(x)$ over $I = (0,\infty)$ (or $[0,\infty)$), given the damping function as in (S.33). In Theorem 2, we have supposed the uniform boundedness of π, but this condition can be removed if we make slight changes of relevant conditions such as replacing "$\psi(\cdot) \in \mathcal{D}(B,\pi)$" with

$$\psi(\cdot) \in \mathcal{D}\left(B^{1/2},\pi\right) \quad \text{and} \quad \sup_{x\in(0,\infty)} B^{1/2}(x)\pi(x) < \infty, \tag{S.35}$$

for example. As seen in the proof of Theorem 2, we need to show that the following object:

$$\sup_{x\in(0,\infty)} (\Delta/Th)\sum\nolimits_{j=1}^{n-1} K\left(\tfrac{X_{j\Delta}-x}{h}\right), \tag{S.36}$$

is bounded (in probability). However, it may not be so if $\pi(x)$ is unbounded. In this case, instead of (S.36), we consider

$$\sup_{x\in(0,\infty)} (\Delta/Th)\sum\nolimits_{j=1}^{n-1} K\left(\tfrac{X_{j\Delta}-x}{h}\right) B^{1/2}(X_{j\Delta}), \tag{S.37}$$

whose O_p-boundedness is guaranteed under the condition that $\sup_{x\in(0,\infty)} B^{1/2}(x)\pi(x) < \infty$, where we note that even for the case with $I = [0,\infty)$ (i.e., the process may attain the point 0), the supremum of x needs to be taken over $(0,\infty)$, instead of $[0,\infty)$, if $\pi(0)$ is unbounded. The same argument applies to Theorem 3, for which the conditions in (S.35) can be used to relax the uniform boundedness of $\pi(x)$.

Under the condition as in (S.35), we can still verify the same convergence rates of the variance effect terms as in Theorems 2-3 even with an unbounded $\pi(x)$ around $x = 0$. However, the boundedness of the end point in general makes the convergence rate of the smoothing bias slower. This observation is summarized as the following remark:

[10] This can be checked by noting the following facts: i) the process is conservative and reversible (see Sections 8-9 in Kent, 1978); ii) the spectrum of its (infinitesimal) generator is discrete and has a gap left to zero, which is given by $\{\lambda_j\}$ with $\lambda_j = -\kappa j$ (see, e.g., p. 334 of KT81); and iii) i) and ii) imply $\{X_s\}$ is geometrically ρ-mixing (see discussions in p. 799 of Hansen and Scheinkman, 1995, as well as those in Hansen *et al.*, 1998). Note that in the case $2\kappa\theta/\sigma^2 > 1$, neither of the boundaries is attracting, and we can also check the geometric mixing property of the process by the same argument or by using, for example, Corollary 5.5 of Chen *et al.* (2010).

Remark 2. **(i)** Let $I = (0, \infty)$ or $[0, \infty)$. Then, given the conditions in (S.35) and the kernel function K satisfying (B2), it holds that as $h \to 0$,

$$\sup_{x \in (0,\infty)} \left| \bar{G}_{n,T}(x) - H(x) \right| = O(h), \tag{S.38}$$

where $H(x) = B(x) \psi(x) \pi(x)$. This slower convergence occurs because we cannot use the symmetricity property of the kernel in the neighborhood of zero (i.e., $\int xK(x)\,dx = 0$) to kill the first-order term of the smoothing bias (therefore, if $I \neq \mathbb{R}$, the use of higher-order kernels does not improve the uniform convergence rate over I). This kind of phenomenon, the so-called boundary bias, is observed if the endpoint of the support is bounded and the symmetric kernel is used (see the arguments in Bouezmarni and Scaillet, 2005), while the boundary bias may be avoided by using asymmetric kernels as in Bouezmarni *et al.* (2005) and Gospodinov and Hirukawa (2012). We note that the supremum is taken over the open set $(0, \infty)$ in (S.38), which is for avoiding the indefiniteness at $x = 0$ when $I = [0, \infty)$ and $\pi(0)$ is unbounded (we may have $[0, \infty)$ in (S.38) if 0 is a point attainable by the process and $\pi(0) < \infty$).
(ii) If we use some special kernel and restrict the domain of x, we can recover the smoothing-bias rate of h^2. That is, we suppose that $K(\cdot)$ is a non-negative valued kernel with bounded support (resp. the normal kernel), then under the conditions in (S.35),

$$\sup_{x \in [r(h),\infty)} \left| \bar{G}_{n,T}(x) - H(x) \right| = O\left(h^2\right), \tag{S.39}$$

as $h \to 0$, where $r(h)\,(\to 0)$ is a trimming sequence with $r(h) = -c_K h$ and $c_K := \inf\{c < 0 : K(c) > 0\}$ (resp. $r(h) = 2h\sqrt{\log(1/h)}$).

We can apply the results (S.38)-(S.39) to obtain the uniform rates of $\hat{\mu}(x)$ and $\hat{\sigma}^2(x)$ when $I = (0, \infty)$ or $[0, \infty)$.

Proof of the Statements in Remark 2. Let $I = (0, \infty)$ or $I = [0, \infty)$. (i) We prove (S.38) here:

$$\begin{aligned}
\left| \bar{G}_{n,T}(x) - H(x) \right| &= \sup_{x \in (0,\infty)} \left| \int_{-x/h}^{\infty} K(q) \left[H(qh + x) - H(x) \right] dq \right| \\
&= \sup_{x \in (0,\infty)} \left| \int_{-x/h}^{\infty} K(q) H'(\tilde{x})\, qh\, dq \right| \\
&\leq \sup_{x \in (0,\infty)} |H'(x)|\, h \int_{-\infty}^{\infty} |qK(q)|\, dq = O(h),
\end{aligned}$$

where we note that $qh + x \in I$ (if $x \in I$ and $q \in (-x/h, \infty)$) for the second inequality, and the third equality holds by the Taylor expansion ($\tilde{x}$ is on the line segment connecting x to $qh + x$ and the expansion is valid for any $x \in (0, \infty)$).
(ii) Suppose that $K(\cdot)$ is a non-negative valued kernel with a bounded support or it is the normal kernel. We prove (S.39) here:

$$\begin{aligned}
&\left| \bar{G}_{n,T}(x) - H(x) \right| \\
&= \sup_{x \in [r(h),\infty)} \left| \int_{-x/h}^{\infty} K(q) \left[H'(x)\, qh + (1/2) H''(\tilde{x}) (qh)^2 \right] dq \right| \\
&\leq h \sup_{x \in (0,\infty)} |H'(x)| \times \sup_{x \in [r(h),\infty)} \left| \int_{-x/h}^{\infty} qK(q)\, dq \right| + \left(h^2/2\right) \int_{-x/h}^{\infty} q^2 |K(q)|\, dq \sup_{x \in (0,\infty)} |H''(x)|,
\end{aligned} \tag{S.40}$$

where we can easily check that the second term on the RHS is $O(h^2)$. If the support of $K(\cdot)$ is bounded and $r(h) = -c_K h (> 0)$ with $c_K = \inf\{c < 0 : K(c) > 0\}$, then the first-order term on the RHS is zero since $\int_{-x/h}^{\infty} qK(q)\,dq = \int_{c_K}^{\infty} qK(q)\,dq = 0$ for any $x \in [-c_K h, \infty)$. If $K(\cdot)$ is the normal kernel and $r(h) = 2h\sqrt{\log(1/h)}$, then

$$\begin{aligned}
\sup_{x\in[r(h),\infty)} \left|\int_{-x/h}^{\infty} qK(q)\,dq\right| &= \sup_{x\in[r(h),\infty)} \left|\int_{x/h}^{\infty} qK(q)\,dq\right| \\
&\leq (2\pi)^{-1/2} \int_{r(h)/h}^{\infty} [q \exp\left\{-q^2/4\right\}] \exp\left\{-q^2/4\right\} dq \\
&\leq (2\pi)^{-1/2} \int_{0}^{\infty} q \exp\left\{-q^2/4\right\} dq \times \exp\{-(r(h)/h)^2/4\} \\
&= O(h),
\end{aligned}$$

where the first equality holds since $\int_{-\infty}^{\infty} qK(q)\,dq = 0$. From these arguments, we can show that the RHS of (S.40) is $O(h^2)$, completing the proof. □

Research Papers 2013

2015-32: Jean-Guy Simonato and Lars Stentoft: Which pricing approach for options under GARCH with non-normal innovations?

2015-33: Nina Munkholt Jakobsen and Michael Sørensen: Efficient Estimation for Diffusions Sampled at High Frequency Over a Fixed Time Interval

2015-34: Wei Wei and Denis Pelletier: A Jump-Diffusion Model with Stochastic Volatility and Durations

2015-35: Yunus Emre Ergemen and Carlos Velasco: Estimation of Fractionally Integrated Panels with Fixed Effects and Cross-Section Dependence

2015-36: Markku Lanne and Henri Nyberg: Nonlinear dynamic interrelationships between real activity and stock returns

2015-37: Markku Lanne and Jani Luoto: Estimation of DSGE Models under Diffuse Priors and Data-Driven Identification Constraints

2015-38: Lorenzo Boldrini and Eric Hillebrand: Supervision in Factor Models Using a Large Number of Predictors

2015-39: Lorenzo Boldrini and Eric Hillebrand: The Forecasting Power of the Yield Curve, a Supervised Factor Model Approach

2015-40: Lorenzo Boldrini: Forecasting the Global Mean Sea Level, a Continuous-Time State-Space Approach

2015:41: Yunus Emre Ergemen and Abderrahim Taamouti: Parametric Portfolio Policies with Common Volatility Dynamics

2015-42: Mikkel Bennedsen: Rough electricity: a new fractal multi-factor model of electricity spot prices

2015-43: Mikkel Bennedsen, Asger Lunde and Mikko S. Pakkanen: Hybrid scheme for Brownian semistationary processes

2015-44: Jonas Nygaard Eriksen: Expected Business Conditions and Bond Risk Premia

2015-45: Kim Christensen, Mark Podolskij, Nopporn Thamrongrat and Bezirgen Veliyev: Inference from high-frequency data: A subsampling approach

2015-46: Asger Lunde, Anne Floor Brix and Wei Wei: A Generalized Schwartz Model for Energy Spot Prices - Estimation using a Particle MCMC Method

2015-47: Annastiina Silvennoinen and Timo Teräsvirta: Testing constancy of unconditional variance in volatility models by misspecification and specification tests

2015-48: Harri Pönkä: The Role of Credit in Predicting US Recessions

2015-49: Palle Sørensen: Credit policies before and during the financial crisis

2015-50: Shin Kanaya: Uniform Convergence Rates of Kernel-Based Nonparametric Estimators for Continuous Time Diffusion Processes: A Damping Function Approach