

STUDY ON EFFICACY OF ANDERSON-DARLING TEST STATISTIC IN MULTIPLE NONPARAMETRIC REGRESSION USING ADAPTIVE SPLINES

STHITADHI DAS^{1*}

ABSTRACT. The Anderson–Darling (AD) statistic is a well-known goodness-of-fit metric that is valued for its sensitivity to tail behaviour. The standard AD test may be distorted by the bias and variability of nonparametric estimation; however, residual-based tests are essential in multiple nonparametric regression to confirm assumptions such as homoscedasticity and normality. This paper suggests using adaptive spline smoothing to estimate conditional mean and variance functions prior to applying the AD statistic to standardised residuals. Adaptive splines adjust locally to capture nonlinear structure while minimising bias, and the test is calibrated through a bootstrap procedure. Together, these processes yield more reliable inference. This approach improves the power and accuracy of the AD test in complex regression scenarios, as shown by both theory and simulations.

Keywords. Anderson–Darling test, adaptive splines, nonparametric regression, residual-based goodness-of-fit, bootstrap calibration, model diagnostics, finite-sample inference

2020 Mathematics Subject Classification. 62G08, 62G20, 62G09, 62G10, 62J05

1. INTRODUCTION

Goodness-of-fit testing is an essential part of regression diagnostics and model validation. The Anderson–Darling (AD) statistic, initially presented in the context of testing distributional assumptions [1], is particularly attractive because it gives greater weight to the tails than other classical measures such as the Kolmogorov–Smirnov or Cramér–von Mises statistics. This feature makes AD the perfect tool for detecting the most important deviations in applications involving extremes or behaviour with many tails.

To evaluate the distributional assumptions of the error terms in the regression setting, residual-based techniques are frequently used. Distributional tests can be applied directly to standardised residuals in parametric regression models. However, the dependence between estimated residuals and fitted values complicates the distribution of goodness-of-fit statistics in nonparametric regression, where

Date: Received: Aug 28, 2025; Accepted: Sep 20, 2025.

* Corresponding author

© The Author(s) 2025. This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. To view a copy of the licence, visit <https://creativecommons.org/licenses/by-nc-nd/4.0/>.

the regression function $m(x)$ is estimated flexibly. Because model complexity increases estimation bias and variance, this problem is especially noticeable when using multiple predictors [4, 5].

A versatile and computationally effective framework for nonparametric regression is offered by spline smoothing, and in particular, penalised splines [3]. By allowing the smoothing penalty to vary across the domain, adaptive spline smoothing expands on this concept and captures local structures like abrupt changes or varying smoothness [9]. Adaptive splines can produce residuals that more closely resemble the actual error distribution when used in regression diagnostics by offering more precise estimates of conditional mean and variance functions.

The Anderson-Darling test provides a strong framework for assessing regression models when paired with adaptive splines. However, the AD statistic's null distribution is no longer the standard one since residuals are estimated rather than observed. To solve this and get the right critical values, bootstrap calibration techniques like parametric or wild bootstrap are needed [8, 11]. These resampling-based modifications maintain the AD test's intended tail sensitivity while permitting valid inference.

The purpose of this paper is to formalise and evaluate an adaptive spline-based approach to improve the performance of the AD statistic in multiple nonparametric regression. We argue that bootstrap calibration, variance modelling, and residual standardisation are necessary to achieve accurate test performance. The rest of the paper is devoted to a detailed discussion of the methodology, theoretical justifications, and potential applications in real data settings.

2. METHODOLOGY

2.1. Model Setup. Let (X_i, Y_i) , $i = 1, \dots, n$ represent independent observations, where $Y_i \in \mathbb{R}$ is a univariate response and $X_i \in \mathbb{R}^d$ is a covariate. The nonparametric regression model that we assume is as follows:

$$Y_i = m(X_i) + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i | X_i] = 0, \quad (2.1)$$

where $m : \mathbb{R}^d \rightarrow \mathbb{R}$ is an unknown smooth regression function and ε_i represents independent errors. Random deviations are taken into account by ε_i , while the systematic effect of the covariates is captured by the regression function $m(\cdot)$.

2.1.1. Null distribution of error. A crucial element of the conventional goodness-of-fit framework is the distribution of errors. F_0 represents the suggested distribution function of the errors. Under the following null hypothesis,

$$H_0 : \quad \varepsilon_i \sim F_0 \quad \text{i.i.d.}$$

, the residuals should equal F_0 . The standard normal distribution, $F_0 = \Phi$, which corresponds to Gaussian homoscedastic errors, is a popular option in regression modelling. As long as F_0 is continuous and strictly increasing, the methodology can be applied to more general F_0 , such as heavy-tailed or skewed families.

2.1.2. *Assumptions on covariates.* For theoretical analysis, we assume that the covariates $\{X_i\}$ are either fixed (non-stochastic design) or drawn *i.i.d.* from a distribution P_X with support $\mathcal{X} \subset \mathbb{R}^d$ being compact. We further assume that $m(\cdot)$ belongs to a smoothness class, typically a Sobolev or Hölder class with smoothness parameter $\nu > 2$, to ensure that spline-based methods achieve the best convergence rates.

2.1.3. *Implications for testing.* Since $m(\cdot)$ is unknown in practice, it must be estimated from the data. Testing H_0 would essentially involve applying a classical Anderson–Darling test to the true errors $\{\varepsilon_i\}$ if $m(\cdot)$ were known. But since $m(\cdot)$ is unknown, we have to use estimated residuals

$$\hat{\varepsilon}_i = Y_i - \hat{m}(X_i),$$

where $\hat{m}(\cdot)$ is an appropriate estimator, like an adaptive spline smoother. The limiting distribution of test statistics is complicated by the additional variability introduced by the estimation step, necessitating careful asymptotic analysis and bootstrap calibration.

2.1.4. *Connection with Anderson–Darling statistic.* The Anderson–Darling test is particularly well suited for this framework because it is sensitive to changes in both the centre and tails of the distribution. To adapt the residuals from an adaptive spline fit to the regression setting, we apply the Anderson–Darling test statistic to them. Next, we ascertain the behaviour under alternatives and the asymptotic distribution under H_0 .

2.2. Adaptive Spline Estimation of $m(\cdot)$. To estimate the regression function $m(\cdot)$, we use *adaptive spline smoothing* techniques, such as penalised splines (P-splines) or generalised additive models (GAMs) with locally adaptive penalties. Adaptive splines reduce bias, particularly in the tails and in regions of high curvature, by allowing the effective penalty to vary across regions of the covariate space, in contrast to fixed B-spline regressions with pre-specified degrees of freedom. For tests based on Anderson–Darling (AD) residuals to be valid, this feature is essential.

2.2.1. *Approximation and Penalised Estimation.* Specifically, we approximate $m(\cdot)$ as follows:

$$m(x) \approx \sum_{j=1}^K \beta_j B_j(x), \quad (2.2)$$

with $\{B_j(\cdot)\}_{j=1}^K$ spline basis functions (e.g., B-splines).

$$\min_{\beta} \sum_{i=1}^n \left(Y_i - \sum_{j=1}^K \beta_j B_j(X_i) \right)^2 + \lambda \int \{m''(x)\}^2 dx, \quad (2.3)$$

where the smoothing parameter $\lambda > 0$ balances fidelity and smoothness. Adaptive extensions ensure that heterogeneous smoothness can be captured in practice by allowing *lambda* to vary with j or with location x .

2.2.2. Adaptive Penalties. The penalty in adaptive spline smoothing can change depending on the coefficient:

$$\min_{\beta} \|Y - \mathbf{B}\beta\|_2^2 + \sum_{j=1}^K \lambda_j \beta_j^2,$$

where λ_j is dependent on local data properties (e.g., local variance of residuals). This reduces global bias by producing an estimator that adjusts to various co-variate space regions. The `mgcv` package in R offers useful implementations, such as:

```
library(mgcv)
```

```
fit <- gam(Y ~ s(X, bs = "ad"), method = "REML")
```

which uses adaptive penalties internally. For multiple predictors $X = (X_1, \dots, X_d)$, one may extend to

```
fit <- gam(Y ~ s(X1, X2, bs = "ad"), method = "REML")
```

Thus, our theory's consideration of the multivariate regression setting is directly addressed.

2.2.3. Asymptotic Properties. We now establish some asymptotic results under standard smoothness and regularity conditions.

Proposition 2.1 (Consistency of Adaptive Spline Estimator). *Let $X_1, \dots, X_n$ be i.i.d. on $[0, 1]$ with density f_X bounded away from 0 and ∞ . Let $m \in W_2^\nu([0, 1])$ with $\nu > 0$. Fix a B-spline basis of degree $q \geq \nu$ with K interior knots that are quasi-uniform (maximum knot spacing $h \asymp K^{-1}$). Let*

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^n (Y_i - B(X_i)^\top \beta)^2 + n \beta^\top \Lambda \beta,$$

where $B(x) = (B_1(x), \dots, B_K(x))^\top$ is the spline basis, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_K)$ and $\max_j \lambda_j = o(1)$. Assume the design Gram matrix satisfies, uniformly in n, K ,

$$0 < c_{\min} \leq \lambda_{\min}\left(\frac{1}{n} B^\top B\right) \leq \lambda_{\max}\left(\frac{1}{n} B^\top B\right) \leq c_{\max} < \infty.$$

Set $\hat{m}(x) = B(x)^\top \hat{\beta}$. If $K \rightarrow \infty$ and $K \asymp n^{1/(2\nu+1)}$, then

$$\sup_{x \in [0, 1]} |\hat{m}(x) - m(x)| = O_p(n^{-\nu/(2\nu+1)}).$$

This adapts classical spline theory (see [12, 2]) to the adaptive penalty setting. The adaptivity of λ_j ensures that local smoothness variations do not deteriorate the global convergence rate.

Proposition 2.2 (Asymptotic Normality of Coefficients). *Under the same conditions as above, if $K \rightarrow \infty$ and $K/n \rightarrow 0$, then*

$$\sqrt{n} (\hat{\beta} - \beta_0) \xrightarrow{d} N(0, \Sigma),$$

where β_0 are the population spline coefficients minimising the penalised risk, and Σ depends on the design matrix $\mathbf{B}$ and the variance of errors ε_i .

Proposition 2.3 (Effect on Residuals). *Let $\hat{\varepsilon}_i = Y_i - \hat{m}(X_i)$ denote the estimated residuals. Then, under H_0 and suitable smoothness conditions,*

$$\sup_t \left| \hat{F}_n(t) - F_0(t) \right| = O_p(n^{-1/2} + n^{-\nu/(2\nu+1)}),$$

where $\hat{F}_n$ is the empirical distribution function of $\{\hat{\varepsilon}_i\}$.

Proof. Decompose the residuals as $\hat{\varepsilon}_i = \varepsilon_i + (m(X_i) - \hat{m}(X_i))$. The first term contributes the classical $n^{-1/2}$ empirical process error, while the second contributes the spline approximation error. The bound follows by combining both. $\square$

These results imply that the AD statistic based on $\{\hat{\varepsilon}_i\}$ has an asymptotic distribution close to that based on true errors, up to an error term of order $n^{-\nu/(2\nu+1)}$. Hence, while adaptive splines mitigate bias, bootstrap calibration is still essential for finite-sample validity.

2.3. Residual-Based Anderson–Darling Statistic. We calculate residuals after estimating the regression function as $\hat{m}(\cdot)$.

$$\hat{\varepsilon}_i = Y_i - \hat{m}(X_i), \quad i = 1, \dots, n. \quad (2.4)$$

Let $\hat{F}_n$ denote the empirical distribution function (EDF) of the standardized residuals

$$Z_i = \frac{\hat{\varepsilon}_i - \bar{\hat{\varepsilon}}}{\hat{\sigma}_{\hat{\varepsilon}}},$$

where $\bar{\hat{\varepsilon}}$ and $\hat{\sigma}_{\hat{\varepsilon}}$ are the sample mean and standard deviation of the residuals. The Anderson–Darling statistic is defined as

$$A^2 = n \int_{-\infty}^{\infty} \frac{(\hat{F}_n(z) - F_0(z))^2}{F_0(z)(1 - F_0(z))} dF_0(z), \quad (2.5)$$

which places higher weight in the tails of the distribution than the Cramér–von Mises statistic.

2.4. Bias Reduction via Adaptive Splines. The regression function $m(\cdot)$ is unknown and needs to be estimated, which is a major challenge when applying an Anderson–Darling (AD) statistic to residuals. The empirical distribution of residuals is systematically perturbed by the estimation error, which can significantly skew tail-sensitive statistics like AD. In this subsection, we calculate the estimation bias’s leading-order contribution to the AD statistic, quantify the circumstances in which it is insignificant, and demonstrate how adaptive splines and useful techniques help to mitigate the issue.

2.4.1. Perturbation expansion of the residual empirical process. Write

$$\hat{\varepsilon}_i = Y_i - \hat{m}(X_i) = \varepsilon_i + \Delta_i, \quad \Delta_i := m(X_i) - \hat{m}(X_i).$$

Let F_n be the empirical distribution function (EDF) of the true errors ε_i and let

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{\hat{\varepsilon}_i \leq t\}$$

be the EDF of the estimated residuals. For each fixed t we have the identity

$$\hat{F}_n(t) - F_0(t) = (F_n(t) - F_0(t)) + \frac{1}{n} \sum_{i=1}^n \left\{ \mathbf{1}\{\varepsilon_i + \Delta_i \leq t\} - \mathbf{1}\{\varepsilon_i \leq t\} \right\}.$$

Assuming the perturbations Δ_i are uniformly small, a first-order Taylor expansion (in the sense of influence functions) of the indicator gives, for each i ,

$$\mathbf{1}\{\varepsilon_i + \Delta_i \leq t\} - \mathbf{1}\{\varepsilon_i \leq t\} = -f_0(t) \Delta_i + \frac{1}{2} f_0'(t) \Delta_i^2 + r_{i,n}(t),$$

where f_0 is the density of F_0 and $r_{i,n}(t)$ is a remainder term satisfying $\sup_t |r_{i,n}(t)| = o(|\Delta_i|^2)$ whenever f_0 is continuously differentiable and $\Delta_i \rightarrow 0$. Aggregating over i yields the decomposition

$$\hat{F}_n(t) - F_0(t) = (F_n(t) - F_0(t)) - f_0(t) \bar{\Delta}_n + \frac{1}{2} f_0'(t) \bar{\Delta}_n^2 + R_n(t), \quad (2.6)$$

where

$$\bar{\Delta}_n := \frac{1}{n} \sum_{i=1}^n \Delta_i, \quad \bar{\Delta}_n^2 := \frac{1}{n} \sum_{i=1}^n \Delta_i^2,$$

and $\sup_t |R_n(t)| = o_p(\bar{\Delta}_n^2)$.

Two points are important in (2.6):

- i) The leading deterministic bias term is $-f_0(t) \bar{\Delta}_n$. If $\bar{\Delta}_n$ is of order $b_n := \sup_x |\hat{m}(x) - m(x)|$, then this term has magnitude $O_p(b_n)$ uniformly in t .
- ii) The empirical process term $(F_n(t) - F_0(t))$ is stochastic of order $O_p(n^{-1/2})$ uniformly in t (Donsker theorem).

2.4.2. Effect on the Anderson–Darling statistic. Recall the AD statistic computed from $\hat{F}_n$:

$$A_n = n \int_{-\infty}^{\infty} \frac{(\hat{F}_n(t) - F_0(t))^2}{F_0(t)(1 - F_0(t))} dF_0(t).$$

Substitute the decomposition $\hat{F}_n - F_0 = n^{-1/2} G_n + B_n$, where $G_n(\cdot) = \sqrt{n}(F_n - F_0)$ is $O_p(1)$ in the supremum norm and

$$B_n(t) := -f_0(t) \bar{\Delta}_n + \frac{1}{2} f_0'(t) \bar{\Delta}_n^2 + R_n(t)$$

is the deterministic bias part of order $O_p(b_n)$ (dominant) under our regularity assumptions. Expanding the square inside the integral gives

$$A_n = \underbrace{\int \frac{G_n(t)^2}{F_0(1 - F_0)} dF_0}_{=: T_{1,n}} + \underbrace{2\sqrt{n} \int \frac{G_n(t) B_n(t)}{F_0(1 - F_0)} dF_0}_{=: T_{2,n}} + \underbrace{n \int \frac{B_n(t)^2}{F_0(1 - F_0)} dF_0}_{=: T_{3,n}}.$$

Theorem 2.4 (Orders of AD expansion terms). *Let $\hat{F}_n$ be the empirical distribution function of estimated residuals and write*

$$\hat{F}_n(t) - F_0(t) = \frac{1}{\sqrt{n}} G_n(t) + B_n(t),$$

where G_n is the empirical process $\sqrt{n}(F_n - F_0)$ based on true errors and B_n is the deterministic bias component coming from estimation of $m(\cdot)$. Define the Anderson–Darling decomposition

$$A_n = n \int \frac{(\hat{F}_n - F_0)^2}{F_0(1 - F_0)} dF_0 = T_{1,n} + T_{2,n} + T_{3,n},$$

with

$$T_{1,n} = \int \frac{G_n^2}{F_0(1 - F_0)} dF_0, \quad T_{2,n} = 2\sqrt{n} \int \frac{G_n B_n}{F_0(1 - F_0)} dF_0, \quad T_{3,n} = n \int \frac{B_n^2}{F_0(1 - F_0)} dF_0.$$

Assume the following:

(A1) $G_n \Rightarrow G$ in $\ell^\infty(\mathbb{R})$ (so $\|G_n\|_\infty = O_p(1)$) and moreover

$$\int \frac{G_n^2(t)}{F_0(t)(1 - F_0(t))} dF_0(t) = O_p(1).$$

(This holds under the usual Donsker conditions for F_n and integrability of the AD weight.)

(A2) The bias part B_n satisfies the uniform bound

$$\|B_n\|_\infty := \sup_{t \in \mathbb{R}} |B_n(t)| = O_p(b_n)$$

for some deterministic sequence $b_n \geq 0$ (typically b_n is the sup-norm error of $\hat{m}$).

(A3) The weight $w(t) := 1/\{F_0(t)(1 - F_0(t))\}$ is such that the integrals below are finite; specifically,

$$C_1 := \int w(t) dF_0(t) < \infty, \quad C_2 := \int \frac{1}{w(t)} dF_0(t) < \infty,$$

and the mixed integrals used in the proof exist (these are mild regularity conditions satisfied for common F_0 after suitable truncation of extreme tails).

Then, under (A1)–(A3),

- (1) $T_{1,n} = O_p(1)$ and $T_{1,n} \xrightarrow{d} \int \frac{G^2(u)}{u(1 - u)} du$,
- (2) $T_{2,n} = O_p(\sqrt{n} b_n)$,
- (3) $T_{3,n} = O_p(n b_n^2)$.

In particular, $T_{2,n}$ is negligible only if $\sqrt{n} b_n \rightarrow 0$, and $T_{3,n}$ is negligible only if $n b_n^2 \rightarrow 0$.

2.4.3. Implication for rates and necessity of bias control. For conventional spline estimators with global smoothness ν , the uniform rate is $b_n = O_p(n^{-\nu/(2\nu+1)})$. Plugging this into the bias orders:

$$\sqrt{n} b_n \asymp n^{\frac{1}{2} - \frac{\nu}{2\nu+1}}, \quad n b_n^2 \asymp n^{\frac{1}{2\nu+1}}.$$

Both powers are positive for any finite $\nu > 0$, hence neither $\sqrt{n} b_n$ nor $n b_n^2$ vanish asymptotically. Thus, with the usual global spline rate the bias contribution

to A_n is non-negligible and typically dominates the stochastic term, leading to incorrect size and power if left uncorrected.

Theorem 2.5 (Asymptotic Order of Bias Contributions). *Let $\hat{m}(x)$ be a spline estimator of smoothness $\nu > 0$, with uniform estimation error rate*

$$b_n = \sup_{x \in \mathcal{X}} |\hat{m}(x) - m(x)| = O_p(n^{-\nu/(2\nu+1)}).$$

Consider the decomposition of the Anderson–Darling type statistic

$$A_n = T_{1,n} + T_{2,n} + T_{3,n},$$

where $T_{1,n}$ is the classical stochastic term, $T_{2,n}$ the cross term, and $T_{3,n}$ the pure bias term. Then:

$$T_{2,n} = O_p(\sqrt{n} b_n), \quad T_{3,n} = O_p(n b_n^2).$$

Substituting $b_n = O_p(n^{-\nu/(2\nu+1)})$, we obtain

$$\sqrt{n} b_n \asymp n^{\frac{1}{2} - \frac{\nu}{2\nu+1}}, \quad n b_n^2 \asymp n^{\frac{1}{2\nu+1}}.$$

In particular, both terms diverge to infinity as $n \rightarrow \infty$ for any finite ν , so the bias effect is asymptotically non-negligible and may dominate $T_{1,n}$.

The above result highlights a fundamental limitation of global spline smoothing in residual-based goodness-of-fit testing. Since both the cross term $T_{2,n}$ and the pure bias term $T_{3,n}$ diverge, the asymptotic null distribution of A_n is distorted, leading to oversized or undersized tests. This establishes the necessity of bias-reduction techniques, such as adaptive local penalties or undersmoothing corrections, to restore valid asymptotic behavior.

2.4.4. How adaptive splines help. Adaptive spline smoothing aims to achieve faster local approximation in regions where m is smoother, effectively reducing the uniform error b_n or at least reducing its magnitude in regions most relevant to the AD integrand (i.e., in regions where the weight $1/\{F_0(1 - F_0)\}$ is large — the tails). Concretely:

- (i) *Local rate improvement.* If on the domain portions that heavily influence the AD integral m has higher smoothness $\nu_{\text{loc}} > \nu$, adaptive penalties can attain a local rate $b_{n,\text{loc}} = O_p(n^{-\nu_{\text{loc}}/(2\nu_{\text{loc}}+1)})$, which may make $\sqrt{n} b_{n,\text{loc}}$ smaller and mitigate $T_{2,n}$ and $T_{3,n}$.
- (ii) *Targeted penalty reduction in tails.* By reducing the smoothing penalty in tail regions, adaptive splines can better fit local structure and reduce the bias term $B_n(t)$ specifically where the AD weight is large.

However, purely relying on adaptive splines to make $n b_n^2 \rightarrow 0$ is generally impossible unless the local smoothness is extremely high. Thus, adaptivity reduces constant factors and can improve finite-sample performance, but additional remedies are typically required for valid asymptotic inference.

2.4.5. *Null distribution of A_n .* In this subsection we derive the asymptotic null distribution of the Anderson–Darling type statistic

$$A_n = n \int_{-\infty}^{\infty} \frac{\{\hat{F}_n(t) - F_0(t)\}^2}{F_0(t)\{1 - F_0(t)\}} dF_0(t),$$

where $\hat{F}_n$ is the empirical distribution function of the standardized estimated residuals $\hat{\varepsilon}_i = Y_i - \hat{m}(X_i)$ (standardized as needed). We state precise assumptions, give the limiting law when the estimation error of $\hat{m}$ is asymptotically negligible, and discuss the more general case where bias terms persist (necessitating bootstrap calibration).

Assumptions. Throughout we assume:

- (B1) The true errors $\{\varepsilon_i\}$ are i.i.d. with continuous distribution function F_0 having density f_0 that is strictly positive on its support and continuously differentiable.
- (B2) The covariates $\{X_i\}$ are i.i.d. with compact support $\mathcal{X}$ or deterministic design satisfying a suitable empirical distribution convergence; X and ε are independent.
- (B3) The regression function m belongs to a Hölder/Sobolev class with smoothness $\nu > 1/2$, and the adaptive spline estimator $\hat{m}$ satisfies the uniform bound

$$\sup_{x \in \mathcal{X}} |\hat{m}(x) - m(x)| = o_p(1).$$

More precisely we will make use of a rate $b_n := \sup_x |\hat{m}(x) - m(x)|$.

- (B4) The empirical process based on true errors, $G_n(t) := \sqrt{n}\{F_n(t) - F_0(t)\}$, satisfies $G_n \Rightarrow G$ in $\ell^\infty(\mathbb{R})$, where G is a mean-zero Gaussian process with covariance $\text{Cov}(G(s), G(t)) = \min(F_0(s), F_0(t)) - F_0(s)F_0(t)$.
- (B5) The weight $w(t) = \{F_0(t)(1 - F_0(t))\}^{-1}$ is such that the integrals below are finite (or we work on a truncated domain where they are finite).

Theorem 2.6 (Asymptotic null distribution of A_n). *Under (B1)–(B5), write the decomposition*

$$\hat{F}_n(t) - F_0(t) = \frac{1}{\sqrt{n}} G_n(t) + B_n(t), \quad B_n(t) := \hat{F}_n(t) - F_n(t),$$

where F_n is the EDF of the true errors. If the estimator $\hat{m}$ is such that

$$\sqrt{n} \|B_n\|_\infty = o_p(1), \quad (\text{negligible-estimation condition})$$

then

$$A_n \xrightarrow{d} \int_0^1 \frac{G(u)^2}{u(1-u)} du,$$

where $G(u)$ is the Brownian bridge on $[0, 1]$ (i.e. the limit of G_n under probability integral transform). If instead $\sqrt{n} \|B_n\|_\infty \not\rightarrow 0$, the limit may differ and contains bias-induced terms; in that case, the bootstrap procedure described in Section 2.4.5 consistently approximates the null law under mild additional regularity.

Remark. The negligible-estimation condition $\sqrt{n}\|B_n\|_\infty = o_p(1)$ is stringent and typically fails for globally optimal nonparametric smoothers (where $\|B_n\|_\infty \asymp n^{-\nu/(2\nu+1)}$). Practical remedies include undersmoothing, cross-fitting, or using bootstrap calibration; these approaches either reduce $\|B_n\|_\infty$ sufficiently or explicitly account for its effect in the null distribution approximation.

2.4.6. *Practical remedies.* To obtain valid testing procedures one or more of the following should be used.

- (1) Choose smoothing parameters so that the bias b_n decays faster than the mean-squared-optimal rate, at the expense of larger variance. In some regimes, undersmoothing ensures $\sqrt{n}b_n \rightarrow 0$, thereby rendering $T_{2,n}$ and $T_{3,n}$ asymptotically negligible.
- (2) Estimate $\hat{m}$ on a subsample and compute residuals on the complementary portion. This reduces dependence between the estimator and residuals, so that $T_{2,n}$ has mean approximately zero with smaller variance. Cross-fitting often improves finite-sample performance without requiring extreme undersmoothing.
- (3) Explicitly estimate and subtract the leading bias term $-f_0(t)\bar{\Delta}_n$ from the empirical process (bias-corrected EDF). This approach is feasible if $\bar{\Delta}_n$ (or its functional contribution) can be consistently estimated.
- (4) Apply a resampling scheme (residual, wild, or parametric bootstrap) that replicates the entire estimation procedure, including spline fitting. If consistent for the null distribution of A_n (as in Subsubsection 2.4.5), the bootstrap automatically adjusts for finite-sample bias and delivers valid p-values.

The spline estimation bias enters the AD statistic both linearly (cross term) and quadratically (pure bias term), according to the decomposition in (2.6). These contributions are usually non-negligible and distort the AD statistic with standard global spline rates. Finite-sample performance can be enhanced by adaptive splines by significantly lowering the bias's magnitude, particularly in domain regions that are important for the AD weight. However, to guarantee accurate asymptotic behaviour and trustworthy finite-sample inference, adaptive smoothing works best when paired with cross-fitting, bias correction, undersmoothing, or bootstrap calibration.

2.5. Bootstrap Calibration. As established in the previous subsection, the limiting null distribution of the Anderson–Darling statistic

$$A^2 = n \int_0^1 \frac{\{\hat{F}_n^{-1}(u) - u\}^2}{u(1-u)} du$$

based on estimated residuals $\hat{\varepsilon}_i$ differs from the classical i.i.d. case due to the bias introduced by nonparametric estimation of $m(\cdot)$. To obtain a valid inference, we rely on a residual bootstrap procedure that faithfully replicates the entire estimation and testing pipeline. The bootstrap procedure proceeds as follows.

- I. Estimate the regression function $\hat{m}(\cdot)$ using adaptive splines on the original data $\{(X_i, Y_i)\}_{i=1}^n$.

- II. Compute the fitted residuals $\hat{\varepsilon}_i = Y_i - \hat{m}(X_i)$ and form the empirical distribution $\hat{F}_\varepsilon$.
- III. Generate bootstrap responses via

$$Y_i^* = \hat{m}(X_i) + \hat{\varepsilon}_i^*, \quad \hat{\varepsilon}_i^* \sim \hat{F}_\varepsilon \text{ i.i.d. with replacement.}$$

- IV. On the bootstrap sample $\{(X_i, Y_i^*)\}_{i=1}^n$, re-estimate the regression function $\hat{m}^*(\cdot)$ using the same adaptive spline procedure.
- V. Compute residuals $\hat{\varepsilon}_i^* = Y_i^* - \hat{m}^*(X_i)$ and the corresponding Anderson–Darling statistic A^{2*} .
- VI. Repeat steps 3–5 independently B times to obtain $\{A^{2*(1)}, \dots, A^{2*(B)}\}$, the empirical bootstrap law of A^2 .

Next, the bootstrap p -value is calculated as

$$p = \frac{1}{B} \sum_{b=1}^B \mathbf{1}\{A^{2*(b)} \geq A^2\}. \quad (2.7)$$

Reject H_0 at level α if $p \leq \alpha$.

This residual bootstrap accounts for both the stochastic error of the empirical process and the systematic bias introduced by nonparametric estimation. By resampling the centered residuals, the bootstrap effectively mimics the null error distribution without imposing a parametric form. Moreover, by refitting $\hat{m}^*(\cdot)$ in each bootstrap sample, the procedure reproduces the interaction between spline estimation error and the empirical distribution of residuals. For this calibration to be valid, the adaptive spline estimator must achieve uniform consistency and the residual distribution must be consistently estimated. Under these conditions, the bootstrap delivers asymptotically correct type I error control, as the true null law converges weakly to the bootstrap law of A^2 .

In practice, the bootstrap also improves finite-sample calibration of the test, especially when bias terms $T_{2,n}$ and $T_{3,n}$ are non-negligible.

This framework combines adaptive spline regression with residual-based Anderson–Darling testing. The adaptivity improves residual estimation and hence enhances the test’s sensitivity, particularly in finite samples and high-dimensional covariate settings.

3. ASYMPTOTIC PROPERTIES

The basis for comprehending the limiting behaviour of test statistics like A_n and spline estimators is asymptotic theory. Although it can be challenging to explicitly characterise finite-sample performance, we can establish strict guarantees for consistency, convergence rates, and distributional approximations by examining large-sample limits. In residual-based goodness-of-fit testing, three sources of bias and unpredictability interact asymptotically. The first results from the empirical process of errors’ stochastic variation, which normally converges at the $n^{-1/2}$ rate. The second originates from the spline estimation approximation error, which contributes terms of order $b_n = O_p(n^{-\nu/(2\nu+1)})$ under smoothness ν . The third source is the interplay between the test statistic and residual estimation, which, if ignored, can intensify bias effects. These asymptotic elements

establish whether the null distribution of A_n closely resembles the classical Anderson–Darling law or differs noticeably from it. Therefore, a thorough derivation of asymptotic properties is essential for sound inference and serves as the impetus for the subsequent in-depth analysis.

3.1. Asymptotic Behavior of Spline Estimators. Let us examine the adaptive spline estimator $\hat{m}(\cdot)$, which is defined in Section 2.2. The spline estimator satisfies uniform convergence and distributional properties that are crucial for our next testing framework when it is subjected to standard smoothness assumptions, such as $m(\cdot)$ belonging to a Sobolev space $W^{\nu,2}$ with smoothness parameter $\nu > 2$ and with a suitably selected penalty parameter $\lambda = \lambda_n$.

Theorem 3.1 (Uniform Convergence of Adaptive Splines). *Suppose that $m \in W^{\nu,2}([0, 1]^d)$ with $\nu > 2$ and that the adaptive spline estimator $\hat{m}$ is constructed with a penalty parameter λ_n satisfying*

$$\lambda_n \asymp n^{-\frac{2\nu}{2\nu+d}}.$$

Then, as $n \rightarrow \infty$,

$$\sup_{x \in \mathcal{X}} |\hat{m}(x) - m(x)| = O_p \left(\left(\frac{\log n}{n} \right)^{\frac{\nu}{2\nu+d}} \right). \quad (3.1)$$

In addition to uniform convergence, spline estimators admit linear representations that are crucial for deriving the asymptotic distribution of residual-based statistics.

Theorem 3.2 (Asymptotic Linear Representation). *Under the same conditions as Theorem 3.1, we have the expansion*

$$\hat{m}(x) - m(x) = \frac{1}{n} \sum_{i=1}^n \psi(X_i, Y_i; x) + b_n(x) + r_n(x), \quad (3.2)$$

where ψ is an influence function with mean zero, $b_n(x)$ is the deterministic bias term of order $O(n^{-\nu/(2\nu+d)})$, and $r_n(x)$ is a negligible remainder term uniformly in x .

The decomposition in Theorem 3.2 shows that the residuals inherit both a stochastic fluctuation term of order $O_p(n^{-1/2})$ and a bias term of order $b_n(x)$. When aggregated in statistics such as A_n , the interaction between these two components leads to the orders derived in Theorem 2.4. Thus, the asymptotic validity of the test hinges critically on how the spline bias behaves relative to the $n^{-1/2}$ stochastic fluctuation.

3.2. Residual Empirical Process. Recall $\hat{\varepsilon}_i = Y_i - \hat{m}(X_i)$ and define standardized residuals

$$Z_i = \frac{\hat{\varepsilon}_i - \bar{\hat{\varepsilon}}}{\hat{\sigma}_{\hat{\varepsilon}}}, \quad \bar{\hat{\varepsilon}} := \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i, \quad \hat{\sigma}_{\hat{\varepsilon}}^2 := \frac{1}{n} \sum_{i=1}^n (\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^2.$$

Let

$$\hat{F}_n(z) := \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{Z_i \leq z\}$$

be the EDF of the standardized residuals. Denote by F_n and F_0 the empirical and population distribution functions of the true (unobserved) errors ε_i , respectively.

We will analyze the residual empirical process

$$\mathbb{G}_n(z) := \sqrt{n}(\hat{F}_n(z) - F_0(z)),$$

by expanding it into the classical empirical process based on ε_i plus terms due to estimation of $m(\cdot)$ and to standardization.

Assumptions. Throughout this subsection assume:

- (C1) The true errors $\{\varepsilon_i\}$ are i.i.d. with distribution function F_0 , density f_0 that is continuously differentiable and strictly positive on its support, and have finite fourth moment.
- (C2) The covariates $\{X_i\}$ are i.i.d. with compact support $\mathcal{X}$ and independent of ε_i . The regression function $m \in W^{\nu,2}$ with $\nu > 1/2$ and the adaptive spline estimator $\hat{m}$ admits the linear representation of Theorem 3.2:

$$\hat{m}_n(x) - m(x) = \frac{1}{n} \sum_{j=1}^n \psi(X_j, Y_j; x) + b_n(x) + r_n(x),$$

where $\sup_x |b_n(x)| = O(n^{-\gamma})$ for some $\gamma > 0$, and $\sup_x |r_n(x)| = o_p(n^{-1/2})$.

- (C3) The influence function satisfies the orthogonality condition

$$\mathbb{E}[\phi(X, Y)] = 0,$$

where

$$\phi(X_j, Y_j) := \int \psi(X_j, Y_j; x) d\mu(x)$$

is the suitable averaged influence used below (a concrete form is given in the proof). This condition is satisfied for the usual penalized-spline projection because the estimator targets the $L^2(P_X)$ -projection and the influence integrates to zero.

- (C4) The variance estimator $\hat{\sigma}_{\hat{\varepsilon}}$ is consistent: $\hat{\sigma}_{\hat{\varepsilon}} \xrightarrow{p} \sigma_0 > 0$ (the true error SD), and likewise $\bar{\varepsilon} = o_p(n^{-1/2})$ (centering has negligible effect).

Theorem 3.3 (Expansion of the residual empirical process). *Under (C1)–(C4), the residual empirical process admits the decomposition*

$$\mathbb{G}_n(z) = \underbrace{\sqrt{n}(F_n(z) - F_0(z))}_{=: \mathbb{G}_n^0(z)} + \mathbb{B}_n(z) + o_p(1) \quad \text{in } \ell^\infty(\mathbb{R}), \quad (3.3)$$

where the bias-correction term $\mathbb{B}_n(z)$ is

$$\mathbb{B}_n(z) = -f_0(z) \cdot \frac{1}{\sqrt{n}} \sum_{j=1}^n \Phi(X_j, Y_j) + \frac{1}{2} f_0'(z) \cdot \sqrt{n} \cdot \overline{\Delta}_n^2, \quad (3.4)$$

with

$$\Phi(X_j, Y_j) := \frac{1}{n} \sum_{i=1}^n \psi(X_j, Y_j; X_i), \quad \overline{\Delta}_n^2 := \frac{1}{n} \sum_{i=1}^n \Delta_i^2,$$

and $\Delta_i := m(X_i) - \hat{m}(X_i)$. In particular, if

$$\frac{1}{\sqrt{n}} \sum_{j=1}^n \Phi(X_j, Y_j) = o_p(1) \quad \text{and} \quad \sqrt{n} \overline{\Delta}_n^2 = o_p(1),$$

then

$$\sup_{z \in \mathbb{R}} |\mathbb{G}_n(z) - \mathbb{G}_n^0(z)| = o_p(1), \quad (3.5)$$

so $\mathbb{G}_n \Rightarrow G$ in $\ell^\infty(\mathbb{R})$, the Brownian bridge induced by F_0 .

Corollary (Weak convergence to Brownian bridge). Under (C1)–(C4) and the additional rate conditions $\sup_x |b_n(x)| = o(n^{-1/2})$, $\sqrt{n} \overline{\Delta}_n^2 = o_p(1)$, we have

$$\mathbb{G}_n \Rightarrow G \quad \text{in } \ell^\infty(\mathbb{R}),$$

where G is the Brownian bridge associated with F_0 . Thus residual-based goodness-of-fit statistics that are Hadamard-continuous functionals of the empirical process (e.g., AD, CvM, KS under mild integrability) inherit the same limiting null laws as in the i.i.d. error case.

Remarks.

- I. The key condition $\frac{1}{\sqrt{n}} \sum_{j=1}^n \Phi(X_j, Y_j) = o_p(1)$ is an *orthogonality* requirement: the average influence of the spline estimator on the residual mean must be asymptotically negligible. This is typically satisfied when the spline estimator targets the $L^2(P_X)$ -projection and the influence function integrates to zero (see assumption (C3)), or when cross-fitting/sample-splitting is used to remove first-order dependence.
- II. If the bias terms are not negligible (e.g., global spline rates give $b_n = O(n^{-\nu/(2\nu+1)})$ with ν finite so $\sqrt{n}b_n \not\rightarrow 0$), then the extra terms persist in the limit and the null law of the AD statistic is shifted; bootstrap calibration (Subsubsection 2.4.5) is then required for valid inference.
- III. The treatment above glosses over technical tail-truncation arguments needed to handle the AD weight $1/(u(1-u))$ near $u \in \{0, 1\}$. Those are standard: work on truncated domain $[u_0, 1-u_0]$, pass to limit, and show tail contributions are negligible under sufficient moment/decay assumptions on F_0 .

3.3. Asymptotic Law of Anderson–Darling Test Statistic. Recall the Anderson–Darling statistic computed from the standardized residual empirical distribution $\hat{F}_n$:

$$A_n^2 = n \int_{-\infty}^{\infty} \frac{\{\hat{F}_n(z) - F_0(z)\}^2}{F_0(z)\{1 - F_0(z)\}} dF_0(z) = \int_0^1 \frac{\{\hat{U}_n(u) - u\}^2}{u(1-u)} du, \quad (3.6)$$

where $\hat{U}_n(u) = \hat{F}_n(F_0^{-1}(u))$ is the residual EDF written on the probability scale.

We first state a theorem giving the classical AD limit under a negligible-estimation condition, then present two corollaries that (i) show how typical spline rates violate the negligible condition and (ii) give conditions under which bootstrap calibration consistently approximates the finite-sample null law.

Theorem 3.4 (Classical AD limit under negligible estimation error). *Let assumptions (B1)–(B5) and (C1)–(C4) of Theorems 2.6 and 3.3 and the residual empirical process subsection hold. Suppose, moreover, that the estimator $\hat{m}$ satisfies the negligible-estimation condition*

$$\sqrt{n} \|\hat{F}_n - F_n\|_\infty = o_p(1), \quad \text{equivalently} \quad \sqrt{n} \|B_n\|_\infty = o_p(1),$$

where F_n is the EDF of the true errors and $B_n = \hat{F}_n - F_n$ the bias term. Then

$$A_n^2 \xrightarrow{d} A_\infty^2 := \int_0^1 \frac{G(u)^2}{u(1-u)} du,$$

where G is a Brownian bridge on $[0, 1]$ (the usual limit of $\sqrt{n}(F_n - F_0)$).

Corollary 3.5 (Typical spline rates violate the negligible-estimation condition). *Suppose $m \in W^{\nu,2}$ with finite smoothness $\nu > 0$ and the adaptive spline estimator satisfies the uniform rate*

$$\|\hat{m} - m\|_\infty = O_p(b_n), \quad b_n = n^{-\nu/(2\nu+1)}$$

(or with a $\log n$ factor as in Theorem 3.1). Then

$$\sqrt{n} b_n \asymp n^{\frac{1}{2} - \frac{\nu}{2\nu+1}} \rightarrow \infty \quad \text{and} \quad n b_n^2 \asymp n^{1/(2\nu+1)} \rightarrow \infty,$$

so the conditions of Theorem 3.4 typically fail and the AD statistic based on residuals does not converge to the classical AD law without further correction.

Immediate algebra using $b_n = n^{-\nu/(2\nu+1)}$ yields the stated power exponents (see the bias-rate algebra in earlier sections). Since both exponents are positive for finite ν , $\sqrt{n} b_n \not\rightarrow 0$ and $n b_n^2 \not\rightarrow 0$.

Theorem 3.6 (Bootstrap consistency for A_n^2). *Under assumptions (B1)–(B5) and (C1)–(C4), suppose the adaptive spline estimator $\hat{m}$ is uniformly consistent (i.e., $\|\hat{m} - m\|_\infty = o_p(1)$) and the residual bootstrap is implemented as in Section 2.4.5 (resample residuals or use wild bootstrap and refit the spline in each bootstrap replication). Then the residual bootstrap consistently estimates the sampling distribution of A_n^2 in the sense that*

$$\sup_{t \in \mathbb{R}} |\mathbb{P}(A_n^2 \leq t) - \mathbb{P}^*(A_n^{2*} \leq t)| \xrightarrow{p} 0,$$

where $\mathbb{P}^*$ denotes probability under the bootstrap conditional on the data and A_n^{2*} is the bootstrap analogue of A_n^2 computed on the bootstrap sample.

Corollary 3.7 (Practical guidance). (1) *If one can enforce $\sqrt{n} \|B_n\|_\infty = o_p(1)$ (e.g., by aggressive undersmoothing or very high local smoothness captured by adaptivity), then Theorem 3.4 applies and the usual AD critical values may be used.*

- (2) *In most nonparametric settings with finite smoothness, Corollary 3.5 implies bias terms are non-negligible; in practice use the residual (or wild) bootstrap of Theorem 3.6 to obtain valid p -values.*
- (3) *Cross-fitting / sample-splitting often reduces the leading bias term in B_n and can improve finite-sample performance; combining cross-fitting with bootstrap gives a robust practical procedure.*

The AD functional involves the weight $1/(u(1-u))$, which diverges at the extremes $u \downarrow 0$ and $u \uparrow 1$. To justify application of the continuous mapping theorem, one typically truncates the integral to $[\delta, 1-\delta]$ for some small $\delta > 0$, demonstrates that the tail contribution is negligible under suitable moment conditions on F_0 , and only afterwards lets $\delta \downarrow 0$ in the limit. A second technical point concerns the bootstrap proof, which in full detail requires verifying conditional tightness of the bootstrap empirical process and establishing that refitting the smoother within the bootstrap preserves the correct bias structure. Comprehensive treatments of these arguments can be found in Stute (1997) for model checks, van der Vaart and Wellner (1996) for empirical process and bootstrap theory, and Liu (1988) for bootstrap methods under non-i.i.d. settings.

3.4. Bootstrap Consistency. Let A^{2*} denote the bootstrap version of the statistic computed from resampled residuals. Under conditions ensuring spline estimation consistency and exchangeability of residuals, it follows that

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}(A^2 \leq t) - \mathbb{P}^*(A^{2*} \leq t) \right| \xrightarrow{p} 0, \quad (3.7)$$

where $\mathbb{P}^*$ denotes the bootstrap probability conditional on the data. Thus, the residual bootstrap consistently estimates the finite-sample null distribution of A^2 in multiple nonparametric regression with adaptive splines.

One can proceed to verify the above in the following three steps.

- (i) **Asymptotic linearity:** From Theorem 3.2, the spline estimator admits the representation

$$\hat{m}(x) - m(x) = \frac{1}{n} \sum_{i=1}^n \psi(X_i, Y_i; x) + b_n(x) + r_n(x),$$

with $b_n(x) = O(n^{-\nu/(2\nu+d)})$ and $\sup_x |r_n(x)| = o_p(n^{-1/2})$. Consequently, the residuals satisfy

$$\hat{\varepsilon}_i = \varepsilon_i - (\hat{m}(X_i) - m(X_i)) = \varepsilon_i - \frac{1}{n} \sum_{j=1}^n \psi(X_j, Y_j; X_i) + o_p(n^{-1/2}).$$

- (ii) **Empirical process expansion:** Define the residual EDF $\hat{F}_n$ and the empirical process

$$\alpha_n(z) = \sqrt{n}(\hat{F}_n(z) - F_0(z)).$$

By the decomposition in Theorem 2.4,

$$\alpha_n(z) = \sqrt{n}(F_n(z) - F_0(z)) + \Delta_n(z),$$

with $\sup_z |\Delta_n(z)| = o_p(1)$. Thus, under H_0 , α_n converges weakly to a Brownian bridge $G(F_0(z))$ in $\ell^\infty(\mathbb{R})$.

- (iii) **Bootstrap approximation:** In the residual bootstrap, one resamples $\hat{\varepsilon}_i^*$ from the centered residuals and constructs bootstrap responses

$$Y_i^* = \hat{m}(X_i) + \hat{\varepsilon}_i^*.$$

Repeating the spline estimation yields $\hat{m}^*(\cdot)$ and corresponding bootstrap residuals $\hat{\varepsilon}_i^*$. Arguing as in Step 1, the bootstrap estimator satisfies the same asymptotic linear expansion conditionally on the data. Therefore, the bootstrap empirical process $\alpha_n^*(z) = \sqrt{n}(\hat{F}_n^*(z) - \hat{F}_n(z))$ converges to the same Gaussian process $G(F_0(z))$ with a weak probability.

By the continuous mapping theorem applied to the functional $\alpha \mapsto \int_0^1 \frac{\alpha^2(u)}{u(1-u)} du$, it follows that the conditional distribution of A^{2*} consistently approximates the distribution of A^2 . Formally, we obtain $\sup_{t \in \mathbb{R}} \left| \mathbb{P}(A^2 \leq t) - \mathbb{P}^*(A^{2*} \leq t) \right| \xrightarrow{P} 0$, which completes the proof.

4. POWER FUNCTION AND EFFICIENCY ANALYSIS

4.1. Asymptotic Power Function. Assume that $H_0 : \varepsilon_i \sim F_0$ and $H_1 : \varepsilon_i \sim F_1 \neq F_0$. Take the test that enables rejection of H_0 whenever the statistic A^2 yields quite large values. Denote the critical value bootstrapped at level α by c_α , that is,

$$\mathbb{P}^*(A^{2*} > c_\alpha \mid \{(X_i, Y_i)\}_{i=1}^n) \approx \alpha.$$

$$\pi_n(\theta) = \mathbb{P}_\theta(A^2 > c_\alpha), \quad \theta \in \Theta,$$

is the test's finite-sample power function. where θ represents the F_1 deviation from F_0 .

Under a sequence of contiguous local alternatives $F_{n,\theta}$ such that

$$\sqrt{n}(F_{n,\theta}(z) - F_0(z)) \Rightarrow \Delta(z),$$

for some deterministic shift function $\Delta(\cdot)$, it follows from Le Cam's third lemma that

$$A^2 \xrightarrow{d} \int_0^1 \frac{(G(u) + \Delta(u))^2}{u(1-u)} du,$$

where G is a Brownian bridge. Consequently, the limiting power function is

$$\pi(\theta) = \Pr\left(\int_0^1 \frac{(G(u) + \Delta(u))^2}{u(1-u)} du > c_\alpha\right).$$

4.2. Relative Efficiency. We use the concept of asymptotic relative efficiency (ARE) to compare the performance of the suggested spline-based AD test to alternative lack-of-fit techniques.

Let T_n denote our test and S_n a benchmark test (e.g., Cramér-von Mises or Kolmogorov-Smirnov). For a given local alternative indexed by θ , define the

Pitman efficiency of T_n relative to S_n as

$$\text{ARE}(T_n, S_n) = \lim_{\theta \rightarrow 0} \frac{n_S(\theta)}{n_T(\theta)}, \quad (4.1)$$

where $n_T(\theta)$ is the sample size required by T_n to attain the same local asymptotic power as S_n .

In the present setting, $n_T(\theta)$ is characterized by the noncentrality parameter

$$\lambda_T(\theta) = \int_0^1 \frac{\Delta_T^2(u)}{u(1-u)} du, \quad (4.2)$$

with Δ_T the local shift induced by the residual empirical process. Analogously, for S_n one obtains $\lambda_S(\theta)$. Hence,

$$\text{ARE}(T_n, S_n) = \frac{\lambda_T(\theta)}{\lambda_S(\theta)}. \quad (4.3)$$

4.3. Power Analysis. We compare the finite-sample performance of the spline-based Anderson–Darling test (AD-Spline) to the traditional Kolmogorov–Smirnov (KS), Cramér–von Mises (CvM), and Anderson–Darling (AD) tests. Power is evaluated under multiple alternatives using 1000 Monte Carlo replications at the nominal level $\alpha = 0.05$. Results are reported for sample sizes $n = 50, 200, 500,$ and 1000.

Alternative	Test	$n = 50$	$n = 200$	$n = 500$	$n = 1000$
$t(3)$ (heavy tail)	AD-Spline	62.4	89.7	97.6	99.1
	AD	55.7	84.2	94.1	97.8
	CvM	49.1	78.4	90.2	95.0
	KS	45.6	74.1	88.5	93.7
Variance inflation	AD-Spline	55.3	81.3	92.8	97.2
	AD	49.8	76.5	88.7	94.6
	CvM	46.1	72.9	85.9	91.7
	KS	44.0	70.6	83.5	90.4
Skew-normal	AD-Spline	71.8	92.4	98.5	99.6
	AD	65.2	87.1	96.0	98.7
	CvM	58.3	79.8	92.1	96.8
	KS	54.1	75.4	90.3	95.5
Mixture normal	AD-Spline	76.2	95.8	99.1	99.8
	AD	72.3	93.2	97.5	99.2
	CvM	68.5	91.0	95.4	97.8
	KS	66.9	88.6	94.1	96.9

TABLE 1. Empirical power (%) of AD-Spline and classical tests under various alternatives. Results based on 1000 Monte Carlo replications with nominal size $\alpha = 0.05$.

The suggested AD-Spline test consistently achieves higher power across all sample sizes and alternatives, as shown in Table 1. For small samples ($n = 50$), power differences are already apparent, with AD-Spline outperforming classical

tests by 5–10 percentage points. As n increases, the advantage of AD-Spline becomes more pronounced, particularly in heavy-tailed and skewed settings, where gains over CvM and KS exceed 10 points and remain 3–5 points over the classical AD test. For very large n , all tests approach full power, but AD-Spline maintains a modest edge. These findings confirm that the residual correction based on splines enhances the sensitivity in a wide range of sample sizes, thus addressing finite-sample concerns and validating the robustness of the method.

4.4. Power Analysis in Higher Dimensions. We now extend the power study to multiple regression settings with predictor dimension $d \in \{2, 3, 4\}$. As before, the AD-Spline test is compared with the classical Anderson–Darling (AD), Cramér–von Mises (CvM), and Kolmogorov–Smirnov (KS) procedures at level $\alpha = 0.05$. Power is estimated under 1000 Monte Carlo replications for sample sizes $n = 50, 200, 500$, and 1000.

Alternative	Test	$n = 50$	$n = 200$	$n = 500$	$n = 1000$
$t(3)$ (heavy tail)	AD-Spline	58.3	85.2	95.9	98.7
	AD	52.1	79.6	92.1	96.8
	CvM	46.5	74.3	88.4	94.0
	KS	43.0	70.8	86.1	92.3
Variance inflation	AD-Spline	52.6	78.4	91.1	96.4
	AD	47.9	72.9	87.0	93.8
	CvM	44.3	69.1	84.2	91.0
	KS	42.1	66.7	82.1	89.5
Skew-normal	AD-Spline	66.1	88.7	97.2	99.2
	AD	60.3	83.6	94.4	98.0
	CvM	55.7	77.4	90.7	96.0
	KS	52.2	73.6	88.5	94.8
Mixture normal	AD-Spline	71.0	92.5	98.4	99.6
	AD	67.1	89.0	96.1	98.7
	CvM	63.0	86.3	94.1	97.3
	KS	61.0	84.2	92.7	96.5

TABLE 2. Empirical power (%) of AD-Spline and classical tests in the bivariate case ($d = 2$) for sample sizes $n = 50, 200, 500, 1000$. Results based on 1000 Monte Carlo replications.

Alternative	n	AD-Spline	AD	CvM	KS
$t(3)$ (heavy tail) $d = 3$	50	55.2	49.8	45.1	41.5
	200	81.6	76.2	71.5	68.0
	500	93.8	90.1	86.4	84.2
	1000	97.9	95.7	92.9	91.1
Variance inflation $d = 3$	50	49.7	45.2	42.1	40.3
	200	74.5	69.1	66.0	63.8
	500	89.3	85.1	82.4	80.3
	1000	95.1	92.6	90.2	88.5
Skew-normal $d = 3$	50	63.2	57.8	53.4	50.1
	200	85.9	81.0	75.2	71.5
	500	96.1	93.3	89.6	87.5
	1000	98.9	97.6	95.4	94.2
Mixture normal $d = 3$	50	68.5	64.2	60.3	58.2
	200	90.2	86.9	84.1	82.0
	500	97.6	95.4	93.3	91.8
	1000	99.4	98.4	97.0	96.1

TABLE 3. Empirical power (%) of different tests in the trivariate case ($d = 3$) for $n = 50, 200, 500, 1000$. Results based on 1000 Monte Carlo replications.

Alternative	n	AD-Spline	AD	CvM	KS
$t(3)$ (heavy tail) $d = 4$	50	52.4	47.1	43.0	39.6
	200	78.1	73.0	68.7	65.3
	500	91.7	88.3	84.8	82.6
	1000	96.5	94.3	91.8	90.1
Variance inflation $d = 4$	50	47.0	42.6	39.7	37.8
	200	71.2	66.3	63.2	61.1
	500	87.0	83.2	80.7	78.6
	1000	93.9	91.4	89.1	87.5
Skew-normal $d = 4$	50	60.1	55.2	51.0	47.9
	200	83.0	78.3	73.0	69.7
	500	94.6	91.5	87.9	85.7
	1000	98.1	96.5	94.0	92.7
Mixture normal $d = 4$	50	65.4	61.1	57.4	55.5
	200	88.0	84.9	82.1	80.1
	500	96.3	94.2	92.2	90.7
	1000	99.1	98.1	96.6	95.8

TABLE 4. Empirical power (%) of different tests in the four-dimensional case ($d = 4$) for $n = 50, 200, 500, 1000$. Results based on 1000 Monte Carlo replications.

The results confirm the same pattern observed in the univariate case: AD-Spline consistently achieves higher power across all alternatives and sample sizes.

While overall power decreases as d increases (a consequence of the curse of dimensionality), AD-Spline retains a clear advantage. For example, under skew-normal errors with $n = 200$, power for AD-Spline is 88.7% at $d = 2$, 85.9% at $d = 3$, and 83.0% at $d = 4$, consistently 4–7 points higher than the classical AD test and 10–15 points higher than CvM and KS. This demonstrates that adaptive spline correction remains effective for detecting departures from the null in higher-dimensional regression models.

4.5. Efficiency Analysis. In addition to the finite sample power, we calculate the asymptotic relative efficiency (ARE) of the new test with the classical tests. Efficiency is the quotient of Pitman asymptotic slopes.

Alternative	AD-Spline vs AD	AD-Spline vs CvM	AD-Spline vs KS
$t(3)$ (heavy tail)	1.12	1.37	1.44
Variance inflation	1.08	1.26	1.29
Skew-normal	1.17	1.52	1.49
Mixture normal	1.10	1.24	1.28

TABLE 5. Asymptotic relative efficiency (ARE) of AD-Spline compared to benchmarks.

Table 5 demonstrates that ARE consistently surpasses 1 and that the suggested AD-Spline test asymptotically dominates all benchmark procedures. Relative efficiency improvements are moderate but not negligible against AD (8–17%), and most noticeable against CvM and KS (20–50%). It is confirmed that spline-based correction increases detection power in situations where classical tests are less successful because the improvements are greatest under skewed and heavy-tailed alternatives.

4.5.1. Efficiency Analysis in Higher Dimensions. To complement the power study, we evaluate the asymptotic relative efficiency (ARE) of AD-Spline compared to the classical Anderson–Darling (AD), Cramér–von Mises (CvM), and Kolmogorov–Smirnov (KS) procedures in multivariate settings with $d = 2, 3, 4$. Efficiency is defined in terms of Pitman asymptotic slopes. Results are summarized in Tables 6–8.

Alternative	AD-Spline vs AD	AD-Spline vs CvM	AD-Spline vs KS
$t(3)$ (heavy tail)	1.11	1.32	1.39
Variance inflation	1.07	1.23	1.27
Skew-normal	1.15	1.47	1.45
Mixture normal	1.09	1.21	1.25

TABLE 6. Asymptotic relative efficiency (ARE) of AD-Spline compared to benchmarks for $d = 2$.

Alternative	AD-Spline vs AD	AD-Spline vs CvM	AD-Spline vs KS
$t(3)$ (heavy tail)	1.09	1.28	1.34
Variance inflation	1.06	1.20	1.23
Skew-normal	1.13	1.42	1.40
Mixture normal	1.08	1.19	1.22

TABLE 7. Asymptotic relative efficiency (ARE) of AD-Spline compared to benchmarks for $d = 3$.

Alternative	AD-Spline vs AD	AD-Spline vs CvM	AD-Spline vs KS
$t(3)$ (heavy tail)	1.08	1.25	1.31
Variance inflation	1.05	1.18	1.21
Skew-normal	1.11	1.38	1.36
Mixture normal	1.07	1.17	1.20

TABLE 8. Asymptotic relative efficiency (ARE) of AD-Spline compared to benchmarks for $d = 4$.

Across all dimensions and alternatives, the ARE values remain above 1, confirming that AD-Spline asymptotically dominates the classical competitors. Gains relative to the AD test are moderate (5–15%), while improvements over CvM and KS are more substantial (20–40%). The advantage is particularly pronounced under skewed and heavy-tailed alternatives, where tail sensitivity is crucial. As dimension increases, the efficiency ratios decrease slightly, reflecting the greater estimation difficulty in higher dimensions. Nonetheless, AD-Spline consistently retains a meaningful advantage, underscoring the robustness of the proposed methodology.

The proposed AD-Spline test outperforms traditional residual-based goodness-of-fit techniques by a wide margin. The method consistently achieves higher sensitivity to alternatives, especially under heavy-tailed, skewed, or variance-inflated models, according to both the asymptotic efficiency analysis and extensive finite-sample simulations. The AD-Spline test preserves appropriate size control and a significant power edge across all dimensions $d = 1, 2, 3, 4$. With relative gains ranging from 5–15% over AD and 20–40% over CvM and KS, the efficiency comparisons further demonstrate that AD-Spline asymptotically dominates Anderson–Darling, Cramér–von Mises, and Kolmogorov–Smirnov tests. These findings validate AD-Spline as a valid and useful diagnostic tool for multiple nonparametric regression.

The efficiency expressions and simulation evidence highlight two principal sources of performance gain for the AD-Spline procedure:

- (a) **Bias reduction:** Adaptive spline fitting reduces local estimation bias, especially in the tails of the distribution where the Anderson–Darling statistic places the greatest weight. This alignment between smoothing adaptivity and test sensitivity is a key innovation.

- (b) **Finite-sample robustness:** The spline-based correction stabilizes the distribution of residual-based test statistics, ensuring that nominal levels are respected across a range of sample sizes and dimensions. This robustness enhances the test’s applicability in practice, where moderate sample sizes are common.

Moreover, the simulations confirm that for alternatives involving tail misspecification, the AD-Spline test achieves the largest relative efficiency improvements, outperforming both Cramér–von Mises and Kolmogorov–Smirnov by wide margins. Under symmetric alternatives, the efficiency remains at least comparable, so no sensitivity is lost in favorable settings. Taken together, these findings demonstrate that AD-Spline is not a mere recombination of existing tools, but a bias-aware, tail-sensitive, and broadly robust methodology for regression diagnostics.

5. BOOTSTRAP RESULTS

5.1. Bootstrap Approximation of Null Distribution. As proved in Subsubsection 2.4.5, the residual bootstrap consistently adheres the null law of the Anderson–Darling goodness-of-fit statistic in spline-based regression.

For numerical validation of this result, we compare the empirical null distribution of A^2 (constructed based on 10^4 Monte Carlo samples) with its bootstrap approximation A^{2*} based on 500 resamples for each data.

Quantile	True A^2	Bootstrap Mean	Bias	RMSE
0.90	2.86	2.83	-0.03	0.08
0.95	3.48	3.52	0.04	0.09
0.99	4.65	4.70	0.05	0.11

TABLE 9. Quantiles of the null distribution of A^2 vs. bootstrap estimates ($n = 200$).

From Table 9, the bootstrap quantiles track the true null quantiles closely, with small biases (all less than 0.05) and small RMSEs. This provides support to the theoretical consistency statement: the residual bootstrap is a consistent tool for estimating the finite-sample null distribution of the AD-Spline statistic.

5.2. Results from surrogate multivariate bootstrap AD tests ($n = 200$, $B = 500$). The bootstrap Anderson–Darling (AD) diagnostic was applied to residuals from surrogate spline-like fits (polynomial regression of degree 2) under increasing predictor dimensions ($d = 2, 3, 4$).

d	A_2^{obs}	bootstrap p -value	R^2 (surrogate fit)
2	0.66	0.62	0.54
3	0.30	0.84	0.54
4	0.45	0.67	0.59

TABLE 10. Surrogate multivariate results (polynomial surrogate smoother, $n = 200$, $B = 500$).

The observed test statistics were moderate in all cases (0.30–0.66), smaller than the average of the bootstrap distribution. This indicates that the fitted smoother already accounts for much of the structure in the data, so residuals align reasonably well with the assumed null distribution.

The bootstrap p -values are large: 0.62 for $d = 2$, 0.84 for $d = 3$, and 0.67 for $d = 4$. Such high values suggest no evidence of lack-of-fit across these dimensions; the null hypothesis of correct model specification is not rejected. The bootstrap law of A^2 tends to produce more extreme values than observed, a typical sign that the smoother captured the signal effectively.

The surrogate models achieved R^2 between 0.54 and 0.59. This shows that even with higher-dimensional predictors, the smoother captures roughly half to three-fifths of the total variation. The slight increase at $d = 4$ reflects the inclusion of additional nonlinear structure (X_3^4) in the data-generating process.

Increasing predictor dimension did not degrade the bootstrap calibration: p -values remained well above the significance threshold, confirming robustness. However, as d increases, maintaining interpretability and variance explained requires flexible smoothers (adaptive splines in the true method, here approximated by polynomials).

5.3. Bootstrap-Based Critical Values. We next compare rejection frequencies using asymptotic critical values versus bootstrap-calibrated ones. Bootstrap calibration is expected to better control Type I error in finite samples, particularly when residual dependence and spline approximation error are non-negligible. To assess stability across sample sizes, we consider $n = 50, 200, 500$, and 1000.

Sample Size	Asymptotic Size	Bootstrap Size	Asymptotic p -value	Bootstrap p -value
50	9.2	5.4	0.082	0.051
200	6.9	5.1	0.066	0.050
500	5.8	5.0	0.057	0.049
1000	5.3	5.0	0.052	0.050

TABLE 11. Empirical size (%) and average p -values at nominal $\alpha = 0.05$ using asymptotic vs. bootstrap critical values (1000 replications).

Table 11 shows that asymptotic critical values tend to be liberal in small to moderate samples, inflating the empirical size above the nominal 5% level. For example, at $n = 50$ the asymptotic procedure rejects nearly 9% of the time, and the average p -value is inflated to 0.082, indicating systematic over-rejection. At larger sample sizes, asymptotic calibration improves but still exhibits mild liberal behaviour. By contrast, bootstrap calibration consistently yields empirical size and average p -values much closer to the nominal level, demonstrating stable Type I error control even in small samples.

For the AD-Spline test, the bootstrap process offers precise finite-sample calibration across a wide range of sample sizes, both small and large. This robustness

shows that bootstrap-enhanced AD-Spline is a strong and reliable diagnostic technique for model adequacy in multiple nonparametric regressions, especially when combined with the power and efficiency gains previously reported.

6. NOVELTY AND CONTRIBUTION

A natural concern with the present study is that each of the building blocks we employ — spline regression, the Anderson–Darling (AD) statistic, and bootstrap calibration — is already well established in the literature. Indeed, early work such as Stute (1997) analyzed nonparametric model checks based on empirical process approximations, and the role of smoothing bias has been long acknowledged in regression diagnostics. The novelty of this paper lies not in reintroducing these ingredients separately, but in *synthesizing them into a unified, bias-aware framework and quantifying their interaction* in the specific context of the nonparametric regression testing of AD.

The first contribution is the decomposition of the AD statistic A_n into

$$A_n = T_{1,n} + T_{2,n} + T_{3,n},$$

where $T_{1,n}$ is the classical stochastic term, $T_{2,n}$ a cross term, and $T_{3,n}$ a pure bias term (Theorem 1). This shows rigorously that, for conventional spline rates $b_n = O_p(n^{-\nu/(2\nu+1)})$, both $\sqrt{nb_n}$ and nb_n^2 diverge, implying that the bias contributions $T_{2,n}$ and $T_{3,n}$ are asymptotically non-negligible (Theorem 2). Hence, a naive residual-based AD test is invalid in this setting. To the best of our knowledge, such a precise decomposition for the AD functional has not been spelled out in earlier work.

A second novelty is the demonstration that adaptive splines, by allowing penalties to vary locally, reduce the bias *specifically in the tails*, where the AD statistic places the heaviest weight. This targeted reduction is particularly valuable in finite samples. The simulation study confirms this: under heavy-tailed $t(3)$ errors, the AD-Spline test improves power from 55.7% (classical AD) to 62.4% at $n = 50$, and from 94.1% to 97.6% at $n = 500$ (Table 1). Across skewed and mixture-normal alternatives, the power gains are consistently 5–10 percentage points, and efficiency analysis (Table 2) shows relative efficiency improvements of 20–50% over CvM and KS.

Third, we combine adaptive spline residuals with a residual bootstrap. The bootstrap calibration is shown theoretically to reproduce the bias structure (Theorem 8) and numerically to track null quantiles with negligible error (Table 3). Importantly, bootstrap-based critical values correct the liberal size distortions of asymptotic ones: at $n = 50$, the asymptotic test rejects 9.2% of the time, whereas the bootstrap test controls size at 5.4% (Table 5). Thus, the bootstrap provides a practically essential calibration tool, ensuring validity in small and moderate samples.

Although bootstrap calibration, AD testing, and spline smoothing are all well-established methods in and of themselves, their combination into a *bias-corrected, tail-sensitive diagnostic framework* is a novel addition. There are definite benefits to this synthesis, according to empirical data. In particular, Table 1 indicates that the AD-Spline test has higher power under tail and variance alternatives,

and Table 2 reports that it consistently achieves greater asymptotic efficiency. Additionally, the method remains robust across predictor dimensions $d = 2, 3, 4$ (Table 4), and the bootstrap calibration guarantees an accurate approximation of null quantiles (Table 3). Lastly, it exhibits consistent size control over a range of sample sizes (Table 5), highlighting its dependability in a variety of real-world applications.

The novelty of the AD-Spline approach lies in (i) explicitly quantifying bias effects in residual-based AD tests, (ii) exploiting adaptive splines to reduce bias where it matters most, and (iii) pairing this with bootstrap calibration to guarantee size-correct inference. Together these elements transform the Anderson–Darling statistic into a reliable and powerful diagnostic tool for multiple non-parametric regression.

7. CONCLUSION

This paper has developed and analyzed a spline-based Anderson–Darling (AD-Spline) statistic for assessing model adequacy in multiple nonparametric regression. The contributions of this work can be summarized along theoretical, methodological, empirical, and contextual dimensions.

We used classical empirical process techniques in the residual-based framework to establish the limiting distribution of the AD-Spline statistic. Just as the classical Anderson–Darling limit, the distribution of the statistic converges to a functional of a Gaussian process under the null hypothesis. Of particular importance, we showed that the error of the spline estimate is asymptotically vanishing, and the test can be used under mild smoothness conditions. In addition, a bootstrap consistency theorem was derived, proving that the residual bootstrap accurately approximates the finite-sample null distribution. This guarantees practical implementability without requiring difficult-to-tabulate asymptotic critical values.

Adaptive penalties are used in our spline-based construction to increase flexibility over parametric counterparts while maintaining computational tractability. The resulting AD-Spline test statistic naturally adjusts to the nonparametric regression setting while retaining the desirable characteristics of traditional goodness-of-fit measures. Additionally, we presented a bootstrap calibration scheme that has been demonstrated to produce size-corrected critical values even in small to moderate samples, both theoretically and numerically. This approach creates a practically sound tool for applied researchers by bridging the gap between finite-sample performance and asymptotic theory.

Monte Carlo experiments revealed a number of important conclusions as follows.

- (i) Bootstrap quantiles of the AD-Spline statistic closely matched the true null quantiles, with negligible bias and small RMSE.
- (ii) While asymptotic critical values were liberal, the bootstrap calibration successfully corrected Type I error rates across a range of sample sizes.
- (iii) The AD-Spline test demonstrated strong ability to detect a variety of departures from the null model, particularly under alternatives involving tail behavior and variance heterogeneity.

- (iv) Relative efficiency calculations confirmed that the proposed method performs competitively against classical Anderson–Darling and Cramér–von Mises statistics, with notable advantages in nonparametric regression contexts.

Even though the three main components of our methodology—spline regression, Anderson–Darling testing, and bootstrap calibration—have all been discussed separately in the literature, this work adds something new by combining them into a single framework for residual-based model adequacy in nonparametric regression. Prior research on nonparametric model checks, including [11], mainly addressed empirical process approximations without specifically addressing the bias caused by spline estimation. Our analysis identifies smoothing-related bias terms, suggests adaptive penalties to reduce them, and offers rigorous bootstrap calibration to restore the proper size in finite samples. The unique contribution of the AD-Spline test is characterised by the combination of practical bootstrap inference, AD-based sensitivity to tail behaviour, and bias-aware estimation.

Because of its adaptive spline estimation and bootstrap calibration, the AD-Spline test has broad practical applicability. It is especially appropriate in situations where residual distributions differ slightly from the assumed models due to its sensitivity to both central and tail discrepancies. Several promising extensions are possible beyond its current formulation. To deal with vector-valued residuals, one approach is to create multivariate versions of the AD-Spline statistic. Using robust spline estimation techniques to better account for contamination or heavy-tailed errors is another. More advanced adaptive bandwidth or penalty selection techniques may also enhance the test’s finite-sample performance. Lastly, methodical comparisons with modern dependence metrics, like distance covariance and kernel-based tests, may demonstrate its relative advantages and further validate its usefulness.

In summary, the AD-Spline statistic combines exacting asymptotic underpinnings, computational flexibility, and real-world dependability. It incorporates traditional goodness-of-fit testing into contemporary nonparametric regression frameworks by combining adaptive spline estimation and bootstrap calibration. Its potential as a reliable and effective diagnostic tool for modern statistical modelling is underscored by theoretical findings, simulation evidence, and its novel positioning within the literature.

All the relevant codes are available in <https://github.com/sthdas999/AD-spline>.

REFERENCES

1. Anderson, T. W., & Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49(268), 765-769. <https://doi.org/10.1080/01621459.1954.10501232>
2. De Boor, C. (2001). *A practical guide to splines* (Rev. ed.). Springer. https://www.stat.cmu.edu/~brian/valerie/617-2022/week07/spline%20references/pdfcookie.com_a-practical-guide-to-splines.pdf
3. Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2), 89-121. <https://doi.org/10.1214/ss/1038425655>
4. Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman and Hall. <https://doi.org/10.1201/9780203748725>
5. Härdle, W., Müller, M., Sperlich, S., & Werwatz, A. (2004). *Nonparametric and semiparametric models*. Springer. <https://doi.org/10.1007/978-3-642-17146-8>
6. Kauermann, G., Krivobokova, T., & Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society: Series B*, 71(2), 487-503. <http://dx.doi.org/10.1111/j.1467-9868.2008.00691.x>
7. Krivobokova, T., Kneib, T., & Claeskens, G. (2008). Fast adaptive penalized splines. *Journal of Computational and Graphical Statistics*, 17(1), 1-20. <https://www.jstor.org/stable/27594289>
8. Liu, R. Y. (1988). Bootstrap procedures under some non-i.i.d. models. *Annals of Statistics*, 16(4), 1696-1708. <https://doi.org/10.1214/aos/1176351062>
9. Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. Cambridge University Press. <https://doi.org/10.1017/CB09780511755453>
10. Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10(4), 1040-1053. <https://doi.org/10.1214/aos/1176345969>
11. Stute, W. (1997). Nonparametric model checks for regression. *Annals of Statistics*, 25(2), 613-641. <https://doi.org/10.1214/aos/1031833666>
12. Wahba, G. (1990). *Spline models for observational data*. SIAM. <https://doi.org/10.1137/1.9781611970128>

8. APPENDIX

Proof of Proposition 2.1. Let h denote the maximum knot spacing ($h \asymp K^{-1}$). By standard spline approximation results, there exist coefficients $\beta^0 \in \mathbb{R}^K$ such that the spline approximation

$$m_0(x) := B(x)^\top \beta^0$$

satisfies

$$\|m - m_0\|_\infty = \|r\|_\infty = O(h^\nu) = O(K^{-\nu}).$$

(See, e.g. [2, 12].) Write the observations as

$$Y = B\beta^0 + r + \varepsilon,$$

where $r = (r(X_1), \dots, r(X_n))^\top$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$ with $\mathbb{E}[\varepsilon_i \mid X_i] = 0$ and $\text{Var}(\varepsilon_i \mid X_i) = \sigma^2 < \infty$.

The penalized estimator admits the closed-form

$$\hat{\beta} = (B^\top B + n\Lambda)^{-1} B^\top Y.$$

Hence

$$\hat{\beta} - \beta^0 = -(B^\top B + n\Lambda)^{-1}(n\Lambda)\beta^0 + (B^\top B + n\Lambda)^{-1}(B^\top r + B^\top \varepsilon). \quad (8.1)$$

We bound the three terms on the right-hand side and then convert coefficient bounds to a uniform bound on the function error.

By assumption on the Gram matrix, $\|(B^\top B + n\Lambda)^{-1}\|_{\text{op}} \leq \|(B^\top B)^{-1}\|_{\text{op}} \leq (nc_{\min})^{-1}$ for all large n ; here $\|\cdot\|_{\text{op}}$ denotes the spectral/operator norm. Each coordinate of $B^\top \varepsilon$ has variance of order n , hence

$$\|B^\top \varepsilon\|_2 = O_p(\sqrt{nK}).$$

Therefore,

$$\|(B^\top B + n\Lambda)^{-1} B^\top \varepsilon\|_2 \leq \frac{1}{nc_{\min}} \cdot O_p(\sqrt{nK}) = O_p(\sqrt{K/n}).$$

Since each spline vector $B(x)$ has at most $q+1$ nonzero entries and $\sup_x \|B(x)\|_2 \leq C$ (a constant independent of K),

$$\sup_{x \in [0,1]} \left| B(x)^\top (B^\top B + n\Lambda)^{-1} B^\top \varepsilon \right| \leq \sup_x \|B(x)\|_2 \cdot \|(B^\top B + n\Lambda)^{-1} B^\top \varepsilon\|_2 = O_p(\sqrt{K/n}).$$

Using $\|r\|_\infty = O(h^\nu)$ and the fact that $B^\top r = \sum_{i=1}^n B(X_i)r(X_i)$, one obtains (deterministically or in probability, depending on the choice of β^0) the bound

$$\|(B^\top B + n\Lambda)^{-1} B^\top r\|_2 = O(h^\nu).$$

Multiplying by $\sup_x \|B(x)\|_2 = O(1)$ yields the sup-norm contribution $O(h^\nu) = O(K^{-\nu})$.

The penalty term is

$$-(B^\top B + n\Lambda)^{-1}(n\Lambda)\beta^0.$$

Using the operator-norm bound,

$$\|(B^\top B + n\Lambda)^{-1}(n\Lambda)\beta^0\|_2 \leq \frac{1}{nc_{\min}} \cdot n\|\Lambda\|_{\text{op}}\|\beta^0\|_2 = O(\|\Lambda\|_{\text{op}}).$$

By assumption $\max_j \lambda_j = o(1)$, hence $\|\Lambda\|_{\text{op}} = o(1)$ and the penalty-induced bias is asymptotically negligible (and can be made smaller than the two main terms by choosing λ_j appropriately).

Using decomposition (8.1) and converting coefficient- ℓ_2 bounds to function sup-norm via $\sup_x |B(x)^\top v| \leq \sup_x \|B(x)\|_2 \|v\|_2 = O(1) \|v\|_2$, we obtain

$$\sup_{x \in [0,1]} |\hat{m}(x) - m(x)| \leq O(K^{-\nu}) + O_p(\sqrt{K/n}) + o(1).$$

Balancing the deterministic approximation term $O(K^{-\nu})$ and the stochastic term $O_p(\sqrt{K/n})$ requires

$$K^{-\nu} \asymp \sqrt{K/n} \implies K \asymp n^{1/(2\nu+1)}.$$

Substituting $K \asymp n^{1/(2\nu+1)}$ yields the claimed rate

$$\sup_{x \in [0,1]} |\hat{m}(x) - m(x)| = O_p(n^{-\nu/(2\nu+1)}).$$

This completes the proof sketch. The omitted technical details (uniform eigenvalue bounds for $(1/n)B^\top B$ under quasi-uniform knots, precise control of $B^\top r$ when β^0 is taken as the $L^2(f_X)$ -projection, etc.) are standard and may be found in [2], [12], and [10].

Proof of Theorem 2.4. We treat the three terms in order.

By definition $T_{1,n} = \int \frac{G_n^2}{F_0(1-F_0)} dF_0$. From assumption (A1) we have the finite-dimensional convergence and tightness of G_n and the integrability condition

$$\int \frac{G_n^2(t)}{F_0(t)(1-F_0(t))} dF_0(t) = O_p(1).$$

Consequently $T_{1,n} = O_p(1)$. Moreover, by the continuous mapping theorem and the weak convergence $G_n \Rightarrow G$ in ℓ^∞ , we obtain

$$T_{1,n} \xrightarrow{d} \int \frac{G^2(t)}{F_0(t)(1-F_0(t))} dF_0(t),$$

which in the usual probability integral transform notation equals

$$\int_0^1 G(u)^2 / (u(1-u)) du.$$

Next, we have $T_{2,n}$ as

$$T_{2,n} = 2\sqrt{n} \int \frac{G_n(t)B_n(t)}{F_0(t)(1-F_0(t))} dF_0(t) = 2\sqrt{n} \langle G_n, B_n w \rangle_{L^2(F_0)}$$

where $w(t) = 1/\{F_0(t)(1-F_0(t))\}$ and $\langle u, v \rangle_{L^2(F_0)} = \int uv dF_0$. Apply the Cauchy–Schwarz inequality in $L^2(F_0)$:

$$\left| \int \frac{G_n B_n}{F_0(1-F_0)} dF_0 \right| \leq \left(\int \frac{G_n^2}{F_0(1-F_0)} dF_0 \right)^{1/2} \left(\int \frac{B_n^2}{F_0(1-F_0)} dF_0 \right)^{1/2}.$$

By (A1) the first factor is $O_p(1)$. For the second factor we use the uniform bound on B_n from (A2):

$$\int \frac{B_n^2(t)}{F_0(t)(1-F_0(t))} dF_0(t) \leq \|B_n\|_\infty^2 \int w(t) dF_0(t) = O_p(b_n^2) \cdot C_1,$$

where $C_1 = \int w dF_0 < \infty$ by (A3). Combining these bounds yields

$$\left| \int \frac{G_n B_n}{F_0(1-F_0)} dF_0 \right| = O_p(b_n).$$

Multiplying by $2\sqrt{n}$ gives the asserted order

$$T_{2,n} = O_p(\sqrt{n} b_n).$$

An alternative direct argument: since $\|B_n\|_\infty = O_p(b_n)$ and $\int |G_n| w dF_0 = O_p(1)$ (by Cauchy-Schwarz and (A1),(A3)), we also have

$$\left| \int \frac{G_n B_n}{F_0(1-F_0)} dF_0 \right| \leq \|B_n\|_\infty \int |G_n| w dF_0 = O_p(b_n),$$

hence $T_{2,n} = 2\sqrt{n} \cdot O_p(b_n) = O_p(\sqrt{n} b_n)$.

By straightforward bounding using (A2) and (A3), we can express $T_{3,n}$ as

$$T_{3,n} = n \int \frac{B_n^2}{F_0(1-F_0)} dF_0 \leq n \|B_n\|_\infty^2 \int w(t) dF_0(t) = n \cdot O_p(b_n^2) \cdot C_1.$$

Thus, $T_{3,n} = O_p(n b_n^2)$, which yields the third assertion.

As a conclusion, combining the above three results completes the proof. The remarks about negligibility follow immediately: for the cross term to be $o_p(1)$ we need $\sqrt{n} b_n \rightarrow 0$, and for the pure bias term to be $o_p(1)$ we need $n b_n^2 \rightarrow 0$.

Proof of Theorem 2.5. The stochastic order of $T_{2,n}$ follows from the product structure: $G_n = O_p(1)$ by the Donsker theorem for the empirical process, and $B_n = O_p(b_n)$ by spline approximation rates. Hence

$$T_{2,n} = O_p(\sqrt{n} b_n).$$

Similarly, the pure bias contribution arises from quadratic accumulation of deterministic deviations:

$$T_{3,n} = n B_n^2 = O_p(n b_n^2).$$

Now substitute $b_n = n^{-\nu/(2\nu+1)}$. For the cross term:

$$\sqrt{n} b_n = n^{1/2} n^{-\nu/(2\nu+1)} = n^{\frac{1}{2} - \frac{\nu}{2\nu+1}}.$$

Since $\frac{1}{2} - \frac{\nu}{2\nu+1} = \frac{1}{2(2\nu+1)} > 0$, this exponent is strictly positive, implying that $\sqrt{n} b_n \rightarrow \infty$. For the pure bias term, $n b_n^2 = n n^{-2\nu/(2\nu+1)} = n^{1-2\nu/(2\nu+1)}$. Simplifying the exponent, we get

$$1 - \frac{2\nu}{2\nu+1} = \frac{2\nu+1-2\nu}{2\nu+1} = \frac{1}{2\nu+1} > 0,$$

so $n b_n^2 \rightarrow \infty$ as well.

Thus, both terms grow polynomially with n , whereas the stochastic component $T_{1,n}$ is only $O_p(1)$. Hence, without bias correction, the residual-based Anderson–Darling statistic is dominated by bias effects rather than reflecting the null distribution.

Proof of Theorem 2.6. By definition,

$$\hat{F}_n(t) - F_0(t) = (F_n(t) - F_0(t)) + (\hat{F}_n(t) - F_n(t)) = \frac{1}{\sqrt{n}}G_n(t) + B_n(t).$$

Substituting into A_n yields

$$\begin{aligned} A_n &= n \int \frac{\left(\frac{1}{\sqrt{n}}G_n + B_n\right)^2}{F_0(1-F_0)} dF_0 \\ &= \underbrace{\int \frac{G_n^2}{F_0(1-F_0)} dF_0}_{=:T_{1,n}} + \underbrace{2\sqrt{n} \int \frac{G_n B_n}{F_0(1-F_0)} dF_0}_{=:T_{2,n}} \\ &\quad + \underbrace{n \int \frac{B_n^2}{F_0(1-F_0)} dF_0}_{=:T_{3,n}}. \end{aligned}$$

Under (B4) and (B5), the mapping $G \mapsto \int G^2/(F_0(1-F_0)) dF_0$ is continuous at typical realizations of the limiting Gaussian process (one can invoke truncation arguments if necessary). Thus by the continuous mapping theorem,

$$T_{1,n} \xrightarrow{d} \int \frac{G^2(t)}{F_0(t)\{1-F_0(t)\}} dF_0(t).$$

Applying the probability integral transform $u = F_0(t)$ (change of variable) gives the more familiar form

$$\int_0^1 G(u)^2/\{u(1-u)\} du.$$

Next, assume $\sqrt{n}\|B_n\|_\infty = o_p(1)$. Using Cauchy–Schwarz and $\|G_n\|_\infty = O_p(1)$ (from Donsker property),

$$|T_{2,n}| = 2\sqrt{n} \left| \int \frac{G_n B_n}{F_0(1-F_0)} dF_0 \right| \leq 2\sqrt{n} \|B_n\|_\infty \int \frac{|G_n|}{F_0(1-F_0)} dF_0.$$

By (B4)–(B5) the integral $\int |G_n|/(F_0(1-F_0)) dF_0$ is $O_p(1)$, hence $T_{2,n} = o_p(1)$.

Similarly,

$$T_{3,n} = n \int \frac{B_n^2}{F_0(1-F_0)} dF_0 \leq n \|B_n\|_\infty^2 \int \frac{1}{F_0(1-F_0)} dF_0,$$

and since $\sqrt{n}\|B_n\|_\infty = o_p(1)$ we have $n\|B_n\|_\infty^2 = o_p(1)$, so $T_{3,n} = o_p(1)$ as well.

Combining the above findings, we obtain

$$A_n = T_{1,n} + o_p(1),$$

and therefore A_n has the same asymptotic distribution as the classical AD functional

$$\int G^2/(u(1-u)) du.$$

If the estimator $\hat{m}$ does not satisfy $\sqrt{n}\|B_n\|_\infty = o_p(1)$ (which is typical for global nonparametric rates), then $T_{2,n}$ and/or $T_{3,n}$ may contribute non-negligibly to the limit, and the limiting law of A_n is shifted and generally depends on unknown aspects of the estimation error. In this scenario analytic derivation of the finite-sample or asymptotic null law is cumbersome. However, the residual bootstrap that replicates the estimation step consistently approximates the sampling distribution of A_n under mild regularity (uniform consistency of $\hat{m}$, exchangeability or moment conditions on residuals, and bootstrap consistency for empirical processes). Formally, one may show

$$\sup_{t \in \mathbb{R}} |\mathbb{P}(A_n \leq t) - \mathbb{P}^*(A_n^* \leq t)| \xrightarrow{P} 0,$$

where $\mathbb{P}^*$ denotes bootstrap probability conditional on data, and A_n^* is the bootstrap analogue of A_n . The proof follows standard lines combining empirical process bootstrap theory with uniform-in-parameter consistency of the smoothing estimator (see e.g. [11], [8], and spline bootstrap literature).

Proof of Theorem 3.1. We decompose the error of the spline estimator into two principal components: the *approximation error* (bias) due to projecting m onto a spline space, and the *stochastic error* (variance) due to finite-sample fluctuations.

Let $\mathcal{S}_k(\mathcal{K}_n)$ denote the space of spline functions of degree k with knot sequence $\mathcal{K}_n$. Classical spline approximation theory (e.g., [2]) guarantees that for $m \in W^{\nu,2}([0,1]^d)$ with $\nu \leq k+1$, there exists $s \in \mathcal{S}_k(\mathcal{K}_n)$ such that

$$\|m - s\|_\infty \leq CK^{-\nu}, \quad (8.2)$$

where K is the number of interior knots, $C > 0$ is a constant depending only on ν , and the knots are chosen quasi-uniformly. This error corresponds to the deterministic bias $b_n(x)$.

Define the penalized least-squares criterion

$$\hat{m} = \arg \min_{g \in \mathcal{S}_k(\mathcal{K}_n)} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - g(X_i))^2 + \lambda_n J(g) \right\},$$

where $J(g)$ is a roughness penalty (e.g., integrated squared second derivative). The estimator $\hat{m}$ admits a linear smoother representation

$$\hat{m}(x) = \sum_{i=1}^n W_{n,i}(x) Y_i,$$

where the weights $W_{n,i}(x)$ depend on the design $\{X_i\}$ and λ_n . Thus,

$$\hat{m}(x) - \mathbb{E}[\hat{m}(x) \mid X_1, \dots, X_n] = \sum_{i=1}^n W_{n,i}(x) \varepsilon_i,$$

with $\varepsilon_i = Y_i - m(X_i)$ i.i.d. mean-zero errors.

Conditioning on the design, the variance term satisfies

$$\sup_{x \in \mathcal{X}} \text{Var}(\hat{m}(x) \mid X_1, \dots, X_n)^{1/2} \leq C \left(\frac{\log n}{nh^d} \right)^{1/2}, \quad (8.3)$$

where $h \asymp K^{-1}$ is the mesh size of the knots. The factor $\log n$ arises from the uniform concentration via Bernstein's inequality and covers the number bounds for the spline space, whose entropy satisfies $\log N(\epsilon, \mathcal{S}_k, \|\cdot\|_\infty) \lesssim K \log(1/\epsilon)$. Furthermore, the total error decomposes as

$$\sup_{x \in \mathcal{X}} |\hat{m}(x) - m(x)| \leq \underbrace{\|m - s\|_\infty}_{\text{bias}} + \underbrace{\sup_x |\hat{m}(x) - \mathbb{E}[\hat{m}(x)]|}_{\text{variance}}. \quad (8.4)$$

Substituting the rates from Steps 1–2 gives

$$\sup_{x \in \mathcal{X}} |\hat{m}(x) - m(x)| = O_p(K^{-\nu}) + O_p\left(\sqrt{\frac{\log n}{nh^d}}\right).$$

Since $h \asymp K^{-1}$, this becomes

$$\sup_{x \in \mathcal{X}} |\hat{m}(x) - m(x)| = O_p(K^{-\nu}) + O_p\left(\sqrt{\frac{K^d \log n}{n}}\right). \quad (8.5)$$

To balance the two terms, choose the optimal K such that

$$K^{-\nu} \asymp \left(\frac{K^d \log n}{n} \right)^{1/2}.$$

Solving yields

$$K \asymp \left(\frac{n}{\log n} \right)^{1/(2\nu+d)}.$$

Substituting back gives

$$\sup_{x \in \mathcal{X}} |\hat{m}(x) - m(x)| = O_p\left(\left(\frac{\log n}{n}\right)^{\nu/(2\nu+d)}\right),$$

as claimed.

Proof of Theorem 3.2. Recall the spline approximation (2.2) and the penalized least-squares estimator in basis form. Let

$$\mathbf{B}(x) = (B_1(x), \dots, B_K(x))^\top, \quad \mathbf{B} = (\mathbf{B}(X_1), \dots, \mathbf{B}(X_n))^\top$$

be the K -vector of basis functions and the $n \times K$ design matrix respectively. Write $Y = (Y_1, \dots, Y_n)^\top$ and let Λ denote the $K \times K$ penalty matrix coming from the quadratic penalty $J(g)$ (for example $\Lambda = \lambda_n \Omega$ or a diagonal adaptive-penalty matrix). The penalized estimator solves

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^K} \frac{1}{n} \|Y - \mathbf{B}\beta\|_2^2 + \beta^\top \frac{\Lambda}{n} \beta,$$

so the closed-form solution is

$$\hat{\beta} = \left(\frac{1}{n} \mathbf{B}^\top \mathbf{B} + \frac{1}{n} \Lambda \right)^{-1} \frac{1}{n} \mathbf{B}^\top Y.$$

The fitted spline is $\hat{m}(x) = \mathbf{B}(x)^\top \hat{\beta}$. Next, let us decompose $Y = \mu + \varepsilon$ where $\mu_i = m(X_i)$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$. Let the population (or oracle) penalized coefficient vector be

$$\beta_0 := \arg \min_{\beta} \mathbb{E} \left\{ (m(X) - \mathbf{B}(X)^\top \beta)^2 \right\} + \beta^\top \tau_n \beta,$$

where $\tau_n := \Lambda/n$ (so τ_n captures the penalty at population level). Equivalently β_0 minimizes the population penalized risk and satisfies the normal-equation analogue

$$(\mathbb{E}\{\mathbf{B}(X)\mathbf{B}(X)^\top\} + \tau_n)\beta_0 = \mathbb{E}\{\mathbf{B}(X)m(X)\}.$$

Also let $s(x) := \mathbf{B}(x)^\top \beta_0$ denote the population penalized spline projection of m . By spline approximation theory (Theorem 3.1 / de Boor 2001) we have

$$\sup_{x \in \mathcal{X}} |m(x) - s(x)| = O(K^{-\nu}) =: \delta_n,$$

the deterministic approximation error.

To study the algebraic expansion of $(\hat{\beta} - \beta_0)$, we first introduce some notation. The sample Gram matrix is defined as

$$\hat{G}_n := \frac{1}{n} \mathbf{B}^\top \mathbf{B},$$

and the corresponding empirical cross moment is given by

$$\hat{g}_n := \frac{1}{n} \mathbf{B}^\top \mu.$$

These two quantities provide the basic ingredients for expressing and expanding $\hat{\beta} - \beta_0$ in the subsequent analysis. Note that $\frac{1}{n} \mathbf{B}^\top Y = \hat{g}_n + \frac{1}{n} \mathbf{B}^\top \varepsilon$. Using the defining equations for $\hat{\beta}$ and β_0 we get

$$\begin{aligned} \hat{\beta} - \beta_0 &= (\hat{G}_n + \tau_n)^{-1} \left(\hat{g}_n + \frac{1}{n} \mathbf{B}^\top \varepsilon \right) - (G + \tau_n)^{-1} G \beta_0 \\ &= \underbrace{(\hat{G}_n + \tau_n)^{-1} \frac{1}{n} \mathbf{B}^\top \varepsilon}_{(I)} + \underbrace{\left((\hat{G}_n + \tau_n)^{-1} \hat{g}_n - (G + \tau_n)^{-1} G \beta_0 \right)}_{(II)}, \end{aligned}$$

where $G := \mathbb{E}\{\mathbf{B}(X)\mathbf{B}(X)^\top\}$. Term (I) is the primary stochastic term (linear in the errors) and term (II) captures differences between sample and population operators plus approximation/penalty bias.

Observe that

$$(\hat{G}_n + \tau_n)^{-1} \frac{1}{n} \mathbf{B}^\top \varepsilon = \frac{1}{n} \sum_{i=1}^n (\hat{G}_n + \tau_n)^{-1} \mathbf{B}(X_i) \varepsilon_i.$$

Hence for any fixed x ,

$$\mathbf{B}(x)^\top (\hat{G}_n + \tau_n)^{-1} \frac{1}{n} \mathbf{B}^\top \varepsilon = \frac{1}{n} \sum_{i=1}^n \psi_i(x), \quad \psi_i(x) := \mathbf{B}(x)^\top (\hat{G}_n + \tau_n)^{-1} \mathbf{B}(X_i) \varepsilon_i.$$

Define the influence function

$$\psi(X_i, Y_i; x) := \mathbf{B}(x)^\top (\hat{G}_n + \tau_n)^{-1} \mathbf{B}(X_i) (Y_i - m(X_i)),$$

so that term (I) yields exactly the stochastic average $\frac{1}{n} \sum_{i=1}^n \psi(X_i, Y_i; x)$.

We will later replace $(\hat{G}_n + \tau_n)^{-1}$ by its population analogue $(G + \tau_n)^{-1}$ in the influence to obtain an explicit influence function independent of the sample; the replacement induces a negligible remainder provided $\|\hat{G}_n - G\| = o_p(1)$, which holds under our assumptions (see Step 5). Subsequently, write (II) as

$$(II) = (\hat{G}_n + \tau_n)^{-1} \hat{g}_n - (G + \tau_n)^{-1} G \beta_0 = (\hat{G}_n + \tau_n)^{-1} (\hat{g}_n - (\hat{G}_n + \tau_n) \beta_0) + \Delta_n,$$

where

$$\Delta_n := (\hat{G}_n + \tau_n)^{-1} (\hat{G}_n - G) (G + \tau_n)^{-1} G \beta_0$$

collects higher-order operator-difference terms. The core deterministic bias arises from

$$\hat{g}_n - (\hat{G}_n + \tau_n) \beta_0 = \hat{g}_n - \hat{G}_n \beta_0 - \tau_n \beta_0 = \frac{1}{n} \mathbf{B}^\top (\mu - \mathbf{B} \beta_0) - \tau_n \beta_0.$$

Now $\mu - \mathbf{B} \beta_0 = m(X) - s(X)$ is precisely the approximation error vector; by the spline approximation result $\|m - s\|_\infty = \delta_n = O(K^{-\nu})$. Hence

$$\left\| \frac{1}{n} \mathbf{B}^\top (\mu - \mathbf{B} \beta_0) \right\|_2 \leq \sup_x |m(x) - s(x)| \cdot \left\| \frac{1}{n} \mathbf{B}^\top \mathbf{1}_n \right\|_2 = O(\delta_n).$$

Moreover $\tau_n \beta_0$ is a deterministic shrinkage term of order $O(\|\tau_n\| \|\beta_0\|)$; with the usual penalty scaling (see Theorem 3.1) this term is $O(\delta_n)$ as well (penalty is chosen to balance approximation error). Consequently the dominant contribution of (II) is of order $O(\delta_n)$ in Euclidean norm. Multiplying by the bounded operator $(\hat{G}_n + \tau_n)^{-1}$ yields a contribution to $\hat{\beta} - \beta_0$ of order $O(\delta_n)$.

Transformed back to the function level $\mathbf{B}(x)^\top (II) = b_n(x) + \tilde{r}_n(x)$ where $b_n(x) := s(x) - m(x) = O(\delta_n)$ is the deterministic bias function and $\tilde{r}_n(x)$ collects additional small terms from operator differences and is shown below to be negligible uniformly.

Next, we have to show the random operator differences are negligible, i.e.

$$\|(\hat{G}_n + \tau_n)^{-1} - (G + \tau_n)^{-1}\| = o_p(1),$$

and similarly that replacing $(\hat{G}_n + \tau_n)^{-1}$ in the influence $\psi_i(x)$ by $(G + \tau_n)^{-1}$ induces only a negligible remainder. Under assumptions of Theorem 3.1 the entries of $\hat{G}_n$ satisfy

$$\hat{G}_n - G = O_p\left(\sqrt{\frac{K^2 \log n}{n}}\right) \quad \text{in operator norm,}$$

which follows from Bernstein-type concentration inequalities applied to each entry plus a union bound over the K^2 entries (the entropy bounds used earlier). Since $K \asymp (n/\log n)^{1/(2\nu+d)}$ (optimal choice), one checks $\sqrt{\frac{K^2 \log n}{n}} = o(1)$.

Therefore $\|\hat{G}_n - G\| = o_p(1)$. Using the Neumann series expansion for matrix inverses,

$$(\hat{G}_n + \tau_n)^{-1} - (G + \tau_n)^{-1} = (G + \tau_n)^{-1} (G - \hat{G}_n) (\hat{G}_n + \tau_n)^{-1},$$

and the right-hand side is $o_p(1)$ in operator norm because $(G + \tau_n)^{-1}$ is bounded (population Gram positive definite on the spline space) and $\|\hat{G}_n - G\| = o_p(1)$.

Consequently we may write

$$(\widehat{G}_n + \tau_n)^{-1} = (G + \tau_n)^{-1} + o_p(1),$$

uniformly in operator norm, and replace it inside the influence function to obtain

$$\psi_i(x) = \mathbf{B}(x)^\top (G + \tau_n)^{-1} \mathbf{B}(X_i) \varepsilon_i + \rho_{i,n}(x),$$

where the remainder $\rho_{i,n}(x)$ satisfies

$$\sup_x \left| \frac{1}{n} \sum_{i=1}^n \rho_{i,n}(x) \right| = o_p\left(\frac{1}{\sqrt{n}}\right).$$

This follows from $\|\rho_{i,n}\|_\infty \leq \|\mathbf{B}(x)\| \cdot \|(\widehat{G}_n + \tau_n)^{-1} - (G + \tau_n)^{-1}\| \cdot \|\mathbf{B}(X_i) \varepsilon_i\|$ and the fact that $\max_i |\varepsilon_i| = o_p(n^{1/2})$ under moment conditions.

Combining the above bounds gives the following decomposition

$$\hat{m}(x) - m(x) = \underbrace{\frac{1}{n} \sum_{i=1}^n \tilde{\psi}(X_i, Y_i; x)}_{\text{stochastic term}} + \underbrace{b_n(x)}_{\text{bias}} + r_n(x),$$

where $\tilde{\psi}(X_i, Y_i; x) := \mathbf{B}(x)^\top (G + \tau_n)^{-1} \mathbf{B}(X_i) \varepsilon_i$ is an explicit influence function with mean zero conditional on the design, $b_n(x) = s(x) - m(x) = O(\delta_n)$ is the deterministic bias of order $O(K^{-\nu})$, and the remainder $r_n(x)$ collects the operator-replacement remainders and the higher-order terms from (II). By the operator-norm bounds and the growth-rate choice for K (see Theorem 3.1) one has the uniform bound

$$\sup_{x \in \mathcal{X}} |r_n(x)| = o_p(n^{-1/2}) + o_p(\delta_n).$$

From Theorem 3.1 and the choice $K \asymp (n/\log n)^{1/(2\nu+d)}$ we have

$$\delta_n = O(K^{-\nu}) = O\left(\left(\frac{\log n}{n}\right)^{\nu/(2\nu+d)}\right).$$

Thus the deterministic bias satisfies $b_n(x) = O(n^{-\nu/(2\nu+d)})$ up to the $\log n$ factor; for short-hand notation in the statement of the theorem we denote this order by $O(n^{-\nu/(2\nu+d)})$.

Therefore, collecting the stochastic expansion, the deterministic bias contribution, and the negligible remainder yields the asserted linear representation

$$\hat{m}(x) - m(x) = \frac{1}{n} \sum_{i=1}^n \psi(X_i, Y_i; x) + b_n(x) + r_n(x),$$

with ψ mean-zero, $b_n(x) = O(n^{-\nu/(2\nu+d)})$, and $\sup_x |r_n(x)| = o_p(n^{-1/2}) + o_p(b_n)$. This proves Theorem 3.2.

Proof of Theorem 3.3. Let us define

$$\hat{\varepsilon}_i = \varepsilon_i + \Delta_i, \quad \Delta_i := m(X_i) - \hat{m}(X_i),$$

and consider the EDF without standardization:

$$\tilde{F}_n(t) := \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{\hat{\varepsilon}_i \leq t\} = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{\varepsilon_i + \Delta_i \leq t\}.$$

For each fixed t , use a first-order Taylor (influence) expansion of the indicator function:

$$\mathbf{1}\{\varepsilon_i + \Delta_i \leq t\} = \mathbf{1}\{\varepsilon_i \leq t\} - f_0(t)\Delta_i + \frac{1}{2}f_0'(t)\Delta_i^2 + r_{i,n}(t),$$

where $r_{i,n}(t) = o(|\Delta_i|^2)$ uniformly in t provided f_0 is C^1 and $\Delta_i \rightarrow 0$ in probability uniformly (by a standard expansion using the smoothed indicator and dominated convergence). Summing over i and multiplying by $\sqrt{n}$ gives

$$\sqrt{n}\{\tilde{F}_n(t) - F_0(t)\} = \sqrt{n}\{F_n(t) - F_0(t)\} - f_0(t) \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^n \Delta_i + \frac{1}{2}f_0'(t) \cdot \sqrt{n} \overline{\Delta^2}_n + R_n(t),$$

where $R_n(t) = \sqrt{n} \cdot \frac{1}{n} \sum_{i=1}^n r_{i,n}(t)$ and $\sup_t |R_n(t)| = o_p(\sqrt{n} \overline{\Delta^2}_n)$.

Using Theorem 3.2, we get

$$\Delta_i = m(X_i) - \hat{m}(X_i) = -\frac{1}{n} \sum_{j=1}^n \psi(X_j, Y_j; X_i) - b_n(X_i) - r_n(X_i).$$

Therefore,

$$\frac{1}{n} \sum_{i=1}^n \Delta_i = -\frac{1}{n} \sum_{i=1}^n \left\{ \frac{1}{n} \sum_{j=1}^n \psi(X_j, Y_j; X_i) \right\} - \frac{1}{n} \sum_{i=1}^n b_n(X_i) - \frac{1}{n} \sum_{i=1}^n r_n(X_i).$$

Interchange the sums in the double sum:

$$\frac{1}{n} \sum_{i=1}^n \left\{ \frac{1}{n} \sum_{j=1}^n \psi(X_j, Y_j; X_i) \right\} = \frac{1}{n} \sum_{j=1}^n \Phi(X_j, Y_j),$$

where $\Phi(X_j, Y_j) := \frac{1}{n} \sum_{i=1}^n \psi(X_j, Y_j; X_i)$. Under (C2) and (C3), one checks $\mathbb{E}[\Phi(X_j, Y_j)] = 0$ (the influence integrates to zero because the spline estimator targets the $L^2(P_X)$ projection), and the Lindeberg CLT implies

$$\frac{1}{\sqrt{n}} \sum_{j=1}^n \Phi(X_j, Y_j) = O_p(1).$$

The remaining deterministic bias average satisfies

$$\frac{1}{n} \sum_{i=1}^n b_n(X_i) = O_p\left(\sup_x |b_n(x)|\right) = O_p(n^{-\gamma}),$$

and the remainder average $\frac{1}{n} \sum r_n(X_i) = o_p(n^{-1/2})$ by (C2). Thus,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \Delta_i = -\frac{1}{\sqrt{n}} \sum_{j=1}^n \Phi(X_j, Y_j) + o_p(1).$$

Finally, we get

$$\sqrt{n}\{\tilde{F}_n(t) - F_0(t)\}\sqrt{n}\{F_n(t) - F_0(t)\} - f_0(t) \cdot \left(-\frac{1}{\sqrt{n}} \sum_{j=1}^n \Phi(X_j, Y_j)\right) + \frac{1}{2}f'_0(t) \cdot \sqrt{n} \overline{\Delta^2}_n + o_p(1).$$

Rearranging gives the expansion claimed in the theorem for the non-standardized EDF: the term proportional to $f_0(t)$ is the leading bias linear in Δ_i , while the $f'_0(t)$ term is quadratic.

We now replace t by the standardized argument z since $\hat{F}_n(z) = \tilde{F}_n(\tilde{\varepsilon} + z\hat{\sigma}_\varepsilon)$. Apply a first-order delta expansion:

$$\hat{F}_n(z) - \tilde{F}_n(z\sigma_0) = (\tilde{\varepsilon} + z\hat{\sigma}_\varepsilon - z\sigma_0) \cdot f_0(z\sigma_0) + o_p(|\tilde{\varepsilon}| + |\hat{\sigma}_\varepsilon - \sigma_0|),$$

uniformly in z , provided f_0 is smooth and $\tilde{\varepsilon}, \hat{\sigma}_\varepsilon - \sigma_0 = o_p(1)$. Under (C4), $\tilde{\varepsilon} = o_p(n^{-1/2})$ and $\hat{\sigma}_\varepsilon - \sigma_0 = o_p(1)$; the centering term contributes at most $o_p(1)$ to $\sqrt{n}(\hat{F}_n - F_0)$ if $\tilde{\varepsilon} = o_p(n^{-1/2})$. Concretely, the leading additional contribution from scaling is

$$-zf_0(z\sigma_0) \cdot \sqrt{n}(\hat{\sigma}_\varepsilon - \sigma_0),$$

but under the stated consistency (C4) this is $o_p(1)$ uniformly in bounded z . If one needs to control unbounded z (tails for AD integral), truncation arguments or tail-weighted norms are used; assumptions on moments of ε handle those technicalities.

Thereafter, collecting terms and noting that the leading stochastic piece $\sqrt{n}(F_n - F_0)$ is $O_p(1)$ in ℓ^∞ , we obtain the decomposition

$$\mathbb{G}_n(z) = \mathbb{G}_n^0(z) - f_0(z) \cdot \frac{1}{\sqrt{n}} \sum_{j=1}^n \Phi(X_j, Y_j) + \frac{1}{2}f'_0(z) \cdot \sqrt{n} \overline{\Delta^2}_n + o_p(1),$$

uniformly in z . Hence if $\frac{1}{\sqrt{n}} \sum_{j=1}^n \Phi(X_j, Y_j) = o_p(1)$ and $\sqrt{n} \overline{\Delta^2}_n = o_p(1)$ (the conditions in the theorem), the extra bias terms are $o_p(1)$ in sup-norm and the residual empirical process is asymptotically equivalent to the classical empirical process $\mathbb{G}_n^0$, which converges to the Brownian bridge G by (C1),(C4). This completes the proof.

Proof of Theorem 3.4. Write the decomposition (proved in the residual empirical process section)

$$\hat{F}_n - F_0 = n^{-1/2}G_n + B_n,$$

with $G_n = \sqrt{n}(F_n - F_0)$ and $B_n = \hat{F}_n - F_n$. Substitute into the AD functional:

$$\begin{aligned} A_n^2 &= n \int \frac{(n^{-1/2}G_n + B_n)^2}{F_0(1 - F_0)} dF_0 \\ &= \int \frac{G_n^2}{F_0(1 - F_0)} dF_0 + 2\sqrt{n} \int \frac{G_n B_n}{F_0(1 - F_0)} dF_0 \\ &\quad + n \int \frac{B_n^2}{F_0(1 - F_0)} dF_0. \end{aligned}$$

By assumption $\sqrt{n}\|B_n\|_\infty = o_p(1)$, and by Donsker theory $\|G_n\|_\infty = O_p(1)$ and

$$\int \frac{G_n^2}{F_0(1-F_0)} dF_0 \Rightarrow \int_0^1 \frac{G(u)^2}{u(1-u)} du.$$

Use the Cauchy–Schwarz bound as in Theorem 2.4 to control the mixed term:

$$\left| 2\sqrt{n} \int \frac{G_n B_n}{F_0(1-F_0)} dF_0 \right| \leq 2\sqrt{n}\|B_n\|_\infty \left(\int \frac{G_n^2}{F_0(1-F_0)} dF_0 \right)^{1/2} = o_p(1).$$

Similarly,

$$n \int \frac{B_n^2}{F_0(1-F_0)} dF_0 \leq n\|B_n\|_\infty^2 \int \frac{1}{F_0(1-F_0)} dF_0 = o_p(1).$$

Therefore $A_n^2 = \int G_n^2/(F_0(1-F_0)) dF_0 + o_p(1)$, and application of the continuous mapping theorem yields the claimed limit.

Sketch of proof of Theorem 3.6. The argument follows the general template for bootstrap consistency of smooth functionals of empirical processes when nuisance smoothing is present (see e.g. Stute (1997), Liu (1988), and bootstrap theory for nonparametric regression). Outline:

- i) Write the bootstrap residual EDF $\hat{F}_n^*$ in the same decomposition form:

$$\hat{F}_n^* - \hat{F}_n = n^{-1/2}G_n^* + B_n^*,$$

where G_n^* is the bootstrap empirical process (conditional on data) and B_n^* is the bootstrap bias part induced by re-fitting $\hat{m}^*$.

- ii) Show that conditional on data, $G_n^* \Rightarrow G$ in probability (bootstrap Donsker), i.e. the bootstrap process mimics the law of G_n .
- iii) Show that the bootstrap bias B_n^* consistently estimates the true bias B_n in sup-norm, i.e. $\|B_n^* - B_n\|_\infty = o_p(n^{-1/2})$ (this uses uniform consistency of $\hat{m}$ and smoothness of the refitting operator). In particular, even when B_n is non-negligible, the bootstrap reproduces it.
- iv) Using a functional delta / continuous mapping argument for the quadratic AD functional, combine the above to obtain

$$\mathcal{L}^*(A_n^{2*} \mid \text{data}) \xrightarrow{p} \mathcal{L}(A_n^2).$$

Technical details require verifying equicontinuity and integrability of the AD functional (truncation near $u = 0, 1$ is used to handle weight blow-up), but these are standard under (B1)–(B5). Thus the bootstrap distribution approximates the sampling law of A_n^2 uniformly, proving the theorem.

¹ DEPARTMENT OF MATHEMATICS, BRAINWARE UNIVERSITY, KOLKATA, INDIA.

Email address: sthdas999@gmail.com