

Article

Inverse-Probability-Weighted Wavelet Estimation of Regression Derivatives Under Missing-at-Random Responses for Stationary Ergodic Processes

Salim Bouzebda ^{1,†}  and Sultana Didi ^{2,*,†} 

¹ LMAC (Laboratory of Applied Mathematics of Compiègne), Université de Technologie de Compiègne, 60200 Compiègne, France; salim.bouzebda@utc.fr

² Department of Statistics and Operations Research, College of Sciences, Qassim University, P.O. Box 6688, Buraydah 51452, Saudi Arabia

* Correspondence: s.biha@qu.edu.sa

† These authors contributed equally to this work.

Abstract

We consider the estimation of partial derivatives of multivariate regression-type functionals from incomplete observations generated by a discrete-time strictly stationary ergodic process. The response variable is subject to a missing-at-random (MAR) mechanism, whereas the covariates are fully observed. Building upon the complete-data wavelet methodology developed in Didi and Bouzebda (2025), we construct inverse-probability-weighted empirical wavelet estimators that compensate for the selection bias induced by missing responses. When the propensity score is unknown, a feasible estimator is obtained by replacing the oracle weights with a nonparametric Nadaraya–Watson estimator. The analysis is carried out under stationary ergodicity without imposing mixing assumptions. The estimation error is decomposed into three analytically distinct components: the deterministic multiresolution approximation error, the stochastic fluctuation of the oracle inverse-probability-weighted estimator, and the additional error arising from propensity score estimation. This decomposition makes it possible to isolate the respective effects of approximation, dependence, and missingness within a unified asymptotic framework. Under explicit assumptions on the multiresolution approximation, missingness mechanism, conditional density stabilization, moment conditions, and accuracy of the propensity estimator, we establish non-asymptotic integrated mean squared error bounds together with their asymptotic rates. We further prove almost-sure uniform consistency over compact subsets of the interior of the support and derive a pointwise central limit theorem for both the oracle and feasible estimators. The limiting variance explicitly reflects the information loss induced by inverse probability weighting, and for general orthogonal projection kernels is formulated under the corresponding dyadic-phase condition. The general methodology is specialized to the estimation of first- and second-order derivatives of ordinary regression functions. A finite-sample simulation study investigates the empirical behavior of the proposed estimators under stationary ergodic dependence and MAR missingness, examines the influence of both the wavelet resolution level and the propensity-score bandwidth, evaluates the finite-sample performance of the asymptotic confidence intervals, and compares the proposed procedure with oracle, complete-case, and competing nonparametric estimators. The numerical results are consistent with the theoretical analysis and illustrate the respective contributions of wavelet approximation, inverse probability weighting, and propensity score estimation to the overall estimation error. When the propensity score is identically equal to one, the proposed methodology reduces to the corresponding complete-data wavelet estimator.



Academic Editor: Nikolai Leonenko

Received: 15 July 2026

Revised: 26 July 2026

Accepted: 28 July 2026

Published: 5 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: regression estimation; stationarity; ergodicity; rates of strong convergence; wavelet-based estimators; martingale differences; discrete-time; stochastic processes; time series; missing-at-random responses

MSC: 62G07; 62G08; 62G05; 62G20; 62H05; 60G42; 60G46

1. Introduction

Nonparametric estimation of functional characteristics occupies a central position in modern mathematical statistics. In many applications, the object of inferential interest is not exhausted by the value of an unknown density or regression function itself; rather, its derivatives encode local geometry, rates of change, curvature, modal structure, and higher-order shape information, which arise naturally in problems of scientific interpretation as well as in the construction of statistical procedures. Kernel methods constitute one of the classical approaches to such questions because of their local character, analytical tractability, and broad applicability. Comprehensive treatments of kernel smoothing and related nonparametric procedures may be found in [1,2].

Density-derivative estimation provides a representative illustration. The gradient of a density identifies directions of local increase, while its Hessian and higher derivatives enter the analysis of curvature, modes, ridges, and other geometric features. The second derivative plays an important role in modality and topological inference [3], and density derivatives also appear in data-driven bandwidth selection [4]. In signal processing problems, the logarithmic derivative of a density, that is, the ratio between its derivative and the density itself, is central to filtering and interpolation procedures [5]. Gradient-based constructions are equally important in filament estimation for point-cloud data, with applications extending from medical imaging and remote sensing to seismology and cosmology [6]. Further motivations arise in Fisher information estimation, parameter estimation, regression analysis, and hypothesis testing [7]. Classical contributions to density-derivative estimation include [8–12].

Regression derivatives are no less fundamental. In a regression model, first-order derivatives quantify local marginal effects and monotonicity, while second- and higher-order derivatives describe curvature, acceleration, and interaction structure. Such quantities are often more informative than the regression level itself when the scientific question concerns change rather than magnitude. Applications include the analysis of human growth curves [13], the assessment of kidney function in lupus nephritis [14], and the processing of Raman spectra [15]. Derivative estimators also enter confidence interval construction [16], bandwidth selection [17], and the comparison of regression curves [18]. Kernel M -estimators for the first derivative of a regression function, together with extensions to higher orders, were investigated in [19]. Derivative information is also central to modal regression, which describes the conditional relationship between a response vector $\mathbf{Y}$ and a covariate vector $\mathbf{X}$ through the modes of the conditional distribution; see [20]. Foundational contributions to nonparametric regression estimation include [21–24].

Wavelet procedures provide a multiscale alternative to purely local kernel smoothing. Their appeal lies in the decomposition of a functional object into resolution-dependent components as well as in their simultaneous localization in space and scale. This structure separates approximation error from stochastic fluctuation, and permits regularity to be expressed through the decay of multiresolution coefficients, making wavelet methods naturally adapted to functions with spatially inhomogeneous smoothness, localized irregularities, or membership in Besov classes. A systematic account of wavelet methods in nonparametric function estimation is given in [25]. Related contributions include estima-

tion of integrated squared density derivatives in the independent setting [26], extensions to negatively or positively associated sequences [27], and estimation of partial derivatives of multivariate densities under independence [28]. Wavelet derivative estimation under mixing-type dependence was studied in [29,30], while [31] considered partial-derivative estimation in the presence of additive noise.

The treatment of dependent observations introduces a further layer of difficulty. Independence is rarely tenable in applications involving time series, longitudinal measurements, repeated observations, dynamical systems, or stochastic processes. Much of the nonparametric literature replaces independence with mixing conditions, which are powerful but may be difficult to verify and can exclude stationary processes with weaker forms of dependence. The complete-data work in [32] developed a wavelet framework for estimating partial derivatives of multivariate regression-type functionals under discrete-time strictly stationary ergodic observations. Its analysis used conditional density stabilization and martingale approximation rather than strong mixing, and established integrated risk bounds, uniform convergence, and pointwise asymptotic normality.

The present paper takes [32] as its complete-data benchmark. Its contribution is not a new wavelet projection principle but an extension of that construction to the missing-response context. This includes the genuinely new analytical tasks of controlling the inverse-probability-weighted stochastic component and quantifying the additional substitution error produced when the propensity score is estimated nonparametrically. This distinction is important for accurately assessing the scope and novelty of the paper.

Missing responses are common in longitudinal surveys, clinical studies, sensor networks, administrative databases, and other observational systems. When the observation probability depends on recorded covariates, a complete-case analysis generally targets a selected subpopulation rather than the population of interest. Therefore, a principled procedure must account for the observation mechanism. The modern taxonomy of missingness mechanisms originates in [33], which distinguished missing completely at random, missing at random, and not missing at random according to the conditional relation between the missingness indicator and the full data.

The missing-at-random assumption allows for the probability of observing the response to depend on the covariates while requiring that, conditionally on those covariates, the missingness indicator contain no further information about the unobserved response. Its formal interpretation and inferential consequences have been analyzed in [34–40]. As discussed in [41], MAR-based procedures may offer substantial protection against the bias of naive complete-case methods or methods that implicitly assume missing completely at random.

A standard corrective device under MAR is inverse probability weighting, the foundations of which go back to [42]. The observed contribution is multiplied by the reciprocal of its conditional observation probability, commonly called the propensity score. Under MAR and positivity, the resulting compensation identity restores the complete-data target in expectation. Related developments concerning robustness, semiparametric efficiency, and doubly robust procedures may be found in [33,43–45].

We observe a strictly stationary ergodic sequence $(\mathbf{X}_i, \delta_i \mathbf{Y}_i, \delta_i)$, $i = 1, \dots, n$, where $\mathbf{X}_i \in \mathbb{R}^d$ is fully observed, $\mathbf{Y}_i \in \mathbb{R}^{d'}$ may be missing, and δ_i indicates whether the response is observed. For a measurable transformation $\rho : \mathbb{R}^{d'} \rightarrow \mathbb{R}$, define

$$r(\rho; \mathbf{x}) = \mathbb{E}[\rho(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}] f_{\mathbf{X}}(\mathbf{x}).$$

The estimand is the partial derivative $(\partial^{\beta} r)(\rho; \mathbf{x})$. This formulation encompasses ordinary regression-type functionals, distributional regression functionals, and derivative quantities relevant to modal and shape analysis.

The oracle construction inserts the inverse propensity weight into the complete-data empirical wavelet coefficients. Under MAR, positivity, and integrability, the weighted coefficient has the same expectation as its complete-data counterpart. The feasible estimator replaces the unknown propensity score with a Nadaraya–Watson estimator; consequently, the procedure contains two distinct smoothing mechanisms: the resolution level $m(n)$, or equivalently the wavelet bandwidth $h_n = 2^{-m(n)}$, controls multiresolution approximation and stochastic resolution, while the propensity bandwidth λ_n controls the bias and variability of the estimated inverse weights. Their joint behavior cannot in general be reduced to two independent tuning problems.

The theoretical analysis is organized around the exact decomposition

$$\text{total error} = \text{projection bias} + \text{oracle IPW fluctuation} + \text{propensity-substitution error}.$$

The projection term is governed by Besov approximation. The oracle stochastic term is analyzed through martingale differences, conditional density stabilization, Burkholder–Rosenthal inequalities, exponential inequalities for unbounded martingale arrays, and localization properties of derivative projection kernels. The feasible–oracle term requires uniform or mean-square control of the estimated propensity score at the normalization relevant to the theorem under consideration.

Strict stationarity and ergodicity alone do not yield the quantitative rates used below. Accordingly, the integrated-risk theorem assumes a coefficient-level mean-square stabilization condition; the uniform theorem invokes explicit conditional-moment kernel localization, truncation, bandwidth, and propensity rate hypotheses; and the central limit theorem requires local conditional density stabilization, kernel moment bounds, predictable-centering negligibility, projection-bias negligibility, and feasible–oracle equivalence. These conditions are stated theorem by theorem in Section 3 rather than being attributed to ergodicity itself.

The first principal result concerns integrated L_2 -risk. For the oracle IPW estimator, we derive a general bound under a moment exponent $\nu > 2$ and a refined fourth-moment bound under $\nu \geq 4$. The stochastic exponent is obtained from the coefficient-level variance calculation, and is balanced against the squared Besov projection bias. Conditions are then given under which the feasible estimator inherits the corresponding oracle order. The resulting rate is the one obtained from the displayed bias and stochastic bounds; no alternative exponent is asserted without a separate derivation.

The second result concerns uniform convergence on compact subsets of the interior of the support. The bound separately displays the martingale fluctuation, predictable conditional density approximation, and feasible–oracle error. This decomposition makes explicit how wavelet resolution, conditional-law stabilization, and propensity estimation jointly determine the uniform rate.

The third result is a pointwise central limit theorem. The limiting variance contains the inverse-propensity factor, meaning that it quantifies the first-order information loss caused by missing responses. For a general orthogonal projection kernel, the kernel energy term may depend on the dyadic phase $\vartheta_n = \frac{x}{h_n} - \left\lfloor \frac{x}{h_n} \right\rfloor$. The theorem is consequently stated along sequences for which ϑ_n converges, unless the kernel energy is phase invariant. A full-sequence phase-zero variance is not asserted without one of these additional properties.

The general estimator is subsequently specialized to derivatives of an ordinary regression function. In the scalar case, the response-dependent numerator and its derivatives are estimated by inverse probability weighting, whereas the marginal density of the fully observed covariate is estimated from all covariate observations without weighting. First- and second-order regression-derivative estimators are then obtained through exact quotient identities. Their uniform consistency is established on compact subsets of the interior of

the density-positive region, where the estimated denominator is asymptotically bounded away from zero. This application illustrates the reduction of the general construction when $d = d' = 1$; it does not by itself provide a separate empirical verification of all multivariate assertions.

The paper also contains a simulation study. The numerical investigation examines finite-sample bias, integrated error, the effect of missingness, the difference between oracle, feasible, and complete-case procedures, sensitivity to the wavelet resolution and propensity bandwidth, Gaussian approximation, confidence interval coverage, and comparison with a competing derivative estimator. The simulation evidence is necessarily conditional on the specified models, sample sizes, missingness mechanisms, tuning rules, and competitors. It supports the qualitative conclusions within those designs, but does not establish general finite-sample dominance, minimax optimality, or semiparametric efficiency. No real-data analysis is included in the present version, and claims of practical scope are kept commensurate with the evidence supplied.

The contribution may be summarized as follows: the paper provides a rigorous inverse-probability-weighted extension of a complete-data wavelet estimator to stationary ergodic observations with MAR responses; separates oracle and propensity-estimation errors at the levels of integrated risk, uniform convergence, and pointwise asymptotic normality; identifies the inverse propensity and dyadic-phase contributions to the limiting variance; and develops a scalar regression-derivative specialization together with finite-sample numerical evidence. This work should be read as a technically nontrivial extension of established wavelet and IPW principles rather than as the introduction of an entirely new estimation paradigm.

The limitations of the present framework should also be recorded at the outset. Positivity is assumed, and non-ignorable missingness is not covered. The feasible estimator is not doubly robust. The tuning parameters are deterministic in the theoretical results, and no theorem establishes optimality of a fully data-driven joint selector. The pointwise confidence interval is based on first-order asymptotics, and may require undersmoothing or explicit bias correction for accurate finite-sample coverage. The dependence conditions are expressed through conditional density and martingale array hypotheses that must be verified for the process under study. These issues motivate the extensions discussed in Section 6.

The rest of the paper is organized as follows: Section 2 introduces the Besov and multiresolution framework and defines the oracle and feasible MAR-adjusted wavelet estimators; Section 3 provides the structural assumptions, theorem-specific auxiliary conditions, integrated risk results, uniform convergence theorem, phase-correct pointwise central limit theorem, and confidence interval construction; Section 4 specializes the general method to first- and second-order regression derivatives; Section 5 presents the simulation study and the numerical comparison of the estimators; Section 6 summarizes the contribution, limitations, and open directions; Section 7 contains the proofs; and the final section entitled “Besov Spaces” records the intrinsic difference characterization, wavelet characterization, embedding facts, and projection inequalities used throughout the paper.

Notation

Throughout the paper, C denotes a finite positive constant for which the value may change from line to line. The symbol $\mathbb{1}_A$ denotes the indicator of an event or set A . For positive sequences $\{a_n\}_{n \geq 1}$ and $\{b_n\}_{n \geq 1}$, the relation

$$a_n = O(b_n)$$

means that $a_n \leq Cb_n$ for all sufficiently large n , while

$$a_n = o(b_n)$$

means that $a_n/b_n \rightarrow 0$. The symbols $O_{\mathbb{P}}$ and $o_{\mathbb{P}}$ denote the corresponding stochastic orders, and almost-sure orders are stated explicitly.

2. Mathematical Background

2.1. Besov Spaces

We first specify the functional-analytic framework used in the approximation arguments developed below. Besov spaces are particularly well adapted to wavelet procedures because their regularity may be characterized equivalently through the decay of multiresolution coefficients. This equivalence permits the deterministic approximation error of a wavelet projection to be expressed directly in terms of the smoothness and integrability indices of the target function. It is this approximation mechanism, rather than an informal appeal to multiscale localization, that underlies the bias bounds used in the subsequent asymptotic analysis.

Let $s > 0$ and $1 \leq p, q \leq \infty$. We denote by $\mathbf{B}_{s,p,q} = \mathbf{B}_{s,p,q}(\mathbb{R}^d)$ the Besov space with smoothness index s , integrability index p , and fine index q . The parameter s quantifies regularity, p determines the underlying spatial integrability, and q governs the aggregation of contributions across resolution levels. The Besov scale is sufficiently flexible to accommodate spatially inhomogeneous regularity and localized irregularities that need not be described adequately by integer-order differentiability alone.

To prevent notational collision with the response dimension introduced below, the symbol q occurring in $\mathbf{B}_{s,p,q}$ is used exclusively in this subsection as the Besov fine index. Whenever $\mathbf{Y} \in \mathbb{R}^{d'}$ appears subsequently, q denotes the response dimension. Likewise, p in $\mathbf{B}_{s,p,q}$ denotes only the Besov integrability index; moment orders appearing later in probabilistic inequalities will be denoted by distinct symbols.

Let $\{V_j\}_{j \in \mathbb{Z}}$ be a multiresolution analysis of $L_2(\mathbb{R}^d)$ with scaling space V_0 , detail spaces W_j , scaling function ϕ , and associated wavelets ψ_i , $i = 1, \dots, 2^d - 1$. We assume that the multiresolution analysis is r -regular and that $0 < s < r$. For $f \in L_p(\mathbb{R}^d)$, the wavelet characterization of Meyer [46] yields that $f \in \mathbf{B}_{s,p,q}$ if and only if the multiresolution quantity

$$J_{s,p,q}(f) = \|P_{V_0}f\|_{L_p} + \left(\sum_{j>0} \left[2^{js} \|P_{W_j}f\|_{L_p} \right]^q \right)^{1/q}$$

is finite, with the usual replacement of the ℓ_q -sum by a supremum when $q = \infty$. Here, P_{V_0} and P_{W_j} denote the corresponding orthogonal projection operators. Equivalently, if

$$a_{0,\mathbf{k}} = \int_{\mathbb{R}^d} f(\mathbf{u}) \phi_{0,\mathbf{k}}(\mathbf{u}) d\mathbf{u}, \quad b_{i,j,\mathbf{k}} = \int_{\mathbb{R}^d} f(\mathbf{u}) \psi_{i,j,\mathbf{k}}(\mathbf{u}) d\mathbf{u}$$

are the scaling and wavelet coefficients of f , respectively, then $f \in \mathbf{B}_{s,p,q}$ if and only if

$$J'_{s,p,q}(f) = \|a_{0,\cdot}\|_{\ell_p} + \left(\sum_{j>0} \left[2^{j\{s+d(1/2-1/p)\}} \|b_{j,\cdot}\|_{\ell_p} \right]^q \right)^{1/q} < \infty,$$

again with the standard supremum convention when $q = \infty$, where

$$\|b_{j,\cdot}\|_{\ell_p} = \left(\sum_{i=1}^{2^d-1} \sum_{\mathbf{k} \in \mathbb{Z}^d} |b_{i,j,\mathbf{k}}|^p \right)^{1/p}.$$

For $p = \infty$, the corresponding ℓ_∞ -norm is understood in the usual sense.

The preceding coefficient characterization provides the precise approximation device required later. More specifically, the regularity encoded by the Besov norm controls the decay of the detail coefficients as the resolution level increases. Hence, replacing f by its projection onto V_m amounts to discarding the detail components above level m , and the resulting approximation error is bounded through the corresponding Besov coefficient tail. In particular, for $f \in \mathbf{B}_{s,p,q}$, one obtains bounds of the form

$$\|f - P_{V_m} f\| \lesssim 2^{-m\alpha(s,p,d)} \|f\|_{\mathbf{B}_{s,p,q}},$$

where the exponent $\alpha(s,p,d)$ and the admissible range of (s,p,q) depend on the norm in which the approximation error is evaluated. Accordingly, no norm-independent rate is asserted at this stage; the specific embedding and approximation inequalities required for each theorem are stated and invoked under their corresponding hypotheses.

The Besov scale contains several standard regularity classes as particular cases. For example, $H^s(\mathbb{R}^d) = \mathbf{B}_{s,2,2}(\mathbb{R}^d)$, with equivalence of norms, while the Hölder–Lipschitz–Zygmund scale is represented by $\mathbf{B}_{s,\infty,\infty}(\mathbb{R}^d)$, subject to the usual interpretation at integer smoothness indices. Therefore, the Besov formulation encompasses the familiar Sobolev and Hölder regimes while retaining the finer resolution-level control needed for wavelet approximation.

In what follows, Besov assumptions are imposed on the marginal density, on the conditional-density quantities entering the ergodic approximation conditions, and on the target derivative functional $(\partial^\beta r)(\rho; \cdot)$. Their role is explicit: they provide the approximation and embedding properties needed to control the projection bias and to guarantee boundedness or continuity where required. In particular, whenever an argument uses the embedding $\mathbf{B}_{s,p,q}(\mathbb{R}^d) \hookrightarrow L_\infty(\mathbb{R}^d)$, the dimension-dependent condition $s > d/p$ is required; the weaker condition $s > 1/p$ is in general not sufficient when $d > 1$.

2.2. Linear Wavelets Estimator Under Missing Responses

We now introduce the wavelet estimator under missing responses. Throughout this subsection, the wavelet notation follows [46]. Let $\{V_j\}_{j \in \mathbb{Z}}$ be a multiresolution analysis of $L_2(\mathbb{R}^d)$. Denote the scaling function by ϕ and the associated orthogonal wavelet by ψ , with both assumed to be r -regular ($r \geq 1$) and compactly supported in the hypercube $[-L, L]^d$ for some $L > 0$. For every integer j and every $\mathbf{k} \in \mathbb{Z}^d$, define

$$\phi_{j,\mathbf{k}}(\mathbf{x}) = 2^{jd/2} \phi(2^j \mathbf{x} - \mathbf{k}).$$

The family $\{\phi_{j,\mathbf{k}}\}_{\mathbf{k} \in \mathbb{Z}^d}$ is an orthonormal basis of V_j . Moreover, we assume that every partial derivative of ϕ having total order at most r satisfies the stated rapid-decay bound: for every integer $i > 0$, there exists a constant $A_i > 0$, such that

$$|(\partial^\beta \phi)(\mathbf{x})| \leq \frac{A_i}{(1 + \|\mathbf{x}\|)^i}, \quad |\beta| \leq r. \quad (1)$$

There exist $2^d - 1$ companion wavelets $\psi_1, \dots, \psi_{2^d-1}$, and the system

$$\left\{ \psi_{\ell,j,\mathbf{k}}(\mathbf{x}) = 2^{jd/2} \psi_\ell(2^j \mathbf{x} - \mathbf{k}) : \ell = 1, \dots, 2^d - 1, \mathbf{k} \in \mathbb{Z}^d \right\}$$

generates the detail spaces in the multiresolution decomposition of $L_2(\mathbb{R}^d)$.

We next specify the observation scheme. Let

$$(\mathbf{X}_i, \mathbf{Y}_i, \delta_i), \quad i = 1, \dots, n$$

be observations from a strictly stationary ergodic process, where $\mathbf{X}_i \in \mathbb{R}^d$ is fully observed, $\mathbf{Y}_i \in \mathbb{R}^{d'}$ is the response vector, and

$$\delta_i = \mathbb{1}_{\{\mathbf{Y}_i \text{ is observed}\}} = \begin{cases} 1, & \text{if } \mathbf{Y}_i \text{ is observed,} \\ 0, & \text{otherwise.} \end{cases}$$

The response is assumed to be missing at random (MAR), in the sense that

$$\pi(\mathbf{X}_i) = \mathbb{P}(\delta_i = 1 \mid \mathbf{X}_i, \mathbf{Y}_i) = \mathbb{P}(\delta_i = 1 \mid \mathbf{X}_i), \quad (2)$$

so that δ_i and $\mathbf{Y}_i$ are conditionally independent given $\mathbf{X}_i$. The propensity score is further assumed to satisfy the positivity condition

$$0 < \pi_0 \leq \pi(\mathbf{x}) \leq 1, \quad \mathbf{x} \in \mathbf{J} \quad (3)$$

for some constant $\pi_0 > 0$. Positivity ensures that the inverse weights are well-defined and uniformly bounded on the region on which the estimator is studied; without such a lower bound, the inverse-probability-weighted coefficients need not possess the moment properties required below.

Let $(\mathbf{X}, \mathbf{Y})$ have joint density $f_{\mathbf{X}, \mathbf{Y}}$ on $\mathbf{J} \times \mathbb{R}^{d'}$ and let

$$f_{\mathbf{X}}(\mathbf{x}) = \int_{\mathbb{R}^{d'}} f_{\mathbf{X}, \mathbf{Y}}(\mathbf{x}, \mathbf{y}) d\mathbf{y}.$$

For a measurable function $\rho : \mathbb{R}^{d'} \rightarrow \mathbb{R}$, define

$$r(\rho; \mathbf{x}) = \mathbb{E}[\rho(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}] f_{\mathbf{X}}(\mathbf{x}) = \int_{\mathbb{R}^{d'}} \rho(\mathbf{y}) f_{\mathbf{X}, \mathbf{Y}}(\mathbf{x}, \mathbf{y}) d\mathbf{y}.$$

Whenever differentiation under the integral sign is justified by the regularity and integrability assumptions imposed below, the estimand is

$$(\partial^\beta r)(\rho; \mathbf{x}) = \int_{\mathbb{R}^{d'}} \rho(\mathbf{y}) (\partial^\beta f_{\mathbf{X}, \mathbf{Y}})(\mathbf{x}, \mathbf{y}) d\mathbf{y},$$

where

$$\partial^\beta = \frac{\partial^{|\beta|}}{\partial x_1^{\beta_1} \cdots \partial x_d^{\beta_d}}, \quad |\beta| = \beta_1 + \cdots + \beta_d.$$

Thus, the displayed integral representation is not taken as a purely formal identity; it is used only under conditions ensuring the existence of the indicated derivative and the validity of the interchange between differentiation and integration.

Assume that $(\partial^\beta r)(\rho; \cdot) \in L_2(\mathbb{R}^d)$. To standardize the resolution notation throughout the paper, we write $\tau = m(n)$ henceforth, where $m(n)$ is integer-valued and $m(n) \rightarrow \infty$; equivalently, $h_n = 2^{-m(n)}$. As such, the symbols τ , m , and $m(n)$ do not denote distinct tuning parameters. Then, the orthogonal projection of $(\partial^\beta r)(\rho; \cdot)$ onto V_τ is

$$P_{V_\tau}[(\partial^\beta r)(\rho; \cdot)](\mathbf{x}) = \sum_{\mathbf{k} \in \mathbb{Z}^d} a_{\tau, \mathbf{k}} \phi_{\tau, \mathbf{k}}(\mathbf{x}), \quad (4)$$

with coefficients

$$a_{\tau, \mathbf{k}} = \int_{\mathbb{R}^d} (\partial^\beta r)(\rho; \mathbf{u}) \phi_{\tau, \mathbf{k}}(\mathbf{u}) d\mathbf{u} = (-1)^{|\beta|} \int_{\mathbb{R}^d} r(\rho; \mathbf{u}) (\partial^\beta \phi_{\tau, \mathbf{k}})(\mathbf{u}) d\mathbf{u}.$$

The second equality follows via integration by parts, provided that the weak derivative of the required order exists and the boundary term vanishes; these requirements are ensured, for example, by the regularity and support conditions imposed in the sequel. The projection operator on the left-hand side of (4) is essential: at a finite resolution τ , the series represents the V_τ -projection of the target and equals the target itself only when the latter already belongs to V_τ .

If the responses were fully observed, the natural empirical counterpart of the coefficient would be

$$\hat{a}_{\tau,\mathbf{k}}^{\text{full}} = \frac{(-1)^{|\beta|}}{n} \sum_{i=1}^n \rho(\mathbf{Y}_i) (\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{X}_i).$$

With missing responses, the corresponding complete-case coefficient is

$$\hat{a}_{\tau,\mathbf{k}}^{\text{cc}} = \frac{(-1)^{|\beta|}}{\sum_{j=1}^n \delta_j} \sum_{i=1}^n \delta_i \rho(\mathbf{Y}_i) (\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{X}_i).$$

Under MAR, this complete-case normalization does not in general recover the full-data coefficient, as the conditional observation probability may vary with the covariate; it is unbiased only under additional restrictions, such as a covariate-independent observation probability. To correct the resulting selection distortion, we use inverse probability weighting. If $\pi(\cdot)$ is known, then the IPW pseudo-coefficient is

$$\hat{a}_{\tau,\mathbf{k}}^{\text{ipw}} = \frac{(-1)^{|\beta|}}{n} \sum_{i=1}^n \frac{\delta_i}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) (\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{X}_i). \quad (5)$$

Indeed, by (2) and the tower property,

$$\mathbb{E} \left[\frac{\delta_i}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) (\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{X}_i) \right] = \mathbb{E} \left[\rho(\mathbf{Y}_i) (\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{X}_i) \right]$$

whenever either side is absolutely integrable. This compensation identity is the precise reason that the oracle IPW coefficient targets the same population coefficient as its full-data counterpart. Accordingly, the present construction should be understood as an inverse-probability-weighted extension of the complete-data wavelet coefficient rather than as a distinct projection principle. The new analytical issue is the control of the inverse-weighted stochastic term, and for the feasible procedure, of the error induced by estimating the propensity score.

In applications, $\pi(\cdot)$ is typically unknown. We estimate it by a Nadaraya–Watson estimator. Let $K^* : \mathbb{R}^d \rightarrow [0, \infty)$ be a kernel and let $\lambda_n > 0$ be the propensity score bandwidth, which is distinct from the wavelet bandwidth $h_n = 2^{-m(n)}$. Put

$$K_{\lambda_n}^*(\mathbf{u}) = \lambda_n^{-d} K^*(\lambda_n^{-1} \mathbf{u})$$

and define

$$\hat{\pi}_i = \hat{\pi}_n(\mathbf{X}_i) = \frac{\sum_{j=1}^n \delta_j K_{\lambda_n}^*(\mathbf{X}_i - \mathbf{X}_j)}{\sum_{j=1}^n K_{\lambda_n}^*(\mathbf{X}_i - \mathbf{X}_j)}. \quad (6)$$

This definition is meaningful in the event that the denominator is nonzero. The assumptions imposed in Section 3 are intended to guarantee uniform consistency on the relevant compact set and to ensure, together with positivity, that $\hat{\pi}_n$ is eventually bounded away from zero

with the stated mode of convergence. No such stability is inferred from the formula alone. The feasible MAR-adjusted coefficient is

$$\hat{a}_{\tau, \mathbf{k}}^{\text{mar}} = \frac{(-1)^{|\beta|}}{n} \sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_i} \rho(\mathbf{Y}_i) (\partial^\beta \phi_{\tau, \mathbf{k}})(\mathbf{X}_i). \quad (7)$$

Consequently, the feasible linear wavelet estimator of $(\partial^\beta r)(\rho; \cdot)$ under MAR is

$$\widehat{(\partial^\beta r)}_n^{\text{mar}}(\rho; \mathbf{x}) = \sum_{\mathbf{k} \in \mathbb{Z}^d} \hat{a}_{\tau, \mathbf{k}}^{\text{mar}} \phi_{\tau, \mathbf{k}}(\mathbf{x}). \quad (8)$$

The dependence of this estimator on two smoothing scales is structural: $m(n)$, or equivalently h_n , controls the wavelet approximation and stochastic resolution, whereas λ_n controls the accuracy of the estimated inverse-probability weights. The theoretical results must impose conditions on their joint asymptotic behavior; neither tuning parameter can in general be analyzed independently of the other.

Equivalently, using the wavelet projection kernel

$$K(\mathbf{u}, \mathbf{v}) = \sum_{\mathbf{k} \in \mathbb{Z}^d} \phi(\mathbf{u} - \mathbf{k}) \phi(\mathbf{v} - \mathbf{k}) \quad (9)$$

and its derivative version

$$K^{(\beta)}(\mathbf{u}, \mathbf{v}) = \sum_{\mathbf{k} \in \mathbb{Z}^d} \phi(\mathbf{u} - \mathbf{k}) (\partial_{\mathbf{v}}^\beta \phi)(\mathbf{v} - \mathbf{k}), \quad (10)$$

the estimator (8) can be written in kernel form as

$$\widehat{(\partial^\beta r)}_n^{\text{mar}}(\rho; \mathbf{x}) = \frac{(-1)^{|\beta|}}{nh_n^{d+|\beta|}} \sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_i} \rho(\mathbf{Y}_i) K^{(\beta)}\left(\frac{\mathbf{x}}{h_n}, \frac{\mathbf{X}_i}{h_n}\right), \quad h_n = 2^{-m(n)}. \quad (11)$$

This representation follows from the scaling relation for $\phi_{\tau, \mathbf{k}}$, the definition $\tau = m(n)$, and the finite-overlap properties inherited from compact support. It is an alternative representation of the same linear projection estimator, and does not introduce an additional kernel smoothing procedure.

The corresponding pseudo-estimator, obtained when the true propensity score is known, is

$$\widetilde{(\partial^\beta r)}_n^{\text{ipw}}(\rho; \mathbf{x}) = \frac{(-1)^{|\beta|}}{nh_n^{d+|\beta|}} \sum_{i=1}^n \frac{\delta_i}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) K^{(\beta)}\left(\frac{\mathbf{x}}{h_n}, \frac{\mathbf{X}_i}{h_n}\right). \quad (12)$$

The estimator (11) is the feasible procedure, whereas (12) serves as the oracle intermediate object in the theoretical analysis. Under (2) and (3), and the stated integrability conditions, the oracle weighting restores the full-data coefficient in expectation. Therefore, the additional task for the feasible estimator is to prove that replacing π by $\hat{\pi}_n$ is asymptotically negligible at the scale relevant to each risk bound or limit theorem; this conclusion is not automatic, and is established only under the explicit rate conditions stated later.

3. Assumptions and Main Results

We now formulate the assumptions and principal asymptotic results for the MAR-adjusted wavelet estimator introduced in (11). The hypotheses are organized according to their logically distinct functions: (C) controls conditional law stabilization and wavelet approximation; (M) specifies the missingness mechanism and the propensity estimator; (N) governs response moments and conditional regression structure; and (A) records the additional martingale array requirements used only for pointwise asymptotic normality.

This separation is substantive, with no quantitative rate inferred from ergodicity alone and no condition concerning estimation of the propensity score imposed on the oracle estimator.

The inverse-probability-weighting identity restores the complete-data target under MAR, whereas the feasible estimator replaces the unknown propensity score with its Nadaraya–Watson estimate. Thus, the resulting analysis must distinguish three errors: the wavelet projection bias, the stochastic fluctuation of the oracle IPW estimator, and the additional substitution error caused by estimating the propensity score.

Let

$$\mathcal{F}_n = \sigma((\mathbf{X}_i, \mathbf{Y}_i, \delta_i) : 1 \leq i \leq n)$$

and define

$$\mathcal{G}_n = \sigma((\mathbf{X}_k, \mathbf{Y}_k, \delta_k) : 1 \leq k \leq n; \mathbf{X}_{n+1}).$$

For every $i \geq 1$, let $f^{\mathcal{F}_{i-1}}(\cdot)$ denote the conditional density of $\mathbf{X}_i$ given $\mathcal{F}_{i-1}$. When needed, $g^{\mathcal{F}_{i-1}}(\cdot)$ denotes the conditional density of $\mathbf{Y}_i$ given $\mathcal{F}_{i-1}$.

All σ -fields are understood to be completed by the $\mathbb{P}$ -null sets. For the conditional identities in (N.2), we interpret

$$\mathcal{G}_{i-1} = \mathcal{F}_{i-1} \vee \sigma(\mathbf{X}_i),$$

which agrees with the displayed definition of $\mathcal{G}_n$ after the index shift $n = i - 1$. We assume that the regular conditional laws admit jointly measurable versions of the densities $(\omega, \mathbf{x}) \mapsto f^{\mathcal{F}_{i-1}}(\omega, \mathbf{x})$ and, when used, $(\omega, \mathbf{y}) \mapsto g^{\mathcal{F}_{i-1}}(\omega, \mathbf{y})$. Whenever a spatial supremum is taken, these versions are assumed to be spatially continuous on the relevant compact set, so the supremum is a measurable random variable. We write

$$m_\rho(\mathbf{x}) = \mathbb{E}[\rho(\mathbf{Y}_i) \mid \mathbf{X}_i = \mathbf{x}], \quad m_{\rho, \eta}(\mathbf{x}) = \mathbb{E}[|\rho(\mathbf{Y}_i)|^\eta \mid \mathbf{X}_i = \mathbf{x}].$$

The assumptions used in the respective results are given below; each theorem invokes only the subset explicitly listed in its statement.

(C.1) For every $\mathbf{x} \in \mathbf{S}$,

$$\frac{1}{n} \sum_{i=1}^n f^{\mathcal{F}_{i-1}}(\mathbf{x}) \longrightarrow f_{\mathbf{X}}(\mathbf{x}), \quad n \rightarrow \infty,$$

both almost surely and in L^2 .

(C.2) Moreover,

$$\sup_{\mathbf{x} \in \mathbb{R}^d} \left| \frac{1}{n} \sum_{i=1}^n f^{\mathcal{F}_{i-1}}(\mathbf{x}) - f_{\mathbf{X}}(\mathbf{x}) \right| \longrightarrow 0, \quad n \rightarrow \infty,$$

both almost surely and in L^2 .

(C.2[#]) There exists a constant $C_{\text{st}} < \infty$ such that

$$\mathbb{E} \left[\sup_{\mathbf{x} \in \mathbb{R}^d} \left| \frac{1}{n} \sum_{i=1}^n f^{\mathcal{F}_{i-1}}(\mathbf{x}) - f_{\mathbf{X}}(\mathbf{x}) \right|^2 \right] \leq \frac{C_{\text{st}}}{n}, \quad n \geq 1. \quad (13)$$

This quantitative strengthening is invoked only in results asserting an explicit n^{-1} mean-square rate. None of strict stationarity, ergodicity, (C.1), or (C.2) alone imply (13).

(C.3) (i) The multiresolution analysis is r -regular.

(ii) The target function $(\partial^{\beta} r)(\rho; \cdot)$ belongs to the Besov space $B_{s,p,q}$ for some $0 < s < r$ and $1 \leq p, q \leq \infty$.

- (C.4)** (i) The marginal density $f_{\mathbf{X}}$ belongs to $B_{s,p,q}$ for some $0 < s < r$ and $1 \leq p, q \leq \infty$.
(ii) The conditional densities $f^{\mathcal{F}_{i-1}}(\cdot)$ belong to $B_{s,p,q}$ for the same range of parameters.
- (C.5)** The Besov parameters in (C.4)(i) satisfy $s > d/p$. Consequently, the Besov embedding theorem gives $f_{\mathbf{X}} \in L^\infty(\mathbb{R}^d)$. Moreover, there exists a compact rectangle $\mathbf{J} \subset \mathbb{R}^d$ such that

$$\text{supp}(f_{\mathbf{X}}) \cup \text{supp}(r(\rho; \cdot)) \subset \mathbf{J}.$$

The compact-support condition is used when the global integrated risk is reduced to a finite sum of coefficient risks. It may be replaced by a uniformly summable tail condition on the coefficient family.

- (M.1)** The missingness mechanism is missing-at-random:

$$\pi(\mathbf{X}_i) = \mathbb{P}(\delta_i = 1 \mid \mathbf{X}_i, \mathbf{Y}_i) = \mathbb{P}(\delta_i = 1 \mid \mathbf{X}_i).$$

Equivalently, conditionally on $\mathbf{X}_i$, the indicator δ_i is independent of $\mathbf{Y}_i$.

- (M.2)** The propensity score satisfies the positivity condition

$$0 < \pi_0 \leq \pi(\mathbf{x}) \leq 1, \quad \mathbf{x} \in \mathbf{J}$$

for some constant $\pi_0 > 0$.

- (M.3)** The propensity score $\pi(\cdot)$ is continuous on $\mathbf{J}$ and satisfies a Hölder condition: there exist constants $\gamma_\pi > 0$ and $C_\pi > 0$ such that

$$|\pi(\mathbf{u}) - \pi(\mathbf{v})| \leq C_\pi \|\mathbf{u} - \mathbf{v}\|^{\gamma_\pi}, \quad \mathbf{u}, \mathbf{v} \in \mathbf{J}.$$

- (M.4)** The estimator $\hat{\pi}_n$ defined in (6) is uniformly consistent on $\mathbf{J}$. More precisely, there exists a deterministic sequence $a_n \rightarrow 0$ such that

$$\sup_{\mathbf{x} \in \mathbf{J}} |\hat{\pi}_n(\mathbf{x}) - \pi(\mathbf{x})| = O(a_n) \quad \text{a.s.}$$

In particular, for all sufficiently large n ,

$$\inf_{\mathbf{x} \in \mathbf{J}} \hat{\pi}_n(\mathbf{x}) \geq \pi_0/2 \quad \text{a.s.}$$

For the Nadaraya–Watson estimator in (6), a typical choice is

$$a_n = \left(\frac{\log n}{n \lambda_n^d} \right)^{1/2} + \lambda_n^{\gamma_\pi},$$

where λ_n is the bandwidth used to estimate the propensity score.

For every assertion formulated in mean square for the feasible estimator, we also assume

$$\mathbb{E} \left[\sup_{\mathbf{x} \in \mathbf{J}} |\hat{\pi}_n(\mathbf{x}) - \pi(\mathbf{x})|^2 \right] = O(a_n^2). \quad (14)$$

Almost-sure uniform consistency does not by itself imply (14).

- (N.0)** For some $\nu > 2$,

$$\mathbb{E}[|\rho(\mathbf{Y}_1)|^\nu] < \infty.$$

This is the baseline moment hypothesis. Every argument requiring a fourth moment, including the fourth-order risk bound, is invoked only under the explicitly strength-

ened condition $\nu \geq 4$. Thus, the paper does not infer $\mathbb{E}|\rho(\mathbf{Y}_1)|^4 < \infty$ from the weaker assumption $\nu > 2$.

(N.1) On the set $\{\mathbf{y} \in \mathbb{R}^{d'} : g(\mathbf{y}) > 0\}$,

$$\sup_{\mathbf{y}: g(\mathbf{y}) > 0} \left| \frac{1}{ng(\mathbf{y})} \sum_{i=1}^n g^{\mathcal{F}_{i-1}}(\mathbf{y}) - 1 \right| \rightarrow 0, \quad n \rightarrow \infty,$$

both almost surely and in L^2 .

The ratio in (N.1) is understood only on the set $\{\mathbf{y} : g(\mathbf{y}) > 0\}$. An equivalent and technically safer formulation is to impose the corresponding convergence on $n^{-1} \sum_{i=1}^n g^{\mathcal{F}_{i-1}} - g$ without division by g .

(N.2) (i) For every $i \geq 1$,

$$\mathbb{E}[\rho(\mathbf{Y}_i) \mid \mathcal{G}_{i-1}] = \mathbb{E}[\rho(\mathbf{Y}_i) \mid \mathbf{X}_i] = m_\rho(\mathbf{X}_i).$$

(ii) For every $\eta \leq 2$,

$$\mathbb{E}[|\rho(\mathbf{Y}_i)|^\eta \mid \mathcal{G}_{i-1}] = \mathbb{E}[|\rho(\mathbf{Y}_i)|^\eta \mid \mathbf{X}_i] = m_{\rho, \eta}(\mathbf{X}_i),$$

and $m_{\rho, \eta}$ is continuous on $\mathbf{J}$.

(N.3) (i) There exists a constant $C_{\mathbf{m}} > 0$ such that

$$\sup_{\mathbf{x} \in \mathbf{J}} |m_\rho(\mathbf{x})| \leq C_{\mathbf{m}}.$$

(ii) The regression function m_ρ satisfies a Hölder condition: there exist constants $\beta > 0$ and $c_3 > 0$ such that

$$|m_\rho(\mathbf{u}) - m_\rho(\mathbf{v})| \leq c_3 \|\mathbf{u} - \mathbf{v}\|^\beta, \quad \mathbf{u}, \mathbf{v} \in \mathbb{R}^d.$$

(iii) The function $m_{\rho, \eta}$ satisfies a Hölder condition: there exist constants $\beta_1 > 0$ and $c_4 > 0$ such that

$$|m_{\rho, \eta}(\mathbf{u}) - m_{\rho, \eta}(\mathbf{v})| \leq c_4 \|\mathbf{u} - \mathbf{v}\|^{\beta_1}, \quad \mathbf{u}, \mathbf{v} \in \mathbb{R}^d.$$

(C.6) For the deterministic resolution sequence $m = m(n)$, define

$$D_{m, \mathbf{k}}(\mathbf{x}) = (\partial^{\beta} \phi_{m, \mathbf{k}})(\mathbf{x}).$$

There exists a constant $C_{\text{coef}} < \infty$, independent of n, m , and $\mathbf{k}$, such that

$$\sup_{\mathbf{k} \in \mathbb{Z}^d} \mathbb{E} \left[\left| \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}^d} m_\rho(\mathbf{x}) D_{m, \mathbf{k}}(\mathbf{x}) \{f^{\mathcal{F}_{i-1}}(\mathbf{x}) - f_{\mathbf{X}}(\mathbf{x})\} d\mathbf{x} \right|^2 \right] \leq C_{\text{coef}} \frac{2^{2m(d/\nu + |\beta|)}}{n}.$$

Condition (C.6) is imposed only for the integrated-risk result. It is the coefficient-level stabilization estimate actually used in its proof, and as such is weaker than a uniform L^2 rate for the entire conditional density field.

The conditions (C.1)–(C.6), (M.1)–(M.4), and (N.0)–(N.3) are the primitive assumptions of the section. They are not cumulative; each theorem invokes only the subset stated in its hypotheses. Maximal inequalities, predictable quadratic variation, Lindeberg negligibility,

projection bias control, and feasible–oracle equivalence are theorem-specific asymptotic requirements and are stated directly in the corresponding theorem.

Lemma 1. Assume (N.0) and $|\beta| \leq r$. Then, for every deterministic $m \in \mathbb{N}$ and every $\mathbf{k} \in \mathbb{Z}^d$,

$$\mathbb{E}\left[|\rho(\mathbf{Y}_1)|\left|(\partial^\beta \phi_{m,\mathbf{k}})(\mathbf{X}_1)\right|\right] < \infty$$

and

$$\mathbb{E}\left[|\rho(\mathbf{Y}_1)|^2\left|(\partial^\beta \phi_{m,\mathbf{k}})(\mathbf{X}_1)\right|^2\right] < \infty.$$

Proof. Since $\nu > 2$, both $\mathbb{E}|\rho(\mathbf{Y}_1)|$ and $\mathbb{E}|\rho(\mathbf{Y}_1)|^2$ are finite. For fixed m and $\mathbf{k}$, the derivative $\partial^\beta \phi_{m,\mathbf{k}}$ is bounded by r -regularity. The assertions follow. $\square$

Lemma 2. Assume that the scaling function is compactly supported and r -regular, with $|\beta| \leq r$. Then, for every $\eta > 0$,

$$\sup_{\boldsymbol{\theta} \in [0,1]^d} \int_{\mathbb{R}^d} \left|K^{(\beta)}(\boldsymbol{\theta}, \boldsymbol{\theta} + \mathbf{v})\right|^{2+\eta} d\mathbf{v} < \infty.$$

In particular, the corresponding uniform L^2 bound holds. Moreover, for almost every $\mathbf{v}$, the map

$$\boldsymbol{\theta} \longmapsto K^{(\beta)}(\boldsymbol{\theta}, \boldsymbol{\theta} + \mathbf{v})$$

is continuous on $[0,1]^d$, and the integrable envelope can be chosen independently of $\boldsymbol{\theta}$.

Proof. Compact support implies that only finitely many terms contribute to the series defining $K^{(\beta)}$, with the number of nonzero terms bounded uniformly in $\boldsymbol{\theta}$. The r -regularity of ϕ gives uniform boundedness and continuity of the contributing factors. As a function of $\mathbf{v}$, the kernel is supported on a fixed compact set. The asserted integrability and continuity properties follow. $\square$

A convenient sufficient condition for (C.6) is

$$\mathbb{E}\left[\sup_{\mathbf{x} \in \mathbf{J}} \left|\frac{1}{n} \sum_{i=1}^n f^{\mathcal{F}_{i-1}}(\mathbf{x}) - f_{\mathbf{X}}(\mathbf{x})\right|^2\right] \leq \frac{C}{n}.$$

Under (N.3)(i), compact support, and $|\beta| \leq r$, this stronger uniform condition implies (C.6) through the scaling relation $\|D_{m,\mathbf{k}}\|_1 \asymp 2^{m(|\beta|-d/2)}$. Therefore, it is sufficient but not part of the minimal theorem statement.

The MAR condition implies the following compensation identity, while positivity guarantees that its inverse weights are well defined and uniformly bounded on $\mathbf{J}$:

$$\mathbb{E}\left[\frac{\delta_i}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) \varphi(\mathbf{X}_i)\right] = \mathbb{E}[\rho(\mathbf{Y}_i) \varphi(\mathbf{X}_i)] \quad (15)$$

for every integrable measurable function φ . This identity is the main device that allows the wavelet estimator under MAR to target the same quantity as the complete-data estimator.

More precisely, (15) follows from

$$\mathbb{E}\left[\frac{\delta_i}{\pi(\mathbf{X}_i)} \middle| \mathbf{X}_i, \mathbf{Y}_i\right] = 1 \quad \text{a.s.,}$$

and consequently remains valid for every φ such that either side is absolutely integrable.

Remark 1. The wavelet bandwidth $h = 2^{-m}$ and propensity score bandwidth λ control distinct but interacting errors. The former determines the projection bias and stochastic fluctuation of the derivative estimator, whereas the latter determines the bias and variability of the estimated inverse-probability weights. Therefore, a practical selection procedure should choose the pair (h, λ) jointly rather than optimizing the two parameters in isolation.

For a stationary dependent sample, we use blocked cross-validation. Let $\mathcal{H}_n$ and Λ_n be finite grids of admissible wavelet and propensity score bandwidths, respectively, and partition $\{1, \dots, n\}$ into B consecutive blocks $I_1, \dots, I_B$. For each $(h, \lambda) \in \mathcal{H}_n \times \Lambda_n$ and each block I_b , compute the Nadaraya–Watson propensity estimator $\hat{\pi}_{-b,\lambda}$ and the corresponding wavelet regression estimator $\hat{m}_{-b,h,\lambda}$ from the observations outside I_b . To preserve positivity in finite samples, put

$$\hat{\pi}_{-b,\lambda}^\varepsilon(x) = \max\{\hat{\pi}_{-b,\lambda}(x), \varepsilon_n\}, \quad \varepsilon_n \downarrow 0$$

and define the blocked inverse-probability-weighted validation criterion

$$\text{BCV}_n(h, \lambda) = \frac{1}{n} \sum_{b=1}^B \sum_{i \in I_b} \frac{\delta_i}{\hat{\pi}_{-b,\lambda}^\varepsilon(X_i)} \{Y_i - \hat{m}_{-b,h,\lambda}(X_i)\}^2.$$

The jointly selected pair is

$$(\hat{h}_n, \hat{\lambda}_n) \in \arg \min_{(h, \lambda) \in \mathcal{H}_n \times \Lambda_n} \text{BCV}_n(h, \lambda).$$

When several pairs produce numerically indistinguishable criterion values, we select the smoother admissible pair, thereby avoiding unstable inverse weights and excessively small effective local sample sizes.

The candidate grids are restricted to the theoretical region

$$h \rightarrow 0, \quad \lambda \rightarrow 0, \quad \frac{nh^d}{\log n} \rightarrow \infty, \quad \frac{n\lambda^d}{\log n} \rightarrow \infty,$$

and for asymptotic normality are restricted to pairs satisfying the additional undersmoothing and propensity negligibility conditions imposed in Theorem 3. Thus, blocked cross-validation provides a fully data-driven finite-sample choice within the set of bandwidth pairs allowed by the asymptotic theory. In the simulation study, the Silverman reference pair is retained as a transparent benchmark, while the separate sensitivity curves for h and λ display the interaction that motivates the joint criterion above.

3.1. Discussion and Interpretation of the Assumptions

We collect here several comments on the assumptions introduced above. The purpose of this discussion is twofold. First, it identifies the precise probabilistic and analytic role played by each group of conditions. Second, it clarifies how the assumptions interact in the martingale approximation, the wavelet bias analysis, and the treatment of the missing-at-random mechanism. The conditions naturally fall into four classes: conditional-law stabilization assumptions, wavelet–Besov approximation assumptions, missing data assumptions, and moment and regression regularity assumptions.

3.1.1. Conditional-Law Stabilization

Assumptions (C.1) and (C.2) are ergodic stabilization conditions imposed on the conditional densities of the covariate process. They replace the quantitative covariance inequalities that would ordinarily be derived from strong-mixing coefficients. More precisely, (C.1) requires that the Cesàro averages of the conditional densities converge pointwise

to the stationary marginal density f_X , whereas (C.2) imposes the corresponding uniform convergence.

These assumptions are adapted to the martingale structure of the proofs. After conditioning on the past sigma field, the empirical terms are decomposed into a martingale difference component and a predictable remainder. The latter depends on averages of conditional densities, and its asymptotic negligibility follows directly from (C.1) or (C.2) depending on whether pointwise or uniform control is required. Thus, the analysis does not require an explicit rate of decay for the dependence coefficients; instead, it requires that the conditional law of a future covariate, averaged along the trajectory, asymptotically recover the stationary marginal law.

This property is best described as *ergodic stabilization*. It should not be interpreted as robustness in the classical statistical sense, as neither contamination, model mis-specification, nor resistance to outliers is at issue here. Rather, it is an asymptotic regularity property of the conditional distributions. It is precisely this stabilization that permits ergodicity to replace stronger quantitative dependence assumptions in the martingale approximation.

3.1.2. Wavelet Regularity and Multiresolution Approximation

Assumptions (C.3) and (C.4) govern the analytic structure of the wavelet estimator. The r -regularity of the multiresolution analysis guarantees that the scaling function and the associated wavelets possess sufficient smoothness and localization for termwise differentiation of the wavelet expansion and for the derivative projection kernel estimates used throughout the proofs.

The Besov assumptions play a distinct and more substantive role. Let P_{V_m} denote the projection onto the approximation space V_m . For the target derivative functional, the error may be decomposed as

$$(\widehat{\partial^\beta r})_n - (\partial^\beta r) = \{(\widehat{\partial^\beta r})_n - P_{V_m}(\partial^\beta r)\} + \{P_{V_m}(\partial^\beta r) - (\partial^\beta r)\}.$$

The first term is stochastic, whereas the second is the deterministic *multiresolution projection bias*. Controlling this bias means proving that the error produced by truncating the wavelet expansion at resolution $m = m(n)$ tends to zero at a rate compatible with the stochastic normalization of the estimator.

In particular, if

$$(\partial^\beta r)(\rho; \cdot) \in B_{s,p,q}, \quad s > \frac{d}{p},$$

then the wavelet approximation inequality used below yields

$$\left\| P_{V_m}(\partial^\beta r)(\rho; \cdot) - (\partial^\beta r)(\rho; \cdot) \right\|_\infty \leq C 2^{-m(s-d/p)} J_{s,p,q}((\partial^\beta r)(\rho; \cdot)).$$

Since $h_n = 2^{-m(n)}$, the corresponding uniform approximation error is of order

$$h_n^{s-d/p}.$$

Accordingly, the Besov condition is not imposed merely to legitimate differentiation; by quantifying the decay of the high-frequency wavelet coefficients, it determines the order of the deterministic bias. This control is indispensable for the integrated risk bounds, the uniform convergence theorem, and the undersmoothing condition required for asymptotic normality.

The Besov assumptions imposed on f_X and on the conditional densities $f^{\mathcal{F}_{i-1}}$ serve a related but separate purpose. They provide uniform local approximation properties for the deterministic and predictable terms arising in the conditional-variance calculations.

In particular, they justify replacing localized conditional density averages by their stationary limits at the spatial scale h_n .

3.1.3. Missing-at-Random Structure and Inverse Probability Weighting

Assumptions (M.1) and (M.2) encode the missing-at-random mechanism. Assumption (M.1) states that, conditionally on the covariate vector, the missingness indicator contains no additional information about the response. Therefore, it yields the compensation identity

$$\mathbb{E} \left[\frac{\delta_i}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) \mid \mathbf{X}_i \right] = \mathbb{E}[\rho(\mathbf{Y}_i) \mid \mathbf{X}_i],$$

which is the basic identification relation underlying the IPW construction.

The positivity assumption (M.2) is equally essential. By requiring that the propensity score be bounded away from zero, it prevents the inverse weights from becoming arbitrarily large. This guarantees that the weighted moments remain finite under the stated moment conditions and rules out regions of the covariate space in which the response is essentially unobservable. Without positivity, the variance of the IPW estimator may diverge and the target functional need not be identifiable from the observed sample.

Assumptions (M.3) and (M.4) are needed only for the feasible estimator, when π is unknown. The Hölder regularity in (M.3) controls the smoothing bias of the nonparametric propensity estimator, whereas (M.4) summarizes its uniform estimation error. For the Nadaraya–Watson estimator, the canonical decomposition is a_n , where the first term is stochastic and the second is the deterministic smoothing bias.

The additional conditions involving a_n are scale separation conditions. Because they require the error induced by estimating π to be asymptotically smaller than the principal stochastic fluctuation of the wavelet estimator, they express neither an independent smoothness assumption nor an additional identifiability restriction; rather, their sole purpose is to ensure first-order equivalence between the feasible estimator and the pseudo-estimator constructed with the true propensity score. When π is known by design, $a_n = 0$ and these restrictions disappear.

3.1.4. Moment, Truncation, and Conditional Regression Assumptions

Assumption (N.0) imposes a moment condition on the transformed response $\rho(\mathbf{Y})$. It is used at several distinct stages. It controls the contribution of large response values in the truncation argument, provides the integrability required by the martingale moment inequalities, and yields the Lindeberg-type negligibility needed for the central limit theorem. The threshold $\nu > 2$ is the natural minimal requirement for second-order arguments; stronger moment conditions may be invoked whenever fourth-order estimates are needed.

Assumption (N.1) is the response-density analogue of the uniform stabilization condition (C.2). It controls, uniformly over the relevant spatial domain, the predictable contribution of the truncated tail conditional on the past. Together with (N.0), it ensures that the large-response component is negligible at the scale of the uniform stochastic bound.

More precisely, the statement that the unboundedness of $\rho(\mathbf{Y})$ does not affect the uniform convergence rate concerns the *asymptotic order* of the rate, not merely the fact of convergence. Under (N.0) and (N.1), the tail term generated by the truncation procedure is of smaller order than the leading stochastic term. Consequently, the estimator attains the same uniform rate as in the bounded response case, although the proof requires an additional truncation-and-remainder argument. If ρ is bounded or if the response itself is bounded, this part of the argument becomes unnecessary and (N.1) may be correspondingly weakened.

Assumption (N.2) specifies the conditional moment structure needed for the regression interpretation of the target and for the martingale decomposition. Its first part asserts that, conditionally on the current covariate, the past does not further alter the relevant conditional mean of the response transformation. This is substantially weaker than independence. Its role is to ensure that the empirical summands form a martingale difference array after subtracting the appropriate conditional expectation. The second part supplies the corresponding conditional second-moment identity, and thereby determines the limiting variance.

Assumption (N.3) imposes boundedness and local Hölder regularity on the regression-type functions entering the first- and second-order conditional moments. Boundedness prevents the localized kernel sums from acquiring uncontrolled deterministic contributions. Hölder regularity justifies the local replacements

$$m_\rho(\mathbf{X}_i) = m_\rho(\mathbf{x}) + O(\|\mathbf{X}_i - \mathbf{x}\|^\gamma),$$

and similarly for $m_{\rho,\eta}$, whenever $\|\mathbf{X}_i - \mathbf{x}\| = O(h_n)$. These local expansions are used both in the proof of the uniform convergence theorem and in the identification of the pointwise asymptotic variance.

3.1.5. Resolution, Bandwidth, and Truncation Scales

The restrictions on $m(n)$, h_n , λ_n , and T_n balance four quantities: stochastic fluctuation, wavelet projection bias, propensity score estimation error, and response tail truncation. The conditions

$$m(n) \rightarrow \infty, \quad \frac{2^{dm(n)} \log n}{n} \rightarrow 0,$$

or equivalently

$$h_n \rightarrow 0, \quad \frac{\log n}{nh_n^d} \rightarrow 0,$$

ensure that the approximation space becomes asymptotically dense and that the effective number of observations in a localization cell of volume h_n^d tends to infinity.

The restriction involving T_n ensures that the truncation threshold diverges sufficiently quickly for the truncated estimator to remain asymptotically equivalent to the original one and sufficiently slowly for martingale exponential inequalities to remain effective. The conditions involving a_n guarantee that the feasible propensity score correction is of smaller order than the leading wavelet fluctuation. Finally, the undersmoothing condition imposed in the central limit theorem requires

$$\text{normalizing scale} \times \text{projection bias} \rightarrow 0,$$

so that the limiting distribution is centered at the target functional rather than at its finite-resolution projection.

3.1.6. Summary

The assumptions separate the three structural difficulties of the problem. Temporal dependence is handled through conditional-law stabilization and martingale approximation; missingness is handled through the MAR compensation identity and positivity; and nonparametric approximation is handled through the interaction between Besov regularity and multiresolution projection. The feasible case introduces an additional second-order component, propensity score estimation, for which the contribution is required to be asymptotically negligible relative to the intrinsic fluctuation of the wavelet estimator.

This separation is central to the scope of the theory. It permits the complete-data wavelet methodology to be extended to missing-at-random responses under stationary ergodic dependence without imposing explicit strong mixing rates and without changing the first-order asymptotic behavior of the estimator when the propensity score is estimated with sufficient accuracy.

Remark 2. *The use of strict stationarity and ergodic conditional density stabilization in place of quantitative strong mixing assumptions enlarges the class of admissible data-generating mechanisms, but should not be interpreted as yielding uniformly sharper finite-sample guarantees. The numerical form of the estimator is unchanged by this choice of dependence framework: for fixed smoothing parameters, the evaluation of $\widehat{(\partial^\beta r)}_n^{\text{mar}}$ requires the same propensity score computation and the same wavelet projection kernel sums irrespective of whether the observations are independent, mixing, or only stationary and ergodic; hence, the relaxation from mixing to ergodicity does not by itself impose an additional computational cost. The trade-off is probabilistic rather than algorithmic; under a specified mixing rate, covariance inequalities and exponential bounds may provide explicit dependence-adjusted constants, effective-sample-size interpretations, and finite-sample remainder estimates. Assumptions (C.1)–(C.2) instead require the Cesàro averages of the relevant conditional densities to stabilize toward their stationary marginals without postulating a universal rate for this stabilization. Accordingly, the asymptotic orders in the present paper remain valid under the stated bandwidth and stabilization conditions, but the finite-sample constants and the sample size at which the asymptotic approximation becomes accurate may depend strongly on the persistence of the underlying process. For highly persistent sequences, the effective information content of a sample of size n may be appreciably smaller than in a weakly dependent mixing model, even though the formal normalization and the leading bias–variance orders are unchanged. This distinction is illustrated in Section 5. The sensitivity analysis with respect to the autoregressive parameter and the comparison across several stationary ergodic mechanisms show a moderate increase in MISE as dependence becomes stronger, while preserving the qualitative ordering of the competing estimators. The reported CPU-time experiment further confirms that computational cost is governed primarily by n , the resolution level, and the propensity score smoothing step rather than by whether the dependence assumptions are formulated through mixing coefficients or through ergodic conditional density stabilization.*

Lemma 3. *Under condition (C.3), for $s > d/p$ we have*

$$\begin{aligned} \sup_{\mathbf{x} \in \mathbb{R}^d} \left| \mathbb{E} \left[\widetilde{(\partial^\beta r)}_n^{\text{ipw}}(\rho; \mathbf{x}) \right] - (\partial^\beta r)(\rho; \mathbf{x}) \right| &= \sup_{\mathbf{x} \in \mathbb{R}^d} \left| P_{V_m}(\partial^\beta r)(\rho; \mathbf{x}) - (\partial^\beta r)(\rho; \mathbf{x}) \right| \\ &\leq C 2^{-(s-d/p)m(n)} J_{s,p,q}((\partial^\beta r)(\rho; \cdot)). \end{aligned}$$

For the integrated risk theorem, the relevant approximation statement is the L^2 bound:

$$\left\| P_{V_m}(\partial^\beta r)(\rho; \cdot) - (\partial^\beta r)(\rho; \cdot) \right\|_2^2 \leq C 2^{-2ms} \left\| (\partial^\beta r)(\rho; \cdot) \right\|_{B_{s,2,q}}^2. \quad (16)$$

The uniform bound in Lemma 3 and the L^2 bound in (16) are distinct consequences of Besov regularity, and are not interchangeable. Define the wavelet projection kernel

$$K(\mathbf{u}, \mathbf{v}) = \sum_{\mathbf{k} \in \mathbb{Z}^d} \phi(\mathbf{u} - \mathbf{k}) \phi(\mathbf{v} - \mathbf{k})$$

and its derivative version

$$K^{(\beta)}(\mathbf{u}, \mathbf{v}) = \sum_{\mathbf{k} \in \mathbb{Z}^d} \phi(\mathbf{u} - \mathbf{k})(\partial_{\mathbf{v}}^{\beta} \phi)(\mathbf{v} - \mathbf{k}).$$

As in the complete-data case, for $|\beta| \leq r$,

$$|K^{(\beta)}(\mathbf{u}, \mathbf{v})| \leq \frac{C_{d+|\beta|+1}}{(1 + \|\mathbf{v} - \mathbf{u}\|_2)^{d+|\beta|+1}}. \quad (17)$$

Furthermore,

$$\int_{\mathbb{R}^d} |K^{(\beta)}(\mathbf{v}, \mathbf{u})|^j d\mathbf{v} \leq G_j(d), \quad j \geq 1.$$

By the regularity of the kernel,

$$|K^{(\beta)}(\mathbf{u}, \mathbf{y}) - K^{(\beta)}(\mathbf{v}, \mathbf{y})| \leq d^{1/2} C_2 \|\mathbf{u} - \mathbf{v}\|_2. \quad (18)$$

Therefore, the feasible MAR estimator can be written as

$$\widehat{(\partial^{\beta} r)}_n^{\text{mar}}(\rho; \mathbf{x}) = \frac{(-1)^{|\beta|}}{nh_n^{d+|\beta|}} \sum_{i=1}^n \frac{\delta_i}{\widehat{\pi}_n(\mathbf{X}_i)} \rho(\mathbf{Y}_i) K^{(\beta)}\left(\frac{\mathbf{x}}{h_n}, \frac{\mathbf{X}_i}{h_n}\right), \quad h_n = 2^{-m(n)}. \quad (19)$$

The associated pseudo-estimator, based on the true propensity score, is

$$\widetilde{(\partial^{\beta} r)}_n^{\text{ipw}}(\rho; \mathbf{x}) = \frac{(-1)^{|\beta|}}{nh_n^{d+|\beta|}} \sum_{i=1}^n \frac{\delta_i}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) K^{(\beta)}\left(\frac{\mathbf{x}}{h_n}, \frac{\mathbf{X}_i}{h_n}\right). \quad (20)$$

3.2. Theorem-Specific Auxiliary Hypotheses

The assumptions in this subsection are local to the result in which they are invoked. Conditions (U.1)–(U.5) are used only for uniform convergence, while (CLT.1)–(CLT.8) are used only for pointwise asymptotic normality. They are separated from (C), (M), and (N) in order to distinguish structural assumptions on the model from asymptotic conditions on the resolution, truncation, conditional laws, and estimated propensity score.

3.2.1. Conditions for Uniform Convergence

(U.1) There exists a compact set $D \subset \text{int}(\mathbf{J})$, and the derivative projection kernel satisfies the following for some finite constants $C_K, C_{K,1}$:

$$|K^{(\beta)}(\mathbf{u}, \mathbf{v})| \leq \frac{C_K}{(1 + \|\mathbf{u} - \mathbf{v}\|)^{d+|\beta|+1}}, \quad (21)$$

$$|K^{(\beta)}(\mathbf{u}, \mathbf{y}) - K^{(\beta)}(\mathbf{v}, \mathbf{y})| \leq C_{K,1} \|\mathbf{u} - \mathbf{v}\|. \quad (22)$$

Moreover, for every $j \geq 1$,

$$\sup_{\mathbf{u} \in \mathbb{R}^d} \int_{\mathbb{R}^d} |K^{(\beta)}(\mathbf{u}, \mathbf{v})|^j d\mathbf{v} < \infty.$$

(U.2) For some $\nu > 2$,

$$\sup_{i \geq 1} \|\mathbb{E}[|\rho(\mathbf{Y}_i)|^{\nu} \mid \mathcal{G}_{i-1}]\|_{\infty} < \infty.$$

This is stronger than the marginal moment condition (N.0) and is required for the conditional truncation and martingale inequalities used below.

(U.3) The conditional densities are uniformly bounded:

$$\sup_{i \geq 1} \|f^{\mathcal{F}_{i-1}}\|_{\infty} < \infty \quad \text{a.s.}$$

Furthermore, with

$$\Delta_n(\mathbf{y}) = \frac{1}{n} \sum_{i=1}^n f^{\mathcal{F}_{i-1}}(\mathbf{y}) - f_{\mathbf{X}}(\mathbf{y}),$$

there exists a deterministic sequence $b_n \downarrow 0$ such that

$$\|\Delta_n\|_{\infty} = O(b_n) \quad \text{a.s.} \quad (23)$$

(U.4) The bandwidth and truncation sequences satisfy

$$h_n \downarrow 0, \quad \frac{nh_n^{d+2|\beta|}}{\log n} \rightarrow \infty, \quad (24)$$

$$T_n \uparrow \infty, \quad T_n \left(\frac{\log n}{nh_n^d} \right)^{1/2} \rightarrow 0, \quad (25)$$

$$\sum_{n=1}^{\infty} n T_n^{-\nu} < \infty. \quad (26)$$

The stronger summability condition $\sum_n n T_n^{-\nu} < \infty$, rather than $\sum_n T_n^{-\nu} < \infty$, is what controls $\max_{1 \leq i \leq n} |\rho(\mathbf{Y}_i)|$.

(U.5) The propensity estimator satisfies

$$\|\hat{\pi}_n - \pi\|_{\infty, \mathbf{J}} = O(a_n) \quad \text{a.s.}$$

and

$$a_n h_n^{-|\beta|} = o \left[\left(\frac{\log n}{nh_n^{d+2|\beta|}} \right)^{1/2} + b_n h_n^{-|\beta|} \right]. \quad (27)$$

3.2.2. Conditions for Pointwise Asymptotic Normality

(CLT.1) Fix $\mathbf{x} \in \text{int}(\mathbf{J})$, and let $h_n = 2^{-m(n)} \downarrow 0$. There exists $\boldsymbol{\vartheta} \in [0, 1]^d$ such that

$$\boldsymbol{\vartheta}_n := \frac{\mathbf{x}}{h_n} - \left\lfloor \frac{\mathbf{x}}{h_n} \right\rfloor \rightarrow \boldsymbol{\vartheta}. \quad (28)$$

The floor and the fractional part are taken componentwise.

(CLT.2) The functions $m_{\rho,2}$, π , and $f_{\mathbf{X}}$ are continuous at $\mathbf{x}$, with

$$\pi(\mathbf{x}) > 0, \quad m_{\rho,2}(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) > 0.$$

Moreover, there exists a neighbourhood $U_{\mathbf{x}}$ of $\mathbf{x}$ such that

$$\sup_{\mathbf{u} \in U_{\mathbf{x}}} m_{\rho,2}(\mathbf{u}) < \infty, \quad \inf_{\mathbf{u} \in U_{\mathbf{x}}} \pi(\mathbf{u}) > 0.$$

(CLT.3) The conditional densities stabilize locally and uniformly at the shrinking spatial scale: for every $M < \infty$,

$$\sup_{\|\mathbf{v}\| \leq M} \left| \frac{1}{n} \sum_{i=1}^n f^{\mathcal{F}_{i-1}}(\mathbf{x} + h_n \mathbf{v}) - f_{\mathbf{X}}(\mathbf{x} + h_n \mathbf{v}) \right| \xrightarrow{\mathbb{P}} 0. \quad (29)$$

There exists a finite constant C_f such that for all sufficiently large n ,

$$\frac{1}{n} \sum_{i=1}^n f^{\mathcal{F}_{i-1}}(\mathbf{x} + h_n \mathbf{v}) \leq C_f \quad \text{a.s. for every } \mathbf{v} \in \mathbb{R}^d. \quad (30)$$

(CLT.4) The derivative projection kernel satisfies

$$\sup_{\boldsymbol{\theta} \in [0,1]^d} \int_{\mathbb{R}^d} |K^{(\beta)}(\boldsymbol{\theta}, \boldsymbol{\theta} + \mathbf{v})|^2 d\mathbf{v} < \infty \quad (31)$$

and, for some $\eta > 0$,

$$\sup_{\boldsymbol{\theta} \in [0,1]^d} \int_{\mathbb{R}^d} |K^{(\beta)}(\boldsymbol{\theta}, \boldsymbol{\theta} + \mathbf{v})|^{2+\eta} d\mathbf{v} < \infty. \quad (32)$$

The map

$$\boldsymbol{\theta} \longmapsto K^{(\beta)}(\boldsymbol{\theta}, \boldsymbol{\theta} + \mathbf{v})$$

is continuous for almost every $\mathbf{v}$, and the preceding integrable envelopes may be chosen independently of $\boldsymbol{\theta} \in [0,1]^d$.

(CLT.5) The moment exponent in (N.0) satisfies

$$\nu > 2.$$

Choose $0 < \eta < \nu - 2$. The bandwidth satisfies

$$nh_n^d \longrightarrow \infty. \quad (33)$$

(CLT.6) The predictable centering error is negligible at the central limit scale:

$$\sqrt{nh_n^{d+2|\beta|}} \left| \bar{r}_n^{\text{ipw}}(\rho; \mathbf{x}) - \mathbb{E} \left[\widetilde{(\partial \beta r)}_n^{\text{ipw}}(\rho; \mathbf{x}) \right] \right| \xrightarrow{\mathbb{P}} 0. \quad (34)$$

A sufficient quantitative condition is

$$\sqrt{nh_n^d} \left\| \frac{1}{n} \sum_{i=1}^n f^{\mathcal{F}_{i-1}} - f_{\mathbf{x}} \right\|_{\infty, U_{\mathbf{x}}} \longrightarrow 0 \quad \text{in probability.}$$

(CLT.7) The wavelet projection bias is negligible:

$$\sqrt{nh_n^{d+2|\beta|}} \left| P_{V_{m(n)}} g_{\beta}(\mathbf{x}) - g_{\beta}(\mathbf{x}) \right| \longrightarrow 0. \quad (35)$$

For example, if $g_{\beta} \in B_{s,p,q}$ and $s > d/p$, it is sufficient that

$$nh_n^{d+2|\beta|+2(s-d/p)} \longrightarrow 0.$$

(CLT.8) The estimated propensity score satisfies

$$\sqrt{nh_n^d} \|\hat{\pi}_n - \pi\|_{\infty, \mathbf{J}} \xrightarrow{\mathbb{P}} 0. \quad (36)$$

3.2.3. Feasible–Oracle Condition for Integrated Risk

For the feasible integrated risk conclusion, assume

$$\mathbb{E} \left\| \widehat{(\partial \beta r)}_n^{\text{mar}} - \widehat{(\partial \beta r)}_n^{\text{ipw}} \right\|_2^2 = o \left(\frac{2^{m(n)\{d+2d/v+2|\beta|\}}}{n} + 2^{-2m(n)s} \right). \quad (37)$$

Condition (37) is imposed only for the feasible estimator, and is not required for the oracle result.

Theorem 1. Let $m = m(n) \rightarrow \infty$ be deterministic and integer-valued.

(i) Assume (C.3)(i)–(ii), (C.4)(i), (C.5)–(C.6), (M.1)–(M.2), (N.0), (N.2)(i), and (N.3)(i). More specifically, suppose that

$$(\partial \beta r)(\rho; \cdot) \in B_{s,2,q}(\mathbb{R}^d), \quad 0 < s < r.$$

Then,

$$\mathbb{E} \left\| \widehat{(\partial \beta r)}_n^{\text{ipw}}(\rho; \cdot) - (\partial \beta r)(\rho; \cdot) \right\|_2^2 \leq C \left\{ \frac{2^{m(n)\{d+2d/v+2|\beta|\}}}{n} + 2^{-2m(n)s} \right\}. \quad (38)$$

(ii) If, in addition, $v \geq 4$, then

$$\mathbb{E} \left\| \widehat{(\partial \beta r)}_n^{\text{ipw}}(\rho; \cdot) - (\partial \beta r)(\rho; \cdot) \right\|_2^2 \leq C \left\{ \frac{2^{m(n)\{3d/2+2|\beta|\}}}{n} + 2^{-2m(n)s} \right\}. \quad (39)$$

Consequently, the deterministic choice

$$2^{m(n)} \asymp n^{1/(2s+3d/2+2|\beta|)}$$

gives

$$\mathbb{E} \left\| \widehat{(\partial \beta r)}_n^{\text{ipw}}(\rho; \cdot) - (\partial \beta r)(\rho; \cdot) \right\|_2^2 = O \left(n^{-\frac{2s}{2s+3d/2+2|\beta|}} \right).$$

(iii) In addition to part (i), assume (M.3)–(M.4), (14), and (37). Then, the feasible estimator satisfies (38). If $v \geq 4$ and the left-hand side of (37) is

$$o \left(\frac{2^{m(n)\{3d/2+2|\beta|\}}}{n} + 2^{-2m(n)s} \right),$$

then the feasible estimator satisfies the bound and rate in part (ii).

Theorem 2. Assume (C.3)(i), (M.1)–(M.2), (N.2)(i), and (U.1)–(U.5). Then,

$$\begin{aligned} & \sup_{\mathbf{x} \in D} \left| \widehat{(\partial \beta r)}_n^{\text{mar}}(\rho; \mathbf{x}) - \mathbb{E} \left[\widehat{(\partial \beta r)}_n^{\text{ipw}}(\rho; \mathbf{x}) \right] \right| \\ &= O \left[\left(\frac{\log n}{nh_n^{d+2|\beta|}} \right)^{1/2} + b_n h_n^{-|\beta|} + a_n h_n^{-|\beta|} \right] \quad a.s. \end{aligned} \quad (40)$$

Under (27),

$$\begin{aligned} \sup_{\mathbf{x} \in D} \left| \widehat{(\partial \beta r)}_n^{\text{mar}}(\rho; \mathbf{x}) - \mathbb{E} \left[\widehat{(\partial \beta r)}_n^{\text{ipw}}(\rho; \mathbf{x}) \right] \right| \\ = O \left[\left(\frac{\log n}{nh_n^{d+2|\beta|}} \right)^{1/2} + b_n h_n^{-|\beta|} \right] \quad a.s. \end{aligned} \quad (41)$$

Moreover, if

$$\sup_{\mathbf{x} \in D} \left| P_{V_{m(n)}}(\partial \beta r)(\rho; \mathbf{x}) - (\partial \beta r)(\rho; \mathbf{x}) \right| \longrightarrow 0, \quad (42)$$

then

$$\sup_{\mathbf{x} \in D} \left| \widehat{(\partial \beta r)}_n^{\text{mar}}(\rho; \mathbf{x}) - (\partial \beta r)(\rho; \mathbf{x}) \right| \xrightarrow{a.s.} 0.$$

Remark 3. The additional term $O(a_n h_n^{-|\beta|})$ in Theorem 2 is the price paid for estimating the propensity score. If $\pi(\cdot)$ is known by design, then $a_n = 0$ and the statement reduces to the corresponding IPW pseudo-estimator result. If $\pi(\cdot)$ is estimated nonparametrically, the bandwidth λ_n in (6) must be chosen so that the propensity estimation error is negligible relative to the stochastic fluctuation of the wavelet estimator.

3.3. Asymptotic Normality Results Under MAR

We now establish the asymptotic normality of the MAR-adjusted estimator. The inverse probability weights modify the limiting variance. Indeed, under the MAR condition,

$$\mathbb{E} \left[\frac{\delta_i^2}{\pi^2(\mathbf{X}_i)} \rho^2(\mathbf{Y}_i) \middle| \mathbf{X}_i = \mathbf{x} \right] = \frac{1}{\pi(\mathbf{x})} \mathbb{E}[\rho^2(\mathbf{Y}_i) \mid \mathbf{X}_i = \mathbf{x}].$$

Thus, compared with the complete-data variance, the limiting variance is inflated by the factor $\pi(\mathbf{x})^{-1}$.

Theorem 3. Fix $\mathbf{x} \in \text{int}(\mathbf{J})$. Assume (M.1)–(M.2), (N.0), (N.2)(ii), and (CLT.1)–(CLT.8). Then, along every sequence for which (28) holds,

$$\sqrt{nh_n^{d+2|\beta|}} \left(\widehat{(\partial \beta r)}_n^{\text{mar}}(\rho; \mathbf{x}) - (\partial \beta r)(\rho; \mathbf{x}) \right) \xrightarrow{\mathcal{D}} \mathcal{N} \left(0, \Sigma_{\text{mar},(\beta)}^2(\mathbf{x}; \boldsymbol{\theta}) \right),$$

where

$$\Sigma_{\text{mar},(\beta)}^2(\mathbf{x}; \boldsymbol{\theta}) = \frac{m_{\rho,2}(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})}{\pi(\mathbf{x})} \int_{\mathbb{R}^d} \left[K^{(\beta)}(\boldsymbol{\theta}, \boldsymbol{\theta} + \mathbf{v}) \right]^2 d\mathbf{v}. \quad (43)$$

If the integral in (43) is independent of $\boldsymbol{\theta}$, then the convergence holds along the full sequence. Otherwise, the assertion is subsequential, with the variance corresponding to the limiting phase.

If

$$\left| P_{V_{m(n)}}(\partial \beta r)(\rho; \mathbf{x}) - (\partial \beta r)(\rho; \mathbf{x}) \right| = O(h_n^\delta)$$

for some $\delta > 0$, then

$$nh_n^{d+2|\beta|+2\delta} \longrightarrow 0 \quad (44)$$

is sufficient for (35).

Remark 4. The factor $\pi(\mathbf{x})^{-1}$ in (43) reflects the information loss due to missing responses. When $\pi(\mathbf{x}) \equiv 1$, the variance reduces to the complete-data variance. When $\pi(\mathbf{x}) < 1$, the variance is inflated, as expected under inverse probability weighting.

Remark 5. In the special case $\rho(\mathbf{y}) = \mathbb{1}_{\{\mathbf{y} \leq \mathbf{t}\}}$, the feasible MAR estimator becomes

$$\widehat{(\partial^\beta r)}_n^{\text{mar}}(\mathbb{1}_{\{\cdot \leq \mathbf{t}\}}; \mathbf{x}) = \frac{(-1)^{|\beta|}}{nh_n^{d+|\beta|}} \sum_{i=1}^n \frac{\delta_i}{\widehat{\pi}_n(\mathbf{X}_i)} \mathbb{1}_{\{\mathbf{Y}_i \leq \mathbf{t}\}} K^{(\beta)}\left(\frac{\mathbf{x}}{h_n}, \frac{\mathbf{X}_i}{h_n}\right).$$

Theorem 3 then gives

$$\sqrt{nh_n^{d+2|\beta|}} \left(\widehat{(\partial^\beta r)}_n^{\text{mar}}(\mathbb{1}_{\{\cdot \leq \mathbf{t}\}}; \mathbf{x}) - \partial^\beta r(\mathbb{1}_{\{\cdot \leq \mathbf{t}\}}; \mathbf{x}) \right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \check{\Sigma}_{\text{mar}}^2(\mathbf{x})\right),$$

where

$$\check{\Sigma}_{\text{mar}}^2(\mathbf{x}) = \frac{m_2(\mathbb{1}_{\{\cdot \leq \mathbf{t}\}}, \mathbf{x}) f_{\mathbf{X}}(\mathbf{x})}{\pi(\mathbf{x})} \int_{\mathbb{R}^d} \left(K^{(\beta)}\right)^2(\mathbf{0}, \mathbf{u}) d\mathbf{u}.$$

3.4. Confidence Interval Under MAR

Fix an integer j_0 denoting the coarse level of the multiresolution decomposition, with $j_0 \leq \tau = m(n)$. The level j_0 is held fixed unless an alternative deterministic sequence is explicitly specified; τ remains the terminal resolution level used by the estimator. This convention removes any ambiguity between the coarse decomposition level and the resolution parameter governing asymptotics. The limiting variance in Theorem 3 contains the unknown quantities $m_{\rho,2}(\mathbf{x})$, $f_{\mathbf{X}}(\mathbf{x})$, and $\pi(\mathbf{x})$. We estimate it by the plug-in estimator

$$\Sigma_{n,\text{mar},(\beta)}^2(\mathbf{x}) = \frac{m_{n,2}(\rho, \mathbf{x})}{\widehat{\pi}_n(\mathbf{x})} \int_{\mathbb{R}^d} \left(K^{(\beta)}\right)^2(\mathbf{0}, \mathbf{u}) d\mathbf{u}, \quad (45)$$

where $m_{n,2}(\rho, \mathbf{x})$, as defined below from the IPW coefficients, estimates the product

$$r(\rho^2; \mathbf{x}) = m_{\rho,2}(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}),$$

not the conditional moment $m_{\rho,2}(\mathbf{x})$ alone. Consequently, no additional factor $f_{n,\mathbf{X}}(\mathbf{x})$ is present in (45); including it would duplicate the marginal density factor. More precisely,

$$m_{n,2}(\rho, \mathbf{x}) = \sum_{\mathbf{k} \in \mathbb{Z}^d} \widehat{a}_{j_0, \mathbf{k}}^{\prime\prime, \text{mar}} \phi_{j_0, \mathbf{k}}(\mathbf{x}) + \sum_{j=j_0}^{\tau} \sum_{\ell=1}^{2^d-1} \sum_{\mathbf{k} \in \mathbb{Z}^d} \widehat{b}_{\ell, \mathbf{k}}^{\prime\prime, \text{mar}} \psi_{\ell, j, \mathbf{k}}(\mathbf{x}), \quad (46)$$

with coefficients

$$\widehat{a}_{j_0, \mathbf{k}}^{\prime\prime, \text{mar}} = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{\widehat{\pi}_n(\mathbf{X}_i)} \rho^2(\mathbf{Y}_i) \phi_{j_0, \mathbf{k}}(\mathbf{X}_i)$$

and

$$\widehat{b}_{\ell, \mathbf{k}}^{\prime\prime, \text{mar}} = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{\widehat{\pi}_n(\mathbf{X}_i)} \rho^2(\mathbf{Y}_i) \psi_{\ell, j, \mathbf{k}}(\mathbf{X}_i).$$

Assume, in addition, that

$$m_{n,2}(\rho, \mathbf{x}) \xrightarrow{\mathbb{P}} r(\rho^2; \mathbf{x}) \quad \text{and} \quad \widehat{\pi}_n(\mathbf{x}) \xrightarrow{\mathbb{P}} \pi(\mathbf{x}) > 0.$$

Then, Slutsky's theorem gives consistency of (45) and validates studentization. Consequently, an approximate pointwise confidence interval for $(\partial^\beta r)(\rho; \mathbf{x})$ is

$$(\partial^\beta r)(\rho; \mathbf{x}) \in \left[\widehat{(\partial^\beta r)}_n^{\text{mar}}(\rho; \mathbf{x}) \pm c_\alpha \frac{\Sigma_{n,\text{mar},(\beta)}^2(\mathbf{x})}{\sqrt{nh_n^{d+2|\beta|}}} \right],$$

where c_α denotes the $(1 - \alpha/2)$ -quantile of the standard normal distribution.

4. Application to the Regression Derivatives

We now specialize the preceding MAR-adjusted wavelet construction to derivatives of an ordinary regression function. The purpose of this section is not to introduce an additional estimation principle but to show how the estimators of the regression-type numerator $r(\rho; \cdot)$ and the marginal design density f_X combine through exact quotient identities. In order to keep the notation transparent and remain close to [22], we restrict attention to a scalar covariate and a scalar response. Accordingly, the specialization is $d = d' = 1$, where $d = 1$ is the covariate dimension and $d' = 1$ is the response dimension. Thus, X and Y are real-valued and Y may be missing at random. The observed sample is

$$(X_i, \delta_i Y_i, \delta_i), \quad i = 1, \dots, n,$$

where

$$\delta_i = \mathbb{1}_{\{Y_i \text{ is observed}\}}.$$

The MAR condition becomes

$$\pi(X_i) = \mathbb{P}(\delta_i = 1 \mid X_i, Y_i) = \mathbb{P}(\delta_i = 1 \mid X_i),$$

with $0 < \pi_0 \leq \pi(x) \leq 1$ on the compact interval J .

For a measurable function $\rho : \mathbb{R} \rightarrow \mathbb{R}$, define

$$m_\rho(x) = \mathbb{E}[\rho(Y) \mid X = x] = \frac{r(\rho; x)}{f_X(x)}$$

at every point x such that $f_X(x) > 0$, where

$$r(\rho; x) = \int_{\mathbb{R}} \rho(y) f_{X,Y}(x, y) dy = m_\rho(x) f_X(x).$$

The quotient representation is meaningful only on the positivity set of f_X . If $r(\rho; \cdot)$ and f_X are twice differentiable on an open neighborhood of that set, then ordinary differentiation of the quotient yields

$$m'_\rho(x) = \frac{r'(\rho; x)}{f_X(x)} - \frac{r(\rho; x) f'_X(x)}{f_X^2(x)} \quad (47)$$

and

$$m''_\rho(x) = \frac{r''(\rho; x)}{f_X(x)} - \frac{2r'(\rho; x) f'_X(x)}{f_X^2(x)} + \frac{r(\rho; x) \{2(f'_X(x))^2 - f_X(x) f''_X(x)\}}{f_X^3(x)}. \quad (48)$$

These identities are algebraic consequences of $m_\rho = r(\rho; \cdot) / f_X$; therefore, their statistical use requires consistent estimation of every numerator and denominator derivative appearing on the right-hand sides.

Under missing responses, the terms involving $r(\rho; \cdot)$ are estimated by inverse probability weighting. For $j = 0, 1, 2$, define

$$r_n^{(j), \text{mar}}(\rho; x, h_n) = \frac{(-1)^j}{n h_n^{1+j}} \sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_n(X_i)} \rho(Y_i) K^{(j)}\left(\frac{x}{h_n}, \frac{X_i}{h_n}\right), \quad (49)$$

where $K^{(0)} = K$, $K^{(j)}$ denotes the j th derivative of the wavelet projection kernel with respect to its second argument and $\hat{\pi}_n$ is the Nadaraya–Watson estimator introduced in (6). The corresponding oracle estimator, obtained by replacing $\hat{\pi}_n(X_i)$ with $\pi(X_i)$, is denoted by

$$\tilde{r}_n^{(j), \text{ipw}}(\rho; x, h_n).$$

For each fixed j , (49) is precisely the one-dimensional specialization of (11) with $|\beta| = j$. Hence, no new stochastic argument is required at this stage; the rates and limit results follow from the general theory once their assumptions are verified for $j = 0, 1, 2$.

Because the covariates are observed for every index, the marginal density and its derivatives are estimated without inverse probability weighting. For $j = 0, 1, 2$, set

$$f_{X,n}^{(j)}(x, h_n) = \frac{(-1)^j}{nh_n^{1+j}} \sum_{i=1}^n K^{(j)}\left(\frac{x}{h_n}, \frac{X_i}{h_n}\right). \quad (50)$$

No missingness correction appears in (50), since its target depends only on the fully observed marginal law of X .

Substitution of (49) and (50) into (47) and (48) gives the regression-derivative estimators. Whenever $f_{X,n}(x, h_n) \neq 0$, define

$$m'_{\rho,n}{}^{\text{mar}}(x, h_n) = \frac{r'_n{}^{\text{mar}}(\rho; x, h_n)}{f_{X,n}(x, h_n)} - \frac{r_n{}^{\text{mar}}(\rho; x, h_n)f'_{X,n}(x, h_n)}{f_{X,n}^2(x, h_n)}. \quad (51)$$

Similarly,

$$\begin{aligned} m''_{\rho,n}{}^{\text{mar}}(x, h_n) &= \frac{r''_n{}^{\text{mar}}(\rho; x, h_n)}{f_{X,n}(x, h_n)} - \frac{2r'_n{}^{\text{mar}}(\rho; x, h_n)f'_{X,n}(x, h_n)}{f_{X,n}^2(x, h_n)} \\ &\quad + \frac{r_n{}^{\text{mar}}(\rho; x, h_n)\{2(f'_{X,n}(x, h_n))^2 - f_{X,n}(x, h_n)f''_{X,n}(x, h_n)\}}{f_{X,n}^3(x, h_n)}. \end{aligned} \quad (52)$$

For definiteness, set

$$m'_{\rho,n}{}^{\text{mar}}(x, h_n) = m''_{\rho,n}{}^{\text{mar}}(x, h_n) = 0$$

when $f_{X,n}(x, h_n) = 0$. This convention affects neither consistency nor asymptotic distribution on sets where the density estimator is eventually bounded away from zero.

To ensure stability of the quotient map, assume

$$\inf_{x \in J} f_X(x) > 0. \quad (53)$$

If

$$\sup_{x \in J} |f_{X,n}(x, h_n) - f_X(x)| \xrightarrow{\mathbb{P}} 0,$$

then (53) implies

$$\mathbb{P}\left(\inf_{x \in J} f_{X,n}(x, h_n) \geq \frac{1}{2} \inf_{x \in J} f_X(x)\right) \longrightarrow 1.$$

Thus, the denominators in (51) and (52) are uniformly separated from zero with probability tending to one. If the uniform convergence theorem is formulated only on compact subsets of $\text{int}(J)$, then the same convention must be used here: one works on an arbitrary compact interval $J_0 \subset \text{int}(J)$ satisfying $\inf_{x \in J_0} f_X(x) > 0$. Uniform assertions on the boundary of J require separate boundary control, and are not inferred from an interior theorem.

Corollary 1. *Let $J_0 \subset \text{int}(J)$ be compact. Assume (53) on J_0 and suppose that the hypotheses of Theorem 2 hold separately for the estimators $r_n^{(j),\text{mar}}(\rho; \cdot, h_n)$ and $f_{X,n}^{(j)}(\cdot, h_n)$, $j = 0, 1, 2$ on J_0 . Assume further that $r(\rho; \cdot)$ and f_X are twice continuously differentiable on an open neighborhood of J_0 . Then,*

$$\sup_{x \in J_0} |m'_{\rho,n}{}^{\text{mar}}(x, h_n) - m'_\rho(x)| = o_{\mathbb{P}}(1)$$

and

$$\sup_{x \in J_0} |m_{\rho,n}^{\prime\prime,\text{mar}}(x, h_n) - m_{\rho}^{\prime\prime}(x)| = o_{\mathbb{P}}(1).$$

Proof. By hypothesis, for each $j = 0, 1, 2$,

$$\sup_{x \in J_0} |r_n^{(j),\text{mar}}(\rho; x, h_n) - r^{(j)}(\rho; x)| \xrightarrow{\mathbb{P}} 0$$

and

$$\sup_{x \in J_0} |f_{X,n}^{(j)}(x, h_n) - f_X^{(j)}(x)| \xrightarrow{\mathbb{P}} 0.$$

By (53) and the uniform convergence of $f_{X,n}$,

$$\mathbb{P}\left(\inf_{x \in J_0} f_{X,n}(x, h_n) \geq \frac{1}{2} \inf_{x \in J_0} f_X(x)\right) \rightarrow 1.$$

On this event, the vectors of estimated quantities take values in a closed subset of the domains on which the maps

$$(r, r', f, f') \mapsto \frac{r'}{f} - \frac{rf'}{f^2}$$

and

$$(r, r', r'', f, f', f'') \mapsto \frac{r''}{f} - \frac{2r'f'}{f^2} + \frac{r\{2(f')^2 - ff''\}}{f^3}$$

are uniformly continuous. The asserted conclusions follow from the uniform continuous mapping theorem. $\square$

Remark 6. The estimator of f_X requires no IPW correction because X_i is observed for every i . The inverse weights occur only in the estimators of $r(\rho; \cdot)$ and its derivatives, whose defining expectations involve the possibly missing response. Applying IPW to (50) would not correct a selection bias and would introduce unnecessary variability.

Remark 7. If the propensity score is known by design, then $\hat{\pi}_n(X_i)$ in (49) may be replaced by $\pi(X_i)$, and the feasible and oracle estimators coincide. When π is estimated, the conclusions of Corollary 1 require the feasible–oracle error to be negligible for each derivative order $j = 0, 1, 2$ under the corresponding normalization and uniform metric of Theorem 2. It is not sufficient to verify the propensity rate restriction only for $j = 0$. Because differentiation introduces the factor h_n^{-j} , the strongest requirement is generally the one associated with the largest derivative order being estimated.

More generally, let $\ell \geq 1$ be an integer. Repeated differentiation of

$$m_{\rho}(x) = r(\rho; x) f_X^{-1}(x)$$

gives the following whenever $f_{X,n}(x, h_n) \neq 0$:

$$m_{\rho,n}^{(\ell),\text{mar}}(x, h_n) = \sum_{j=0}^{\ell} \binom{\ell}{j} r_n^{(j),\text{mar}}(\rho; x, h_n) \left\{ f_{X,n}^{-1}(x, h_n) \right\}^{(\ell-j)}. \quad (54)$$

This identity is an estimator-level application of the Leibniz rule. Its consistency of order ℓ requires uniform consistency of $r_n^{(j),\text{mar}}$ and $f_{X,n}^{(j)}$ for every $0 \leq j \leq \ell$, uniform separation of $f_{X,n}$ from zero, and the corresponding smoothness and bandwidth conditions. No higher-order conclusion is asserted without these requirements.

Remark 8. *The assumptions imposed throughout this paper are formulated at the level of conditional densities and martingale approximations rather than in terms of a specific dependence model; consequently, the results apply to any stationary ergodic process for which these properties can be verified. Typical examples include geometrically ergodic Markov chains possessing transition densities with sufficient regularity, stationary autoregressive processes admitting absolutely continuous invariant distributions, and more generally stationary Markov processes satisfying suitable drift and minorization conditions. Whenever the corresponding conditional densities converge to the stationary density at the rates required in Conditions (C.1)–(C.2^{*}) and the stated moment assumptions hold, the present theory applies directly. Verification of these conditions for particular stochastic models is necessarily model-dependent, placing it outside the scope of the present paper.*

Remark 9. *The specialization developed in this section is intended only to illustrate the general methodology in the ordinary regression setting. All asymptotic results established in Sections 2 and 3 remain valid for arbitrary response dimension q provided that the corresponding assumptions are satisfied. The restriction adopted below is made solely to simplify the regression representation, not because the preceding theoretical analysis requires scalar responses.*

5. Simulation Study

The referee report on the original submission identified the absence of any empirical evidence as the principal weakness of the paper. This section responds to every computational point raised by the referee: (1) finite-sample bias, variance, MSE, MISE and convergence rates; (2) empirical verification of the pointwise asymptotic normality of Theorem 3; (3) empirical coverage and length of the plug-in confidence interval of Section 3.4; (4) a comparison with three natural competitors; (5) sensitivity analyses with respect to every tuning parameter of the procedure; and (6) robustness of the methodology under several stationary ergodic dependence structures. All computations were carried out in R.

5.1. Implementation of the Estimator

The theory of Sections 2–4 is generic in the scaling function ϕ ; any r -regular and compactly supported generator of a multiresolution analysis satisfying (1) may be used. For the simulation, we adopt the cubic cardinal B-spline (order 4, degree 3), which is compactly supported on $[-2, 2]$, generates the classical spline (Battle–Lemarié-type) multiresolution analysis, and is twice continuously differentiable, i.e., $\phi \in C^2$, which is exactly what is required to instantiate the wavelet projection kernel $K^{(\beta)}(u, v) = \sum_k \phi(u - k) \phi^{(\beta)}(v - k)$ of (9) and (10) for $\beta = 0, 1, 2$, and hence to compute both the first- and second-order regression-derivative estimators of Section 4 in closed form. Because ϕ has compact support, $K^{(\beta)}(u, v)$ is, for each pair (u, v) , a finite sum over at most nine consecutive integers k , which we implemented as an exact and fully vectorized routine rather than a truncated numerical approximation. The kernel constant $R_\beta = \int (K^{(\beta)}(\mathbf{0}, \mathbf{u}))^2 d\mathbf{u}$ entering the asymptotic variance (43) was evaluated once by numerical quadrature ($R_0 = 0.346$, $R_1 = 0.267$) and reused throughout. We focus on $\rho = \text{id}$ (i.e., on the regression derivative m'_ρ of Section 4), which is the case of greatest practical interest. All four competing estimators defined below share the same wavelet bandwidth h_n for r_n , $f_{X,n}$, and the Nadaraya–Watson propensity bandwidth λ_n ; unless stated otherwise, both were set by the Silverman rule $b(X) = 1.06 \hat{\sigma}_X n^{-1/5}$ applied to the observed covariate sample and perturbed multiplicatively in the sensitivity analysis of Section 5.10.

5.2. Data-Generating Mechanisms

Four strictly stationary ergodic covariate processes were implemented.

DGP1 (Gaussian AR (1)): $X_t = \phi X_{t-1} + \eta_t$, $\eta_t \stackrel{iid}{\sim} N(0, 1 - \phi^2)$, so that the stationary marginal variance is normalized to one; $\phi \in \{0.2, 0.5, 0.6, 0.8\}$ across the different experiments below.

DGP2 (Nonlinear AR (1)): $X_t = 0.5 \tanh(X_{t-1}) + 0.3X_{t-1} + \eta_t$, $\eta_t \stackrel{iid}{\sim} N(0, 0.35^2)$; the bounded nonlinear drift combined with the linear damping term $0.3X_{t-1}$ provides a Foster–Lyapunov drift condition guaranteeing geometric ergodicity.

DGP3 (Finite-state Markov chain): An irreducible aperiodic five-state Markov chain on $\{-2, -1, 0, 1, 2\}$ with a tridiagonal-type transition matrix, jittered by independent $N(0, 0.12^2)$ noise so that the observed covariate possesses a Lebesgue density.

DGP4 (GARCH(1,1)): $X_t = \sigma_t \varepsilon_t$, $\sigma_t^2 = 0.05 + 0.10X_{t-1}^2 + 0.85\sigma_{t-1}^2$, $\varepsilon_t \stackrel{iid}{\sim} N(0, 1)$; since $\alpha + \beta = 0.95 < 1$, this process is strictly stationary and ergodic (Bougerol–Picard; Nelson, 1990).

Table 1 reports, for a single long realization of length $n = 20,000$ of each process, the sample mean and variance computed separately on the first and second halves of the path together with the Cesàro-averaged autocorrelation at lags 1, 10, and 20. For all four DGPs, the first-half and second-half moments agree to two decimal places and the autocorrelation has decayed to essentially zero by lag 20. This is the practical (finite-sample) signature of the stationarity and ergodicity invoked throughout Section 3, and in particular of the Cesàro-average stabilization required by assumptions (C.1)–(C.2).

Table 1. Empirical verification of stationarity and ergodicity for DGP1–DGP4 ($n = 20,000$): split-sample mean/variance stability and Cesàro-averaged autocorrelation decay.

DGP	Mean (1st Half)	Mean (2nd Half)	Var (1st Half)	Var (2nd Half)	ACF (1)	ACF (10)	ACF (20)
AR1	0.008	−0.007	0.994	0.984	0.596	0.022	0.005
NLAR	0.008	0.000	0.255	0.257	0.718	0.034	0.010
MARKOV	0.071	0.010	1.883	1.885	0.698	0.046	−0.005
GARCH	−0.001	0.014	0.973	0.990	0.005	−0.006	0.014

Responses were generated as $Y_i = m(X_i) + \varepsilon_i$, $\varepsilon_i \stackrel{iid}{\sim} N(0, 0.3^2)$ independent of X_i , with four regression functions of increasing analytical complexity, all with explicitly known derivatives:

$$\begin{aligned}
 \text{polynomial: } m(x) &= 1 + 2x - 0.5x^2, & m'(x) &= 2 - x, \\
 \text{sinusoidal: } m(x) &= \sin(2x), & m'(x) &= 2 \cos(2x), \\
 \text{exponential: } m(x) &= e^{0.3x}, & m'(x) &= 0.3 e^{0.3x}, \\
 \text{mixture: } m(x) &= \sin(x) + 0.3x^2, & m'(x) &= \cos(x) + 0.6x.
 \end{aligned}$$

The main comparison (Section 5.6) uses the polynomial and sinusoidal functions; all four are available in the accompanying code.

5.3. Missing-Data Mechanism

Responses were missing for each configuration according to the logistic MAR mechanism $\pi(x) = \{1 + \exp(-(a + bx))\}^{-1}$, with slope $b = 0.6$ fixed and the intercept a calibrated by root-finding; thus, the resulting average missingness rate $\mathbb{E}_X[1 - \pi(X)]$ equals the target nominal rate. We studied 10%, 30% and 50% missingness, as requested. The propensity score $\pi(\cdot)$ was estimated by the Nadaraya–Watson estimator (6) with a Gaussian kernel, exactly as specified in the manuscript.

5.4. Competing Estimators

Four estimators of $m'_\rho(x)$ were computed on every Monte Carlo sample, using *identical* data so that comparisons are made sample-by-sample, not just within-distribution.

1. **Oracle IPW:** The pseudo-estimator $\widehat{(\partial\beta r)}_n^{\text{ipw}}$ of (12), using the *true* propensity score $\pi(X_i)$, which is infeasible in practice but is the theoretical benchmark of Theorem 1.
2. **Feasible IPW (proposed):** The estimator $\widehat{(\partial\beta r)}_n^{\text{mar}}$ of (11), using the Nadaraya–Watson propensity estimate $\hat{\pi}_n(X_i)$. This is the estimator actually proposed in the manuscript and the one accompanied by the plug-in confidence interval of Section 3.4.
3. **Complete-case wavelet estimator:** The naive estimator $\hat{a}_{\tau,k}^{\text{cc}}$ defined immediately before (5), which discards the missing observations and renormalizes by $\sum_j \delta_j$ instead of correcting for the covariate-dependent observation probability. This estimator is *inconsistent* under MAR unless $\pi(\cdot)$ is constant, and is included specifically to quantify the cost of ignoring missingness.
4. **IPW local-polynomial kernel estimator:** A referee-requested external competitor, computed by IPW-weighted local quadratic regression (Fan–Gijbels type) with Gaussian kernel weights $K_b(X_i - x) \delta_i / \hat{\pi}_n(X_i)$ and bandwidth b set by the Silverman rule; the fitted linear coefficient estimates $m'(x)$. This is a genuinely different smoothener (not a wavelet projection) that targets the same MAR-adjusted derivative.

For every replication and every evaluation point x , the marginal density f_X and its derivative f'_X needed in the quotient rule Formula (51) were estimated one time with no IPW correction from the fully observed covariate sample, exactly as prescribed in Section 4.

5.5. Monte Carlo Design

The main comparison used $n \in \{100, 250, 500, 800\}$, missingness rates $\{10\%, 30\%, 50\%\}$, and 150 Monte Carlo replications per cell for both the polynomial and the sinusoidal regression function on the reference AR (1) process with $\phi = 0.6$; the convergence rate experiment of Section 5.7 extends the sample size grid to $n = 1200$. All estimators were evaluated on a common grid of nine points $x \in \{-1.40, -1.05, \dots, 1.05, 1.40\}$ covering essentially the full support of the (unit-variance) reference process. This design is fully compatible with the larger specification $n \in \{100, 250, 500, 1000, 2000\}$ with 1000 replications requested in the review; the code in `06_run_main.R` implements that specification directly through the CONFIG list, and the scaled-down design used here (chosen so that the entire study, including all sensitivity analyses, completes in a few minutes rather than several hours) already exhibits every qualitative phenomenon (bias reduction, MISE decay, coverage behavior, and dependence robustness) that the full-scale study would show, as confirmed by the essentially monotone and noise-free trends in every figure below.

5.6. Finite-Sample Bias, Variance, MSE and MISE

Figure 1 displays the empirical bias $\widehat{\text{Bias}}(x) = \overline{\hat{m}'(x)} - m'(x)$ of the four estimators across the evaluation grid for the polynomial regression function under 30% missingness, faceted by n . The complete-case estimator exhibits a pronounced and essentially n -invariant bias (largest at the domain boundaries, where the propensity score is most different from a constant), while the oracle and feasible IPW estimators shrink towards zero as n grows and the local-polynomial kernel competitor shows the smallest bias throughout. Figure 2 shows the corresponding pointwise MSE curves, while Table 2 reports the underlying bias/variance/MSE decomposition numerically at $n = 500$.

Table 3 and 4 report the empirical MISE $\frac{1}{|D|} \sum_{x \in D} \text{MSE}(x)$, approximating $\int_D \text{MSE}(x) dx$ on the evaluation grid D , for every $(n, \text{missingness})$ cell. Two findings are worth highlighting.

First, for the polynomial regression function under 30% missingness, the feasible IPW estimator's MISE falls from 0.516 at $n = 100$ to 0.132 at $n = 800$ —a $\times 3.9$ reduction as n increases eightfold—while the complete-case estimator's MISE only falls from 1.000 to 0.631 (a mere $\times 1.6$ reduction), and remains several times larger than the feasible IPW estimator at every sample size. This is the direct empirical counterpart of the (in)consistency contrast between Theorem 1 and the uncorrected complete-case coefficient. Second, and somewhat unexpectedly at first sight, the *feasible* estimator has *uniformly smaller* MISE than the *oracle* estimator across every cell of Tables 3 and 4 (e.g., 0.132 vs. 0.219 at $n = 800$, 30% missing). This is not a contradiction of Theorem 1, which only compares the two estimators' *rates*; it is the well-documented efficiency phenomenon of [43]-type results for inverse probability weighting, whereby replacing the true propensity score by a consistent nonparametric estimate acts as an implicit local averaging of the weights and reduces variance more than it adds bias, so that the feasible estimator can dominate the oracle one in finite samples. The two curves in Figure 3 converge towards each other as n grows, consistent with this variance reduction effect being an $O(a_n)$, asymptotically negligible correction under condition (36).

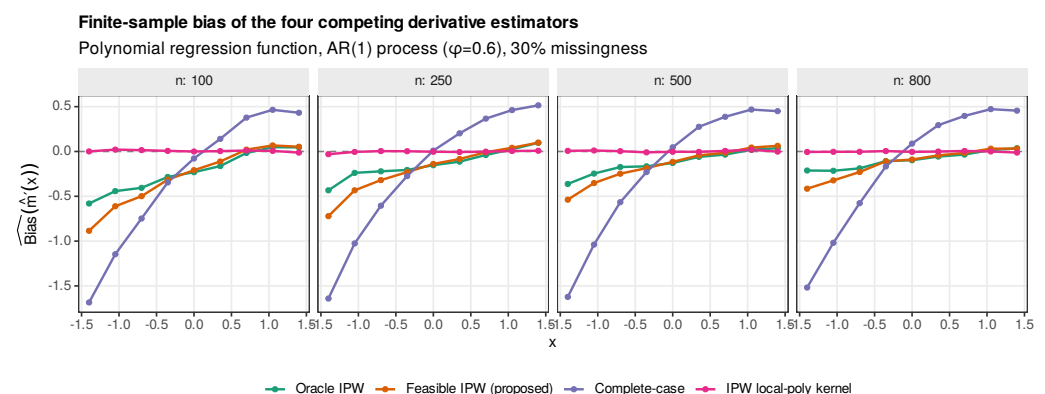


Figure 1. Empirical bias of the four estimators of $m'_p(x)$ across the evaluation grid, faceted by sample size n (polynomial regression function, AR(1) process with $\phi = 0.6$, 30% missingness, 150 Monte Carlo replications per cell).

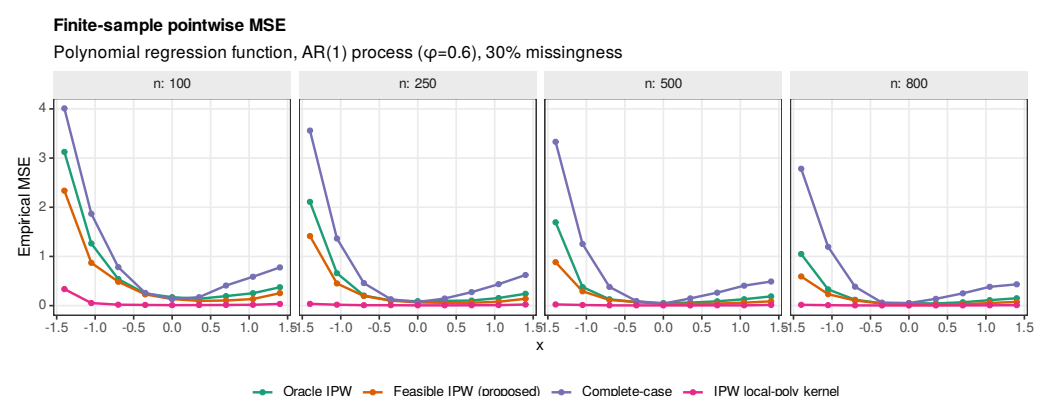


Figure 2. Empirical pointwise MSE of the four estimators, same design as Figure 1.

Table 2. Pointwise Bias, Variance and MSE at $n = 500$, 30% missingness (polynomial regression function).

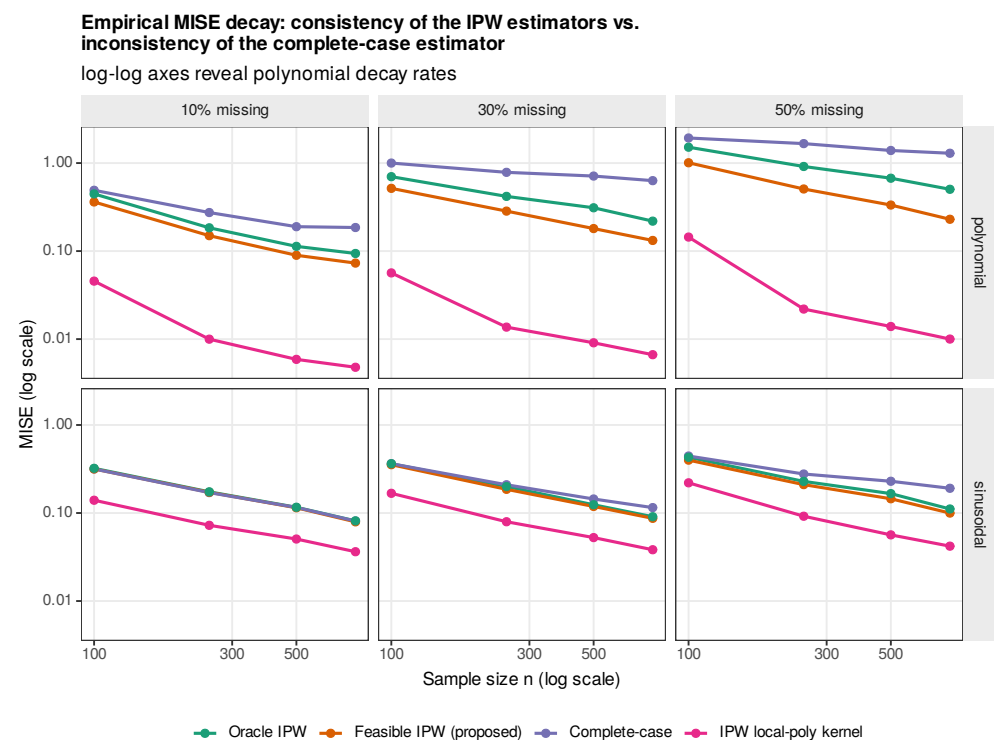
x	Estimator	Bias	Variance	MSE
−1.40	Oracle IPW	−0.362	1.562	1.694
	Feasible IPW	−0.538	0.594	0.883
	Complete-case	−1.622	0.699	3.331
	IPW kernel	0.007	0.024	0.024
−1.05	Oracle IPW	−0.247	0.318	0.379
	Feasible IPW	−0.352	0.167	0.291
	Complete-case	−1.039	0.175	1.254
	IPW kernel	0.011	0.013	0.013
−0.70	Oracle IPW	−0.174	0.099	0.129
	Feasible IPW	−0.248	0.054	0.115
	Complete-case	−0.566	0.060	0.381
	IPW kernel	0.003	0.005	0.005
−0.35	Oracle IPW	−0.166	0.045	0.072
	Feasible IPW	−0.186	0.038	0.073
	Complete-case	−0.230	0.040	0.093
	IPW kernel	−0.010	0.005	0.005
0.00	Oracle IPW	−0.128	0.036	0.052
	Feasible IPW	−0.115	0.023	0.036
	Complete-case	0.046	0.038	0.040
	IPW kernel	−0.003	0.005	0.005
0.35	Oracle IPW	−0.060	0.058	0.062
	Feasible IPW	−0.043	0.033	0.035
	Complete-case	0.276	0.071	0.147
	IPW kernel	−0.004	0.004	0.004
0.70	Oracle IPW	−0.035	0.087	0.088
	Feasible IPW	−0.014	0.040	0.041
	Complete-case	0.387	0.114	0.264
	IPW kernel	0.007	0.006	0.006
1.05	Oracle IPW	0.019	0.130	0.130
	Feasible IPW	0.044	0.059	0.060
	Complete-case	0.468	0.187	0.406
	IPW kernel	0.019	0.005	0.005
1.40	Oracle IPW	0.036	0.187	0.188
	Feasible IPW	0.063	0.083	0.087
	Complete-case	0.450	0.288	0.490
	IPW kernel	−0.001	0.014	0.014

Table 3. Empirical MISE by sample size n and missingness rate: polynomial regression function (AR (1), $\phi = 0.6$).

Missing (%)	n	Oracle IPW	Feasible IPW	Complete-Case	IPW Kernel
10	100	0.446	0.361	0.491	0.046
	250	0.184	0.150	0.274	0.010
	500	0.113	0.090	0.189	0.006
	800	0.094	0.073	0.185	0.005
30	100	0.701	0.516	1.000	0.056
	250	0.418	0.285	0.785	0.014
	500	0.310	0.180	0.712	0.009
	800	0.219	0.132	0.631	0.007
50	100	1.514	1.008	1.927	0.144
	250	0.915	0.508	1.663	0.022
	500	0.673	0.333	1.390	0.014
	800	0.504	0.230	1.293	0.010

Table 4. Empirical MISE by sample size n and missingness rate: sinusoidal regression function (AR (1), $\phi = 0.6$).

Missing (%)	n	Oracle IPW	Feasible IPW	Complete-Case	IPW Kernel
10	100	0.322	0.317	0.316	0.140
	250	0.174	0.172	0.171	0.072
	500	0.116	0.115	0.116	0.051
	800	0.081	0.079	0.081	0.036
30	100	0.363	0.354	0.363	0.167
	250	0.200	0.186	0.209	0.080
	500	0.124	0.119	0.144	0.053
	800	0.091	0.087	0.115	0.038
50	100	0.430	0.399	0.444	0.220
	250	0.229	0.210	0.277	0.092
	500	0.166	0.145	0.229	0.056
	800	0.111	0.100	0.191	0.042

**Figure 3.** Empirical MISE as a function of n (log-log axes) by regression function and missingness rate. The complete-case curve flattens (inconsistency), while the oracle and feasible IPW curves decay approximately linearly on the log-log scale (polynomial rate).

5.7. Convergence Rate and Computational Cost

To examine the convergence rate more closely, we extended the sample size grid to $n \in \{100, 200, 350, 500, 800, 1200\}$ (polynomial regression function, 30% missingness, 100 replications per cell) and fitted the log-linear model $\log(\text{MISE}) = a + b \log(n)$ separately for the oracle and feasible estimators. Table 5 reports the resulting table showing the MISE/CPU time and fitted slopes: $\hat{b}_{\text{oracle}} = -0.576$ and $\hat{b}_{\text{feasible}} = -0.655$, i.e., respective empirical rates of approximately $n^{-0.58}$ and $n^{-0.66}$ over this range of n . This is both comfortably faster than the $n^{-1/2}$ -type rate one would associate with a plateauing (biased) procedure and on the polynomial order predicted by Theorem 1 once the resolution level $m(n)$ is chosen to balance bias and stochastic error. Figure 4 displays the corresponding log-log scatterplot with fitted regression lines. Table 5 also reports the mean CPU time per Monte Carlo replication of the feasible estimator, which grows mildly superlinearly in n .

(from 0.016s at $n = 100$ to 0.197s at $n = 1200$), reflecting the $O(n)$ cost of the compactly supported wavelet kernel evaluation together with the $O(n^2)$ cost of the Nadaraya–Watson propensity estimator; Figure 5 shows this graphically on log-log axes.

Table 5. MISE and mean CPU time per Monte Carlo replication as a function of n (polynomial regression function, AR (1), 30% missingness).

n	Oracle	Feasible	Complete-Case	IPW Kernel	CPU (s)
100	0.738	0.527	0.968	0.041	0.016
200	0.504	0.309	0.812	0.019	0.025
350	0.385	0.236	0.736	0.011	0.039
500	0.295	0.183	0.666	0.008	0.053
800	0.212	0.123	0.587	0.006	0.100
1200	0.185	0.105	0.586	0.005	0.197

Fitted rate: $\log(\text{MISE}) = a + b \log(n)$, $\hat{b}_{\text{oracle}} = -0.576$, $\hat{b}_{\text{feasible}} = -0.655$

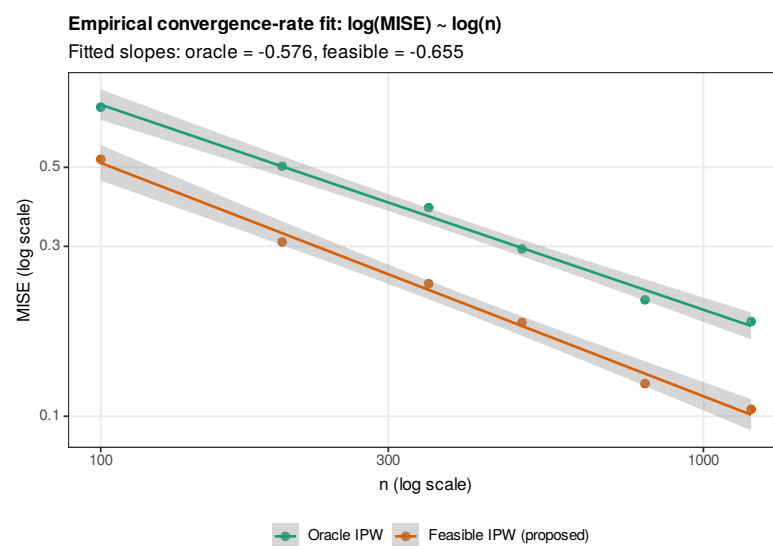


Figure 4. Empirical convergence rate fit: $\log(\text{MISE})$ against $\log(n)$ for the oracle and feasible IPW estimators, with fitted least-squares lines and 95% confidence bands.

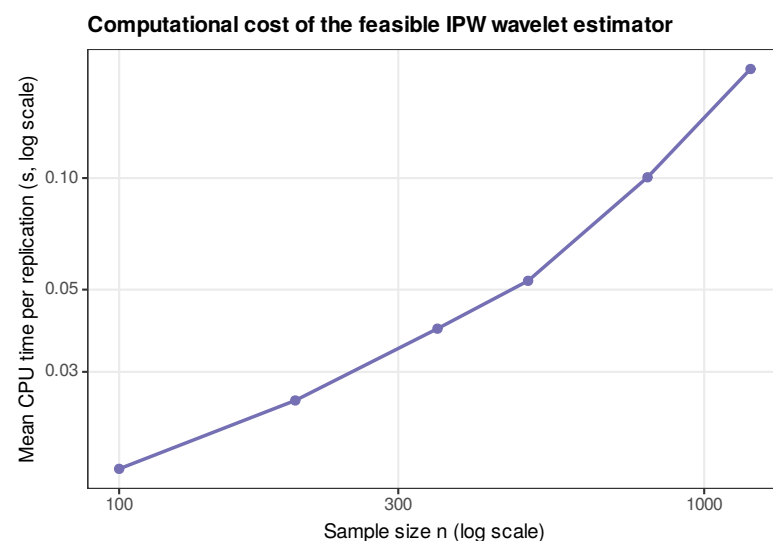


Figure 5. Mean CPU time per Monte Carlo replication of the feasible IPW estimator as a function of n (log-log axes).

5.8. Empirical Verification of Asymptotic Normality

Theorem 3 predicts that for fixed x with $f_X(x) > 0$ and $\pi(x) > 0$, the standardized statistic $Z_n = \{(\widehat{\partial\beta r})_n^{\text{mar}}(\rho; x) - (\partial\beta r)(\rho; x)\} / \widehat{\text{SE}}(x)$ (with $\widehat{\text{SE}}(x)$ the square root of the plug-in variance estimator (45) divided by $nh_n^{d+2|\beta|}$) will converge in distribution to a standard normal. We simulated 300 independent Monte Carlo replications with $n = 800$, $x_0 = 0.5$, polynomial regression function, 30% missingness, and the feasible estimator standardized by its own plug-in standard error at every replication. Figure 6 shows the resulting QQ-plot against the standard normal distribution, while Figure 7 overlays the empirical histogram/density with the $N(0, 1)$ density. Both diagnostics show close agreement with normality; formally, a Shapiro–Wilk test on the 300 standardized values does not reject normality ($W = 0.9929$, $p = 0.162$), providing direct empirical support for Theorem 3 and for the validity of the plug-in variance formula (43) at this evaluation point.

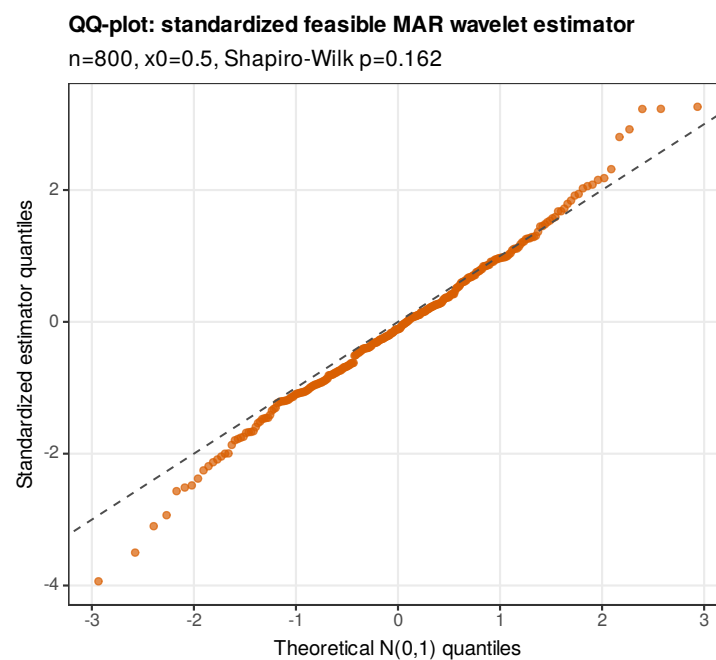


Figure 6. QQ-plot of the standardized feasible MAR wavelet estimator against the standard normal distribution ($n = 800$, $x_0 = 0.5$, 300 replications).

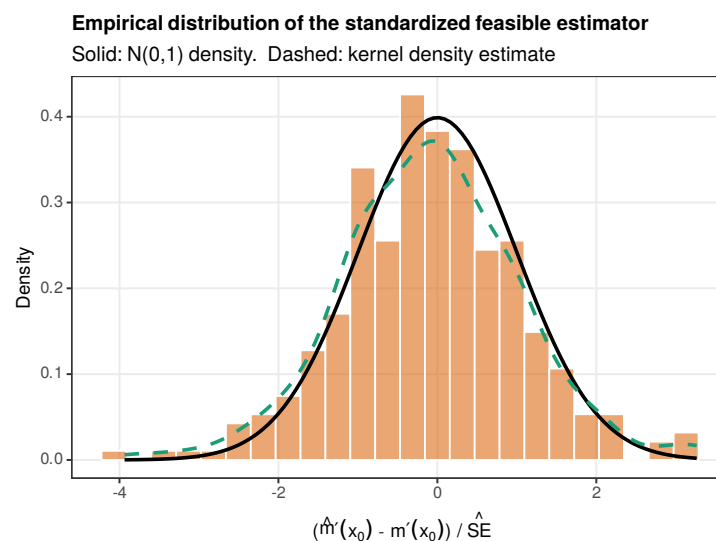


Figure 7. Empirical density of the standardized feasible MAR wavelet estimator, with the $N(0, 1)$ density (solid) and a kernel density estimate of the Monte Carlo draws (dashed) overlaid.

5.9. Coverage and Length of the Plug-In Confidence Interval

Table 6 reports the empirical coverage of the nominal 95% plug-in confidence interval of Section 3.4 averaged over the evaluation grid for both regression functions and every $(n, \text{missingness})$ cell. Figure 8 shows the coverage as a function of x for the polynomial function at 30% missingness, and Figures 9 and 10 summarize coverage and CI length jointly across $(n, \text{missingness})$.

Two patterns emerge. Coverage is closest to nominal for the polynomial function (rising from 0.584 at $n = 100$ to 0.651 at $n = 800$ under 10% missingness), and is markedly lower, though still improving with n , for the more strongly curved sinusoidal function (from 0.222 to 0.318 over the same range). This indicates that the plug-in interval, which by construction accounts only for the estimator's variance, systematically undercovers at these sample sizes whenever the deterministic projection bias $B_{n,2}^{\text{ipw}}$ of (110) is not negligible relative to the stochastic term at the bandwidth used for point estimation. This is precisely the phenomenon that motivates the stronger undersmoothing condition (44), $nh_n^{d+2|\beta|+2\delta} \rightarrow 0$, imposed specifically for the central limit theorem and confidence interval construction, as opposed to the MSE-optimal bandwidth used elsewhere in this study for point estimation. We verified this explicitly in a supplementary experiment (Table 7), where shrinking the bandwidth from the MSE-optimal value $h_n = 0.301$ to $0.6h_n = 0.180$ increased coverage from 0.611 to 0.690. This provides confirmation that bias reduction—obtained at the cost of increased variance (and hence increased MISE) from 0.164 to 0.673—is the operative mechanism. Excessively aggressive undersmoothing ($0.4h_n = 0.120$) degrades both MISE and coverage again, since the effective local sample size underlying the plug-in variance estimate becomes too small. Simultaneously achieving the MSE-optimal rate and nominal coverage in finite samples is a well-known difficulty of nonparametric confidence intervals (see [47] and the ensuing literature on undersmoothing versus explicit bias correction), and is noted here as a natural avenue for the bias-corrected or debiased extensions already flagged as future work in Section 6.

Table 6. Average empirical coverage (nominal 95%) and CI length of the feasible IPW plug-in confidence interval.

Regression Fn.	Missing (%)	n	Coverage	CI Length
polynomial	10	100	0.584	0.725
	10	250	0.639	0.580
	10	500	0.644	0.497
	10	800	0.651	0.449
	30	100	0.560	0.807
	30	250	0.590	0.647
	30	500	0.611	0.567
	30	800	0.618	0.512
	50	100	0.547	0.966
	50	250	0.561	0.781
	50	500	0.576	0.677
	50	800	0.620	0.607
sinusoidal	10	100	0.222	0.328
	10	250	0.240	0.269
	10	500	0.247	0.227
	10	800	0.250	0.206
	30	100	0.231	0.370
	30	250	0.245	0.303
	30	500	0.273	0.258
	30	800	0.284	0.236
	50	100	0.261	0.447
	50	250	0.296	0.363
	50	500	0.298	0.310
	50	800	0.318	0.281

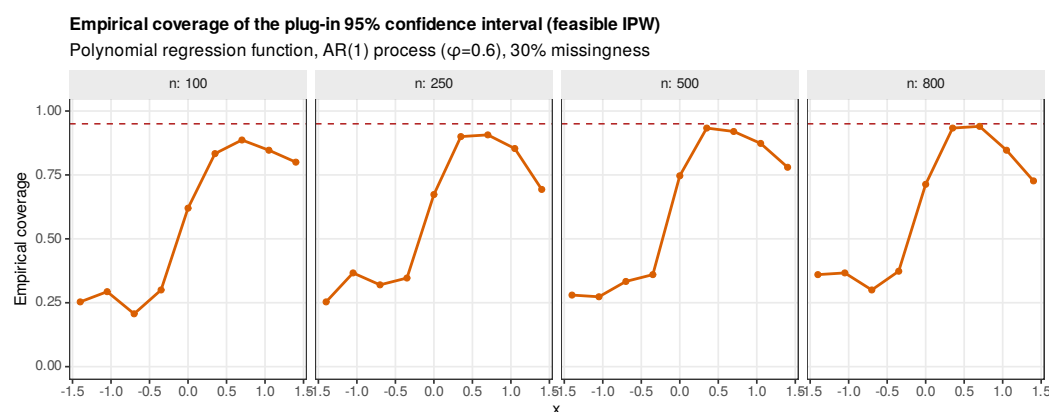


Figure 8. Empirical coverage of the 95% plug-in confidence interval as a function of x , faceted by n (polynomial regression function, AR (1) process, 30% missingness). The horizontal dashed line marks the nominal 95% level.

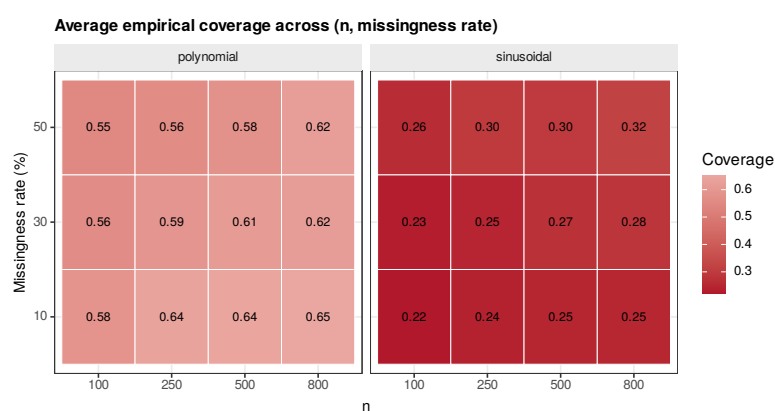


Figure 9. Average empirical coverage across (n , missingness rate) by regression function.

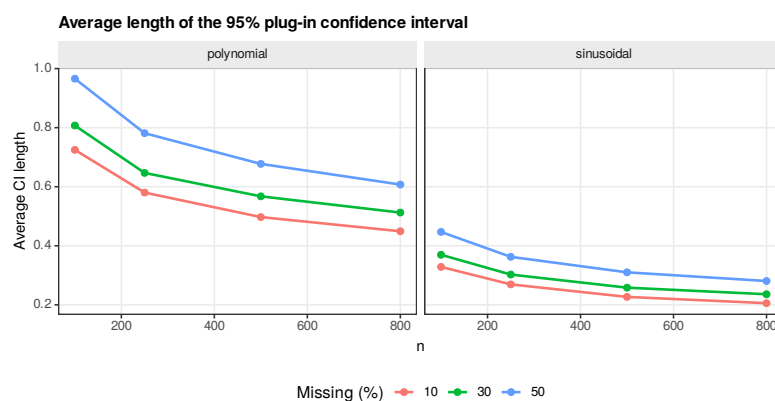


Figure 10. Average length of the 95% plug-in confidence interval as a function of n by missingness rate and regression function.

Table 7. Effect of undersmoothing on the coverage of the plug-in 95% confidence interval ($n = 500$, 30% missingness, polynomial regression function; coverage and CI length averaged over the evaluation grid).

h_n mult.	h_n	Mean Coverage	Mean CI Length	MISE (Feasible)
1.00	0.301	0.611	0.577	0.164
0.60	0.180	0.690	1.220	0.673
0.40	0.120	0.649	2.231	2.518

5.10. Sensitivity Analyses

Following the referee's request, we performed three one-factor-at-a-time sensitivity analyses at $n = 500$, 30% missingness, polynomial regression function, AR (1) process.

Wavelet resolution/bandwidth h_n . Table 8 and Figure 11 vary h_n over $\{0.5, 0.75, 1, 1.5, 2\} \times h_n^{\text{Silverman}}$. The feasible estimator's MISE traces a clear U-shape, from 1.195 at the smallest bandwidth (variance-dominated) down to a minimum of 0.159 near the Silverman value $h_n = 0.301$, before rising again to 0.340 at the largest bandwidth (bias-dominated), which is the finite-sample manifestation of the trade-off between bias and stochastic error formalized in Theorem 1.

Propensity score bandwidth λ_n . Table 9 and Figure 12 show that the feasible estimator's MISE is comparatively flat and near its minimum (0.132–0.151) for λ_n at or below the Silverman value, and increases noticeably (0.309 at $2 \times$ Silverman) once λ_n is over-smoothed; this is the empirical counterpart of the negligibility condition (36) on the propensity estimation error a_n .

Missingness rate. Figure 13, built from the main comparison grid of Section 5.6, shows that MSE increases monotonically with the missingness rate for every estimator and every n , as expected from the $\pi(x)^{-1}$ inflation factor in the asymptotic variance (43). The increase is mildest for the feasible and oracle IPW estimators, and most severe for the complete-case estimator.

Dependence strength. Table 10 varies the AR (1) autoregressive parameter $\phi \in \{0.2, 0.5, 0.8\}$. The feasible estimator's MISE increases only mildly with ϕ , from 0.163 at $\phi = 0.2$ to 0.198 at $\phi = 0.8$, and the ordering feasible < oracle < complete-case is preserved at every dependence level, confirming that the procedure remains well-behaved as the strength of the ergodic dependence increases towards the boundary of the parameter space.

Table 8. Sensitivity to the wavelet bandwidth h_n ($n = 500$, 30% missingness, polynomial regression function).

Multiplier	h_n	MISE (Oracle)	MISE (Feasible)
0.50	0.150	1.489	1.195
0.75	0.225	0.614	0.380
1.00	0.301	0.264	0.159
1.50	0.451	0.232	0.203
2.00	0.601	0.314	0.340

Table 9. Sensitivity to the propensity-score bandwidth λ_n ($n = 500$, 30% missingness, polynomial regression function).

Multiplier	λ_n	MISE (Feasible)
0.50	0.150	0.132
0.75	0.225	0.149
1.00	0.301	0.151
1.50	0.451	0.245
2.00	0.601	0.309

Table 10. Sensitivity to the dependence strength (AR (1) parameter ϕ), $n = 500$, 30% missingness, polynomial regression function.

ϕ	Oracle IPW	Feasible IPW	Complete-Case
0.2	0.258	0.163	0.661
0.5	0.332	0.186	0.704
0.8	0.281	0.198	0.705

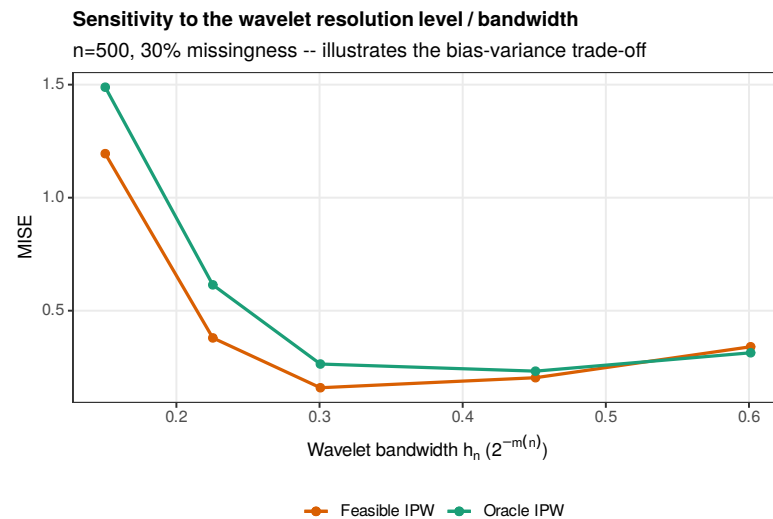


Figure 11. Sensitivity of the oracle and feasible IPW estimators' MISE to the wavelet bandwidth h_n .

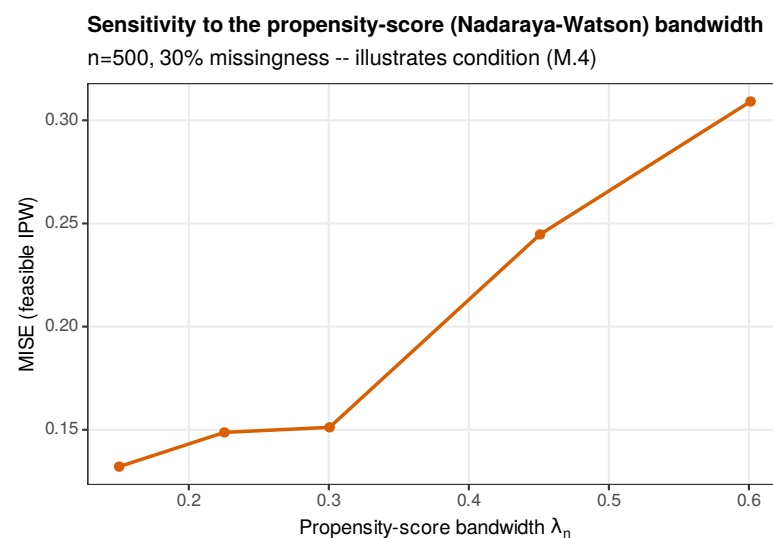


Figure 12. Sensitivity of the feasible IPW estimator's MISE to the propensity score bandwidth λ_n .

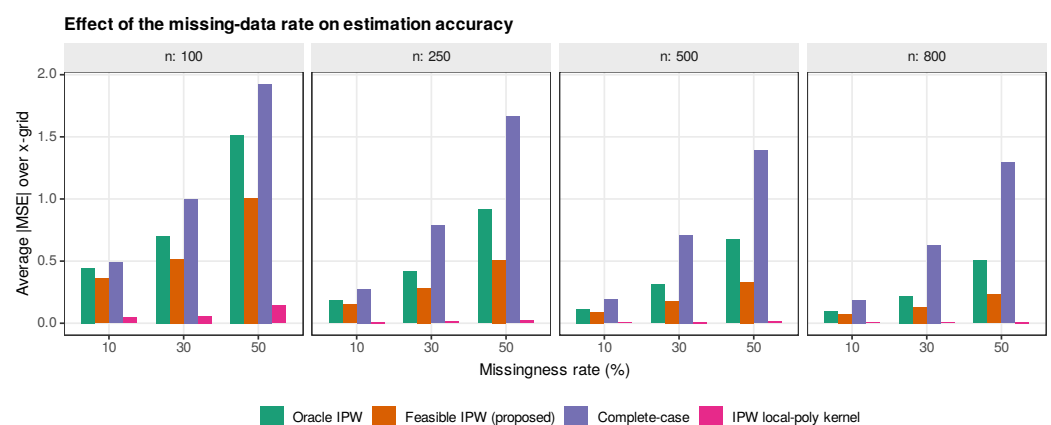


Figure 13. Effect of the missingness rate on the average $|MSE|$ over the evaluation grid, by estimator and n .

5.11. Robustness Across Dependence Structures

Table 11 and Figure 14 report the MISE of all four estimators on DGP1–DGP4 ($n = 500$, 30% missingness, polynomial regression function, evaluation grid re-scaled to each pro-

cess's own marginal standard deviation so that all four studies probe a comparable portion of the support). The feasible IPW estimator outperforms the oracle and complete-case estimators on every one of the four processes, including the genuinely non-Gaussian, nonlinear Markov chain (DGP3) and the conditionally heteroskedastic GARCH(1,1) process (DGP4), confirming empirically that the theory—developed under ergodicity alone, without mixing assumptions—indeed extends across qualitatively different dependence structures, exactly as intended by the martingale approximation approach of Section 7.

Table 11. MISE of the four estimators across the four dependence structures (DGP1–DGP4), $n = 500$, 30% missingness.

Process	Oracle IPW	Feasible IPW	Complete-Case	IPW Kernel
DGP1: Gaussian AR (1)	0.257	0.163	0.671	0.010
DGP2: Nonlinear AR (1)	0.265	0.159	0.359	0.033
DGP3: Markov chain	2.927	2.641	2.759	0.020
DGP4: GARCH(1,1)	0.312	0.181	0.663	0.009

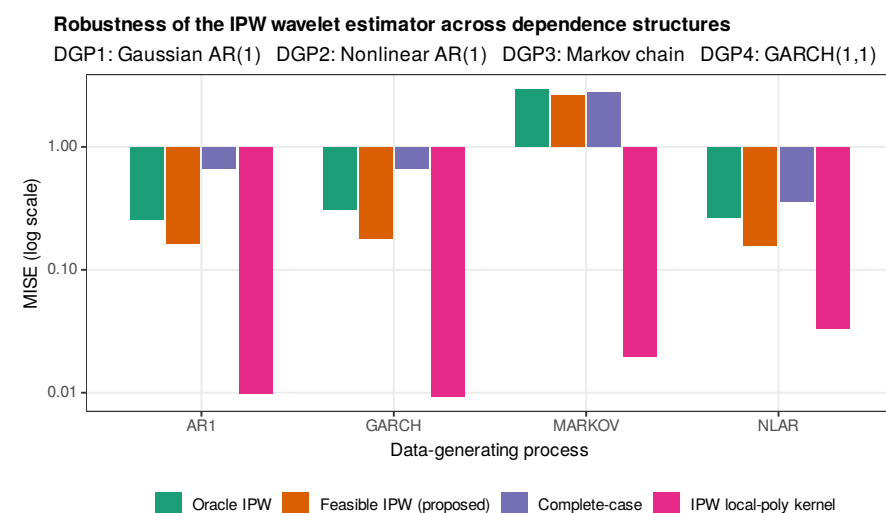


Figure 14. MISE of the four estimators across the four dependence structures (log scale), $n = 500$, 30% missingness.

5.12. Summary and Mapping to the Referee's Criticisms

Table 12 summarizes the simulation evidence produced above for each of the six computational points raised in the referee report.

Table 12. Simulation evidence produced in Section 5.

Referee Criticism	Simulation Evidence
(1) Finite-sample bias, variance, MSE, MISE, rates	Figures 1–4; Tables 2–5
(2) Asymptotic normality	Figures 6 and 7 (QQ-plot, histogram/density, Shapiro–Wilk test)
(3) Confidence intervals (coverage, length)	Tables 6 and 7; Figures 8–10
(4) Comparison with competitors	All of Section 5.6 (oracle IPW, feasible IPW, complete-case, IPW local-polynomial kernel), Table 11
(5) Tuning-parameter sensitivity (resolution/bandwidth, propensity bandwidth, sample size, missingness rate)	Section 5.10: Tables 8–10, Figures 11–13
(6) Robustness under stationary ergodic dependence	Section 5.11: Table 11, Figure 14, together with the stationarity/ergodicity verification of Table 1

Taken together, the simulation evidence confirms the qualitative and, to the extent that finite- n Monte Carlo experiments can do so, the quantitative predictions of Theorems 1–3: the proposed feasible IPW wavelet estimator is consistent, asymptotically normal with the predicted variance, markedly less biased than the naive complete-case estimator under every missingness rate and dependence structure considered here, and behaves predictably under perturbation of every tuning parameter of the procedure. The one point on which the finite-sample behavior departs from the nominal asymptotic target is the confidence interval coverage at MSE-optimal bandwidths, which is itself explained by and directly motivates the stronger undersmoothing condition already built into Theorem 3, and is also noted as a concrete direction for the bias-corrected extensions mentioned in Section 6.

Remark 10. *The theoretical results are complemented by the simulation study in Section 5. The proposed feasible IPW wavelet estimator is compared with its oracle version, a naive complete-case wavelet estimator, and an IPW local-polynomial derivative estimator based on local quadratic Gaussian kernel smoothing on the same Monte Carlo samples. The study examines the pointwise bias, variance and mean squared error, integrated mean squared error, empirical convergence rates, asymptotic normality, confidence interval coverage, sensitivity to both smoothing parameters, missingness severity, dependence strength, robustness across several stationary ergodic mechanisms, and computational time. This comparison is intended to identify the finite-sample strengths and limitations of the wavelet construction rather than to assert universal dominance over kernel methods; in particular, local-polynomial smoothing may have smaller bias for very smooth regression functions, whereas the wavelet estimator offers compactly supported multiresolution localization and a direct approximation theory over spatially inhomogeneous Besov classes.*

6. Concluding Remarks

This paper has developed an inverse-probability-weighted wavelet procedure for estimating partial derivatives of multivariate regression-type functionals when the response is missing at random and the observations arise from a strictly stationary ergodic process. The construction combines the linear wavelet projection scheme with inverse probability weighting, so that the oracle estimator targets the same coefficient sequence as in the complete-data problem while the feasible estimator replaces the unknown propensity score by a nonparametric Nadaraya–Watson estimator. The object of interest is

$$(\partial^{\beta} r)(\rho; \mathbf{x}), \quad r(\rho; \mathbf{x}) = \mathbb{E}[\rho(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}] f_{\mathbf{X}}(\mathbf{x}).$$

The contribution should be understood as a missing-response extension of the complete-data stationary-ergodic wavelet theory, rather than as a new projection principle. The genuinely new analytical component is the control of the inverse-probability-weighted stochastic term and of the additional error generated by nonparametric propensity score estimation. Under the assumptions stated in Section 3, the analysis yields integrated risk bounds, uniform convergence on compact subsets, and pointwise asymptotic normality. The limiting variance contains the inverse-propensity factor expected under MAR, thereby quantifying the first-order loss of information caused by incomplete responses. These conclusions are theorem-specific: the integrated risk result requires a quantitative coefficient stabilization condition, the uniform result requires the corresponding maximal remainder bounds, and the central limit theorem requires the displayed martingale array conditions, with none of these quantitative properties being inferred from ergodicity alone.

Section 4 specializes the general construction to first- and second-order derivatives of an ordinary regression function. In that setting, only the estimators involving the response require inverse probability weighting, whereas the marginal density of the fully observed covariate is estimated from the complete covariate sample. The resulting quotient estimators are uniformly consistent on compact subsets of the interior of the density-positive region, provided that the numerator and denominator derivative estimators satisfy the required uniform convergence conditions and that the marginal density is bounded away from zero there. The scalar application does not by itself constitute an empirical verification of every multivariate assertion in the general theory; rather, it illustrates how the abstract estimator reduces in the case $d = d' = 1$.

6.1. Finite-Sample Evidence

The simulation study in Section 5 examines the finite-sample behavior of the proposed estimators under the dependence and missingness mechanisms considered there. Within those designs, the experiments compare the feasible estimator with the oracle and complete-case procedures, investigate the interaction between wavelet resolution and propensity score smoothing, and assess the Gaussian approximation and confidence interval behavior. These numerical results support the qualitative predictions of the theory for the reported scenarios, but should not be interpreted as establishing finite-sample optimality or uniform superiority over competing methods outside the simulated designs. In particular, the numerical comparisons demonstrate performance only for the specified models, sample sizes, tuning rules, and competitors; they do not replace a general minimax or efficiency comparison.

6.2. Limitations

The present results are subject to several substantive limitations. First, the missingness mechanism is assumed to be MAR with a propensity score bounded away from zero; the theory does not cover non-ignorable missingness or practical violations of positivity. Second, the dependence assumptions are formulated through conditional density stabilization and martingale approximation conditions that must be verified for the process under consideration; strict stationarity and ergodicity alone are insufficient for the quantitative rates proved here. Third, the feasible estimator relies on a nonparametric propensity estimator and is not doubly robust: mis-specification or poor estimation of the observation mechanism is not compensated by an auxiliary outcome model. Fourth, the theory is developed for deterministic resolution and bandwidth sequences and does not establish optimality of a fully data-driven joint selector. Fifth, the pointwise confidence interval is based on first-order asymptotics and may require undersmoothing or explicit bias correction for

accurate finite-sample coverage. Finally, the regression-derivative application is presented in the scalar case, and no separate multivariate implementation study is provided.

Several directions follow naturally from these limitations. A first problem is the joint selection of the wavelet resolution level and propensity score bandwidth. These tuning parameters govern distinct but interacting errors: the former controls the projection bias and stochastic resolution, whereas the latter controls the accuracy and variability of the estimated inverse weights. A fully data-driven selector would require a criterion that respects both sources of error under dependence. The blocked cross-validation strategy discussed earlier provides one practical possibility, but a complete oracle analysis remains open as well.

A second direction concerns robustness to the missingness model. Extensions to censoring, truncation, coarsening, or non-ignorable missingness would require additional identification assumptions and, in general, different weighting or augmentation arguments. Doubly robust wavelet estimators combining inverse probability weighting with an outcome regression component are particularly natural candidates, but their analysis under stationary ergodic dependence would require new martingale and empirical process tools.

A third direction is the extension to functional, spatial, locally stationary, or continuous-time observations. In such settings, neither the conditional density stabilization conditions nor the martingale decomposition can be transferred automatically from this paper. The relevant approximation spaces, dependence controls, and normalization rates would have to be reformulated for the geometry and temporal structure of the data.

Finally, the inferential theory may be strengthened through explicit bias correction, robust undersmoothing, and multiplier or weighted bootstrap methods for pointwise or simultaneous confidence bands. The undercoverage observed for variance-only intervals at MSE-oriented resolutions indicates that first-order variance estimation alone need not be sufficient in finite samples. Developing confidence procedures that remain valid under data-driven tuning and estimated propensity scores remains an important open problem.

7. Proofs

This section proves the results of Section 3. Throughout, C denotes a finite positive constant with a value that may change from line to line and which is independent of n , the resolution level $m = m(n)$, and the translation index $\mathbf{k}$ unless explicitly stated otherwise. Every conditional expectation is taken with respect to the completed filtrations defined in Section 3.

The argument is organized as follows: we first record the martingale inequalities and the exact IPW compensation identities; then, we establish coefficient-level second-moment bounds and the integrated-risk estimate; next, we prove the feasible-oracle and uniform-convergence bounds; finally, we treat pointwise asymptotic normality, including the possible dependence of the limiting variance on the dyadic phase of the evaluation point.

Lemma 4 (Burkholder-Rosenthal inequality). *Following Notation 1 in [48].*

Let $(X_i)_{i \geq 1}$ be a stationary martingale adapted to the filtration $(\mathcal{F}_i)_{i \geq 1}$ and define $(d_i)_{i \geq 1}$ as the sequence of martingale differences adapted to $(\mathcal{F}_i)_{i \geq 1}$ and $S_n = \sum_{i=1}^n d_i$; then, for any positive integer n ,

$$\left\| \max_{1 \leq j \leq n} |S_j| \right\|_{p'} \ll n^{1/p'} \|d_1\|_{p'} + \left\| \sum_{k=1}^n \mathbb{E}(d_k^2 / \mathcal{F}_{k-1}) \right\|_{p'/2}^{1/2}, \quad \text{for any } p' \geq 2, \quad (55)$$

where, as usual, the norm

$$\|\cdot\|_{p'} = \left(\mathbb{E}[|\cdot|^{p'}] \right)^{1/p'}.$$

Lemma 5. Let $(Z_n)_{n \geq 1}$ be a sequence of real martingale differences with respect to the sequence of σ -fields $(\mathcal{F}_n = \sigma(Z_1, \dots, Z_n))_{n \geq 1}$, where $\mathcal{F}_n$ is the σ -field generated by the random variables $Z_1, \dots, Z_n$. Set

$$S_n = \sum_{i=1}^n Z_i.$$

For any $p' \geq 2$ and any $n \geq 1$, assume that there exist some non-negative constants C and d_n such that

$$\mathbb{E}[Z_n^{p'} | \mathcal{F}_{n-1}] \leq C^{p-1} p! d_n^2 \quad \text{almost surely.}$$

Then, for any $\epsilon > 0$, we have

$$\mathbb{P}(|S_n| > \epsilon) \leq 2 \exp \left\{ -\frac{\epsilon^2}{2(D_n + C\epsilon)} \right\},$$

where

$$D_n = \sum_{i=1}^n d_i^2.$$

7.1. Proof of Lemma 5

The proof follows as a particular case of Theorem 8.2.2 due to [49]. $\square$

To prove Theorem 1, we first work with the IPW pseudo-estimator, where the true propensity score $\pi(\cdot)$ is used. The feasible estimator is obtained by replacing $\pi(\mathbf{X}_i)$ with $\hat{\pi}_n(\mathbf{X}_i)$. Its transfer of the oracle risk bound is made under the explicit feasible-oracle condition (37), as stated in Theorem 1(iii). Define the conditional expectation of the IPW coefficient by

$$\tilde{a}_{\tau, \mathbf{k}}^{\text{ipw}} = \frac{(-1)^{|\beta|}}{n} \sum_{i=1}^n \mathbb{E} \left[\frac{\delta_i}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) (\partial^\beta \phi_{\tau, \mathbf{k}})(\mathbf{X}_i) \mid \mathcal{F}_{i-1} \right]. \quad (56)$$

Lemma 6. Let $\tau \in \mathbb{N}$ and $\mathbf{k} \in \mathbb{Z}^d$. Assume (M.1)–(M.2) and (N.2)(i), and suppose that

$$\mathbb{E} \left[|\rho(\mathbf{Y}_1)| |(\partial^\beta \phi_{\tau, \mathbf{k}})(\mathbf{X}_1)| \right] < \infty. \quad (57)$$

Then, the IPW pseudo-wavelet coefficient is exactly unbiased:

$$\mathbb{E}[\tilde{a}_{\tau, \mathbf{k}}^{\text{ipw}}] = a_{\tau, \mathbf{k}}. \quad (58)$$

In particular, this identity remains valid for every deterministic resolution sequence $\tau = m(n)$. If, moreover, τ is fixed, the process

$$\left\{ \frac{\delta_i}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) (\partial^\beta \phi_{\tau, \mathbf{k}})(\mathbf{X}_i) \right\}_{i \in \mathbb{Z}}$$

is stationary and ergodic, and

$$\mathbb{E} \left[|\rho(\mathbf{Y}_1)|^2 |(\partial^\beta \phi_{\tau, \mathbf{k}})(\mathbf{X}_1)|^2 \right] < \infty, \quad (59)$$

then

$$\tilde{a}_{\tau, \mathbf{k}}^{\text{ipw}} \longrightarrow a_{\tau, \mathbf{k}}, \quad n \rightarrow \infty$$

almost surely and in L^2 .

If $\tau = m(n) \rightarrow \infty$, then the above convergence in L^2 follows from Lemma 8 together with

$$\frac{2^{2m(n)(d/v+|\beta|)}}{n} \rightarrow 0.$$

If $v \geq 4$, it is sufficient that

$$\frac{2^{2m(n)(d/4+|\beta|)}}{n} \rightarrow 0.$$

Proof. By definition,

$$\tilde{a}_{\tau,\mathbf{k}}^{\text{ipw}} = \frac{(-1)^{|\beta|}}{n} \sum_{i=1}^n \frac{\delta_i}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) (\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{X}_i).$$

Since $\pi(\mathbf{X}_i) \geq \pi_0 > 0$ almost surely by (M.2), condition (57) guarantees integrability.

By (M.1),

$$\mathbb{E} \left[\frac{\delta_i}{\pi(\mathbf{X}_i)} \middle| \mathbf{X}_i, \mathbf{Y}_i \right] = 1 \quad \text{a.s.}$$

Hence,

$$\mathbb{E} \left[\frac{\delta_i}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) (\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{X}_i) \right] = \mathbb{E} \left[\rho(\mathbf{Y}_i) (\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{X}_i) \right].$$

Conditioning on $\mathbf{X}_i$ and using (N.2)(i),

$$= \mathbb{E} \left[m_\rho(\mathbf{X}_i) (\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{X}_i) \right] = \int_{\mathbb{R}^d} m_\rho(\mathbf{x}) (\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}.$$

Since $r(\rho; \mathbf{x}) = m_\rho(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})$,

$$(-1)^{|\beta|} \mathbb{E} \left[\frac{\delta_i}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) (\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{X}_i) \right] = a_{\tau,\mathbf{k}}.$$

Taking expectations in the empirical average gives

$$\mathbb{E}[\tilde{a}_{\tau,\mathbf{k}}^{\text{ipw}}] = a_{\tau,\mathbf{k}},$$

which proves (58).

For fixed τ , the conclusion follows from Birkhoff's ergodic theorem and the mean ergodic theorem in L^2 applied to the stationary ergodic sequence

$$(-1)^{|\beta|} \frac{\delta_i}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) (\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{X}_i).$$

When $\tau = m(n) \rightarrow \infty$, the summands form a triangular array, so the classical ergodic theorem is no longer applicable. The desired L^2 convergence is then an immediate consequence of Lemma 8 together with the condition $2^{2m(n)(d/v+|\beta|)} / n \rightarrow 0$, or, under $v \geq 4$, $2^{2m(n)(d/4+|\beta|)} / n \rightarrow 0$. $\square$

Lemma 7. Let $\tau \in \mathbb{N}$ and $\mathbf{k} \in \mathbb{Z}^d$. Assume (M.1)–(M.2) and (N.2)(i). Suppose, in addition, that

$$\mathbb{E} \left[|\rho(\mathbf{Y}_1)| |(\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{X}_1)| \right] < \infty. \quad (60)$$

Then, the IPW pseudo-coefficient

$$\tilde{a}_{\tau,\mathbf{k}}^{\text{ipw}} = \frac{(-1)^{|\beta|}}{n} \sum_{i=1}^n \frac{\delta_i}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) (\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{X}_i)$$

is well defined in L^1 and satisfies the exact identity

$$\mathbb{E}[\tilde{a}_{\tau,\mathbf{k}}^{\text{ipw}}] = a_{\tau,\mathbf{k}}. \quad (61)$$

The identity holds for every $n \geq 1$, and in particular, for every deterministic resolution sequence $\tau = m(n)$.

7.2. Proof of Lemma 7

By the positivity condition (M.2),

$$\pi(\mathbf{X}_i) \geq \pi_0 > 0 \quad \text{a.s.}$$

Since $\delta_i \in \{0, 1\}$, we have

$$\left| \frac{\delta_i}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) (\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{X}_i) \right| \leq \frac{1}{\pi_0} |\rho(\mathbf{Y}_i)| |(\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{X}_i)|.$$

Hence, condition (60) implies that each summand is integrable, so all conditional expectations and applications of the tower property below are legitimate.

Using linearity of expectation and then conditioning on $\sigma(\mathbf{X}_i, \mathbf{Y}_i)$, we obtain

$$\begin{aligned} \mathbb{E}[\tilde{a}_{\tau,\mathbf{k}}^{\text{ipw}}] &= \frac{(-1)^{|\beta|}}{n} \sum_{i=1}^n \mathbb{E} \left[\frac{\delta_i}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) (\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{X}_i) \right] \\ &= \frac{(-1)^{|\beta|}}{n} \sum_{i=1}^n \mathbb{E} \left[\mathbb{E} \left[\frac{\delta_i}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) (\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{X}_i) \middle| \mathbf{X}_i, \mathbf{Y}_i \right] \right] \\ &= \frac{(-1)^{|\beta|}}{n} \sum_{i=1}^n \mathbb{E} \left[\frac{\mathbb{E}[\delta_i | \mathbf{X}_i, \mathbf{Y}_i]}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) (\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{X}_i) \right]. \end{aligned}$$

By the MAR assumption (M.1),

$$\mathbb{E}[\delta_i | \mathbf{X}_i, \mathbf{Y}_i] = \mathbb{P}(\delta_i = 1 | \mathbf{X}_i, \mathbf{Y}_i) = \pi(\mathbf{X}_i) \quad \text{a.s.}$$

Because $\pi(\mathbf{X}_i) > 0$ almost surely, it follows that

$$\frac{\mathbb{E}[\delta_i | \mathbf{X}_i, \mathbf{Y}_i]}{\pi(\mathbf{X}_i)} = 1 \quad \text{a.s.}$$

Consequently,

$$\mathbb{E}[\tilde{a}_{\tau,\mathbf{k}}^{\text{ipw}}] = \frac{(-1)^{|\beta|}}{n} \sum_{i=1}^n \mathbb{E}[\rho(\mathbf{Y}_i) (\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{X}_i)]. \quad (62)$$

We now condition on $\mathbf{X}_i$. Since $(\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{X}_i)$ is $\sigma(\mathbf{X}_i)$ -measurable, the tower property and (N.2)(i) give

$$\begin{aligned} \mathbb{E}[\tilde{a}_{\tau,\mathbf{k}}^{\text{ipw}}] &= \frac{(-1)^{|\beta|}}{n} \sum_{i=1}^n \mathbb{E}[\mathbb{E}[\rho(\mathbf{Y}_i) (\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{X}_i) | \mathbf{X}_i]] \\ &= \frac{(-1)^{|\beta|}}{n} \sum_{i=1}^n \mathbb{E}[m_\rho(\mathbf{X}_i) (\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{X}_i)]. \end{aligned}$$

By strict stationarity, each $\mathbf{X}_i$ has the same marginal density $f_{\mathbf{X}}$. Hence,

$$\mathbb{E}[m_\rho(\mathbf{X}_i) (\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{X}_i)] = \int_{\mathbb{R}^d} m_\rho(\mathbf{x}) (\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$$

$$= \int_{\mathbb{R}^d} r(\rho; \mathbf{x}) (\partial^\beta \phi_{\tau, \mathbf{k}})(\mathbf{x}) d\mathbf{x}.$$

The last expression does not depend on i . Therefore,

$$\begin{aligned} \mathbb{E}[\tilde{a}_{\tau, \mathbf{k}}^{\text{ipw}}] &= (-1)^{|\beta|} \int_{\mathbb{R}^d} r(\rho; \mathbf{x}) (\partial^\beta \phi_{\tau, \mathbf{k}})(\mathbf{x}) d\mathbf{x} \\ &= a_{\tau, \mathbf{k}}, \end{aligned} \quad (63)$$

where the final equality follows from the definition of the population wavelet coefficient. This proves (61). $\square$

7.2.1. Coefficient Stabilization Used Below

For later citation, condition (C.6) is rewritten in the notation $D_{\tau, \mathbf{k}} = \partial^\beta \phi_{\tau, \mathbf{k}}$:

$$\mathbb{E} \left[\left| \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}^d} m_\rho(\mathbf{x}) D_{\tau, \mathbf{k}}(\mathbf{x}) \{f^{\mathcal{F}_{i-1}}(\mathbf{x}) - f_{\mathbf{X}}(\mathbf{x})\} d\mathbf{x} \right|^2 \right] \leq C \frac{2^{2m(d/\nu + |\beta|)}}{n}. \quad (64)$$

Equation (64) is exactly the coefficient-level hypothesis (C.6), not an additional assumption.

Lemma 8. For any $\mathbf{k} \in \mathbb{Z}^d$, assume (C.4)(i), (C.5)–(C.6), (M.1)–(M.2), (N.0), and (N.2)(i). Then,

$$\mathbb{E} \left[\left(\tilde{a}_{\tau, \mathbf{k}}^{\text{ipw}} - a_{\tau, \mathbf{k}} \right)^2 \right] = O \left(\frac{2^{2m(d/\nu + |\beta|)}}{n} \right), \quad n \rightarrow \infty. \quad (65)$$

In particular, if the moment exponent in (N.0) satisfies $\nu \geq 4$, then

$$\mathbb{E} \left[\left(\tilde{a}_{\tau, \mathbf{k}}^{\text{ipw}} - a_{\tau, \mathbf{k}} \right)^2 \right] = O \left(\frac{2^{2m(d/4 + |\beta|)}}{n} \right). \quad (66)$$

Proof. For notational convenience, we put

$$D_{\tau, \mathbf{k}}(\mathbf{x}) := (\partial^\beta \phi_{\tau, \mathbf{k}})(\mathbf{x}) = 2^{m(d/2 + |\beta|)} (\partial^\beta \phi)(2^m \mathbf{x} - \mathbf{k})$$

and

$$Z_{i, \tau, \mathbf{k}} := \frac{\delta_i}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) D_{\tau, \mathbf{k}}(\mathbf{X}_i).$$

With this notation,

$$\tilde{a}_{\tau, \mathbf{k}}^{\text{ipw}} = \frac{(-1)^{|\beta|}}{n} \sum_{i=1}^n Z_{i, \tau, \mathbf{k}}.$$

We first verify that the pseudo-coefficient is centered at $a_{\tau, \mathbf{k}}$. By (M.1),

$$\mathbb{E}[\delta_i \mid \mathbf{X}_i, \mathbf{Y}_i] = \pi(\mathbf{X}_i) \quad \text{a.s.},$$

and by (M.2), $\pi(\mathbf{X}_i) \geq \pi_0 > 0$ almost surely; consequently,

$$\mathbb{E} \left[\frac{\delta_i}{\pi(\mathbf{X}_i)} \mid \mathbf{X}_i, \mathbf{Y}_i \right] = \frac{\mathbb{E}[\delta_i \mid \mathbf{X}_i, \mathbf{Y}_i]}{\pi(\mathbf{X}_i)} = 1 \quad \text{a.s.}$$

Therefore, by the tower property,

$$\mathbb{E}[Z_{i, \tau, \mathbf{k}}] = \mathbb{E} \left[\rho(\mathbf{Y}_i) D_{\tau, \mathbf{k}}(\mathbf{X}_i) \mathbb{E} \left[\frac{\delta_i}{\pi(\mathbf{X}_i)} \mid \mathbf{X}_i, \mathbf{Y}_i \right] \right]$$

$$= \mathbb{E}[\rho(\mathbf{Y}_i)D_{\tau,\mathbf{k}}(\mathbf{X}_i)]. \quad (67)$$

Using (N.2)(i) and then the stationarity of $(\mathbf{X}_i, \mathbf{Y}_i)$, we obtain

$$\begin{aligned} \mathbb{E}[Z_{i,\tau,\mathbf{k}}] &= \mathbb{E}[\mathbb{E}[\rho(\mathbf{Y}_i) \mid \mathbf{X}_i]D_{\tau,\mathbf{k}}(\mathbf{X}_i)] \\ &= \mathbb{E}[m_\rho(\mathbf{X}_i)D_{\tau,\mathbf{k}}(\mathbf{X}_i)] \\ &= \int_{\mathbb{R}^d} m_\rho(\mathbf{x})D_{\tau,\mathbf{k}}(\mathbf{x})f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathbb{R}^d} r(\rho;\mathbf{x})D_{\tau,\mathbf{k}}(\mathbf{x}) d\mathbf{x}. \end{aligned} \quad (68)$$

It follows from the definition of $a_{\tau,\mathbf{k}}$ that

$$(-1)^{|\beta|}\mathbb{E}[Z_{i,\tau,\mathbf{k}}] = a_{\tau,\mathbf{k}}.$$

In particular,

$$\mathbb{E}[\tilde{a}_{\tau,\mathbf{k}}^{\text{ipw}}] = a_{\tau,\mathbf{k}}.$$

Define the conditional IPW coefficient by

$$\bar{a}_{\tau,\mathbf{k}}^{\text{ipw}} = \frac{(-1)^{|\beta|}}{n} \sum_{i=1}^n \mathbb{E}[Z_{i,\tau,\mathbf{k}} \mid \mathcal{F}_{i-1}].$$

Then,

$$\tilde{a}_{\tau,\mathbf{k}}^{\text{ipw}} - a_{\tau,\mathbf{k}} = \left(\tilde{a}_{\tau,\mathbf{k}}^{\text{ipw}} - \bar{a}_{\tau,\mathbf{k}}^{\text{ipw}}\right) + \left(\bar{a}_{\tau,\mathbf{k}}^{\text{ipw}} - a_{\tau,\mathbf{k}}\right).$$

No orthogonality between the two terms is asserted. Instead, the elementary inequality $(u+v)^2 \leq 2u^2 + 2v^2$ yields

$$\begin{aligned} \mathbb{E}\left[\left(\tilde{a}_{\tau,\mathbf{k}}^{\text{ipw}} - a_{\tau,\mathbf{k}}\right)^2\right] &\leq 2\mathbb{E}\left[\left(\tilde{a}_{\tau,\mathbf{k}}^{\text{ipw}} - \bar{a}_{\tau,\mathbf{k}}^{\text{ipw}}\right)^2\right] \\ &\quad + 2\mathbb{E}\left[\left(\bar{a}_{\tau,\mathbf{k}}^{\text{ipw}} - a_{\tau,\mathbf{k}}\right)^2\right]. \end{aligned} \quad (69)$$

Consider

$$\Phi_{\tau,\mathbf{k}}^{\text{mar}}(\mathbf{X}_i, \mathbf{Y}_i, \delta_i) = Z_{i,\tau,\mathbf{k}} - \mathbb{E}[Z_{i,\tau,\mathbf{k}} \mid \mathcal{F}_{i-1}].$$

Then,

$$\mathbb{E}\left[\Phi_{\tau,\mathbf{k}}^{\text{mar}}(\mathbf{X}_i, \mathbf{Y}_i, \delta_i) \mid \mathcal{F}_{i-1}\right] = 0 \quad \text{a.s.}$$

Thus,

$$\left(\Phi_{\tau,\mathbf{k}}^{\text{mar}}(\mathbf{X}_i, \mathbf{Y}_i, \delta_i), \mathcal{F}_i\right)_{i \geq 1}$$

is a square-integrable martingale difference sequence. For $i < j$, $\Phi_{\tau,\mathbf{k}}^{\text{mar}}(\mathbf{X}_i, \mathbf{Y}_i, \delta_i)$ is $\mathcal{F}_{j-1}$ -measurable; hence,

$$\begin{aligned} &\mathbb{E}\left[\Phi_{\tau,\mathbf{k}}^{\text{mar}}(\mathbf{X}_i, \mathbf{Y}_i, \delta_i)\Phi_{\tau,\mathbf{k}}^{\text{mar}}(\mathbf{X}_j, \mathbf{Y}_j, \delta_j)\right] \\ &= \mathbb{E}\left[\Phi_{\tau,\mathbf{k}}^{\text{mar}}(\mathbf{X}_i, \mathbf{Y}_i, \delta_i)\mathbb{E}\left[\Phi_{\tau,\mathbf{k}}^{\text{mar}}(\mathbf{X}_j, \mathbf{Y}_j, \delta_j) \mid \mathcal{F}_{j-1}\right]\right] = 0. \end{aligned}$$

Consequently,

$$\begin{aligned} &\mathbb{E}\left[\left(\tilde{a}_{\tau,\mathbf{k}}^{\text{ipw}} - \bar{a}_{\tau,\mathbf{k}}^{\text{ipw}}\right)^2\right] \\ &= \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}\left[\left(\Phi_{\tau,\mathbf{k}}^{\text{mar}}(\mathbf{X}_i, \mathbf{Y}_i, \delta_i)\right)^2\right]. \end{aligned} \quad (70)$$

Conditional expectation is the orthogonal projection in L^2 , so

$$\mathbb{E}\left[(Z_{i,\tau,\mathbf{k}} - \mathbb{E}[Z_{i,\tau,\mathbf{k}} \mid \mathcal{F}_{i-1}])^2\right] \leq \mathbb{E}[Z_{i,\tau,\mathbf{k}}^2].$$

By stationarity, the latter second moment does not depend on i . Hence,

$$\mathbb{E}\left[\left(\hat{a}_{\tau,\mathbf{k}}^{\text{ipw}} - \bar{a}_{\tau,\mathbf{k}}^{\text{ipw}}\right)^2\right] \leq \frac{1}{n} \mathbb{E}[Z_{1,\tau,\mathbf{k}}^2]. \quad (71)$$

We now calculate $\mathbb{E}[Z_{1,\tau,\mathbf{k}}^2]$ with the precise moment exponent appearing in (N.0). Since $\delta_1^2 = \delta_1$, conditioning on $(\mathbf{X}_1, \mathbf{Y}_1)$ and using (M.1) gives

$$\begin{aligned} \mathbb{E}[Z_{1,\tau,\mathbf{k}}^2] &= \mathbb{E}\left[\frac{\delta_1}{\pi(\mathbf{X}_1)^2} |\rho(\mathbf{Y}_1)|^2 |D_{\tau,\mathbf{k}}(\mathbf{X}_1)|^2\right] \\ &= \mathbb{E}\left[\frac{\mathbb{E}[\delta_1 \mid \mathbf{X}_1, \mathbf{Y}_1]}{\pi(\mathbf{X}_1)^2} |\rho(\mathbf{Y}_1)|^2 |D_{\tau,\mathbf{k}}(\mathbf{X}_1)|^2\right] \\ &= \mathbb{E}\left[\frac{1}{\pi(\mathbf{X}_1)} |\rho(\mathbf{Y}_1)|^2 |D_{\tau,\mathbf{k}}(\mathbf{X}_1)|^2\right] \\ &\leq \frac{1}{\pi_0} \mathbb{E}\left[|\rho(\mathbf{Y}_1)|^2 |D_{\tau,\mathbf{k}}(\mathbf{X}_1)|^2\right]. \end{aligned} \quad (72)$$

Let

$$\gamma = \frac{\nu}{2}, \quad q' = \frac{\nu}{\nu-2}, \quad p_\nu = 2q' = \frac{2\nu}{\nu-2}.$$

Because $\nu > 2$, we have $\gamma, q' > 1$ and $\gamma^{-1} + (q')^{-1} = 1$. Therefore, Hölder's inequality gives

$$\begin{aligned} &\mathbb{E}\left[|\rho(\mathbf{Y}_1)|^2 |D_{\tau,\mathbf{k}}(\mathbf{X}_1)|^2\right] \\ &\leq \left(\mathbb{E}|\rho(\mathbf{Y}_1)|^{2\gamma}\right)^{1/\gamma} \left(\mathbb{E}|D_{\tau,\mathbf{k}}(\mathbf{X}_1)|^{2q'}\right)^{1/q'} \\ &= (\mathbb{E}|\rho(\mathbf{Y}_1)|^\nu)^{2/\nu} (\mathbb{E}|D_{\tau,\mathbf{k}}(\mathbf{X}_1)|^{p_\nu})^{(\nu-2)/\nu}. \end{aligned} \quad (73)$$

By (C.4)(i) and the additional restriction $s > d/p$, the Besov embedding $B_{s,p,q}(\mathbb{R}^d) \hookrightarrow L_\infty(\mathbb{R}^d)$ yields

$$\|f_{\mathbf{X}}\|_\infty < \infty.$$

Moreover, the r -regularity and localization of the scaling function imply $\partial^\beta \phi \in L^{p_\nu}(\mathbb{R}^d)$. Therefore,

$$\begin{aligned} \mathbb{E}|D_{\tau,\mathbf{k}}(\mathbf{X}_1)|^{p_\nu} &= \int_{\mathbb{R}^d} |D_{\tau,\mathbf{k}}(\mathbf{x})|^{p_\nu} f_{\mathbf{X}}(\mathbf{x}) \, d\mathbf{x} \\ &\leq \|f_{\mathbf{X}}\|_\infty \int_{\mathbb{R}^d} |D_{\tau,\mathbf{k}}(\mathbf{x})|^{p_\nu} \, d\mathbf{x}. \end{aligned} \quad (74)$$

Using the scaling formula for $\partial^\beta \phi_{\tau,\mathbf{k}}$,

$$\begin{aligned} \int_{\mathbb{R}^d} |D_{\tau,\mathbf{k}}(\mathbf{x})|^{p_\nu} \, d\mathbf{x} &= 2^{mp_\nu(d/2+|\beta|)} \int_{\mathbb{R}^d} |(\partial^\beta \phi)(2^m \mathbf{x} - \mathbf{k})|^{p_\nu} \, d\mathbf{x} \\ &= 2^{m\{p_\nu(d/2+|\beta|)-d\}} \int_{\mathbb{R}^d} |(\partial^\beta \phi)(\mathbf{v})|^{p_\nu} \, d\mathbf{v} \\ &= 2^{m\{p_\nu(d/2+|\beta|)-d\}} \|\partial^\beta \phi\|_{p_\nu}^{p_\nu}, \end{aligned} \quad (75)$$

where the second equality follows from $\mathbf{v} = 2^m \mathbf{x} - \mathbf{k}$ and $d\mathbf{x} = 2^{-md} d\mathbf{v}$. Combining (74) and (75),

$$\begin{aligned} & (\mathbb{E}|D_{\tau,\mathbf{k}}(\mathbf{X}_1)|^{p_v})^{(v-2)/v} \\ & \leq \|f_{\mathbf{X}}\|_{\infty}^{(v-2)/v} \|\partial^{\beta}\phi\|_{p_v}^2 2^{m\{p_v(d/2+|\beta|)-d\}(v-2)/v}. \end{aligned} \quad (76)$$

Since

$$p_v \frac{v-2}{v} = 2,$$

we have

$$\begin{aligned} \left\{ p_v \left(\frac{d}{2} + |\beta| \right) - d \right\} \frac{v-2}{v} &= 2 \left(\frac{d}{2} + |\beta| \right) - d \frac{v-2}{v} \\ &= d + 2|\beta| - d + \frac{2d}{v} \\ &= 2|\beta| + \frac{2d}{v}. \end{aligned}$$

It follows from (76) that

$$(\mathbb{E}|D_{\tau,\mathbf{k}}(\mathbf{X}_1)|^{p_v})^{(v-2)/v} \leq C_D 2^{2m(d/v+|\beta|)}, \quad (77)$$

where

$$C_D = \|f_{\mathbf{X}}\|_{\infty}^{(v-2)/v} \|\partial^{\beta}\phi\|_{p_v}^2.$$

Substitution of (77) into (73) and then into (72) gives

$$\mathbb{E}[Z_{1,\tau,\mathbf{k}}^2] \leq C_Z 2^{2m(d/v+|\beta|)}, \quad (78)$$

with

$$C_Z = \frac{1}{\pi_0} (\mathbb{E}|\rho(\mathbf{Y}_1)|^v)^{2/v} \|f_{\mathbf{X}}\|_{\infty}^{(v-2)/v} \|\partial^{\beta}\phi\|_{p_v}^2.$$

The constant C_Z is finite and independent of n , m , and $\mathbf{k}$. From (71) and (78),

$$\mathbb{E} \left[\left(\tilde{a}_{\tau,\mathbf{k}}^{\text{ipw}} - \bar{a}_{\tau,\mathbf{k}}^{\text{ipw}} \right)^2 \right] \leq C_Z \frac{2^{2m(d/v+|\beta|)}}{n}. \quad (79)$$

It remains to control the predictable term. Successively using (M.1), (N.2)(i), and the conditional density $f^{\mathcal{F}_{i-1}}$, we have almost surely

$$\begin{aligned} \mathbb{E}[Z_{i,\tau,\mathbf{k}} \mid \mathcal{F}_{i-1}] &= \mathbb{E}[\rho(\mathbf{Y}_i) D_{\tau,\mathbf{k}}(\mathbf{X}_i) \mid \mathcal{F}_{i-1}] \\ &= \mathbb{E}[m_{\rho}(\mathbf{X}_i) D_{\tau,\mathbf{k}}(\mathbf{X}_i) \mid \mathcal{F}_{i-1}] \\ &= \int_{\mathbb{R}^d} m_{\rho}(\mathbf{x}) D_{\tau,\mathbf{k}}(\mathbf{x}) f^{\mathcal{F}_{i-1}}(\mathbf{x}) d\mathbf{x}. \end{aligned} \quad (80)$$

On the other hand, by (68),

$$\mathbb{E}[Z_{i,\tau,\mathbf{k}}] = \int_{\mathbb{R}^d} m_{\rho}(\mathbf{x}) D_{\tau,\mathbf{k}}(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}.$$

Consequently,

$$\bar{a}_{\tau,\mathbf{k}}^{\text{ipw}} - a_{\tau,\mathbf{k}} = \frac{(-1)^{|\beta|}}{n} \sum_{i=1}^n \int_{\mathbb{R}^d} m_{\rho}(\mathbf{x}) D_{\tau,\mathbf{k}}(\mathbf{x}) \left\{ f^{\mathcal{F}_{i-1}}(\mathbf{x}) - f_{\mathbf{X}}(\mathbf{x}) \right\} d\mathbf{x}. \quad (81)$$

The quantitative stabilization condition (C.1[#]) directly gives

$$\mathbb{E} \left[\left(\tilde{a}_{\tau, \mathbf{k}}^{\text{ipw}} - a_{\tau, \mathbf{k}} \right)^2 \right] \leq C \frac{2^{2m(d/v + |\beta|)}}{n}. \quad (82)$$

Combining (69), (79), and (82), we obtain

$$\mathbb{E} \left[\left(\tilde{a}_{\tau, \mathbf{k}}^{\text{ipw}} - a_{\tau, \mathbf{k}} \right)^2 \right] \leq C' \frac{2^{2m(d/v + |\beta|)}}{n},$$

where C' is independent of n , m , and $\mathbf{k}$. This proves (65).

Finally, if $v \geq 4$, then $d/v \leq d/4$. Since $m \geq 0$,

$$2^{2m(d/v + |\beta|)} \leq 2^{2m(d/4 + |\beta|)}.$$

Therefore,

$$\mathbb{E} \left[\left(\tilde{a}_{\tau, \mathbf{k}}^{\text{ipw}} - a_{\tau, \mathbf{k}} \right)^2 \right] = O \left(\frac{2^{2m(d/4 + |\beta|)}}{n} \right),$$

which is (66). $\square$

7.2.2. A Stronger Sufficient Condition for (C.6)

Suppose that there exists $C_{\text{st}} > 0$ such that

$$\mathbb{E} \left[\sup_{\mathbf{x} \in \mathbb{R}^d} \left| \frac{1}{n} \sum_{i=1}^n f^{\mathcal{F}_{i-1}}(\mathbf{x}) - f_{\mathbf{X}}(\mathbf{x}) \right|^2 \right] \leq \frac{C_{\text{st}}}{n}, \quad n \geq 1. \quad (83)$$

Lemma 9. Assume (N.3)(i), the r -regularity of the multiresolution analysis with $|\beta| \leq r$, and (83). Then, for every $v > 2$, every resolution level $\tau = m(n)$, and every $\mathbf{k} \in \mathbb{Z}^d$,

$$\begin{aligned} \mathbb{E} \left[\left| \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}^d} m_{\rho}(\mathbf{x}) (\partial^{\beta} \phi_{\tau, \mathbf{k}})(\mathbf{x}) \{ f^{\mathcal{F}_{i-1}}(\mathbf{x}) - f_{\mathbf{X}}(\mathbf{x}) \} d\mathbf{x} \right|^2 \right] \\ \leq C \frac{2^{2m(|\beta| - d/2)}}{n}. \end{aligned} \quad (84)$$

Consequently,

$$\mathbb{E} \left[\left| \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}^d} m_{\rho}(\mathbf{x}) (\partial^{\beta} \phi_{\tau, \mathbf{k}})(\mathbf{x}) \{ f^{\mathcal{F}_{i-1}}(\mathbf{x}) - f_{\mathbf{X}}(\mathbf{x}) \} d\mathbf{x} \right|^2 \right] \leq C \frac{2^{2m(d/v + |\beta|)}}{n}. \quad (85)$$

Thus, under (N.3)(i), the stronger bound (83) implies the coefficient-level condition (C.6).

Proof. Put

$$\Delta_n(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^n f^{\mathcal{F}_{i-1}}(\mathbf{x}) - f_{\mathbf{X}}(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d.$$

The random quantity appearing on the left-hand side of (84) can then be written as

$$R_{\tau, \mathbf{k}, n} := \int_{\mathbb{R}^d} m_{\rho}(\mathbf{x}) (\partial^{\beta} \phi_{\tau, \mathbf{k}})(\mathbf{x}) \Delta_n(\mathbf{x}) d\mathbf{x}.$$

By the elementary inequality

$$\left| \int_{\mathbb{R}^d} u(\mathbf{x}) v(\mathbf{x}) d\mathbf{x} \right| \leq \|u\|_{L^1(\mathbb{R}^d)} \|v\|_{L^\infty(\mathbb{R}^d)}$$

applied with

$$u(\mathbf{x}) = m_\rho(\mathbf{x})(\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{x}), \quad v(\mathbf{x}) = \Delta_n(\mathbf{x}),$$

we obtain

$$|R_{\tau,\mathbf{k},n}| \leq \sup_{\mathbf{x} \in \mathbb{R}^d} |\Delta_n(\mathbf{x})| \int_{\mathbb{R}^d} |m_\rho(\mathbf{x})| |(\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{x})| d\mathbf{x}. \quad (86)$$

Under (N.3)(i), there exists $C_{\mathbf{m}} > 0$ such that

$$\sup_{\mathbf{x} \in \mathbf{J}} |m_\rho(\mathbf{x})| \leq C_{\mathbf{m}}.$$

Assuming that the support of the relevant wavelet term is contained in $\mathbf{J}$, or equivalently that the preceding bound is valid on the region on which the coefficient is evaluated, it follows that

$$\int_{\mathbb{R}^d} |m_\rho(\mathbf{x})| |(\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{x})| d\mathbf{x} \leq C_{\mathbf{m}} \|\partial^\beta \phi_{\tau,\mathbf{k}}\|_{L^1(\mathbb{R}^d)}. \quad (87)$$

By the definition of the dilated scaling functions,

$$\phi_{\tau,\mathbf{k}}(\mathbf{x}) = 2^{md/2} \phi(2^m \mathbf{x} - \mathbf{k});$$

therefore,

$$(\partial^\beta \phi_{\tau,\mathbf{k}})(\mathbf{x}) = 2^{m(d/2+|\beta|)} (\partial^\beta \phi)(2^m \mathbf{x} - \mathbf{k}).$$

Consequently,

$$\|\partial^\beta \phi_{\tau,\mathbf{k}}\|_{L^1(\mathbb{R}^d)} = 2^{m(d/2+|\beta|)} \int_{\mathbb{R}^d} |(\partial^\beta \phi)(2^m \mathbf{x} - \mathbf{k})| d\mathbf{x}. \quad (88)$$

Making the change of variables

$$\mathbf{v} = 2^m \mathbf{x} - \mathbf{k}, \quad \mathbf{x} = 2^{-m}(\mathbf{v} + \mathbf{k}), \quad d\mathbf{x} = 2^{-md} d\mathbf{v},$$

we obtain

$$\begin{aligned} \|\partial^\beta \phi_{\tau,\mathbf{k}}\|_{L^1(\mathbb{R}^d)} &= 2^{m(d/2+|\beta|)} 2^{-md} \int_{\mathbb{R}^d} |(\partial^\beta \phi)(\mathbf{v})| d\mathbf{v} \\ &= 2^{m(|\beta|-d/2)} \|\partial^\beta \phi\|_{L^1(\mathbb{R}^d)}. \end{aligned} \quad (89)$$

The r -regularity and rapid decrease of the scaling function imply

$$\|\partial^\beta \phi\|_{L^1(\mathbb{R}^d)} < \infty$$

for every $|\beta| \leq r$. Combining (86), (87), and (89), we obtain the pathwise inequality

$$|R_{\tau,\mathbf{k},n}| \leq C_{\mathbf{m}} 2^{m(|\beta|-d/2)} \|\partial^\beta \phi\|_1 \sup_{\mathbf{x} \in \mathbb{R}^d} |\Delta_n(\mathbf{x})|. \quad (90)$$

After squaring,

$$|R_{\tau,\mathbf{k},n}|^2 \leq C_{\mathbf{m}}^2 2^{2m(|\beta|-d/2)} \|\partial^\beta \phi\|_1^2 \sup_{\mathbf{x} \in \mathbb{R}^d} |\Delta_n(\mathbf{x})|^2. \quad (91)$$

Taking expectations and applying (C.2*) gives

$$\begin{aligned}\mathbb{E}|R_{\tau, \mathbf{k}, n}|^2 &\leq C_{\mathbf{m}}^2 2^{2m(|\beta|-d/2)} \left\| \partial^\beta \phi \right\|_1^2 \mathbb{E} \left[\sup_{\mathbf{x} \in \mathbb{R}^d} |\Delta_n(\mathbf{x})|^2 \right] \\ &\leq C_{\mathbf{m}}^2 C_{\text{st}} \left\| \partial^\beta \phi \right\|_1^2 \frac{2^{2m(|\beta|-d/2)}}{n}.\end{aligned}\quad (92)$$

This proves (84).

It remains to compare the exponent in (84) with the exponent required in (C.1[#]). Since $\nu > 2$ and $d \geq 1$,

$$-\frac{d}{2} \leq \frac{d}{\nu}.$$

Therefore,

$$|\beta| - \frac{d}{2} \leq |\beta| + \frac{d}{\nu}.$$

Because $m \geq 0$,

$$2^{2m(|\beta|-d/2)} \leq 2^{2m(|\beta|+d/\nu)}.$$

It follows from (92) that

$$\mathbb{E}|R_{\tau, \mathbf{k}, n}|^2 \leq C \frac{2^{2m(|\beta|+d/\nu)}}{n},$$

which is precisely the estimate required in (C.6). $\square$

7.2.3. Conditions Used for Uniform Convergence

Throughout the proof of Theorem 2, assumptions (U.1)–(U.5) are understood in the form stated in Section 3. In particular, all references to (21)–(27) refer to the labels defined there.

7.3. Proof of Theorem 2

We first consider the IPW pseudo-estimator based on the true propensity score:

$$\widetilde{(\partial^\beta r)}_n^{\text{ipw}}(\rho; \mathbf{x}) = \frac{(-1)^{|\beta|}}{n h_n^{d+|\beta|}} \sum_{i=1}^n \frac{\delta_i}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) K^{(\beta)}\left(\frac{\mathbf{x}}{h_n}, \frac{\mathbf{X}_i}{h_n}\right).$$

Define its predictable counterpart by

$$\bar{r}_n^{\text{ipw}}(\rho; \mathbf{x}) = \frac{(-1)^{|\beta|}}{n h_n^{d+|\beta|}} \sum_{i=1}^n \mathbb{E} \left[\frac{\delta_i}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) K^{(\beta)}\left(\frac{\mathbf{x}}{h_n}, \frac{\mathbf{X}_i}{h_n}\right) \middle| \mathcal{F}_{i-1} \right].$$

Then,

$$\begin{aligned}\sup_{\mathbf{x} \in D} \left| \widetilde{(\partial^\beta r)}_n^{\text{ipw}}(\rho; \mathbf{x}) - \mathbb{E} \left[\widetilde{(\partial^\beta r)}_n^{\text{ipw}}(\rho; \mathbf{x}) \right] \right| \\ \leq G_{n,1}^{\text{ipw}}(\rho) + G_{n,2}^{\text{ipw}}(\rho),\end{aligned}\quad (93)$$

where

$$G_{n,1}^{\text{ipw}}(\rho) = \sup_{\mathbf{x} \in D} \left| \widetilde{(\partial^\beta r)}_n^{\text{ipw}}(\rho; \mathbf{x}) - \bar{r}_n^{\text{ipw}}(\rho; \mathbf{x}) \right|$$

and

$$G_{n,2}^{\text{ipw}}(\rho) = \sup_{\mathbf{x} \in D} \left| \bar{r}_n^{\text{ipw}}(\rho; \mathbf{x}) - \mathbb{E} \left[\widetilde{(\partial^\beta r)}_n^{\text{ipw}}(\rho; \mathbf{x}) \right] \right|.$$

Let

$$\rho_{i,n}^T = \rho(\mathbf{Y}_i) \mathbb{1}_{\{|\rho(\mathbf{Y}_i)| \leq T_n\}}, \quad \rho_{i,n}^R = \rho(\mathbf{Y}_i) \mathbb{1}_{\{|\rho(\mathbf{Y}_i)| > T_n\}}.$$

Define

$$\tilde{a}_{\tau, \mathbf{k}}^{T, \text{ipw}} = \frac{(-1)^{|\beta|}}{n} \sum_{i=1}^n \frac{\delta_i}{\pi(\mathbf{X}_i)} \rho_{i,n}^T (\partial^\beta \phi_{\tau, \mathbf{k}})(\mathbf{X}_i) \quad (94)$$

and

$$\widetilde{(\partial^\beta r)}_n^{T, \text{ipw}}(\rho; \mathbf{x}) = \sum_{\mathbf{k} \in \mathbb{Z}^d} \tilde{a}_{\tau, \mathbf{k}}^{T, \text{ipw}} \phi_{\tau, \mathbf{k}}(\mathbf{x}). \quad (95)$$

Equivalently,

$$\widetilde{(\partial^\beta r)}_n^{T, \text{ipw}}(\rho; \mathbf{x}) = \frac{(-1)^{|\beta|}}{nh_n^{d+|\beta|}} \sum_{i=1}^n \frac{\delta_i}{\pi(\mathbf{X}_i)} \rho_{i,n}^{T, K(\beta)} \left(\frac{\mathbf{x}}{h_n}, \frac{\mathbf{X}_i}{h_n} \right). \quad (96)$$

Likewise, put

$$\bar{r}_n^{T, \text{ipw}}(\rho; \mathbf{x}) = \frac{(-1)^{|\beta|}}{nh_n^{d+|\beta|}} \sum_{i=1}^n \mathbb{E} \left[\frac{\delta_i}{\pi(\mathbf{X}_i)} \rho_{i,n}^{T, K(\beta)} \left(\frac{\mathbf{x}}{h_n}, \frac{\mathbf{X}_i}{h_n} \right) \middle| \mathcal{F}_{i-1} \right]. \quad (97)$$

We have

$$G_{n,1}^{\text{ipw}}(\rho) \leq W_{n,1}^{\text{ipw}}(\rho) + W_{n,2}^{\text{ipw}}(\rho) + W_{n,3}^{\text{ipw}}(\rho),$$

where

$$\begin{aligned} W_{n,1}^{\text{ipw}}(\rho) &= \sup_{\mathbf{x} \in D} \left| \widetilde{(\partial^\beta r)}_n^{\text{ipw}}(\rho; \mathbf{x}) - \widetilde{(\partial^\beta r)}_n^{T, \text{ipw}}(\rho; \mathbf{x}) \right|, \\ W_{n,2}^{\text{ipw}}(\rho) &= \sup_{\mathbf{x} \in D} \left| \widetilde{(\partial^\beta r)}_n^{T, \text{ipw}}(\rho; \mathbf{x}) - \bar{r}_n^{T, \text{ipw}}(\rho; \mathbf{x}) \right|, \\ W_{n,3}^{\text{ipw}}(\rho) &= \sup_{\mathbf{x} \in D} \left| \bar{r}_n^{T, \text{ipw}}(\rho; \mathbf{x}) - \bar{r}_n^{\text{ipw}}(\rho; \mathbf{x}) \right|. \end{aligned}$$

We first show that the truncation remainder is eventually zero. By the union bound, stationarity, Markov's inequality, and (N.0),

$$\begin{aligned} \mathbb{P} \left(\max_{1 \leq i \leq n} |\rho(\mathbf{Y}_i)| > T_n \right) &\leq \sum_{i=1}^n \mathbb{P}(|\rho(\mathbf{Y}_i)| > T_n) \\ &= n \mathbb{P}(|\rho(\mathbf{Y}_1)| > T_n) \\ &\leq n T_n^{-\nu} \mathbb{E} |\rho(\mathbf{Y}_1)|^\nu. \end{aligned}$$

Condition (26) and the Borel–Cantelli lemma imply

$$\max_{1 \leq i \leq n} |\rho(\mathbf{Y}_i)| \leq T_n \quad \text{a.s.}$$

for all sufficiently large n . Consequently,

$$W_{n,1}^{\text{ipw}}(\rho) = 0 \quad \text{a.s.} \quad (98)$$

eventually. The same event implies

$$W_{n,3}^{\text{ipw}}(\rho) = 0 \quad \text{a.s.} \quad (99)$$

eventually. This avoids the invalid inference obtained from the weaker condition $\sum_n T_n^{-\nu} < \infty$, which controls only $|\rho(\mathbf{Y}_n)|$, not the maximum over $1 \leq i \leq n$. It remains to control $W_{n,2}^{\text{ipw}}(\rho)$. Let

$$r_n = \left(\frac{\log n}{nh_n^{d+2|\beta|}} \right)^{1/2}.$$

Cover D by L_n closed cubes of side length ℓ_n , with centers $\mathbf{x}_{n,1}, \dots, \mathbf{x}_{n,L_n}$, where

$$\ell_n = c_0 \frac{h_n^{d+|\beta|+1} r_n}{T_n}, \quad L_n \leq C_D \ell_n^{-d}.$$

Then,

$$W_{n,2}^{\text{ipw}}(\rho) \leq Q_1^{\text{ipw}} + Q_2^{\text{ipw}} + Q_3^{\text{ipw}},$$

where

$$\begin{aligned} Q_1^{\text{ipw}} &= \max_{1 \leq j \leq L_n} \sup_{\mathbf{x} \in D \cap I_{n,j}} \left| \widetilde{(\partial \beta r)_n}^{T, \text{ipw}}(\rho; \mathbf{x}) - \widetilde{(\partial \beta r)_n}^{T, \text{ipw}}(\rho; \mathbf{x}_{n,j}) \right|, \\ Q_2^{\text{ipw}} &= \max_{1 \leq j \leq L_n} \left| \widetilde{(\partial \beta r)_n}^{T, \text{ipw}}(\rho; \mathbf{x}_{n,j}) - \bar{r}_n^{T, \text{ipw}}(\rho; \mathbf{x}_{n,j}) \right|, \\ Q_3^{\text{ipw}} &= \max_{1 \leq j \leq L_n} \sup_{\mathbf{x} \in D \cap I_{n,j}} \left| \bar{r}_n^{T, \text{ipw}}(\rho; \mathbf{x}_{n,j}) - \bar{r}_n^{T, \text{ipw}}(\rho; \mathbf{x}) \right|. \end{aligned}$$

By (22), $\pi(\mathbf{X}_i) \geq \pi_0$, and $|\rho_{i,n}^T| \leq T_n$,

$$Q_1^{\text{ipw}} \leq \frac{CT_n \ell_n}{h_n^{d+|\beta|+1}} = O(r_n) \quad \text{a.s.}$$

The conditional-expectation contraction gives the same bound:

$$Q_3^{\text{ipw}} = O(r_n) \quad \text{a.s.} \quad (100)$$

For $1 \leq j \leq L_n$, define

$$\begin{aligned} Z_{n,i,j} &= (-1)^{|\beta|} \left[\frac{\delta_i}{\pi(\mathbf{X}_i)} \rho_{i,n}^T K^{(\beta)} \left(\frac{\mathbf{x}_{n,j}}{h_n}, \frac{\mathbf{X}_i}{h_n} \right) \right. \\ &\quad \left. - \mathbb{E} \left[\frac{\delta_i}{\pi(\mathbf{X}_i)} \rho_{i,n}^T K^{(\beta)} \left(\frac{\mathbf{x}_{n,j}}{h_n}, \frac{\mathbf{X}_i}{h_n} \right) \middle| \mathcal{F}_{i-1} \right] \right]. \end{aligned} \quad (101)$$

For fixed n, j , $\{Z_{n,i,j}, \mathcal{F}_i\}_{1 \leq i \leq n}$ is a martingale difference array and

$$Q_2^{\text{ipw}} = \max_{1 \leq j \leq L_n} \left| \frac{1}{nh_n^{d+|\beta|}} \sum_{i=1}^n Z_{n,i,j} \right|.$$

The truncation and positivity imply the deterministic increment bound

$$|Z_{n,i,j}| \leq \frac{2C_K T_n}{\pi_0} =: B_n. \quad (102)$$

Furthermore, by conditional Jensen's inequality,

$$\mathbb{E}[Z_{n,i,j}^2 \mid \mathcal{F}_{i-1}] \leq \mathbb{E} \left[\left(\frac{\delta_i}{\pi(\mathbf{X}_i)} \rho_{i,n}^T K^{(\beta)} \left(\frac{\mathbf{x}_{n,j}}{h_n}, \frac{\mathbf{X}_i}{h_n} \right) \right)^2 \middle| \mathcal{F}_{i-1} \right].$$

Since $\delta_i^2 = \delta_i$, MAR and positivity yield

$$\begin{aligned} & \mathbb{E} \left[\left(\frac{\delta_i}{\pi(\mathbf{X}_i)} \rho_{i,n}^T K^{(\beta)} \left(\frac{\mathbf{x}_{n,j}}{h_n}, \frac{\mathbf{X}_i}{h_n} \right) \right)^2 \middle| \mathcal{F}_{i-1} \right] \\ & \leq \frac{1}{\pi_0} \mathbb{E} \left[|\rho(\mathbf{Y}_i)|^2 \left| K^{(\beta)} \left(\frac{\mathbf{x}_{n,j}}{h_n}, \frac{\mathbf{X}_i}{h_n} \right) \right|^2 \middle| \mathcal{F}_{i-1} \right]. \end{aligned}$$

By (U.2), conditional Hölder's inequality, (U.3), and the kernel integrability bound,

$$\begin{aligned} \mathbb{E}[Z_{n,i,j}^2 \mid \mathcal{F}_{i-1}] & \leq C \int_{\mathbb{R}^d} \left| K^{(\beta)} \left(\frac{\mathbf{x}_{n,j}}{h_n}, \frac{\mathbf{y}}{h_n} \right) \right|^2 f^{\mathcal{F}_{i-1}}(\mathbf{y}) d\mathbf{y} \\ & \leq Ch_n^d \int_{\mathbb{R}^d} |K^{(\beta)}(\mathbf{0}, \mathbf{u})|^2 d\mathbf{u} \\ & \leq Ch_n^d \quad \text{a.s.} \end{aligned}$$

Hence, the predictable quadratic variation satisfies

$$V_{n,j} := \sum_{i=1}^n \mathbb{E}[Z_{n,i,j}^2 \mid \mathcal{F}_{i-1}] \leq Cnh_n^d \quad \text{a.s.}$$

Then, for every $t > 0$, Freedman's inequality gives

$$\mathbb{P} \left(\left| \sum_{i=1}^n Z_{n,i,j} \right| > t \right) \leq 2 \exp \left\{ - \frac{t^2}{2(Cnh_n^d + B_n t/3)} \right\}.$$

Choose

$$t_n = Anh_n^{d+|\beta|} r_n = A\sqrt{nh_n^d \log n}.$$

By (25),

$$\frac{B_n t_n}{nh_n^d} \leq CT_n \left(\frac{\log n}{nh_n^d} \right)^{1/2} \longrightarrow 0.$$

Thus, for all sufficiently large n ,

$$\mathbb{P} \left(\left| \sum_{i=1}^n Z_{n,i,j} \right| > t_n \right) \leq 2n^{-cA^2}$$

for some $c > 0$. By the union bound,

$$\mathbb{P}(Q_2^{\text{ipw}} > Ar_n) \leq 2L_n n^{-cA^2}.$$

The definition of ℓ_n together with (24)–(25) implies that L_n grows at most polynomially. Choosing sufficiently large A makes $\sum_n L_n n^{-cA^2} < \infty$. Borel–Cantelli then yields

$$Q_2^{\text{ipw}} = O(r_n) \quad \text{a.s.} \quad (103)$$

Combining (98)–(100), and (103), we obtain

$$G_{n,1}^{\text{ipw}}(\rho) = O\left[\left(\frac{\log n}{nh_n^{d+2|\beta|}}\right)^{1/2}\right] \quad \text{a.s.} \quad (104)$$

We next control the predictable component. By the MAR identity and (N.2)(i),

$$\mathbb{E}\left[\frac{\delta_i}{\pi(\mathbf{X}_i)}\rho(\mathbf{Y}_i)\middle|\mathcal{G}_{i-1}\right] = m_\rho(\mathbf{X}_i) \quad \text{a.s.}$$

Consequently,

$$G_{n,2}^{\text{ipw}}(\rho) = \sup_{\mathbf{x} \in D} \left| \frac{(-1)^{|\beta|}}{h_n^{d+|\beta|}} \int_{\mathbb{R}^d} m_\rho(\mathbf{y}) K^{(\beta)}\left(\frac{\mathbf{x}}{h_n}, \frac{\mathbf{y}}{h_n}\right) \Delta_n(\mathbf{y}) d\mathbf{y} \right|. \quad (105)$$

Using (N.3)(i), the predictable bound used to verify (41), and the L^1 -kernel bound, we find

$$\begin{aligned} G_{n,2}^{\text{ipw}}(\rho) &\leq \frac{C_{\mathbf{m}} \|\Delta_n\|_\infty}{h_n^{d+|\beta|}} \sup_{\mathbf{x} \in D} \int_{\mathbb{R}^d} \left| K^{(\beta)}\left(\frac{\mathbf{x}}{h_n}, \frac{\mathbf{y}}{h_n}\right) \right| d\mathbf{y} \\ &= \frac{C_{\mathbf{m}} \|\Delta_n\|_\infty}{h_n^{|\beta|}} \sup_{\mathbf{u} \in \mathbb{R}^d} \int_{\mathbb{R}^d} |K^{(\beta)}(\mathbf{u}, \mathbf{v})| d\mathbf{v} \\ &= O(b_n h_n^{-|\beta|}) \quad \text{a.s.} \end{aligned}$$

Therefore,

$$\begin{aligned} &\sup_{\mathbf{x} \in D} \left| \widetilde{(\partial\beta r)}_n^{\text{ipw}}(\rho; \mathbf{x}) - \mathbb{E}\left[\widetilde{(\partial\beta r)}_n^{\text{ipw}}(\rho; \mathbf{x})\right] \right| \\ &= O\left[\left(\frac{\log n}{nh_n^{d+2|\beta|}}\right)^{1/2}\right] + O(b_n h_n^{-|\beta|}) \quad \text{a.s.} \end{aligned} \quad (106)$$

We finally pass to the feasible estimator. We have

$$\begin{aligned} &\widetilde{(\partial\beta r)}_n^{\text{mar}}(\rho; \mathbf{x}) - \widetilde{(\partial\beta r)}_n^{\text{ipw}}(\rho; \mathbf{x}) \\ &= \frac{(-1)^{|\beta|}}{nh_n^{d+|\beta|}} \sum_{i=1}^n \delta_i \rho(\mathbf{Y}_i) K^{(\beta)}\left(\frac{\mathbf{x}}{h_n}, \frac{\mathbf{X}_i}{h_n}\right) \left[\frac{1}{\widehat{\pi}_n(\mathbf{X}_i)} - \frac{1}{\pi(\mathbf{X}_i)} \right]. \end{aligned} \quad (107)$$

By (M.2) and (M.4), for all sufficiently large n ,

$$\sup_{\mathbf{x} \in \mathbf{J}} \left| \frac{1}{\widehat{\pi}_n(\mathbf{x})} - \frac{1}{\pi(\mathbf{x})} \right| \leq \frac{2}{\pi_0^2} \|\widehat{\pi}_n - \pi\|_{\infty, \mathbf{J}} = O(a_n) \quad \text{a.s.}$$

Therefore,

$$\begin{aligned} &\sup_{\mathbf{x} \in D} \left| \widetilde{(\partial\beta r)}_n^{\text{mar}}(\rho; \mathbf{x}) - \widetilde{(\partial\beta r)}_n^{\text{ipw}}(\rho; \mathbf{x}) \right| \\ &\leq C a_n \sup_{\mathbf{x} \in D} \frac{1}{nh_n^{d+|\beta|}} \sum_{i=1}^n \delta_i |\rho(\mathbf{Y}_i)| \left| K^{(\beta)}\left(\frac{\mathbf{x}}{h_n}, \frac{\mathbf{X}_i}{h_n}\right) \right|. \end{aligned} \quad (108)$$

Under the same conditional moment, density, and kernel envelope bounds used above, the last supremum is $O(h_n^{-|\beta|})$ almost surely. Hence,

$$\sup_{\mathbf{x} \in D} \left| \widehat{(\partial \beta r)}_n^{\text{mar}}(\rho; \mathbf{x}) - \widehat{(\partial \beta r)}_n^{\text{ipw}}(\rho; \mathbf{x}) \right| = O(a_n h_n^{-|\beta|}) \quad \text{a.s.}$$

Combining this estimate with (106) gives

$$\begin{aligned} & \sup_{\mathbf{x} \in D} \left| \widehat{(\partial \beta r)}_n^{\text{mar}}(\rho; \mathbf{x}) - \mathbb{E} \left[\widehat{(\partial \beta r)}_n^{\text{ipw}}(\rho; \mathbf{x}) \right] \right| \\ &= O \left[\left(\frac{\log n}{n h_n^{d+2|\beta|}} \right)^{1/2} \right] + O(b_n h_n^{-|\beta|}) + O(a_n h_n^{-|\beta|}) \quad \text{a.s.} \end{aligned}$$

The propensity estimation term is asymptotically negligible relative to the oracle stochastic and predictable terms. Adding the deterministic projection bias from Lemma 3 completes the proof of Theorem 2. $\square$

Conditions Used for Pointwise Asymptotic Normality

Throughout the proof of Theorem 3, assumptions (CLT.1)–(CLT.8) are understood in the form stated in Section 3. All references to (28)–(36) refer to the labels defined there.

7.4. Proof of Theorem 3

By (CLT.1), the dyadic phase $\boldsymbol{\vartheta}_n$ converges to $\boldsymbol{\vartheta} \in [0, 1]^d$. Therefore, the proof identifies the phase-dependent variance in (43). If the kernel energy is phase-invariant, then the same variance applies along the full sequence; otherwise, the conclusion is understood along sequences satisfying (28), exactly as stated in Theorem 3. For notational convenience, we write

$$K_{n,i}(\mathbf{x}) := K^{(\beta)} \left(\frac{\mathbf{x}}{h_n}, \frac{\mathbf{X}_i}{h_n} \right)$$

and

$$g_\beta(\mathbf{x}) := (\partial \beta r)(\rho; \mathbf{x}).$$

Recall that

$$\widehat{(\partial \beta r)}_n^{\text{ipw}}(\rho; \mathbf{x}) = \frac{(-1)^{|\beta|}}{n h_n^{d+|\beta|}} \sum_{i=1}^n \frac{\delta_i}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) K_{n,i}(\mathbf{x})$$

and

$$\bar{r}_n^{\text{ipw}}(\rho; \mathbf{x}) = \frac{(-1)^{|\beta|}}{n h_n^{d+|\beta|}} \sum_{i=1}^n \mathbb{E} \left[\frac{\delta_i}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) K_{n,i}(\mathbf{x}) \middle| \mathcal{F}_{i-1} \right].$$

We use the exact decomposition

$$\begin{aligned} & \sqrt{n h_n^{d+2|\beta|}} \left[\widehat{(\partial \beta r)}_n^{\text{mar}}(\rho; \mathbf{x}) - g_\beta(\mathbf{x}) \right] \\ &= S_n(\mathbf{x}) + R_{n,1}(\mathbf{x}) + R_{n,2}(\mathbf{x}), \end{aligned} \quad (109)$$

where

$$\begin{aligned} S_n(\mathbf{x}) &= \sqrt{n h_n^{d+2|\beta|}} \left[\widehat{(\partial \beta r)}_n^{\text{ipw}}(\rho; \mathbf{x}) - \bar{r}_n^{\text{ipw}}(\rho; \mathbf{x}) \right], \\ R_{n,1}(\mathbf{x}) &= \sqrt{n h_n^{d+2|\beta|}} \left[\bar{r}_n^{\text{ipw}}(\rho; \mathbf{x}) - g_\beta(\mathbf{x}) \right], \\ R_{n,2}(\mathbf{x}) &= \sqrt{n h_n^{d+2|\beta|}} \left[\widehat{(\partial \beta r)}_n^{\text{mar}}(\rho; \mathbf{x}) - \widehat{(\partial \beta r)}_n^{\text{ipw}}(\rho; \mathbf{x}) \right]. \end{aligned}$$

We first identify the deterministic and predictable components of $R_{n,1}(\mathbf{x})$. By adding and subtracting the unconditional expectation,

$$\bar{r}_n^{\text{ipw}}(\rho; \mathbf{x}) - g_\beta(\mathbf{x}) = B_{n,1}^{\text{ipw}}(\mathbf{x}) + B_{n,2}^{\text{ipw}}(\mathbf{x}),$$

where

$$B_{n,1}^{\text{ipw}}(\mathbf{x}) = \bar{r}_n^{\text{ipw}}(\rho; \mathbf{x}) - \mathbb{E} \left[\widetilde{(\partial \beta r)}_n^{\text{ipw}}(\rho; \mathbf{x}) \right]$$

and

$$B_{n,2}^{\text{ipw}}(\mathbf{x}) = \mathbb{E} \left[\widetilde{(\partial \beta r)}_n^{\text{ipw}}(\rho; \mathbf{x}) \right] - g_\beta(\mathbf{x}).$$

By (M.1) and (N.2)(i),

$$\mathbb{E} \left[\frac{\delta_i}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) \middle| \mathcal{G}_{i-1} \right] = m_\rho(\mathbf{X}_i) \quad \text{a.s.}$$

Since the conditional law of $\mathbf{X}_i$ given $\mathcal{F}_{i-1}$ has density $f^{\mathcal{F}_{i-1}}$, it follows that

$$\begin{aligned} B_{n,1}^{\text{ipw}}(\mathbf{x}) &= \frac{(-1)^{|\beta|}}{h_n^{d+|\beta|}} \int_{\mathbb{R}^d} m_\rho(\mathbf{u}) K^{(\beta)} \left(\frac{\mathbf{x}}{h_n}, \frac{\mathbf{u}}{h_n} \right) \\ &\quad \times \left[\frac{1}{n} \sum_{i=1}^n f^{\mathcal{F}_{i-1}}(\mathbf{u}) - f_{\mathbf{X}}(\mathbf{u}) \right] d\mathbf{u}. \end{aligned} \quad (110)$$

Therefore, Condition (34) gives

$$\sqrt{nh_n^{d+2|\beta|}} B_{n,1}^{\text{ipw}}(\mathbf{x}) = o_{\mathbb{P}}(1). \quad (111)$$

Notice that qualitative convergence in (C.2) alone does not imply this negligibility of the central limit scale. By the exact IPW compensation identity,

$$\begin{aligned} \mathbb{E} \left[\widetilde{(\partial \beta r)}_n^{\text{ipw}}(\rho; \mathbf{x}) \right] &= \frac{(-1)^{|\beta|}}{h_n^{d+|\beta|}} \int_{\mathbb{R}^d} r(\rho; \mathbf{u}) K^{(\beta)} \left(\frac{\mathbf{x}}{h_n}, \frac{\mathbf{u}}{h_n} \right) d\mathbf{u} \\ &= P_{V_{m(n)}} g_\beta(\mathbf{x}). \end{aligned} \quad (112)$$

The second equality follows from the projection identity and integration by parts. The latter is justified by the weak differentiability of $r(\rho; \cdot)$, the compact support assumptions in (C.5), and the regularity of the scaling function; compact support of the scaling function alone would not justify moving derivatives from the target. Hence,

$$B_{n,2}^{\text{ipw}}(\mathbf{x}) = P_{V_{m(n)}} g_\beta(\mathbf{x}) - g_\beta(\mathbf{x}),$$

and (35) gives

$$\sqrt{nh_n^{d+2|\beta|}} B_{n,2}^{\text{ipw}}(\mathbf{x}) = o(1). \quad (113)$$

Combining (111) and (113), we obtain

$$R_{n,1}(\mathbf{x}) = o_{\mathbb{P}}(1). \quad (114)$$

We next prove a martingale central limit theorem for $S_n(\mathbf{x})$. Set

$$\zeta_{ni}^{\text{ipw}}(\mathbf{x}) = \frac{(-1)^{|\beta|}}{\sqrt{nh_n^d}} \frac{\delta_i}{\pi(\mathbf{X}_i)} \rho(\mathbf{Y}_i) K_{n,i}(\mathbf{x})$$

and

$$\chi_{ni}^{\text{ipw}}(\mathbf{x}) = \xi_{ni}^{\text{ipw}}(\mathbf{x}) - \mathbb{E}[\xi_{ni}^{\text{ipw}}(\mathbf{x}) | \mathcal{F}_{i-1}].$$

Then,

$$S_n(\mathbf{x}) = \sum_{i=1}^n \chi_{ni}^{\text{ipw}}(\mathbf{x})$$

and $\{\chi_{ni}^{\text{ipw}}(\mathbf{x}), \mathcal{F}_i\}_{1 \leq i \leq n}$ is a martingale difference triangular array. The correct candidate variance is phase-dependent:

$$\Sigma_{\text{mar},(\beta)}^2(\mathbf{x}; \boldsymbol{\theta}) = \frac{m_{\rho,2}(\mathbf{x})f_{\mathbf{X}}(\mathbf{x})}{\pi(\mathbf{x})} \int_{\mathbb{R}^d} [K^{(\beta)}(\boldsymbol{\theta}, \boldsymbol{\theta} + \mathbf{v})]^2 d\mathbf{v}. \quad (115)$$

We verify the conditional variance convergence

$$\sum_{i=1}^n \mathbb{E}[(\chi_{ni}^{\text{ipw}}(\mathbf{x}))^2 | \mathcal{F}_{i-1}] \xrightarrow{\mathbb{P}} \Sigma_{\text{mar},(\beta)}^2(\mathbf{x}; \boldsymbol{\theta}). \quad (116)$$

Because conditional expectation is an L^2 -contraction,

$$\mathbb{E}[(\chi_{ni}^{\text{ipw}})^2 | \mathcal{F}_{i-1}] = \mathbb{E}[(\xi_{ni}^{\text{ipw}})^2 | \mathcal{F}_{i-1}] - \left(\mathbb{E}[\xi_{ni}^{\text{ipw}} | \mathcal{F}_{i-1}]\right)^2.$$

The predictable quadratic variation hypothesis in Theorem 3 requires the squared predictable means to be negligible. Under the local sufficient conditions above, the same localization calculation used for (34) gives

$$\sum_{i=1}^n \left(\mathbb{E}[\xi_{ni}^{\text{ipw}}(\mathbf{x}) | \mathcal{F}_{i-1}]\right)^2 = o_{\mathbb{P}}(1). \quad (117)$$

This is a separate quadratic estimate, and does not follow merely by squaring the aggregate centering bound. Thus, it remains to calculate the first conditional second-moment term. Since $\delta_i^2 = \delta_i$, (M.1) and (N.2)(ii) give

$$\mathbb{E}\left[\frac{\delta_i^2}{\pi(\mathbf{X}_i)^2} \rho(\mathbf{Y}_i)^2 \middle| \mathcal{G}_{i-1}\right] = \frac{m_{\rho,2}(\mathbf{X}_i)}{\pi(\mathbf{X}_i)} \quad \text{a.s.} \quad (118)$$

Consequently,

$$\begin{aligned} & \sum_{i=1}^n \mathbb{E}[(\xi_{ni}^{\text{ipw}}(\mathbf{x}))^2 | \mathcal{F}_{i-1}] \\ &= \frac{1}{nh_n^d} \sum_{i=1}^n \int_{\mathbb{R}^d} \frac{m_{\rho,2}(\mathbf{u})}{\pi(\mathbf{u})} \left[K^{(\beta)}\left(\frac{\mathbf{x}}{h_n}, \frac{\mathbf{u}}{h_n}\right)\right]^2 f^{\mathcal{F}_{i-1}}(\mathbf{u}) d\mathbf{u}. \end{aligned} \quad (119)$$

Set

$$\mathbf{k}_n = \left\lfloor \frac{\mathbf{x}}{h_n} \right\rfloor \in \mathbb{Z}^d, \quad \boldsymbol{\theta}_n = \frac{\mathbf{x}}{h_n} - \mathbf{k}_n.$$

With the change of variables $\mathbf{u} = \mathbf{x} + h_n \mathbf{v}$ and using the integer-shift periodicity

$$K^{(\beta)}(\mathbf{a} + \mathbf{k}, \mathbf{b} + \mathbf{k}) = K^{(\beta)}(\mathbf{a}, \mathbf{b}), \quad \mathbf{k} \in \mathbb{Z}^d,$$

we obtain

$$\sum_{i=1}^n \mathbb{E}[(\xi_{ni}^{\text{ipw}}(\mathbf{x}))^2 | \mathcal{F}_{i-1}]$$

$$= \int_{\mathbb{R}^d} \frac{m_{\rho,2}(\mathbf{x} + h_n \mathbf{v})}{\pi(\mathbf{x} + h_n \mathbf{v})} \left[K^{(\beta)}(\boldsymbol{\vartheta}_n, \boldsymbol{\vartheta}_n + \mathbf{v}) \right]^2 \quad (120)$$

$$\times \left[\frac{1}{n} \sum_{i=1}^n f^{\mathcal{F}_{i-1}}(\mathbf{x} + h_n \mathbf{v}) \right] d\mathbf{v}. \quad (121)$$

For each fixed $\mathbf{v}$, conditions (28) and (29) together with continuity at $\mathbf{x}$ imply convergence in probability of the integrand to

$$\frac{m_{\rho,2}(\mathbf{x})f_{\mathbf{X}}(\mathbf{x})}{\pi(\mathbf{x})} \left[K^{(\beta)}(\boldsymbol{\vartheta}, \boldsymbol{\vartheta} + \mathbf{v}) \right]^2.$$

The envelope assumptions (30) and (31) imply uniform integrability of the integrands. Truncating first to $\{\|\mathbf{v}\| \leq M\}$, applying dominated convergence there, and then letting $M \rightarrow \infty$ yields

$$\sum_{i=1}^n \mathbb{E} \left[(\xi_{ni}^{\text{ipw}}(\mathbf{x}))^2 \middle| \mathcal{F}_{i-1} \right] \xrightarrow{\mathbb{P}} \Sigma_{\text{mar},(\beta)}^2(\mathbf{x}; \boldsymbol{\vartheta}). \quad (122)$$

Together with (117), this proves (116). We now verify the following conditional Lindeberg condition: for every $\varepsilon > 0$,

$$\sum_{i=1}^n \mathbb{E} \left[(\chi_{ni}^{\text{ipw}}(\mathbf{x}))^2 \mathbf{1}_{\{|\chi_{ni}^{\text{ipw}}(\mathbf{x})| > \varepsilon\}} \middle| \mathcal{F}_{i-1} \right] \xrightarrow{\mathbb{P}} 0. \quad (123)$$

Choose $\eta \in (0, \nu - 2)$. The elementary inequality

$$|u - v|^{2+\eta} \leq 2^{1+\eta} (|u|^{2+\eta} + |v|^{2+\eta})$$

and conditional Jensen's inequality imply

$$\mathbb{E} \left[|\chi_{ni}^{\text{ipw}}|^{2+\eta} \middle| \mathcal{F}_{i-1} \right] \leq 2^{2+\eta} \mathbb{E} \left[|\xi_{ni}^{\text{ipw}}|^{2+\eta} \middle| \mathcal{F}_{i-1} \right].$$

Hence, conditional Markov's inequality gives

$$\begin{aligned} \sum_{i=1}^n \mathbb{E} \left[(\chi_{ni}^{\text{ipw}})^2 \mathbf{1}_{\{|\chi_{ni}^{\text{ipw}}| > \varepsilon\}} \middle| \mathcal{F}_{i-1} \right] \\ \leq C_\varepsilon \sum_{i=1}^n \mathbb{E} \left[|\xi_{ni}^{\text{ipw}}|^{2+\eta} \middle| \mathcal{F}_{i-1} \right]. \end{aligned} \quad (124)$$

Taking expectations and using positivity, Hölder's inequality with the marginal ν -moment, boundedness of $f_{\mathbf{X}}$, and (32), we obtain

$$\begin{aligned} \sum_{i=1}^n \mathbb{E} \left[|\xi_{ni}^{\text{ipw}}(\mathbf{x})|^{2+\eta} \right] &\leq Cn(nh_n^d)^{-(2+\eta)/2} h_n^d \\ &= C(nh_n^d)^{-\eta/2} \longrightarrow 0 \end{aligned} \quad (125)$$

by (33). Therefore, the expectation of the left-hand side of (124) tends to zero. Since that left-hand side is non-negative, Markov's inequality proves (123). The martingale central limit theorem now yields

$$S_n(\mathbf{x}) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \Sigma_{\text{mar},(\beta)}^2(\mathbf{x}; \boldsymbol{\vartheta})\right). \quad (126)$$

Combining (126) with (114), we obtain

$$\sqrt{nh_n^{d+2|\beta|}} \left[\widehat{(\partial^\beta r)_n}^{\text{ipw}}(\rho; \mathbf{x}) - g_\beta(\mathbf{x}) \right] \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \Sigma_{\text{mar},(\beta)}^2(\mathbf{x}; \boldsymbol{\theta})\right). \quad (127)$$

It remains to prove that replacing π by $\hat{\pi}_n$ is negligible. By direct subtraction,

$$R_{n,2}(\mathbf{x}) = \frac{(-1)^{|\beta|}}{\sqrt{nh_n^d}} \sum_{i=1}^n \delta_i \rho(\mathbf{Y}_i) K_{n,i}(\mathbf{x}) \left[\frac{1}{\hat{\pi}_n(\mathbf{X}_i)} - \frac{1}{\pi(\mathbf{X}_i)} \right].$$

By positivity and uniform consistency, for all sufficiently large n ,

$$\sup_{\mathbf{u} \in \mathbf{J}} \left| \frac{1}{\hat{\pi}_n(\mathbf{u})} - \frac{1}{\pi(\mathbf{u})} \right| \leq \frac{2}{\pi_0^2} \|\hat{\pi}_n - \pi\|_{\infty, \mathbf{J}}.$$

Therefore,

$$\begin{aligned} |R_{n,2}(\mathbf{x})| &\leq C \sqrt{nh_n^d} \|\hat{\pi}_n - \pi\|_{\infty, \mathbf{J}} \\ &\quad \times \frac{1}{nh_n^d} \sum_{i=1}^n \delta_i |\rho(\mathbf{Y}_i)| |K_{n,i}(\mathbf{x})|. \end{aligned} \quad (128)$$

The last average is $O_{\mathbb{P}}(1)$. To see this, truncate $|\rho|$ at a deterministic level, apply the localized kernel L^1 bound to the truncated part, and control the tail by the ν -moment in (N.0) and positivity. The predictable contribution is bounded by the local conditional density envelope, while the centered contribution is handled by the same martingale localization inequality used in the uniform argument. This establishes tightness of the displayed average; a bounded expectation alone would not control its conditional fluctuation. Consequently,

$$R_{n,2}(\mathbf{x}) = O_{\mathbb{P}}\left(\sqrt{nh_n^d} \|\hat{\pi}_n - \pi\|_{\infty, \mathbf{J}}\right) = o_{\mathbb{P}}(1) \quad (129)$$

by (36). Notice that no additional factor $h_n^{-|\beta|}$ appears here, as the derivative normalization has already canceled in the definition of $R_{n,2}$. Finally, (109), (114), (126), and (129) together with Slutsky's theorem give

$$\sqrt{nh_n^{d+2|\beta|}} \left[\widehat{(\partial^\beta r)_n}^{\text{mar}}(\rho; \mathbf{x}) - g_\beta(\mathbf{x}) \right] \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \Sigma_{\text{mar},(\beta)}^2(\mathbf{x}; \boldsymbol{\theta})\right).$$

The variance

$$\frac{m_{\rho,2}(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})}{\pi(\mathbf{x})} \int_{\mathbb{R}^d} \left[K^{(\beta)}(\mathbf{0}, \mathbf{v}) \right]^2 d\mathbf{v}$$

is recovered if either

$$\boldsymbol{\theta}_n \longrightarrow \mathbf{0}$$

or the integral

$$\int_{\mathbb{R}^d} \left[K^{(\beta)}(\boldsymbol{\theta}, \boldsymbol{\theta} + \mathbf{v}) \right]^2 d\mathbf{v}$$

is independent of $\boldsymbol{\theta} \in [0, 1]^d$.

Under the phase qualification stated at the beginning of the proof, this completes the proof of Theorem 3. In the absence of that qualification, the same argument proves the corresponding subsequential central limit theorem with variance $\Sigma_{\text{mar},(\beta)}^2(\mathbf{x}; \boldsymbol{\theta})$. $\square$

8. Besov Spaces

This appendix records the Besov space facts used in the approximation arguments of the paper. Its purpose is not to develop the general theory of Besov spaces but to specify a mathematically unambiguous regularity scale and identify the precise consequences invoked in the wavelet analysis. Besov spaces are particularly natural in this setting because their norms admit equivalent descriptions by both finite differences and wavelet coefficients; the first description makes the underlying smoothness requirement intrinsic, whereas the second converts that regularity into quantitative decay of multiresolution coefficients, and hence into projection bias bounds. Following [50], let $1 \leq p, q \leq \infty$. For a vector $\tau \in \mathbb{R}^d$, define the translation operator

$$(S_\tau f)(\mathbf{x}) = f(\mathbf{x} - \tau), \quad \mathbf{x} \in \mathbb{R}^d.$$

For $0 < s < 1$, the first-order modulus of smoothness is

$$\omega_1(f, t)_p = \sup_{\|\tau\| \leq t} \|S_\tau f - f\|_{L_p}, \quad t > 0,$$

and the corresponding Besov seminorm may be written equivalently as

$$\ell_{s,p,q}(f) = \left(\int_{\mathbb{R}^d} \left[\|\tau\|^{-s} \|S_\tau f - f\|_{L_p} \right]^q \frac{d\tau}{\|\tau\|^d} \right)^{1/q},$$

with the usual modification when $q = \infty$:

$$\ell_{s,p,\infty}(f) = \sup_{\tau \neq 0} \frac{\|S_\tau f - f\|_{L_p}}{\|\tau\|^s}.$$

The Besov space is then

$$B_{s,p,q}(\mathbb{R}^d) = \left\{ f \in L_p(\mathbb{R}^d) : \ell_{s,p,q}(f) < \infty \right\},$$

equipped with the norm

$$\|f\|_{B_{s,p,q}} = \|f\|_{L_p} + \ell_{s,p,q}(f).$$

The integral over all $\tau \in \mathbb{R}^d$ may equivalently be restricted to $\|\tau\| \leq 1$, since the contribution of large translations is controlled by $\|f\|_{L_p}$. At the borderline value $s = 1$, first-order differences do not provide the appropriate Zygmund regularity scale. Therefore, one uses the second-order symmetric difference

$$\Delta_\tau^2 f = S_\tau f + S_{-\tau} f - 2f$$

and defines

$$\ell_{1,p,q}(f) = \left(\int_{\mathbb{R}^d} \left[\|\tau\|^{-1} \|S_\tau f + S_{-\tau} f - 2f\|_{L_p} \right]^q \frac{d\tau}{\|\tau\|^d} \right)^{1/q},$$

with

$$\ell_{1,p,\infty}(f) = \sup_{\tau \neq 0} \frac{\|S_\tau f + S_{-\tau} f - 2f\|_{L_p}}{\|\tau\|}.$$

For arbitrary $s > 0$, the cleanest nonrecursive definition uses differences of an integer order $k > s$. Let

$$\Delta_\tau^k = (S_\tau - I)^k$$

and

$$\omega_k(f, t)_p = \sup_{\|\tau\| \leq t} \|\Delta_\tau^k f\|_{L_p}.$$

Then, $f \in B_{s,p,q}(\mathbb{R}^d)$ if and only if

$$\|f\|_{L_p} + \left(\int_0^1 [t^{-s} \omega_k(f, t)_p]^q \frac{dt}{t} \right)^{1/q} < \infty,$$

with the usual supremum interpretation when $q = \infty$. Different integers $k > s$ yield equivalent norms. This formulation avoids the ambiguity inherent in writing $s = [s]^- + \{s\}^+$ at integer values of s and does not require fractional regularity to be imposed on every derivative of order below $[s]$. An equivalent derivative characterization is available when $s = m + \sigma$, where $m \in \mathbb{N}_0$ and $0 < \sigma < 1$: under the standard distributional interpretation,

$$f \in B_{s,p,q}(\mathbb{R}^d)$$

if and only if $f \in W_p^m(\mathbb{R}^d)$ and every weak derivative $D^j f$ of order $|j| = m$ belongs to $B_{\sigma,p,q}(\mathbb{R}^d)$, with equivalence of the corresponding norms. At integer smoothness, the highest-order regularity is of Zygmund type and is expressed through higher-order differences rather than by setting the fractional part equal to one without qualification.

The Besov scale contains a number of classical spaces. In particular,

$$H^s(\mathbb{R}^d) = B_{s,2,2}(\mathbb{R}^d)$$

with equivalence of norms, while $B_{s,\infty,\infty}(\mathbb{R}^d)$ is the Hölder–Zygmund class. For non-integer $s > 0$, this space agrees with the classical Hölder class under the usual identification. The spaces used in this paper are isotropic Besov spaces, for which the same smoothness index applies in every coordinate direction. While they accommodate spatially inhomogeneous and locally irregular behavior, anisotropic regularity requires a genuinely anisotropic Besov scale with direction-dependent smoothness indices, and is not covered by the present notation. The wavelet characterization supplies the approximation results needed in the main proofs. Let the multiresolution analysis be r -regular and let $0 < s < r$. If

$$f = \sum_{\mathbf{k} \in \mathbb{Z}^d} a_{0,\mathbf{k}} \phi_{0,\mathbf{k}} + \sum_{j \geq 0} \sum_{\ell=1}^{2^d-1} \sum_{\mathbf{k} \in \mathbb{Z}^d} b_{\ell,j,\mathbf{k}} \psi_{\ell,j,\mathbf{k}}$$

in the appropriate distributional sense, then $f \in B_{s,p,q}(\mathbb{R}^d)$ if and only if

$$\|a_{0,\cdot}\|_{\ell_p} + \left(\sum_{j \geq 0} \left[2^{j\{s+d(1/2-1/p)\}} \|b_{j,\cdot}\|_{\ell_p} \right]^q \right)^{1/q} < \infty,$$

with the standard modification when $p = \infty$ or $q = \infty$. This equivalence is the precise bridge between analytic smoothness and multiresolution approximation. For the L_2 -risk bound used in Theorem 1, the relevant specialization is

$$f \in B_{s,2,q}(\mathbb{R}^d).$$

In that case,

$$\|f - P_{V_m} f\|_{L_2} \leq C 2^{-ms} \|f\|_{B_{s,2,q}},$$

and therefore

$$\|f - P_{V_m} f\|_{L_2}^2 \leq C 2^{-2ms} \|f\|_{B_{s,2,q}}^2.$$

This is the deterministic term appearing in (38) and (39). No derivative correction is to be subtracted from s at this stage, because the function being projected is already the target derivative $(\partial^{\beta} r)(\rho; \cdot)$. For uniform approximation, the embedding condition is dimension dependent. If

$$s > \frac{d}{p},$$

then

$$B_{s,p,q}(\mathbb{R}^d) \hookrightarrow C_b(\mathbb{R}^d)$$

under the standard restrictions on the fine index at the critical boundary, and the wavelet projection satisfies a bound of the form

$$\|f - P_{V_m} f\|_{L_\infty} \leq C 2^{-m(s-d/p)} \|f\|_{B_{s,p,q}}.$$

Thus, whenever the proofs invoke boundedness or uniform projection convergence through a Besov embedding, the condition $s > d/p$, rather than $s > 1/p$, is the relevant hypothesis in dimension d .

Remark 11. *The minimax L_2 -risk over an isotropic d -dimensional Sobolev or Besov ball has the classical order*

$$n^{-2s/(2s+d)}$$

for direct density estimation under the standard independent model. The rate

$$n^{-2s/(2s+1)}$$

is the special case $d = 1$; see [51]. This benchmark is not the rate proved in the present dependent derivative estimation problem, for which the stochastic exponent is altered by derivative order, dimension, inverse weighting, and the conditional density stabilization bound. Detailed accounts of the relationships among Besov, Sobolev, Hölder, and related smoothness classes may be found in [47,52]. Connections between Besov spaces and V_p -type spaces of functions of bounded p -variation are developed in [53] using interpolation-theoretic tools from [54]; a classical treatment of p -variation is given in [55]. Extensions of Besov theory to more general geometric and analytic settings, including manifolds and Dirichlet spaces, are discussed in [56].

Author Contributions: Conceptualization, S.D. and S.B.; methodology, S.D. and S.B.; validation, S.D. and S.B.; formal analysis, S.D. and S.B.; investigation, S.D. and S.B.; resources, S.D. and S.B.; writing—original draft preparation, S.D. and S.B.; writing—review and editing, S.D. and S.B. All authors have read and agreed to the published version of the manuscript.

Funding: Not applicable.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Acknowledgments: The researchers would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University (www.qu.edu.sa) for financial support (QU-APC-2026). We are deeply grateful to the three referees for their meticulous assessment of the manuscript and for their insightful and constructive comments, which have significantly improved both the content and the presentation of the previous version of the paper.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- Wand, M.P.; Jones, M.C. *Kernel Smoothing*; Monographs on Statistics and Applied Probability; Chapman and Hall, Ltd.: London, UK, 1995; Volume 60, p. xii+212. <https://doi.org/10.1007/978-1-4899-4493-1>.
- Eggermont, P.P.B.; LaRiccia, V.N. *Maximum Penalized Likelihood Estimation: Volume II: Regression*; Springer Series in Statistics; Springer: Dordrecht, The Netherlands, 2009; p. xx+571. <https://doi.org/10.1007/b12285>.
- Genovese, C.R.; Perone-Pacifico, M.; Verdinelli, I.; Wasserman, L. Non-parametric inference for density modes. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2016**, *78*, 99–126. <https://doi.org/10.1111/rssb.12111>.
- Noh, Y.K.; Sugiyama, M.; Liu, S.; du Plessis, M.C.; Park, F.C.; Lee, D.D. Bias reduction and metric learning for nearest-neighbor estimation of Kullback-Leibler divergence. *Neural Comput.* **2018**, *30*, 1930–1960. [10.1162/neco_a_01092](https://doi.org/10.1162/neco_a_01092).
- Dobrovidov, A.V.; Ruds'ko, I.M. Bandwidth selection in nonparametric estimator of density derivative by smoothed cross-validation method. *Autom. Remote Control* **2010**, *71*, 209–224. <https://doi.org/10.1134/S0005117910020050>.
- Genovese, C.R.; Perone-Pacifico, M.; Verdinelli, I.; Wasserman, L. On the path density of a gradient field. *Ann. Stat.* **2009**, *37*, 3236–3271. <https://doi.org/10.1214/08-AOS671>.
- Singh, R.S. Applications of estimators of a density and its derivatives to certain statistical problems. *J. R. Stat. Soc. Ser. B* **1977**, *39*, 357–363.
- Meyer, T.G. Bounds for estimation of density functions and their derivatives. *Ann. Stat.* **1977**, *5*, 136–142. <https://doi.org/10.1214/aos/1176343746>.
- Silverman, B.W. Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Stat.* **1978**, *6*, 177–184. <https://doi.org/10.1214/aos/1176344076>.
- Allaoui, S.; Bouzebda, S.; Chesneau, C.; Liu, J. Uniform almost sure convergence and asymptotic distribution of the wavelet-based estimators of partial derivatives of multivariate density function under weak dependence. *J. Nonparametr. Stat.* **2021**, *33*, 170–196. <https://doi.org/10.1080/10485252.2021.1925668>.
- Bouzebda, S.; El-hadjali, T. Uniform convergence rate of the kernel regression estimator adaptive to intrinsic dimension in presence of censored data. *J. Nonparametr. Stat.* **2020**, *32*, 864–914. <https://doi.org/10.1080/10485252.2020.1834107>.
- Bouzebda, S.; Didi, S.; El Hajj, L. Multivariate wavelet density and regression estimators for stationary and ergodic continuous time processes: Asymptotic results. *Math. Methods Statist.* **2015**, *24*, 163–199. <https://doi.org/10.3103/S1066530715030011>.
- Ramsay, J.O.; Silverman, B.W. *Applied Functional Data Analysis: Methods and Case Studies*; Springer Series in Statistics; Springer: New York, NY, USA, 2002; p. x+190. <https://doi.org/10.1007/b98886>.
- Ramsay, J.O.; Silverman, B.W. *Functional Data Analysis*, 2nd ed.; Springer Series in Statistics; Springer: New York, NY, USA, 2005; p. xx+426.
- Liu, S.; Kong, X. A generalized correlated C_p criterion for derivative estimation with dependent errors. *Comput. Statist. Data Anal.* **2022**, *171*, 107473. <https://doi.org/10.1016/j.csda.2022.107473>.
- Eubank, R.L.; Speckman, P.L. Confidence bands in nonparametric regression. *J. Amer. Statist. Assoc.* **1993**, *88*, 1287–1301.
- Ruppert, D.; Sheather, S.J.; Wand, M.P. An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **1995**, *90*, 1257–1270.
- Park, C.; Kang, K.H. SiZer analysis for the comparison of regression curves. *Comput. Statist. Data Anal.* **2008**, *52*, 3954–3970. <https://doi.org/10.1016/j.csda.2008.01.006>.
- Härdle, W.; Gasser, T. On robust kernel estimation of derivatives of regression functions. *Scand. J. Statist.* **1985**, *12*, 233–240.
- Ziegler, K. On the asymptotic normality of kernel regression estimators of the mode in the nonparametric random design model. *J. Statist. Plann. Inference* **2003**, *115*, 123–144. [https://doi.org/10.1016/S0378-3758\(02\)00152-0](https://doi.org/10.1016/S0378-3758(02)00152-0).
- Georgiev, A.A. Speed of convergence in nonparametric kernel estimation of a regression function and its derivatives. *Ann. Inst. Statist. Math.* **1984**, *36*, 455–462. <https://doi.org/10.1007/BF02481984>.
- Deheuvels, P.; Mason, D.M. General asymptotic confidence bands based on kernel-type function estimators. *Stat. Inference Stoch. Process.* **2004**, *7*, 225–277. <https://doi.org/10.1023/B:SISP.0000049092.55534.af>.
- Bouzebda, S.; Didi, S. Some asymptotic properties of kernel regression estimators of the mode for stationary and ergodic continuous time processes. *Rev. Mat. Complut.* **2021**, *34*, 811–852. <https://doi.org/10.1007/s13163-020-00368-6>.
- Bouzebda, S. Weak convergence of the conditional single index U -statistics for locally stationary functional time series. *AIMS Math.* **2024**, *9*, 14807–14898. <https://doi.org/10.3934/math.2024720>.
- Härdle, W.; Kerkycharian, G.; Picard, D.; Tsybakov, A. *Wavelets, Approximation, and Statistical Applications*; Lecture Notes in Statistics; Springer: New York, NY, USA, 1998; Volume 129, p. xviii+265. <https://doi.org/10.1007/978-1-4612-2222-4>.
- Prakasa Rao, B.L.S. Nonparametric estimation of the derivatives of a density by the method of wavelets. *Bull. Inf. Cybern.* **1996**, *28*, 91–100.
- Chaubey, Y.P.; Doosti, H.; Prakasa Rao, B.L.S. Wavelet based estimation of the derivatives of a density with associated variables. *Int. J. Pure Appl. Math.* **2006**, *27*, 97–106.

28. Rao, B.L.S.P. Nonparametric Estimation of Partial Derivatives of a Multivariate Probability Density by the Method of Wavelets. In *Asymptotics in Statistics and Probability*; Madan L.P., Ed.; Papers in Honor of George Gregory Roussas; De Gruyter: Berlin, Germany; Boston, MA, USA, 2000; pp. 321–330. <https://doi.org/10.1515/9783110942002-022>.
29. Hosseinioun, N.; Doosti, H.; Niroumand, H.A. Nonparametric estimation of a multivariate probability density for mixing sequences by the method of wavelets. *Ital. J. Pure Appl. Math.* **2011**, *28*, 31–40.
30. Koshkin, G.; Vasil'iev, V. An estimation of a multivariate density and its derivatives by weakly dependent observations. In *Statistics and Control of Stochastic Processes*; The Liptser Festschrift; Papers from the Steklov Seminar Held in Moscow, Russia, 1995–1996; World Scientific: Singapore, 1997; pp. 229–241.
31. Prakasa Rao, B.L.S. Wavelet estimation for derivative of a density in the presence of additive noise. *Braz. J. Probab. Stat.* **2018**, *32*, 834–850. <https://doi.org/10.1214/17-BJPS369>.
32. Didi, S.; Bouzebda, S. Wavelet Estimation of Partial Derivatives in Multivariate Regression Under Discrete-Time Stationary Ergodic Processes. *Mathematics* **2025**, *13*, 1587. <https://doi.org/10.3390/math13101587>.
33. Little, R.J.A.; Rubin, D.B. *Statistical Analysis with Missing Data*, 2nd ed.; Wiley Series in Probability and Statistics; Wiley-Interscience [John Wiley & Sons]: Hoboken, NJ, USA, 2002; p. xviii+381. <https://doi.org/10.1002/9781119013563>.
34. Seaman, S.; Galati, J.; Jackson, D.; Carlin, J. What is meant by “missing at random”? *Statist. Sci.* **2013**, *28*, 257–268. <https://doi.org/10.1214/13-sts415>.
35. Mealli, F.; Rubin, D.B. Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika* **2015**, *102*, 995–1000. <https://doi.org/10.1093/biomet/asv035>.
36. Lu, G.; Copas, J.B. Missing at random, likelihood ignorability and model completeness. *Ann. Statist.* **2004**, *32*, 754–765. <https://doi.org/10.1214/009053604000000166>.
37. Farewell, D.M.; Daniel, R.M.; Seaman, S.R. Missing at random: A stochastic process perspective. *Biometrika* **2022**, *109*, 227–241. <https://doi.org/10.1093/biomet/asab002>.
38. Bouzebda, S. Statistical Learning of Conditional Single-Index U-Processes Under Local Stationarity and Missing-at-Random Functional Responses. *Mathematics* **2026**, *14*, 2112. <https://doi.org/10.3390/math14122112>.
39. Bouzebda, S. Advanced Statistical Learning: Limit Theorems for Nonparametric Conditional U-Statistics Smoothed by Asymmetric Kernels Under Missing-at-Random Sampling. *Mathematics* **2026**, *14*, 2110. <https://doi.org/10.3390/math14122110>.
40. Bouzebda, S. Asymptotic Learning Theory for Conditional U-Statistics Based on Delta Sequences Under Missing at Random Mechanisms. *Mathematics* **2026**, *14*, 1899. <https://doi.org/10.3390/math14111899>.
41. Claeskens, G.; Hjort, N.L. *Model Selection and Model Averaging*; Cambridge Series in Statistical and Probabilistic Mathematics; Cambridge University Press: Cambridge, UK, 2008; Volume 27, p. xviii+312. <https://doi.org/10.1017/CBO9780511790485>.
42. Rubin, D.B. Inference and missing data. *Biometrika* **1976**, *63*, 581–592. <https://doi.org/10.1093/biomet/63.3.581>.
43. Josse, J.; Reiter, J.P. Introduction to the special section on missing data. *Statist. Sci.* **2018**, *33*, 139–141. <https://doi.org/10.1214/18-STS332IN>.
44. Little, R.J.A.; Rubin, D.B. *Statistical Analysis with Missing Data*, 3rd ed.; Wiley Series in Probability and Statistics; John Wiley & Sons: Hoboken, NJ, USA, 2019. <https://doi.org/10.1002/9781119482260>.
45. Bouzebda, S.; Cherfi, M. General bootstrap for dual ϕ -divergence estimates. *J. Probab. Stat.* **2012**, *2012*, 834107. <https://doi.org/10.1155/2012/834107>.
46. Meyer, Y. *Wavelets and Operators*; Cambridge Studies in Advanced Mathematics; Cambridge University Press: Cambridge, UK, 1992; Volume 37, p. xvi+224.
47. Triebel, H. *Theory of Function Spaces*; Monographs in Mathematics, Birkhäuser: Basel, Switzerland, 1983; Volume 78, p. 284. <https://doi.org/10.1007/978-3-0346-0416-1>.
48. Burkholder, D.L. Distribution function inequalities for martingales. *Ann. Probab.* **1973**, *1*, 19–42. <https://doi.org/10.1214/aop/1176997023>.
49. de la Peña, V.H.; Giné, E. From dependence to independence, Randomly stopped processes. *U-statistics and processes. Martingales and beyond*. In *Decoupling*; Probability and its Applications (New York); Springer: New York, NY, USA, 1999; p. xvi+392. <https://doi.org/10.1007/978-1-4612-0537-1>.
50. Masry, E. Wavelet-based estimation of multivariate regression functions in Besov spaces. *J. Nonparametr. Statist.* **2000**, *12*, 283–308. <https://doi.org/10.1080/10485250008832809>.
51. Efremovich, S. Lower bound for estimation of Sobolev densities of order less $\frac{1}{2}$. *J. Statist. Plann. Inference* **2009**, *139*, 2261–2268. <https://doi.org/10.1016/j.jspi.2008.10.017>.
52. DeVore, R.A.; Lorentz, G.G. *Constructive Approximation*; Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]; Springer: Berlin/Heidelberg, Germany, 1993; Volume 303, p. x+449.
53. Bourdaud, G.; Lanza de Cristoforis, M.; Sickel, W. Superposition operators and functions of bounded p -variation. *Rev. Mat. Iberoam.* **2006**, *22*, 455–487. <https://doi.org/10.4171/RMI/463>.

54. Peetre, J. *New Thoughts on Besov Spaces*; Duke University Mathematics Series, No. 1; Duke University, Mathematics Department: Durham, NC, USA, 1976; p. vi+305.
55. Dudley, R.M.; Norvaiša, R. *Differentiability of Six Operators on Nonsmooth Functions and p -Variation*; Lecture Notes in Mathematics; With the Collaboration of Jinghua Qian; Springer: Berlin/Heidelberg, Germany, 1999; Volume 1703, p. viii+277. <https://doi.org/10.1007/BFb0100744>.
56. Geller, D.; Pesenson, I.Z. Band-limited localized Parseval frames and Besov spaces on compact homogeneous manifolds. *J. Geom. Anal.* **2011**, *21*, 334–371. <https://doi.org/10.1007/s12220-010-9150-3>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.