Uniform Estimation and Inference for Nonparametric Partitioning-Based M-Estimators

Matias D. Cattaneo[*] Yingjie Feng[†] Boris Shigida[‡]

September 1, 2025

Abstract

This paper presents uniform estimation and inference theory for a large class of nonparametric partitioning-based M-estimators. The main theoretical results include: (i) uniform consistency for convex and non-convex objective functions; (ii) rate-optimal uniform Bahadur representations; (iii) rate-optimal uniform (and mean square) convergence rates; (iv) valid strong approximations and feasible uniform inference methods; and (v) extensions to functional transformations of underlying estimators. Uniformity is established over both the evaluation point of the nonparametric functional parameter and a Euclidean parameter indexing the class of loss functions. The results also account explicitly for the smoothness degree of the loss function (if any), and allow for a possibly non-identity (inverse) link function. We illustrate the theoretical and methodological results in four examples: quantile regression, distribution regression, L_p regression, and Logistic regression. Many other possibly non-smooth, nonlinear, generalized, robust M-estimation settings are covered by our results. We provide detailed comparisons with the existing literature and demonstrate substantive improvements: we achieve the best (in some cases optimal) known results under improved (in some cases minimal) requirements in terms of regularity conditions and side rate restrictions. The supplemental appendix reports complementary technical results that may be of independent interest, including a novel uniform strong approximation result based on Yurinskii's coupling.

Keywords: nonparametric estimation and inference, series methods, partitioning estimators, quantile regression, nonlinear regression, robust regression, generalized linear models, uniform distribution theory.

[*]Department of Operations Research and Financial Engineering, Princeton University.
[†]School of Economics and Management, Tsinghua University.
[‡]Department of Operations Research and Financial Engineering, Princeton University.

Contents

1 Introduction **1**
- 1.1 Partitioning-Based Methodology . . . 1
- 1.2 Summary of Contributions . . . 3
- 1.3 Related Literature . . . 4
- 1.4 Notation . . . 5

2 Examples **5**

3 Assumptions **7**

4 Consistency **10**
- 4.1 Convex Loss Function . . . 10
- 4.2 Non-Convex Loss Function . . . 11
- 4.3 Weaker Conditions for Special Cases . . . 12
- 4.4 Comparison with Existing Results . . . 12

5 Bahadur Representation and Convergence Rates **13**
- 5.1 Rates of Convergence . . . 14
- 5.2 Comparison with Existing Results . . . 15

6 Strong Approximation **17**
- 6.1 Comparison with Existing Results . . . 19

7 Feasible Plug-in Uniform Inference **20**

8 Verification of Assumptions in Examples **22**
- 8.1 Example 1: Generalized Conditional Quantile Regression . . . 22
- 8.2 Example 2: Generalized Conditional Distribution Regression . . . 23
- 8.3 Example 3: Generalized L_p Regression . . . 23
- 8.4 Example 4: Logistic Regression . . . 24

9 Other Parameters of Interest **24**

10 Conclusion **26**

A Introduction to supplemental material **30**
- A.1 Notation . . . 30
- A.2 Organization . . . 31
- A.3 Comparison to prior literature . . . 33

B Full assumptions **33**
- B.1 Simplifying assumption on loss function . . . 37

C Frequently used lemmas **42**

D Consistency **53**
D.1 Convex case . . . 53
D.2 Nonconvex case . . . 56
D.3 Weaker conditions for special cases . . . 60

E Bahadur representation **63**
E.1 Proof: convex case . . . 64
E.2 Proof: general case . . . 74
E.3 Rates of convergence . . . 77

F Strong approximation **79**
F.1 Yurinskii coupling . . . 79
F.2 Main Result . . . 88

G Uniform inference **94**
G.1 Plug-in approximation . . . 94
G.2 Confidence bands . . . 102
G.3 Kolmogorov-Smirnov distance bound . . . 105

H Examples **111**
H.1 Quantile regression . . . 111
H.2 Distribution regression . . . 116
H.3 L_p regression . . . 120
H.4 Logistic regression . . . 123
H.5 Examples of basis functions . . . 126

I Other parameters of interest **127**

J Simulation Evidence **129**

List of important statements

1 Assumption (Quasi-uniform partition) . . . 2
2 Assumption (Local basis) . . . 2
1 Example (Generalized Conditional Quantile Regression) . . . 6
2 Example (Generalized Conditional Distribution Regression) . . . 6
3 Example (Generalized L_p Regression) . . . 6
4 Example (Generalized Linear Models) . . . 7
3 Assumption (Data Generating Process) . . . 7
4 Assumption (Loss Function) . . . 8
5 Assumption (Simplified Setup) . . . 9
6 Assumption (Approximation Error) . . . 10
1 Lemma (Consistency, convex case) . . . 10
2 Lemma (Consistency, non-convex case) . . . 11
1 Theorem (Bahadur representation) . . . 13
1 Corollary (Uniform Rate of Convergence) . . . 14
2 Corollary (Mean Square Rate of Convergence) . . . 14
2 Theorem (Strong approximation) . . . 18
3 Theorem (Feasible Plug-in Strong Approximation) . . . 21
1 Proposition (Quantile Regression) . . . 22
2 Proposition (Distribution Regression) . . . 23
3 Proposition (L_p Regression) . . . 23
4 Proposition (Logit Estimation) . . . 24
B.1 Assumption (Quasi-uniform partition) . . . 34
B.2 Assumption (Local basis) . . . 35
B.3 Assumption (Data generating process) . . . 35
B.4 Assumption (Loss function) . . . 35
B.5 Assumption (Complexity of $\{\psi(\cdot, \cdot; \boldsymbol{q})\}$.) . . . 36
B.6 Assumption (Approximation error) . . . 37
B.7 Assumption (Estimator of the Gram matrix) . . . 37
B.8 Assumption (Additional complexity assumption) . . . 37
B.9 Proposition (Simplified setup) . . . 37
E.1 Theorem (Bahadur representation) . . . 63
E.16 Corollary (Uniform rate of convergence) . . . 77
E.17 Corollary (Mean square rate of convergence) . . . 77
F.1 Theorem (Conditional Strassen's theorem) . . . 79
F.2 Theorem (Conditional Yurinskii coupling, d-norm) . . . 79
F.4 Theorem (Strong approximation) . . . 88
G.1 Theorem (Plug-in approximation) . . . 94
G.7 Theorem (Confidence bands) . . . 102
G.11 Theorem (Kolmogorov-Smirnov Distance) . . . 105
H.1 Proposition (Verification of Assumption B.4 for quantile regression) . . . 111
H.2 Proposition (Verification of the conditions of Lemma D.3) . . . 115
H.3 Proposition (Verification of assumptions for distribution regression) . . . 116
H.5 Proposition (Verification of Assumption B.4 for L_p regression) . . . 120
H.6 Proposition (Verification of Assumption B.4 and others for logistic regression) . . . 123
H.9 Proposition (Verifying Assumptions B.2 and B.6) . . . 127
I.1 Theorem (Other parameters of interest) . . . 127

List of lemmas

1	Lemma (Consistency, convex case)	10
2	Lemma (Consistency, non-convex case)	11
B.10	Lemma (Class G_1)	38
B.11	Lemma (Class G_2)	38
B.12	Lemma (Class G_3)	39
B.13	Lemma (Class G_4)	40
B.14	Lemma (Class G_5)	40
C.1	Lemma (Second moment bound of the max of sub-Gaussian random variables)	43
C.2	Lemma (Boundedness of conditional expectation in probability implies unconditional boundedness in probability)	43
C.3	Lemma (Converging to zero in conditional probability is the same as converging to zero in probability)	43
C.4	Lemma (Permanence properties of the uniform entropy bound)	43
C.5	Lemma (Maximal inequality for Gaussian vectors)	44
C.6	Lemma (A maximal inequality for i.n.i.d. empirical processes)	44
C.7	Lemma (Maximal inequalities for Gaussian processes)	45
C.8	Lemma (Closeness in probability implies closeness of conditional quantiles)	45
C.9	Lemma (Anti-concentration for suprema of separable Gaussian processes)	45
C.10	Lemma (The existence of $\mu(\cdot)$)	45
C.11	Lemma (Gram matrix)	46
C.12	Lemma (Uniform convergence: variance)	48
C.14	Lemma (Projection of approximation error)	50
D.1	Lemma (Consistency, convex case)	53
D.2	Lemma (Preparation for consistency in the nonconvex case)	56
D.3	Lemma (Consistency, nonconvex case)	57
D.4	Lemma (Consistency, convex case, unconnected basis functions)	60
D.5	Lemma (Consistency in the nonconvex case: unconnected basis)	61
D.6	Lemma (Consistency: strongly convex and strongly smooth loss case)	62
E.3	Lemma (Uniform concentration)	68
E.4	Lemma (Bonding variance)	70
E.5	Lemma (Complexity of class $\mathcal{G}$)	71
E.6	Lemma (Bounding T_1)	72
E.7	Lemma (Bounding T_2)	73
E.8	Lemma (Bounding T_3)	74
E.10	Lemma (Bounding variance)	75
E.11	Lemma (Complexity of class $\mathcal{G}$)	75
E.12	Lemma (Uniform concentration in $\mathcal{G}$)	75
E.13	Lemma	75
E.14	Lemma	75
E.15	Lemma	76
F.3	Lemma (Yurinskii coupling: K-dimensional process)	81
F.6	Lemma (Asymptotic variance)	89
F.7	Lemma (Closeness of linear terms)	90
F.8	Lemma (Hats off)	93
G.3	Lemma (Solution to the program)	96
G.4	Lemma (Consistency of $\widehat{\boldsymbol{\Sigma}}^{+}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$)	96

G.5 Lemma (The matrices $\bar{\mathbf{\Sigma}}^{\delta_n}$ and $\widehat{\mathbf{\Sigma}}^{\delta_n}$ are close) . . . 97
G.6 Lemma (Bounding the supremum of the Gaussian process) . . . 100
G.8 Lemma (Closeness rates) . . . 102
G.9 Lemma (First sequence) . . . 103
G.10 Lemma (Second sequence) . . . 103
G.12 Lemma . . . 105
G.13 Lemma (The matrices $\bar{\mathbf{\Sigma}}^{\delta_n}$ and $\mathbf{\Sigma}^{\delta_n}$ are close) . . . 109
H.7 Lemma (Class G_1) . . . 124
H.8 Lemma (Class $\mathcal{G}_3$) . . . 125

1 Introduction

Let $(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \cdots, (y_n, \boldsymbol{x}_n)$ be independent and identically distributed copies of the random vector $(Y, \boldsymbol{X}) \in \mathcal{Y} \times \mathcal{X} \subseteq \mathbb{R} \times \mathbb{R}^d$. Given a loss function $\rho\colon \mathcal{Y} \times \mathcal{E} \times \mathcal{Q} \to \mathbb{R}$ with $\mathcal{E} \subseteq \mathbb{R}$ an open connected set and $\mathcal{Q} \subseteq \mathbb{R}$ a connected compact set, and $\eta\colon \mathbb{R} \to \mathcal{E}$ a strictly monotonic transformation function, consider the functional parameter $\mu_0\colon \mathcal{X} \times \mathcal{Q} \to \mathbb{R}$ satisfying

$$\mu_0(\cdot, q) \in \operatorname*{arg\,min}_{\mu \in \mathscr{M}} \mathbb{E}\big[\rho\big(y_i, \eta\big(\mu(\boldsymbol{x}_i)\big); q\big)\big], \tag{1.1}$$

where the minimization is over the space of measurable functions from $\mathcal{X}$ to $\mathbb{R}$. In particular, we assume that the (local) minimum is achieved, which is true in most cases. This setup covers settings of interest in nonparametric statistics, econometrics, and data science, including generalized linear models, robust nonlinear regression, and generalized conditional quantile regression. In practice, the parameter of interest may be μ_0 itself, or otherwise specific transformations thereof such as $\eta(\mu_0(\cdot, \cdot))$ or its partial derivatives. This paper presents uniform over $\mathcal{X} \times \mathcal{Q}$ estimation and inference results for μ_0, and transformations thereof, based on nonparametric partitioning-based M-estimation.

1.1 Partitioning-Based Methodology

The series (or sieve) nonparametric partitioning-based M-estimator is

$$\widehat{\mu}(\boldsymbol{x}, q) = \boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}} \widehat{\boldsymbol{\beta}}(q), \qquad \widehat{\boldsymbol{\beta}}(q) \in \operatorname*{arg\,min}_{\boldsymbol{b} \in \mathcal{B}} \sum_{i=1}^{n} \rho\big(y_i, \eta(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}} \boldsymbol{b}); q\big), \tag{1.2}$$

where $\mathcal{B} \subseteq \mathbb{R}^K$ is the feasible set of the optimization problem, and $\boldsymbol{x} \mapsto \boldsymbol{p}(\boldsymbol{x}) = \boldsymbol{p}(\boldsymbol{x}; \Delta, m) = \big(p_1(\boldsymbol{x}; \Delta, m), \ldots, p_K(\boldsymbol{x}; \Delta, m)\big)^{\mathsf{T}}$ is a dictionary of K locally supported basis functions of order m based on a quasi-uniform partition $\Delta = \{\delta_l : 1 \le l \le \kappa\}$ containing a collection of open disjoint polyhedra in $\mathcal{X}$ such that the closure of their union covers $\mathcal{X}$.

[24] give a textbook introduction to the partitioning-based estimation literature. In this nonparametric framework, the *Haar* basis

$$\boldsymbol{p}_{\mathtt{H}}(\boldsymbol{x}) = \big(\mathbb{1}(\boldsymbol{x} \in \delta_1), \ldots, \mathbb{1}(\boldsymbol{x} \in \delta_\kappa)\big)^{\mathsf{T}}$$

corresponds to the canonical basis with $m = 1$ and $K = \kappa$, and is an essential building block for the construction of other basis functions. Since the Haar basis is "unconnected" across cells (i.e., each basis function is supported on a single cell), the resulting estimator in (1.2) reduces to K separate M-estimators, each only using observations with $\boldsymbol{x}_i \in \delta_k$, for $k = 1, \ldots, K$. For estimation and inference, "small" cells decrease bias but increase variance, while "large" cells have the opposite effect.

A natural generalization is the *piecewise polynomial* basis

$$\boldsymbol{p}_{\mathtt{P}}(\boldsymbol{x}) = \boldsymbol{p}_{\mathtt{H}}(\boldsymbol{x}) \otimes \boldsymbol{r}_m(\boldsymbol{x}),$$

where the vector $\boldsymbol{r}_m(\boldsymbol{x})$ contains the unique terms of an $(m-1)$th-degree polynomial expansion based on $\boldsymbol{x}$ and $\otimes$ is the Kronecker product operator, and thus $K = \frac{(m+d-1)!}{(m-1)!d!}\kappa$. The resulting piecewise polynomial fit within each cell gives more flexible approximation, thus decreasing bias, but the estimation approach remains unconnected.

Since the piecewise polynomial estimator may be discontinuous over $\mathcal{X}$, it is sometimes preferred to impose smoothness restrictions across cells: for example, if the partition Δ admits a tensor product representation with equal number of partitions along the d axes, then the *Splines* basis is

$$\boldsymbol{p}_{\mathtt{S}}(\boldsymbol{x}) = \otimes_{k=1}^{d} \boldsymbol{T}_s \boldsymbol{p}_{\mathtt{P}}(\boldsymbol{e}_k^{\mathsf{T}}\boldsymbol{x}),$$

where $\boldsymbol{e}_k$ denotes the k-th unit vector ($1 \leq k \leq d$), and $\boldsymbol{T}_s$ denotes a transformation matrix that ensures the estimator $\boldsymbol{x} \mapsto \widehat{\mu}(\boldsymbol{x}, q)$ is $(s-1)$-times continuously differentiable ($s < m$) over $\mathcal{X}$, and thus $K = ((m-s)\kappa^{1/d} + s)^d$. Due to the global smoothness restrictions, the spline basis is no longer unconnected, and the resulting estimator in (1.2) cannot be reduced to separate local estimators. Other spline constructions on more general partitioning schemes are available, and compactly supported wavelets are yet another example of a local basis constructed recursively out of the Haar basis. See [4], [10], [11], and [15] for more discussion on these and other basis of approximation. Furthermore, partitioning-based estimation naturally arises in the recursive partitioning literature [20, 43].

To enable good statistical performance, we need to restrict the partition of $\mathcal{X}$, and the local basis constructed on it. The first assumption concerns the regularity of the cells in the partition. Let $a_n \lesssim b_n$ denote $\limsup_{n\to\infty} |a_n/b_n| < \infty$.

Assumption 1 (Quasi-uniform partition)**.** *The ratio of the sizes of inscribed and circumscribed balls of each $\delta \in \Delta$ is bounded away from zero uniformly in $\delta \in \Delta$, and*

$$\frac{\max\{\mathrm{diam}(\delta) : \delta \in \Delta\}}{\min\{\mathrm{diam}(\delta) : \delta \in \Delta\}} \lesssim 1$$

where $\mathrm{diam}(\delta)$ denotes the diameter of δ. Further, for $h = \max\{\mathrm{diam}(\delta) : \delta \in \Delta\}$, assume $h = o(1)$ and $\log(1/h) \lesssim \log n$, as $n \to \infty$.

Assumption 1 requires the partition Δ be quasi-uniform: the elements in the partition Δ do not differ too much in size asymptotically. As a consequence, we can use the maximum diameter h as a universal measure of mesh sizes. The next assumption requires the basis be "locally supported", non-collinear, and bounded in a proper sense. A function $p(\cdot)$ on $\mathcal{X}$ is said to be *active* on $\delta \in \Delta$ if it is not identically zero on δ; we also employ standard multi-index notation (see Section 1.4 for details).

Assumption 2 (Local basis)**.**

(i) *For each basis function p_k, $k = 1, \ldots, K$, the union of elements of Δ on which p_k is active is a connected set, denoted by $\mathcal{H}_k$. For all $k = 1, \ldots, K$, both the number of elements of $\mathcal{H}_k$ and the number of basis functions which are active on $\mathcal{H}_k$ are bounded by a constant.*

(ii) *For any $\boldsymbol{a} = (a_1, \ldots, a_K)^{\mathsf{T}} \in \mathbb{R}^K$, $a_k^2 h^d \lesssim \boldsymbol{a}^{\mathsf{T}} \int_{\mathcal{H}_k} \boldsymbol{p}(\boldsymbol{x})\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}} d\boldsymbol{x}\, \boldsymbol{a}$ for $k = 1, \ldots, K$.*

(iii) *Let $|\boldsymbol{v}| < m$. There exists an integer $\varsigma \in [|\boldsymbol{v}|, m)$ such that, for all $\boldsymbol{\varsigma}, |\boldsymbol{\varsigma}| \leq \varsigma$,*

$$h^{-|\boldsymbol{\varsigma}|} \lesssim \inf_{\delta\in\Delta} \inf_{\boldsymbol{x}\in\mathrm{cl}(\delta)} \left\|\boldsymbol{p}^{(\boldsymbol{\varsigma})}(\boldsymbol{x})\right\| \leq \sup_{\delta\in\Delta} \sup_{\boldsymbol{x}\in\mathrm{cl}(\delta)} \left\|\boldsymbol{p}^{(\boldsymbol{\varsigma})}(\boldsymbol{x})\right\| \lesssim h^{-|\boldsymbol{\varsigma}|},$$

where $\mathrm{cl}(\delta)$ is the closure of δ.

In Assumption 2, condition (i) implies that each basis function in $\boldsymbol{p}(\boldsymbol{x})$ is supported by a region consisting of a finite number of cells in Δ (independent of n). Then, as $\kappa \to \infty$, all basis functions are locally supported relative to the whole support of the data. Condition (ii) can be read as "non-collinearity" of the basis functions in $\boldsymbol{p}(\boldsymbol{x})$. Since local support condition has been imposed, it suffices to require the basis functions are not too collinear "locally". Condition (iii) controls the magnitude of the local basis in a uniform sense.

Assumptions 1 and 2 implicitly relate the number of approximating series terms, the number of cells in Δ, and the maximum mesh size: $K \asymp \kappa \asymp h^{-d}$, where $a_n \asymp b_n$ means $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Under appropriate assumptions on the statistical model (Assumption 3 in Section 3), the parameter m will control how well μ_0 can be approximated by linear combinations of the local basis (via Assumption 6 in Section 3). We consider large sample approximations where d and m are fixed constants, and $\kappa \to \infty$ (and thus $K \asymp h^{-d} \to \infty$) as $n \to \infty$. As a consequence, appropriate choices of Δ and $\boldsymbol{p}(\cdot)$ will enable valid nonparametric approximations of μ_0, and transformations thereof, in large samples.

In practice, the parameter (1.1) and its associated plug-in estimator (1.2) may not be unique (e.g., when the objective function is not convex and hence several local minima may exist). In such cases the interpretation of the estimator and its probability limit may depend on the specific (algorithmic) implementation used. This paper does not study these additional complications, but rather assumes that the estimator (1.2) has been computed, and then relies on the assumptions in Section 3 concerning the data generating process to study the large sample statistical properties of the partitioning-based M-estimator and transformations thereof.

1.2 Summary of Contributions

The objective function in (1.2) may not be convex. To address this challenge, we first provide primitive conditions for uniform over $\mathcal{X} \times \mathcal{Q}$ (and mean square) consistency of the partitioning-based estimator $\widehat{\mu}(\boldsymbol{x}, q)$, taking explicitly into account whether the loss function $\theta \mapsto \rho(y, \eta(\theta); q)$ is convex: setting $\mathcal{B} = \mathbb{R}^K$ if it is convex, or otherwise $\mathcal{B} = \{\boldsymbol{b} \in \mathbb{R}^K : \|\boldsymbol{b}\|_\infty \leq R\}$ for some large enough fixed constant $R > 0$, we establish $\sup_{q\in\mathcal{Q}} \|\widehat{\boldsymbol{\beta}}(q) - \boldsymbol{\beta}_0(q)\|_\infty = o_{\mathbb{P}}(1)$, where $\|\cdot\|_\infty$ denotes the ℓ^∞-norm, and $\boldsymbol{\beta}_0 \colon \mathcal{Q} \to \mathbb{R}^K$ denotes coefficients such that $\boldsymbol{\beta}_0(q)^\mathsf{T}\boldsymbol{p}$ approximates μ_0 well enough uniformly over $\mathcal{X} \times \mathcal{Q}$ (Assumption 6 in Section 3). For the non-convex case, the resulting "fixed box" constrained optimization is arguably a mild assumption in practice, and may be justified in theory under different regularity conditions. These results are presented in Section 4.

Taking the uniform consistency of the partitioning-based estimator as given, and hence being agnostic about the shape of the objective function and other optimization-related aspects, we establish three theoretical results for the partitioning-based series M-estimator in (1.2):

(i) rate-optimal Bahadur representation uniformly over $\mathcal{X} \times \mathcal{Q}$,

(ii) rate-optimal convergence rates in mean square and uniformly over $\mathcal{X} \times \mathcal{Q}$, and

(iii) valid strong approximation and feasible distribution theory uniformly over $\mathcal{X} \times \mathcal{Q}$.

These results allow for a large class of possibly non-smooth loss functions. In addition, we precisely characterize how the degrees of smoothness of ρ and η affect the order of the remainder in the uniform Bahadur representation for $\widehat{\mu}$, its convergence rates, and the validity of the associated uniform inference procedures. Results (i) and (ii) are presented in Section 5, while results (iii) are presented in Sections 6 and 7.

Section 2 introduces four examples: *Generalized Conditional Quantile Regression*, *Generalized Conditional Distribution Regression*, *Generalized L_p Regression Estimation*, and *Maximum Likelihood Logistic Regression*. These examples are used to both motivate our high-level assumptions and demonstrate the broad applicability of our uniform estimation and inference results. Our most general results cover other applications such as nonparametric partitioning-based (quasi-maximum likelihood) Poisson regression, censored and truncated regression, as well as Tukey and Huber regression. Section 3 presents the slightly simplified high-level technical assumptions used throughout the paper, but their most general form is given in the supplemental appendix to streamline the presentation. Section 8 demonstrates how our general sufficient conditions are verified for each of our motivating examples.

Section 9 discusses how our results can be extended to cover other parameters of interest, while Section 10 concludes. The supplemental appendix reports simulation evidence, collects all the technical proofs, presents other theoretical results that may be of independent interest. In particular, our more general theoretical results (i) allow for $\mathcal{Q}$ to be a set of vectors rather than scalars, which can be useful in other examples beyond those studied in this paper; and (ii) consider more complex (VC-type) classes of loss and transformation functions, thereby covering a broader class of settings than those studied herein, but at the cost of additional, cumbersome notation and technicalities. In addition, the supplemental appendix presents new strong approximation results for a class of K-dimensional linear stochastic processes indexed by $\mathcal{X}\times\mathcal{Q}$ under standard complexity and smoothness conditions, leveraging a conditional Strassen's Theorem [16, 36] and generalizing prior Yurinskii's coupling results in the literature [3, 42].

1.3 Related Literature

Our paper contributes to the literature on nonparametric curve estimation and inference, focusing in particular on series (or sieve) partitioning-based methods. See, for example, [23] and [24] for textbook introductions. This literature is mature and well-developed for the special case of a square loss function $\rho(y, \eta(\theta); q) = (y - \theta)^2$ with identity transformation $\eta(u) = u$, and hence not a function of $q \in \mathcal{Q}$. See, for example, [4], [10], [11], [9], [15], [27], [44] for pointwise and uniform over $\mathcal{X}$ estimation and inference results at different levels of generality, and with increasingly weaker technical conditions. This strand of the literature explicitly exploits the special structure, which leads to a closed-form solution of the estimator in (1.2), and hence results are often obtained under minimal assumptions and technical regularity conditions. To be more precise, up to polylog(n) terms and mild regularity conditions, [11] show that the minimal requirement $K/n \to 0$ is (necessary and) sufficient for rate-optimal convergence rates for any $d \geq 1$, and for strong approximations uniformly over $\mathcal{X}$ when $d = 1$. They also establish valid strong approximations uniformly over $\mathcal{X}$ for $d > 1$ under the requirement $K^3/n \to 0$, up to polylog(n) terms and mild regularity conditions.

Despite aiming for generality, that is, allowing for a large class of loss functions with different levels of smoothness and a non-identity transformation function, this paper establishes rate-optimal uniform over *both* $\mathcal{X}$ and $\mathcal{Q}$ estimation results under the same minimal assumption $K/n \to 0$ for unconnected bases, and under the slightly stronger assumption $K^2/n \to 0$ for general partitioning-based estimators. Furthermore, we establish valid uniform over *both* $\mathcal{X}$ and $\mathcal{Q}$ inference under the same condition $K^3/n \to 0$, leveraging a new strong approximation result given in the supplemental appendix. Compared to the prior literature focusing on the special case of square loss and identity transformation, we are able to achieve the same best known (in some cases rate-optimal) estimation and inference results, under the same (in some cases minimal) side rate restrictions and conditions on the partitioning-based method (Assumptions 1 and 2). Furthermore, our results on uniform consistency disentangling convex and non-convex loss functions (Section 4), rate-optimal uniform Bahadur representation and convergence capturing explicitly the smoothness degree of the loss function (Section 5), and uniform feasible inference validity (Sections 6 and 7), are necessarily new relative to prior work studying the special case of least square partitioning-based methods.

Going beyond square loss and identity transformation, there are only a handful of results available in the literature. The closest antecedent is [3], who consider nonparametric conditional quantile series regression estimation and inference uniformly over $\mathcal{X} \times \mathcal{Q}$ with $\eta(u) = u$, and under the side rate restriction $K^4/n \to 0$, up to $\text{polylog}(n)$ terms, and other regularity conditions. As a comparison, for the special case of nonparametric quantile regression (Example 1 below), this paper allows for a non-identity (inverse) link function $\eta(\cdot)$, and establishes convergence rates under the minimal condition $K/n \to 0$ for piecewise polynomials, and the improved condition $K^2/n \to 0$ for connected bases, while for uniform inference we require the weaker condition $K^3/n \to 0$, in all cases up to $\text{polylog}(n)$ terms. We also weaken other assumptions imposed in [3]: see Example 1 in Sections 2 and 8.1, and Sections 5.2 and 6.1. On the other hand, it is worth noting that [3] also consider generic dimension-increasing covariates, while our paper focuses exclusively on partitioning-based local basis.

Our contributions can also be compared to recent work on nonparametric M-estimation employing other smoothing techniques. For example, [30] considers local polynomial methods, and [38] considers smoothing spline methods. Sections 5.2 and 6.1 give a more detailed comparison with prior literature, and explain precisely how our general results are either on par with or improve upon prior work. In a nutshell, we present estimation and inference results for partitioning-based M-estimators that (i) allow for a large class of possibly non-smooth loss functions, (ii) are uniformly valid over both $\mathcal{X}$ and $\mathcal{Q}$, (iii) achieve the best known (in some cases rate-optimal) convergence rates, and (iv) require substantially weaker (in some cases minimal) side rate restrictions and regularity conditions. Our results, for example, permit the use of Haar M-estimators for estimation and inference, which were ruled out by prior work.

1.4 Notation

We employ standard notation in probability, statistics and empirical process theory [5, 22, 28, 41]. For any vector $\boldsymbol{a} = (a_1, \cdots, a_M) \in \mathbb{R}^M$, we write $\|\boldsymbol{a}\| = (\sum_{j=1}^M a_j^2)^{1/2}$ and $\|\boldsymbol{a}\|_\infty = \max_{1 \le j \le M} |a_j|$. For any real function f depending on d variables $(t_1, \ldots, t_d)$ and any vector $\boldsymbol{v} = (v_1, \cdots, v_d)$ of nonnegative integers, denote $f^{(\boldsymbol{v})} = \frac{\partial^{|\boldsymbol{v}|}}{\partial t_1^{v_1} \ldots \partial t_d^{v_d}} f$ where $|\boldsymbol{v}| = \sum_{k=1}^d v_j$. For functions that depend on $(\boldsymbol{x}, q)$, the multi-index derivative notation is taken with respect to the first argument $\boldsymbol{x}$, unless otherwise noted. We say a function f is α-Hölder on a set $\mathcal{I}$ if for some constant $C > 0$ and $\alpha > 0$, $|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)| \le C\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^\alpha$ for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{I}$. For any two numbers a and b, $a \vee b = \max\{a, b\}$, and $a \wedge b = \min\{a, b\}$. Let $\mathbb{E}_n[g(x_i)] = \frac{1}{n}\sum_{i=1}^n g(x_i)$ and $\mathbb{G}_n[g(x_i)] = \frac{1}{\sqrt{n}}\sum_{i=1}^n (g(x_i) - \mathbb{E}[g(x_i)])$. For sequences, $a_n = O(b_n)$ or $a_n \lesssim b_n$ denotes $\limsup_n |a_n/b_n|$ is finite, $a_n = O_{\mathbb{P}}(b_n)$ denotes $\limsup_{\epsilon \to \infty} \limsup_{n \to \infty} \mathbb{P}[|a_n/b_n| \ge \epsilon] = 0$, $a_n = o(b_n)$ denotes $a_n/b_n \to 0$, and $a_n = o_{\mathbb{P}}(b_n)$ denotes $a_n/b_n \to_{\mathbb{P}} 0$, where $\to_{\mathbb{P}}$ is convergence in probability. Limits are taken as $n \to \infty$, and the dependence on n is often suppressed, e. g. $K = K_n$. Also, we say a random variable ξ is sub-Gaussian conditional on $\boldsymbol{X}$ if for some constant $\sigma^2 > 0$, $\mathbb{P}(|\xi| \ge t | \boldsymbol{X} = \boldsymbol{x}) \le 2\exp(-t^2/\sigma^2)$ for all $t \ge 0$ and $\boldsymbol{x} \in \mathcal{X}$.

2 Examples

We discuss four motivating examples of interest covered by our theoretical results. Section 8 demonstrates how our high-level assumptions, introduced in Section 3, are verified for these examples in order to obtain uniform estimation and inference results; the supplemental appendix collects omitted details.

Our first example generalizes the work of [3], who studies the large sample properties of nonparametric conditional quantile series regression with $\eta(u) = u$. We allow for non-identity transformation under substantially weaker technical conditions.

Example 1 (Generalized Conditional Quantile Regression). *The quantile regression loss function is*

$$\rho(y, \eta; q) = (q - \mathbb{1}(y < \eta))(y - \eta),$$

where $q \in \mathcal{Q} = [\varepsilon_0, 1 - \varepsilon_0]$ denotes the quantile position, with $\varepsilon_0 > 0$. Then, $\eta(\mu_0(\boldsymbol{x}, q))$ is the q-th conditional quantile function of Y given $\boldsymbol{X} = \boldsymbol{x}$, and the partitioning-based quantile regression estimator is $\eta(\widehat{\mu}(\boldsymbol{x}, q))$ as defined in (1.2). In the classical case, $\eta(\cdot)$ is the identity function, but our theory accommodates other transformations. Interest lies on the quantile process estimator $(\eta(\widehat{\mu}(\boldsymbol{x}, q)) : (\boldsymbol{x}, q) \in \mathcal{X} \times \mathcal{Q})$, which can be used to characterize heterogeneous effects of covariates on the outcome distribution and to conduct specification testing. See Section 8.1 for our main results, and Section 9 for results on transformations. ▲

[18] obtains large sample estimation and inference results for parametric (K fixed) generalized conditional distribution regression, and applies them to counterfactual analysis and causal inference. The following example discusses a nonparametric partitioning-based generalized conditional distribution regression estimator.

Example 2 (Generalized Conditional Distribution Regression). *Non-linear least squares conditional distribution regression employs*

$$\rho(y, \eta; q) = (\mathbb{1}(y \leq q) - \eta)^2,$$

where we can, for example, use the complementary log-log link $\eta(\theta) = 1 - \exp(-\exp(\theta))$. Estimands of interest are $\eta(\mu_0(\boldsymbol{x}, q))$, which corresponds to the conditional distribution function of Y given $\boldsymbol{X} = \boldsymbol{x}$ (i.e., $F_{Y|\boldsymbol{X}}(q|\boldsymbol{x}) = \mathbb{E}[\mathbb{1}(Y \leq q)|\boldsymbol{X} = \boldsymbol{x}]$), and derivatives thereof. Uniform estimation and inference results based on $(\eta(\widehat{\mu}(\boldsymbol{x}, q)) : (\boldsymbol{x}, q) \in \mathcal{X} \times \mathcal{Q})$ are useful for a variety of purposes, including heterogeneous treatment effect estimation and specification testing. See Section 8.2 for our main results, and Section 9 for transformations. ▲

[31] studies L_p regression estimation with identity transformation $\eta(\cdot)$ in a parametric setting (K fixed). The next example considers a class of nonparametric partitioning-based generalized L_p regression with $p \in [1, 2]$, covering the full interpolation between nonparametric generalized median regression ($p = 1$) and nonparametric nonlinear least squares regression ($p = 2$), with possibly non-identity $\eta(\cdot)$.

Example 3 (Generalized L_p Regression). *The (possibly nonlinear) L_p regression estimator is defined by taking*

$$\rho(y, \eta) = |y - \eta|^p,$$

for a fixed $p > 0$. In particular, $p = 2$ leads to nonlinear least squares, and $p = 1$ leads to generalized least absolute deviations, when a non-identity transformation function $\eta(\cdot)$ is used. The estimand of interest is usually the transformed regression function $\eta(\mu_0(\boldsymbol{x}))$, which needs to be interpreted in context. Our general theory applies to any choice $p \in [1, 2]$, delivering uniform estimation and inference methods based on $(\widehat{\mu}(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X})$, and transformations thereof. See Section 8.3 for our main results, and Section 9 for transformations. ▲

The final example considers (nonparametric) Generalized Linear Models [35]. For specificity, we focus on (quasi-)maximum likelihood logistic regression, but our results cover many other examples within this class such as regression models with limited dependent variables (e.g., Poisson, fractional, censored and truncation regression).

Example 4 (Generalized Linear Models). *The classical logistic regression model, or binary classification with sigmoid (inverse) link, employs*

$$\rho(y,\eta) = -y\log\eta - (1-y)\log(1-\eta), \qquad \eta(\theta) = \exp(\theta)/(1+\exp(\theta)),$$

with $\mathcal{Y} = \{0,1\}$. The estimand $\eta(\mu_0(\boldsymbol{x}))$ characterizes the conditional probability of $Y = 1$ given $\boldsymbol{X} = \boldsymbol{x}$. See Section 8.4 for uniform estimation and inference methods based on $(\widehat{\mu}(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X})$, and Section 9 for transformations thereof. Furthermore, our results also cover other related quasi-maximum likelihood (and non-linear least squares) problems such as fractional regression where $\mathcal{Y} = [0,1]$. ▲

The four examples illustrate distinct settings from a technical perspective. In Example 1 uniformity over $\mathcal{X} \times \mathcal{Q}$ is of interest, and the loss function is non-smooth as a function of $\boldsymbol{x} \in \mathcal{X}$ but smooth as a function of $q \in \mathcal{Q}$. Example 2 is the antithesis of Example 1 because now the loss function is smooth as a function of $\boldsymbol{x} \in \mathcal{X}$ and non-smooth as a function of $q \in \mathcal{Q}$, while uniformity over $\mathcal{X} \times \mathcal{Q}$ is still of interest. In Example 3 only uniformity over $\mathcal{X}$ is of interest because $q \in \mathcal{Q}$ is not present in the loss function, but its smoothness depends on $p \in [0,1]$; the a. e. derivative of $\eta \mapsto \rho(y,\eta)$ ranges from discontinuous ($p = 1$), to Hölder continuous ($p \in (1,2)$), to linear ($p = 2$). Likewise, Example 4 only involves uniformity over $\mathcal{X}$ because $q \in \mathcal{Q}$ is not present in the loss function, but now the loss function is smooth and well-behaved; this last example serves as a benchmark for our theoretical development. All of the examples above have a convex loss function when $\eta(u) = u$, but can be non-convex when $\eta(\cdot)$ is not the identity function.

Our theoretical results cover other examples. For instance, Tukey and Huber regression are popular methods in robust statistics, and our theory allows for their generalizations to nonparametric partitioning-based uniform estimation and inference. Specifically, Tukey regression employs the loss function $\rho(y,\eta;q) = q^2(1-[1-(y-\eta)^2/q^2]^3)\mathbb{1}(|y-\eta| \le q) + q^2\mathbb{1}(|y-\eta| > q)$, while Huber regression uses the loss function $\rho(y,\eta;q) = (y-\eta)^2\mathbb{1}(|y-\eta| \le q) + q(2|y-\eta| - q)\mathbb{1}(|y-\eta| > q)$, where q is treated as a tuning parameter that balances the robustness and the bias of the estimation. We do not discuss these and other examples for brevity.

3 Assumptions

Our theoretical work proceeds under Assumptions 1 and 2 on the partitioning-based estimation framework, three assumptions concerning the data generating process and the loss function, and a final assumption linking the statistical model and partition-based approximation.

Assumption 3 (Data Generating Process).

(i) *$((y_i, \boldsymbol{x}_i) \colon 1 \le i \le n)$ is a random sample satisfying (1.1).*

(ii) *The distribution of $\boldsymbol{x}_i$ admits a Lebesgue density $f_X(\cdot)$ which is continuous and bounded away from zero on support $\mathcal{X} \subset \mathbb{R}^d$, where $\mathcal{X}$ is the closure of an open, connected and bounded set.*

(iii) *The conditional distribution of y_i given $\boldsymbol{x}_i$ admits a conditional density $f_{Y|X}(y|\boldsymbol{x})$ with support $\mathcal{Y}_{\boldsymbol{x}}$ with respect to some (sigma-finite) measure $\mathfrak{M}$, and $\sup_{\boldsymbol{x}\in\mathcal{X}} \sup_{y\in\mathcal{Y}_{\boldsymbol{x}}} f_{Y|X}(y \mid \boldsymbol{x}) < \infty$.*

(iv) *$\boldsymbol{x} \mapsto \mu_0(\boldsymbol{x}, q)$ is $m \ge 1$ times continuously differentiable for every $q \in \mathcal{Q}$, $\boldsymbol{x} \mapsto \mu_0(\boldsymbol{x}, q)$ and its derivatives of order no greater than m are bounded uniformly over $(\boldsymbol{x}, q) \in \mathcal{X} \times \mathcal{Q}$, and*

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \sup_{q_1\neq q_2} \frac{|\mu_0(\boldsymbol{x}, q_1) - \mu_0(\boldsymbol{x}, q_2)|}{|q_1 - q_2|} \lesssim 1.$$

Assumption 3 imposes standard conditions from the nonparametric regression literature, including basic support and smoothness restrictions. Minimal additional regularity is imposed to accommodate uniformity over $q \in \mathcal{Q}$, and different types of conditional distributions of $Y|\boldsymbol{X}$ (e. g., absolutely continuous, discrete or mixed) are allowed. We consider continuously distributed covariates for simplicity, but with additional notation, and by appropriate modification of our assumptions and proof, it is possible to accommodate $\boldsymbol{x}_i$ with continuous and discrete components.

The next assumption requires regularity conditions on the loss and transformation functions. Define $B_q(\boldsymbol{x}) = \{\zeta : |\zeta - \mu_0(\boldsymbol{x}, q)| \leq r\}$ for some fixed (small enough) constant $r > 0$, which is a "ball" around the true value $\mu_0(\boldsymbol{x}, q)$ with radius r.

Assumption 4 (Loss Function)**.**

(i) *Let $\mathcal{Q} \subseteq \mathbb{R}$ be a connected compact set. For each $q \in \mathcal{Q}$, $y \in \mathcal{Y}$, and some open connected subset $\mathcal{E}$ of $\mathbb{R}$ not depending on y, $\eta \mapsto \rho(y, \eta; q)$ is absolutely continuous on closed bounded intervals within $\mathcal{E}$, and admits an a. e. derivative $\psi(y, \eta; q)$.*

(ii) *The first-order optimality condition $\mathbb{E}[\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, q)); q)|\boldsymbol{x}_i] = 0$ holds; the function $\mathbb{E}[\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, q)); q)^2|\boldsymbol{x}_i = \boldsymbol{x}]$ is continuous in both arguments $(\boldsymbol{x}, q) \in \mathcal{X} \times \mathcal{Q}$, bounded away from zero, and Lipschitz in q uniformly in $\boldsymbol{x}$; there is a positive measurable envelope function $\overline{\psi}(\boldsymbol{x}_i, y_i)$ such that $\sup_{q \in \mathcal{Q}} |\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, q)); q)| \leq \overline{\psi}(\boldsymbol{x}_i, y_i)$ with $\sup_{\boldsymbol{x} \in \mathcal{X}} \mathbb{E}[\overline{\psi}(\boldsymbol{x}_i, y_i)^\nu|\boldsymbol{x}_i = \boldsymbol{x}] < \infty$ for some $\nu > 2$.*

(iii) *$\eta(\cdot)$ is strictly monotonic and twice continuously differentiable. Furthermore, for some fixed constant $\alpha \in (0, 1]$, for any $(\boldsymbol{x}, q) \in \mathcal{X} \times \mathcal{Q}$, and a pair of points $\zeta_1, \zeta_2 \in B_q(\boldsymbol{x})$, $\psi(\cdot)$ satisfies the following (constants hidden in $\lesssim$ do not depend on $\boldsymbol{x}$, q, ζ_1, ζ_2):*

- *if $\mathfrak{M}$ is the Lebesgue measure, then*

$$\sup_{\lambda \in [0,1]} \sup_{y \notin [\eta(\zeta_1) \wedge \eta(\zeta_2), \eta(\zeta_1) \vee \eta(\zeta_2)]} |\psi(y, \eta(\zeta_1 + \lambda(\zeta_2 - \zeta_1)); q) - \psi(y, \eta(\zeta_2); q)| \lesssim |\zeta_1 - \zeta_2|^\alpha,$$
$$\sup_{\lambda \in [0,1]} \sup_{y \in [\eta(\zeta_1) \wedge \eta(\zeta_2), \eta(\zeta_1) \vee \eta(\zeta_2)]} |\psi(y, \eta(\zeta_1 + \lambda(\zeta_2 - \zeta_1)); q) - \psi(y, \eta(\zeta_2); q)| \lesssim 1;$$

- *if $\mathfrak{M}$ is* not *the Lebesgue measure, then*

$$\sup_{\lambda \in [0,1]} \sup_{y \in \mathcal{Y}} |\psi(y, \eta(\zeta_1 + \lambda(\zeta_2 - \zeta_1)); q) - \psi(y, \eta(\zeta_2); q)| \lesssim |\zeta_1 - \zeta_2|^\alpha.$$

(iv) *$\Psi(\boldsymbol{x}, \eta; q) = \mathbb{E}[\psi(y_i, \eta; q)|\boldsymbol{x}_i = \boldsymbol{x}]$ is twice continuously differentiable with respect to η,*

$$\sup_{\boldsymbol{x} \in \mathcal{X}, q \in \mathcal{Q}} \sup_{\zeta \in B_q(\boldsymbol{x})} |\Psi_k(\boldsymbol{x}, \eta(\zeta); q)| < \infty, \quad \Psi_k(\boldsymbol{x}, \eta; q) = \frac{\partial^k}{\partial \eta^k} \Psi(\boldsymbol{x}, \eta; q), \quad k = 1, 2,$$

and

$$\inf_{\boldsymbol{x} \in \mathcal{X}, q \in \mathcal{Q}} \inf_{\zeta \in B_q(\boldsymbol{x})} \Psi_1(\boldsymbol{x}, \eta(\zeta); q)(\eta^{(1)}(\zeta))^2 > 0.$$

Assumption 4 is carefully crafted to accommodate all the examples discussed in Section 2, and many others. Part (i) allows for different degrees of smoothness in the loss function, assuming only absolute continuity (with respect to η). It also makes clear that q is scalar, which is assumed only to simplify the notation; see the supplemental appendix for the general case $\mathcal{Q} \subseteq \mathbb{R}^{d_\mathcal{Q}}$ with $d_\mathcal{Q} \geq 1$.

Part (ii) formalizes the idea that $\mu_0(\boldsymbol{x}, q)$ may not be a unique (global) minimizer in (1.1), and consequently it is only required to be a root of the (conditional) first-order condition; the rest of the assumptions in that part are mild regularity conditions. In some applications, $\mu_0(\boldsymbol{x}, q)$ can be the unique minimizer; see, for example, [33], [34], and references therein.

Part (iii) of Assumption 4 imposes additional structure on the a. e. first derivative of the loss function, allowing for all types of outcome data (discrete, mixed, and continuous) and rescalings emerging in some of the motivating examples. Importantly, this part characterizes precisely the role of (Hölder) smoothness, which is controlled by the parameter $\alpha \in (0, 1]$. We illustrate the full power of this general assumption in Section 8, where $\alpha = 1$ in Examples 1 and 2, $\alpha = p - 1$ in Example 3 when $p > 1$, and $\alpha = 1$ in Example 4. The strict monotonicity condition on $\eta(\cdot)$ is satisfied by usual (inverse) link functions used in generalized linear models. Finally, part (iv) of Assumption 4 collects mild regularity conditions on the smoothed-out a. e. derivative of the loss function.

Assumptions 3 and 4 have restricted basic aspects of the statistical model, imposing standard support, moment, and smoothness conditions, in addition to other minimal structure required on the loss and transformation functions. These conditions are sufficient for pointwise estimation and inference, but more is needed for uniform over $\mathcal{X} \times \mathcal{Q}$ results. In the supplemental appendix, our theoretical results are established under one more condition that governs the complexity of the loss function and related function classes. To avoid a long list of complexity bounds, we present a more restrictive but simpler assumption motivated by the examples discussed in Section 2: we consider a loss function $\rho(y, \eta; q)$ that can be expressed as a linear combination of certain "simple" functions. See Section B.1 for omitted details.

Assumption 5 (Simplified Setup)**.**

(i) *$q \mapsto \mu_0(\boldsymbol{x}, q)$ is non-decreasing.*

(ii) *$\rho(y, \eta; q) = \sum_{j=1}^{4} \rho_j(y, \eta; q)$, where the functions $\rho_j(\cdot)$ are of the following types:*

- *Type I: $\rho_1(y, \eta; q) = (f_1(y) + D_1\eta)\mathbb{1}(y \le \eta)$,*
- *Type II: $\rho_2(y, \eta; q) = (f_2(y) + D_2\eta)\mathbb{1}(y \le q)$,*
- *Type III: $\rho_3(y, \eta; q) = (f_3(y) + D_3\eta)q$,*
- *Type IV: $\rho_4(y, \eta; q) = \mathcal{T}(y, \eta)$,*

where $f_j(\cdot)$ are fixed continuous functions, D_j are universal constants, and $\mathcal{T}(\cdot)$ is a fixed continuous function.

(iii) *$\eta \mapsto \mathbb{E}[\tau(y_i, \eta)|\boldsymbol{x}_i = \boldsymbol{x}]$ is differentiable, where $\tau(y, \eta) = \frac{\partial}{\partial \eta}\mathcal{T}(y, \eta)$, $\tau(y, \eta)$ and $\frac{\partial}{\partial \eta}\mathbb{E}[\tau(y_i, \eta)|\boldsymbol{x}_i = \boldsymbol{x}]$ are continuous in their arguments and α-Hölder continuous ($\alpha \in (0, 1]$) in η for η in any fixed compact subset of $\mathcal{E}$ with the Hölder constants independent of $(y, \boldsymbol{x})$.*

(iv) *$\sup_{q \in \mathcal{Q}} |\tau(y, \eta(\mu_0(\boldsymbol{x}, q)))| \le \bar{\tau}(\boldsymbol{x}, y)$ with $\sup_{\boldsymbol{x} \in \mathcal{X}} \mathbb{E}[\bar{\tau}(\boldsymbol{x}_i, y_i)^\nu|\boldsymbol{x}_i = \boldsymbol{x}] < \infty$ for some $\nu > 2$.*

(v) *If $D_1 \neq 0$, then $F_{Y|X}$ is differentiable with a Lebesgue density $f_{Y|X}$, and $f_{Y|X}$ is continuous in both arguments and $y \mapsto f_{Y|X}(y|\boldsymbol{x})$ is continuously differentiable.*

This third assumption imposes an additional weak monotonicity condition on $\mu_0(\boldsymbol{x}, q)$ as a function of q, which is compatible with all the examples in Section 2. Finally, the key restriction emerging from Assumption 5 is on the structure of the loss function, which allows for linear combinations of smooth loss functions of y and η, and non-smooth loss functions involving indicator functions of either y and η, or y and q. These restrictions are still general enough to cover the four motivating examples: the loss function in Example 1 is a combination of Type I and Type III functions with f_1 and f_3 being linear functions of y; the loss function in Example 2 is a combination of Type II and Type IV functions; the loss function in Example 3 is of Type IV for $p > 1$, and of the same type as Example 1 when $p = 1$ (median regression); and the loss function in Example 4 is usually a Type IV function. See Section 8 for details.

Our final assumption concerns the approximation power of the basis $\boldsymbol{p}(\cdot)$ in connection with the underlying functional parameter.

Assumption 6 (Approximation Error)**.** *There exists a vector of coefficients $\boldsymbol{\beta}_0(q) \in \mathbb{R}^K$ such that for all $\boldsymbol{\varsigma}$ satisfying $|\boldsymbol{\varsigma}| \leq \varsigma$ in Assumption 2,*

$$\sup_{q\in\mathcal{Q}}\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\mu_0^{(\boldsymbol{\varsigma})}(\boldsymbol{x},q) - \boldsymbol{\beta}_0(q)^\mathsf{T}\boldsymbol{p}^{(\boldsymbol{\varsigma})}(\boldsymbol{x})\right| \lesssim h^{m-|\boldsymbol{\varsigma}|}.$$

The vector $\boldsymbol{\beta}_0(q)$ can be viewed as a pseudo-true value, and does not have to be unique. The existence of such $\boldsymbol{\beta}_0(q)$ can be established using approximation theory or related methods, and necessarily depends on the specific underlying structure of the statistical model (determining $\mu_0(\boldsymbol{x}, q)$) and the partitioning-based method (determining $\boldsymbol{p}^{(\boldsymbol{\varsigma})}(\boldsymbol{x})$). For standard local bases, the assumption can be verified by imposing smoothness conditions on $(\boldsymbol{x}, q) \mapsto \mu_0^{(\boldsymbol{\varsigma})}(\boldsymbol{x}, q)$ (see Assumption 3(iv)). For more discussion, see [4], [3], [10], [11], [15], [27], and references therein.

4 Consistency

We show that the partitioning-based M-estimator is consistent, which is the starting point for establishing its main point estimation and inference asymptotic properties. We endeavor to impose the weakest possible conditions, which requires careful consideration of the specific shape of the loss function in (1.2): we consider two cases, either the loss function $\rho(y, \eta(\theta); q)$ is convex with respect to θ or not; in the latter case, we will have to restrict the feasibility region $\mathcal{B}$.

4.1 Convex Loss Function

For the case of convex $\theta \mapsto \rho(y, \eta(\theta); q)$, consistency can be established for general unconstrained estimators ($\mathcal{B} = \mathbb{R}^K$ in (1.2)) under mild conditions. The proof is deferred to the supplemental appendix (Lemma D.1).

Lemma 1 (Consistency, convex case)**.** *Suppose that Assumptions 1–6 hold, $\rho(y, \eta(\theta); q)$ is convex with respect to θ with left or right derivative $\psi(y, \eta(\theta); q)\eta^{(1)}(\theta)$, $\mathcal{B} = \mathbb{R}^K$ in (1.2), and $m > d/2$. Furthermore, assume that one of the following two conditions holds:*

(i) $\frac{(\log n)^{\frac{\nu}{\nu-1}}}{nh^{\frac{2\nu}{\nu-1}d}} = o(1)$, *or*

(ii) $\frac{(\log n)^{3/2}}{nh^{2d}} = o(1)$ *and $\overline{\psi}(\boldsymbol{x}_i, y_i)$ is sub-Gaussian conditional on $\boldsymbol{x}_i$.*

Then

$$\sup_{q\in\mathcal{Q}}\left\|\widehat{\boldsymbol{\beta}}(q) - \boldsymbol{\beta}_0(q)\right\| = o_\mathbb{P}(1), \tag{4.1}$$

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\sup_{q\in\mathcal{Q}}\left|\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x},q) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x},q)\right| = o_\mathbb{P}(h^{-|\boldsymbol{v}|}), \tag{4.2}$$

$$\sup_{q\in\mathcal{Q}}\int\left(\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x},q) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x},q)\right)^2 f_X(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = o_\mathbb{P}(h^{d-2|\boldsymbol{v}|}). \tag{4.3}$$

Lemma 1 shows that the function estimator $\widehat{\mu}$ is uniform-in-q consistent for the true value μ_0 in both L_2-norm and sup-norm over $\mathcal{X}$, whereas for the derivative estimator $\widehat{\mu}^{(\boldsymbol{v})}$ (with $|\boldsymbol{v}| > 0$) the lemma only provides a bound on its deviation from the estimand $\mu_0^{(\boldsymbol{v})}$. Technically, all we need from this lemma to establish the Bahadur representation later is the uniform-in-q consistency of the coefficients estimator $\widehat{\boldsymbol{\beta}}(q)$ for the pseudo-true coefficients $\boldsymbol{\beta}_0(q)$ in sup-norm, i.e., $\|\widehat{\boldsymbol{\beta}}(q)-\boldsymbol{\beta}_0(q)\|_\infty = o_{\mathbb{P}}(1)$, which is immediate from (4.1), the uniform-in-q consistency in the Euclidean norm.

Two kinds of rate restrictions are imposed in Lemma 1, depending on the moment condition assumed for the generalized residual $\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, q)); q)$. In the best case when the residual has a sub-Gaussian envelope, we need $1/(nh^{2d}) \asymp K^2/n = o(1)$, up to polylog$(n)$ terms, while in the worst case when the envelope of $\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, q)); q)$ has a bounded ν-th moment with ν close to 2, we roughly need $1/(nh^{4d}) \asymp K^4/n = o(1)$, up to polylog$(n)$ terms.

A feature of Lemma 1 is that *no* constraints are imposed on the coefficients in the optimization procedure, which allows the estimation space to be, for example, piecewise polynomials. In contrast, many studies of series (or sieve) methods restrict the functions in the estimation space to satisfy certain smoothness conditions, e.g., Lipschitz continuity, to derive the uniform consistency [e.g., 19].

4.2 Non-Convex Loss Function

Consider the case when the loss $\rho(y, \eta(\theta); q)$ is possibly non-convex with respect to θ. This setting naturally arises, for example, in nonlinear regression when $\rho(y, \eta(\theta); q) = (y - \eta(\theta))^2$ with $\eta(\cdot)$ non-identity: while $\eta \mapsto \rho(y, \eta; q)$ is a square loss function, hence convex, introducing a transformation function η such as the (inverse) logistic link will often make $\theta \mapsto \rho(y, \eta(\theta); q)$ non-convex.

A proof of consistency for the unconstrained estimator in (1.2) with a non-convex loss function is not available, but we are able to establish consistency of a regularized M-estimator. Specifically, we add a *fixed* "box" constraint: for some fixed constant $R > 0$,

$$\widehat{\boldsymbol{\beta}}(q) \in \underset{\|\boldsymbol{b}\|_\infty \le R}{\arg\min} \sum_{i=1}^n \rho(y_i, \eta(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\boldsymbol{b}); q).$$

In the supplemental appendix we show that the pseudo-true coefficients $\boldsymbol{\beta}_0(q)$ from Assumption 6 are bounded in sup-norm by a universal constant: $\sup_{q\in\mathcal{Q}} \|\boldsymbol{\beta}_0(q)\|_\infty \lesssim 1$ (because $\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(q)$ has to be close to $\mu_0(\boldsymbol{x}, q)$ which is uniformly bounded). Therefore, we can always choose a sufficiently large constant R in the optimization procedure, making the box constraint set contain $\boldsymbol{\beta}_0(q)$ as an interior point. The following lemma, proven in the supplemental appendix (Lemma D.3), establishes consistency of the constrained estimator.

Lemma 2 (Consistency, non-convex case)**.** *Suppose that Assumptions 1–6 hold, $\mathcal{B} = \{\boldsymbol{b} \in \mathbb{R}^K : \|\boldsymbol{b}\|_\infty \le R\}$ with $R \ge 2\sup_{q\in\mathcal{Q}} \|\boldsymbol{\beta}_0(q)\|_\infty$ in (1.2), $m > d/2$, and that there exists some constant $c > 0$ such that $\inf \Psi_1(\boldsymbol{x}, \zeta; q) > c$, where the infimum is over $\boldsymbol{x} \in \mathcal{X}$, $q \in \mathcal{Q}$, ζ between $\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})$ and $\eta(\mu_0(\boldsymbol{x}, q))$, and $\boldsymbol{\beta} \in \mathcal{B}$. Furthermore, assume one of the following two conditions holds:*

(i) $\frac{(\log n)^{\frac{\nu}{\nu-1}}}{nh^{\frac{2\nu}{\nu-1}d}} = o(1)$, *or*

(ii) $\frac{(\log n)^{3/2}}{nh^{2d}} = o(1)$ *and $\overline{\psi}(\boldsymbol{x}_i, y_i)$ is sub-Gaussian conditional on $\boldsymbol{x}_i$.*

Then (4.1), (4.2), and (4.3) hold.

Compared to Lemma 1, two additional restrictions are imposed in this lemma. The first one, $R \geq 2\sup_{q\in\mathcal{Q}} \|\boldsymbol{\beta}_0(q)\|_\infty$, can be theoretically justified by Lemma D.2 in the supplemental appendix, and in practice a large enough R is recommended. The other restriction concerns a lower bound for Ψ_1, which implies that the (population) loss function is strongly convex in a neighborhood of the true value $\eta(\mu_0(\boldsymbol{x}, q))$, making the (constrained) minimizer well defined. (This condition does *not* break because of the shape of $\eta(\cdot)$, in contrast with the convexity of $\rho(y, \eta(\theta); q)$ in θ.) The other conditions in this lemma are the same as those in the convex case, and thus all improvements discussed before also apply to this second consistency result.

4.3 Weaker Conditions for Special Cases

In the supplemental appendix we provide additional consistency results for two special cases:

- the loss function is strongly convex and smooth (i.e., the second "derivative" of $\rho(y, \eta(\cdot); q)$ is bounded and bounded away from zero), or
- an unconnected basis (i.e., each basis function is supported on a single cell of Δ) is employed.

The first case covers the usual square loss function with identity transformation. The second case covers partitioning-based M-estimation using the (Haar and) piecewise polynomial basis. Notably, in these two special cases, the consistency result $\|\widehat{\boldsymbol{\beta}}(q) - \boldsymbol{\beta}_0(q)\|_\infty = o_{\mathbb{P}}(1)$ is established for any m and d, so the requirement $m > d/2$ imposed in Lemmas 1 and 2 is not needed. Furthermore, in these two cases, it only requires the minimal side rate restrictions $1/(nh^d) \asymp K/n = o(1)$ in the sub-Gaussian case, and $1/(nh^{\frac{\nu}{\nu-1}d}) \asymp K^{\frac{\nu}{\nu-1}}/n = o(1)$ in the bounded ν-th moment case, up to polylog(n) terms. See Section D in the supplemental appendix for more details.

4.4 Comparison with Existing Results

These restrictions imposed in Lemmas 1 and 2, and the associated results in the supplemental appendix for special cases, are either comparable to or improve upon the existing literature. In the special case of square loss function and identity transformation, uniform (over $\boldsymbol{x} \in \mathcal{X}$) consistency of the partitioning-based estimator is essentially automatic due to the intrinsic closed-form and linearity of the estimator. Nevertheless, compared to the best known result in that strand of the literature [10, 4, 11], our general results are essentially on par in terms of side rate restrictions and regularity conditions. For example, in the sub-Gaussian case, and up to polylog(n) terms, the best side rate restriction in that literature requires $1/(nh^d) \asymp K/n = o(1)$, while Lemmas 1 and 2 require $1/(nh^{2d}) \asymp K^2/n = o(1)$, and our improved results in the supplemental appendix for unconnected bases require $1/(nh^d) \asymp K/n = o(1)$. Therefore, our results are on par with the best available results for series-based least squares regression [10, 4, 11], despite being able to cover a large class of M-estimation settings such as piecewise-polynomial-based quantile, nonlinear, or robust regression.

In the case of quantile regression with tensor-product B-splines, Corollary 1 of [3] implies that the L_2-consistency (4.3) can be obtained under $1/(nh^{2d}) \asymp K^2/n = o(1)$, and their Corollary 2 implies that the uniform consistency (4.2) can be obtained under $1/(nh^{4d}) \asymp K^4/n = o(1)$. In contrast, since the generalized residual from quantile regression has a sub-Gaussian envelope, we only require $1/(nh^{2d}) \asymp K^2/n = o(1)$ to establish both kinds of consistency. Moreover, when an unconnected basis is used, or when the loss function is strongly convex and smooth (e.g., the square loss for mean regression), we establish consistency under the weakest possible restriction: $1/(nh^d) \asymp K/n = o(1)$, up to polylog$(n)$ terms. It remains an open question whether it is possible to establish consistency under the weakest condition $1/(nh^d) \asymp K/n = o(1)$ for general partitioning-based M-estimators.

5 Bahadur Representation and Convergence Rates

The Bahadur representation is

$$\mathsf{L}^{(\boldsymbol{v})}(\boldsymbol{x}, q) = -\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{Q}_{0,q}^{-1} \mathbb{E}_n \big[\boldsymbol{p}(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i, q))\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, q)); q)\big]$$

where

$$\boldsymbol{Q}_{0,q} = \mathbb{E}\big[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}} \Psi_1(\boldsymbol{x}_i, \eta(\mu_0(\boldsymbol{x}_i, q)); q)[\eta^{(1)}(\mu_0(\boldsymbol{x}_i, q))]^2\big].$$

The following theorem takes the sup-norm consistency of the coefficient estimators $\widehat{\boldsymbol{\beta}}(q)$ as a high-level assumption, and thus avoids imposing any of the specific sufficient conditions discussed in Section 4. The proof is provided in the supplemental appendix (Theorem E.1).

Theorem 1 (Bahadur representation)**.** *Suppose that Assumptions 1–6 hold. Furthermore, assume the following four conditions:*

(i) $\sup_{q\in\mathcal{Q}} \big\|\widehat{\boldsymbol{\beta}}(q) - \boldsymbol{\beta}_0(q)\big\|_\infty = o_{\mathbb{P}}(1)$;

(ii) *there exists a fixed constant* $c > 0$ *such that* $\{\boldsymbol{b} \in \mathbb{R}^K : \|\boldsymbol{b} - \boldsymbol{\beta}_0(q)\|_\infty \leq c, q \in \mathcal{Q}\} \subseteq \mathcal{B}$;

(iii) $\frac{(\log n)^{d+2}}{nh^d} = o(1)$;

(iv) *either* $\frac{(\log n)^d}{n^{1-2/\nu}h^d} = o(1)$ *or* $\overline{\psi}(\boldsymbol{x}_i, y_i)$ *is sub-Gaussian conditional on* $\boldsymbol{x}_i$.

Then

$$\sup_{q\in\mathcal{Q}} \sup_{\boldsymbol{x}\in\mathcal{X}} \big|\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x}, q) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x}, q) - \mathsf{L}^{(\boldsymbol{v})}(\boldsymbol{x}, q)\big| \lesssim_{\mathbb{P}} h^{-|\boldsymbol{v}|}\Big(\frac{(\log n)^d}{nh^d}\Big)^{\frac{1}{2}+(\frac{\alpha}{2}\wedge\frac{1}{4})} \log n + h^{m-|\boldsymbol{v}|}. \tag{5.1}$$

If, in addition, $\sup_{y\in\mathcal{Y},q\in\mathcal{Q}} |\psi(y, \eta(\zeta_1); q) - \psi(y, \eta(\zeta_2); q)| \lesssim |\zeta_1 - \zeta_2|^\alpha$ *without any restrictions on* y *in Assumption 4(iii), then*

$$\sup_{q\in\mathcal{Q}} \sup_{\boldsymbol{x}\in\mathcal{X}} \big|\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x}, q) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x}, q) - \mathsf{L}^{(\boldsymbol{v})}(\boldsymbol{x}, q)\big| \lesssim_{\mathbb{P}} h^{-|\boldsymbol{v}|}\Big(\frac{(\log n)^d}{nh^d}\Big)^{\frac{1+\alpha}{2}} \log n + h^{m-|\boldsymbol{v}|}. \tag{5.2}$$

The Bahadur representation (5.1) applies to the case where the "derivative" $\psi(\cdot, \cdot; q)$ of the loss function may be discontinuous. One typical example is quantile regression (Example 1), where the "derivative" $\psi(y, \eta; q) = \mathbb{1}(y - \eta < 0) - q$, as a function of $(y - \eta)$, is piecewise constant with a jump at zero. In this case we can let $\alpha = 1$, and (5.1) implies that the order of the remainder in the Bahadur representation for partitioning-based quantile regression is $O(h^{-|\boldsymbol{v}|}(nh^d)^{-3/4} + h^{m-|\boldsymbol{v}|})$, up to polylog($n$) terms. Another example is L_p regression with $p \in (1, 2)$, Example 3, where the derivative of the loss function is $\psi(y, \eta) \equiv \psi(y - \eta) = p|y - \eta|^{p-1}\mathrm{sgn}(\eta - y)$ with $\mathrm{sgn}(\cdot)$ denoting the sign function. As a function of $(y - \eta)$, $\psi(\cdot)$ is α-Hölder on $[0, \infty)$ or $(-\infty, 0]$ for all $\alpha \in (0, p - 1]$ but not for $\alpha > p - 1$. Thus, (5.1) applies with the order of the remainder depending on p, which is the same as that for quantile regression when $p \geq 3/2$.

On the other hand, the Bahadur representation (5.2) applies to the case where the "derivative" of the loss is a continuous function of $(y-\eta)$. Nonlinear least squares regression (Example 2 and quasi-maximum likelihood estimation of generalized linear models (Example 4) fall into this category with the Hölder parameter $\alpha = 1$. In such cases, (5.2) implies that the order of the remainder in the Bahadur representation is $O(h^{-|\boldsymbol{v}|}(nh^d)^{-1} + h^{m-|\boldsymbol{v}|})$, up to polylog$(n)$ terms, which is a tighter upper bound than that implied by (5.1). See Section 8 and the supplemental appendix for more details.

In both cases, the remainder of the Bahadur representation consists of two terms. The last term $h^{m-|\boldsymbol{v}|}$ corresponds to the error from approximating the function μ_0 using the partitioning basis (cf. Assumption 6), whereas the first term arises from the (potential) nonlinearity underlying the M-estimation, and reflects explicitly the role of non-smoothness of the loss function. When the "derivative" of the loss function has discontinuity points, the order of the remainder in (5.1) is greater than that in the continuous case (5.2); with a smaller Hölder parameter α, the order of the remainder in both cases could increase.

5.1 Rates of Convergence

The uniform Bahadur representations (Theorem 1) can be used to establish convergence rates for the general partitioning-based M-estimators. We focus first on uniform convergence over $\boldsymbol{x} \in \mathcal{X}$ and $q \in \mathcal{Q}$.

Corollary 1 (Uniform Rate of Convergence)**.** *Suppose that Assumptions 1–6 and the four conditions (i)–(iv) in Theorem 1 hold. Furthermore, assume one of the following two conditions holds:*

(i) $\frac{(\log n)^{d+\frac{d+1}{\alpha\wedge 0.5}}}{nh^d} = O(1)$, *and* $h^{(\alpha\wedge 0.5)m}(\log n)^{0.5d} = O(1)$, *or*

(ii) *the additional condition for* (5.2) *holds,* $\frac{(\log n)^{d+\frac{d+1}{\alpha}}}{nh^d} = O(1)$, *and* $h^{\alpha m}(\log n)^{0.5d} = O(1)$.

Then

$$\sup_{q\in\mathcal{Q}}\sup_{\boldsymbol{x}\in\mathcal{X}} \left|\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x}, q) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x}, q)\right| \lesssim_{\mathbb{P}} h^{-|\boldsymbol{v}|}\sqrt{\frac{\log n}{nh^d}} + h^{m-|\boldsymbol{v}|}. \tag{5.3}$$

By setting $h \asymp \left(\frac{\log n}{n}\right)^{\frac{1}{2m+d}}$, Corollary 1 implies that the partitioning-based M-estimator can achieve the uniform convergence rate $\left(\frac{\log n}{n}\right)^{\frac{m}{2m+d}}$. This matches the optimal rate of convergence in sup-norm for nonparametric estimators of the conditional mean [40] and conditional quantiles [14]. In this sense, the rate of convergence in Corollary 1 is optimal and cannot be further improved at our level of generality.

Theorem 1 can also be used to obtain the mean square convergence rate of the partitioning-based M-estimator uniformly-in-q.

Corollary 2 (Mean Square Rate of Convergence)**.** *Suppose that Assumptions 1–6 and the four conditions (i)–(iv) in Theorem 1 hold. Furthermore, assume one of the following two conditions holds:*

(i) $\frac{(\log n)^{d+\frac{d+2}{\alpha\wedge 0.5}}}{nh^d} = o(1)$ *and* $h^{(\alpha\wedge 0.5)m}(\log n)^{\frac{d+1}{2}} = o(1)$, *or*

(ii) *the additional condition for* (5.2) *holds,* $\frac{(\log n)^{d+\frac{d+2}{\alpha}}}{nh^d} = o(1)$, *and* $h^{\alpha m}(\log n)^{\frac{d+1}{2}} = o(1)$.

Then

$$\sup_{q\in\mathcal{Q}}\int_{\mathcal{X}} \left(\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x}, q) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x}, q)\right)^2 f_X(\boldsymbol{x})d\boldsymbol{x} \lesssim_{\mathbb{P}} \frac{1}{nh^{d+2|\boldsymbol{v}|}} + h^{2(m-|\boldsymbol{v}|)}. \tag{5.4}$$

By setting $h \asymp n^{-\frac{1}{2m+d}}$, Corollary 2 implies that the partitioning-based M-estimator can also achieve the L_2 convergence rate $n^{-\frac{m}{2m+d}}$, uniformly over $\mathcal{Q}$, thereby matching the optimal rate of convergence in L_2-norm for nonparametric estimators of conditional means [39] and conditional quantiles [14].

The convergence rates in (5.3) and (5.4) capture two contributions: the first term reflects the variance of the estimator, while the second term arises from the error of approximating the unknown μ_0 by the partitioning basis. In the case of Corollary 2, it is possible to further leverage Theorem 1 to obtain a precise first-order asymptotic approximation for the integrated mean square error of the partitioning-based M-estimator, uniformly over $\mathcal{Q}$, which in turn could be used to develop plug-in asymptotically optimal rules for selecting $K \asymp h^{-d}$. See, for example, Theorem 4.2 in [11] for a similar result in the special case of square loss and identity transformation functions. We do not pursue this result here for brevity.

5.2 Comparison with Existing Results

To our knowledge, this paper is the first to establish uniformly valid Bahadur representations for partitioning-based M-estimators at the level of generality allowed in Theorem 1, and the implied convergence rates in Corollaries 1 and 2. The restriction on the tuning parameter h required by the theorem is seemingly minimal: when the envelope of the generalized residual $\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, q)); q)$ is sub-Gaussian (or its ν-th moment is bounded with a large ν), we roughly only need $1/(nh^d) \asymp K/n = o(1)$, up to polylog$(n)$ terms. Having noted this, verification of the high-level consistency assumption $\|\widehat{\boldsymbol{\beta}}(q) - \boldsymbol{\beta}_0(q)\|_\infty = o_{\mathbb{P}}(1)$ in the sub-Gaussian case may require a more stringent condition on h, as discussed in Section 4. In the best scenario (e.g., an unconnected basis is used), the minimal restriction $1/(nh^d) \asymp K/n = o(1)$ suffices, while in the worst scenario we need at most $1/(nh^{2d}) \asymp K^2/n = o(1)$, up to polylog$(n)$ terms.

The rest of this section discusses precisely how our results improve on prior literature.

Mean Regression

The usual mean regression is a special case of our general setup where $\rho(\cdot, \cdot)$ is the square loss, $\eta(\cdot)$ is the identity link, and $\mathcal{Q}$ is a singleton. Bahadur representations for this special case were established by [4] and [11]. Since the derivative of the square loss for mean regression is linear, the first term in (5.1) or (5.2) does not show up in the uniform linearization of least squares series estimators. See, for example, Lemma SA-4.2 of [11]; $R_{1n,q}$ defined therein has been implicitly included in the leading variance term in (5.2) above. Theorem 1 substantially extends these prior results to other nonlinear settings, under minimal additional conditions.

Finally, Corollaries 1 and 2 demonstrate the convergence rate optimality of general partitioning-based series M-estimation, recovering in particular known results for mean regression [4, 11] under essentially the same minimal conditions.

Quantile Regression

Theorem 1 improves upon prior theoretical results for nonparametric series quantile regression estimators. The most recent advance in this literature is due to [3], which establishes a uniform linear approximation for general series-based quantile regression estimators. In comparison, we exploit the "local support" feature of the partitioning basis, and make improvements in (at least) four aspects. To summarize these improvements without additional cumbersome notation, we set $\boldsymbol{v} = \mathbf{0}$ and ignore the smoothing bias h^m in the Bahadur approximation remainders.

First, [3] shows that the order of the remainder in the Bahadur representation is $O((nh^d)^{-3/4}h^{-d/2})$, up to polylog(n) terms (see proofs of Theorem 2 and Corollary 2 therein for details). In contrast, Theorem 1 implies that the remainder in the Bahadur representation for partitioning-based quantile regression estimators is $O((nh^d)^{-3/4})$, up to polylog(n) terms, which is not only a much tighter bound but also matches the optimal parametric bound when taking nh^d as the effective sample size.

Second, the rate restriction $1/(nh^{4d}) \asymp K^4/n = o(1)$ is required for B-spline-based estimators in [3]. In contrast, the restriction on h in Theorem 1 depends on the tail behavior of the generalized residuals and becomes weaker as ν gets larger. In the best case (the residuals have a sub-Gaussian envelope) we only need the seemingly weakest restriction $1/(nh^d) \asymp K/n = o(1)$, up to polylog($n$) terms, along with the consistency condition for $\widehat{\boldsymbol{\beta}}(q)$. Recall that in the sub-Gaussian scenario we need at worst $1/(nh^{2d}) \asymp K^2/n = o(1)$, up to polylog($n$) terms, to satisfy the consistency requirement.

Third, the restriction $h^{m-d} = o(n^{-\varepsilon})$ for some $\varepsilon > 0$ in [3] implicitly requires the smoothness m of the conditional quantile function be greater than the dimensionality d of the covariates. In contrast, the proof of Theorem 1 does not need such a restriction, though a weaker condition $m > d/2$ might be needed to verify the consistency condition on $\widehat{\boldsymbol{\beta}}(q)$; see Lemmas 1 and 2. Furthermore, when an unconnected basis (e.g., piecewise polynomials) is used for approximation, the condition $m > d/2$ is unnecessary for consistency, and thus we have *no* constraint on the relation between smoothness m and dimensionality d; see Section 4.3.

Fourth, compared to [3], we allow for a possibly non-identity link. Introducing a link function may lead to non-convexity of the loss $\rho(y, \eta(\theta); q)$ with respect to θ, making the usual proof strategies for consistency and Bahadur representation under convexity inapplicable. For example, non-convex quantile regression is covered in Theorem 1 by virtue of our general consistency results in Lemma 2.

All of the aforementioned improvements are practically relevant. For example, they accommodate univariate quantile regression using the piecewise constant basis with the IMSE-optimal choice of the mesh size h (in this case $h \asymp n^{-1/3}$ and $m = d$), which was theoretical excluded in prior literature.

Finally, Corollaries 1 and 2 establish the optimal rate of convergence for general partitioning-based series M-estimators, which substantially improve on prior work on quantile series regression in particular. More specifically, the conditions on the mesh size h, the smoothness m, and the dimensionality d in both corollaries are weaker than in prior work. In the best case (e.g., an unconnected basis is used and a sub-Gaussian envelope for residuals exists), we only require the seemingly minimal restriction $1/(nh^d) \asymp K/n = o(1)$, up to polylog($n$) terms, and an arbitrary relation between m and d is permitted. For comparison, in the special case of quantile regression, [3] shows that series estimators can achieve the fastest possible uniform-in-q rate in both L_2-norm and sup-norm (see Comments 3 and 4 therein), but under more stringent conditions: η is the identity function, $m > d$, and $1/(nh^{4d}) \asymp K^4/n = o(n^{-\varepsilon})$ for some $\varepsilon > 0$ (see their Corollary 2). Such conditions exclude, e.g., the IMSE-optimal choice of h for the Haar basis when $d = 1$ (since $h \asymp n^{-1/3}$ and $m = d = 1$), or piecewise linear fit when $d = 2$ (since $m = d = 2$).

Other Nonparametric Smoothing Methods

[30] establishes a similar Bahadur representation for kernel-based M-estimators using weakly stationary time series data. They consider a special case of our setup in Assumption 4: their loss function class $\mathcal{Q}$ is a singleton, η is an identity function, and the "derivative" of the loss can be written as $\psi(y, \eta) \equiv \psi(y - \eta)$ and is assumed to be piecewise Lipschitz continuous. In a compa-

rable cross-sectional context with $\alpha = 1$ and $\boldsymbol{v} = \mathbf{0}$, the order of the remainder in our Bahadur representation (5.1) is $O((nh^d)^{-3/4})$, up to polylog(n) terms, and thus Theorem 1 matches their approximation error up to a minor difference in $\log n$ terms. Taking nh^d to be the effective sample size, the approximation rate can not be further improved at this level of generality, and hence Theorem 1 establishes that the partitioning-based series M-estimator in (1.2) can achieve the same best Bahadur approximation as local polynomial kernel methods, up to polylog(n) terms.

Furthermore, compared to [30] or other similar contributions in the literature, Theorem 1 exhibits (at least) two novel features. First, the Bahadur representations (5.1) and (5.2) hold uniformly not only over the evaluation point $x \in \mathcal{X}$, but also over the loss function index $q \in \mathcal{Q}$, which may be important, for example, to study simultaneous quantile regression where the *entire* conditional quantile process may be of interest. Second, Theorem 1 also covers the more general setup where the "derivative" function may exhibit different degrees of smoothness, reflected by discontinuity points and/or the Hölder parameter α, or admits a more complex structure so that $\psi(y, \eta; q)$ cannot be written as $\psi(y - \eta; q)$. Thus, we cover more examples such as distribution regression (Example 2) and L_p regression with $p \in (1, 2)$ (Example 3). Finally, [30] does not discuss convergence rates as we do in Corollaries 1 and 2.

In the context of nonparametric penalized smoothing spline methods, [38] also establishes a uniform Bahadur representation (and other results) that can be compared to Theorem 1. However, their paper imposes more stringent assumptions and hence cover a smaller class of settings: using our notation, they assume that (i) $\mathcal{Q}$ is a singleton so their uniformity is only over $\mathcal{X}$; (ii) $d = 1$ so they consider only scalar covariate $\boldsymbol{x}_i$; and (iii) $\rho(\cdot, \eta(\cdot))$ is smooth so they rule out many important examples such as quantile regression, and Tukey and Huber regression. Furthermore, their results do not take explicit advantage of specific moment and boundedness conditions, or the structure of the nonparametric estimator, and instead impose the generic side condition $nh^2 \to \infty$, which is comparable to our condition $K^2/n \to \infty$, up to polylog(n) terms. Most importantly, in the closest comparable case ($d = 1$, $\alpha = 1$, and $\boldsymbol{v} = \mathbf{0}$), and only focusing on the variance component for simplicity, the order of the remainder in their uniform Bahadur representation (a combination of Theorem 3.4 and Lemma 3.1 in [38]) is $O((nh)^{-1}h^{-(6m-1)/(4m)})$, while (5.2) in Theorem 1 gives the optimal result $O((nh)^{-1})$, thereby demonstrating a substantial improvement over their result. Finally, as for convergence rates, Proposition 3.3 in [38] and our Corollary 2 are essentially equivalent, both delivering optimal mean square convergence. They do not explicitly discuss uniform convergence rates as we do in Corollary 1.

6 Strong Approximation

The uniform Bahadur representations in Theorem 1 can also be leveraged to establish uniform distribution theory for $\widehat{\mu}^{(\boldsymbol{v})}$. The infeasible conditional variance of the estimator can be written as

$$\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x}, q) = \boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}} \bar{\boldsymbol{Q}}_q^{-1} \bar{\boldsymbol{\Sigma}}_q \bar{\boldsymbol{Q}}_q^{-1} \boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x}),$$

where

$$\bar{\boldsymbol{Q}}_q = \mathbb{E}_n\big[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}} \Psi_1(\boldsymbol{x}_i, \eta(\mu_0(\boldsymbol{x}_i, q)); q)[\eta^{(1)}(\mu_0(\boldsymbol{x}_i, q))]^2\big], \qquad \text{and}$$
$$\bar{\boldsymbol{\Sigma}}_q = \mathbb{E}_n\big[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}} \mathbb{E}[\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, q)); q)^2 | \boldsymbol{x}_i][\eta^{(1)}(\mu_0(\boldsymbol{x}_i, q))]^2\big]$$

Accordingly, we define a feasible variance estimator as

$$\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x}, q) = \boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}} \widehat{\boldsymbol{Q}}_q^{-1} \widehat{\boldsymbol{\Sigma}}_q \widehat{\boldsymbol{Q}}_q^{-1} \boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x}),$$

where $\widehat{\boldsymbol{Q}}_q$ and $\widehat{\boldsymbol{\Sigma}}_q$ are some estimators of $\bar{\boldsymbol{Q}}_q$ and $\bar{\boldsymbol{\Sigma}}_q$, respectively, which are consistent in a sense described below. Therefore, $\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x}, q)$ is an estimator of the infeasible conditional variance $\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x}, q)$.

Statistical inference on $\mu_0^{(\boldsymbol{v})}$ usually relies on the following t-statistic process:

$$T(\boldsymbol{x}, q) = \frac{\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x}, q) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x}, q)}{\sqrt{\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x}, q)/n}}, \qquad (\boldsymbol{x}, q) \in \mathcal{X} \times \mathcal{Q},$$

where we drop the dependence of $T(\cdot)$ on $\boldsymbol{v}$ for simplicity.

Employing Theorem 1, or more precise arguments under slightly weaker conditions, it is easy to show that $T(\boldsymbol{x}, q)$ converges in distribution to $\mathsf{N}(0, 1)$ for each $(\boldsymbol{x}, q) \in \mathcal{X} \times \mathcal{Q}$. However, the stochastic process $(T(\boldsymbol{x}, q) : (\boldsymbol{x}, q) \in \mathcal{X} \times \mathcal{Q})$ is generally not asymptotically tight and, therefore, does not converge weakly in $\ell^\infty(\mathcal{X} \times \mathcal{Q})$, where $\ell^\infty(\mathcal{X} \times \mathcal{Q})$ denotes the set of all (uniformly) bounded real functions on $\mathcal{X} \times \mathcal{Q}$ equipped with uniform norm [41]. Nevertheless, we can construct a Gaussian process, in a possibly enlarged probability space, that approximates the entire process $T(\cdot)$ sufficiently fast, which can then be used to approximate the finite sample distribution of the function M-estimator $\widehat{\mu}^{(\boldsymbol{v})}(\cdot)$.

More precisely, under some mild consistency conditions on $\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x}, q)$, our Theorem 1 guarantees that $\sup_{q \in \mathcal{Q}} \sup_{\boldsymbol{x} \in \mathcal{X}} \big|T(\boldsymbol{x}, q) - t(\boldsymbol{x}, q)\big| \to_{\mathbb{P}} 0$ sufficiently fast, where

$$t(\boldsymbol{x}, q) = -\frac{\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^\mathsf{T} \bar{\boldsymbol{Q}}_q^{-1}}{\sqrt{\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x}, q)}} \mathbb{G}_n\big[\boldsymbol{p}(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i, q))\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, q)); q)\big].$$

It follows that, conditional on $\{\boldsymbol{x}_i\}_{i=1}^n$, the randomness of $t(\boldsymbol{x}, q)$ comes exclusively from the K-dimensional vector $\mathbb{G}_n[\boldsymbol{p}(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i, q))\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, q)); q)]$. Thus, our proof strategy is to further "discretize" this vector with respect to $q \in \mathcal{Q}$, and then apply Yurinskii's coupling [42] to construct a (conditional) Gaussian process that is close to the original t-statistic process $T(\boldsymbol{x}, q)$ uniformly over both $\boldsymbol{x} \in \mathcal{X}$ and $q \in \mathcal{Q}$. Our construction leverages a conditional Strassen's theorem [16, Theorem B.2] to generalize prior coupling results [3, Lemma 36]. See Section F in the supplemental appendix for details.

Our strong approximation approach is formalized in the next theorem. We employ high-level conditions to ease the exposition, but those conditions can be verified using Corollaries 1 and 2, and Theorem 1, as well as using the more general results in the supplemental appendix. Let $r_{\mathtt{UC}}$, $r_{\mathtt{BR}}$, $r_{\mathtt{VC}}$, and $r_{\mathtt{SA}}$ be positive non-random sequences as $n \to \infty$. The proof is available in the supplemental appendix (Theorem F.4).

Theorem 2 (Strong approximation)**.** *Suppose that Assumptions 1–6 with $\nu \geq 3$ hold. Furthermore, assume the following four conditions hold:*

(i) $\sup_{q \in \mathcal{Q}} \sup_{\boldsymbol{x} \in \mathcal{X}} |\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x}, q) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x}, q)| \lesssim_{\mathbb{P}} h^{-|\boldsymbol{v}|} r_{\mathtt{UC}}$.

(ii) $\sup_{q \in \mathcal{Q}} \sup_{\boldsymbol{x} \in \mathcal{X}} \big|\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x}, q) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x}, q) - \mathsf{L}^{(\boldsymbol{v})}(\boldsymbol{x}, q)\big| \lesssim_{\mathbb{P}} h^{-|\boldsymbol{v}|} r_{\mathtt{BR}}$*, with* $\frac{\log n}{nh^d} \lesssim r_{\mathtt{BR}}$.

(iii) $\sup_{q \in \mathcal{Q}} \sup_{\boldsymbol{x} \in \mathcal{X}} \big|\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x}, q) - \bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x}, q)\big| \lesssim_{\mathbb{P}} h^{-2|\boldsymbol{v}|-d} r_{\mathtt{VC}}$*, with* $r_{\mathtt{VC}} = o(1)$.

(iv) $\mathbb{E}\big[\big|\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, q)); q)\eta^{(1)}(\mu_0(\boldsymbol{x}_i, q)) - \psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, \tilde{q})); \tilde{q})\eta^{(1)}(\mu_0(\boldsymbol{x}_i, \tilde{q}))\big|^2 \big| \boldsymbol{x}_i\big] \lesssim |q - \tilde{q}|$*, for all* $q, \tilde{q} \in \mathcal{Q}$.

Then (provided the probability space is rich enough) there exists a stochastic process $(Z(\boldsymbol{x}, q) : (\boldsymbol{x}, q) \in \mathcal{X} \times \mathcal{Q})$ such that, conditional on $\boldsymbol{X}_n = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$, Z is a mean-zero Gaussian process with $\mathbb{E}[Z(\boldsymbol{x}, q)Z(\tilde{\boldsymbol{x}}, \tilde{q})|\boldsymbol{X}_n] = \mathbb{E}[t(\boldsymbol{x}, q)t(\tilde{\boldsymbol{x}}, \tilde{q})|\boldsymbol{X}_n]$ for all $(\boldsymbol{x}, q), (\tilde{\boldsymbol{x}}, \tilde{q}) \in \mathcal{X} \times \mathcal{Q}$, and

$$\sup_{q\in\mathcal{Q}} \sup_{\boldsymbol{x}\in\mathcal{X}} \left|T(\boldsymbol{x}, q) - Z(\boldsymbol{x}, q)\right| \lesssim_{\mathbb{P}} r_{\mathtt{SA}} + \sqrt{nh^d}(r_{\mathtt{UC}}\, r_{\mathtt{VC}} + r_{\mathtt{BR}}).$$

where $r_{\mathtt{SA}} = o(1)$ is any positive sequence satisfying

$$\Big(\frac{1}{nh^{3d}}\Big)^{\frac{1}{10}} \sqrt{\log n} + \frac{\log n}{\sqrt{n^{1-2/\nu}h^d}} = o(r_{\mathtt{SA}}).$$

Furthermore, if $\overline{\psi}(\boldsymbol{x}_i, y_i)$ is sub-Gaussian conditional on $\boldsymbol{x}_i$, then the same result holds with any positive sequence $r_{\mathtt{SA}} = o(1)$ satisfying

$$\Big(\frac{1}{nh^{3d}}\Big)^{\frac{1}{10}} \sqrt{\log n} + \frac{(\log n)^{3/2}}{\sqrt{nh^d}} = o(r_{\mathtt{SA}}).$$

The speed of strong approximation in Theorem 2 is determined by four factors: the uniform convergence rate $r_{\mathtt{UC}}$, the order of the remainder in the Bahadur representation $r_{\mathtt{BR}}$, the convergence rate $r_{\mathtt{VC}}$ of the variance estimator $\widehat{\Omega}_{\boldsymbol{v}}$, and the strong approximation rate $r_{\mathtt{SA}}$. Therefore, our strong approximation results are established at a high level of generality, building on our prior theoretical results: Corollary 1 for $r_{\mathtt{UC}}$, and Theorem 1 for $r_{\mathtt{BR}}$, while $r_{\mathtt{VC}}$ is a high-level condition that needs to be verified on a case-by-case basis. See Sections 7 and 8 for more discussion.

With respect to the strong approximation rate, Theorem 2 lays down two versions of lower bounds on $r_{\mathtt{SA}}$, depending on the tail behavior of the generalized residuals. Such restrictions may not be optimal, but are still weak enough to cover almost all partition size choices commonly used in practice. In particular, the restriction on $r_{\mathtt{SA}}$ in Theorem 2 allows for the MSE-optimal choice $h \asymp n^{-\frac{1}{2m+d}}$ in all cases except the unidimensional Haar basis approximation ($m = d = 1$); there is also room for undersmoothing in order to make the smoothing bias negligible in all cases but $m = d = 1$. The strong approximation for one dimensional partitioning-based series estimators in the special case of square loss and identity transformation functions was studied in [11, 9] via a different coupling strategy, which delivered tighter approximation results allowing for an MSE-optimal choice of h. We conjecture those techniques could be adapted to cover the case $m = d = 1$ for general partitioning-based M-estimator in (1.2), but we do not pursue this line of research here because it would require a different theoretical treatment.

6.1 Comparison with Existing Results

Theorem 2 is the first to establish strong approximation results for general partitioning-based M-estimators at the level of generality considered in this paper. In the prior literature, similar results are usually available only in specific scenarios such as least squares regression or quantile regression. To be more precise, in the least squares context ($\mathcal{Q}$ is a singleton), [11] establishes uniform inference theory for univariate regression ($d = 1$) and multivariate regression ($d > 1$) separately via different strong approximation methods. In particular, when $d > 1$, the same Yurinskii coupling technique is employed to obtain strong approximation for t-statistic processes, leading to similar rate restrictions on h. Theorem 2 is a substantial generalization of results therein, not just covering other loss functions, but also providing distributional approximation uniformly over the loss function index $q \in \mathcal{Q}$.

In the quantile regression context, [3] provides two strong approximations for general series-based estimators. When B-splines are used, their first strategy relies on a pivotal coupling, imposing $1/(nh^{10d}) = o(n^{-\varepsilon})$ and $h^{m-d} = o(n^{-\varepsilon})$ for some constant $\varepsilon > 0$ (see Theorem 11 therein), while the second strategy uses a Gaussian coupling (as in this paper), imposing $1/(nh^{4d \vee (2+3d)}) = o(n^{-\varepsilon})$ and $h^{m-d} = o(n^{-\varepsilon})$ (see Theorem 12 and Comment 13 therein). In comparison, our Theorem 2 requires weaker conditions on the tuning parameter h and the relation between the smoothness m and the dimensionality d. Specifically, we assume $1/(nh^{3d}) = o(1)$ up to $\text{polylog}(n)$ terms for a valid approximation. This improvement is practically relevant: for example, it allows for Gaussian approximation of linear-spline-based univariate quantile regression estimators with the MSE-optimal mesh size $h \asymp n^{-1/5}$. In addition, our general strategy to verify the consistency condition on $\widehat{\boldsymbol{\beta}}(q)$ in Theorem 1 only requires $m > d/2$ (not required for the special case of unconnected basis), which is weaker than $m > d$ as implicitly assumed in [3]. In practice, this improvement can accommodate, for example, the use of cubic splines for trivariate quantile regression.

Finally, [38] establishes uniform inference results for nonparametric penalized smoothing spline M-estimators. As mentioned before, their work is more specialized because they assume that $d = 1$, $\mathcal{Q}$ is a singleton, and $\rho(\cdot, \eta(\cdot))$ is smooth. Furthermore, their approach to constructing valid confidence bands and related uniform inference methods relies on approximating the suprema of the stochastic process directly via extreme value theory [38, Theorem 5.1], which leads to substantially slower approximation rates and requires stronger assumption and side rate restrictions; see [4] and [11] for more discussion in the context of nonparametric least squares series estimation. In contrast, Theorem 2, and our related uniform inference methods, provide a pre-asymptotic approximation with better finite sample properties, faster approximation rates, and weaker regularity conditions.

7 Feasible Plug-in Uniform Inference

The (conditional) Gaussian process $(Z(\boldsymbol{x}, q) : (\boldsymbol{x}, q) \in \mathcal{X} \times \mathcal{Q})$ in Theorem 2 is still infeasible since its covariance structure contains unknowns. This section establishes the validity of a generic *plug-in method* to construct a feasible version of $Z(\boldsymbol{x}, q)$. We later employ this result in Section 8 to develop feasible uniform inference in the context of our motivating examples.

The core idea behind the plug-in method is to estimate the covariance structure of $Z(\boldsymbol{x}, q)$ and then simulate its feasible version $\widehat{Z}(\boldsymbol{x}, q)$, a Gaussian process conditional on the data. If the covariance estimate converges to the true covariance sufficiently fast, $\widehat{Z}(\boldsymbol{x}, q)$ will be "close" to a copy of $Z(\boldsymbol{x}, q)$. The covariance structure of the process $Z(\boldsymbol{x}, q)$ in Theorem 2 is

$$\mathbb{E}[Z(\boldsymbol{x}, q) Z(\tilde{\boldsymbol{x}}, \tilde{q}) | \boldsymbol{X}_n] = \frac{\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}} \bar{\boldsymbol{Q}}_q^{-1} \bar{\boldsymbol{\Sigma}}_{q,\tilde{q}} \bar{\boldsymbol{Q}}_{\tilde{q}}^{-1} \boldsymbol{p}^{(\boldsymbol{v})}(\tilde{\boldsymbol{x}})}{\sqrt{\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x}, q) \bar{\Omega}_{\boldsymbol{v}}(\tilde{\boldsymbol{x}}, \tilde{q})}},$$

for all $(\boldsymbol{x}, q), (\tilde{\boldsymbol{x}}, \tilde{q}) \in \mathcal{X} \times \mathcal{Q}$, where

$$\bar{\boldsymbol{\Sigma}}_{q,\tilde{q}} = \mathbb{E}_n\big[S_{q,\tilde{q}}(\boldsymbol{x}_i) \eta^{(1)}(\mu_0(\boldsymbol{x}_i, q)) \eta^{(1)}(\mu_0(\boldsymbol{x}_i, \tilde{q})) \boldsymbol{p}(\boldsymbol{x}_i) \boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\big]$$

with

$$S_{q,\tilde{q}}(\boldsymbol{x}) = \mathbb{E}\big[\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, q)); q) \psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, \tilde{q})); \tilde{q}) \big| \boldsymbol{x}_i = \boldsymbol{x}\big].$$

Given context-specific estimates $\widehat{\boldsymbol{Q}}_q$ and $\boldsymbol{x} \mapsto \widehat{S}_{q,\tilde{q}}(\boldsymbol{x})$, we can put

$$\widehat{\boldsymbol{\Sigma}}_{q,\tilde{q}} = \mathbb{E}_n\big[\widehat{S}_{q,\tilde{q}}(\boldsymbol{x}_i) \eta^{(1)}(\widehat{\mu}(\boldsymbol{x}_i, q)) \eta^{(1)}(\widehat{\mu}(\boldsymbol{x}_i, \tilde{q})) \boldsymbol{p}(\boldsymbol{x}_i) \boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\big],$$

and $\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x}, q) = \boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}} \widehat{\boldsymbol{Q}}_q^{-1} \widehat{\boldsymbol{\Sigma}}_{q,q} \widehat{\boldsymbol{Q}}_q^{-1} \boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})$ as above. Section 8 illustrates how the estimates $\widehat{\boldsymbol{Q}}_q$ and $\boldsymbol{x} \mapsto \widehat{S}_{q,\tilde{q}}(\boldsymbol{x})$ can be constructed in specific examples. Then, a feasible Gaussian approximation $\widehat{Z}(\boldsymbol{x}, q)$ can be constructed as a mean-zero Gaussian process conditional on the data $\boldsymbol{D}_n = ((y_1, \boldsymbol{x}_1), \cdots, (y_n, \boldsymbol{x}_n))$ with conditional covariance structure

$$\mathbb{E}[\widehat{Z}(\boldsymbol{x}, q)\widehat{Z}(\tilde{\boldsymbol{x}}, \tilde{q})|\boldsymbol{D}_n] = \frac{\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}} \widehat{\boldsymbol{Q}}_q^{-1} \widehat{\boldsymbol{\Sigma}}_{q,\tilde{q}} \widehat{\boldsymbol{Q}}_{\tilde{q}}^{-1} \boldsymbol{p}^{(\boldsymbol{v})}(\tilde{\boldsymbol{x}})}{\sqrt{\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x}, q)\widehat{\Omega}_{\boldsymbol{v}}(\tilde{\boldsymbol{x}}, \tilde{q})}}, \tag{7.1}$$

for all $(\boldsymbol{x}, q), (\tilde{\boldsymbol{x}}, \tilde{q}) \in \mathcal{X} \times \mathcal{Q}$.

The following theorem establishes the validity of the plug-in approach.

Theorem 3 (Feasible Plug-in Strong Approximation)**.** *Suppose the assumptions in Theorem 2 hold. Furthermore, assume the following three conditions hold:*

(i) $\|\widehat{\boldsymbol{Q}}_q - \bar{\boldsymbol{Q}}_q\|_\infty \lesssim_{\mathbb{P}} h^d r_{\mathtt{Q}}$ *and* $\|\widehat{\boldsymbol{Q}}_q^{-1}\|_\infty \lesssim_{\mathbb{P}} h^{-d}$, *with* $r_{\mathtt{Q}} = o(1)$.

(ii) $S_{q,\tilde{q}}(\boldsymbol{x})$ *is continuous for all* $q, \tilde{q}$, *and* $\sup_{\boldsymbol{x},q,q_1 \neq q_2} \frac{|S_{q,q_1}(\boldsymbol{x}) - S_{q,q_2}(\boldsymbol{x})|}{|q_1 - q_2|} \lesssim 1$.

(iii) $\sup_{q,\tilde{q},\boldsymbol{x}} |\widehat{S}_{q,\tilde{q}}(\boldsymbol{x}) - S_{q,\tilde{q}}(\boldsymbol{x})| \lesssim_{\mathbb{P}} r_{\mathtt{S}}$, *with* $r_{\mathtt{S}} = o(1)$.

Then (provided the probability space is rich enough) there exists a mean-zero Gaussian process, conditional on $\boldsymbol{D}_n$, $(\widehat{Z}(\boldsymbol{x}, q) : (\boldsymbol{x}, q) \in \mathcal{X} \times \mathcal{Q})$ *satisfying* (7.1) *and*

$$\sup_{q \in \mathcal{Q}} \sup_{\boldsymbol{x} \in \mathcal{X}} \left|\widehat{Z}(\boldsymbol{x}, q) - Z^\star(\boldsymbol{x}, q)\right| \lesssim_{\mathbb{P}} [(r_{\mathtt{UC}} + r_{\mathtt{S}})^{1/4} + r_{\mathtt{Q}} + r_{\mathtt{VC}}]\sqrt{\log n},$$

where $(Z^\star(\boldsymbol{x}, q) : (\boldsymbol{x}, q) \in \mathcal{X} \times \mathcal{Q})$ *is a process such that, conditional on* $\boldsymbol{X}_n$, $Z(\cdot)$ *and* $Z^\star(\cdot)$ *have the same (conditional) distribution, and* $Z^\star(\cdot)$ *is (conditionally) independent of* $(y_1, \ldots, y_n)$.

Once we have a feasible process $\widehat{Z}(\boldsymbol{x}, q)$ that is "close" to a copy of $Z(\boldsymbol{x}, q)$ uniformly over $\mathcal{X} \times \mathcal{Q}$, conditional on the data, then $\widehat{Z}(\boldsymbol{x}, q)$ can be used to conduct inference on the entire function $\mu_0(\boldsymbol{x}, q)$, and functionals thereof. For example, our strong approximation results can be converted to convergence of the Kolmogorov distance between the distributions of $\sup_{q \in \mathcal{Q}} \sup_{\boldsymbol{x} \in \mathcal{X}} |T(\boldsymbol{x}, q)|$ and its feasible Gaussian approximation $\sup_{q \in \mathcal{Q}} \sup_{\boldsymbol{x} \in \mathcal{X}} |\widehat{Z}(\boldsymbol{x}, q)|$. See Theorem G.11 in the supplemental appendix for the formal result.

Furthermore, Theorem G.7 in the supplemental appendix establishes the asymptotic validity of the uniform confidence band for $\mu_0^{(\boldsymbol{v})}$ given by

$$\mathsf{CB}_{1-\alpha}(\boldsymbol{x}, q) = \left[\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x}, q) \pm \mathfrak{c}_{1-\alpha}\sqrt{\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x}, q)/n}\right] \tag{7.2}$$

with $\mathfrak{c}_{1-\alpha}$ satisfying

$$\mathbb{P}\Big(\sup_{q \in \mathcal{Q}} \sup_{\boldsymbol{x} \in \mathcal{X}} |\widehat{Z}(\boldsymbol{x}, q)| \leq \mathfrak{c}_{1-\alpha}\Big|\boldsymbol{D}_n\Big) = 1 - \alpha + o_{\mathbb{P}}(1),$$

provided the smoothing (or misspecification) bias relative to the standard error of the estimator is small, which could be achieved by undersmoothing, bias correction [25], simply ignoring the bias [26], robust bias correction [7, 8], or the Lepskii's method [32, 6], among other possibilities. Thus, under regularity conditions, the confidence band (7.2) covers $\mu_0^{(\boldsymbol{v})}$ with probability approximately $1 - \alpha$ in large samples, that is,

$$\lim_{n \to \infty} \mathbb{P}\big(\mu_0^{(\boldsymbol{v})}(\boldsymbol{x}, q) \in \mathsf{CB}_{1-\alpha}(\boldsymbol{x}, q), \;\text{ for all } (\boldsymbol{x}, q) \in \mathcal{X} \times \mathcal{Q}\big) = 1 - \alpha.$$

8 Verification of Assumptions in Examples

We verify assumptions in our four motivating examples (Section 2). Assumptions 1 and 2 concern the partitioning-based methodology itself, Assumption 3 are already primitive conditions on the data generating process, and Assumption 6 can be verified for usual local bases (e.g., piecewise polynomials and splines) when we assume the functional parameter μ_0, such as the conditional quantile function in Example 1 or conditional distribution function in Example 2, is smooth enough (see Assumption 3(iv) for details). Thus, we focus attention on two major issues that remain: (i) how the high-level conditions imposed in Assumptions 4 and 5, and Condition (iv) in Theorem 2, can be verified under intuitive primitive assumptions; and (ii) how to implement uniform inference based on our theory in Section 7.

8.1 Example 1: Generalized Conditional Quantile Regression

This example considers generalized conditional quantile regression with a possibly non-identity link: $\rho(y, \eta; q) = (q - \mathbb{1}(y < \eta))(y - \eta)$, where $q \in \mathcal{Q}$ denotes the quantile position. Thus, let $\eta(\mu_0(\boldsymbol{x}, q))$ be the conditional q-quantile of Y given $\boldsymbol{X} = \boldsymbol{x}$; we verify in the supplemental appendix that such μ_0 solves (1.1). For this example, the following simple proposition, proven in the supplemental appendix (Proposition H.1), gives sufficient conditions to verify the general Assumptions 4 and 5, and Condition (iv) in Theorem 2.

Proposition 1 (Quantile Regression)**.** *Suppose Assumption 3 holds with $\mathcal{Q} = [\varepsilon_0, 1 - \varepsilon_0]$ for some $\varepsilon_0 \in (0, 0.5)$, the loss is $\rho(y, \eta; q) = (q - \mathbb{1}(y < \eta))(y - \eta)$, the first moment of Y is finite $\mathbb{E}[|Y|] < \infty$. Assume further that $\eta(\cdot) \colon \mathbb{R} \to \mathcal{E}$ is strictly monotonic and twice continuously differentiable with $\mathcal{E}$ an open connected subset of $\mathbb{R}$ containing the conditional q-quantile of $Y|\boldsymbol{X} = \boldsymbol{x}$, given by $\eta(\mu_0(\boldsymbol{x}, q))$ for all $(\boldsymbol{x}, q)$; $f_{Y|X}(\eta(\mu_0(\boldsymbol{x}, q))|\boldsymbol{x})$ is bounded away from zero uniformly over $q \in \mathcal{Q}$ and $\boldsymbol{x} \in \mathcal{X}$, and the derivative of $y \mapsto f_{Y|X}(y|\boldsymbol{x})$ is continuous and bounded in absolute value from above uniformly over $y \in \mathcal{Y}_{\boldsymbol{x}}$ and $\boldsymbol{x} \in \mathcal{X}$. Then Assumptions 4–5 and Condition (iv) in Theorem 2 hold.*

The additional conditions in this proposition are primitive and easy-to-interpret, only restricting the conditional density of Y given $\boldsymbol{X}$ to be bounded and smooth in a mild sense. Our assumptions are on par with or are weaker than those imposed in [3], despite the high level of generality of our theoretical results.

We can implement uniform inference following the plug-in method described in Section 7. In this context $S_{q,\tilde{q}}(\boldsymbol{x}) = q \wedge \tilde{q} - q\tilde{q}$ is known and constant in $\boldsymbol{x}$, so a natural plug-in estimator of $\bar{\boldsymbol{\Sigma}}_{q,\tilde{q}}$ is

$$\widehat{\boldsymbol{\Sigma}}_{q,\tilde{q}} = (q \wedge \tilde{q} - q\tilde{q})\mathbb{E}_n\big[\eta^{(1)}(\widehat{\mu}(\boldsymbol{x}_i, q))\eta^{(1)}(\widehat{\mu}(\boldsymbol{x}_i, \tilde{q}))\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\big].$$

On the other hand, the matrix $\bar{\boldsymbol{Q}}_q$

$$\bar{\boldsymbol{Q}}_q = \mathbb{E}_n\big[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}} f_{Y|X}(\eta(\mu_0(\boldsymbol{x}_i, q))|\boldsymbol{x}_i)[\eta^{(1)}(\mu_0(\boldsymbol{x}_i, q))]^2\big]$$

depends on the unknown conditional density $f_{Y|X}$, and a plug-in estimator is not immediately available. However, many estimation strategies have been proposed in the literature [29]. We do not recommend a particular choice, but rather any estimator satisfying the mild convergence rate requirement in Condition (iii) of Theorem 2 may be used.

8.2 Example 2: Generalized Conditional Distribution Regression

The loss function is $\rho(y, \eta; q) = (\mathbb{1}(y \le q) - \eta)^2$ with a possibly non-identity inverse link function $\eta(\cdot)$. The derivative function is $\psi(y, \eta; q) = -2(\mathbb{1}(y \le q) - \eta)$. The following proposition, proven in the supplemental appendix (Proposition H.3), verifies our high-level assumptions under mild regularity conditions on the conditional distribution function of Y given $\boldsymbol{X}$.

Proposition 2 (Distribution Regression)**.** *Let $\mathcal{Q} = [-A, A]$ for some $A > 0$. Suppose that Assumption 3 holds, the loss is $\rho(y, \eta; q) = (\mathbb{1}(y \le q) - \eta)^2$, $\eta(\cdot)\colon \mathbb{R} \to (0, 1)$ is strictly monotonic and twice continuously differentiable, $\boldsymbol{x} \mapsto F_{Y|X}(q|\boldsymbol{x})$ is a continuous function, and $F_{Y|X}(q|\boldsymbol{x}) = \eta(\mu_0(\boldsymbol{x}, q))$ lies in a compact subset of $(0, 1)$ for all $q \in \mathcal{Q}$ and $\boldsymbol{x} \in \mathcal{X}$ (this subset does not depend on q and $\boldsymbol{x}$). Then Assumptions 4–5 and Condition (iv) in Theorem 2 hold.*

The implementation of uniform inference follows the plug-in method described in Section 7. To construct the prerequisite estimators, in this case $S_{q,\tilde{q}}(\boldsymbol{x}_i) = 4F_{Y|X}(q \wedge \tilde{q}|\boldsymbol{x}_i)\big(1 - F_{Y|X}(q \vee \tilde{q}|\boldsymbol{x}_i)\big)$. Therefore, a simple plug-in estimator of $\bar{\boldsymbol{\Sigma}}_{q,\tilde{q}}$ is

$$\widehat{\boldsymbol{\Sigma}}_{q,\tilde{q}} = 4\mathbb{E}_n\big[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\eta(\widehat{\mu}(\boldsymbol{x}_i, q \wedge \tilde{q}))(1 - \eta(\widehat{\mu}(\boldsymbol{x}_i, q \vee \tilde{q})))\eta^{(1)}(\widehat{\mu}(\boldsymbol{x}_i, q))\eta^{(1)}(\widehat{\mu}(\boldsymbol{x}_i, \tilde{q}))\big].$$

In addition, a plug-in estimator of the matrix $\bar{\boldsymbol{Q}}_q$ is $\widehat{\boldsymbol{Q}}_q = 2\mathbb{E}_n[(\eta^{(1)}(\widehat{\mu}(\boldsymbol{x}_i, q)))^2\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}]$.

8.3 Example 3: Generalized L_p Regression

The loss function is $\rho(y, \eta) = |y - \eta|^p$, $p \in (1, 2]$ with a possibly non-identity link. The case $p = 1$ is equivalent to quantile (median) regression discussed previously. The derivative function is $\psi(y, \eta) \equiv \psi(y - \eta) = p|y - \eta|^{p-1}\operatorname{sgn}(\eta - y)$. In this example the family $\mathcal{Q}$ of the loss functions is a singleton, and hence the dependence on the index q can be dropped to simplify notation.

The following proposition, proven in the supplemental appendix (Proposition H.5), provides a set of simple regularity conditions that ensure our general theory can be applied to study generalized L_p regression estimation and inference.

Proposition 3 (L_p Regression)**.** *Suppose that Assumption 3 holds with the loss function $\rho(y, \eta) = |y - \eta|^p$, $p \in (1, 2]$, and $\eta(\cdot)\colon \mathbb{R} \to \mathcal{E}$ is strictly monotonic and twice continuously differentiable with $\mathcal{E}$ an open connected subset of $\mathbb{R}$. Denoting by a_l and a_r the left and right ends of $\mathcal{E}$ respectively (possibly $\pm\infty$), assume that $\int_{\mathbb{R}} \psi(y; a_l) f_{Y|X}(y|x) dy < 0$ if a_l is finite, and $\int_{\mathbb{R}} \psi(y; a_r) f_{Y|X}(y|x) dy > 0$ if a_r is finite. Also assume that $\mathbb{E}[|Y|^{\nu(p-1)}] < \infty$ for some $\nu > 2$, and that $\boldsymbol{x} \mapsto f_{Y|X}(y|\boldsymbol{x})$ is continuous for any $y \in \mathcal{Y}$. In addition, assume that $\eta \mapsto \int_{\mathbb{R}} |\eta - y|^{p-1}\operatorname{sgn}(\eta - y) f_{Y|X}(y|\boldsymbol{x})\,\mathrm{d}y$ is twice continuously differentiable with derivatives $\frac{\mathrm{d}^j}{\mathrm{d}\eta^j}\int_{\mathbb{R}} |\eta - y|^{p-1}\operatorname{sgn}(\eta - y) f_{Y|X}(y|\boldsymbol{x})\,\mathrm{d}y = \int_{\mathbb{R}} |\eta - y|^{p-1}\operatorname{sgn}(\eta - y)\frac{\partial^j}{\partial y^j} f_{Y|X}(y|\boldsymbol{x})\,\mathrm{d}y$ for $j \in \{1, 2\}$. Moreover, the function $\int_{\mathbb{R}} |\eta(\zeta) - y|^{p-1}\operatorname{sgn}(\eta(\zeta) - y)\frac{\partial}{\partial y} f_{Y|X}(y|\boldsymbol{x})\,\mathrm{d}y$ is bounded and bounded away from zero uniformly over $\boldsymbol{x} \in \mathcal{X}$ and $\zeta \in B(\boldsymbol{x})$ with $B(\boldsymbol{x}) = \{\zeta : |\zeta - \mu_0(\boldsymbol{x})| \le r\}$ for some $r > 0$, and the function $\int_{\mathbb{R}} |\eta(\zeta) - y|^{p-1}\operatorname{sgn}(\eta(\zeta) - y)\frac{\partial^2}{\partial y^2} f_{Y|X}(y|\boldsymbol{x})\,\mathrm{d}y$ is bounded in absolute value uniformly over $\boldsymbol{x} \in \mathcal{X}$ and $\zeta \in B(\boldsymbol{x})$. Then Assumptions 4–5 and Condition (iv) in Theorem 2 hold.*

For implementation, we follow the plug-in method in Section 7. Since $\mathcal{Q}$ is a singleton, dependence on q can be dropped. Direct plug-in choices for estimating the prerequisite matrices take the form

$$\widehat{\boldsymbol{Q}} = \mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\widehat{\Psi}_{1,i}[\eta^{(1)}(\widehat{\mu}(\boldsymbol{x}_i))]^2] \quad \text{and} \quad \widehat{\boldsymbol{\Sigma}} = \mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\psi(\widehat{\epsilon}_i)^2[\eta^{(1)}(\widehat{\mu}(\boldsymbol{x}_i))]^2],$$

where $\widehat{\epsilon}_i = y_i - \eta(\widehat{\mu}(\boldsymbol{x}_i))$ and $\widehat{\Psi}_{1,i}$ is some estimator of the function $\Psi_1(\boldsymbol{x}_i, \eta(\mu_0(\boldsymbol{x}_i)))$. In L_p regression with $p \in (1,2]$, $\Psi_1(\boldsymbol{x}, \eta) = p(p-1)\mathbb{E}[|Y-\eta|^{p-2}\mathrm{sgn}(\eta - Y)|\boldsymbol{X} = \boldsymbol{x}]$, and therefore a simple plug-in choice is $\widehat{\Psi}_{1,i} = p(p-1)|y_i - \eta(\widehat{\mu}(\boldsymbol{x}_i))|^{p-2}\mathrm{sgn}(\eta(\widehat{\mu}(\boldsymbol{x}_i)) - y_i)$. As an alternative, bootstrap-based inference could be used.

8.4 Example 4: Logistic Regression

For this final example, the loss function is $\rho(y,\eta) = -y\log\eta - (1-y)\log(1-\eta)$, the inverse link function is $\eta(\theta) = 1/(1+e^{-\theta})$, and the derivative function is $\psi(y,\eta) = -y/\eta + (1-y)/(1-\eta)$, and the loss function does not depend on $q \in \mathcal{Q}$. The following proposition, proven in the supplemental appendix (Proposition H.6), gives simple primitive conditions verifying the high-level assumptions for our general theoretical results.

Proposition 4 (Logit Estimation)**.** *Suppose that Assumption 3 holds with the loss function $\rho(y,\eta) = -y\log\eta - (1-y)\log(1-\eta)$ and the inverse link $\eta(\theta) = 1/(1+e^{-\theta})$; $\mathcal{Y} = \{0,1\}$; $\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x})$ is continuous and lies in the interval $(0,1)$ for all $\boldsymbol{x} \in \mathcal{X}$. Then Assumptions 4–5 and Condition (iv) in Theorem 2 hold.*

It is easy to construct a feasible Gaussian process $\widehat{Z}(\boldsymbol{x})$ conditional on the data $\boldsymbol{D}_n$ with covariance structure (7.1). Standard choices are

$$\widehat{\boldsymbol{Q}} = \mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\widehat{\eta}_i(1-\widehat{\eta}_i)] \qquad \text{and} \qquad \widehat{\boldsymbol{\Sigma}} = \mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\widehat{\epsilon}_i^2],$$

where $\widehat{\eta}_i = \eta(\widehat{\mu}(\boldsymbol{x}_i))$ and $\widehat{\epsilon}_i = y_i - \widehat{\eta}_i$. See Section 7 for more discussion.

9 Other Parameters of Interest

We focused on uniform estimation and inference for the unknown function μ_0 and derivatives thereof. However, the parameter of interest may be other linear or nonlinear transformations of μ_0. For example, in generalized linear models usually the goal is to estimate the function $\eta(\mu_0(\boldsymbol{x},q))$, or the marginal effect of a covariate on that function $\frac{\partial}{\partial x_k}\eta(\mu_0(\boldsymbol{x},q)) = \eta^{(1)}(\mu_0(\boldsymbol{x},q))\mu_0^{(\boldsymbol{e}_k)}(\boldsymbol{x},q)$. Furthermore, in treatment effect and causal inference settings [1, and references therein], interest often lies in differences of such estimands across two or more subgroups: for two treatment levels $j=1,2$, $\eta(\mu_2(\boldsymbol{x},q)) - \eta(\mu_1(\boldsymbol{x},q))$ can be interpreted as a mean, quantile, or other conditional (on $(\boldsymbol{x},q)$) treatment effect, where $\mu_j(\boldsymbol{x},q)$ is estimated using separately the subsample of, say, control $(j=1)$ and treated $(j=2)$ units. Our results can be applied to all these cases of practical interest with minimal additional effort.

We showcase the generality of our theory by briefly discussing uniform inference on the transformed function $\eta(\mu_0(\boldsymbol{x},q))$, its first derivative, and differences thereof across subgroups. Given the partitioning-based M-estimators $\widehat{\mu}(\boldsymbol{x},q)$ and $\widehat{\mu}_j(\boldsymbol{x},q)$, $j=1,2$, where $\widehat{\mu}_j$ is constructed using only data from the subsample j of the full sample, we can immediately plug in to form the desired estimators.

- *Level Estimator*: $\eta(\widehat{\mu}(\boldsymbol{x},q))$.
- *Marginal Effect Estimator*: $\eta^{(1)}(\widehat{\mu}(\boldsymbol{x},q))\widehat{\mu}^{(\boldsymbol{e}_k)}(\boldsymbol{x},q)$.
- *Conditional Treatment Effect Estimator*: $\eta(\widehat{\mu}_1(\boldsymbol{x},q)) - \eta(\widehat{\mu}_2(\boldsymbol{x},q))$.

Uniform consistency of the three estimators follows from uniform consistency of $\widehat{\mu}(\boldsymbol{x},q)$ (Corollary 1) because the transformation function η is twice continuously differentiable. A Bahadur representation for each of the transformation estimators can be established via Theorem 1 and a Taylor expansion. For example, for the level estimator,

$$\sup_{q\in\mathcal{Q}}\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\eta(\widehat{\mu}(\boldsymbol{x},q))-\eta(\mu_0(\boldsymbol{x},q))-\mathsf{L}_{\mathtt{LE}}(\boldsymbol{x},q)\right|\lesssim_{\mathbb{P}} r_{\mathtt{LE}}$$

with

$$\mathsf{L}_{\mathtt{LE}}(\boldsymbol{x},q)=-\eta^{(1)}(\mu_0(\boldsymbol{x},q))\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{Q}_{0,q}^{-1}\mathbb{E}_n\big[\boldsymbol{p}(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i,q))\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,q));q)\big],$$

and for the marginal effect of the kth covariate,

$$\sup_{q\in\mathcal{Q}}\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\eta^{(1)}(\widehat{\mu}(\boldsymbol{x},q))\widehat{\mu}^{(\boldsymbol{e}_k)}(\boldsymbol{x},q)-\eta^{(1)}(\mu_0(\boldsymbol{x},q))\mu_0^{(\boldsymbol{e}_k)}(\boldsymbol{x},q)-\mathsf{L}_{\mathtt{ME}}(\boldsymbol{x},q)\right|\lesssim_{\mathbb{P}} r_{\mathtt{ME}}$$

with

$$\mathsf{L}_{\mathtt{ME}}(\boldsymbol{x},q)=-\eta^{(1)}(\mu_0(\boldsymbol{x},q))\boldsymbol{p}^{(\boldsymbol{e}_k)}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{Q}_{0,q}^{-1}\mathbb{E}_n\big[\boldsymbol{p}(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i,q))\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,q));q)\big],$$

where the approximation remainders from the Taylor expansion, and their uniform rates $r_{\mathtt{LE}}$ and $r_{\mathtt{ME}}$, are precisely characterized in the supplemental appendix (Theorem I.1). The conditional treatment effect estimator is simply a difference of two level estimators, each employing a disjoint sub-sample, and therefore it follows directly that

$$\sup_{q\in\mathcal{Q}}\sup_{\boldsymbol{x}\in\mathcal{X}}\left|(\eta(\widehat{\mu}_2(\boldsymbol{x},q))-\eta(\widehat{\mu}_1(\boldsymbol{x},q)))-(\eta(\mu_2(\boldsymbol{x},q))-\eta(\mu_1(\boldsymbol{x},q)))-\mathsf{L}_{\mathtt{CTE}}(\boldsymbol{x},q)\right|\lesssim_{\mathbb{P}} r_{\mathtt{LE}}$$

with $\mathsf{L}_{\mathtt{CTE}}(\boldsymbol{x},q)=\mathsf{L}_{\mathtt{LE},2}(\boldsymbol{x},q)-\mathsf{L}_{\mathtt{LE},1}(\boldsymbol{x},q)$ with $\mathsf{L}_{\mathtt{LE},j}(\boldsymbol{x},q)$ denoting the Bahadur approximation $\mathsf{L}_{\mathtt{LE}}(\boldsymbol{x},q)$ but when only using the sub-sample j.

Given the uniform Bahadur representations for each of the transformation estimators, strong approximations of their corresponding t-statistic processes can be constructed as in Section 6. For example, conditional on $\boldsymbol{X}_n$, the stochastic process $(\mathsf{L}_{\mathtt{LE}}(\boldsymbol{x},q):(\boldsymbol{x},q)\in\mathcal{X}\times\mathcal{Q})$ has mean zero and variance $|\eta^{(1)}(\mu_0(\boldsymbol{x},q))|^2\bar{\Omega}_{\mathbf{0}}(\boldsymbol{x},q)/n$. Then, applying our strong approximation strategy, we can construct a conditional Gaussian process $Z_{\mathtt{LE}}(\boldsymbol{x},q)$ that approximates the t-statistic process of $\eta(\widehat{\mu}(\boldsymbol{x},q))$:

$$\sup_{q\in\mathcal{Q}}\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\frac{\eta(\widehat{\mu}(\boldsymbol{x},q))-\eta(\mu_0(\boldsymbol{x},q))}{|\eta^{(1)}(\mu_0(\boldsymbol{x},q))|\sqrt{\bar{\Omega}_{\mathbf{0}}(\boldsymbol{x},q)/n}}-Z_{\mathtt{LE}}(\boldsymbol{x},q)\right|\lesssim_{\mathbb{P}} r_{\mathtt{SALE}}$$

with strong approximation rate $r_{\mathtt{SALE}}$ as in Theorem 2. Similarly, we can also construct a conditional Gaussian process $Z_{\mathtt{ME}}(\boldsymbol{x},q)$ that approximates the t-statistic process of the marginal effect estimator $\frac{\partial}{\partial x_k}\eta(\widehat{\mu}(\boldsymbol{x},q))$:

$$\sup_{q\in\mathcal{Q}}\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\frac{\eta^{(1)}(\widehat{\mu}(\boldsymbol{x},q))\widehat{\mu}^{(\boldsymbol{e}_k)}(\boldsymbol{x},q)-\eta^{(1)}(\mu_0(\boldsymbol{x},q))\mu_0^{(\boldsymbol{e}_k)}(\boldsymbol{x},q)}{|\eta^{(1)}(\mu_0(\boldsymbol{x},q))|\sqrt{\bar{\Omega}_{\boldsymbol{e}_k}(\boldsymbol{x},q)/n}}-Z_{\mathtt{ME}}(\boldsymbol{x},q)\right|\lesssim_{\mathbb{P}} r_{\mathtt{SAME}}$$

with strong approximation rate $r_{\mathtt{SAME}}$ as in Theorem 2. These results are formalized in the supplemental appendix (Theorem I.1). An analogous result holds for the conditional treatment effect estimator.

Finally, for implementation we can construct feasible processes to approximate $Z_{\mathtt{LE}}(\boldsymbol{x},q)$ and $Z_{\mathtt{ME}}(\boldsymbol{x},q)$ via the plug-in method discussed in Section 7, and illustrated in Section 8, which then can be employed to approximate the distributions of the entire level process $(\eta(\widehat{\mu}(\boldsymbol{x},q)):(\boldsymbol{x},q)\in\mathcal{X}\times\mathcal{Q})$, marginal effect process $(\frac{\partial}{\partial x_k}\eta(\widehat{\mu}(\boldsymbol{x},q)):(\boldsymbol{x},q)\in\mathcal{X}\times\mathcal{Q})$, and conditional treatment effect process $(\eta(\widehat{\mu}_2(\boldsymbol{x},q))-\eta(\widehat{\mu}_1(\boldsymbol{x},q)):(\boldsymbol{x},q)\in\mathcal{X}\times\mathcal{Q})$.

10 Conclusion

This paper investigated the asymptotic properties of a large class of nonparametric partitioning-based M-estimators, allowing for different degrees of non-smoothness in the loss function and a possibly non-identity monotonic transformation function. Our main theoretical results include uniform consistency for convex and non-convex objective functions, uniform Bahadur representations with optimal remainder under appropriate conditions, uniform and mean square convergence rates achieving optimal approximation under appropriate conditions, uniform strong approximation methods under general conditions, and uniform inference methods via plug-in approximations. We illustrated our general theory with four examples, and demonstrated how our results improve on prior literature, in many cases requiring minimal side rate restrictions on tuning parameters and achieving rate-optimal approximation rates. The supplemental appendix collects further theoretical results and generalizations that may be of independent interest. In future work, we plan to investigate optimal tuning parameter selection, including random partitioning schemes, and the validity of bootstrap-based approximations.

Acknowledgments

We thank Richard Crump, Max Farrell, Will Underwood, and Rae Yu for helpful comments and discussions. We also thank the Co-Editor, Associate Editor, and two reviewers for their comments. Cattaneo gratefully acknowledges financial support from the National Science Foundation through grants DMS-2210561 and SES-2241575. Feng gratefully acknowledges financial support from the National Natural Science Foundation of China (NSFC) through grants 72203122.

References

[1] Abadie, A. and Cattaneo, M. D. (2018). "Econometric Methods for Program Evaluation," *Annual Review of Economics*, *10*, 465–503.

[2] Aliprantis, C. D. and Border, K. C. (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Berlin, Heidelberg: Springer, 3rd edition, DOI: http://dx.doi.org/10.1007/3-540-29587-9.

[3] Belloni, A., Chernozhukov, V., Chetverikov, D., and Fernandez-Val, I. (2019). "Conditional Quantile Processes based on Series or Many Regressors," *Journal of Econometrics*, *213*(1), 4–29.

[4] Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). "Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results," *Journal of Econometrics*, *186*(2), 345–366.

[5] Bhatia, R. (2013). *Matrix analysis*, *169*: Springer Science & Business Media.

[6] Birgé, L. (2001). "An alternative point of view on Lepski's method," *Lecture Notes – Monograph Series*, *36*, 113–133.

[7] Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2018). "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference," *Journal of the American Statistical Association*, *113*(522), 767–779.

[8] Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2022). "Coverage Error Optimal Confidence Intervals for Local Polynomial Regression," *Bernoulli*, *28*(4), 2998–3022.

[9] Cattaneo, M. D., Crump, R. K., Farrell, M. H., and Feng, Y. (2024a). "On Binscatter," *American Economic Review*, *114*(5), 1488–1514.

[10] Cattaneo, M. D. and Farrell, M. H. (2013). "Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators," *Journal of Econometrics*, *174*(2), 127–143.

[11] Cattaneo, M. D., Farrell, M. H., and Feng, Y. (2020). "Large Sample Properties of Partitioning-Based Series Estimators," *Annals of Statistics*, *48*(3), 1718–1741.

[12] Cattaneo, M. D., Feng, Y., and Underwood, W. G. (2024b). "Uniform Inference for Kernel Density Estimators with Dyadic Data," *Journal of the American Statistical Association*.

[13] Cattaneo, M. D., Masini, R. P., and Underwood, W. G. (2024c). "Yurinskii's Coupling for Martingales," *arXiv preprint arXiv:2210.00362*.

[14] Chaudhuri, P. (1991). "Global Nonparametric Estimation of Conditional Quantile Functions and Their Derivatives," *Journal of Multivariate Analysis*, *39*(2), 246–269.

[15] Chen, X. and Christensen, T. M. (2015). "Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions," *Journal of Econometrics*, *188*(2), 447–465.

[16] Chen, X. and Kato, K. (2020). "Jackknife multiplier bootstrap: finite sample approximations to the U-process supremum with applications," *Probability Theory and Related Fields*, *176*, 1097–1163.

[17] Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). "Anti-Concentration and Honest Adaptive Confidence Bands," *Annals of Statistics*, *42*(5), 1787–1818.

[18] Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013). "Inference on counterfactual distributions," *Econometrica*, *81*(6), 2205–2268.

[19] Chernozhukov, V., Imbens, G. W., and Newey, W. K. (2007). "Instrumental variable estimation of nonseparable models," *Journal of Econometrics*, *139*(1), 4–14.

[20] Devroye, L., Györfi, L., and Lugosi, G. (2013). *A Probabilistic Theory of Pattern Recognition*, *31*: Springer Science & Business Media.

[21] Donald, S. G. and Hsu, Y.-C. (2014). "Estimation and inference for distribution functions and quantile functions in treatment effect models," *Journal of Econometrics*, *178*, 383–397.

[22] Dudley, R. M. (2014). *Uniform central limit theorems*, *142*: Cambridge University Press, URL: https://www.cambridge.org/core/books/uniform-central-limit-theorems/08936B317287871C6F78A0735B1DD96D.

[23] Eggermont, P. P. B. and LaRiccia, V. N. (2009). *Maximum Penalized Likelihood Estimation: Regression*, New York, NY: Springer.

[24] Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*: Springer-Verlag.

[25] Hall, P. (1992). "Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density," *Annals of Statistics*, *20*(2), 675–694.

[26] Hall, P. and Kang, K.-H. (2001). "Bootstrapping nonparametric density estimators with empirically chosen bandwidths," *Annals of Statistics*, *29*(5), 1443–1468.

[27] Huang, J. Z. (2003). "Local Asymptotics for Polynomial Spline Regression," *Annals of Statistics*, *31*(5), 1600–1635.

[28] Kallenberg, O. (2021). *Foundations of Modern Probability*, Probability Theory and Stochastic Modelling: Springer Cham, 3rd edition, XII, 946.

[29] Koenker, R. (2005). *Quantile regression*, *38*: Cambridge university press.

[30] Kong, E., Linton, O., and Xia, Y. (2013). "Global Bahadur representation for nonparametric censored regression quantiles and its applications," *Econometric Theory*, *29*(5), 941–968.

[31] Lai, P. and Lee, S. M. S. (2005). "An Overview of Asymptotic Properties of L_p Regression Under General Classes of Error Distributions," *Journal of the American Statistical Association*, *100*(470), 446–458.

[32] Lepskii, O. V. (1992). "Asymptotically minimax adaptive estimation. I: Upper bounds. Optimally adaptive estimates," *Theory of Probability & its Applications*, *36*(4), 682–697.

[33] Lin, W. and Kulasekera, K. (2007). "Identifiability of single-index models and additive-index models," *Biometrika*, *94*(2), 496–501.

[34] Matzkin, R. L. (2007). "Nonparametric identification," *Handbook of econometrics*, *6*, 5307–5368.

[35] McCullagh, P. and Nelder, J. A. (2019). *Generalized Linear Models*: Routledge.

[36] Monrad, D. and Philipp, W. (1991). "Nearby variables with nearby conditional laws and a strong approximation theorem for Hilbert space valued martingales," *Probability Theory and Related Fields*, *88*(3), 381–404.

[37] Schumaker, L. L. (2007). *Spline Functions: Basic Theory*, Cambridge Mathematical Library, Cambridge: Cambridge University Press, 3rd edition, URL: `https://www.cambridge.org/core/books/spline-functions-basic-theory/843475201223F90091FFBDDCBF210BFB`, DOI: `http://dx.doi.org/10.1017/CBO9780511618994`.

[38] Shang, Z. and Cheng, G. (2013). "Local and global asymptotic inference in smoothing spline models," *Annals of Statistics*, *41*(5), 2608–2638.

[39] Stone, C. J. (1980). "Optimal Rates of Convergence for Nonparametric Estimators," *Annals of Statistics*, 1348–1360.

[40] Stone, C. J. (1982). "Optimal Global Rates of Convergence for Nonparametric Regression," *Annals of Statistics*, 1040–1053.

[41] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*, Springer Series in Statistics: Springer New York.

[42] Yurinskii, V. V. (1978). "On the error of the Gaussian approximation for convolutions," *Theory of Probability & its Applications*, *22*(2), 236–247.

[43] Zhang, H. and Singer, B. H. (2010). *Recursive Partitioning and Applications*: Springer.

[44] Zhou, S., Shen, X., and Wolfe, D. (1998). "Local Asymptotics for Regression Splines and Confidence Regions," *Annals of Statistics*, *26*(5), 1760–1782.

A Introduction to supplemental material

Let $\mathcal{Q} \subset \mathbb{R}^{d_{\mathcal{Q}}}$, $\mathcal{X} \subset \mathbb{R}^d$ be fixed compact sets, where $d_{\mathcal{Q}}$ and d are positive integers. (In the paper, $d_{\mathcal{Q}} = 1$ was set only for simplicity.) Suppose that $\big((y_i, \boldsymbol{x}_i)\big)_{1 \le i \le n}$ is a random sample, where $y_i \in \mathcal{Y} \subset \mathbb{R}$ is a scalar response variable, $\boldsymbol{x}_i$ is a d-dimensional covariate with values in $\mathcal{X}$. Let $\rho(\cdot, \cdot; \boldsymbol{q})$ be a loss function parametrized by $\boldsymbol{q} \in \mathcal{Q}$ (Borel-measurable in all three arguments), and let $\eta(\cdot)\colon \mathbb{R} \to \mathcal{E}$ be a strictly monotonic continuously differentiable transformation function. (More detailed assumptions on $\rho(\cdot, \cdot; \boldsymbol{q})$ and $\eta(\cdot)$ will be provided below.) We fix a function $\mu_0(\cdot, \cdot)\colon \mathcal{X} \times \mathcal{Q} \to \mathbb{R}$, Borel-measurable in both arguments and such that

$$\mu_0(\cdot, \boldsymbol{q}) \in \operatorname*{arg\,min}_{\mu} \mathbb{E}[\rho(y_1, \eta(\mu(\boldsymbol{x}_1)); \boldsymbol{q})], \tag{A.1}$$

where the argmin is over the space of Borel functions $\mathcal{X} \to \mathbb{R}$. In particular, we assume that the minimum is finite, and such a minimizer exists.

Our main goal is to conduct uniform (over $\mathcal{X} \times \mathcal{Q}$) estimation and inference for μ_0, and transformations thereof, employing the partitioning-based series M-estimator

$$\widehat{\mu}(\boldsymbol{x}, \boldsymbol{q}) = \boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}} \widehat{\boldsymbol{\beta}}(\boldsymbol{q}), \qquad \widehat{\boldsymbol{\beta}}(\boldsymbol{q}) \in \operatorname*{arg\,min}_{\boldsymbol{b} \in \mathcal{B}} \sum_{i=1}^{n} \rho(y_i, \eta(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}} \boldsymbol{b}); \boldsymbol{q}), \tag{A.2}$$

where $\mathcal{B} \subseteq \mathbb{R}^K$ is the feasible set of the optimization problem, and

$$\boldsymbol{x} \mapsto \boldsymbol{p}(\boldsymbol{x}) = \boldsymbol{p}(\boldsymbol{x}; \Delta, m) = \big(p_1(\boldsymbol{x}; \Delta, m), \dots, p_K(\boldsymbol{x}; \Delta, m)\big)^{\mathsf{T}}$$

is a dictionary of K locally supported basis functions of order m based on a quasi-uniform partition $\Delta = \{\delta_l : 1 \le l \le \kappa\}$ containing a collection of open disjoint polyhedra in $\mathcal{X}$ such that the closure of their union covers $\mathcal{X}$. The m parameter controls how well μ_0 can be approximated by linear combinations of the basis (Assumption B.6 below); the partition being quasi-uniform intuitively means that the largest size of a cell cannot get asymptotically bigger than the smallest one (Assumption B.1 below). We consider large sample approximations where d and m are fixed constants, and $\kappa \to \infty$ (and thus $K \to \infty$) as $n \to \infty$. Prior literature is discussed in the paper.

A.1 Notation

For any real function f depending on d variables $(t_1, \dots, t_d)$ and any vector $\boldsymbol{v}$ of nonnegative integers, denote

$$f^{(\boldsymbol{v})} := \partial^{\boldsymbol{v}} f := \frac{\partial^{|\boldsymbol{v}|}}{\partial t_1^{v_1} \dots \partial t_d^{v_d}} f.$$

the multi-indexed partial derivative of f, where $|\boldsymbol{v}| = \sum_{k=1}^{d} v_j$. A derivative of order zero is the function itself, so if $v_i = 0$, the ith partial differentiation is ignored. For functions that depend on $(\boldsymbol{x}, \boldsymbol{q})$, the multi-index derivative notation is taken with respect to the first argument $\boldsymbol{x}$, unless otherwise noted.

We will denote $N(\mathcal{F}, \rho, \varepsilon)$ the ε-covering number of a class $\mathcal{F}$ with respect to a semi-metric ρ defined on it.

For a function $f\colon S \to \mathbb{R}$ the set $\{(x, t) \in S \times \mathbb{R} : t < f(x)\}$ is called the *subgraph* of f. A class $\mathcal{F}$ of measurable functions from $\mathcal{S}$ to $\mathbb{R}$ is called a *VC-subgraph class* or *VC-class* if the collections of all subgraphs of functions in $\mathcal{F}$ is a VC-class of sets in $S \times \mathbb{R}$, which means that for some finite

m no set of size m is shattered by it. In this case, the smallest such m is called the VC-index of $\mathcal{F}$. See [41] for details.

For a measurable function $f\colon S \to \mathbb{R}$ on a measurable space $(S, \mathcal{S})$, a probability measure $\mathbb{Q}$ on this space and some $q \geq 1$, define the $(\mathbb{Q}, q)$-*norm* of f as $\|f\|_{\mathbb{Q},q}^q = \mathbb{E}_{X\sim\mathbb{Q}}[f(X)^q]$.

We will say that a class of measurable functions $\mathcal{F}$ from any set S to $\mathbb{R}$ has a measurable *envelope* F if $F\colon S \to \mathbb{R}$ is such a measurable function that $|f(s)| \leq F(s)$ for all $s \in S$ and all $f \in \mathcal{F}$. We will say that this class satisfies the *uniform entropy bound* with envelope F and real constants $A \geq e$ and $V \geq 1$ if

$$\sup_{\mathbb{Q}} N(\mathcal{F}, \|\cdot\|_{\mathbb{Q},2}, \varepsilon\|F\|_{\mathbb{Q},2}) \leq \left(\frac{A}{\varepsilon}\right)^V \tag{A.3}$$

for all $0 < \varepsilon \leq 1$, where the supremum is taken over all finite discrete probability measures $\mathbb{Q}$ with $\|F\|_{\mathbb{Q},2} > 0$, $\|\cdot\|_{\mathbb{Q},2}$ denotes the $(\mathbb{Q}, 2)$-norm.

We will say that an $\mathbb{R}$-valued random variable ξ is σ^2-*sub-Gaussian*, where $\sigma^2 > 0$, if

$$\mathbb{P}\{|\xi| \geq t\} \leq 2\exp\{-t^2/\sigma^2\} \quad \text{for all } t \geq 0. \tag{A.4}$$

We will denote by $\boldsymbol{D}_n$ the random vector of all the data $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$, $\mathbf{X}_n$ for $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\boldsymbol{y}_n$ for $\{y_i\}_{i=1}^n$.

For two random elements Z_1 and Z_2, the notation $Z_1 \perp\!\!\!\perp Z_2$ means they are independent, and the notation $Z_1 \underset{\xi}{\perp\!\!\!\perp} Z_2$ means they are independent given a third random element ξ.

The notation $Z_1 \stackrel{d}{=} Z_2$ means the two random elements Z_1 and Z_2 have the same laws.

If we say the probability space is *rich enough*, it means that it admits a randomization variable, that is, a random variable distributed uniformly on $[0, 1]$ independent of the data. Since this one random variable can be replicated (see, e. g., Lemma 4.21 in [28]), we can find a random variable distributed uniformly on $[0, 1]$ independent of the data and such random variables previously used, whenever the argument requires.

Finally, we will use the following notations:

$$\bar{\mathbf{Q}}_{\boldsymbol{q}} := \mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\Psi_1(\boldsymbol{x}_i, \eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q})); \boldsymbol{q})\eta^{(1)}(\mu_0(\boldsymbol{x}_i, \boldsymbol{q}))^2] \tag{A.5}$$

$$\mathbf{Q}_{0,\boldsymbol{q}} := \mathbb{E}[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\Psi_1(\boldsymbol{x}_i, \eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q})); \boldsymbol{q})\eta^{(1)}(\mu_0(\boldsymbol{x}_i, \boldsymbol{q}))^2] \tag{A.6}$$

$$\boldsymbol{\Sigma}_{0,\boldsymbol{q}} := \mathbb{E}[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q})); \boldsymbol{q})^2\eta^{(1)}(\mu_0(\boldsymbol{x}_i, \boldsymbol{q}))^2] \tag{A.7}$$

$$\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q}} := \mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\sigma_{\boldsymbol{q}}^2(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i, \boldsymbol{q}))^2] \tag{A.8}$$

$$\Omega_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q}) := \boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^\mathsf{T}\mathbf{Q}_{0,\boldsymbol{q}}^{-1}\boldsymbol{\Sigma}_{0,\boldsymbol{q}}\mathbf{Q}_{0,\boldsymbol{q}}^{-1}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x}) \tag{A.9}$$

$$\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q}) := \boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^\mathsf{T}\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q}}\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x}), \tag{A.10}$$

where $\psi(\cdot, \cdot; \cdot)$, $\Psi_1(\cdot, \cdot; \cdot)$, $\sigma_\cdot^2(\cdot)$ are defined in Assumption B.4.

Many mathematical notations are clickable and link to their definitions; for example, clicking on $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q}}$ should lead to (A.8).

A.2 Organization

Appendix B collects the general assumptions used through this supplemental appendix. In subsequent sections, we will list which assumptions are required for each lemma, proposition, or theorem. Note that Assumption B.4 is weaker than the assumptions on the loss function described in Section 3 of the main paper. In particular, a complexity condition on $\{\psi(\cdot, \cdot; \boldsymbol{q})\}$ is absent in Section 3 because the main paper contains a stronger but simpler Assumption 5. Thus, the setup in this supplement is more general that in the main paper; this fact is formally shown by Proposition B.9. The correspondence of assumptions given in the paper and the versions given in the supplement is as follows:

$$
\begin{aligned}
&\text{Assumption 1} \Leftrightarrow \text{Assumption B.1},\\
&\text{Assumption 2} \Leftrightarrow \text{Assumption B.2},\\
&\text{Assumption 3} \Leftrightarrow \text{Assumption B.3},\\
&\text{Assumption 4} \Leftrightarrow \text{Assumption B.4},\\
&\text{Assumption 5} \Rightarrow \text{Assumptions B.5 and B.8} \quad \text{(by Proposition B.9)},\\
&\text{Assumption 6} \Leftrightarrow \text{Assumption B.6}.
\end{aligned}
$$

Appendix C records some well-known results and tools in probability and stochastic process theory that will be useful for our theoretical analysis, and also presents a collections of lemmas that are used repeatedly in many subsequent arguments throughout this supplement.

The rest of the supplement presents our main theoretical results, which encompass and generalize the simplified results presented in the main paper (to improve exposition). In the rest of this subsection, we explain how the two set of theoretical results relate to each other. All equivalences below are given up to the fact that the supplement contains weaker complexity Assumptions B.5 and B.8, which are not present in the paper because the simplified setup (Assumption 5) is used.

Appendix D presents all our consistency results, categorized based on whether the objective function is convex or not. Additional results of theoretical and methodological interest, such as the consistency of partitioning-based M-estimators in special cases (e.g., an unconnected basis or a strongly convex and smooth loss function), which were not formally reported in the paper, are also presented. The matching of consistency results given in the paper with the versions given in the supplement is as follows:

$$
\begin{aligned}
&\text{Lemma 1} \Leftrightarrow \text{Lemma D.1},\\
&\text{Lemma 2} \Leftrightarrow \text{Lemma D.3}.
\end{aligned}
$$

Appendix E proves our main Bahadur representation result and its corollaries, matched as follows to the results in the paper:

$$
\begin{aligned}
&\text{Theorem 1} \Leftrightarrow \text{Theorem E.1},\\
&\text{Corollary 1} \Leftarrow \text{Corollary E.16},\\
&\text{Corollary 2} \Leftrightarrow \text{Corollary E.17}.
\end{aligned}
$$

Appendix F develops strong approximation results using a generalized conditional Yurinskii's coupling approach. First, Appendix F.1 gives general results that may be of independent theoretical interest: they provide generalizations of, and in some cases complement, prior coupling results established in [3], [12], [13], and references therein. Second, Appendix F.2 deploys those results to the setting of interest in our paper to verify our main result Theorem F.4 that confirms Theorem 2 in the paper:

$$\text{Theorem 2} \Leftarrow \text{Theorem F.4 (see also Remark F.5)}.$$

Appendix G discusses results related to the implementation of uniform inference, formally showing the validity of the plug-in approximation method and confidence bands described in the paper.

In particular, the following correspondence holds:

$$\text{Theorem 3} \Leftarrow \text{Theorem G.1 (see also Remark G.2)}.$$

Appendix H discusses in detail the verification of our high-level assumptions for each of the four motivating examples in the paper, proving Propositions 1 through 4 in the paper along with verifying more general complexity assumptions discussed above. The correspondence between the results in the paper and the results in this section is as follows:

$$\begin{aligned}
\text{Proposition 1} &\Leftarrow \text{Proposition H.1},\\
\text{Proposition 2} &\Leftarrow \text{Proposition H.3},\\
\text{Proposition 3} &\Leftarrow \text{Proposition H.5},\\
\text{Proposition 4} &\Leftarrow \text{Proposition H.6}.
\end{aligned}$$

Appendix I discusses other parameters of potential interest such as the level and marginal effect functions, which formalizes the claims made in Section 9 of the paper.

Appendix J is devoted to simulation runs with synthetic data, covering pointwise and uniform inference.

A.3 Comparison to prior literature

Our general M-estimation theoretical results are on par or improve over prior literature partitioning-based nonparametric least squares regression [4, 10, 11, 9, 15, 27, 44], and on series nonparametric quantile regression [3]. Both strands of the literature assume an identity transformation function $\eta(\cdot)$. In addition, we also compare to the closely related setting of nonparametric smoothing spline nonlinear regression [38].

The main paper gives a detailed discussion of our improvements when discussing the theoretical results for consistency (Section 4), Bahadur representation (Section 5), and strong approximation (Section 6). To complement that discussion, we present three tables that summarize the comparison between prior literature and this paper.

Table 1 compares the best known results for mean-square (L_2) and uniform (L_∞) consistency of series and partitioning-based estimators. In each case considered in this table, rate-optimal L_2 and L_∞ convergence rates are achievable, but with different side rate conditions on h, m, and d. The table shows that our results either achieve the best (minimal) restrictions in the literature for the special case of square loss and identity transformation functions, or improve upon prior results in the general M-estimation case.

Table 2 compares the best known results for uniform Bahadur representation for series and partitioning-based estimators. It again demonstrates that the results in our paper either achieve the best (minimal) restrictions in the literature for the special case of square loss and identity transformation functions, or improve upon prior literature in the general M-estimation case.

Finally, Table 3 compares the best known results for uniform strong approximation for series and partitioning-based estimators. Our results achieve the best (minimal) restrictions in the literature for the special case of square loss and identity transformation functions when $d > 1$, and improve upon prior literature in the general M-estimation case.

B Full assumptions

This section collects the assumptions used throughout the supplemental appendix. These assumptions are weaker than (i.e, implied by) the assumptions imposed in the paper.

Table 1: Comparison to prior literature: Consistency (Section 4)

	Side rate restriction	m v.s. d restriction	Uniformity
Mean regression	$nh^d \to \infty$	None	$\boldsymbol{x}$ only
Quantile regression, L_2-consistency	$nh^{2d} \to \infty$	$m > d$	$\boldsymbol{x}$, and q (scalar)
Quantile regression, L_∞-consistency	$nh^{4d} \to \infty$	$m > d$	$\boldsymbol{x}$, and q (scalar)
This paper, L_2- and L_∞-consistency			
General partitioning basis	$nh^{2d} \to \infty$	$m > d/2$	$\boldsymbol{x}$, and $\boldsymbol{q}$
Strongly convex and smooth loss function	$nh^d \to \infty$	None	$\boldsymbol{x}$, and $\boldsymbol{q}$
Unconnected partitioning basis	$nh^d \to \infty$	None	$\boldsymbol{x}$, and $\boldsymbol{q}$

Notes: (i) m is the smoothness of $\boldsymbol{x} \mapsto \mu_0(\boldsymbol{x}, \boldsymbol{q})$ and order of the basis; (ii) d is the dimension of $\boldsymbol{x}_i$; and (iii) poly-log-n and similar additional terms are not reported to simplify the exposition.

Table 2: Comparison to prior literature: Uniform Bahadur representation (Section 5)

	Side rate restriction	m v.s. d restriction	remainder	Uniformity
Mean regression	$nh^d \to \infty$	None	$\frac{1}{nh^d}$ (optimal)	$\boldsymbol{x}$ only
Quantile regression	$nh^{4d} \to \infty$	$m > d$	$\frac{1}{(nh^d)^{3/4}h^{d/2}}$	$\boldsymbol{x}$, and q (scalar)
Smooth loss + smoothing spline	$nh^{2d} \to \infty$	$d = 1$	$\frac{1}{nh^{10m-1}}$	x, and q (scalar)
This paper (assuming consistency)				
Smooth weak derivative ($\alpha = 1$)	$nh^d \to \infty$	None	$\frac{1}{(nh^d)^{3/4}}$ (optimal)	$\boldsymbol{x}$, and $\boldsymbol{q}$
Non-smooth week derivative ($\alpha \in (0,1)$)	$nh^d \to \infty$	None	See Theorem 1	$\boldsymbol{x}$, and $\boldsymbol{q}$

Notes: (i) m is the smoothness of $\boldsymbol{x} \mapsto \mu_0(\boldsymbol{x}, \boldsymbol{q})$ and order of the basis; (ii) d is the dimension of $\boldsymbol{x}_i$; and (iii) poly-log-n and similar additional terms are not reported to simplify the exposition.

Table 3: Comparison to prior literature: Strong approximation (Section 6)

	Side rate restriction	m v.s. d restriction	Uniformity
Mean regression			
$d = 1$	$nh^d \to \infty$	None	$\boldsymbol{x}$ only
$d \geq 1$	$nh^{3d} \to \infty$	None	$\boldsymbol{x}$ only
Quantile regression ($d \geq 1$)	$nh^{4d \vee (2+3d)} \to \infty$	$m > d$	$\boldsymbol{x}$, and q (scalar)
This paper (assuming consistency, $d \geq 1$)	$nh^{3d} \to \infty$	None	$\boldsymbol{x}$, and $\boldsymbol{q}$

Notes: (i) m is the smoothness of $\boldsymbol{x} \mapsto \mu_0(\boldsymbol{x}, \boldsymbol{q})$ and order of the basis; (ii) d is the dimension of $\boldsymbol{x}_i$; and (iii) poly-log-n and similar additional terms are not reported to simplify the exposition.

Assumption B.1 (Quasi-uniform partition)**.** *The ratio of the sizes of inscribed and circumscribed balls of each $\delta \in \Delta$ is bounded away from zero uniformly in $\delta \in \Delta$, and*

$$\frac{\max\{\mathrm{diam}(\delta) : \delta \in \Delta\}}{\min\{\mathrm{diam}(\delta) : \delta \in \Delta\}} \lesssim 1$$

where $\mathrm{diam}(\delta)$ *denotes the diameter of* δ*. Further, for* $h = \max\{\mathrm{diam}(\delta) : \delta \in \Delta\}$*, assume* $h = o(1)$

and $\log(1/h) \lesssim \log(n)$.

Assumption B.2 (Local basis)**.**

(i) For each basis function p_k, $k = 1, \dots, K$, the union of elements of Δ on which p_k is active is a connected set, denoted by $\mathcal{H}_k$. For all $k = 1, \dots, K$, both the number of elements of $\mathcal{H}_k$ and the number of basis functions which are active on $\mathcal{H}_k$ are bounded by a constant.

(ii) For any $\boldsymbol{a} = (a_1, \dots, a_K)^\mathsf{T} \in \mathbb{R}^K$

$$\boldsymbol{a}^\mathsf{T} \int_{\mathcal{H}_k} \boldsymbol{p}(\boldsymbol{x}; \Delta, m) \boldsymbol{p}(\boldsymbol{x}; \Delta, m)^\mathsf{T} d\boldsymbol{x}\ \boldsymbol{a} \gtrsim a_k^2 h^d, \quad k = 1, \dots, K.$$

(iii) Let $|\boldsymbol{v}| < m$. There exists an integer $\varsigma \in [|\boldsymbol{v}|, m)$ such that, for all $\boldsymbol{\varsigma}, |\boldsymbol{\varsigma}| \le \varsigma$,

$$h^{-|\boldsymbol{\varsigma}|} \lesssim \inf_{\delta \in \Delta} \inf_{\boldsymbol{x} \in \mathrm{cl}(\delta)} \|\boldsymbol{p}^{(\boldsymbol{\varsigma})}(\boldsymbol{x}; \Delta, m)\| \le \sup_{\delta \in \Delta} \sup_{\boldsymbol{x} \in \mathrm{cl}(\delta)} \|\boldsymbol{p}^{(\boldsymbol{\varsigma})}(\boldsymbol{x}; \Delta, m)\| \lesssim h^{-|\boldsymbol{\varsigma}|}$$

where $\mathrm{cl}(\delta)$ is the closure of δ.

Assumption B.2 implicitly relates the number of basis functions and the maximum mesh size: $K \asymp h^{-d} = J^d$.

For the following assumption and throughout the document, when speaking of the conditional distribution of y_1 given $\boldsymbol{x}_1$, or its functionals (like conditional moments or quantiles), we mean one fixed regular variant of such a distribution satisfying all the assumptions listed.

Assumption B.3 (Data generating process)**.**

(i) $\big((y_i, \boldsymbol{x}_i)\big)_{1 \le i \le n}$ is a random sample satisfying (A.1).

(ii) The distribution of $\boldsymbol{x}_i$ admits a Lebesgue density $f_X(\cdot)$ which is continuous and bounded away from zero on support $\mathcal{X} \subset \mathbb{R}^d$, where $\mathcal{X}$ is the closure of an open, connected and bounded set.

(iii) The conditional distribution of y_i given $\boldsymbol{x}_i$ admits a conditional density $f_{Y|X}(y|\boldsymbol{x})$ with support $\mathcal{Y}_{\boldsymbol{x}}$ with respect to some (sigma-finite) measure $\mathfrak{M}$, and $\sup_{\boldsymbol{x} \in \mathcal{X}} \sup_{y \in \mathcal{Y}_{\boldsymbol{x}}} f_{Y|X}(y|\boldsymbol{x}) < \infty$.

(iv) $\boldsymbol{x} \mapsto \mu_0(\boldsymbol{x}, \boldsymbol{q})$ is $m \ge 1$ times continuously differentiable for every $\boldsymbol{q} \in \mathcal{Q}$, $\boldsymbol{x} \mapsto \mu_0(\boldsymbol{x}, \boldsymbol{q})$ and its derivatives of order no greater than m are bounded uniformly over $(\boldsymbol{x}, \boldsymbol{q}) \in \mathcal{X} \times \mathcal{Q}$, and

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \sup_{\boldsymbol{q}_1 \ne \boldsymbol{q}_2} \frac{|\mu_0(\boldsymbol{x}, \boldsymbol{q}_1) - \mu_0(\boldsymbol{x}, \boldsymbol{q}_2)|}{\|\boldsymbol{q}_1 - \boldsymbol{q}_2\|} \lesssim 1.$$

For the following assumption and throughout the document, we fix some (small enough) $r > 0$ and denote

$$B_{\boldsymbol{q}}(\boldsymbol{x}) := \{\zeta \colon |\zeta - \mu_0(\boldsymbol{x}, \boldsymbol{q})| \le r\}, \tag{B.1}$$

i. e. we will work with a fixed neighborhood of $\mu_0(\boldsymbol{x}, \boldsymbol{q})$.

Assumption B.4 (Loss function)**.**

(i) Let $\mathcal{Q} \subset \mathbb{R}^{d_\mathcal{Q}}$ be a connected compact set. For each $\boldsymbol{q} \in \mathcal{Q}$ and $y \in \mathcal{Y}$, and some open connected subset $\mathcal{E}$ of $\mathbb{R}$ not depending on y, $\eta \mapsto \rho(y, \eta; \boldsymbol{q})$ is absolutely continuous on closed bounded intervals within $\mathcal{E}$, and admits an a. e. derivative $\psi(y, \eta; \boldsymbol{q})$, Borel measurable as a function of $(y, \eta, \boldsymbol{q})$.

(ii) The first-order optimality condition $\mathbb{E}[\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q})); \boldsymbol{q}) \,|\, \boldsymbol{x}_i] = 0$ holds; the function

$$\sigma^2_{\boldsymbol{q}}(\boldsymbol{x}) := \mathbb{E}\big[\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q})); \boldsymbol{q})^2 \,\big|\, \boldsymbol{x}_i = \boldsymbol{x}\big]$$

is continuous in both arguments $(\boldsymbol{x}, \boldsymbol{q}) \in \mathcal{X} \times \mathcal{Q}$, bounded away from zero, and Lipschitz in $\boldsymbol{q}$ uniformly in $\boldsymbol{x}$; there is a positive measurable envelope function $\overline{\psi}(\boldsymbol{x}_i, y_i)$ such that $\sup_{\boldsymbol{q} \in \mathcal{Q}} |\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q})); \boldsymbol{q})| \le \overline{\psi}(\boldsymbol{x}_i, y_i)$ with

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \mathbb{E}\big[\overline{\psi}(\boldsymbol{x}_i, y_i)^\nu \,\big|\, \boldsymbol{x}_i = \boldsymbol{x}\big] < \infty \quad \textit{for some } \nu > 2.$$

(iii) The function $\eta(\cdot)\colon \mathbb{R} \to \mathcal{E}$ is strictly monotonic and twice continuously differentiable; for some fixed constant $\alpha \in (0,1]$, for any $(\boldsymbol{x},\boldsymbol{q}) \in \mathcal{X} \times \mathcal{Q}$, and a pair of points $\zeta_1, \zeta_2 \in B_{\boldsymbol{q}}(\boldsymbol{x})$, $\psi(\cdot,\cdot;\boldsymbol{q})$ satisfies the following (constants hidden in $\lesssim$ do not depend on $\boldsymbol{x},\boldsymbol{q},\zeta_1,\zeta_2$):

- *if $\mathfrak{M}$ in Assumption B.3(iii) is the Lebesgue measure, then*

$$\begin{aligned} &\sup_{\lambda\in[0,1]} \sup_{y\notin[\eta(\zeta_1)\wedge\eta(\zeta_2),\eta(\zeta_1)\vee\eta(\zeta_2)]} \big|\psi(y,\eta(\zeta_1+\lambda(\zeta_2-\zeta_1));\boldsymbol{q}) - \psi(y,\eta(\zeta_2);\boldsymbol{q})\big| \lesssim |\zeta_1-\zeta_2|^{\alpha}, \\ &\sup_{\lambda\in[0,1]} \sup_{y\in[\eta(\zeta_1)\wedge\eta(\zeta_2),\eta(\zeta_1)\vee\eta(\zeta_2)]} \big|\psi(y,\eta(\zeta_1+\lambda(\zeta_2-\zeta_1));\boldsymbol{q}) - \psi(y,\eta(\zeta_2);\boldsymbol{q})\big| \lesssim 1; \end{aligned} \tag{B.2}$$

- *if $\mathfrak{M}$ is* not *the Lebesgue measure, then*

$$\sup_{\lambda\in[0,1]} \sup_{y\in\mathcal{Y}} \big|\psi(y,\eta(\zeta_1+\lambda(\zeta_2-\zeta_1));\boldsymbol{q}) - \psi(y,\eta(\zeta_2);\boldsymbol{q})\big| \lesssim |\zeta_1-\zeta_2|^{\alpha}. \tag{B.3}$$

(iv) $\Psi(\boldsymbol{x},\eta;\boldsymbol{q}) := \mathbb{E}[\psi(y_i,\eta;\boldsymbol{q}) \,|\, \boldsymbol{x}_i = \boldsymbol{x}]$ is twice continuously differentiable with respect to η,

$$\sup_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{q}\in\mathcal{Q}} \sup_{\zeta\in B_{\boldsymbol{q}}(\boldsymbol{x})} |\Psi_k(\boldsymbol{x},\eta(\zeta);\boldsymbol{q})| < \infty, \quad \Psi_k(\boldsymbol{x},\eta;\boldsymbol{q}) := \frac{\partial^k}{\partial\eta^k}\Psi(\boldsymbol{x},\eta;\boldsymbol{q}), \quad k=1,2, \tag{B.4}$$

and

$$\inf_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{q}\in\mathcal{Q}} \inf_{\zeta\in B_{\boldsymbol{q}}(\boldsymbol{x})} \Psi_1(\boldsymbol{x},\eta(\zeta);\boldsymbol{q})\eta^{(1)}(\zeta)^2 > 0.$$

In addition, we make the following assumption on the complexity of $\{\psi(\cdot,\cdot;\boldsymbol{q})\}$. It is much more general than Assumption 5 in the main paper (Proposition B.9).

Assumption B.5 (Complexity of $\{\psi(\cdot,\cdot;\boldsymbol{q})\}$.)**.** *For any fixed $r>0$ and $c>0$, $l\in\{1,\dots,K\}$, the classes of functions*

$$\begin{aligned} \mathcal{G}_1 &:= \big\{\mathcal{X}\times\mathcal{Y} \ni (\boldsymbol{x},y) \mapsto \psi(y,\eta(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta});\boldsymbol{q}) - \psi(y,\eta(\mu_0(\boldsymbol{x},\boldsymbol{q}));\boldsymbol{q})\colon \\ &\qquad \|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})\|_\infty \le r, \boldsymbol{q}\in\mathcal{Q}\big\} \\ \mathcal{G}_2 &:= \big\{\mathcal{X}\times\mathcal{Y} \ni (\boldsymbol{x},y) \mapsto \psi(y,\eta(\mu_0(\boldsymbol{x},\boldsymbol{q}));\boldsymbol{q})\colon \boldsymbol{q}\in\mathcal{Q}\big\}, \\ \mathcal{G}_3 &:= \big\{\mathcal{X}\times\mathcal{Y} \ni (\boldsymbol{x},y) \mapsto \\ &\qquad [\psi(y,\eta(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta});\boldsymbol{q}) - \psi(y,\eta(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})]\mathbb{1}(\boldsymbol{x}\in\mathcal{N}_{[c\log n]}(\delta))\colon \\ &\qquad \|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})\|_\infty \le r, \delta\in\Delta, \boldsymbol{q}\in\mathcal{Q}\big\}, \\ \mathcal{G}_4 &:= \big\{\mathcal{X} \ni \boldsymbol{x} \mapsto \Psi_1(\boldsymbol{x},\eta(\mu_0(\boldsymbol{x},\boldsymbol{q}));\boldsymbol{q}) : \boldsymbol{q}\in\mathcal{Q}\big\}, \\ \mathcal{G}_5 &:= \big\{\mathcal{X}\times\mathcal{Y} \ni (\boldsymbol{x},y) \mapsto \\ &\qquad p_l(\boldsymbol{x})[\psi(y,\eta(\mu_0(\boldsymbol{x},\boldsymbol{q}));\boldsymbol{q}) - \psi(y,\eta(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})] : \boldsymbol{q}\in\mathcal{Q}\big\} \end{aligned}$$

satisfy the uniform entropy bound (A.3) with respective envelopes and constants as follows:

$$\begin{aligned} &\bar{G}_1 \lesssim 1, && A_1 \lesssim 1, && V_1 \lesssim K \asymp h^{-d}; \\ &\bar{G}_2(\boldsymbol{x},y) \lesssim \overline{\psi}(\boldsymbol{x},y), && A_2 \lesssim 1, && V_2 \lesssim 1; \\ &\bar{G}_3 \lesssim 1, && A_3 \lesssim 1, && V_3 \lesssim \log^d n; \\ &\bar{G}_4 \lesssim 1, && A_4 \lesssim 1, && V_4 \lesssim 1; \\ &\bar{G}_5 \lesssim 1, && A_5 \lesssim 1, && V_5 \lesssim 1, \end{aligned}$$

where for $s\in\mathbb{Z}\cap[0,\infty)$, $\mathcal{N}_s(\delta)$ denotes the s-neighborhood of cell $\delta\in\Delta$ which is the union of all cells $\delta'\in\Delta$ reachable from some point of δ in no more than s steps (following a continuous path).

Assumption B.6 (Approximation error)**.** *There exists a vector of coefficients $\boldsymbol{\beta}_0(\boldsymbol{q}) \in \mathbb{R}^K$ such that for all $\boldsymbol{\varsigma}$ satisfying $|\boldsymbol{\varsigma}| \le \varsigma$ in Assumption B.2,*

$$\sup_{\boldsymbol{q}\in\mathcal{Q},\boldsymbol{x}\in\mathcal{X}} \left|\mu_0^{(\boldsymbol{\varsigma})}(\boldsymbol{x},\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})^\mathsf{T}\boldsymbol{p}^{(\boldsymbol{\varsigma})}(\boldsymbol{x};\Delta,m)\right| \lesssim h^{m-|\boldsymbol{\varsigma}|}.$$

In particular, this requires $\sup_{\boldsymbol{q},\boldsymbol{x}}\left|\mu_0^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})^\mathsf{T}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x};\Delta,m)\right| \lesssim h^{m-|\boldsymbol{v}|}$.

Assumption B.7 (Estimator of the Gram matrix)**.** *$\widehat{\mathbf{Q}}_{\boldsymbol{q}}$ is an estimator of the matrix $\bar{\mathbf{Q}}_{\boldsymbol{q}}$ such that $\left\|\widehat{\mathbf{Q}}_{\boldsymbol{q}} - \bar{\mathbf{Q}}_{\boldsymbol{q}}\right\|_\infty \lesssim_\mathbb{P} h^d r_{\mathtt{Q}}$ and $\left\|\widehat{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\right\|_\infty \lesssim_\mathbb{P} h^{-d}$, where $r_{\mathtt{Q}} = o(1)$.*

Finally, we state the following technical addition to Assumption B.5 needed for Theorem E.1. As proven in Proposition B.9, it is also automatically true under Assumption 5 in the main paper.

Assumption B.8 (Additional complexity assumption)**.** *For any fixed $c > 0$, $\gamma > 0$, and a positive sequence $\varepsilon_n \to 0$ the class of functions*

$$\begin{aligned}\Big\{(\boldsymbol{x},y) \mapsto \int_{-\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{v}}^{0} &\left[\psi(y,\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta})+t);\boldsymbol{q}) - \psi(y,\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\right] \\ &\times \eta^{(1)}(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta})+t)\,\mathrm{d}t \cdot \mathbb{1}(\boldsymbol{x}\in\mathcal{N}_{[c\log n]}(\delta)) : \\ \|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})\|_\infty \le \gamma, \|\boldsymbol{v}\|_\infty \le \varepsilon_n, \delta\in\Delta, \boldsymbol{q}\in\mathcal{Q}\Big\}\end{aligned}$$

with envelope ε_n multiplied by a large enough constant (not depending on n), satisfies the uniform entropy bound (A.3) with $A \lesssim 1/\varepsilon_n$, $V \lesssim \log^d n$, where the constants in $\lesssim$ do not depend on n.

B.1 Simplifying assumption on loss function

As discussed in the paper, it is possible to impose a simplifying assumption on the loss function, which is motivated by the examples considered. More specifically, we aim at simplifying the general complexity Assumptions B.5 and B.8.

Proposition B.9 (Simplified setup)**.** *Assume the following conditions.*

(i) $q \mapsto \mu_0(\boldsymbol{x},q)$ is non-decreasing.

(ii) $\rho(y,\eta;q) = \sum_{j=1}^4 \rho_j(y,\eta;q)$, where the functions $\rho_j(y,\eta;q)$ are of the following types:

- *Type I: $\rho_1(y,\eta;q) = (f_1(y) + D_1\eta)\mathbb{1}\{y \le \eta\}$,*
- *Type II: $\rho_2(y,\eta;q) = (f_2(y) + D_2\eta)\mathbb{1}\{y \le q\}$,*
- *Type III: $\rho_3(y,\eta;q) = (f_3(y) + D_3\eta)q$,*
- *Type IV: $\rho_4(y,\eta;q) = \mathcal{T}(y,\eta)$,*

where $f_j(\cdot)$ are fixed continuous functions, D_j are universal constants, and $\mathcal{T}(y;\eta) : \mathcal{Y}\times\mathcal{E} \to \mathbb{R}$ is a measurable function, differentiable in η for any fixed y with a derivative $\tau(y,\eta) := \frac{\partial}{\partial\eta}\mathcal{T}(y,\eta)$.

(iii) $\mathbb{E}[\tau(y_i,\eta)\,|\,\boldsymbol{x}_i = \boldsymbol{x}]$ is also differentiable in η for any $\boldsymbol{x}$. The functions

$$\begin{aligned}&(y,\eta) \mapsto \tau(y,\eta), \text{ and} \\ &(\boldsymbol{x},\eta) \mapsto \frac{\partial}{\partial\eta}\mathbb{E}[\tau(y_i,\eta)\,|\,\boldsymbol{x}_i = \boldsymbol{x}]\end{aligned}$$

are continuous in their arguments, and α-Hölder continuous (with $0 < \alpha \le 1$) in η on $\mathcal{Y}\times\mathcal{K}$ and $\mathcal{X}\times\mathcal{K}$ respectively, where $\mathcal{K}$ is any fixed compact subset of $\mathcal{E}$ (with the Hölder constant possibly depending on $\mathcal{K}$, but on y or $\boldsymbol{x}$).

(iv) Assumptions B.1 to B.4 hold with $\overline{\psi}(\boldsymbol{x}, y) = \overline{\tau}(\boldsymbol{x}, y) + |D_1| + |D_2| + |D_3| \max_{q\in\mathcal{Q}} |q|$, *where* $\overline{\tau}(\boldsymbol{x}, y)$ *is a measurable envelope of* $\{(\boldsymbol{x}, y) \mapsto \tau(y, \eta(\mu_0(\boldsymbol{x}, q)))\}$.

(v) If D_1 *is nonzero,* $F_{Y|X}$ *is differentiable and* $f_{Y|X}$ *is its derivative (in particular,* $\mathfrak{M}$ *is Lebesgue measure),* $(\boldsymbol{x}, \eta) \mapsto f_{Y|X}(\eta|\boldsymbol{x})$ *is continuous in both arguments and continuously differentiable in* η.

Then Assumptions B.5 and B.8 also hold.

We will prove this now.

In this case

$$\psi(y, \eta; q) = D_1\mathbb{1}\{y \le \eta\} + D_2\mathbb{1}\{y \le q\} + D_3q + \tau(y, \eta).$$

B.1.1 Verifying Assumption B.5

Lemma B.10 (Class G_1)**.** *The class* $\mathcal{G}_1$ *described in Assumption B.5 with a large enough constant envelope satisfies the uniform entropy bound (A.3) with* $A \lesssim 1$, $V \lesssim K$.

Proof of Lemma B.10. It is shown in the proof of Proposition H.1 (replacing $<$ with $\le$ does not change the argument) that for any fixed $r > 0$ the class

$$\{(\boldsymbol{x}, y) \mapsto \mathbb{1}\{y \le \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})\} - \mathbb{1}\{y \le \eta(\mu_0(\boldsymbol{x}, q))\} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, q \in \mathcal{Q}\}$$

with envelope 2 satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim K$.

Next, assume that the infinity-norms of $\boldsymbol{\beta}$ and $\widetilde{\boldsymbol{\beta}}$ lie in a fixed bounded interval, and let $q, \tilde{q} \in \mathcal{Q}$. Note that by α-Hölder continuity of $\tau(y, \cdot)$ on compacta

$$\begin{aligned}
&\left|\tau(y, \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})) - \tau(y; \eta(\mu_0(\boldsymbol{x}, q))) - \tau(y, \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\widetilde{\boldsymbol{\beta}})) + \tau(y, \eta(\mu_0(\boldsymbol{x}, \tilde{q})))\right| \\
&\quad\lesssim \left|\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}) - \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\widetilde{\boldsymbol{\beta}})\right|^\alpha + \left|\eta(\mu_0(\boldsymbol{x}, q)) - \eta(\mu_0(\boldsymbol{x}, \tilde{q}))\right|^\alpha \\
&\quad\lesssim \|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|_\infty^\alpha + |q - \tilde{q}|^\alpha,
\end{aligned}$$

where the constants in $\lesssim$ do not depend on $\boldsymbol{\beta}$, $\widetilde{\boldsymbol{\beta}}$, q, $\tilde{q}$, and we used that $\eta(\cdot)$ on a fixed bounded interval is Lipschitz, and $q \mapsto \mu_0(\boldsymbol{x}, q)$ is Lipschitz (uniformly in $\boldsymbol{x}$). Again by α-Hölder continuity for any fixed $r > 0$ the class

$$\{(\boldsymbol{x}, y) \mapsto \tau(y, \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})) - \tau(y, \eta(\mu_0(\boldsymbol{x}, q))) : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, q \in \mathcal{Q}\}$$

has a constant envelope. It follows that it satisfies the uniform entropy bound (A.3) with $A \lesssim 1$, $V \lesssim K$.

Combining these results concludes the proof of Lemma B.10 by Lemma C.4. ☐

Lemma B.11 (Class G_2)**.** *The class* $\mathcal{G}_2$ *described in Assumption B.5 with envelope* $\overline{\psi}(\boldsymbol{x}, y)$ *satisfies the uniform entropy bound (A.3) with* $A \lesssim 1$, $V \lesssim 1$.

Proof of Lemma B.11. The class

$$\{\mathcal{X} \times \mathcal{Y} \ni (\boldsymbol{x}, y) \mapsto \tau(y, \eta(\mu_0(\boldsymbol{x}, q))) : q \in \mathcal{Q}\}$$

with envelope $\overline{\tau}(\boldsymbol{x}, y)$ satisfies the uniform entropy bound (A.3) with $A \lesssim 1$, $V \lesssim 1$ by α-Hölder continuity and since $\eta(\cdot)$ on a fixed bounded interval is Lipschitz, $q \mapsto \mu_0(\boldsymbol{x}, q)$ is Lipschitz (uniformly in $\boldsymbol{x}$).

The class

$$\{\mathcal{X} \times \mathcal{Y} \ni (\boldsymbol{x}, y) \mapsto \mathbb{1}\{y \le \eta(\mu_0(\boldsymbol{x}, q))\} : q \in \mathcal{Q}\}$$

with envelope 1 satisfies the uniform entropy bound (A.3) with $A \lesssim 1$, $V \lesssim 1$ because $\mu_0(\boldsymbol{x}, q)$ is nondecreasing in q, see the proof of Proposition H.1.

The class

$$\big\{\mathcal{X} \times \mathcal{Y} \ni (\boldsymbol{x}, y) \mapsto \mathbb{1}\{y \le q\} \colon q \in \mathcal{Q}\big\}$$

with envelope 1 satisfies the uniform entropy bound (A.3) with $A \lesssim 1$, $V \lesssim 1$ because it is VC-subgraph with a constant index, (cf. the proof of Proposition H.3).

The class

$$\big\{\mathcal{X} \times \mathcal{Y} \ni (\boldsymbol{x}, y) \mapsto q \colon q \in \mathcal{Q}\big\}$$

with envelope $\max_{q \in \mathcal{Q}} |q|$ satisfies the uniform entropy bound (A.3) with $A \lesssim 1$, $V \lesssim 1$ because it is VC-subgraph with a constant index (as a subclass of a one-dimensional space of functions, namely constants in $\boldsymbol{x}, y$).

It is left to apply Lemma C.4, concluding the proof of Lemma B.11. $\square$

Lemma B.12 (Class G_3)**.** *The class $\mathcal{G}_3$ described in Assumption B.5 with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$, $V \lesssim \log^d n$.*

Proof of Lemma B.12. Fix $\delta \in \Delta$ and some large enough $R > 0$. Note that if $\mathbb{1}\big\{\boldsymbol{x} \in \mathcal{N}_{[c \log n]}(\delta)\big\}$ and $\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}$ are both nonzero, $\boldsymbol{\beta}$ must lie in a vector subspace $\mathcal{B}_\delta$ of $\mathbb{R}^K$ of dimension $O(\log^d n)$. For any positive and small enough ε, the class of vectors $\big\{\boldsymbol{\beta} \in \mathcal{B}_\delta, \|\boldsymbol{\beta}\|_\infty < R\big\}$ has an infinity-norm ε-net $\bar{\mathcal{B}}_\delta^\varepsilon$ such that

$$\log|\bar{\mathcal{B}}_\delta^\varepsilon| \lesssim \log^d n \log(C/\varepsilon),$$

where C is some positive constant.

By α-Hölder continuity of $\tau(y, \cdot)$ on compacta and since $\eta(\cdot)$ on a compact is Lipschitz, this means that the class of bounded (by a constant not depending on n) functions

$$\begin{aligned}\mathcal{G}_{3,\delta} := \big\{&(\boldsymbol{x}, y) \mapsto \big[\tau\big(y, \eta(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta})\big) - \tau\big(y, \eta(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}_0(q))\big)\big]\mathbb{1}\{\boldsymbol{x} \in \mathcal{N}_{[c \log n]}(\delta)\} : \\ &\|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, q \in \mathcal{Q}\big\}\end{aligned}$$

has a covering number bound

$$\log N(\mathcal{G}_{3,\delta}, \text{sup-norm}, C'\varepsilon^\alpha) \lesssim \log^d n \log(C/\varepsilon),$$

where C' is some other positive constant. This means that also

$$\log N(\mathcal{G}_{3,\delta}, \text{sup-norm}, \varepsilon) \lesssim \log^d n \log(C''/\varepsilon)$$

for some other positive constant C''. Finally, from this we can conclude that $\mathcal{G}_{3,\delta}$ with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim \log^d n$. Therefore, the union of $O(h^{-d})$ such classes

$$\begin{aligned}\big\{&(\boldsymbol{x}, y) \mapsto \big\{\tau\big(y, \eta(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta})\big) - \tau\big(y, \eta(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}_0(q))\big)\big\}\mathbb{1}\{\boldsymbol{x} \in \mathcal{N}_{[c \log n]}(\delta)\} : \\ &\|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, q \in \mathcal{Q}, \delta \in \Delta\big\}\end{aligned}$$

with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim \log^d n$, see (E.34).

Next, since $\eta(\cdot)$ is monotonic, the functions $\boldsymbol{x} \mapsto \boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}$ with $\boldsymbol{\beta} \in \mathcal{B}_\delta$ form a vector space of $O(\log^d n)$ dimension, and $\boldsymbol{x} \mapsto \mathbb{1}\{\boldsymbol{x} \in \mathcal{N}_{[c \log n]}(\delta)\}$ is one fixed function, the class

$$\big\{(\boldsymbol{x}, y) \mapsto \mathbb{1}\{y \le \eta(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta})\}\mathbb{1}\{\boldsymbol{x} \in \mathcal{N}_{[c \log n]}(\delta)\} : \boldsymbol{\beta} \in \mathbb{R}^K\big\}$$

is VC with an index $O(\log^d n)$. Therefore, it satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim \log^d n$. By Lemma C.4, a subclass of the difference of two such classes

$$\big\{(\boldsymbol{x}, y) \mapsto \big(\mathbb{1}\{y \le \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})\} - \mathbb{1}\{y \le \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(q))\}\big)\mathbb{1}\{\boldsymbol{x} \in \mathcal{N}_{[c\log n]}(\delta)\} :$$
$$\|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, q \in \mathcal{Q}\big\}$$

for fixed $\delta \in \Delta$ also satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim \log^d n$. Therefore, the union of $O(h^{-d})$ such classes

$$\big\{(\boldsymbol{x}, y) \mapsto \{\mathbb{1}\{y \le \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})\} - \mathbb{1}\{y \le \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(q))\}\}\mathbb{1}\{\boldsymbol{x} \in \mathcal{N}_{[c\log n]}(\delta)\} :$$
$$\|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, \delta \in \Delta, q \in \mathcal{Q}\big\}$$

with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim \log^d n$, see (E.34).

It is left to apply Lemma C.4 once again, concluding the proof of Lemma B.12. ☐

Lemma B.13 (Class G_4)**.** *The class $\mathcal{G}_4$ described in Assumption B.5 with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$, $V \lesssim 1$.*

Proof of Lemma B.13. In this case

$$\Psi(\boldsymbol{x}, \eta; q) = \mathbb{E}[\tau(y_i, \eta) \,|\, \boldsymbol{x}_i = \boldsymbol{x}] + D_1 F_{Y|X}(\eta|\boldsymbol{x}) + D_2 F_{Y|X}(q|\boldsymbol{x}) + D_3 q$$

and

$$\Psi_1(\boldsymbol{x}, \eta; q) = \frac{\partial}{\partial \eta}\mathbb{E}[\tau(y_i, \eta) \,|\, \boldsymbol{x}_i = \boldsymbol{x}] + D_1 f_{Y|X}(\eta|\boldsymbol{x}).$$

By assumption, if η lies in a fixed compact, this function of $\boldsymbol{x}, \eta$ is bounded (by continuity). Moreover,

$$\begin{aligned}&\big|\Psi_1(\boldsymbol{x}, \eta(\mu_0(\boldsymbol{x}, q)); q) - \Psi_1(\tilde{q}, \boldsymbol{x}; q)\eta(\mu_0(\boldsymbol{x}, \tilde{q}))\big| \\ &\quad\lesssim |\eta(\mu_0(\boldsymbol{x}, q)) - \eta(\mu_0(\boldsymbol{x}, \tilde{q}))|^\alpha + |\eta(\mu_0(\boldsymbol{x}, q)) - \eta(\mu_0(\boldsymbol{x}, \tilde{q}))| \overset{\text{(a)}}{\lesssim} |q - \tilde{q}|^\alpha + |q - \tilde{q}|,\end{aligned}$$

where in (a) we used that $\eta(\cdot)$ on compacta is Lipschitz and $q \mapsto \mu_0(\boldsymbol{x}, q)$ is uniformly Lipschitz in q. The result of Lemma B.13 follows. ☐

Lemma B.14 (Class G_5)**.** *The class $\mathcal{G}_5$ described in Assumption B.5 with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$, $V \lesssim 1$.*

Proof of Lemma B.14. Take $R > 0$ fixed and large enough so that $\|\boldsymbol{\beta}_{0,q}\|_\infty \le R$ for all q and n. Note that for $p_l(\boldsymbol{x})$ and $\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}$ to be nonzero at the same time, $\boldsymbol{\beta}$ must lie in a fixed vector subspace $\mathcal{B}_l$ of $\mathbb{R}^K$ of bounded dimension. For $0 < \varepsilon < 1$, the class of vectors $\{\boldsymbol{\beta} \in \mathcal{B}_l, \|\boldsymbol{\beta}\|_\infty \le R\}$ has an infinity-norm ε-net $\bar{\mathcal{B}}_l^\varepsilon$ such that

$$\log\big|\bar{\mathcal{B}}_l^\varepsilon\big| \lesssim \log(C/\varepsilon),$$

where C is some positive constant.

By α-Hölder continuity of $\tau(y, \cdot)$ on compacta, since $\eta(\cdot)$ on a compact is Lipschitz and $\mu_0(\cdot, q)$ is uniformly Lipschitz in q, this means that the class of bounded (by a constant not depending on n) functions

$$\mathcal{G}_5' := \big\{(\boldsymbol{x}, y) \mapsto p_l(\boldsymbol{x})\big(\tau(y, \eta(\mu_0(\boldsymbol{x}, q))) - \tau(y, \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(q)))\big) : q \in \mathcal{Q}\big\}$$

has a covering number bound

$$\log N(\mathcal{G}_5', \text{sup-norm}, C'\varepsilon^\alpha) \lesssim \log(C/\varepsilon),$$

where C' is some other positive constant. It follows that this class with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim 1$.

As in the proof of Lemma B.11, the class

$$\big\{\mathcal{X} \times \mathcal{Y} \ni (\boldsymbol{x}, y) \mapsto \mathbb{1}\{y \le \eta(\mu_0(\boldsymbol{x}, q))\} \colon q \in \mathcal{Q}\big\}$$

with envelope 1 satisfies the uniform entropy bound (A.3) with $A \lesssim 1$, $V \lesssim 1$. Since $\eta(\cdot)$ is monotonic, the functions $\boldsymbol{x} \mapsto \boldsymbol{\beta}^\mathsf{T}\boldsymbol{p}(\boldsymbol{x})$ with $\boldsymbol{\beta} \in \mathcal{B}_l$ form a vector space, and $p_l(\boldsymbol{x})$ is one fixed function, we have that the class

$$\big\{\mathcal{X} \times \mathcal{Y} \ni (\boldsymbol{x}, y) \mapsto p_l(\boldsymbol{x})\mathbb{1}\{y \le \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})\} : \boldsymbol{\beta} \in \mathbb{R}^K\big\}$$

is VC with a bounded index. Then it also satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim 1$.

It is left to apply Lemma C.4, concluding the proof of Lemma B.14. □

B.1.2 Verifying Assumption B.8

Suppose $\theta_1, \theta_2, \theta \in \mathbb{R}$ and $\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\theta} \in \mathbb{R}$ all lie in a fixed compact interval. Then

$$\begin{aligned}
&\mathcal{T}(y, \eta(\theta_1)) - \mathcal{T}(y, \eta(\theta_2)) - [\eta(\theta_1) - \eta(\theta_2)]\tau(y, \eta(\theta)) \\
&\quad - \mathcal{T}(y, \eta(\tilde{\theta}_1)) + \mathcal{T}(y, \eta(\tilde{\theta}_2)) + [\eta(\tilde{\theta}_1) - \eta(\tilde{\theta}_2)]\tau(y, \eta(\tilde{\theta})) \\
&= \tau(y, \zeta_{1,y})[\eta(\theta_1) - \eta(\tilde{\theta}_1)] - \tau(y, \zeta_{2,y})[\eta(\theta_2) - \eta(\tilde{\theta}_2)] \\
&\quad - [\eta(\theta_1) - \eta(\tilde{\theta}_1) + \eta(\tilde{\theta}_1) - \eta(\theta_2)]\tau(y, \eta(\theta)) + [\eta(\tilde{\theta}_1) - \eta(\theta_2) + \eta(\theta_2) - \eta(\tilde{\theta}_2)]\tau(y, \eta(\tilde{\theta})) \\
&= [\tau(y, \zeta_{1,y}) - \tau(y, \eta(\theta))] \cdot [\eta(\theta_1) - \eta(\tilde{\theta}_1)] - [\tau(y, \zeta_{2,y}) - \tau(y, \eta(\tilde{\theta}))] \cdot [\eta(\theta_2) - \eta(\tilde{\theta}_2)] \\
&\quad - [\eta(\tilde{\theta}_1) - \eta(\theta_2)]\tau(y, \eta(\theta)) + [\eta(\tilde{\theta}_1) - \eta(\theta_2)]\tau(y, \eta(\tilde{\theta})) \\
&= \underbrace{[\tau(y, \zeta_{1,y}) - \tau(y, \eta(\theta))]}_{\lesssim 1} \cdot \underbrace{[\eta(\theta_1) - \eta(\tilde{\theta}_1)]}_{\lesssim |\theta_1 - \tilde{\theta}_1|} - \underbrace{[\tau(y, \zeta_{2,y}) - \tau(y, \eta(\tilde{\theta}))]}_{\lesssim 1} \cdot \underbrace{[\eta(\theta_2) - \eta(\tilde{\theta}_2)]}_{\lesssim |\theta_2 - \tilde{\theta}_2|} \\
&\quad + \underbrace{[\eta(\tilde{\theta}_1) - \eta(\theta_2)]}_{\lesssim 1} \underbrace{[\tau(y, \eta(\tilde{\theta})) - \tau(y, \eta(\theta))]}_{\lesssim |\theta - \tilde{\theta}|^\alpha} \lesssim |\theta_1 - \tilde{\theta}_1| + |\theta_2 - \tilde{\theta}_2| + |\theta - \tilde{\theta}|^\alpha
\end{aligned}$$

for some $\zeta_{1,y}$ between $\eta(\theta_1)$ and $\eta(\tilde{\theta}_1)$, $\zeta_{2,y}$ between $\eta(\theta_2)$ and $\eta(\tilde{\theta}_2)$, where we used the α-Hölder continuity of $\tau(y, \cdot)$ on a fixed compact and the Lipschitzness of $\eta(\cdot)$ on a compact. This means that the class of functions

$$\begin{aligned}
\mathcal{G}' := \Big\{(\boldsymbol{x}, y) \mapsto \Big(&\mathcal{T}\big(y, \eta\big(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta})\big)\big) - \mathcal{T}\big(y, \eta\big(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta} - \boldsymbol{v})\big)\big) \\
&- \big[\eta\big(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta})\big) - \eta\big(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta} - \boldsymbol{v})\big)\big]\tau\big(y, \eta\big(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(q)\big)\big)\Big) \\
&\times \mathbb{1}\{\boldsymbol{x} \in \mathcal{N}_{[c \log n]}(\delta)\} : q \in \mathcal{Q}, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, \|\boldsymbol{v}\|_\infty \le \varepsilon_n\Big\}
\end{aligned}$$

has a covering number bound

$$\log N(\mathcal{G}', \text{sup-norm}, \varepsilon) \lesssim \log^d n \log\left(\frac{C_1}{\varepsilon}\right), \tag{B.5}$$

for all small enough positive ε, and C_1 is some positive constant (not depending on n), where we used that all $\boldsymbol{\beta}_0(q)$, $\boldsymbol{\beta}_0(q)+\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0(q)+\boldsymbol{\beta}-\boldsymbol{v}$ must lie in a vector space of dimension $O(\log^d n)$ if $\mathbb{1}\{\boldsymbol{x}\in\mathcal{N}_{[c\log n]}(\delta)\}$ is not zero. Applying the mean-value theorem to

$$\mathcal{T}\big(y,\eta\big(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta})\big)\big)-\mathcal{T}\big(y,\eta\big(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta}-\boldsymbol{v})\big)\big)$$

and α-Hölder continuity again, we see that class $\mathcal{G}'$ has an envelope which is ε_n multiplied by a large enough constant C_2. Replacing ε with $C_2\varepsilon\varepsilon_n$ (for large enough n this is small enough) in (B.5), we get a covering number bound

$$\log N(\mathcal{G}',\text{sup-norm},C_2\varepsilon\varepsilon_n)\lesssim\log^d n\log\bigg(\frac{C_1}{C_2\varepsilon\varepsilon_n}\bigg).$$

It follows that class $\mathcal{G}'$ satisfies the uniform entropy bound (A.3) with $A\lesssim 1/\varepsilon_n$ and $V\lesssim\log^d n$.

Therefore, the union of $O(h^{-d})$ such classes

$$\begin{aligned}\mathcal{G}':=\Big\{&\Big(\mathcal{T}\big(y,\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta}))\big)-\mathcal{T}\big(y,\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta}-\boldsymbol{v}))\big)\\&-\big[\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta}))-\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta}-\boldsymbol{v}))\big]\tau(y,\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(q)))\Big)\\&\times\mathbb{1}\{\boldsymbol{x}\in\mathcal{N}_{[c\log n]}(\delta)\}:q\in\mathcal{Q},\|\boldsymbol{\beta}-\boldsymbol{\beta}_0(q)\|_\infty\le r,\|\boldsymbol{v}\|_\infty\le\varepsilon_n,\delta\in\Delta\Big\},\end{aligned}$$

also with envelope ε_n multiplied by a large enough constant, satisfies the uniform entropy bound (A.3) with $A\lesssim 1/\varepsilon_n$ and $V\lesssim\log^d n$, by the same argument as (E.34).

Next,

$$\begin{aligned}&[f_1(y)+D_1\eta(\theta_1)]\mathbb{1}\{y\le\eta(\theta_1)\}-[f_1(y)+D_1\eta(\theta_2)]\mathbb{1}\{y\le\eta(\theta_2)\}\\&\quad-[\eta(\theta_1)-\eta(\theta_2)]D_1\mathbb{1}\{y\le\eta(\theta)\}\\&=f_1(y)[\mathbb{1}\{y\le\eta(\theta_1)\}-\mathbb{1}\{y\le\eta(\theta_2)\}]+D_1\eta(\theta_1)\mathbb{1}\{y\le\eta(\theta_1)\}-D_1\eta(\theta_2)\mathbb{1}\{y\le\eta(\theta_2)\}\\&\quad-D_1[\eta(\theta_1)-\eta(\theta_2)]\mathbb{1}\{y\le\eta(\theta)\}.\end{aligned}$$

It is proven by the same argument as in the proof of Proposition H.1 that the class

$$\begin{aligned}\Big\{(\boldsymbol{x},y)\mapsto&\Big(\big[f_1(y)+D_1\eta\big(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta})\big)\big]\mathbb{1}\big\{y\le\eta\big(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta})\big)\big\}\\&-\big[f_1(y)+D_1\eta\big(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta}-\boldsymbol{v})\big)\big]\mathbb{1}\{y\le\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta}-\boldsymbol{v}))\}\\&-\big[\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta}))-\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta}-\boldsymbol{v}))\big]D_1\mathbb{1}\{y\le\eta(\theta)\}\Big)\\&\times\mathbb{1}\{\boldsymbol{x}\in\mathcal{N}_{[c\log n]}(\delta)\}:q\in\mathcal{Q},\|\boldsymbol{\beta}-\boldsymbol{\beta}_0(q)\|_\infty\le r,\|\boldsymbol{v}\|_\infty\le\varepsilon_n,\delta\in\Delta\Big\}\end{aligned}$$

satisfies the uniform entropy bound (A.3) with $A\lesssim 1/\varepsilon_n$ and $V\lesssim\log^d n$.

The terms $(f_2(y)+D_2\eta)\mathbb{1}\{y\le q\}$ and $(f_3(y)+D_3\eta)q$ play no role in this verification because they cancel out in the class described in Assumption B.8.

It is left to apply Lemma C.4.

The proof of Proposition B.9 is finished.

C Frequently used lemmas

We collect several lemmas that will be used multiple times throughout this supplemental appendix. Lemmas C.1 to C.9 are well-known facts, so we provide either brief proofs or references to the literature.

Lemma C.1 (Second moment bound of the max of sub-Gaussian random variables)**.** *Let $n \geq 3$ and $\xi_1, \dots, \xi_n$ be σ^2-sub-Gaussian random variables (not necessarily independent). Then*

$$\mathbb{E}\Big[\max_{1\leq i\leq n} \xi_i^2\Big]^{1/2} \leq C_U \sigma \sqrt{\log n},$$

where C_U is a universal constant.

Proof. If p is an even positive integer, $\mathbb{E}[\xi_i^p] \leq 3\sigma^p p(p/2)^{p/2}$. Then

$$\begin{aligned}\mathbb{E}\Big[\max_{1\leq i\leq n} \xi_i^2\Big]^{1/2} &\leq \mathbb{E}\Big[\max_{1\leq i\leq n} \xi_i^p\Big]^{1/p} \leq \left(\sum_{i=1}^n \mathbb{E}\big[\xi_i^p\big]\right)^{1/p} \lesssim n^{1/p} \cdot \sigma \cdot p^{1/p}\sqrt{p} \\ &\lesssim \sigma n^{1/p}\sqrt{p} \qquad \text{using } p^{1/p} \leq 2.\end{aligned}$$

It is left to take $p = p_n$ such that $\ln n \leq p \leq 2\ln n$. □

Lemma C.2 (Boundedness of conditional expectation in probability implies unconditional boundedness in probability)**.** *Let X_n be a sequence of integrable random variables, $\boldsymbol{D}_n$ a sequence of random vectors, r_n a sequence of positive numbers. If $\mathbb{E}[|X_n| \,|\, \boldsymbol{D}_n] \lesssim_{\mathbb{P}} r_n$, then $|X_n| \lesssim_{\mathbb{P}} r_n$.*

Proof. Take any sequence of positive numbers $\gamma_n \to \infty$. By Markov's inequality,

$$\mathbb{P}\{|X_n| > \gamma_n r_n \,|\, \boldsymbol{D}_n\} \leq \frac{\mathbb{E}[|X_n| \,|\, \boldsymbol{D}_n]}{\gamma_n r_n} \lesssim_{\mathbb{P}} \frac{1}{\gamma_n} = o(1).$$

In other words, the sequence of random variables $\mathbb{P}\{|X_n| > \gamma_n r_n \,|\, \boldsymbol{D}_n\}$ converges to zero in probability. By dominated convergence (in probability), the sequence of numbers $\mathbb{P}\{|X_n| > \gamma_n r_n\}$ converges to zero. Since it is true for any positive sequence $\gamma_n \to \infty$, this means $|X_n| = O_{\mathbb{P}}(r_n)$. □

Lemma C.3 (Converging to zero in conditional probability is the same as converging to zero in probability)**.** *Let X_n be a sequence of random variables, $\boldsymbol{D}_n$ a sequence of random vectors. The following are equivalent:*

(i) *for any $\varepsilon > 0$, we have $\mathbb{P}\{|X_n| > \varepsilon \,|\, \boldsymbol{D}_n\} = o_{\mathbb{P}}(1)$;*

(ii) *$|X_n| = o_{\mathbb{P}}(1)$.*

Proof. The implication (i) ⇒ (ii) follows from dominated convergence in probability. To prove the converse, take any $\varepsilon, \gamma > 0$. By Markov's inequality,

$$\mathbb{P}\{\mathbb{P}\{|X_n| > \varepsilon \,|\, \boldsymbol{D}_n\} > \gamma\} \leq \frac{\mathbb{P}\{|X_n| > \varepsilon\}}{\gamma} \to 0,$$

so by definition $\mathbb{P}\{|X_n| > \varepsilon \,|\, \boldsymbol{D}_n\} = o_{\mathbb{P}}(1)$. □

Lemma C.4 (Permanence properties of the uniform entropy bound)**.** *Let $\mathcal{F}$ and $\mathcal{G}$ be two classes of measurable functions from $S \to \mathbb{R}$ on a measurable space $(S, \mathcal{S})$ with strictly positive measurable envelopes F and G respectively. Then the uniform entropy numbers of $\mathcal{F}\mathcal{G} = \{fg : f \in \mathcal{F}, g \in \mathcal{G}\}$ satisfy*

$$\begin{aligned}&\sup_{\mathbb{Q}} \log N\big(\mathcal{F}\mathcal{G}, \|\cdot\|_{\mathbb{Q},2}, \varepsilon\|FG\|_{\mathbb{Q},2}\big) \\ &\quad \leq \sup_{\mathbb{Q}} \log N\left(\mathcal{F}, \|\cdot\|_{\mathbb{Q},2}, \frac{\varepsilon\|F\|_{\mathbb{Q},2}}{2}\right) + \sup_{\mathbb{Q}} \log N\left(\mathcal{G}, \|\cdot\|_{\mathbb{Q},2}, \frac{\varepsilon\|G\|_{\mathbb{Q},2}}{2}\right)\end{aligned}$$

for all $\varepsilon > 0$. Also, the uniform entropy numbers of $\mathcal{F} + \mathcal{G} = \{f + g : f \in \mathcal{F}, g \in \mathcal{G}\}$ satisfy

$$
\begin{aligned}
&\sup_{\mathbb{Q}} \log N\big(\mathcal{F} + \mathcal{G}, \|\cdot\|_{\mathbb{Q},2}, \varepsilon\|F + G\|_{\mathbb{Q},2}\big) \\
&\quad\leq \sup_{\mathbb{Q}} \log N\bigg(\mathcal{F}, \|\cdot\|_{\mathbb{Q},2}, \frac{\varepsilon\|F\|_{\mathbb{Q},2}}{2}\bigg) + \sup_{\mathbb{Q}} \log N\bigg(\mathcal{G}, \|\cdot\|_{\mathbb{Q},2}, \frac{\varepsilon\|G\|_{\mathbb{Q},2}}{2}\bigg)
\end{aligned}
$$

for all $\varepsilon > 0$. In both cases, $\mathbb{Q}$ ranges over all finitely-discrete probability measures.

Proof. This lemma is well-known. See, for example, [41]. □

Lemma C.5 (Maximal inequality for Gaussian vectors)**.** *Take $n \geq 2$. Let $X_i \sim \mathcal{N}(0, \sigma_i^2)$ for $1 \leq i \leq n$ (not necessarily independent), with $\sigma_i^2 \leq \sigma^2$. Then*

$$
\begin{aligned}
\mathbb{E}\Big[\max_{1\leq i\leq n} X_i\Big] &\leq \sigma\sqrt{2\log n}, \\
\mathbb{E}\Big[\max_{1\leq i\leq n} |X_i|\Big] &\leq 2\sigma\sqrt{\log n}.
\end{aligned}
$$

If $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are positive semi-definite $n \times n$ matrices and $\boldsymbol{n} \sim \mathcal{N}(0, \mathbf{I}_n)$, then

$$
\mathbb{E}[\|\mathbf{\Sigma}_1^{1/2}\boldsymbol{n} - \mathbf{\Sigma}_2^{1/2}\boldsymbol{n}\|_\infty] \leq 2\sqrt{\log n}\|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\|_2^{1/2}.
$$

If further $\mathbf{\Sigma}_1$ is positive definite, then

$$
\mathbb{E}[\|\mathbf{\Sigma}_1^{1/2}\boldsymbol{n} - \mathbf{\Sigma}_2^{1/2}\boldsymbol{n}\|_\infty] \leq \sqrt{\log n}\lambda_{\min}(\mathbf{\Sigma}_1)^{-1/2}\|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\|_2.
$$

Proof. See Lemma SA31 in [12]. □

Lemma C.6 (A maximal inequality for i.n.i.d. empirical processes)**.** *Let $X_1, \ldots, X_n$ be independent but not necessarily identically distributed (i.n.i.d.) random variables taking values in a measurable space $(S, \mathcal{S})$. Denote the joint distribution of $X_1, \ldots, X_n$ by $\mathbb{P}$ and the marginal distribution of X_i by $\mathbb{P}_i$, and let $\overline{\mathbb{P}} = n^{-1}\sum_i \mathbb{P}_i$.*

Let $\mathcal{F}$ be a class of Borel measurable functions from S to $\mathbb{R}$ which is pointwise measurable (i.e. it contains a countable subclass which is dense under pointwise convergence), and satisfying the uniform entropy bound (A.3) with parameters A and V. Let F be a strictly positive measurable envelope function for $\mathcal{F}$ (i.e. $|f(s)| \leq |F(s)|$ for all $f \in \mathcal{F}$ and $s \in S$). Suppose that $\|F\|_{\overline{\mathbb{P}},2} < \infty$. Let $\sigma > 0$ satisfy $\sup_{f\in\mathcal{F}} \|f\|_{\overline{\mathbb{P}},2} \leq \sigma \leq \|F\|_{\overline{\mathbb{P}},2}$ and $M = \max_{1\leq i\leq n} F(X_i)$.

For $f \in \mathcal{F}$ define the empirical process

$$
G_n(f) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (f(X_i) - \mathbb{E}[f(X_i)]).
$$

Then we have

$$
\mathbb{E}\Big[\sup_{f\in\mathcal{F}}|G_n(f)|\Big] \lesssim \sigma\sqrt{V\log\big(A\|F\|_{\overline{\mathbb{P}},2}/\sigma\big)} + \frac{\|M\|_{\mathbb{P},2}V\log\big(A\|F\|_{\overline{\mathbb{P}},2}/\sigma\big)}{\sqrt{n}},
$$

where $\lesssim$ is up to a universal constant.

Proof. See Lemmas SA24 and SA25 in [12]. □

Lemma C.7 (Maximal inequalities for Gaussian processes)**.** *Let Z be a separable mean-zero Gaussian process indexed by $x \in \mathcal{X}$. Recall that Z is separable for example if $\mathcal{X}$ is Polish and Z has continuous trajectories. Define its covariance structure on $\mathcal{X} \times \mathcal{X}$ by $\Sigma(x, x') := \mathbb{E}[Z(x)Z(x')]$, and the corresponding semimetric on $\mathcal{X}$ by*

$$\rho(x, x') := \mathbb{E}[(Z(x) - Z(x'))^2]^{1/2} = \big(\Sigma(x, x) - 2\Sigma(x, x') + \Sigma(x', x')\big)^{1/2}.$$

Let $N(\mathcal{X}, \rho, \varepsilon)$ denote the ε-covering number of $\mathcal{X}$ with respect to the semimetric ρ. Define $\sigma := \sup_x \Sigma(x, x)^{1/2}$.

Then there exists a universal constant $C > 0$ such that for any $\delta > 0$,

$$\mathbb{E}\Big[\sup_{x\in\mathcal{X}} |Z(x)|\Big] \le C\sigma + C\int_0^{2\sigma} \sqrt{\log N(\mathcal{X}, \rho, \varepsilon)}\,\mathrm{d}\varepsilon,$$
$$\mathbb{E}\Big[\sup_{\rho(x,x')\le\delta} |Z(x) - Z(x')|\Big] \le C\int_0^{\delta} \sqrt{\log N(\mathcal{X}, \rho, \varepsilon)}\,\mathrm{d}\varepsilon.$$

Proof. This lemma is well-known. See, for example, [41]. □

Lemma C.8 (Closeness in probability implies closeness of conditional quantiles)**.** *Let X_n and Y_n be random variables and $\boldsymbol{D}_n$ be a random vector. Let $F_{X_n}(x|\boldsymbol{D}_n)$ and $F_{Y_n}(x|\boldsymbol{D}_n)$ denote the conditional distribution functions, and $F^{-1}_{X_n}(x|\boldsymbol{D}_n)$ and $F^{-1}_{Y_n}(x|\boldsymbol{D}_n)$ denote the corresponding conditional quantile functions. If $|X_n - Y_n| = o_{\mathbb{P}}(r_n)$, then there exists a sequence of positive numbers $\nu_n \to 0$, depending on r_n, such that w. p. a. 1*

$$F^{-1}_{X_n}(p|\boldsymbol{D}_n) \le F^{-1}_{Y_n}(p + \nu_n|\boldsymbol{D}_n) + r_n \quad \text{and} \quad F^{-1}_{Y_n}(p|\boldsymbol{D}_n) \le F^{-1}_{X_n}(p + \nu_n|\boldsymbol{D}_n) + r_n$$

for all $p \in (\nu_n, 1 - \nu_n)$.

Proof. See Lemma 13 in [3]. □

Lemma C.9 (Anti-concentration for suprema of separable Gaussian processes)**.** *Let $X = (X_t)_{t\in T}$ be a mean-zero separable Gaussian process indexed by a semimetric space T such that $\mathbb{E}[X_t^2] = 1$ for all $t \in T$. Then for any $\varepsilon > 0$,*

$$\sup_{u\in\mathbb{R}} \mathbb{P}\bigg\{\bigg|\sup_{t\in T} |X_t| - u\bigg| \le \varepsilon\bigg\} \le 4\varepsilon\bigg(\mathbb{E}\bigg[\sup_{t\in T} |X_t|\bigg] + 1\bigg).$$

Proof. See Corollary 2.1 in [17]. □

The following lemma appears to be new to the literature at the level of generality considered. It guarantees the existence and gives some properties of the main estimand considered in this paper.

Lemma C.10 (The existence of $\mu(\cdot)$)**.** *Suppose Assumptions B.3(i) and B.3(iii) hold. We will suppress the dependence of $\rho(\cdot, \cdot)$ on $\boldsymbol{q}$ in this lemma because the result can be applied separately for each $\boldsymbol{q}$. Assume $\eta \mapsto \rho(y, \eta)$ is convex on $\mathcal{E}$, $\mathcal{E}$ is an open connected subset of $\mathbb{R}$, $\psi(y, \cdot)$ is the left or right derivative of $\rho(y, \cdot)$ (in particular, it is a subgradient: $(\eta_1 - \eta_0)\psi(y, \eta_0) \le \rho(y, \eta_1) - \rho(y, \eta_0)$), and $\psi(y, \eta)$ is strictly increasing in η for any fixed $y \in \mathcal{Y}$. Assume the real inverse link function $\eta(\cdot)\colon \mathbb{R} \to \mathcal{E}$ is strictly monotonic and two times continuously differentiable.*

Denoting a_l and a_r the left and right ends of $\mathcal{E}$ respectively (possibly $\pm\infty$), assume that for each $\boldsymbol{x} \in \mathcal{X}$ the expectation $\mathbb{E}[\psi(y_i, \zeta) \,|\, \boldsymbol{x}_i = \boldsymbol{x}]$ is negative for (real deterministic) ζ in a neighborhood of a_l, positive for ζ in a neighborhood of a_r, and continuous in ζ (in particular, it crosses zero).

Then for each $\boldsymbol{x} \in \mathcal{X}$ the number $\min\{\zeta : \mathbb{E}[\psi(y_i,\zeta)\,|\,\boldsymbol{x}_i = \boldsymbol{x}] \geq 0\}$ *exists and belongs to $\mathcal{E}$. Moreover,*

$$\begin{aligned}\mu_0(\boldsymbol{x}) &:= \eta^{-1}\big(\min\{\zeta : \mathbb{E}[\psi(y_i,\zeta)\,|\,\boldsymbol{x}_i = \boldsymbol{x}] \geq 0\}\big)\\ &= \eta^{-1}\big(\min\{\zeta : \mathbb{E}[\psi(y_i,\zeta)\,|\,\boldsymbol{x}_i = \boldsymbol{x}] = 0\}\big)\end{aligned} \tag{C.1}$$

defines a Borel-measurable function such that for all $\boldsymbol{x} \in \mathcal{X}$

$$\mu_0(\boldsymbol{x}) \in \operatorname*{arg\,min}_{\zeta\in\mathbb{R}} \mathbb{E}[\rho(y_i,\eta(\zeta))\,|\,\boldsymbol{x}_i = \boldsymbol{x}].$$

If $\mathcal{Q}$ is not a singleton, applying this result for each $\boldsymbol{q} \in \mathcal{Q}$ gives a function $\mu_0(\boldsymbol{x},\boldsymbol{q})$ which is Borel in $\boldsymbol{x}$ for each fixed $\boldsymbol{q}$. Measurability in $\boldsymbol{q}$ is not asserted by this lemma.

Proof. The conditions ensure that $\min\{\zeta : \mathbb{E}[\psi(y_i,\zeta)\,|\,\boldsymbol{x}_i = \boldsymbol{x}] \geq 0\}$ exists and belongs to $\mathcal{E}$ by continuity.

So the function $\mu_0(\boldsymbol{x})$ is well-defined. It is Borel because η^{-1} is continuous and

$$\{\boldsymbol{x} : \eta(\mu_0(\boldsymbol{x})) > a\} = \{\boldsymbol{x} : \mathbb{E}[\psi(y_i,a)\,|\,\boldsymbol{x}_i = \boldsymbol{x}] < 0\}$$

is a Borel set (the equality of the two sets is true because $\zeta \mapsto \psi(y,\zeta)$ is strictly increasing).

For any $\zeta \in \mathbb{R}$, using $(\eta(\zeta) - \eta(\mu_0(\boldsymbol{x})))\psi(y,\eta(\mu_0(\boldsymbol{x}))) \leq \rho(y,\eta(\zeta)) - \rho(y,\eta(\mu_0(\boldsymbol{x})))$, we have

$$\begin{aligned}0 &= (\eta(\zeta) - \eta(\mu_0(\boldsymbol{x})))\,\mathbb{E}\big[\psi(y_i,a)\,\big|\,\boldsymbol{x}_i = \boldsymbol{x}\big]\big|_{a=\eta(\mu_0(\boldsymbol{x}))}\\ &\leq \mathbb{E}[\rho(y,\eta(\zeta))\,|\,\boldsymbol{x}_i = \boldsymbol{x}] - \mathbb{E}[\rho(y,a)\,|\,\boldsymbol{x}_i = \boldsymbol{x}]|_{a=\eta(\mu_0(\boldsymbol{x}))},\end{aligned}$$

so $\mu_0(\boldsymbol{x})$ is indeed the argmin. ☐

The following lemma establishes basic properties of the "Gram" (or Hessian, depending on the perspective) matrix generated by the partitioning-based M-estimator.

Lemma C.11 (Gram matrix)**.** *Suppose Assumptions B.1 to B.6 hold. Then*

$$h^d \lesssim \inf_{\boldsymbol{q}} \lambda_{\min}(\mathbf{Q}_{0,\boldsymbol{q}}) \leq \sup_{\boldsymbol{q}} \lambda_{\max}(\mathbf{Q}_{0,\boldsymbol{q}}) \lesssim h^d, \tag{C.2}$$

$$\sup_{\boldsymbol{q}} \big\|\mathbf{Q}_{0,\boldsymbol{q}}^{-1}\big\|_\infty \lesssim h^{-d}. \tag{C.3}$$

If, in addition, $\frac{\log(1/h)}{nh^d} = o(1)$, *then uniformly over $\boldsymbol{q} \in \mathcal{Q}$*

$$\sup_{\boldsymbol{q}}\big\{\big\|\bar{\mathbf{Q}}_{\boldsymbol{q}} - \mathbf{Q}_{0,\boldsymbol{q}}\big\|_\infty \vee \big\|\bar{\mathbf{Q}}_{\boldsymbol{q}} - \mathbf{Q}_{0,\boldsymbol{q}}\big\|\big\} \lesssim_{\mathbb{P}} h^d\Big(\frac{\log(1/h)}{nh^d}\Big)^{1/2}, \tag{C.4}$$

$$h^d \lesssim \inf_{\boldsymbol{q}} \lambda_{\min}(\bar{\mathbf{Q}}_{\boldsymbol{q}}) \leq \sup_{\boldsymbol{q}} \lambda_{\max}(\bar{\mathbf{Q}}_{\boldsymbol{q}}) \lesssim h^d \quad \textit{w. p. a. 1}, \tag{C.5}$$

$$\sup_{\boldsymbol{q}} \big\|\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\big\|_\infty \lesssim h^{-d} \quad \textit{w. p. a. 1}, \tag{C.6}$$

$$\sup_{\boldsymbol{q}}\big\{\big\|\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1} - \mathbf{Q}_{0,\boldsymbol{q}}^{-1}\big\|_\infty \vee \big\|\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1} - \mathbf{Q}_{0,\boldsymbol{q}}^{-1}\big\|\big\} \lesssim_{\mathbb{P}} h^{-d}\Big(\frac{\log(1/h)}{nh^d}\Big)^{1/2}. \tag{C.7}$$

For some positive integer $L_n \lesssim 1/h$, the rows and columns of $\bar{\mathbf{Q}}_{\boldsymbol{q}}$ and its inverse can be numbered by multi-indices $(\boldsymbol{\alpha},a) = (\alpha_1,\dots,\alpha_d,a)$ and $(\boldsymbol{\beta},b) = (\beta_1,\dots,\beta_d,b)$, where

$$\boldsymbol{\alpha},\boldsymbol{\beta} \in \{1,\dots,L_n\}^d, \quad a \in \{1,\dots,T_{n,\boldsymbol{\alpha}}\}, b \in \{1,\dots,T_{n,\boldsymbol{\beta}}\}, \quad T_{n,\boldsymbol{\alpha}},T_{n,\boldsymbol{\beta}} \lesssim 1,$$

in the following way. First, $\bar{\mathbf{Q}}_{\boldsymbol{q}}$ has a multi-banded structure:

$$[\bar{\mathbf{Q}}_{\boldsymbol{q}}]_{(\boldsymbol{\alpha},a),(\boldsymbol{\beta},b)} = 0 \quad \textit{if } \|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_\infty > C, \tag{C.8}$$

for some constant $C > 0$ (not depending on n). Second, with probability approaching one

$$\sup_{\boldsymbol{q}} \big|[\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}]_{(\boldsymbol{\alpha},a),(\boldsymbol{\beta},b)}\big| \lesssim h^{-d} \varrho^{\|\boldsymbol{\alpha}-\boldsymbol{\beta}\|_\infty} \tag{C.9}$$

for some constant $\varrho \in (0,1)$ (not depending on n).

The same results hold with $\mathbf{Q}_{0,\boldsymbol{q}}$ replaced by $\boldsymbol{\Sigma}_{0,\boldsymbol{q}}$ and $\bar{\mathbf{Q}}_{\boldsymbol{q}}$ replaced by $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q}}$.

Finally, the same results hold with $\mathbf{Q}_{0,\boldsymbol{q}}$ replaced by $\mathbb{E}[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}]$ and $\bar{\mathbf{Q}}_{\boldsymbol{q}}$ replaced by $\mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}]$.

Proof. The last claim of the lemma, corresponding to the case

$$\Psi_1(\boldsymbol{x}_i, \eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q})); \boldsymbol{q})\eta^{(1)}(\mu_0(\boldsymbol{x}_i, \boldsymbol{q}))^2 \equiv 1,$$

is Lemma SA-2.1 in [11]. The properties (C.8) and (C.9) are not explicitly stated but follow from the proof.

In the general case, by Assumption B.4(iv) $\Psi_1(\boldsymbol{x}_i, \eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q})); \boldsymbol{q})\eta^{(1)}(\mu_0(\boldsymbol{x}_i, \boldsymbol{q}))^2$ is bounded and bounded away from zero uniformly over i, n and $\boldsymbol{q}$, so (C.2) and (C.5) follow from the previous case. The additional $\Psi_1(\boldsymbol{x}_i, \eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q})); \boldsymbol{q})\eta^{(1)}(\mu_0(\boldsymbol{x}_i))^2$ term does not influence the multi-banded structure of the matrices, so (C.3), (C.6), (C.8), (C.9) remain true by the same argument as in the previous case. The inequalities

$$\begin{aligned}
\big\|\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1} - \mathbf{Q}_{0,\boldsymbol{q}}^{-1}\big\|_\infty &\le \big\|\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\big\|_\infty \cdot \big\|\bar{\mathbf{Q}}_{\boldsymbol{q}} - \mathbf{Q}_{0,\boldsymbol{q}}\big\|_\infty \cdot \big\|\mathbf{Q}_{0,\boldsymbol{q}}^{-1}\big\|_\infty, \\
\big\|\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1} - \mathbf{Q}_{0,\boldsymbol{q}}^{-1}\big\| &\le \big\|\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\big\| \cdot \big\|\bar{\mathbf{Q}}_{\boldsymbol{q}} - \mathbf{Q}_{0,\boldsymbol{q}}\big\| \cdot \big\|\mathbf{Q}_{0,\boldsymbol{q}}^{-1}\big\|
\end{aligned}$$

show that (C.7) follows from norm bounds (C.2), (C.3), (C.5), (C.6) and concentration (C.4).

Now we prove Eq. (C.4).

Define the class of functions

$$\mathcal{G} := \big\{\boldsymbol{x} \mapsto p_k(\boldsymbol{x})p_l(\boldsymbol{x})\Psi_1(\boldsymbol{x}, \eta(\mu_0(\boldsymbol{x}, \boldsymbol{q})); \boldsymbol{q})\eta^{(1)}(\mu_0(\boldsymbol{x}, \boldsymbol{q}))^2 \colon 1 \le k, l \le K, \boldsymbol{q} \in \mathcal{Q}\big\}.$$

We will now prove that the class $\mathcal{G}$ with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim h^{-2d}$ and $V \lesssim 1$. By Assumption B.5, the class

$$\{\boldsymbol{x} \mapsto \Psi_1(\boldsymbol{x}, \eta(\mu_0(\boldsymbol{x}, \boldsymbol{q})); \boldsymbol{q}) \colon \boldsymbol{q} \in \mathcal{Q}\}$$

with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim 1$. Since it is also true of the class $\big\{\eta^{(1)}\big(\mu_0(\boldsymbol{x}, \boldsymbol{q})\big)^2 \colon \boldsymbol{q} \in \mathcal{Q}\big\}$ because $\eta^{(1)}\big(\mu_0(\boldsymbol{x}, \boldsymbol{q})\big)^2$ is Lipschitz is $\boldsymbol{q}$, by Lemma C.4 the class

$$\big\{\boldsymbol{x} \mapsto \Psi_1(\boldsymbol{x}, \eta(\mu_0(\boldsymbol{x}, \boldsymbol{q})); \boldsymbol{q})\eta^{(1)}(\mu_0(\boldsymbol{x}, \boldsymbol{q}))^2 \colon \boldsymbol{q} \in \mathcal{Q}\big\}$$

with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim 1$. The class $\{\boldsymbol{x} \mapsto p_k(\boldsymbol{x})p_l(\boldsymbol{x}) \colon 1 \le k, l \le K\}$ with a large enough constant envelope just contains K^2 functions, so it also satisfies the uniform entropy bound (A.3) with $A \lesssim h^{-2d}$ and $V = 1$, where we used $K \asymp h^{-d}$. By Lemma C.4, combining these facts proves the claim about the complexity of $\mathcal{G}$.

Moreover, class $\mathcal{G}$ satisfies the following variance bound:

$$\sup_{g\in\mathcal{G}} \mathbb{E}\big[g(\boldsymbol{x}_i)^2\big] \lesssim h^d,$$

which follows from the fact that the class is bounded by a large enough constant and the Lebesgue measure of the support of $p_k(\boldsymbol{x})p_l(\boldsymbol{x})$ shrinks (uniformly over k,l) at the rate h^d.

Applying Lemma C.6, we see that

$$\begin{aligned}
\sup_{g\in\mathcal{G}}\bigg|\frac{1}{n}\sum_{i=1}^n g(\boldsymbol{x}_i) - \mathbb{E}[g(\boldsymbol{x}_i)]\bigg| &\lesssim_{\mathbb{P}} h^d\bigg(\frac{\log(1/h)}{nh^d}\bigg)^{1/2} + \frac{\log(1/h)}{n} \\
&\lesssim h^d\bigg(\frac{\log(1/h)}{nh^d}\bigg)^{1/2} \qquad \text{since } \tfrac{\log(1/h)}{nh^d} = o(1).
\end{aligned}$$

So we have shown $\max_{k,l}\big|(\bar{\mathbf{Q}}_{\boldsymbol{q}} - \mathbf{Q}_{0,\boldsymbol{q}})_{k,l}\big| \lesssim_{\mathbb{P}} h^d\Big(\frac{\log(1/h)}{nh^d}\Big)^{1/2}$. Since each row and column of $\bar{\mathbf{Q}}_{\boldsymbol{q}} - \mathbf{Q}_{0,\boldsymbol{q}}$ has a bounded number of nonzero entries, this implies

$$\|\bar{\mathbf{Q}}_{\boldsymbol{q}} - \mathbf{Q}_{0,\boldsymbol{q}}\|_\infty \asymp \|\bar{\mathbf{Q}}_{\boldsymbol{q}} - \mathbf{Q}_{0,\boldsymbol{q}}\|_1 \lesssim_{\mathbb{P}} h^d\bigg(\frac{\log(1/h)}{nh^d}\bigg)^{1/2}.$$

To conclude the proof of Eq. (C.4), it is left to use the inequality

$$\|\bar{\mathbf{Q}}_{\boldsymbol{q}} - \mathbf{Q}_{0,\boldsymbol{q}}\| \le \sqrt{\|\bar{\mathbf{Q}}_{\boldsymbol{q}} - \mathbf{Q}_{0,\boldsymbol{q}}\|_1\|\bar{\mathbf{Q}}_{\boldsymbol{q}} - \mathbf{Q}_{0,\boldsymbol{q}}\|_\infty}.$$

The claim about $\boldsymbol{\Sigma}_{0,\boldsymbol{q}}$ and $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q}}$ is proven analogously, using that $\sigma^2_{\boldsymbol{q}}(\boldsymbol{x})$ is bounded and bounded away from zero and Lipschitz in $\boldsymbol{q}$ by Assumption B.4(ii).

Lemma C.11 is proven. □

The following Lemmas C.12 and C.14 are needed for the proof of the Bahadur representation (Theorem E.1), Corollaries E.16 and E.17 and a version of the consistency result (Lemma D.6).

Lemma C.12 (Uniform convergence: variance)**.** *Suppose Assumptions B.1 to B.6 hold. If*

$$\frac{\big[\log(1/h)\big]^{\nu/(\nu-2)}}{nh^{\nu d/(\nu-2)}} = o(1), \text{ or} \tag{C.10}$$

$$\overline{\psi}(\boldsymbol{x}_i, y_i) \text{ is } \sigma^2\text{-sub-Gaussian conditionally on } \boldsymbol{x}_i \text{ and } \frac{\log n\log(1/h)}{nh^d} = o(1), \tag{C.11}$$

then

$$\sup_{\boldsymbol{q}\in\mathcal{Q},\boldsymbol{x}\in\mathcal{X}}\big|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}}\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\mathbb{E}_n\big[\boldsymbol{p}(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}));\boldsymbol{q})\big]\big| \lesssim_{\mathbb{P}} h^{-|\boldsymbol{v}|}\bigg(\frac{\log(1/h)}{nh^d}\bigg)^{1/2};$$

and the same inequality is true with $\bar{\mathbf{Q}}_{\boldsymbol{q}}$ replaced by $\mathbf{Q}_{0,\boldsymbol{q}}$.

Proof. By Assumption B.2, $\sup_{\boldsymbol{x}\in\mathcal{X}}\|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\| \lesssim h^{-|\boldsymbol{v}|}$; by Lemma C.11, $\|\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\|_\infty + \|\mathbf{Q}_{0,\boldsymbol{q}}^{-1}\|_\infty \lesssim_{\mathbb{P}} h^{-d}$. Therefore, it is enough to show

$$\sup_{\boldsymbol{q}\in\mathcal{Q}}\big\|\mathbb{E}_n\big[\boldsymbol{p}(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}));\boldsymbol{q})\big]\big\|_\infty \lesssim_{\mathbb{P}} \sqrt{\frac{h^d\log(1/h)}{n}}. \tag{C.12}$$

Define the function class

$$\mathcal{G} := \big\{(\boldsymbol{x}, y) \mapsto p_l(\boldsymbol{x})\eta^{(1)}(\mu_0(\boldsymbol{x}, \boldsymbol{q}))\psi(y, \eta(\mu_0(\boldsymbol{x}, \boldsymbol{q})); \boldsymbol{q}) : 1 \le l \le K, \boldsymbol{q} \in \mathcal{Q}\big\}.$$

We will now control the complexity of $\mathcal{G}$. Introduce some more classes of functions:

$$\begin{aligned}
\mathcal{W}_1 &:= \{(\boldsymbol{x}, y) \mapsto p_l(\boldsymbol{x}) : 1 \le l \le K\},\\
\mathcal{W}_2 &:= \{(\boldsymbol{x}, y) \mapsto \eta^{(1)}(\mu_0(\boldsymbol{x}, \boldsymbol{q})) : \boldsymbol{q} \in \mathcal{Q}\},\\
\mathcal{W}_3 &:= \{(\boldsymbol{x}, y) \mapsto \psi(y, \eta(\mu_0(\boldsymbol{x}, \boldsymbol{q})); \boldsymbol{q}) : \boldsymbol{q} \in \mathcal{Q}\}.
\end{aligned}$$

$\mathcal{W}_1$ with a large enough constant envelope contains K fixed measurable functions, so it satisfies the uniform entropy bound (A.3) with $A \lesssim h^{-d}$ and $V = 1$. $\mathcal{W}_2$ with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A, V \lesssim 1$ because $\mu_0(\boldsymbol{x}, \boldsymbol{q})$ is bounded uniformly over $\boldsymbol{x}, \boldsymbol{q}$ and Lipschitz in $\boldsymbol{q}$, $\eta^{(1)}$ on a fixed bounded interval is Lipschitz. $\mathcal{W}_3$ with envelope $\overline{\psi}(\boldsymbol{x}, y)$ satisfies the uniform entropy bound (A.3) with $A, V \lesssim 1$ by Assumption B.5. By Lemma C.4, $\mathcal{G}$ with envelope $\overline{\psi}(\boldsymbol{x}, y)$ multiplied by a large enough constant satisfies the uniform entropy bound (A.3) with $A \lesssim h^{-d}$ and $V \lesssim 1$.

Moreover, class $\mathcal{G}$ satisfies the following variance bound:

$$\sup_{g \in \mathcal{G}} \mathbb{E}\big[g(\boldsymbol{x}_i, y_i)^2\big] \lesssim h^d.$$

Indeed, for a fixed $i \in \{1, \dots, n\}$

$$\sup_{g \in \mathcal{G}} \mathbb{E}\big[g(\boldsymbol{x}_i, y_i)^2\big] \lesssim \sup_l \mathbb{E}\Big[p_l(\boldsymbol{x}_i)^2 \mathbb{E}\Big[\overline{\psi}(\boldsymbol{x}_i, y_i)^2 \,\Big|\, \boldsymbol{x}_i\Big]\Big] \lesssim \sup_l \mathbb{E}\big[p_l(\boldsymbol{x}_i)^2\big] \lesssim h^d.$$

Finally, under (C.10)

$$\begin{aligned}
\mathbb{E}\Big[\max_{1 \le i \le n} |\overline{\psi}(\boldsymbol{x}_i, y_i)|^2\Big]^{1/2} &\le \mathbb{E}\Big[\max_{1 \le i \le n} |\overline{\psi}(\boldsymbol{x}_i, y_i)|^\nu\Big]^{1/\nu} \le \mathbb{E}\Bigg[\sum_{i=1}^n |\overline{\psi}(\boldsymbol{x}_i, y_i)|^\nu\Bigg]^{1/\nu}\\
&= \Bigg(\sum_{i=1}^n \mathbb{E}\big[\overline{\psi}(\boldsymbol{x}_i, y_i)^\nu\big]\Bigg)^{1/\nu} \lesssim \Bigg(\sum_{i=1}^n 1\Bigg)^{1/\nu} = n^{1/\nu},
\end{aligned}$$

and under (C.11)

$$\mathbb{E}\Big[\max_{1 \le i \le n} |\overline{\psi}(\boldsymbol{x}_i, y_i)|^2\Big]^{1/2} \lesssim \sqrt{\log n}$$

by Lemma C.1.

Applying Lemma C.6, we obtain (C.12) since

$$\begin{aligned}
\frac{n^{1/\nu} \log(1/h)}{n} &= \sqrt{\frac{h^d \log(1/h)}{n}} \cdot \sqrt{\frac{\log(1/h)}{n^{1-2/\nu} h^d}} = \sqrt{\frac{h^d \log(1/h)}{n}} \cdot o(1), \text{ and}\\
\frac{\sqrt{\log n}\log(1/h)}{n} &= \sqrt{\frac{h^d \log(1/h)}{n}} \cdot \sqrt{\frac{\log n \log(1/h)}{n h^d}} = \sqrt{\frac{h^d \log(1/h)}{n}} \cdot o(1).
\end{aligned}$$

Lemma C.12 is proven. □

Remark C.13. Using

$$\frac{n^{1/\nu} \log(1/h)}{n} = o(h^d) \Leftrightarrow \frac{[\log(1/h)]^{\nu/(\nu-1)}}{n h^{\nu d/(\nu-1)}} = o(1) \tag{C.13}$$

instead of

$$\frac{n^{1/\nu}\log(1/h)}{n} = \sqrt{\frac{h^d\log(1/h)}{n}}\cdot o(1)$$

in the argument leads to the bound

$$\sup_{\boldsymbol{q}\in\mathcal{Q}}\big\|\mathbb{E}_n\big[\boldsymbol{p}(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}));\boldsymbol{q})\big]\big\|_\infty = o_{\mathbb{P}}(h^d)$$

under just the condition (C.13) which is slightly weaker than (C.10).

The following lemma gives control on the projection approximation error.

Lemma C.14 (Projection of approximation error)**.** *Suppose Assumptions B.1 to B.6 hold. If*

$$\frac{\big[\log(1/h)\big]^{\nu/(\nu-2)}}{nh^{\nu d/(\nu-2)}} = o(1),\ \textit{or} \tag{C.14}$$

$$\overline{\psi}(\boldsymbol{x}_i,y_i)\ \textit{is}\ \sigma^2\textit{-sub-Gaussian conditionally on}\ \boldsymbol{x}_i\ \textit{and}\ \frac{\log n\log(1/h)}{nh^d} = o(1),$$

then

$$\begin{aligned}\sup_{\boldsymbol{q}\in\mathcal{Q},\boldsymbol{x}\in\mathcal{X}}&\big|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}}\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\mathbb{E}_n\big[\boldsymbol{p}(\boldsymbol{x}_i)\big\{\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}));\boldsymbol{q})\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))\\ &\quad-\psi(y_i,\eta(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}))\big\}\big]\big|\\ &\lesssim_{\mathbb{P}} h^{m-|\boldsymbol{v}|}+h^{(\alpha\wedge(1/2))m-|\boldsymbol{v}|}\left(\frac{\log(1/h)}{nh^d}\right)^{1/2}+\frac{\log(1/h)}{nh^{|\boldsymbol{v}|+d}}.\end{aligned}$$

Proof. Denote

$$\begin{aligned}A_{1,\boldsymbol{q}}(\boldsymbol{x}_i,y_i) &:= \psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}));\boldsymbol{q})\big\{\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))-\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}))\big\},\\ A_{2,\boldsymbol{q}}(\boldsymbol{x}_i,y_i) &:= \big\{\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}));\boldsymbol{q})-\psi(y_i,\eta(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\big\}\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q})).\end{aligned}$$

By Assumption B.2, $\sup_{\boldsymbol{x}\in\mathcal{X}}\|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\|\lesssim h^{-|\boldsymbol{v}|}$; by Lemma C.11, $\|\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\|_\infty\lesssim_{\mathbb{P}} h^{-d}$. Therefore, it is enough to show

$$\begin{aligned}\sup_{\boldsymbol{q}\in\mathcal{Q}}\Big\|\mathbb{E}_n\Big[&\boldsymbol{p}(\boldsymbol{x}_i)\Big\{\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}));\boldsymbol{q})\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))\\ &-\psi(y_i,\eta(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}))\Big\}\Big]\Big\|_\infty\\ &\lesssim_{\mathbb{P}} h^{d+m}+h^{\frac{d}{2}+(\alpha\wedge(1/2))m}\left(\frac{\log(1/h)}{n}\right)^{1/2}+\frac{\log(1/h)}{n}.\end{aligned}$$

We will do this by showing the three bounds

$$\sup_{\boldsymbol{q}\in\mathcal{Q}}\big\|\mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)A_{1,\boldsymbol{q}}(\boldsymbol{x}_i,y_i)]\big\|_\infty\lesssim_{\mathbb{P}} h^{\frac{d}{2}+m}\left(\frac{\log(1/h)}{n}\right)^{1/2}, \tag{C.15}$$

$$\sup_{\boldsymbol{q}\in\mathcal{Q}}\big\|\mathbb{E}_n\big[\mathbb{E}\big[\boldsymbol{p}(\boldsymbol{x}_i)A_{2,\boldsymbol{q}}(\boldsymbol{x}_i,y_i)\,\big|\,\boldsymbol{x}_i\big]\big]\big\|_\infty\lesssim_{\mathbb{P}} h^{d+m}, \tag{C.16}$$

$$\sup_{\boldsymbol{q}\in\mathcal{Q}}\big\|\mathbb{E}_n\big[\boldsymbol{p}(\boldsymbol{x}_i)(A_{2,\boldsymbol{q}}(\boldsymbol{x}_i,y_i)-\mathbb{E}[A_{2,\boldsymbol{q}}(\boldsymbol{x}_i,y_i)\,|\,\boldsymbol{x}_i])\big]\big\|_\infty\lesssim_{\mathbb{P}} h^{\frac{d}{2}+(\alpha\wedge(1/2))m}\sqrt{\frac{\log(1/h)}{n}}+\frac{\log(1/h)}{n}. \tag{C.17}$$

To show (C.15), consider the class of functions

$$\{(\boldsymbol{x}, y) \mapsto p_l(\boldsymbol{x}) A_{1,\boldsymbol{q}}(\boldsymbol{x}, y) : 1 \le l \le K, \boldsymbol{q} \in \mathcal{Q}\}.$$

Note that $\sup_{\boldsymbol{q},\boldsymbol{x}} |\eta^{(1)}(\mu_0(\boldsymbol{x}, \boldsymbol{q})) - \eta^{(1)}(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T} \boldsymbol{\beta}_0(\boldsymbol{q}))| \lesssim h^m$ by Assumption B.6. (C.15) follows by the same concentration argument as in Lemma C.12.

To show (C.16), note that

$$\begin{aligned}
&\mathbb{E}\Big[\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q})); \boldsymbol{q}) - \psi(y_i, \eta(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T} \boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q}) \,\Big|\, \boldsymbol{x}_i\Big] \\
&\quad = -\Psi(\boldsymbol{x}_i, \eta(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T} \boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q}) = \\
&\quad = \Psi_1(\boldsymbol{x}_i, \zeta; \boldsymbol{q}) \eta^{(1)}(\tilde{\zeta}) \{\mu_0(\boldsymbol{x}_i, \boldsymbol{q}) - \boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T} \boldsymbol{\beta}_0(\boldsymbol{q})\},
\end{aligned}$$

where ζ is between $\eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q}))$ and $\eta(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T} \boldsymbol{\beta}_0(\boldsymbol{q}))$, $\tilde{\zeta}$ is between $\mu_0(\boldsymbol{x}_i, \boldsymbol{q})$ and $\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T} \boldsymbol{\beta}_0(\boldsymbol{q})$. By Assumption B.4(iv) and B.6, it follows that a. s.

$$\sup_{\boldsymbol{q} \in \mathcal{Q}} |\mathbb{E}[\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q})); \boldsymbol{q}) - \psi(y_i, \eta(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T} \boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q}) \,|\, \boldsymbol{x}_i]| \lesssim h^m.$$

Since $\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T} \boldsymbol{\beta}_0(\boldsymbol{q}))$ is bounded, (C.16) follows by applying Lemma C.6 to the class $\{\boldsymbol{x} \mapsto |p_l(\boldsymbol{x})|, 1 \le l \le K\}$.

It is left to show (C.17).

Consider the class of functions

$$\mathcal{G} := \{(\boldsymbol{x}, y) \mapsto p_l(\boldsymbol{x}) A_{2,\boldsymbol{q}}(\boldsymbol{x}, y) : 1 \le l \le K, \boldsymbol{q} \in \mathcal{Q}\}.$$

We will now control the complexity of $\mathcal{G}$. Introduce some more classes of functions:

$$\begin{aligned}
\mathcal{W}_{1,l} &:= \{(\boldsymbol{x}, y) \mapsto p_l(\boldsymbol{x})[\psi(y, \eta(\mu_0(\boldsymbol{x}, \boldsymbol{q})); \boldsymbol{q}) - \psi(y, \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T} \boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q})] : \boldsymbol{q} \in \mathcal{Q}\}, \\
\mathcal{W}_1 &:= \bigcup_{l=1}^K \mathcal{W}_{1,l}, \\
\mathcal{W}_{2,l} &:= \{(\boldsymbol{x}, y) \mapsto \eta^{(1)}(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T} \boldsymbol{\beta}_0(\boldsymbol{q})) \mathbb{1}\{\boldsymbol{x} \in \operatorname{supp} p_l\} : \boldsymbol{q} \in \mathcal{Q}\}, \\
\mathcal{W}_2 &:= \bigcup_{l=1}^K \mathcal{W}_{2,l} = \{(\boldsymbol{x}, y) \mapsto \eta^{(1)}(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T} \boldsymbol{\beta}_0(\boldsymbol{q})) \mathbb{1}\{\boldsymbol{x} \in \operatorname{supp} p_l\} : 1 \le l \le K, \boldsymbol{q} \in \mathcal{Q}\}.
\end{aligned}$$

By Assumption B.5, for l fixed $\mathcal{W}_{1,l}$ with a large enough constant envelope (not depending on l) satisfies the uniform entropy bound (A.3) with $A, V \lesssim 1$ (not depending on l). This immediately implies that $\mathcal{W}_1$ with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim h^{-d}$ and $V \lesssim 1$.

For l fixed, $\mathcal{W}_{2,l}$ is a product of a (bounded) subclass of $\{\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T} \boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathcal{B}_l\}$, where $\mathcal{B}_l$ is a vector space of dimension $O(1)$ (not depending on l), and a fixed function. By Lemma 2.6.15 in [41], $\{\boldsymbol{p}(\boldsymbol{x})^\mathsf{T} \boldsymbol{\beta} : \boldsymbol{\beta} \in \mathcal{B}_l\}$ is VC with a bounded index. Therefore, since $\eta^{(1)}$ on a bounded interval is Lipschitz, $\mathcal{W}_{2,l}$ with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A, V \lesssim 1$. This immediately implies that $\mathcal{W}_2$ with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim h^{-d}$ and $V \lesssim 1$.

By Lemma C.4, it follows from the above that $\mathcal{G}$ with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim h^{-d}$ and $V \lesssim 1$.

Next, we will show that class $\mathcal{G}$ satisfies the following variance bound:

$$\sup_{g\in\mathcal{G}} \mathbb{E}_n[\mathbb{V}[g(\boldsymbol{x}_i, y_i) \,|\, \boldsymbol{x}_i]] \lesssim h^{d+(2\alpha\wedge 1)m} \quad \text{w. p. a. 1.} \tag{C.18}$$

We will prove (C.18) under the assumption that $\mathfrak{M}$ is Lebesgue measure, so (B.2) holds; the argument under (B.3) is similar (and leads to an even stronger variance bound), so it is omitted. For y outside the closed segment between $\eta(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}))$ and $\eta(\mu_0(\boldsymbol{x}, \boldsymbol{q}))$,

$$\begin{aligned}
&|\psi(y, \eta(\mu_0(\boldsymbol{x}, \boldsymbol{q})); \boldsymbol{q}) - \psi(y, \eta(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q})| \\
&\quad \lesssim (\overline{\psi}(\boldsymbol{x}, y) + 1) \cdot |\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}) - \mu_0(\boldsymbol{x}, \boldsymbol{q})|^{\alpha} \\
&\quad \lesssim (\overline{\psi}(\boldsymbol{x}, y) + 1) h^{\alpha m},
\end{aligned} \tag{C.19}$$

where in (C.19) we used Assumption B.6. For y between $\eta(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}))$ and $\eta(\mu_0(\boldsymbol{x}, \boldsymbol{q}))$ inclusive,

$$\begin{aligned}
&|\psi(y, \eta(\mu_0(\boldsymbol{x}, \boldsymbol{q})); \boldsymbol{q}) - \psi(y, \eta(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q})| \lesssim \overline{\psi}(\boldsymbol{x}, y) \cdot |\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}) - \mu_0(\boldsymbol{x}, \boldsymbol{q})| + 1 \\
&\quad \lesssim \overline{\psi}(\boldsymbol{x}, y) h^{m} + 1,
\end{aligned} \tag{C.20}$$

where in (C.20) we again used Assumption B.6.

In the chain below, to avoid cluttering notation we will use $[a, b]$ to denote the closed segment between a and b regardless of their ordering (a more standard notation is $[a \wedge b, a \vee b]$). Using that $\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}))$ is also bounded uniformly over $\boldsymbol{x} \in \mathcal{X}$, we have a. s.

$$\begin{aligned}
&\mathbb{E}\big[A_{2,\boldsymbol{q}}(\boldsymbol{x}_i, y_i)^2 \,\big|\, \boldsymbol{x}_i\big] \\
&\quad = \mathbb{E}\big[A_{2,\boldsymbol{q}}(\boldsymbol{x}_i, y_i)^2 \mathbb{1}\big\{y_i \notin \big[\eta(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q})), \eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q}))\big]\big\} \,\big|\, \boldsymbol{x}_i\big] \\
&\quad\quad + \mathbb{E}\big[A_{2,\boldsymbol{q}}(\boldsymbol{x}_i, y_i)^2 \mathbb{1}\big\{y_i \in \big[\eta(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q})), \eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q}))\big]\big\} \,\big|\, \boldsymbol{x}_i\big] \\
&\quad \lesssim h^{2\alpha m} \mathbb{E}\big[(\overline{\psi}(\boldsymbol{x}_i, y_i)^2 + 1) \mathbb{1}\big\{y_i \notin \big[\eta(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q})), \eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q}))\big]\big\} \,\big|\, \boldsymbol{x}_i\big] \\
&\quad\quad + h^{2m} \mathbb{E}\big[\overline{\psi}(\boldsymbol{x}_i, y_i)^2 \mathbb{1}\big\{y_i \in \big[\eta(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q})), \eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q}))\big]\big\} \,\big|\, \boldsymbol{x}_i\big] \\
&\quad\quad + \mathbb{P}\big\{y_i \in \big[\eta(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q})), \eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q}))\big] \,\big|\, \boldsymbol{x}_i\big\} \\
&\quad \le h^{2\alpha m} \mathbb{E}\Big[(\overline{\psi}(\boldsymbol{x}_i, y_i)^2 + 1) \,\Big|\, \boldsymbol{x}_i\Big] + h^{2m} \mathbb{E}\Big[\overline{\psi}(\boldsymbol{x}_i, y_i)^2 \,\Big|\, \boldsymbol{x}_i\Big] \\
&\quad\quad + \mathbb{P}\Big\{y_i \in \big[\eta(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q})), \eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q}))\big] \,\Big|\, \boldsymbol{x}_i\Big\} \\
&\quad \lesssim h^{2\alpha m} + h^{2m} + |\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}) - \mu_0(\boldsymbol{x}_i, \boldsymbol{q})| \qquad \text{(C.21)} \\
&\quad \lesssim h^{2\alpha m} + h^{2m} + h^{m} \qquad \text{(C.22)} \\
&\quad \lesssim h^{(2\alpha\wedge 1)m},
\end{aligned}$$

where in (C.21) we used that by Assumption B.3(iii) the conditional density of $y_i|\boldsymbol{x}_i$ is bounded and Assumption B.4(ii), in (C.22) we used Assumption B.6.

Therefore, uniformly over l and $\boldsymbol{q}$

$$\begin{aligned}
&\mathbb{E}_n[\mathbb{V}[g(\boldsymbol{x}_i, y_i) \,|\, \boldsymbol{x}_i]] \le \mathbb{E}_n\big[\mathbb{E}[g(\boldsymbol{x}_i, y_i)^2 \,|\, \boldsymbol{x}_i]\big] = \mathbb{E}_n\big[p_l(\boldsymbol{x}_i)^2 \mathbb{E}[A_{2,\boldsymbol{q}}(\boldsymbol{x}_i, y_i)^2 \,|\, \boldsymbol{x}_i]\big] \\
&\quad \lesssim h^{(2\alpha\wedge 1)m} \mathbb{E}_n[p_l(\boldsymbol{x}_i)^2] \\
&\quad \le h^{(2\alpha\wedge 1)m} \|\mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}]\| \\
&\quad \lesssim h^{d+(2\alpha\wedge 1)m} \quad \text{w. p. a. 1,}
\end{aligned} \tag{C.23}$$

where in (C.23) we used Lemma C.11. We have proven (C.18).

Applying Lemma C.6 conditionally on $\{\boldsymbol{x}_i\}_{i=1}^n$, on an event with probability approaching one, we get (C.17), and the proof of Lemma C.14 is finished. □

Remark C.15. Inspecting the argument, one can note that instead of (C.14) the condition

$$\frac{\log n}{nh^d} = o(1)$$

is enough for Eqs. (C.16) and (C.17).

D Consistency

We first study the convex case, and then move on to the non-convex case.

D.1 Convex case

The following lemma gives our most general result for a convex objective function. This is Lemma 1 in the main paper.

Lemma D.1 (Consistency, convex case)**.** *Suppose Assumptions B.1 to B.6 hold, $\rho(y, \eta(\theta); \boldsymbol{q})$ is convex with respect to θ with left or right derivative $\psi(y, \eta(\theta); \boldsymbol{q})\eta^{(1)}(\theta)$, $\mathcal{B} = \mathbb{R}^K$ in (A.2), and $m > d/2$. Furthermore, assume that one of the following two conditions holds:*

$$\frac{\bigl[\log(1/h)\bigr]^{\frac{\nu}{\nu-1}}}{nh^{\frac{2\nu}{\nu-1}d}} = o(1), \ \text{or} \tag{D.1}$$

$$\overline{\psi}(\boldsymbol{x}_i, y_i) \text{ is } \sigma^2\text{-sub-Gaussian conditionally on } \boldsymbol{x}_i \text{ and } \frac{\sqrt{\log n}\log(1/h)}{nh^{2d}} = o(1). \tag{D.2}$$

Then

$$\sup_{\boldsymbol{q}\in\mathcal{Q}} \bigl\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})\bigr\| = o_{\mathbb{P}}(1), \tag{D.3}$$

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \sup_{\boldsymbol{q}\in\mathcal{Q}} \bigl|\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x}, \boldsymbol{q}) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x}, \boldsymbol{q})\bigr| = o_{\mathbb{P}}(h^{-|\boldsymbol{v}|}), \tag{D.4}$$

$$\sup_{\boldsymbol{q}\in\mathcal{Q}} \Bigl(\int_{\mathcal{X}} \bigl(\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x}, \boldsymbol{q}) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x}, \boldsymbol{q})\bigr)^2 f_X(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\Bigr)^{1/2} = o_{\mathbb{P}}\bigl(h^{d/2-|\boldsymbol{v}|}\bigr). \tag{D.5}$$

Proof. First, note that (D.4) follows from (D.3) since uniformly over $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{q} \in \mathcal{Q}$

$$\begin{aligned}
&\bigl|\widehat{\boldsymbol{\beta}}(\boldsymbol{q})^{\mathsf{T}} \boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x}) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x}, \boldsymbol{q})\bigr| \\
&\quad\le \bigl|\widehat{\boldsymbol{\beta}}(\boldsymbol{q})^{\mathsf{T}} \boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x}) - \boldsymbol{\beta}_0(\boldsymbol{q})^{\mathsf{T}} \boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\bigr| + \bigl|\boldsymbol{\beta}_0(\boldsymbol{q})^{\mathsf{T}} \boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x}) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x}, \boldsymbol{q})\bigr| \\
&\quad\lesssim \bigl\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})\bigr\|_\infty \cdot \bigl\|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\bigr\|_1 + \bigl|\boldsymbol{\beta}_0(\boldsymbol{q})^{\mathsf{T}} \boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x}) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x}, \boldsymbol{q})\bigr| \\
&\quad\lesssim \bigl\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})\bigr\|_\infty \cdot h^{-|\boldsymbol{v}|} + \bigl|\boldsymbol{\beta}_0(\boldsymbol{q})^{\mathsf{T}} \boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x}) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x}, \boldsymbol{q})\bigr| && \text{by Assumption B.2} \\
&\quad\lesssim \bigl\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})\bigr\|_\infty \cdot h^{-|\boldsymbol{v}|} + h^{m-|\boldsymbol{v}|} && \text{by Assumption B.6,}
\end{aligned}$$

where we used that only a bounded number of elements in $\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})$ are nonzero. Similarly, (D.5) follows from (D.3) since

$$
\begin{aligned}
&\big\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q})^{\mathsf{T}}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x}) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})\big\|_{L_2(X)} \\
&\quad\le \big\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q})^{\mathsf{T}}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x}) - \boldsymbol{\beta}_0(\boldsymbol{q})^{\mathsf{T}}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\big\|_{L_2(X)} + \big\|\boldsymbol{\beta}_0(\boldsymbol{q})^{\mathsf{T}}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x}) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x},q)\big\|_{L_2(X)} \\
&\quad\le \big\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q})^{\mathsf{T}}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x}) - \boldsymbol{\beta}_0(\boldsymbol{q})^{\mathsf{T}}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\big\|_{L_2(X)} + \sup_{\boldsymbol{x}\in\mathcal{X}}\big|\boldsymbol{\beta}_0(\boldsymbol{q})^{\mathsf{T}}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x}) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})\big| \\
&\quad\lesssim \big\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q})^{\mathsf{T}}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x}) - \boldsymbol{\beta}_0(\boldsymbol{q})^{\mathsf{T}}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\big\|_{L_2(X)} + h^{m-|\boldsymbol{v}|} \qquad \text{by Assumption B.6} \\
&\quad= \Big(\big(\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})\big)^{\mathsf{T}}\mathbb{E}\big[\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x}_i)\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x}_i)^{\mathsf{T}}\big]\big(\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})\big)\Big)^{1/2} + h^{m-|\boldsymbol{v}|} \\
&\quad\le \lambda_{\max}\big(\mathbb{E}\big[\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x}_i)\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x}_i)^{\mathsf{T}}\big]\big)^{1/2}\big\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})\big\| + h^{m-|\boldsymbol{v}|} \\
&\quad\le h^{d/2-|\boldsymbol{v}|}\cdot\big\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})\big\| + h^{m-|\boldsymbol{v}|} = o_{\mathbb{P}}\big(h^{d/2-|\boldsymbol{v}|}\big) + h^{m-|\boldsymbol{v}|} = o_{\mathbb{P}}\big(h^{d/2-|\boldsymbol{v}|}\big).
\end{aligned}
$$

uniformly over $\boldsymbol{q} \in \mathcal{Q}$, where by $\|g(\boldsymbol{x})\|_{L_2(X)}$ we denote $\big(\int_{\mathcal{X}} g(\boldsymbol{x})^2 f_X(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\big)^{1/2}$ for simplicity. In the last equality we used $m > d/2$ again. We also used that the largest eigenvalue of $\mathbb{E}\big[\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x}_i)\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x}_i)^{\mathsf{T}}\big]$ is bounded from above by $h^{d-2|\boldsymbol{v}|}$ up to a multiplicative coefficient, which is proven by the same argument as for $\boldsymbol{v} = 0$ in Lemma C.11 in combination with Assumption B.2.

It is left to prove (D.3). Fix a sufficiently small $\gamma > 0$. Denote for $i \in \{1,\dots,n\}$ and $\boldsymbol{\alpha} \in \mathcal{S}^{K-1}$

$$
\delta_{\boldsymbol{q},i}(\boldsymbol{\alpha}) := \boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{p}(\boldsymbol{x}_i)\psi(y_i, \eta(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \gamma\boldsymbol{\alpha})); \boldsymbol{q})\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \gamma\boldsymbol{\alpha})).
$$

Since $\mathbb{E}_n\big[\psi(y_i, \eta(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\beta}); \boldsymbol{q})\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\beta})\boldsymbol{p}(\boldsymbol{x}_i)\big]$ is a subgradient of the convex (by Assumption B.4(iii)) objective function $\mathbb{E}_n[\rho(y_i, \eta(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\beta}); \boldsymbol{q})]$ of $\boldsymbol{\beta}$, the strategy is to show that

$$
\inf_{\boldsymbol{q},\boldsymbol{\alpha}} \mathbb{E}_n[\delta_{\boldsymbol{q},i}(\boldsymbol{\alpha})] > 0 \quad \text{with probability approaching 1,} \tag{D.6}
$$

which is enough to prove Lemma D.1 by convexity.

To implement this, we will show

$$
\inf_{\boldsymbol{q},\boldsymbol{\alpha}} \mathbb{E}_n[\mathbb{E}[\delta_{\boldsymbol{q},i}(\boldsymbol{\alpha}) \,|\, \boldsymbol{x}_i]] \gtrsim \inf_{\boldsymbol{\alpha}} \mathbb{E}_n[\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\alpha}] + o_{\mathbb{P}}(h^d) \quad \text{and} \tag{D.7}
$$

$$
\sup_{\boldsymbol{q},\boldsymbol{\alpha}}\big|\mathbb{E}_n[\delta_{\boldsymbol{q},i}(\boldsymbol{\alpha}) - \mathbb{E}[\delta_{\boldsymbol{q},i}(\boldsymbol{\alpha}) \,|\, \boldsymbol{x}_i]]\big| \lesssim_{\mathbb{P}} \begin{cases} \sqrt{\frac{\log(1/h)}{n}} + \frac{\log(1/h)}{n^{1-1/\nu}h^d} = o(h^d), \\ \sqrt{\frac{\log(1/h)}{n}} + \frac{\sqrt{\log n}\log(1/h)}{nh^d} = o(h^d) \end{cases} \tag{D.8}
$$

under (D.1) and (D.2) respectively (proof below), and conclude

$$
\begin{aligned}
&\inf_{\boldsymbol{q},\boldsymbol{\alpha}} \mathbb{E}_n[\delta_{\boldsymbol{q},i}(\boldsymbol{\alpha})] \ge \\
&\quad\ge \inf_{\boldsymbol{q},\boldsymbol{\alpha}} \mathbb{E}_n\big[\mathbb{E}[\delta_{\boldsymbol{q},i}(\boldsymbol{\alpha}) \,|\, \mathbf{X}_n]\big] - \sup_{\boldsymbol{q},\boldsymbol{\alpha}}\big|\mathbb{E}_n\big[\delta_{\boldsymbol{q},i}(\boldsymbol{\alpha}) - \mathbb{E}[\delta_{\boldsymbol{q},i}(\boldsymbol{\alpha}) \,|\, \mathbf{X}_n]\big]\big| \\
&\quad\gtrsim \inf_{\boldsymbol{\alpha}} \mathbb{E}_n[\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\alpha}] + o_{\mathbb{P}}(h^d),
\end{aligned}
$$

which gives (D.6) by Lemma C.11.

We will now prove (D.7). By Assumption B.4(iv),

$$\begin{aligned}\mathbb{E}[\delta_{\boldsymbol{q},i}(\boldsymbol{\alpha}) \,|\, \boldsymbol{x}_i] &= \boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}(\boldsymbol{x}_i)\Psi(\boldsymbol{x}_i, \eta(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \gamma\boldsymbol{\alpha})); \boldsymbol{q})\\ &\quad \times \eta^{(1)}(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \gamma\boldsymbol{\alpha}))\\ &= \boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}(\boldsymbol{x}_i)\Psi_1(\boldsymbol{x}_i, \xi_{\boldsymbol{q},i}; \boldsymbol{q})(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \gamma\boldsymbol{\alpha}) - \mu_0(\boldsymbol{x}_i, \boldsymbol{q}))\\ &\quad \times \eta^{(1)}(\zeta_{\boldsymbol{q},i})\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \gamma\boldsymbol{\alpha}))\\ &\gtrsim \boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\boldsymbol{\alpha} - C\sup_{\boldsymbol{q}\in\mathcal{Q},\boldsymbol{x}\in\mathcal{X}}|\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}) - \mu_0(\boldsymbol{x}, \boldsymbol{q})| \cdot |\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}(\boldsymbol{x}_i)|\end{aligned}$$

almost surely uniformly over $\boldsymbol{q}$, where C is some positive constant not depending on n or i, $\xi_{\boldsymbol{q},i}$ is between $\eta(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q})+\gamma\boldsymbol{\alpha}))$ and $\eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q}))$, $\zeta_{\boldsymbol{q},i}$ between $\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q})+\gamma\boldsymbol{\alpha})$ and $\mu_0(\boldsymbol{x}_i, \boldsymbol{q})$. We used that $\Psi(\boldsymbol{x}_i, \eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q})); \boldsymbol{q}) = 0$, γ is small enough, $\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})$ is (for large enough n) uniformly close to $\mu_0(\boldsymbol{x}, \boldsymbol{q})$ by Assumption B.6 and $\eta(\cdot)$ is strictly monotonic by Assumption (B.4)(iii) giving the positivity of the product $\eta^{(1)}(\zeta_{\boldsymbol{q},i})\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \gamma\boldsymbol{\alpha}))$.

Again using the uniform approximation bound $\sup_{\boldsymbol{q}\in\mathcal{Q},\boldsymbol{x}\in\mathcal{X}}|\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}) - \mu_0(\boldsymbol{x}, \boldsymbol{q})| \lesssim h^m$ by Assumption B.6, we obtain

$$\sup_{\boldsymbol{q}\in\mathcal{Q},\boldsymbol{x}\in\mathcal{X}}|\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}) - \mu_0(\boldsymbol{x}, \boldsymbol{q})| \cdot \mathbb{E}_n\big[|\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}(\boldsymbol{x}_i)|\big] \lesssim_{\mathbb{P}} h^{m+d/2} \tag{D.9}$$

since $\mathbb{E}_n\big[|\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}(\boldsymbol{x}_i)|\big] \lesssim \mathbb{E}_n\big[(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}(\boldsymbol{x}_i))^2\big]^{1/2} \lesssim_{\mathbb{P}} h^{d/2}$ by Lyapunov's inequality and Lemma C.11. Note that since $m > d/2$, $h^{m+d/2} = o(h^d)$. (D.7) is proven.

We will now prove (D.8). Define the function class

$$\mathcal{G}_1 := \{(\boldsymbol{x}_i, y_i) \mapsto \delta_{\boldsymbol{q},i}(\boldsymbol{\alpha})\colon \|\boldsymbol{\alpha}\| = 1, \boldsymbol{q} \in \mathcal{Q}\}.$$

By Assumption B.4(iii), B.2(iii) and Assumption B.6, for γ small enough

$$\big|\psi(y_i, \eta(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \gamma\boldsymbol{\alpha})); \boldsymbol{q}) - \psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q})); \boldsymbol{q})\big| \lesssim 1 + \overline{\psi}(\boldsymbol{x}_i, y_i).$$

Recalling the envelope condition in Assumption B.4(ii) and that $\mu_0(\cdot, \boldsymbol{q})$ is bounded by Assumption B.3(iv), we see that $\sup_{g\in\mathcal{G}_1}|g| \lesssim 1 + \overline{\psi}(\boldsymbol{x}_i, y_i)$, which means that under (D.1)

$$\begin{aligned}\mathbb{E}\Big[\max_{1\le i\le n}|\delta_{\boldsymbol{q},i}(\boldsymbol{\alpha})|^2 \,\Big|\, \mathbf{X}_n\Big]^{1/2} &\le \mathbb{E}\Big[\max_{1\le i\le n}|\delta_{\boldsymbol{q},i}(\boldsymbol{\alpha})|^\nu \,\Big|\, \mathbf{X}_n\Big]^{1/\nu}\\ &\le \mathbb{E}\Big[\sum_{i=1}^n|\delta_{\boldsymbol{q},i}(\boldsymbol{\alpha})|^\nu \,\Big|\, \mathbf{X}_n\Big]^{1/\nu} \lesssim \Big(\sum_{i=1}^n \mathbb{E}\big[\big(1 + \overline{\psi}(\boldsymbol{x}_i, y_i)\big)^\nu \,\big|\, \mathbf{X}_n\big]\Big)^{1/\nu}\\ &\lesssim \Big(\sum_{i=1}^n 1\Big)^{1/\nu} = n^{1/\nu} \quad \text{a. s.}\end{aligned}$$

and under (D.2) by Lemma C.1

$$\mathbb{E}\Big[\max_{1\le i\le n}|\delta_{\boldsymbol{q},i}(\boldsymbol{\alpha})|^2 \,\Big|\, \mathbf{X}_n\Big]^{1/2} \lesssim \sqrt{\log n} \quad \text{a. s.}$$

By similar considerations $\sup_{g\in\mathcal{G}_1}\mathbb{E}_n\big[\mathbb{E}[g^2 \,|\, \boldsymbol{x}_i]\big] \lesssim \mathbb{E}_n[\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\boldsymbol{\alpha}] \lesssim h^d$ w. p. a. 1, where the last inequality holds by Lemma C.11.

By Assumption B.5, the class

$$\{(\boldsymbol{x}_i, y_i) \mapsto \psi(y_i, \eta(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\boldsymbol{\beta}); \boldsymbol{q})\colon \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(\boldsymbol{q})\| \le \gamma, \boldsymbol{q} \in \mathcal{Q}\}$$

with envelope $1+\overline{\psi}(\boldsymbol{x}_i, y_i)$ multiplied by a constant has a uniform entropy bound (A.3) with $A \lesssim 1$, $V \lesssim K \asymp h^{-d}$. Moreover, the class

$$\{(\boldsymbol{x}_i, y_i) \mapsto \boldsymbol{\alpha}^\mathsf{T} \boldsymbol{p}(\boldsymbol{x}_i) \colon \|\boldsymbol{\alpha}\| = 1\}$$

has a constant envelope and is VC with index no more than $K+2$ by Lemma 2.6.15 in [41], which means it satisfies the uniform entropy bound (A.3) with $A \lesssim 1$, $V \lesssim K \asymp h^{-d}$. Similarly, the same is true of

$$\{(\boldsymbol{x}_i, y_i) \mapsto \boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T} \boldsymbol{\beta} \colon \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(\boldsymbol{q})\| \le \gamma, \boldsymbol{q} \in \mathcal{Q}\}$$

and therefore of

$$\{(\boldsymbol{x}_i, y_i) \mapsto \eta^{(1)}(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T} \boldsymbol{\beta}) \colon \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(\boldsymbol{q})\| \le \gamma, \boldsymbol{q} \in \mathcal{Q}\}$$

since $\eta^{(1)}(\cdot)$ on a bounded interval is Lipschitz. By Lemma C.4, we conclude that $\mathcal{G}_1$ satisfies the uniform entropy bound (A.3) with envelope $1+\overline{\psi}(\boldsymbol{x}_i, y_i)$ multiplied by a constant, $A \lesssim 1$ and $V \lesssim K \asymp h^{-d}$.

Applying the maximal inequality Lemma C.6, we obtain (D.8). □

D.2 Nonconvex case

Our next goal is to prove the consistency result Lemma D.3 for the nonconvex case. We will need the following lemma.

Lemma D.2 (Preparation for consistency in the nonconvex case)**.** *Suppose Assumptions B.1 to B.4 hold. Then the infinity norm of $\boldsymbol{\beta}_0(\boldsymbol{q})$ is bounded:*

$$\sup_{\boldsymbol{q} \in \mathcal{Q}} \|\boldsymbol{\beta}_0(\boldsymbol{q})\|_\infty \lesssim 1. \tag{D.10}$$

Moreover, for any $R > 0$, there is a positive constant $C_1 = C_1(R)$ depending only on R such that for any $\boldsymbol{x} \in \mathcal{X}$

$$\sup_{\boldsymbol{q} \in \mathcal{Q}} \sup_{\|\boldsymbol{\beta}\|_\infty \le R} \left|\rho(y, \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T} \boldsymbol{\beta}); \boldsymbol{q}) - \rho(y, \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T} \boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q})\right| \le C_1(1 + \overline{\psi}(\boldsymbol{x}, y)). \tag{D.11}$$

Proof. We prove (D.10) first. By Assumption B.6 ($\boldsymbol{\beta}_0(\boldsymbol{q})^\mathsf{T} \boldsymbol{p}(\boldsymbol{x})$ is close to $\mu_0(\boldsymbol{x}, \boldsymbol{q})$) and Assumption B.3(iv) ($\mu_0(\boldsymbol{x}, \boldsymbol{q})$ is uniformly bounded), $|\boldsymbol{\beta}_0(\boldsymbol{q})^\mathsf{T} \boldsymbol{p}(\boldsymbol{x})|$ is bounded uniformly over $\boldsymbol{q} \in \mathcal{Q}$ and $\boldsymbol{x} \in \mathcal{X}$. By Assumption B.2, we can bound the kth coordinate of $\boldsymbol{\beta}_0(\boldsymbol{q})$

$$\begin{aligned}
|(\boldsymbol{\beta}_0(\boldsymbol{q}))_k| &\lesssim h^{-d/2} \left(\int_{\mathcal{H}_k} \left(\boldsymbol{\beta}_0(\boldsymbol{q})^\mathsf{T} \boldsymbol{p}(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x}\right)^{1/2} \\
&\le h^{-d/2} \sup_{\boldsymbol{x} \in \mathcal{X}} |\boldsymbol{\beta}_0(\boldsymbol{q})^\mathsf{T} \boldsymbol{p}(\boldsymbol{x})| \cdot (\operatorname{Leb} \mathcal{H}_k)^{1/2} \lesssim \sup_{\boldsymbol{x} \in \mathcal{X}} |\boldsymbol{\beta}_0(\boldsymbol{q})^\mathsf{T} \boldsymbol{p}(\boldsymbol{x})| \lesssim 1,
\end{aligned}$$

where the constants in $\lesssim$ do not depend on k.

Now we prove (D.11). Note that

$$\begin{aligned}
&\rho(y, \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T} \boldsymbol{\beta}); \boldsymbol{q}) - \rho(y, \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T} \boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q}) \\
&\quad = \int_0^{\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta} - \boldsymbol{\beta}_0(\boldsymbol{q}))} (\psi(y, \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T} \boldsymbol{\beta}_0(\boldsymbol{q}) + t); \boldsymbol{q}) - \psi(y, \eta(\mu_0(\boldsymbol{x}, \boldsymbol{q})); \boldsymbol{q})) \\
&\qquad \times \eta^{(1)}(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T} \boldsymbol{\beta}_0(\boldsymbol{q}) + t) \, \mathrm{d}t \\
&\qquad + \psi(y, \eta(\mu_0(\boldsymbol{x}, \boldsymbol{q})); \boldsymbol{q}) \int_0^{\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta} - \boldsymbol{\beta}_0(\boldsymbol{q}))} \eta^{(1)}(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T} \boldsymbol{\beta}_0(\boldsymbol{q}) + t) \, \mathrm{d}t
\end{aligned}$$

By Assumption B.4(iii), we have a bound

$$|\psi(y, \eta(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}) + t); \boldsymbol{q}) - \psi(y, \eta(\mu_0(\boldsymbol{x}, \boldsymbol{q})); \boldsymbol{q})| \lesssim \overline{\psi}(\boldsymbol{x}, y) + 1.$$

Since both $\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q})$ and $\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}$ lie in a fixed compact interval (not depending on $\boldsymbol{x}$ or $\boldsymbol{q}$), $\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}) + t)$ is uniformly bounded in absolute value. This means that for any $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{q} \in \mathcal{Q}$, $\|\boldsymbol{\beta}\|_\infty \le R$, we have for some positive constants C_2 and C_1 depending only on R

$$\begin{aligned}
&|\rho(y, \eta(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}); \boldsymbol{q}) - \rho(y, \eta(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q})| \le C_2(1 + \overline{\psi}(\boldsymbol{x}, y)) \cdot |\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0(\boldsymbol{q}))| \\
&\quad \le C_2(1 + \overline{\psi}(\boldsymbol{x}, y)) \cdot \|\boldsymbol{p}(\boldsymbol{x})\|_1 \cdot \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(\boldsymbol{q})\|_\infty \\
&\quad \le C_2(1 + \overline{\psi}(\boldsymbol{x}, y)) \cdot \|\boldsymbol{p}(\boldsymbol{x})\|_1 \cdot (\|\boldsymbol{\beta}\|_\infty + \|\boldsymbol{\beta}_0(\boldsymbol{q})\|_\infty) \\
&\quad \le C_1(1 + \overline{\psi}(\boldsymbol{x}, y)),
\end{aligned}$$

concluding the proof. □

We are now ready to prove a general consistency result for an estimator under constraints $\|\boldsymbol{\beta}\|_\infty \le R$ for some large enough constant R. This is Lemma 2 in the paper.

Lemma D.3 (Consistency, nonconvex case)**.** *Assume the following conditions.*

(i) Assumptions B.1 to B.4 and B.6 hold.

(ii) $m > d/2$.

(iii) The following rate condition holds:

$$\frac{\left[\log(1/h)\right]^{\frac{\nu}{\nu-1}}}{nh^{\frac{2\nu}{\nu-1}d}} = o(1), \text{ or} \tag{D.12}$$

$$\overline{\psi}(\boldsymbol{x}_i, y_i) \text{ is } \sigma^2\text{-sub-Gaussian conditionally on } \boldsymbol{x}_i \text{ and } \frac{\sqrt{\log n}\log(1/h)}{nh^{2d}} = o(1). \tag{D.13}$$

(iv) $\mathcal{B} = \{\boldsymbol{\beta} \in \mathbb{R}^K : \|\boldsymbol{\beta}\|_\infty \le R\}$ *in (A.2), where* $R > 0$ *is a fixed number (not depending on* n*) such that* $\sup_{\boldsymbol{q}\in\mathcal{Q}}\|\boldsymbol{\beta}_0(\boldsymbol{q})\|_\infty \le R/2$ *(existing by Lemma D.2).*

(v) There is a positive constant c *such that we have* $\inf \Psi_1(\boldsymbol{x}, \zeta; \boldsymbol{q}) > c$*, where the infimum is over* $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{q} \in \mathcal{Q}$, ζ *between* $\eta(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta})$ *and* $\eta(\mu_0(\boldsymbol{x}, \boldsymbol{q}))$*, and* $\boldsymbol{\beta} \in \mathcal{B}$.

(vi) The class of functions

$$\{(\boldsymbol{x}, y) \mapsto \rho(y, \eta(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}); \boldsymbol{q}) - \rho(y, \eta(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q}) : \|\boldsymbol{\beta}\|_\infty \le R, \boldsymbol{q} \in \mathcal{Q}\}$$

with envelope $C_1(1 + \overline{\psi}(\boldsymbol{x}, y))$ *(by Lemma D.2) satisfies the uniform entropy bound (A.3) with* $A \lesssim 1$ *and* $V \lesssim K$.

Then Eqs. (D.3) to (D.5) hold.

Proof. For $\boldsymbol{\beta}$ satisfying the constraint $\|\boldsymbol{\beta}\|_\infty \le R$, define

$$\begin{aligned}
\delta_{\boldsymbol{q},i}(\boldsymbol{\beta}) &:= \rho(y_i, \eta(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\beta}); \boldsymbol{q}) - \rho(y_i, \eta(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q}) \\
&= \int_0^{\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q}))} \psi(y_i, \eta(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}) + t); \boldsymbol{q})\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}) + t)\,\mathrm{d}t.
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{E}[\delta_{\boldsymbol{q},i}(\boldsymbol{\beta}) \mid \boldsymbol{x}_i] &= \int_0^{\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}(\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q}))} \Psi(\boldsymbol{x}_i, \eta(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})+t); \boldsymbol{q})\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})+t)\,\mathrm{d}t \\
&\overset{\text{(a)}}{=} \int_0^{\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}(\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q}))} \Psi_1(\boldsymbol{x}_i, \xi_{\boldsymbol{q},i,t}; \boldsymbol{q}) \\
&\quad \times \eta^{(1)}(\zeta_{\boldsymbol{q},i,t})\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})+t)\{\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}) - \mu_0(\boldsymbol{x}_i,\boldsymbol{q})+t\}\,\mathrm{d}t \\
&\overset{\text{(b)}}{\geq} C_4\{\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}(\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q}))\}^2 - C_3 \sup_{\boldsymbol{q},\boldsymbol{x}}|\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}) - \mu_0(\boldsymbol{x},\boldsymbol{q})| \cdot |\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}(\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q}))| \\
&\overset{\text{(c)}}{\geq} C_4\{\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}(\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q}))\}^2 - C_5 h^m |\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}(\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q}))|,
\end{aligned}$$

with some positive constants C_4 and C_5 (depending on R), where in (a) we used

$$\begin{aligned}
\Psi(\boldsymbol{x}_i, \eta(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})+t); \boldsymbol{q}) &= \Psi(\boldsymbol{x}_i, \eta(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})+t); \boldsymbol{q}) - \Psi(\boldsymbol{x}_i, \eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q})); \boldsymbol{q}) \\
&= \Psi_1(\boldsymbol{x}_i, \xi_{\boldsymbol{q},i,t}; \boldsymbol{q})\{\eta(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})+t) - \eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))\} \\
&= \Psi_1(\boldsymbol{x}_i, \xi_{\boldsymbol{q},i,t}; \boldsymbol{q})\eta^{(1)}(\zeta_{\boldsymbol{q},i,t})\{\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}) - \mu_0(\boldsymbol{x}_i,\boldsymbol{q})+t\}
\end{aligned}$$

for some $\xi_{\boldsymbol{q},i,t}$ between $\eta(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})+t)$ and $\eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))$, and some $\zeta_{\boldsymbol{q},i,t}$ between $\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})+t$ and $\mu_0(\boldsymbol{x}_i,\boldsymbol{q})$ by the mean-value theorem applied twice; in (b) we used Condition (v), Assumption B.4, in particular that $\eta(\cdot)$ is strictly monotonic giving the positivity of the product $\eta^{(1)}(\zeta_{\boldsymbol{q},i,t})\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})+t)$; in (c) we used Assumption B.6. By Lyapunov's inequality, $\mathbb{E}_n[|\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}(\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q}))|] \leq \mathbb{E}_n[(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}(\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})))^2]^{1/2}$. We conclude

$$\begin{aligned}
&\mathbb{E}_n\big[\mathbb{E}[\delta_{\boldsymbol{q},i}(\boldsymbol{\beta}) \mid \boldsymbol{x}_i]\big] \\
&\quad \geq C_4(\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q}))^\mathsf{T}\mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}](\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})) - C_5 h^m \mathbb{E}_n[(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}(\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})))^2]^{1/2} \\
&\quad \overset{\text{(a)}}{\geq} C_6 h^d \|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})\|^2 - C_7 h^{m+d/2}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})\|
\end{aligned}$$

with probability approaching one for some other positive constants C_6 and C_7 (depending on R), where (a) is by Lemma C.11.

Fix $\varepsilon > 0$ smaller than $R/2$. In this case

$$\{\boldsymbol{\beta} : \|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})\| \leq \varepsilon\} \subset \{\boldsymbol{\beta} : \|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})\|_\infty \leq \varepsilon\} \subset \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_\infty \leq R\}$$

because

$$\|\boldsymbol{\beta}\|_\infty \leq \|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})\|_\infty + \|\boldsymbol{\beta}_0(\boldsymbol{q})\|_\infty \leq \|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})\|_\infty + R/2.$$

Define the class of functions

$$\begin{aligned}
\mathcal{G} := \Big\{&(\boldsymbol{x}, y) \mapsto \|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})\|^{-1}(\rho(y, \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}); \boldsymbol{q}) - \rho(y, \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q})) : \\
&\|\boldsymbol{\beta}\|_\infty \leq R, \|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})\| > \varepsilon, \boldsymbol{q} \in \mathcal{Q}\Big\}.
\end{aligned}$$

It is a product of a subclass of the class

$$\{(\boldsymbol{x}, y) \mapsto a : 0 < a < 1/\varepsilon\}$$

with envelope $1/\varepsilon$, obviously satisfying the uniform entropy bound (A.3) with $A \lesssim 1$, $V \lesssim 1$, and a subclass of the class

$$\{(\boldsymbol{x}, y) \mapsto \rho(y, \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})) - \rho(y, \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}))) : \|\boldsymbol{\beta}\|_\infty \leq R, \boldsymbol{q} \in \mathcal{Q}\}$$

with envelope $C_1(1+\overline{\psi}(\boldsymbol{x},y))$, satisfying the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim K$ by the conditions of Lemma D.3.

By Lemma C.4, class $\mathcal{G}$ with envelope $C_1/\varepsilon(1+\overline{\psi}(\boldsymbol{x},y))$ satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim K$.

Next, under (D.12)

$$\begin{aligned}
&\mathbb{E}\big[\max_{1\le i\le n}(C_1/\varepsilon)^2(1+\overline{\psi}(\boldsymbol{x},y))^2 \,\big|\, \mathbf{X}_n\big]^{1/2}\\
&\quad\le \mathbb{E}\big[\max_{1\le i\le n}\big(C_1/\varepsilon\big)^{\nu}\big(1+\overline{\psi}(\boldsymbol{x},y)\big)^{\nu} \,\big|\, \mathbf{X}_n\big]^{1/\nu}\\
&\quad\le \mathbb{E}\Big[\sum_{i=1}^{n}\big(C_1/\varepsilon\big)^{\nu}\big(1+\overline{\psi}(\boldsymbol{x},y)\big)^{\nu} \,\big|\, \mathbf{X}_n\Big]^{1/\nu}\\
&\quad\lesssim \frac{1}{\varepsilon}\bigg(\sum_{i=1}^{n}\mathbb{E}\big[(1+\overline{\psi}(\boldsymbol{x}_i,y_i))^{\nu} \,\big|\, \mathbf{X}_n\big]\bigg)^{1/\nu}\\
&\quad\lesssim \frac{1}{\varepsilon}\bigg(\sum_{i=1}^{n}1\bigg)^{1/\nu} = \frac{n^{1/\nu}}{\varepsilon} \quad \text{a.s.}
\end{aligned}$$

with constants in $\lesssim$ depending on R but not on n or ε, and under (D.13) by Lemma C.1

$$\mathbb{E}\bigg[\max_{1\le i\le n}(C_1/\varepsilon)^2(1+\overline{\psi}(\boldsymbol{x},y))^2 \,\bigg|\, \mathbf{X}_n\bigg]^{1/2} \lesssim \frac{\sqrt{\log n}}{\varepsilon} \quad \text{a.s.}$$

Moreover,

$$\begin{aligned}
&\mathbb{E}_n\big[\mathbb{E}[g(\boldsymbol{x}_i,y_i)^2 \mid \boldsymbol{x}_i]\big]\\
&\quad\le \|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})\|^{-2}C_2^2\mathbb{E}_n\big[\big(\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q}))\big)^2\mathbb{E}[(1+\overline{\psi}(\boldsymbol{x}_i,y_i))^2 \mid \boldsymbol{x}_i]\big]\\
&\quad\lesssim \|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})\|^{-2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q}))^{\mathsf{T}}\mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}](\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q}))\\
&\quad\overset{\text{(a)}}{\lesssim} h^d,
\end{aligned}$$

where (a) is by Lemma C.11.

By Lemma C.6, we have

$$\sup_{g\in\mathcal{G}}\big|\mathbb{E}\big[g(\boldsymbol{x}_i,y_i)-\mathbb{E}[g(\boldsymbol{x}_i,y_i)\mid \boldsymbol{x}_i]\big]\big| \lesssim_{\mathbb{P}} \begin{cases}\sqrt{\frac{\log(1/h)}{n}}+\frac{\log(1/h)}{n^{1-1/\nu}h^d}=o(h^d),\\ \sqrt{\frac{\log(1/h)}{n}}+\frac{\sqrt{\log n}\log(1/h)}{nh^d}=o(h^d)\end{cases}$$

under (D.12) and (D.13) respectively (since ε is fixed).

Combining, we infer from the previous results that with probability approaching one for all $\|\boldsymbol{\beta}\|_\infty \le R$, $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})\| > \varepsilon$, $\boldsymbol{q}\in\mathcal{Q}$

$$\begin{aligned}
\mathbb{E}_n[\delta_{\boldsymbol{q},i}(\boldsymbol{\beta})] &\ge C_6h^d\|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})\|^2 - C_7h^{m+d/2}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})\| + \|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})\|\cdot o(h^d)\\
&= \|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})\|\cdot\{C_6h^d\|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})\| - C_7h^{m+d/2}+o(h^d)\}\\
&> \|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})\|\cdot\{C_6\varepsilon h^d - C_7h^{m+d/2}+o(h^d)\}\\
&\overset{\text{(a)}}{=} \|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})\|\cdot h^d\{C_6\varepsilon+o(1)\} > 0,
\end{aligned}$$

where in (a) we used $m > d/2$.

It follows that the constrained minimizer under the constraint $\|\boldsymbol{\beta}\|_\infty \le R$ has to lie inside the ball $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0(\boldsymbol{q})\| \le \varepsilon$ for all $\boldsymbol{q} \in \mathcal{Q}$ with probability approaching one. Since ε was arbitrary smaller than $R/2$, it is equivalent to $\sup_{\boldsymbol{q}\in\mathcal{Q}}\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})\| = o_{\mathbb{P}}(1)$. Equations (D.4) and (D.5) follow (as in Lemma D.1). □

D.3 Weaker conditions for special cases

Lemmas D.4 and D.5 consider the special case of unconnected basis functions.

Lemma D.4 (Consistency, convex case, unconnected basis functions)**.** *Assume the following.*

(i) Assumptions B.1 to B.6 hold.

(ii) $\rho(y, \eta(\theta); \boldsymbol{q})$ is convex with respect to θ, and $\psi(y, \eta(\theta); \boldsymbol{q})\eta^{(1)}(\theta)$ is its left or right derivative, and $\mathcal{B} = \mathbb{R}^K$ in (A.2).

(iii) For all $k \in \{1, \dots, K\}$ the kth basis function $p_k(\cdot)$ is only active on one of the cells of Δ.

(iv) The rate of convergence of h to zero is restricted by

$$\frac{\big[\log(1/h)\big]^{\frac{\nu}{\nu-1}}}{nh^{\frac{\nu}{\nu-1}d}} = o(1), \text{ or}$$

$$\overline{\psi}(\boldsymbol{x}_i, y_i) \text{ is } \sigma^2\text{-sub-Gaussian conditionally on } \boldsymbol{x}_i \text{ and } \frac{\sqrt{\log n}\,\log(1/h)}{nh^d} = o(1).$$

Then Eqs. (D.3) *to* (D.5) *hold.*

Proof. As in Lemma D.1, Eqs. (D.4) and (D.5) follow from Eq. (D.3).

For $l \in \{1, \dots, \kappa\}$, the number M_l of basis functions in $\boldsymbol{p}(\cdot)$ which are active on the lth cell of Δ is bounded by a constant. Denote the vector of such basis functions $\boldsymbol{p}_l := (p_{l,1}, \dots, p_{l,M_l})^{\mathsf{T}}$. Define the matrices $\mathbf{Q}_{0,\boldsymbol{q},l}$ and $\bar{\mathbf{Q}}_{\boldsymbol{q},l}$ as before with $\boldsymbol{p}$ replaced by $\boldsymbol{p}_l$ (for different l, the dimensions of these square matrices may vary but are bounded from above). By a simple modification of the argument in Lemma C.11, the analogues of (C.2) and (C.5) continue to hold: uniformly over $\boldsymbol{q} \in \mathcal{Q}$ and l

$$\begin{aligned} &h^d \lesssim \lambda_{\min}(\mathbf{Q}_{0,\boldsymbol{q},l}) \le \lambda_{\max}(\mathbf{Q}_{0,\boldsymbol{q},l}) \lesssim h^d, \\ &h^d \lesssim \lambda_{\min}(\bar{\mathbf{Q}}_{\boldsymbol{q},l}) \le \lambda_{\max}(\bar{\mathbf{Q}}_{\boldsymbol{q},l}) \lesssim h^d \quad \text{w. p. a. 1.} \end{aligned} \tag{D.14}$$

By the assumption of the lemma, we can write $\boldsymbol{\beta}_0(\boldsymbol{q}) = (\boldsymbol{\beta}_{0,\boldsymbol{q},1}, \dots, \boldsymbol{\beta}_{0,\boldsymbol{q},\kappa})^{\mathsf{T}}$, where $\boldsymbol{\beta}_{0,\boldsymbol{q},l}$ is a subvector of dimension M_l corresponding to the elements in $\boldsymbol{p}$ active on the lth cell.

Fix a sufficiently small $\gamma > 0$. Denote for $l \in \{1, \dots, \kappa\}$, $i \in \{1, \dots, n\}$ and $\boldsymbol{\alpha}_l \in \mathcal{S}^{M_l - 1}$

$$\delta_{\boldsymbol{q},i,l}(\boldsymbol{\alpha}_l) := \boldsymbol{\alpha}_l^{\mathsf{T}} \boldsymbol{p}_l(\boldsymbol{x}_i) \psi(y_i, \eta(\boldsymbol{p}_l(\boldsymbol{x}_i)^{\mathsf{T}}(\boldsymbol{\beta}_{0,\boldsymbol{q},l} + \gamma\boldsymbol{\alpha}_l)); \boldsymbol{q})\eta^{(1)}(\boldsymbol{p}_l(\boldsymbol{x}_i)^{\mathsf{T}}(\boldsymbol{\beta}_{0,\boldsymbol{q},l} + \gamma\boldsymbol{\alpha}_l)).$$

Proceeding in the same way as in Lemma D.1, we will show

$$\inf_{\boldsymbol{q},\boldsymbol{\alpha}_l,l} \mathbb{E}_n[\delta_{\boldsymbol{q},i,l}(\boldsymbol{\alpha}_l)] > 0 \quad \text{with probability approaching 1,} \tag{D.15}$$

which is again enough to prove the lemma by convexity. It will follow that with probability approaching one the minimizer $\widehat{\boldsymbol{\beta}}_{\boldsymbol{q},l}$ of $\mathbb{E}_n\big[\rho(y_i, \eta(\boldsymbol{p}_l(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\beta}_l); \boldsymbol{q})\big]$ with respect to $\boldsymbol{\beta}_l$ has to lie inside the ball $\|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\boldsymbol{q},l}\| \le \gamma$, and in particular inside the cube $\|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\boldsymbol{q},l}\|_\infty \le \gamma$. But note that $\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) = \big(\widehat{\boldsymbol{\beta}}_{\boldsymbol{q},1}^{\mathsf{T}}, \dots, \widehat{\boldsymbol{\beta}}_{\boldsymbol{q},\bar{k}}^{\mathsf{T}}\big)^{\mathsf{T}}$. So, with probability approaching one, for all $\boldsymbol{q} \in \mathcal{Q}$, we have $\big\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})\big\|_\infty \le \gamma$. Since γ was arbitrary (small enough), it is equivalent to $\sup_{\boldsymbol{q}\in\mathcal{Q}}\big\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})\big\|_\infty = o_{\mathbb{P}}(1)$.

Equation (D.15) is proven analogously to the corresponding argument in Lemma D.1. The class of functions

$$\mathcal{G}_1 := \big\{(\boldsymbol{x}_i, y_i) \mapsto \delta_{\boldsymbol{q},i,l}(\boldsymbol{\alpha}_l) \colon \boldsymbol{\alpha}_l \in \mathcal{S}^{M_l-1}, l \in \{1, \dots, \kappa\}, \boldsymbol{q} \in \mathcal{Q}\big\}$$

now satisfies the uniform entropy bound (A.3) with $A \lesssim \kappa \asymp h^{-d}$ (since there are κ different values of l) and $V \lesssim 1$ (since the vectors $\boldsymbol{\alpha}_l$ are of bounded dimensions). The bound $\sup_{l,\boldsymbol{\alpha}_l} \mathbb{E}_n\big[|\boldsymbol{\alpha}_l^\mathsf{T} \boldsymbol{p}_l(\boldsymbol{x}_i)|\big] \lesssim h^d$ with probability approaching one can be proven without assuming $m > d/2$ by using

$$\sup_{l,\boldsymbol{\alpha}_l} \mathbb{E}_n\big[|\boldsymbol{\alpha}_l^\mathsf{T} \boldsymbol{p}_l(\boldsymbol{x}_i)|\big] \le \sup_{l,\boldsymbol{\alpha}_l} \|\boldsymbol{\alpha}_l\|_\infty \mathbb{E}_n\big[\|\boldsymbol{p}_l(\boldsymbol{x}_i)\|_1\big] \lesssim h^d \quad \text{w. p. a. 1},$$

since the dimension of $\boldsymbol{p}_l(\cdot)$ is uniformly bounded. □

Lemma D.5 (Consistency in the nonconvex case: unconnected basis)**.** *Assume that for all $k \in \{1, \dots, K\}$ the kth basis function $p_k(\cdot)$ is only active on one of the cells of Δ, and define $\boldsymbol{p}_l(\cdot)$, M_l, $\boldsymbol{\beta}_{0,\boldsymbol{q},l}$ as in the proof of Lemma D.4. Further, assume the conditions of Lemma D.3 with Condition (ii) removed, Condition (iii) replaced by*

$$\frac{\big[\log(1/h)\big]^{\frac{\nu}{\nu-1}}}{n h^{\frac{\nu}{\nu-1} d}} = o(1), \ \text{or} \tag{D.16}$$

$$\overline{\psi}(\boldsymbol{x}_i, y_i) \text{ is } \sigma^2\text{-sub-Gaussian conditionally on } \boldsymbol{x}_i \text{ and } \frac{\sqrt{\log n}\,\log(1/h)}{n h^d} = o(1), \tag{D.17}$$

and Condition (vi) replaced by the following: the class

$$\{(\boldsymbol{x}, y) \mapsto \rho(y, \eta(\boldsymbol{p}_l(\boldsymbol{x})^\mathsf{T} \boldsymbol{\beta}_l)) - \rho(y, \eta(\boldsymbol{p}_l(\boldsymbol{x})^\mathsf{T} \boldsymbol{\beta}_{0,\boldsymbol{q},l})) : \|\boldsymbol{\beta}_l\|_\infty \le R, \boldsymbol{q} \in \mathcal{Q}, l \in \{1, \dots, \kappa\}\}$$

satisfies the uniform entropy bound (A.3) with $A \lesssim \kappa \asymp h^{-d}$ and $V \lesssim 1$. Then Eqs. (D.3) to (D.5) hold.

Proof. As in Lemma D.1, Eqs. (D.4) and (D.5) follow from Eq. (D.3).

Define matrices $\mathbf{Q}_{0,\boldsymbol{q},l}$ and $\bar{\mathbf{Q}}_{\boldsymbol{q},l}$ as in the proof of Lemma D.4, and recall that the asymptotic bounds on their eigenvalues are the same as in the general (not restricted to one cell) case, i. e. (D.14) holds.

For any M_l-dimensional vector $\boldsymbol{\beta}_l$ satisfying the constraint $\|\boldsymbol{\beta}_l\|_\infty \le R$, define

$$\begin{aligned}\delta_{\boldsymbol{q},i,l}(\boldsymbol{\beta}_l) &:= \rho(y_i, \eta(\boldsymbol{p}_l(\boldsymbol{x}_i)^\mathsf{T} \boldsymbol{\beta}_l); \boldsymbol{q}) - \rho(y_i, \eta(\boldsymbol{p}_l(\boldsymbol{x}_i)^\mathsf{T} \boldsymbol{\beta}_{0,\boldsymbol{q},l}); \boldsymbol{q}) \\ &= \int_0^{\boldsymbol{p}_l(\boldsymbol{x}_i)^\mathsf{T}(\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\boldsymbol{q},l})} \psi(y_i, \eta(\boldsymbol{p}_l(\boldsymbol{x}_i)^\mathsf{T} \boldsymbol{\beta}_{0,\boldsymbol{q},l} + t); \boldsymbol{q}) \eta^{(1)}(\boldsymbol{p}_l(\boldsymbol{x}_i)^\mathsf{T} \boldsymbol{\beta}_{0,\boldsymbol{q},l} + t)\, \mathrm{d}t.\end{aligned}$$

By the same argument as in the proof of Lemma D.3, we have

$$\begin{aligned}&\mathbb{E}_n\big[\mathbb{E}[\delta_{\boldsymbol{q},i,l}(\boldsymbol{\beta}_l)|\boldsymbol{x}_i]\big] \\ &\quad \ge C_8 (\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\boldsymbol{q},l})^\mathsf{T} \mathbb{E}_n[\boldsymbol{p}_l(\boldsymbol{x}_i) \boldsymbol{p}_l(\boldsymbol{x}_i)^\mathsf{T}] (\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\boldsymbol{q},l}) - C_9 h^m \mathbb{E}_n\big[\big|\boldsymbol{p}_l(\boldsymbol{x}_i)^\mathsf{T} (\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\boldsymbol{q},l})\big|\big] \\ &\quad \overset{\text{(a)}}{\ge} C_8 h^d \|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\boldsymbol{q},l}\|^2 - C_{10} h^{m+d} \|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\boldsymbol{q},l}\|\end{aligned}$$

with probability approaching one for some positive constants C_8, C_9 and C_{10} (depending on R, but not on $\boldsymbol{q}$ or l), where in (a) we used $\mathbb{E}_n\big[\big|\boldsymbol{p}_l(\boldsymbol{x}_i)^\mathsf{T} (\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\boldsymbol{q},l})\big|\big] \le \|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\boldsymbol{q},l}\|_\infty \mathbb{E}_n\big[\|\boldsymbol{p}_l(\boldsymbol{x}_i)\|_1\big] \lesssim \|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\boldsymbol{q},l}\|_\infty h^d \le \|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\boldsymbol{q},l}\|_\infty h^d$ with probability approaching one since the dimension of $\boldsymbol{p}_l(\cdot)$ is bounded.

Next, proceeding with the same concentration argument as in Lemma D.3, we will obtain that with probability approaching one for all $l \in \{1, \dots, \kappa\}$, $\boldsymbol{q} \in \mathcal{Q}$, $\|\boldsymbol{\beta}_l\|_\infty \leq R$, $\|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\boldsymbol{q},l}\| > \varepsilon$,

$$\begin{aligned}
\mathbb{E}_n[\delta_{\boldsymbol{q},i,l}(\boldsymbol{\beta}_l)] &\geq C_8 h^d \|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\boldsymbol{q},l}\|^2 - C_{10} h^{m+d} \|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\boldsymbol{q},l}\| + \|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\boldsymbol{q},l}\| \cdot o(h^d) \\
&= \|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\boldsymbol{q},l}\| \cdot \big\{C_8 h^d \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(\boldsymbol{q})\| - C_{10} h^{m+d} + o(h^d)\big\} \\
&> \|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\boldsymbol{q},l}\| \cdot \big\{C_8 \varepsilon h^d - C_{10} h^{m+d} + o(h^d)\big\} \\
&\overset{\text{(a)}}{=} \|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\boldsymbol{q},l}\| \cdot h^d \{C_8 \varepsilon + o(1)\} > 0.
\end{aligned}$$

It follows that the constrained minimizer $\widehat{\boldsymbol{\beta}}_{\boldsymbol{q},\mathrm{constr},l}$ of $\mathbb{E}_n\big[\rho(y_i, \eta(\boldsymbol{p}_l(\boldsymbol{x}_i)^\mathsf{T} \boldsymbol{\beta}_l); \boldsymbol{q})\big]$ with respect to $\boldsymbol{\beta}_l$ under the constraint $\|\boldsymbol{\beta}_l\|_\infty \leq R$ has to lie inside the ball $\|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\boldsymbol{q},l}\| \leq \varepsilon$, and in particular inside the cube $\|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\boldsymbol{q},l}\|_\infty \leq \varepsilon$. But this optimization can be solved separately for all l, i.e. $\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) = \big(\widehat{\boldsymbol{\beta}}_{\boldsymbol{q},\mathrm{constr},1}, \dots, \widehat{\boldsymbol{\beta}}_{\boldsymbol{q},\mathrm{constr},\bar{k}}\big)^\mathsf{T}$. So, with probability approaching one, for all $\boldsymbol{q} \in \mathcal{Q}$, we have $\big\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})\big\|_\infty \leq \varepsilon$. Since ε was arbitrary smaller than $R/2$, it is equivalent to $\sup_{\boldsymbol{q} \in \mathcal{Q}} \big\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})\big\|_\infty = o_\mathbb{P}(1)$. □

The following lemma considers the special case of strongly convex and strongly smooth loss function.

Lemma D.6 (Consistency: strongly convex and strongly smooth loss case)**.** *Assume the following conditions.*

(i) Assumptions B.1 to B.6 hold.

(ii) The rate of convergence of h to zero is restricted by

$$\frac{\big[\log(1/h)\big]^{\frac{\nu}{\nu-1}}}{n h^{\frac{\nu}{\nu-1} d}} = o(1), \text{ or}$$

$$\overline{\psi}(\boldsymbol{x}_i, y_i) \text{ is } \sigma^2\text{-sub-Gaussian conditionally on } \boldsymbol{x}_i \text{ and } \frac{\log n \log(1/h)}{n h^d} = o(1).$$

(iii) The function $\eta \mapsto \psi(y, \eta; \boldsymbol{q})$ is continuously differentiable on $\mathbb{R}$ (for all $y, \boldsymbol{q}$), and there exist fixed (not depending on n, $\boldsymbol{q}$ or θ) numbers λ, Λ such that

$$0 < \lambda \leq \frac{\partial}{\partial \theta}\big(\psi(y_i, \eta(\theta); \boldsymbol{q}) \eta^{(1)}(\theta)\big) \leq \Lambda,$$

and $\mathcal{B} = \mathbb{R}^K$.

Then Eqs. (D.3) to (D.5) hold.

Proof. As in Lemma D.1, Eqs. (D.4) and (D.5) follow from Eq. (D.3).

Denote for $\boldsymbol{\beta} \in \mathbb{R}^K$

$$G_n(\boldsymbol{\beta}) := \mathbb{E}_n[\psi(y_i, \eta(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T} \boldsymbol{\beta}); \boldsymbol{q}) \eta^{(1)}(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T} \boldsymbol{\beta}) \boldsymbol{p}(\boldsymbol{x}_i)],$$

which is the gradient of the convex (by Assumption B.4(iii)) function $\mathbb{E}_n[\rho(y_i, \eta(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T} \boldsymbol{\beta}); \boldsymbol{q})]$ of $\boldsymbol{\beta}$. By definition of $\widehat{\boldsymbol{\beta}}(\boldsymbol{q})$ and differentiability, $G_n\big(\widehat{\boldsymbol{\beta}}(\boldsymbol{q})\big) = 0$. By the mean value theorem,

$$G_n(\boldsymbol{\beta}_0(\boldsymbol{q})) = G_n(\boldsymbol{\beta}_0(\boldsymbol{q})) - G_n\big(\widehat{\boldsymbol{\beta}}(\boldsymbol{q})\big) = \mathbb{E}_n[\mu_i \boldsymbol{p}(\boldsymbol{x}_i) \boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}]\big(\boldsymbol{\beta}_0(\boldsymbol{q}) - \widehat{\boldsymbol{\beta}}(\boldsymbol{q})\big), \tag{D.18}$$

where

$$\mu_i := \frac{\partial}{\partial \theta}\big(\psi(y_i, \eta(\theta); \boldsymbol{q}) \eta^{(1)}(\theta)\big)\Big|_{\theta = \widetilde{\theta}_i} \quad \text{for some } \widetilde{\theta}_i \text{ between } \boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T} \boldsymbol{\beta}_0(\boldsymbol{q}) \text{ and } \boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T} \widehat{\boldsymbol{\beta}}(\boldsymbol{q}).$$

By the assumption of the lemma, $0 < \lambda \le \mu_i \le \Lambda$. Therefore, for any vector $\boldsymbol{a} \in \mathbb{R}^K$

$$\begin{aligned}\lambda \cdot \boldsymbol{a}^\mathsf{T} \mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}]\boldsymbol{a} &\le \boldsymbol{a}^\mathsf{T} \mathbb{E}_n[\mu_i \boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}]\boldsymbol{a} \\ &= \mathbb{E}_n[\mu_i (\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\boldsymbol{a})^2] \le \Lambda \cdot \boldsymbol{a}^\mathsf{T} \mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}]\boldsymbol{a}.\end{aligned}$$

Moreover, the matrix

$$\mathbb{E}_n[\mu_i \boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}]$$

has the same multi-banded structure as

$$\mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}].$$

That means that by the same argument as that in Lemma C.11 we have

$$\left\|\mathbb{E}_n[\mu_i \boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}]^{-1}\right\|_\infty \lesssim_\mathbb{P} h^{-d}. \tag{D.19}$$

It is shown in the proofs of Lemma C.12 and Lemma C.14 (see also Remarks C.13 and C.15) that

$$\left\|G_n(\boldsymbol{\beta}_0(\boldsymbol{q}))\right\|_\infty = o_\mathbb{P}(h^d). \tag{D.20}$$

From (D.18)

$$\|\boldsymbol{\beta}_0(\boldsymbol{q}) - \widehat{\boldsymbol{\beta}}(\boldsymbol{q})\|_\infty \le \left\|\mathbb{E}_n[\mu_i \boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}]^{-1}\right\|_\infty \cdot \|G_n(\boldsymbol{\beta}_0(\boldsymbol{q}))\|_\infty,$$

which in combination with (D.19) and (D.20) gives

$$\|\boldsymbol{\beta}_0(\boldsymbol{q}) - \widehat{\boldsymbol{\beta}}(\boldsymbol{q})\|_\infty = o_\mathbb{P}(1)$$

uniformly over $\boldsymbol{q} \in \mathcal{Q}$. □

E Bahadur representation

We will now prove our first main result, the novel Bahadur representation which is Theorem 1 in the paper. Recall the notation

$$\mathsf{L}^{(\boldsymbol{v})}(\boldsymbol{x}, \boldsymbol{q}) := -\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^\mathsf{T} \mathbf{Q}_{0,\boldsymbol{q}}^{-1} \mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i, \boldsymbol{q}))\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q})); \boldsymbol{q})]. \tag{E.1}$$

Theorem E.1 (Bahadur representation)**.** *Suppose Assumptions B.1 to B.6 hold. Furthermore, assume the following conditions:*

(i) $\sup_{\boldsymbol{q}\in\mathcal{Q}}\left\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})\right\|_\infty = o_\mathbb{P}(1)$;

(ii) there exists a constant $c > 0$ *such that* $\{\boldsymbol{\beta} \in \mathbb{R}^K : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(\boldsymbol{q})\|_\infty \le c, \boldsymbol{q} \in \mathcal{Q}\} \subseteq \mathcal{B}$;

(iii) $\frac{\log^{d+2} n}{nh^d} = o(1)$;

(iv) either $\frac{(h^{-1}\log n)^{\frac{\nu}{\nu-2}d}}{n} = o(1)$, *or* $\overline{\psi}(\boldsymbol{x}_i, y_i)$ *is* σ^2*-sub-Gaussian conditionally on* $\boldsymbol{x}_i$;

(v) either $\theta \mapsto \rho(y, \eta(\theta); \boldsymbol{q})$ *is convex with left or right derivative* $\psi(y, \eta(\theta); \boldsymbol{q})\eta^{(1)}(\theta)$ *and* $\mathcal{B} = \mathbb{R}^K$, *or the additional complexity Assumption B.8 holds.*

(a) Then

$$\sup_{\boldsymbol{q}\in\mathcal{Q},\boldsymbol{x}\in\mathcal{X}}\left|\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x}, \boldsymbol{q}) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x}, \boldsymbol{q}) - \mathsf{L}^{(\boldsymbol{v})}(\boldsymbol{x}, \boldsymbol{q})\right| \lesssim_\mathbb{P} h^{-|\boldsymbol{v}|} r_{\mathsf{BR}} \tag{E.2}$$

with

$$r_{\mathsf{BR}} := \left(\frac{\log^d n}{nh^d}\right)^{1/2+(\alpha/2\wedge 1/4)} \log n + h^{(\alpha\wedge 1/2)m}\left(\frac{\log^{d+1} n}{nh^d}\right)^{1/2} + h^m. \tag{E.3}$$

(b) If, in addition to the previous conditions, (B.3) holds (without any restrictions on y), then

$$\sup_{\boldsymbol{q}\in\mathcal{Q},\boldsymbol{x}\in\mathcal{X}}\left|\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})-\mu_0^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})-\mathsf{L}^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})\right|\lesssim_{\mathbb{P}} h^{-|\boldsymbol{v}|}\tilde{r}_{\mathtt{BR}} \tag{E.4}$$

with

$$\tilde{r}_{\mathtt{BR}}:=\left(\frac{\log^d n}{nh^d}\right)^{(1+\alpha)/2}\log n+h^{\alpha m}\left(\frac{\log^{d+1} n}{nh^d}\right)^{1/2}+h^m. \tag{E.5}$$

Remark E.2. As will be clear in the proof, the matrix $\mathbf{Q}_{0,\boldsymbol{q}}^{-1}$ in Eqs. (E.2) and (E.4) can be replaced by $\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}$.

E.1 Proof: convex case

We will now prove Theorem E.1 under the assumption that $\theta\mapsto\rho(y,\eta(\theta);\boldsymbol{q})$ is convex with left or right derivative $\psi(y,\eta(\theta);\boldsymbol{q})\eta^{(1)}(\theta)$ and $\mathcal{B}=\mathbb{R}^K$. We only show the (a) part, since the argument for (b) is very similar with minor changes in obvious places.

Notation In this proof, we will denote

$$\mathbb{G}_n^i[g(\boldsymbol{x}_i,y_i)]:=\sqrt{n}\mathbb{E}_n\big[g(\boldsymbol{x}_i,y_i)-\mathbb{E}[g(\boldsymbol{x}_i,y_i)\,|\,\boldsymbol{x}_i]\big].$$

E.1.1 Strategy

By $\sup_{\boldsymbol{x}}|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})|\lesssim h^{-|\boldsymbol{v}|}$ (Assumption B.2), $|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q})-\mu_0^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})|\lesssim h^{m-|\boldsymbol{v}|}$ (Assumption B.6), and Lemma C.14, the inequality

$$\begin{aligned}\sup_{\boldsymbol{q}\in\mathcal{Q},\boldsymbol{x}\in\mathcal{X}}&\big|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}}\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\mu_0^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})\\&+\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}}\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}));\boldsymbol{q})]\big|\\&\lesssim_{\mathbb{P}} h^{-|\boldsymbol{v}|}r_{\mathtt{BR}}\end{aligned} \tag{E.6}$$

is implied by

$$\sup_{\boldsymbol{q}}\big\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\beta}_0(\boldsymbol{q})-\bar{\boldsymbol{\beta}}_{\boldsymbol{q}}\big\|_\infty\lesssim_{\mathbb{P}}\left(\frac{\log^d n}{nh^d}\right)^{1/2+(\alpha/2\wedge 1/4)}\log n+h^{(\alpha\wedge 1/2)m}\left(\frac{\log^{d+1} n}{nh^d}\right)^{1/2}, \tag{E.7}$$

where we put

$$\begin{aligned}&\boldsymbol{p}_i:=\boldsymbol{p}(\boldsymbol{x}_i),\\&\bar{\boldsymbol{\beta}}_{\boldsymbol{q}}:=-\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\mathbb{E}_n\big[\boldsymbol{p}_i\eta^{(1)}(\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}))\psi(y_i,\eta(\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\big]\end{aligned} \tag{E.8}$$

to declutter notation, and noted that for $h<1$

$$h^{(\alpha\wedge 1/2)m}\left(\frac{\log^{d+1} n}{nh^d}\right)^{1/2}\lesssim\left(\frac{\log^d n}{nh^d}\right)^{1/2+(\alpha/2\wedge 1/4)}\log n+h^m. \tag{E.9}$$

Indeed, if $\alpha\le 1/2$, either $\left(\frac{\log^{d+1} n}{nh^d}\right)^{1/2}\le h^m$, in which case (for $h<1$)

$$h^{\alpha m}\left(\frac{\log^{d+1} n}{nh^d}\right)^{1/2}\le\left(\frac{\log^{d+1} n}{nh^d}\right)^{1/2}\le h^m,$$

or $\left(\frac{\log^{d+1} n}{nh^d}\right)^{1/2} > h^m$, in which case

$$h^{\alpha m}\left(\frac{\log^{d+1} n}{nh^d}\right)^{1/2} \leq \left(\frac{\log^{d+1} n}{nh^d}\right)^{(1+\alpha)/2} \leq \left(\frac{\log^{d} n}{nh^d}\right)^{(1+\alpha)/2} \log n.$$

If $\alpha > 1/2$, (E.9) follows from $ab \lesssim a^2 + b^2$.

From (E.6), (E.2) follows, because $\sup_{\boldsymbol{x}} \|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\|_1 \lesssim h^{-|\boldsymbol{v}|}$ and

$$\begin{aligned}
&\sup_{\boldsymbol{q}} \big\|(\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1} - \mathbf{Q}_{0,\boldsymbol{q}}^{-1})\mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}));\boldsymbol{q})]\big\|_\infty \\
&\quad \lesssim \sup_{\boldsymbol{q}} \|\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1} - \mathbf{Q}_{0,\boldsymbol{q}}^{-1}\|_\infty \sup_{\boldsymbol{q}} \big\|\mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}));\boldsymbol{q})]\big\|_\infty \\
&\quad \overset{\text{(a)}}{\lesssim_{\mathbb{P}}} h^{-d}\left(\frac{\log(1/h)}{nh^d}\right)^{1/2}\left(\frac{h^d \log(1/h)}{n}\right)^{1/2} = \frac{\log(1/h)}{nh^d} = o(r_{\mathtt{BR}}),
\end{aligned} \tag{E.10}$$

where (a) is by Lemma C.11 and Eq. (C.12). So, we will be showing (E.7).

Note that for any vector $\boldsymbol{\alpha}$

$$\mathbb{E}_n\big[\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}_i\psi(y_i,\eta(\boldsymbol{p}_i^\mathsf{T}(\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\alpha}));\boldsymbol{q})\eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}(\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\alpha}))\big] \leq 0, \tag{E.11}$$

because $g(\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\alpha}) := \mathbb{E}_n\big[\psi(y_i,\eta(\boldsymbol{p}_i^\mathsf{T}(\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\alpha}));\boldsymbol{q})\eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}(\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\alpha}))\boldsymbol{p}_i\big]$ is a subgradient of the function $f(\boldsymbol{\beta}) := \mathbb{E}_n\big[\rho(y_i,\eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta});\boldsymbol{q})\big]$ at $\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\alpha}$, and $\widehat{\boldsymbol{\beta}}(\boldsymbol{q})$ is the minimizer of this function, giving $\boldsymbol{\alpha}^\mathsf{T} g\big(\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\alpha}\big) \leq f(\widehat{\boldsymbol{\beta}}(\boldsymbol{q})) - f(\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\alpha}) \leq 0$. Apply (E.11) with a particular vector $\boldsymbol{\alpha}_{\boldsymbol{q}}$ that will be chosen later, and decompose

$$\begin{aligned}
0 &\geq \mathbb{E}_n\big[\boldsymbol{\alpha}_{\boldsymbol{q}}^\mathsf{T}\boldsymbol{p}_i\psi(y_i,\eta(\boldsymbol{p}_i^\mathsf{T}(\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\alpha}_{\boldsymbol{q}}));\boldsymbol{q})\eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}(\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\alpha}_{\boldsymbol{q}}))\big] \\
&= \mathbb{E}_n\big[\boldsymbol{\alpha}_{\boldsymbol{q}}^\mathsf{T}\boldsymbol{p}_i\{\psi(y_i,\eta(\boldsymbol{p}_i^\mathsf{T}(\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\alpha}_{\boldsymbol{q}}));\boldsymbol{q}) - \psi(y_i,\eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\}\eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}(\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\alpha}_{\boldsymbol{q}}))\big] \\
&\quad + T_1 + \mathbb{E}_n\big[\boldsymbol{\alpha}_{\boldsymbol{q}}^\mathsf{T}\boldsymbol{p}_i\psi(y_i,\eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}))\big],
\end{aligned} \tag{E.12}$$

where

$$\begin{aligned}
T_1 &:= \mathbb{E}_n\big[\boldsymbol{\alpha}_{\boldsymbol{q}}^\mathsf{T}\boldsymbol{p}_i\psi(y_i,\eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}(\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\alpha}_{\boldsymbol{q}}))\big] \\
&\quad - \mathbb{E}_n\big[\boldsymbol{\alpha}_{\boldsymbol{q}}^\mathsf{T}\boldsymbol{p}_i\psi(y_i,\eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}))\big]
\end{aligned} \tag{E.13}$$

will be bounded later. Define for simplicity

$$\begin{aligned}
\delta_{\boldsymbol{q},i}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\alpha}) &:= \boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}_i\big\{\psi(y_i,\eta(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{\alpha}));\boldsymbol{q}) - \psi(y_i,\eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\big\} \\
&\qquad \times \eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{\alpha})),
\end{aligned}$$

so that

$$\text{(E.12)} = \mathbb{E}_n\big[\delta_{\boldsymbol{q},i}(\bar{\boldsymbol{\beta}}_{\boldsymbol{q}},\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\beta}_0(\boldsymbol{q})-\bar{\boldsymbol{\beta}}_{\boldsymbol{q}},\boldsymbol{\alpha}_{\boldsymbol{q}})\big].$$

Now, add and subtract the conditional mean of this term, continuing

$$0 \geq n^{-1/2}\mathbb{G}_n^i\big[\delta_{\boldsymbol{q},i}(\bar{\boldsymbol{\beta}}_{\boldsymbol{q}},\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\beta}_0(\boldsymbol{q})-\bar{\boldsymbol{\beta}}_{\boldsymbol{q}},\boldsymbol{\alpha}_{\boldsymbol{q}})\big] \tag{E.14}$$

$$+ \mathbb{E}_n\big[\mathbb{E}\big[\delta_{\boldsymbol{q},i}(\bar{\boldsymbol{\beta}}_{\boldsymbol{q}},\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\beta}_0(\boldsymbol{q})-\bar{\boldsymbol{\beta}}_{\boldsymbol{q}},\boldsymbol{\alpha}_{\boldsymbol{q}}) \,\big|\, \boldsymbol{x}_i\big]\big] \tag{E.15}$$

$$+ T_1 + \mathbb{E}_n\big[\boldsymbol{\alpha}_{\boldsymbol{q}}^\mathsf{T}\boldsymbol{p}_i\psi(y_i,\eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}))\big].$$

The difference (E.14) will be shown to be small by the usual concentration argument. Using Taylor expansion, we will show that the term in (E.15) is close to $\boldsymbol{\alpha}_{\boldsymbol{q}}^\mathsf{T}\widetilde{\mathbf{Q}}_{\boldsymbol{q}}(\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\beta}_0(\boldsymbol{q}))$ with the matrix

$$\widetilde{\mathbf{Q}}_{\boldsymbol{q}} := \mathbb{E}_n\big[\boldsymbol{p}_i\boldsymbol{p}_i^\mathsf{T}\Psi_1(\boldsymbol{x}_i,\eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}))^2\big], \tag{E.16}$$

that is, we will bound

$$T_2 := \mathbb{E}_n\big[\mathbb{E}\big[\delta_{\boldsymbol{q},i}(\bar{\boldsymbol{\beta}}_{\boldsymbol{q}}, \widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q}) - \bar{\boldsymbol{\beta}}_{\boldsymbol{q}}, \boldsymbol{\alpha}_{\boldsymbol{q}}) \,\big|\, \boldsymbol{x}_i\big]\big] - \boldsymbol{\alpha}_{\boldsymbol{q}}^{\mathsf{T}} \widetilde{\mathbf{Q}}_{\boldsymbol{q}}(\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})). \tag{E.17}$$

To deal with the remaining term $\boldsymbol{\alpha}_{\boldsymbol{q}}^{\mathsf{T}} \widetilde{\mathbf{Q}}_{\boldsymbol{q}}(\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q}))$, start with replacing the approximation $\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q})$ in the definition of $\widetilde{\mathbf{Q}}_{\boldsymbol{q}}$ with the true function $\mu_0(\boldsymbol{x}_i, \boldsymbol{q})$ leaving us with $\bar{\mathbf{Q}}_{\boldsymbol{q}}$; the error introduced by this operation

$$T_3 := \boldsymbol{\alpha}_{\boldsymbol{q}}^{\mathsf{T}}(\widetilde{\mathbf{Q}}_{\boldsymbol{q}} - \bar{\boldsymbol{Q}}_{\boldsymbol{q}})(\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})) \tag{E.18}$$

will be bounded. Next, write

$$\begin{aligned}
&\boldsymbol{\alpha}_{\boldsymbol{q}}^{\mathsf{T}} \bar{\mathbf{Q}}_{\boldsymbol{q}}(\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})) + \mathbb{E}_n\big[\boldsymbol{\alpha}_{\boldsymbol{q}}^{\mathsf{T}} \boldsymbol{p}_i \psi(y_i, \eta(\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q}) \eta^{(1)}(\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}))\big] \\
&\quad = \boldsymbol{\alpha}_{\boldsymbol{q}}^{\mathsf{T}} \bar{\mathbf{Q}}_{\boldsymbol{q}}(\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q}) - \bar{\boldsymbol{\beta}}_{\boldsymbol{q}}).
\end{aligned}$$

We obtained

$$0 \geq n^{-1/2}\mathbb{G}_n^i\big[\delta_{\boldsymbol{q},i}(\bar{\boldsymbol{\beta}}_{\boldsymbol{q}}, \widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q}) - \bar{\boldsymbol{\beta}}_{\boldsymbol{q}}, \boldsymbol{\alpha}_{\boldsymbol{q}})\big] + T_1 + T_2 + T_3 + \boldsymbol{\alpha}_{\boldsymbol{q}}^{\mathsf{T}} \bar{\mathbf{Q}}_{\boldsymbol{q}}(\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q}) - \bar{\boldsymbol{\beta}}_{\boldsymbol{q}}). \tag{E.19}$$

At this point, it would be convenient if we could choose $\boldsymbol{\alpha}_{\boldsymbol{q}}$ so that the last term $\boldsymbol{\alpha}_{\boldsymbol{q}}^{\mathsf{T}} \bar{\mathbf{Q}}_{\boldsymbol{q}}(\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q}) - \bar{\boldsymbol{\beta}}_{\boldsymbol{q}})$ is proportional (with a positive coefficient) to $\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q}) - \bar{\boldsymbol{\beta}}_{\boldsymbol{q}}\|_\infty$, because then in combination with all the other terms being small in absolute value, we can conclude that $\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q}) - \bar{\boldsymbol{\beta}}_{\boldsymbol{q}}\|_\infty$ is small in absolute value as well (otherwise it is impossible to obtain a nonnegative quantity). This is essentially what we will do, with one caveat: for the bounds on the other terms to work out, it is helpful if $\boldsymbol{\alpha}_{\boldsymbol{q}}$ is very sparse. Following these considerations, we introduce another vector $\bar{\boldsymbol{\alpha}}_{\boldsymbol{q}}$ proportional to $\big[\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\big]_{\boldsymbol{k},k} \operatorname{sign}\big((\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q}) - \bar{\boldsymbol{\beta}}_{\boldsymbol{q}})_{\boldsymbol{k},k}\big)$ (with a positive coefficient), where $(\boldsymbol{k}, k)$ is the index of the largest component of $\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q}) - \bar{\boldsymbol{\beta}}_{\boldsymbol{q}}$ (in the blockwise index notation as in Lemma C.11), and $\big[\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\big]_{\boldsymbol{k},k}$ is the $(\boldsymbol{k}, k)$th row of $\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}$: this way, $\bar{\boldsymbol{\alpha}}_{\boldsymbol{q}}^{\mathsf{T}} \bar{\mathbf{Q}}_{\boldsymbol{q}}(\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q}) - \bar{\boldsymbol{\beta}}_{\boldsymbol{q}})$ is indeed proportional to $\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q}) - \bar{\boldsymbol{\beta}}_{\boldsymbol{q}}\|_\infty$. Then we choose $\boldsymbol{\alpha}_{\boldsymbol{q}}$ to be a sparse vector that is close enough to $\bar{\boldsymbol{\alpha}}_{\boldsymbol{q}}$ for an appropriate bound on

$$T_4 := (\boldsymbol{\alpha}_{\boldsymbol{q}} - \bar{\boldsymbol{\alpha}}_{\boldsymbol{q}})^{\mathsf{T}} \bar{\boldsymbol{Q}}_{\boldsymbol{q}}(\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q}) - \bar{\boldsymbol{\beta}}_{\boldsymbol{q}}), \tag{E.20}$$

which can be done because of the structure of the rows of $\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}$, specifically the exponential decay in Eq. (C.9).

E.1.2 Main argument

We will now give some specifics. To show a precise bound in probability in (E.2) (as opposed to a $o_{\mathbb{P}}(\cdot)$ bound), it will be convenient to multiply $r_{\mathtt{BR}}$ by another positive sequence γ_n that arbitrarily slowly diverges to infinity:

$$\mathfrak{r}_{2,n} := r_{\mathtt{BR}}\gamma_n. \tag{E.21}$$

We will also put

$$r_{1,n} := \left[\left(\frac{\log^d n}{nh^d}\right)^{1/2} + h^m\right]\gamma_n^{1/(1 + \alpha \wedge (1/2))}, \tag{E.22}$$

so that in particular

$$\left(\frac{\log^d n}{nh^d}\right)^{1/2} + h^m = o(r_{1,n}) \tag{E.23}$$

and by Lemmas C.12 and C.14, we have on the event $\mathcal{A}_0$ of $1-o(1)$ probability

$$\sup_{\boldsymbol{q}\in\mathcal{Q}}\|\bar{\boldsymbol{\beta}}_{\boldsymbol{q}}\|_\infty \le r_{1,n}. \tag{E.24}$$

Let γ_n diverge slowly enough so that $r_{1,n}+\mathfrak{r}_{2,n}=o(1)$; in particular, $\sup_{\boldsymbol{q}\in\mathcal{Q}}\|\bar{\boldsymbol{\beta}}_{\boldsymbol{q}}\|_\infty$ is smaller than any positive constant with probability approaching one.

Fix two small enough constants $c_1, c_2>0$ (the restrictions on which will be discussed later). Since c_1 is a constant, with probability approaching one both $\sup_{\boldsymbol{q}\in\mathcal{Q}}\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\beta}_0(\boldsymbol{q})\|_\infty \le c_1/2$ (by consistency) and $\sup_{\boldsymbol{q}\in\mathcal{Q}}\|\bar{\boldsymbol{\beta}}_{\boldsymbol{q}}\|_\infty \le c_1/2$ (as just discussed); hence, with probability approaching one $\sup_{\boldsymbol{q}\in\mathcal{Q}}\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\beta}_0(\boldsymbol{q})-\bar{\boldsymbol{\beta}}_{\boldsymbol{q}}\|_\infty \le c_1$. This means that the probability of the event $\bigcap_{\boldsymbol{q}\in\mathcal{Q}}\{\widehat{\boldsymbol{\beta}}(\boldsymbol{q})\in\mathcal{O}_{\boldsymbol{q}}\}$ approaches one, where we use the partitioning

$$\mathcal{O}_{\boldsymbol{q}} := \bigcup_{\ell=-\infty}^{\bar{L}_n}\mathcal{O}_{\boldsymbol{q},\ell}, \quad \mathcal{O}_{\boldsymbol{q},\ell} := \left\{\boldsymbol{\beta}\in\mathbb{R}^K : 2^{\ell-1}\mathfrak{r}_{2,n} \le \|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\boldsymbol{q})-\bar{\boldsymbol{\beta}}_{\boldsymbol{q}}\|_\infty < 2^\ell\mathfrak{r}_{2,n}\right\}$$

with $\bar{L}_n$ defined as the smallest integer such that $2^{\bar{L}_n}\mathfrak{r}_{2,n}\ge c_1$. (The sequence $\bar{L}_n$ diverges to infinity.)

Put

$$\bar{\boldsymbol{\alpha}}_{\boldsymbol{q}} = c_2 2^L\mathfrak{r}_{2,n}h^d\left[\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\right]_{\boldsymbol{k},k}\operatorname{sign}\left((\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\beta}_0(\boldsymbol{q})-\bar{\boldsymbol{\beta}}_{\boldsymbol{q}})_{\boldsymbol{k},k}\right),$$

where $(\boldsymbol{k},k)$ is such an index that $|(\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\beta}_0(\boldsymbol{q})-\bar{\boldsymbol{\beta}}_{\boldsymbol{q}})_{\boldsymbol{k},k}| = \|\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\beta}_0(\boldsymbol{q})-\bar{\boldsymbol{\beta}}_{\boldsymbol{q}}\|_\infty$ (as we already discussed above) and L is a constant integer chosen later. The vector $\bar{\boldsymbol{\alpha}}_{\boldsymbol{q}}$ is not sparse, but the components decay exponentially with the "distance" to the multi-index $(\boldsymbol{k},k)$ according to Eq. (C.9). Therefore, we can zero out all components except a logarithmic neighborhood around $(\boldsymbol{k},k)$ of this vector and control the error introduced by this operation: specifically, take $\boldsymbol{\alpha}_{\boldsymbol{q}}\in\mathbb{R}^K$ with components $v_{\boldsymbol{q},\boldsymbol{j},j}=\bar{v}_{\boldsymbol{q},\boldsymbol{j},j}$ for $\|\boldsymbol{j}-\boldsymbol{k}\|_\infty \le c_3\log n$ and zero otherwise, where c_3 is some constant. On the event

$$\mathcal{A}_1 := \left\{\sup_{\delta\in\Delta}\mathbb{E}_n[\mathbb{1}(\boldsymbol{x}_i\in\delta)] \le C_{11}h^d\right\} \tag{E.25}$$

of $1-o(1)$ probability (for some large enough constant $C_{11}>0$), we have

$$\|\bar{\mathbf{Q}}_{\boldsymbol{q}}\|_\infty \lesssim h^d,$$

so on $\mathcal{A}_1\cap\left\{\widehat{\boldsymbol{\beta}}(\boldsymbol{q})\in\mathcal{O}_{\boldsymbol{q},\ell}\right\}$

$$|T_4| \lesssim (2^L\mathfrak{r}_{2,n})h^d(2^\ell\mathfrak{r}_{2,n})n^{-c_4} = o(h^d 2^{L+\ell}\mathfrak{r}_{2,n}^2),$$

where $c_4>0$ is a constant (depending on the constant controlling the neighborhood size c_3).

Bounding T_1, T_2, T_3 is deferred to Lemmas E.6 to E.8. Specifically, on an event $\mathcal{A}:=\mathcal{A}_0\cap\mathcal{A}_1\cap\mathcal{A}_2$ with probability $1-o(1)$, using Lemma E.6 (where $\mathcal{A}_2$ is defined) and the restriction

$$r_{1,n}^2 = o(\mathfrak{r}_{2,n}), \tag{E.26}$$

we bound

$$|T_1| = O(h^d r_{1,n}(r_{1,n}+2^\ell\mathfrak{r}_{2,n}+2^L\mathfrak{r}_{2,n})2^L\mathfrak{r}_{2,n}) = o(h^d 2^{L+\ell}\mathfrak{r}_{2,n}^2).$$

Using Lemma E.7, we bound also on $\left\{\widehat{\boldsymbol{\beta}}(\boldsymbol{q})\in\mathcal{O}_{\boldsymbol{q},\ell}\right\}\cap\mathcal{A}$ with $L\le\ell\le\bar{L}_n$:

$$|T_2| \le \frac{c_2}{8}h^d 2^{L+\ell}\mathfrak{r}_{2,n}^2. \tag{E.27}$$

Indeed,

$$\begin{aligned}2^L\mathfrak{r}_{2,n}h^d\big(r_{1,n}+2^\ell\mathfrak{r}_{2,n}+2^L\mathfrak{r}_{2,n}\big)^2 &\lesssim 2^L\mathfrak{r}_{2,n}h^d r_{1,n}^2+2^L\mathfrak{r}_{2,n}h^d2^{2\ell}\mathfrak{r}_{2,n}^2\\ &\overset{\text{(a)}}{=} 2^L\mathfrak{r}_{2,n}h^d2^{2\ell}\mathfrak{r}_{2,n}^2+o(h^d2^L\mathfrak{r}_{2,n}^2),\end{aligned}$$

where in (a) we used $r_{1,n}^2=o(\mathfrak{r}_{2,n})$ again. Since $2^\ell\mathfrak{r}_{2,n}\le 2c_1$, we can therefore ensure

$$\big|T_2+\boldsymbol{\alpha}_{\boldsymbol{q}}^\top\widetilde{\mathbf{Q}}_{\boldsymbol{q}}\boldsymbol{\alpha}_{\boldsymbol{q}}\big|\le\frac{c_2}{16}h^d2^{L+\ell}\mathfrak{r}_{2,n}^2$$

by taking c_1 sufficiently small relative to c_2. Next, taking c_2 sufficiently small (in particular, making c_2^2 much smaller than c_2), we can upper-bound $\lambda_{\max}(\widetilde{\mathbf{Q}}_{\boldsymbol{q}})\|\boldsymbol{\alpha}_{\boldsymbol{q}}\|^2$ by $(c_2/16)h^d2^{L+\ell}\mathfrak{r}_{2,n}^2$ as well, giving (E.27).

Finally, using Lemma E.8, we bound on $\mathcal{A}$

$$|T_3|=O(h^{m+d}(2^L\mathfrak{r}_{2,n})(r_{1,n}+2^\ell\mathfrak{r}_{2,n}))\overset{\text{(a)}}{=}o(h^d2^{L+\ell}\mathfrak{r}_{2,n}^2),$$

where in (a) we used

$$h^m r_{1,n}=o(\mathfrak{r}_{2,n}). \tag{E.28}$$

Combining, we have

$$|T_1|+|T_2|+|T_3|+|T_4|\le\frac{c_2}{4}h^d2^{L+\ell}\mathfrak{r}_{2,n}^2.$$

Since

$$\bar{\boldsymbol{\alpha}}_{\boldsymbol{q}}^\top\bar{\boldsymbol{Q}}_{\boldsymbol{q}}(\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\beta}_0(\boldsymbol{q})-\bar{\boldsymbol{\beta}}_{\boldsymbol{q}})=c_22^L\mathfrak{r}_{2,n}h^d\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\beta}_0(\boldsymbol{q})-\bar{\boldsymbol{\beta}}_{\boldsymbol{q}}\|_\infty\ge c_22^L\mathfrak{r}_{2,n}h^d2^{\ell-1}\mathfrak{r}_{2,n},$$

we conclude from (E.19) that on $\big\{\widehat{\boldsymbol{\beta}}(\boldsymbol{q})\in\mathcal{O}_{\boldsymbol{q},\ell}\big\}\cap\mathcal{A}$, with $L\le\ell\le\bar{L}_n$,

$$\begin{aligned}-n^{-1/2}\mathbb{G}_n^i\big[\delta_{\boldsymbol{q},i}(\bar{\boldsymbol{\beta}}_{\boldsymbol{q}},\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\beta}_0(\boldsymbol{q})-\bar{\boldsymbol{\beta}}_{\boldsymbol{q}},\boldsymbol{\alpha}_{\boldsymbol{q}})\big]&\ge c_22^L\mathfrak{r}_{2,n}h^d2^{\ell-1}\mathfrak{r}_{2,n}-\frac{c_2}{4}h^d2^{L+\ell}\mathfrak{r}_{2,n}^2\\ &=\frac{c_2}{4}h^d2^{L+\ell}\mathfrak{r}_{2,n}^2.\end{aligned}$$

We will prove that the probability of this event is small enough, by using a concentration argument.

Lemma E.3 (Uniform concentration)**.** *Define*

$$\begin{aligned}\mathcal{V}&:=\big\{\boldsymbol{\alpha}\in\mathbb{R}^K:\exists\ d\text{-dimensional multi-index }\boldsymbol{k},\\ &\qquad |v_{\boldsymbol{\ell},l}|\le\varrho^{\|\boldsymbol{k}-\boldsymbol{l}\|_\infty}2^L\mathfrak{r}_{2,n}\text{ for }\|\boldsymbol{k}-\boldsymbol{l}\|_\infty\le c_3\log n\text{ and }v_{\boldsymbol{\ell},l}=0\text{ otherwise}\big\},\\ \mathcal{H}_1&:=\{\boldsymbol{\beta}\in\mathbb{R}^K:\|\boldsymbol{\beta}\|_\infty\le r_{1,n}\},\\ \mathcal{H}_{2,\ell}&:=\{\boldsymbol{\beta}\in\mathbb{R}^K:\|\boldsymbol{\beta}\|_\infty\le 2^\ell\mathfrak{r}_{2,n}\},\end{aligned}$$

where ϱ is the constant from Lemma C.11. On the event $\mathcal{A}_1$ defined in Eq. (E.25) *we have*

$$\begin{aligned}&\mathbb{E}\Big[\sup_{\boldsymbol{q}\in\mathcal{Q},\boldsymbol{\beta}_1\in\mathcal{H}_1,\boldsymbol{\beta}_2\in\mathcal{H}_{2,\ell},\boldsymbol{\alpha}\in\mathcal{V}}\big|\mathbb{E}_n[\delta_{\boldsymbol{q},i}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\alpha})-\mathbb{E}[\delta_{\boldsymbol{q},i}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\alpha})|\boldsymbol{x}_i]]\big|\ \Big|\ \{\boldsymbol{x}_i\}_{i=1}^n\Big]\\ &\quad\le C_{12}\bigg(\frac{h^{d/2}2^L\mathfrak{r}_{2,n}(r_{1,n}+2^\ell\mathfrak{r}_{2,n}+2^L\mathfrak{r}_{2,n})^{\alpha\wedge(1/2)}}{\sqrt{n}}\log^{(d+1)/2}n+\frac{2^L\mathfrak{r}_{2,n}\log^{d+1}n}{n}\bigg),\end{aligned}\tag{E.29}$$

where the constant C_{12} does not depend on L or ℓ.

Appendix E.1.3 is devoted to the proof of this fact.

On $\{\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) \in \mathcal{O}_{\boldsymbol{q},\ell}\} \cap \mathcal{A}$, we have (as long as c_2 is small enough)

$$\boldsymbol{\alpha}_{\boldsymbol{q}} \in \mathcal{V}, \quad \bar{\boldsymbol{\beta}}_{\boldsymbol{q}} \in \mathcal{H}_1, \quad \widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q}) - \bar{\boldsymbol{\beta}}_{\boldsymbol{q}} \in \mathcal{H}_{2,\ell}.$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\Big\{\sup_{\boldsymbol{q}\in\mathcal{Q}}\big\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q}) - \bar{\boldsymbol{\beta}}_{\boldsymbol{q}}\big\|_\infty \geq 2^L \mathfrak{r}_{2,n} \,\Big|\, \{\boldsymbol{x}_i\}_{i=1}^n\Big\}\\
&\quad= \mathbb{P}\Big\{\bigcap_{\boldsymbol{q}\in\mathcal{Q}} \bigcup_{\ell=L+1}^{\infty} \{\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) \in \mathcal{O}_{\boldsymbol{q},\ell}\} \,\Big|\, \{\boldsymbol{x}_i\}_{i=1}^n\Big\}\\
&\quad= \mathbb{P}\Big\{\bigcap_{\boldsymbol{q}\in\mathcal{Q}} \bigcup_{\ell=L+1}^{\bar{L}_n} \{\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) \in \mathcal{O}_{\boldsymbol{q},\ell}\} \cap \mathcal{A} \,\Big|\, \{\boldsymbol{x}_i\}_{i=1}^n\Big\} + o_{\mathbb{P}}(1)\\
&\quad\leq \mathbb{P}\Big\{\bigcup_{\ell=L+1}^{\bar{L}_n} \Big\{\sup_{\boldsymbol{q}\in\mathcal{Q},\boldsymbol{\beta}_1\in\mathcal{H}_1,\boldsymbol{\beta}_2\in\mathcal{H}_{2,\ell},\boldsymbol{\alpha}\in\mathcal{V}} \big|n^{-1/2}\mathbb{G}_n^i\big[\delta_{\boldsymbol{q},i}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\alpha})\big]\big| \geq \frac{c_2}{4} h^d 2^{L+\ell}\mathfrak{r}_{2,n}^2\Big\} \,\Big|\, \{\boldsymbol{x}_i\}_{i=1}^n\Big\}\mathbb{1}(\mathcal{A}_1) + o_{\mathbb{P}}(1)\\
&\quad\leq \sum_{\ell=L+1}^{\bar{L}_n} \mathbb{P}\Big\{\sup_{\boldsymbol{q}\in\mathcal{Q},\boldsymbol{\beta}_1\in\mathcal{H}_1,\boldsymbol{\beta}_2\in\mathcal{H}_{2,\ell},\boldsymbol{\alpha}\in\mathcal{V}} \big|n^{-1/2}\mathbb{G}_n^i\big[\delta_{\boldsymbol{q},i}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\alpha})\big]\big| \geq \frac{c_2}{4} h^d 2^{L+\ell}\mathfrak{r}_{2,n}^2 \,\Big|\, \{\boldsymbol{x}_i\}_{i=1}^n\Big\}\mathbb{1}(\mathcal{A}_1) + o_{\mathbb{P}}(1)\\
&\quad\leq \sum_{\ell=L+1}^{\bar{L}_n} \Big(\frac{c_2}{4} h^d 2^{L+\ell}\mathfrak{r}_{2,n}^2\Big)^{-1} \mathbb{E}\Big[\sup_{\boldsymbol{q}\in\mathcal{Q},\boldsymbol{\beta}_1\in\mathcal{H}_1,\boldsymbol{\beta}_2\in\mathcal{H}_{2,\ell},\boldsymbol{\alpha}\in\mathcal{V}} \big|n^{-1/2}\mathbb{G}_n^i\big[\delta_{\boldsymbol{q},i}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\alpha})\big]\big| \,\Big|\, \{\boldsymbol{x}_i\}_{i=1}^n\Big]\mathbb{1}(\mathcal{A}_1) + o_{\mathbb{P}}(1)\\
&\quad\leq \sum_{\ell=L+1}^{\bar{L}_n} C_{12}\Big(\frac{c_2}{4} h^d 2^{L+\ell}\mathfrak{r}_{2,n}^2\Big)^{-1} \bigg(\frac{h^{d/2}2^L\mathfrak{r}_{2,n}(r_{1,n} + 2^\ell\mathfrak{r}_{2,n} + 2^L\mathfrak{r}_{2,n})^{\alpha\wedge(1/2)}}{\sqrt{n}}\log^{(d+1)/2} n + \frac{2^L\mathfrak{r}_{2,n}\log^{d+1} n}{n}\bigg)\\
&\quad\lesssim \sum_{\ell=L+1}^{\bar{L}_n} 2^{-\ell}\mathfrak{r}_{2,n}^{-1}\bigg(\frac{r_{1,n}^{\alpha\wedge(1/2)} + \big(2^\ell\mathfrak{r}_{2,n}\big)^{\alpha\wedge(1/2)}}{\sqrt{nh^d}}\log^{(d+1)/2} n + \frac{\log^{d+1} n}{nh^d}\bigg),\\
&\quad\overset{\text{(a)}}{\lesssim} \sum_{\ell=L+1}^{\bar{L}_n} \big(2^{-\ell} + 2^{\alpha\wedge(1/2)-\ell}\big) \lesssim \sum_{\ell=L+1}^{\infty} 2^{-\ell} = 2^{-L},
\end{aligned}$$

where in (a) we used

$$\frac{r_{1,n}^{\alpha\wedge(1/2)}}{\sqrt{nh^d}}\log^{(d+1)/2} n \lesssim \mathfrak{r}_{2,n}, \tag{E.30}$$

$$\frac{\log^{(d+1)/2} n}{\sqrt{nh^d}} \lesssim \mathfrak{r}_{2,n}^{1-\alpha\wedge(1/2)}, \tag{E.31}$$

$$\frac{\log^{d+1} n}{nh^d} \lesssim \mathfrak{r}_{2,n}. \tag{E.32}$$

We conclude that

$$\sup_{\boldsymbol{q}\in\mathcal{Q}}\big\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q}) - \bar{\boldsymbol{\beta}}_{\boldsymbol{q}}\big\|_\infty \lesssim_{\mathbb{P}} \mathfrak{r}_{2,n}.$$

Since γ_n was arbitrarily slowly diverging, it follows that in fact

$$\sup_{\boldsymbol{q}\in\mathcal{Q}}\big\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q}) - \bar{\boldsymbol{\beta}}_{\boldsymbol{q}}\big\|_\infty \lesssim_{\mathbb{P}} r_{\mathtt{BR}}.$$

E.1.3 Concentration argument (proof of Lemma E.3)

We will first argue that $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ can be considered effectively low-dimensional (of polylog n dimension).

Note that $\delta_{\boldsymbol{q},i}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\alpha}) \neq 0$ only if $\boldsymbol{\alpha}^\mathsf{T} \boldsymbol{p}_i \neq 0$. For each $\boldsymbol{\alpha} \in \mathcal{V}$, let $\mathcal{J}_{\boldsymbol{\alpha}} := \{j : v_j \neq 0\}$. By construction, the cardinality of $\mathcal{J}_{\boldsymbol{\alpha}}$ is bounded by $(2c_3 \log n + 1)^d$. We have $\delta_{\boldsymbol{q},i}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\alpha}) \neq 0$ only if $p_j(\boldsymbol{x}_i) \neq 0$ for some $j \in \mathcal{J}_{\boldsymbol{\alpha}}$, which happens only if $\boldsymbol{x}_i \in \mathcal{I}_{\boldsymbol{\alpha}}$ where

$$\mathcal{I}_{\boldsymbol{\alpha}} := \bigcup \{\delta \in \Delta \colon \delta \cap \operatorname{supp} p_j \neq \varnothing \text{ for some } j \in \mathcal{J}_{\boldsymbol{\alpha}}\}.$$

The family $\mathcal{I}_{\boldsymbol{\alpha}}$ includes at most $c_5(c_3 \log n)^d$ cells. Moreover, at most $c_6(c_3 \log n)^d$ basis functions in $\boldsymbol{p}$ have supports overlapping with $\mathcal{I}_{\boldsymbol{\alpha}}$. Denote the set of indices of such basis functions by $\bar{\mathcal{J}}_{\boldsymbol{\alpha}}$. Based on the above observations, we have $\delta_{\boldsymbol{q},i}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\alpha}) = \delta_{\boldsymbol{q},i}(\boldsymbol{\beta}_{1,\bar{\mathcal{J}}_{\boldsymbol{\alpha}}}, \boldsymbol{\beta}_{2,\bar{\mathcal{J}}_{\boldsymbol{\alpha}}}, \boldsymbol{\alpha})$, where

$$\begin{aligned}
&\delta_{\boldsymbol{q},i}(\boldsymbol{\beta}_{1,\bar{\mathcal{J}}_{\boldsymbol{\alpha}}}, \boldsymbol{\beta}_{2,\bar{\mathcal{J}}_{\boldsymbol{\alpha}}}, \boldsymbol{\alpha}) \\
&\quad := \sum_{j \in \mathcal{J}_{\boldsymbol{\alpha}}} p_{i,j} v_j \Bigg[\psi\Bigg(y_i, \eta\Bigg(\sum_{l \in \bar{\mathcal{J}}_{\boldsymbol{\alpha}}} p_{i,l}(\beta_{0,\boldsymbol{q},l} + \beta_{1,l} + \beta_{2,l}) - \sum_{j \in \mathcal{J}_{\boldsymbol{\alpha}}} p_{i,j} v_j\Bigg); \boldsymbol{q}\Bigg) \\
&\qquad - \psi\Bigg(y_i, \eta\Bigg(\sum_{l \in \bar{\mathcal{J}}_{\boldsymbol{\alpha}}} p_{i,l}\beta_{0,\boldsymbol{q},l}\Bigg); \boldsymbol{q}\Bigg)\Bigg] \\
&\qquad \times \eta^{(1)}\Bigg(\sum_{l \in \bar{\mathcal{J}}_{\boldsymbol{\alpha}}} p_{i,l}(\beta_{0,\boldsymbol{q},l} + \beta_{1,l} + \beta_{2,l}) - \sum_{j \in \mathcal{J}_{\boldsymbol{\alpha}}} p_{i,j} v_j\Bigg) \mathbb{1}\{\boldsymbol{x}_i \in \mathcal{I}_{\boldsymbol{\alpha}}\}.
\end{aligned} \tag{E.33}$$

Accordingly, for $\widetilde{\boldsymbol{\beta}}_1 \in \mathbb{R}^{c_6(c_3 \log n)^d}$, $\widetilde{\boldsymbol{\beta}}_2 \in \mathbb{R}^{c_6(c_3 \log n)^d}$, define the function class

$$\mathcal{G} := \big\{(\boldsymbol{x}_i, y_i) \mapsto \delta_{\boldsymbol{q},i}\big(\widetilde{\boldsymbol{\beta}}_1, \widetilde{\boldsymbol{\beta}}_2, \boldsymbol{\alpha}\big) : \boldsymbol{q} \in \mathcal{Q}, \boldsymbol{\alpha} \in \mathcal{V}, \big\|\widetilde{\boldsymbol{\beta}}_1\big\|_\infty \leq r_{1,n}, \big\|\widetilde{\boldsymbol{\beta}}_2\big\|_\infty \leq 2^\ell \mathfrak{r}_{2,n}\big\}.$$

We will now bound $\sup_{g \in \mathcal{G}} |\mathbb{E}_n[g(\boldsymbol{x}_i, y_i)] - \mathbb{E}\big[g(\boldsymbol{x}_i, y_i) \,\big|\, \boldsymbol{x}_i\big]|$. As usual, the strategy is to check conditions of Lemma C.6.

Lemma E.4 (Bonding variance)**.** *On $\mathcal{A}_1$, class $\mathcal{G}$ satisfies the following variance bound:*

$$\sup_{g \in \mathcal{G}} \mathbb{E}_n[\mathbb{V}[g(\boldsymbol{x}_i, y_i) \mid \boldsymbol{x}_i]] \lesssim 2^{2L} \mathfrak{r}_{2,n}^2 h^d (r_{1,n} + 2^\ell \mathfrak{r}_{2,n} + 2^L \mathfrak{r}_{2,n})^{(2\alpha) \wedge 1}.$$

Proof. We will now proceed under the assumption that $\mathfrak{M}$ is Lebesgue measure, so (B.2) holds; the argument under (B.3) is similar (and leads to an even stronger variance bound), so it is omitted.

By the same argument as in the proof of Lemma C.14, using $\big|\psi(y, \eta(p(\boldsymbol{x})^\mathsf{T} \boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q})\big| \lesssim \overline{\psi}(\boldsymbol{x}, y) + 1$, for y_i *outside* the closed segment between the two points

$$\eta\Bigg(\sum_{l \in \bar{\mathcal{J}}_v} p_{i,l}\beta_{0,\boldsymbol{q},l}\Bigg), \quad \text{and} \quad \eta\Bigg(\sum_{l \in \bar{\mathcal{J}}_v} p_{i,l}(\beta_{0,\boldsymbol{q},l} + \beta_{1,l} + \beta_{2,l}) - \sum_{j \in \mathcal{J}_v} p_{i,j} v_j\Bigg),$$

we have

$$\begin{aligned}
&\Bigg|\psi\Bigg(y_i, \eta\Bigg(\sum_{l \in \bar{\mathcal{J}}_v} p_{i,l}(\beta_{0,\boldsymbol{q},l} + \beta_{1,l} + \beta_{2,l}) - \sum_{j \in \mathcal{J}_v} p_{i,j} v_j\Bigg); \boldsymbol{q}\Bigg) - \psi\Bigg(y_i, \eta\Bigg(\sum_{l \in \bar{\mathcal{J}}_v} p_{i,l}\beta_{0,\boldsymbol{q},l}\Bigg); \boldsymbol{q}\Bigg)\Bigg| \\
&\quad \lesssim \big(\overline{\psi}(\boldsymbol{x}_i, y_i) + 1\big)(r_{1,n} + 2^\ell \mathfrak{r}_{2,n} + 2^L \mathfrak{r}_{2,n})^\alpha,
\end{aligned}$$

and for y_i *in* this segment we have

$$\Big|\psi\Big(y_i,\eta\Big(\sum_{l\in\bar{\mathcal{J}}_v} p_{i,l}(\beta_{0,\boldsymbol{q},l}+\beta_{1,l}+\beta_{2,l})-\sum_{j\in\mathcal{J}_v} p_{i,j}v_j\Big);\boldsymbol{q}\Big)-\psi\Big(y_i,\eta\Big(\sum_{l\in\bar{\mathcal{J}}_v} p_{i,l}\beta_{0,\boldsymbol{q},l}\Big);\boldsymbol{q}\Big)\Big|$$
$$\lesssim(\overline{\psi}(\boldsymbol{x}_i,y_i)+1)(r_{1,n}+2^{\ell}\mathfrak{r}_{2,n}+2^{L}\mathfrak{r}_{2,n})+1$$

uniformly over $\boldsymbol{q}$.

By construction, for each $\boldsymbol{\alpha}\in\mathcal{V}$, there exists some $\boldsymbol{k}_{\boldsymbol{\alpha}}$ such that $|v_{\boldsymbol{l},l}|\le\varrho^{\|\boldsymbol{l}-\boldsymbol{k}_{\boldsymbol{\alpha}}\|_\infty}2^L\mathfrak{r}_{2,n}$ if $\|\boldsymbol{l}-\boldsymbol{k}_{\boldsymbol{\alpha}}\|_\infty\le M_n$, and otherwise $v_{\boldsymbol{l},l}=0$. The above facts imply(cf. the proof of Lemma C.14) that for any $\boldsymbol{x}_i\in\delta\subset\mathcal{I}_{\boldsymbol{\alpha}}$,

$$\mathbb{V}[\delta_{\boldsymbol{q},i}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\alpha})\,|\,\boldsymbol{x}_i]\lesssim 2^{2L}\mathfrak{r}_{2,n}^2(r_{1,n}+2^{\ell}\mathfrak{r}_{2,n}+2^{L}\mathfrak{r}_{2,n})^{(2\alpha)\wedge 1}\sum_{(\boldsymbol{l},l)\in\mathcal{L}_\delta}\varrho^{2\|\boldsymbol{l}-\boldsymbol{k}_{\boldsymbol{\alpha}}\|_\infty}$$
$$\text{for}\quad \mathcal{L}_\delta:=\{(\boldsymbol{l},l)\colon\ \operatorname{supp}p_{\boldsymbol{l},l}\cap\delta\neq\varnothing\}.$$

In addition, since $\delta_{\boldsymbol{q},i}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\alpha})\neq 0$ only if $\boldsymbol{x}_i\in\mathcal{I}_{\boldsymbol{\alpha}}$, for all $g\in\mathcal{G}$,

$$\begin{aligned}\mathbb{E}_n[\mathbb{V}[g(\boldsymbol{x}_i,y_i)\,|\,\boldsymbol{x}_i]]&\lesssim 2^{2L}\mathfrak{r}_{2,n}^2(r_{1,n}+2^{\ell}\mathfrak{r}_{2,n}+2^{L}\mathfrak{r}_{2,n})^{(2\alpha)\wedge 1}\sum_{\delta\subset\mathcal{I}_{\boldsymbol{\alpha}}}\mathbb{E}_n[\mathbb{1}(\boldsymbol{x}_i\in\delta)]\sum_{(\boldsymbol{l},l)\in\mathcal{L}_\delta}\varrho^{2\|\boldsymbol{l}-\boldsymbol{k}_{\boldsymbol{\alpha}}\|_\infty}\\&=2^{2L}\mathfrak{r}_{2,n}^2(r_{1,n}+2^{\ell}\mathfrak{r}_{2,n}+2^{L}\mathfrak{r}_{2,n})^{(2\alpha)\wedge 1}\sum_{\boldsymbol{l},l}\varrho^{2\|\boldsymbol{l}-\boldsymbol{k}_{\boldsymbol{\alpha}}\|_\infty}\sum_{\delta\in\mathcal{L}^\star_{\boldsymbol{l},l}}\mathbb{E}_n[\mathbb{1}(\boldsymbol{x}_i\in\delta)]\end{aligned}$$
$$\text{for}\quad\mathcal{L}^\star_{\boldsymbol{l},l}:=\{\delta\subset\mathcal{I}_{\boldsymbol{\alpha}}\colon\ \operatorname{supp}p_{\boldsymbol{l},l}\cap\delta\neq\varnothing\}.$$

Note that $\mathcal{L}^\star_{\boldsymbol{l},l}$ contains a bounded number of elements. Then on $\mathcal{A}_1$,

$$\begin{aligned}\sup_{g\in\mathcal{G}}\mathbb{E}_n[\mathbb{V}[g(\boldsymbol{x}_i,y_i)\,|\,\boldsymbol{x}_i]]&\lesssim 2^{2L}\mathfrak{r}_{2,n}^2h^d(r_{1,n}+2^{\ell}\mathfrak{r}_{2,n}+2^{L}\mathfrak{r}_{2,n})^{(2\alpha)\wedge 1}\sum_{\boldsymbol{l},l}\varrho^{2\|\boldsymbol{l}-\boldsymbol{k}_{\boldsymbol{\alpha}}\|_\infty}\\&\lesssim 2^{2L}\mathfrak{r}_{2,n}^2h^d(r_{1,n}+2^{\ell}\mathfrak{r}_{2,n}+2^{L}\mathfrak{r}_{2,n})^{(2\alpha)\wedge 1}\sum_{\boldsymbol{l}}\varrho^{2\|\boldsymbol{l}-\boldsymbol{k}_{\boldsymbol{\alpha}}\|_\infty}\qquad\text{since } l \text{ is bounded}\\&\lesssim 2^{2L}\mathfrak{r}_{2,n}^2h^d(r_{1,n}+2^{\ell}\mathfrak{r}_{2,n}+2^{L}\mathfrak{r}_{2,n})^{(2\alpha)\wedge 1}\sum_{\boldsymbol{t}\in\mathbb{Z}^d}\varrho^{2\|\boldsymbol{t}\|_\infty}\lesssim 2^{2L}\mathfrak{r}_{2,n}^2h^d(r_{1,n}+2^{\ell}\mathfrak{r}_{2,n}+2^{L}\mathfrak{r}_{2,n})^{(2\alpha)\wedge 1},\end{aligned}$$

concluding the proof of Lemma E.4. □

Lemma E.5 (Complexity of class $\mathcal{G}$)**.** *Class $\mathcal{G}$ with envelope $2^L\mathfrak{r}_{2,n}$ multiplied by a large enough constant satisfies the uniform entropy bound (A.3) with $A\lesssim 1$ and $V\lesssim\log^d n$.*

Proof of Lemma E.5. First, indeed $\sup_{\boldsymbol{x},y}\sup_{g\in\mathcal{G}}|g(\boldsymbol{x},y)|\lesssim 2^L\mathfrak{r}_{2,n}$.

Next, the class of functions $\mathcal{W}_1:=\{(\boldsymbol{x}_i,y_i)\mapsto\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}(\boldsymbol{x}_i)\colon\boldsymbol{\alpha}\in\mathcal{V}\}$ is a union of $O(h^{-d})$ classes $\mathcal{W}_{1,\boldsymbol{k}}:=\{(\boldsymbol{x}_i,y_i)\mapsto\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}(\boldsymbol{x}_i)\colon\boldsymbol{\alpha}\in\mathcal{V}_{\boldsymbol{k}}\}$, where

$$\mathcal{V}_{\boldsymbol{k}}=\big\{\boldsymbol{\alpha}\in\mathbb{R}^K:|v_{\boldsymbol{\ell},l}|\le\varrho^{\|\boldsymbol{k}-\boldsymbol{\ell}\|_\infty}2^L\mathfrak{r}_{2,n}\text{ for }\|\boldsymbol{k}-\boldsymbol{\ell}\|_\infty\le M_n\text{ and }v_{\boldsymbol{\ell},l}=0\text{ otherwise}\big\}.$$

Since $\mathcal{W}_{1,\boldsymbol{k}}$ is a subclass of a vector space of functions of dimension $O(\log^d n)$, by Lemma 2.6.15 in [41] it is VC with index $O(\log^d n)$. This implies that $\mathcal{W}_{1,\boldsymbol{k}}$ with envelope $O(2^L\mathfrak{r}_{2,n})$ satisfies the uniform entropy bound (A.3) with $A\lesssim 1$ and $V\lesssim\log^d n$. Since there are $O(h^{-d})$ such classes and $\log(1/h)\lesssim\log n$, using the chain

$$O(h^{-d})\Big(\frac{A}{\varepsilon}\Big)^{O(\log^d n)}=e^{O(\log n)}\Big(\frac{A}{\varepsilon}\Big)^{O(\log^d n)}\le\Big(\frac{A}{\varepsilon}\Big)^{O(\log n)+O(\log^d n)}=\Big(\frac{A}{\varepsilon}\Big)^{O(\log^d n)}\tag{E.34}$$

(recall that $A \geq e$), we get that $\mathcal{W}_1$ also satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim \log^d n$.

By Assumption B.5, the class of functions

$$\begin{aligned}\mathcal{W}_2 := \big\{&(\boldsymbol{x}_i, y_i) \mapsto \\ &[\psi(y_i, \eta(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{\alpha})); \boldsymbol{q}) - \psi(y_i, \eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q})]\mathbb{1}\{\boldsymbol{x}_i \in \mathcal{I}_{\boldsymbol{\alpha}}\}\colon \\ &\boldsymbol{\beta}_1 \in \mathcal{H}_1, \boldsymbol{\beta}_2 \in \mathcal{H}_{2,\ell}, \boldsymbol{\alpha} \in \mathcal{V}, \boldsymbol{q} \in \mathcal{Q}\big\}\end{aligned}$$

with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim \log^d n$.

The class of functions

$$\begin{aligned}\mathcal{W}_3 := \big\{&(\boldsymbol{x}_i, y_i) \mapsto \eta^{(1)}(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{\alpha}))\mathbb{1}\{\boldsymbol{x}_i \in \mathcal{I}_{\boldsymbol{\alpha}}\}\colon \\ &\boldsymbol{\beta}_1 \in \mathcal{H}_1, \boldsymbol{\beta}_2 \in \mathcal{H}_{2,\ell}, \boldsymbol{\alpha} \in \mathcal{V}, \boldsymbol{q} \in \mathcal{Q}\big\}\end{aligned}$$

is a subset of the union over $\delta \in \Delta$ of classes (for some fixed positive constants c and r, n large enough)

$$\mathcal{W}_{3,\delta} := \big\{(\boldsymbol{x}_i, y_i) \mapsto \eta^{(1)}(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\boldsymbol{\beta})\mathbb{1}\{\boldsymbol{x}_i \in \mathcal{N}_{[c\log n]}(\delta)\}\colon \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(\boldsymbol{q})\|_\infty \leq r, \boldsymbol{q} \in \mathcal{Q}\big\}.$$

Note that $\boldsymbol{\beta}$ can be assumed to lie in a fixed vector space $\mathcal{B}_\delta$ of dimension $\dim \mathcal{B}_\delta = O(\log^d n)$. Again applying Lemma 2.6.15 in [41] and noting that $\eta^{(1)}$ on a fixed (bounded) interval is Lipschitz, we have that $\mathcal{W}_{3,\delta}$ with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim \log^d n$. Similarly to the argument for $\mathcal{W}_1$, this implies that the same is true for $\mathcal{W}_3$.

Applying Lemma C.4 concludes the proof of Lemma E.5. $\square$

Proof of Lemma E.3. Apply Lemma C.6 conditionally on $\{\boldsymbol{x}_i\}_{i=1}^n$ on $\mathcal{A}_1$. $\square$

E.1.4 Bounding T_1, T_2, T_3

Lemma E.6 (Bounding T_1)**.** *There exists an event $\mathcal{A}_2$ whose probability converges to one such that on $\mathcal{A}_2$*

$$\begin{aligned}\sup_{\boldsymbol{q} \in \mathcal{Q}, \boldsymbol{\beta}_1 \in \mathcal{H}_1, \boldsymbol{\beta}_2 \in \mathcal{H}_{2,\ell}, \boldsymbol{\alpha} \in \mathcal{V}} &|\mathbb{E}_n[\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}_i\psi(y_i, \eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q})\eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{\alpha}))] \\ &- \mathbb{E}_n[\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}_i\psi(y_i, \eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q})\eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}))]| \lesssim h^d r_{1,n}(r_{1,n} + 2^\ell \mathfrak{r}_{2,n} + 2^L \mathfrak{r}_{2,n})2^L \mathfrak{r}_{2,n}.\end{aligned}$$

Proof of Lemma E.6. Note that

$$\begin{aligned}&\mathbb{E}_n\big[\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}_i\psi(y_i, \eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q})\eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{\alpha}))\big] \\ &\quad - \mathbb{E}_n[\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}_i\psi(y_i, \eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q})\eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}))] \\ &\quad = \boldsymbol{\alpha}^\mathsf{T}\mathbb{E}_n[\psi(y_i, \eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q})\eta^{(2)}(\xi_{\boldsymbol{q},i})\boldsymbol{p}_i\boldsymbol{p}_i^\mathsf{T}](\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{\alpha}) \\ &\quad \lesssim h^d r_{1,n}(r_{1,n} + 2^\ell \mathfrak{r}_{2,n} + 2^L \mathfrak{r}_{2,n})2^L \mathfrak{r}_{2,n},\end{aligned}$$

where $\xi_{\boldsymbol{q},i}$ is between $\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{\alpha})$ and $\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})$. The bound holds on the event

$$\mathcal{A}_2 := \big\{\sup\big\|\mathbb{E}_n[\psi(y_i, \eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q})\eta^{(2)}(\xi_{\boldsymbol{q},i})\boldsymbol{p}_i\boldsymbol{p}_i^\mathsf{T}]\big\|_\infty \leq h^d r_{1,n}\big\},$$

where the supremum is over $\boldsymbol{\beta}_1 \in \mathcal{H}_1$, $\boldsymbol{\beta}_2 \in \mathcal{H}_{2,\ell}$, $\boldsymbol{\alpha} \in \mathcal{V}$, $\boldsymbol{q} \in \mathcal{Q}$ and $\xi_{\boldsymbol{q},i}$ between $\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{\alpha})$ and $\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})$. By the same argument as Lemmas C.12 and C.14, $\mathbb{P}\{\mathcal{A}_2\} \to 1$. $\square$

Lemma E.7 (Bounding T_2)**.** *On $\mathcal{A}_1$, we have*

$$\sup_{\boldsymbol{q}\in\mathcal{Q},\boldsymbol{\beta}_1\in\mathcal{H}_1,\boldsymbol{\beta}_2\in\mathcal{H}_{2,\ell},\boldsymbol{\alpha}\in\mathcal{V}}\left|\mathbb{E}_n[\mathbb{E}[\delta_{\boldsymbol{q},i}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\alpha})|\boldsymbol{x}_i]]-\boldsymbol{\alpha}^\mathsf{T}\widetilde{\mathbf{Q}}_{\boldsymbol{q}}(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+\boldsymbol{\alpha}^\mathsf{T}\widetilde{\mathbf{Q}}_{\boldsymbol{q}}\boldsymbol{\alpha}\right| \\ \lesssim 2^L\mathfrak{r}_{2,n}h^d(r_{1,n}+2^\ell\mathfrak{r}_{2,n}+2^L\mathfrak{r}_{2,n})^2,$$

and in addition for all $\boldsymbol{q}\in\mathcal{Q}$, $\boldsymbol{\alpha}\in\mathcal{V}$

$$|\boldsymbol{\alpha}^\mathsf{T}\widetilde{\mathbf{Q}}_{\boldsymbol{q}}\boldsymbol{\alpha}|\lesssim h^d\|\boldsymbol{\alpha}\|^2.$$

Proof of Lemma E.7. First, on $\mathcal{A}_1$ the largest eigenvalue of $\widetilde{\mathbf{Q}}_{\boldsymbol{q}}$ is bounded by h^d up to a constant factor (uniformly in $\boldsymbol{q}$):

$$\begin{aligned}
\lambda_{\max}\big(\widetilde{\mathbf{Q}}_{\boldsymbol{q}}\big) &= \sup_{\|\boldsymbol{\alpha}\|=1}\boldsymbol{\alpha}^\mathsf{T}\widetilde{\mathbf{Q}}_{\boldsymbol{q}}\boldsymbol{\alpha}=\sup_{\|\boldsymbol{\alpha}\|=1}\mathbb{E}_n\big[(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}_i)^2\Psi_1(\boldsymbol{x}_i,\eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}))^2\big]\\
&\le \sup_{\|\boldsymbol{\alpha}\|=1}\mathbb{E}_n\big[(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}_i)^2\big|\Psi_1(\boldsymbol{x}_i,\eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\big|\eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}))^2\big]\\
&\lesssim \sup_{\|\boldsymbol{\alpha}\|=1}\mathbb{E}_n\big[(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}_i)^2\big]
\end{aligned}$$

(because $\big|\Psi_1(\boldsymbol{x}_i,\eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\big|\eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}))^2\lesssim 1$ by Assumptions B.4 and B.6)

$$\begin{aligned}
&= \sup_{\|\boldsymbol{\alpha}\|=1}\sum_{l=1}^{\kappa}\mathbb{E}_n\big[(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}_i)^2\mathbb{1}\{\boldsymbol{x}_i\in\delta_l\}\big]\\
&\lesssim \sup_{\|\boldsymbol{\alpha}\|=1}\sum_{l=1}^{\kappa}\left(\sum_{k=1}^{M_l}\alpha_{l,k}^2\right)\mathbb{E}_n\big[\mathbb{1}\{\boldsymbol{x}_i\in\delta_l\}\big]
\end{aligned}$$

(because $\sup_{\boldsymbol{x}\in\delta_l}(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}(\boldsymbol{x}))^2\lesssim\sum_{k=1}^{M_l}\alpha_{l,k}^2$, where $\{\alpha_{l,k}\}_{k=1}^{M_l}$ are the components of $\boldsymbol{\alpha}$ corresponding to the M_l basis functions supported on δ_l)

$$\begin{aligned}
&\le \sup_{\delta\in\Delta}\mathbb{E}_n\big[\mathbb{1}\{\boldsymbol{x}_i\in\delta\}\big]\sup_{\|\boldsymbol{\alpha}\|=1}\sum_{l=1}^{\kappa}\left(\sum_{k=1}^{M_l}\alpha_{l,k}^2\right)\\
&\lesssim \sup_{\delta\in\Delta}\mathbb{E}_n\big[\mathbb{1}\{\boldsymbol{x}_i\in\delta\}\big]\sup_{\|\boldsymbol{\alpha}\|=1}\|\boldsymbol{\alpha}\|^2=\sup_{\delta\in\Delta}\mathbb{E}_n\big[\mathbb{1}\{\boldsymbol{x}_i\in\delta\}\big].
\end{aligned}$$

Next, by Taylor expansion,

$$\begin{aligned}
&\mathbb{E}[\delta_{\boldsymbol{q},i}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\alpha})|\boldsymbol{x}_i]\\
&\quad=\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}_i[\Psi(\boldsymbol{x}_i,\eta(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{\alpha}));\boldsymbol{q})-\Psi(\boldsymbol{x}_i,\eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})]\\
&\qquad\times\eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{\alpha}))\\
&\quad=\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}_i\big[\Psi_1(\boldsymbol{x}_i,\eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\big\{\eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}))\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{\alpha})+\\
&\qquad\quad+(1/2)\eta^{(2)}(\xi_{\boldsymbol{q},i})(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{\alpha}))^2\big\}\\
&\qquad+(1/2)\Psi_2(\boldsymbol{x}_i,\tilde{\xi}_{\boldsymbol{q},i};\boldsymbol{q})\big\{\eta(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{\alpha}))-\eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}))\big\}^2\big]\\
&\qquad\times\eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{\alpha}))
\end{aligned}$$

for some $\xi_{\boldsymbol{q},i}$ between $\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})$ and $\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{\alpha})$, $\tilde{\xi}_{\boldsymbol{q},i}$ is between $\eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}))$ and $\eta(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{\alpha}))$. This gives

$$\mathbb{E}_n[\mathbb{E}[\delta_{\boldsymbol{q},i}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\alpha})|\boldsymbol{x}_i]]=\boldsymbol{\alpha}^\mathsf{T}\widetilde{\mathbf{Q}}_{\boldsymbol{q}}(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)-\boldsymbol{\alpha}^\mathsf{T}\widetilde{\mathbf{Q}}_{\boldsymbol{q}}\boldsymbol{\alpha}+\mathrm{I}+\mathrm{II}+\mathrm{III},$$

where for some $\check{\xi}_{\boldsymbol{q},i}$ between $\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})$ and $\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{\alpha})$

$$
\begin{aligned}
\mathrm{I} &:= \mathbb{E}_n\big[\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}_i\Psi_1(\boldsymbol{x}_i, \eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q})\eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}))\eta^{(2)}(\check{\xi}_{\boldsymbol{q},i})(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{\alpha}))^2\big],\\
\mathrm{II} &:= \frac{1}{2}\mathbb{E}_n\big[\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}_i\Psi_1(\boldsymbol{x}_i, \eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q})\eta^{(2)}(\xi_{\boldsymbol{q},i})\\
&\quad \times \eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{\alpha}))(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{\alpha}))^2\big],\\
\mathrm{III} &:= \frac{1}{2}\mathbb{E}_n\big[\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}_i\Psi_2(\boldsymbol{x}_i, \tilde{\xi}_{\boldsymbol{q},i}; \boldsymbol{q})\{\eta(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{\alpha})) - \eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}))\}^2\\
&\qquad \times \eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{\alpha}))\big]
\end{aligned}
$$

and

$$
\begin{aligned}
\mathrm{I} &\lesssim 2^L\mathfrak{r}_{2,n}h^d(r_{1,n} + 2^\ell\mathfrak{r}_{2,n} + 2^L\mathfrak{r}_{2,n})^2,\\
\mathrm{II} &\lesssim 2^L\mathfrak{r}_{2,n}h^d(r_{1,n} + 2^\ell\mathfrak{r}_{2,n} + 2^L\mathfrak{r}_{2,n})^2,\\
\mathrm{III} &\lesssim 2^L\mathfrak{r}_{2,n}h^d(r_{1,n} + 2^\ell\mathfrak{r}_{2,n} + 2^L\mathfrak{r}_{2,n})^2
\end{aligned}
$$

on the event $\mathcal{A}_1$. □

Lemma E.8 (Bounding T_3)**.** *For the matrix $\widetilde{\mathbf{Q}}_{\boldsymbol{q}}$ defined in Eq. (E.16), we have the following bound on the event $\mathcal{A}_1$:*

$$
\sup_{\boldsymbol{q}\in\mathcal{Q},\boldsymbol{\beta}_1\in\mathcal{H}_1,\boldsymbol{\beta}_2\in\mathcal{H}_{2,\ell},\boldsymbol{\alpha}\in\mathcal{V}}\big|\boldsymbol{\alpha}^\mathsf{T}(\widetilde{\mathbf{Q}}_{\boldsymbol{q}} - \bar{\mathbf{Q}}_{\boldsymbol{q}})(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2)\big| \lesssim h^{m+d}2^L\mathfrak{r}_{2,n}(r_{1,n} + 2^\ell\mathfrak{r}_{2,n}). \tag{E.35}
$$

Proof of Lemma E.8. By the same logic as in Lemma C.11, we have on $\mathcal{A}_1$

$$
\|\bar{\mathbf{Q}}_{\boldsymbol{q}} - \widetilde{\mathbf{Q}}_{\boldsymbol{q}}\|_\infty \vee \|\bar{\mathbf{Q}}_{\boldsymbol{q}} - \widetilde{\mathbf{Q}}_{\boldsymbol{q}}\| \lesssim h^m h^d
$$

uniformly over $\boldsymbol{q}$ with probability approaching one. This gives (E.35), proving Lemma E.8. □

We have now proved the deferred lemmas, and the proof of Theorem E.1 is concluded.

Remark E.9 (Rate restrictions)*.* The rates in the proof are determined by four restrictions: Eqs. (E.23), (E.26), (E.30) and (E.31). Equation (E.32) follows from Eqs. (E.23) and (E.30). Equation (E.28) follows from Eq. (E.23) and Eq. (E.26).

E.2 Proof: general case

We will state another version of the Bahadur Representation theorem, where $\theta \mapsto \rho(y, \eta(\theta); \boldsymbol{q})$ is not assumed to be convex. Here, we will use the additional complexity Assumption B.8.

The argument is almost the same as for Theorem E.1, so we will only describe the changes that need to be made.

The setup is the same as in the convex case except the definition of $\delta_{\boldsymbol{q},i}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\alpha})$ is replaced with

$$
\begin{aligned}
&\delta_{\boldsymbol{q},i}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\alpha})\\
&\quad := \rho(y_i, \eta(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2)); \boldsymbol{q}) - \rho(y_i, \eta(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{\alpha})); \boldsymbol{q})\\
&\qquad - \big[\eta(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2)) - \eta(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{\alpha}))\big]\psi(y_i, \eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q})\\
&\quad = \int_{-\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\alpha}}^0 \big[\psi(y_i, \eta(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + t); \boldsymbol{q}) - \psi(y_i, \eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q})\big]\\
&\qquad \times \eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + t)\,\mathrm{d}t,
\end{aligned}
$$

and (E.33) is changed to fit this definition.

Instead of (E.11), we have for any vector $\boldsymbol{\alpha}$

$$\mathbb{E}_n\big[\rho(y_i, \eta(\boldsymbol{p}_i^\mathsf{T}\widehat{\boldsymbol{\beta}}(\boldsymbol{q})); \boldsymbol{q}) - \rho(y_i, \eta(\boldsymbol{p}_i^\mathsf{T}(\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\alpha})); \boldsymbol{q})\big] \le 0$$

by the definition of $\widehat{\boldsymbol{\beta}}(\boldsymbol{q})$ as long as $\|\boldsymbol{\alpha}\|_\infty$ is small enough (so that $\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\alpha}$ satisfies the constraints). Accordingly, the definition of T_1 becomes

$$\begin{aligned}T_1 &:= \mathbb{E}_n\big[(\eta(\boldsymbol{p}_i^\mathsf{T}\widehat{\boldsymbol{\beta}}(\boldsymbol{q})) - \eta(\boldsymbol{p}_i(\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\alpha}_{\boldsymbol{q}})))\psi(y_i, \eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q})\big]\\ &\quad - \mathbb{E}_n\big[\boldsymbol{\alpha}_{\boldsymbol{q}}^\mathsf{T}\boldsymbol{p}_i\psi(y_i, \eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q})\eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}))\big].\end{aligned}$$

Lemma E.10 (Bounding variance)**.** *On $\mathcal{A}_1$ as in Lemma E.4, class $\mathcal{G}$ satisfies the following variance bound:*

$$\sup_{g\in\mathcal{G}} \mathbb{E}_n[\mathbb{V}[g(\boldsymbol{x}_i, y_i) \mid \boldsymbol{x}_i]] \lesssim 2^{2L}\mathfrak{r}_{2,n}^2 h^d(r_{1,n} + 2^\ell\mathfrak{r}_{2,n} + 2^L\mathfrak{r}_{2,n})^{(2\alpha)\wedge 1}.$$

Proof of Lemma E.10. This is proven by the same argument as in the proof of Theorem E.1. ☐

Lemma E.11 (Complexity of class $\mathcal{G}$)**.** *Class $\mathcal{G}$ with envelope $2^L\mathfrak{r}_{2,n}$ multiplied by a large enough constant satisfies the uniform entropy bound (A.3) with $A \lesssim (2^L\mathfrak{r}_{2,n})^{-1}$ and $V \lesssim \log^d n$.*

Note that A is not constant in this statement but it will not matter since $\log((2^L\mathfrak{r}_{2,n})^{-1}) \lesssim \log n$.

Proof of Lemma E.11. This is a directly assumed in Assumption B.8. ☐

Lemma E.12 (Uniform concentration in $\mathcal{G}$)**.** *On the event $\mathcal{A}_1$, (E.29) holds.*

Proof of Lemma E.12. This is proven by the same argument as in the proof of Theorem E.1. ☐

Lemma E.13. *For $\widetilde{\mathbf{Q}}_{\boldsymbol{q}} := \mathbb{E}_n\big[\boldsymbol{p}_i\boldsymbol{p}_i^\mathsf{T}\Psi_1(\boldsymbol{x}_i, \eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})); \boldsymbol{q})\eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}))^2\big]$, (E.35) holds.*

Proof of Lemma E.13. This is proven by the same argument as in the proof of Theorem E.1. ☐

Lemma E.14. *On $\mathcal{A}_1$, we have*

$$\begin{aligned}&\sup_{\boldsymbol{q}\in\mathcal{Q},\boldsymbol{\beta}_1\in\mathcal{H}_1,\boldsymbol{\beta}_2\in\mathcal{H}_{2,\ell},\boldsymbol{\alpha}\in\mathcal{V}} \big|\mathbb{E}_n\big[\mathbb{E}[\delta_{\boldsymbol{q},i}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\alpha})|\boldsymbol{x}_i]\big] - \boldsymbol{\alpha}^\mathsf{T}\widetilde{\mathbf{Q}}_{\boldsymbol{q}}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + \boldsymbol{\alpha}^\mathsf{T}\widetilde{\mathbf{Q}}_{\boldsymbol{q}}\boldsymbol{\alpha}\big|\\ &\quad\lesssim 2^L\mathfrak{r}_{2,n}h^d(r_{1,n} + 2^\ell\mathfrak{r}_{2,n} + 2^L\mathfrak{r}_{2,n})^2,\end{aligned}$$

and in addition for all $\boldsymbol{q} \in \mathcal{Q}$, $\boldsymbol{\alpha} \in \mathcal{V}$

$$|\boldsymbol{\alpha}^\mathsf{T}\widetilde{\mathbf{Q}}_{\boldsymbol{q}}\boldsymbol{\alpha}| \lesssim h^d\|\boldsymbol{\alpha}\|^2.$$

Proof of Lemma E.14. First, on $\mathcal{A}_1$ the largest eigenvalue of $\widetilde{\mathbf{Q}}_{\boldsymbol{q}}$ is bounded by h^d up to a constant factor (uniformly in $\boldsymbol{q}$), which is proven in Lemma E.7.

Next, by the Taylor expansion,

$$
\begin{aligned}
&\mathbb{E}[\delta_{\boldsymbol{q},i}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\alpha})|\boldsymbol{x}_i] \\
&\quad = \int_{-\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\alpha}}^{0} \big[\Psi(\boldsymbol{x}_i,\eta(\boldsymbol{p}_i^{\mathsf{T}}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+t);\boldsymbol{q}) - \Psi(\boldsymbol{x}_i,\eta(\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\big] \\
&\qquad \times \eta^{(1)}(\boldsymbol{p}_i^{\mathsf{T}}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+t)\,\mathrm{d}t \\
&\quad = \int_{-\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\alpha}}^{0} \big[\Psi_1(\boldsymbol{x}_i,\eta(\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\{\eta^{(1)}(\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}))(\boldsymbol{p}_i^{\mathsf{T}}(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+t) \\
&\qquad\quad + \frac{1}{2}\eta^{(2)}(\xi_{\boldsymbol{q},i,t})(\boldsymbol{p}_i^{\mathsf{T}}(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+t)^2\} \\
&\qquad + \frac{1}{2}\Psi_2(\boldsymbol{x}_i,\tilde{\xi}_{\boldsymbol{q},i,t};\boldsymbol{q})\{\eta(\boldsymbol{p}_i^{\mathsf{T}}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+t) - \eta(\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}))\}^2\big] \\
&\qquad \times \eta^{(1)}(\boldsymbol{p}_i^{\mathsf{T}}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+t)\,\mathrm{d}t
\end{aligned}
$$

for some $\xi_{\boldsymbol{q},i,t}$ between $\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q})$ and $\boldsymbol{p}_i^{\mathsf{T}}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+t$, $\tilde{\xi}_{\boldsymbol{q},i,t}$ between $\eta(\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}))$ and $\eta(\boldsymbol{p}_i^{\mathsf{T}}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+t)$. This gives

$$
\mathbb{E}_n[\mathbb{E}[\delta_{\boldsymbol{q},i}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\alpha})|\boldsymbol{x}_i]] = \boldsymbol{\alpha}^{\mathsf{T}}\widetilde{\mathbf{Q}}_{\boldsymbol{q}}(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2) - \boldsymbol{\alpha}^{\mathsf{T}}\widetilde{\mathbf{Q}}_{\boldsymbol{q}}\boldsymbol{\alpha} + \mathrm{I} + \mathrm{II} + \mathrm{III},
$$

where for some $\check{\xi}_{\boldsymbol{q},i,t}$ between $\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q})$ and $\boldsymbol{p}_i^{\mathsf{T}}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+t$ again

$$
\begin{aligned}
\mathrm{I} &:= \mathbb{E}_n\left[\int_{-\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\alpha}}^{0} \Psi_1(\boldsymbol{x}_i,\eta(\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\eta^{(1)}(\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}))\eta^{(2)}(\check{\xi}_{\boldsymbol{q},i})(\boldsymbol{p}_i^{\mathsf{T}}(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+t)^2\,\mathrm{d}t\right], \\
\mathrm{II} &:= \frac{1}{2}\mathbb{E}_n\left[\int_{-\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\alpha}}^{0} \Psi_1(\boldsymbol{x}_i,\eta(\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\eta^{(2)}(\xi_{\boldsymbol{q},i,t})\right. \\
&\qquad \left.\times \eta^{(1)}(\boldsymbol{p}_i^{\mathsf{T}}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+t)(\boldsymbol{p}_i^{\mathsf{T}}(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+t)^2\right]\mathrm{d}t, \\
\mathrm{III} &:= \frac{1}{2}\mathbb{E}_n\left[\int_{-\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\alpha}}^{0} \Psi_2(\boldsymbol{x}_i,\tilde{\xi}_{\boldsymbol{q},i,t};\boldsymbol{q})\{\eta(\boldsymbol{p}_i^{\mathsf{T}}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+t) - \eta(\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}))\}^2\right. \\
&\qquad \left.\times \eta^{(1)}(\boldsymbol{p}_i^{\mathsf{T}}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+t)\,\mathrm{d}t\right]
\end{aligned}
$$

and

$$
\begin{aligned}
\mathrm{I} &\lesssim 2^L\mathfrak{r}_{2,n}h^d(r_{1,n}+2^\ell\mathfrak{r}_{2,n}+2^L\mathfrak{r}_{2,n})^2, \\
\mathrm{II} &\lesssim 2^L\mathfrak{r}_{2,n}h^d(r_{1,n}+2^\ell\mathfrak{r}_{2,n}+2^L\mathfrak{r}_{2,n})^2, \\
\mathrm{III} &\lesssim 2^L\mathfrak{r}_{2,n}h^d(r_{1,n}+2^\ell\mathfrak{r}_{2,n}+2^L\mathfrak{r}_{2,n})^2
\end{aligned}
$$

on the event $\mathcal{A}_1$. □

Lemma E.15. *There exists an event $\mathcal{A}_2$ whose probability converges to one such that on $\mathcal{A}_2$*

$$
\begin{aligned}
&\sup_{\boldsymbol{q}\in\mathcal{Q},\boldsymbol{\beta}_1\in\mathcal{H}_1,\boldsymbol{\beta}_2\in\mathcal{H}_{2,\ell},\boldsymbol{\alpha}\in\mathcal{V}} \Big|\mathbb{E}_n[\psi(y_i,\eta(\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q}) \\
&\qquad \times (\eta(\boldsymbol{p}_i^{\mathsf{T}}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)) - \eta(\boldsymbol{p}_i^{\mathsf{T}}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{\alpha})))] \\
&\qquad - \mathbb{E}_n[\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{p}_i\psi(y_i,\eta(\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\eta^{(1)}(\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q}))]\Big| \lesssim h^d r_{1,n}(r_{1,n}+2^\ell\mathfrak{r}_{2,n}+2^L\mathfrak{r}_{2,n})2^L\mathfrak{r}_{2,n}.
\end{aligned}
$$

Proof of Lemma E.15. By the Taylor expansion,

$$\begin{aligned}
&\big|\mathbb{E}_n[\psi(y_i,\eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})(\eta(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2))-\eta(\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{\alpha})))] \\
&\quad -\mathbb{E}_n[\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{p}_i\psi(y_i,\eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\eta^{(1)}(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}))]\big| \\
&\quad = \big|\mathbb{E}_n\big[\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\alpha}\psi(y_i,\eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\{\eta^{(2)}(\xi_{\boldsymbol{q},i})\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{\alpha})+(1/2)\eta^{(2)}(\tilde{\xi}_{\boldsymbol{q},i})\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\alpha}\}\big]\big| \\
&\quad \lesssim h^d r_{1,n}(r_{1,n}+2^\ell\mathfrak{r}_{2,n}+2^L\mathfrak{r}_{2,n})2^L\mathfrak{r}_{2,n},
\end{aligned}$$

where $\xi_{\boldsymbol{q},i}$ is between $\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{\alpha})$ and $\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q})$, $\tilde{\xi}_{\boldsymbol{q},i}$ between $\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)$ and $\boldsymbol{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\boldsymbol{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{\alpha})$. The bound holds on the event $\mathcal{A}_2 := \mathcal{A}_2'\cap\mathcal{A}_2''$, where

$$\mathcal{A}_2' := \Big\{\sup_{\boldsymbol{\beta}_1\in\mathcal{H}_1,\boldsymbol{\beta}_2\in\mathcal{H}_{2,\ell},\boldsymbol{\alpha}\in\mathcal{V},\boldsymbol{q}\in\mathcal{Q}}\big\|\mathbb{E}_n[\psi(y_i,\eta(\boldsymbol{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\boldsymbol{q}));\boldsymbol{q})\eta^{(2)}(\xi_{\boldsymbol{q},i})\boldsymbol{p}_i\boldsymbol{p}_i^\mathsf{T}]\big\|_\infty \le h^d r_{1,n}\Big\},$$

and $\mathcal{A}_2''$ is defined the same way as $\mathcal{A}_2'$ with $\xi_{\boldsymbol{q},i}$ replaced by $\tilde{\xi}_{\boldsymbol{q},i}$.

By the same argument as Lemma C.12, $\mathbb{P}\{\mathcal{A}_2\}\to 1$. □

E.3 Rates of convergence

Corollary E.16 (Uniform rate of convergence)**.**

(a) If the conditions of Theorem E.1(a) hold, then

$$\sup_{\boldsymbol{q}\in\mathcal{Q},\boldsymbol{x}\in\mathcal{X}}\big|\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})-\mu_0^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})\big| \lesssim_\mathbb{P} h^{-|\boldsymbol{v}|}\bigg[\bigg(\frac{\log^d n}{nh^d}\bigg)^{1/2}\log n + h^m\bigg] \tag{E.36}$$

(b) If the conditions of Theorem E.1(a) hold and

$$[\log n]^{(d+1)/(\alpha\wedge 1/2)+d} = O(nh^d), \quad h^{(\alpha\wedge 1/2)m}\log^{d/2} n = O(1), \tag{E.37}$$

then

$$\sup_{\boldsymbol{q}\in\mathcal{Q},\boldsymbol{x}\in\mathcal{X}}\big|\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})-\mu_0^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})\big| \lesssim_\mathbb{P} h^{-|\boldsymbol{v}|}\bigg[\bigg(\frac{\log n}{nh^d}\bigg)^{1/2} + h^m\bigg]. \tag{E.38}$$

(c) If the conditions of Theorem E.1(b) hold and

$$[\log n]^{(d+1)/\alpha+d} = O(nh^d), \quad h^{\alpha m}\log^{d/2} n = O(1),$$

then (E.38) is also true.

Proof. By Theorem E.1, Lemma C.12 and triangle inequality,

$$\begin{aligned}
&\sup_{\boldsymbol{q}\in\mathcal{Q},\boldsymbol{x}\in\mathcal{X}}\big|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^\mathsf{T}\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\mu_0^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})\big| \\
&\quad \lesssim_\mathbb{P} h^{-|\boldsymbol{v}|}\bigg[\bigg(\frac{\log(1/h)}{nh^d}\bigg)^{1/2} + \bigg(\frac{\log^d n}{nh^d}\bigg)^{1/2+\alpha/2\wedge 1/4}\log n + h^{(\alpha\wedge 1/2)m}\bigg(\frac{\log^{d+1} n}{nh^d}\bigg)^{1/2} + h^m\bigg].
\end{aligned}$$

Using $\log(1/h)\lesssim\log n$ and simplifying the right-hand size, we obtain (E.36). Additional restrictions (E.37) allow us to get a slightly stronger result (E.38). □

Corollary E.17 (Mean square rate of convergence)**.**

(a) *If the conditions of Theorem E.1(a) hold and*

$$[\log n]^{(d+2)/(\alpha\wedge 1/2)+d} = o(nh^d), \quad h^{(\alpha\wedge 1/2)m}\log^{(d+1)/2} n = o(1), \tag{E.39}$$

then

$$\sup_{\boldsymbol{q}\in\mathcal{Q}}\Big(\int_{\mathcal{X}}\big|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}}\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})\big|^2 f_X(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\Big)^{1/2} \lesssim_{\mathbb{P}} h^{-|\boldsymbol{v}|}\Big[\frac{1}{\sqrt{nh^d}} + h^m\Big]. \tag{E.40}$$

(b) *If the conditions of Theorem E.1(b) hold and*

$$[\log n]^{(d+2)/\alpha+d} = o(nh^d), \quad h^{\alpha m}\log^{(d+1)/2} n = o(1),$$

then (E.40) *is also true.*

Proof. To prove (E.40), note that

$$\begin{aligned}
&\sup_{\boldsymbol{q}}\int_{\mathcal{X}}\big|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}}\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}_0(\boldsymbol{q})\big|^2 f_X(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\\
&\quad= \sup_{\boldsymbol{q}}\big(\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})\big)^{\mathsf{T}}\mathbb{E}[\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}}]\big(\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})\big) \overset{\text{(a)}}{\lesssim} h^{d-2|\boldsymbol{v}|}\big\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})\big\|^2,
\end{aligned}$$

where inequality (a) is true because the largest eigenvalue of $\mathbb{E}[\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x}_i)\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x}_i)^{\mathsf{T}}]$ is bounded from above by $h^{d-2|\boldsymbol{v}|}$ up to a multiplicative coefficient, which is proven by the same argument as for $\boldsymbol{v} = 0$ in Lemma C.11 in combination with Assumption B.2.

It is left to prove

$$\big\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})\big\| \lesssim_{\mathbb{P}} \frac{1}{h^d\sqrt{n}}. \tag{E.41}$$

By the triangle inequality,

$$\begin{aligned}
\big\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q})\big\| &\le \big\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q}) + \mathbf{Q}_{0,\boldsymbol{q}}^{-1}\mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}));\boldsymbol{q})]\big\|\\
&\quad+ \big\|\mathbf{Q}_{0,\boldsymbol{q}}^{-1}\mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}));\boldsymbol{q})]\big\|.
\end{aligned} \tag{E.42}$$

To bound the second term in (E.42) on the right-hand side, consider the expectation

$$\begin{aligned}
&\mathbb{E}\big[\big\|\mathbb{E}_n\big[\boldsymbol{p}(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}));\boldsymbol{q})\big]\big\|^2\big]\\
&\quad= \frac{1}{n^2}\sum_{i=1}^n\mathbb{E}\big[\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))^2\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}));\boldsymbol{q})^2\|\boldsymbol{p}(\boldsymbol{x}_i)\|^2\big]\\
&\quad= \frac{1}{n^2}\sum_{i=1}^n\mathbb{E}\big[\sigma_{\boldsymbol{q}}^2(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))^2\|\boldsymbol{p}(\boldsymbol{x}_i)\|^2\big] \overset{\text{(a)}}{\lesssim} \frac{1}{n^2}\sum_{i=1}^n 1 = \frac{1}{n},
\end{aligned}$$

where in (a) we used uniform boundedness of $\sigma_{\boldsymbol{q}}^2(\boldsymbol{x})$ by Assumption B.4(ii), uniform boundedness of $\mu_0(\boldsymbol{x},\boldsymbol{q})$ and $\|\boldsymbol{p}(\boldsymbol{x})\|$. By Markov's inequality and Lemma C.11, this immediately implies

$$\big\|\mathbf{Q}_{0,\boldsymbol{q}}^{-1}\mathbb{E}_n\big[\boldsymbol{p}(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}));\boldsymbol{q})\big]\big\| \lesssim_{\mathbb{P}} \|\mathbf{Q}_{0,\boldsymbol{q}}^{-1}\|\cdot\frac{1}{\sqrt{n}} \lesssim \frac{1}{h^d\sqrt{n}}.$$

Concerning the first term in (E.42), it is proven in Theorem E.1 (see Eq. (E.10)) that

$$\big\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \boldsymbol{\beta}_0(\boldsymbol{q}) + \mathbf{Q}_{0,\boldsymbol{q}}^{-1}\mathbb{E}_n\big[\boldsymbol{p}(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}));\boldsymbol{q})\big]\big\|_\infty \lesssim_{\mathbb{P}} r_{\mathtt{BR}}$$

for $r_{\mathtt{BR}}$ defined in (E.3), so that

$$\begin{aligned}&\left\|\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\boldsymbol{\beta}_0(\boldsymbol{q})+\mathbf{Q}_{0,\boldsymbol{q}}^{-1}\mathbb{E}_n\big[\boldsymbol{p}(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}));\boldsymbol{q})\big]\right\|\\&\lesssim_{\mathbb{P}}\sqrt{K}r_{\mathtt{BR}}\lesssim\frac{r_{\mathtt{BR}}}{h^{d/2}}\overset{\text{(a)}}{=}o\left(\frac{1}{h^d\sqrt{n}}\right),\end{aligned}$$

where in equality (a) we used $r_{\mathtt{BR}}=o\big(1/\sqrt{nh^d}\big)$ under the assumptions. This concludes the proof of (E.41). $\square$

F Strong approximation

Appendix F.1 collects results that may be of independent theoretical interest. Specifically, we use the conditional Strassen's theorem Theorem F.1 to prove conditional Yurinskii coupling in d-norm, Theorem F.2. We then use it to prove Lemma F.3, a version of Gaussian approximation for a K-dimensional empirical process. Appendix F.2 proves the main result of this appendix, Theorem F.4, which is essentially a corollary of Lemma F.3 after some additional technical work done in Lemmas F.6 to F.8.

F.1 Yurinskii coupling

The three theorems, and their proofs, in this subsection are self-contained, and hence all variables, functions, and stochastic processes, should be treated as defined within each of the theorems and their proofs, and independently of all other statements elsewhere in the supplemental appendix.

The following theorem is due to [36]. We use the statement from [16] making it explicit that the supremum over Borel sets may not be a random variable (a direct proof may also be found in that work). Let (S,d) be a Polish space (where d is its metric), and $\mathcal{B}(S)$ its Borel sigma-algebra.

Theorem F.1 (Conditional Strassen's theorem)**.** *Let X be a random variable defined on some probability space $(\Omega,\mathcal{F},\mathbb{P})$ and with values in some Polish (S,d). Let $\mathcal{J}$ be a contably generated sub-sigma algebra of $\mathcal{F}$ and assume that this probability space is rich enough: there exists a random variable U that is independent of the sigma-algebra $\mathcal{J}\vee\sigma(X)$. Let $\mathcal{B}(S)\times\Omega\ni(A,\omega)\mapsto G(A|\mathcal{J})(\omega)$ be a regular conditional distribution on $\mathcal{B}(S)$, i. e., for each $A\in\mathcal{B}(S)$, $G(A|\mathcal{J})$ is $\mathcal{J}$-measurable, and for each $\omega\in\Omega$, $G(\cdot|\mathcal{J})(\omega)$ is a probability measure on $\mathcal{B}(S)$. Suppose that for some nonnegative numbers α and β*

$$\mathbb{E}^*\Big[\sup_{A\in\mathcal{B}}\{\mathbb{P}\{X\in A\,|\,\mathcal{J}\}-G(A^\alpha|\mathcal{J})\}\Big]\le\beta,$$

where $\mathbb{E}^*$ denotes outer expectation. Then on this probability space there exists an S-valued random element Y such that $G(\cdot|\mathcal{J})$ is its regular conditional distribution given $\mathcal{J}$ and $\mathbb{P}\{d(X,Y)>\alpha\}\le\beta$.

The following theorem is a conditional version of Lemma 39 in [3]; its proof carefully leverages Theorem F.1. See also [13] for a related, but different, conditional Yurinskii's coupling result.

Theorem F.2 (Conditional Yurinskii coupling, d-norm)**.** *Let a random vector $\boldsymbol{C}$ with values in $\mathbb{R}^m$, sequence of random vectors $\{\boldsymbol{\xi}_i\}_{i=1}^n$ with values in $\mathbb{R}^k$ and sequence of random vectors $\{\boldsymbol{g}_i\}_{i=1}^n$ with values in $\mathbb{R}^k$ be defined on the same probability space and be such that the all the $2n$ vectors $\{\boldsymbol{\xi}_i\}_{i=1}^n\cup\{\boldsymbol{g}_i\}_{i=1}^n$ are independent conditionally on $\boldsymbol{C}$, are mean zero conditionally on $\boldsymbol{C}$ and for each $i\in\{1,\dots,n\}$ the distribution of $\boldsymbol{g}_i$ conditionally on $\boldsymbol{C}$ is $\mathcal{N}(0,\mathbb{V}[\boldsymbol{\xi}_i\,|\,\boldsymbol{C}])$. Assume that this probability space is rich enough: there exists a random variable U that is independent of $\{\boldsymbol{\xi}_i\}_{i=1}^n\cup\{\boldsymbol{g}_i\}_{i=1}^n$. Denote*

$$\boldsymbol{S}:=\boldsymbol{\xi}_1+\ldots+\boldsymbol{\xi}_n,\quad\boldsymbol{T}:=\boldsymbol{g}_1+\ldots+\boldsymbol{g}_n,$$

and let

$$\beta := \sum_{i=1}^{n} \mathbb{E}\big[\|\boldsymbol{\xi}_i\|^2\|\boldsymbol{\xi}_i\|_d\big] + \sum_{i=1}^{n} \mathbb{E}\big[\|\boldsymbol{g}_i\|^2\|\boldsymbol{g}_i\|_d\big]$$

be finite. Then for each $\delta > 0$, on this probability space there exists a random vector $\boldsymbol{T}'$ such that $\mathbb{P}_{\boldsymbol{T}|\sigma(\boldsymbol{C})}(\cdot)$ is its regular conditional distribution given $\sigma(\boldsymbol{C})$, and

$$\mathbb{P}\big\{\|\boldsymbol{S} - \boldsymbol{T}'\|_d > 3\delta\big\} \le \min_{t\ge 0}\bigg(2\mathbb{P}\{\|\boldsymbol{Z}\|_d > t\} + \frac{\beta}{\delta^3}t^2\bigg),$$

where $\boldsymbol{Z} \sim \mathcal{N}(0, I_k)$.

Proof. By the conditional Strassen's theorem, it is enough to show

$$\mathbb{E}^*\Big[\sup_{A\in\mathcal{B}(\mathbb{R}^k)}\big(\mathbb{P}\{\boldsymbol{S}\in A\,|\,\sigma(\boldsymbol{C})\} - \mathbb{P}_{\boldsymbol{T}|\sigma(\boldsymbol{C})}(A^{3\delta})\big)\Big] \le \min_{t\ge 0}\bigg(2\mathbb{P}\{\|\boldsymbol{Z}\|_d > t\} + \frac{\beta}{\delta^3}t^2\bigg),$$

or, equivalently, for any $t > 0$

$$\mathbb{E}^*\Big[\sup_{A\in\mathcal{B}(\mathbb{R}^k)}\big(\mathbb{P}\{\boldsymbol{S}\in A\,|\,\boldsymbol{C}\} - \mathbb{P}\{\boldsymbol{T}\in A^{3\delta}\,|\,\boldsymbol{C}\}\big)\Big] \le 2\mathbb{P}\{\|bZ\|_d > t\} + \frac{\beta}{\delta^3}t^2. \tag{F.1}$$

Fix $t > 0$ and $A \in \mathcal{B}(\mathbb{R}^k)$. Let $f\colon \mathbb{R}^k \to \mathbb{R}$ be the same as in the proof of Lemma 39 in [3], namely, it is such that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^k$,

$$\begin{aligned}
&\big|f(\boldsymbol{x}+\boldsymbol{y}) - f(\boldsymbol{x}) - \boldsymbol{y}^\mathsf{T}\nabla f(\boldsymbol{x}) - (1/2)\boldsymbol{y}^\mathsf{T}\nabla^2 f(\boldsymbol{x})\boldsymbol{y}\big| \le \frac{\|\boldsymbol{y}\|^2\|\boldsymbol{y}\|_d}{\sigma^2\delta},\\
&(1-\epsilon)1\{\boldsymbol{x}\in A\} \le f(\boldsymbol{x}) \le \epsilon + (1-\epsilon)1\{\boldsymbol{x}\in A^{3\delta}\},
\end{aligned}$$

with $\sigma := \delta/t$ and $\epsilon := \mathbb{P}\{\|\boldsymbol{Z}\|_d > t\}$.

Then note that

$$\begin{aligned}
\mathbb{P}\{\boldsymbol{S}\in A\,|\,\boldsymbol{C}\} &= \mathbb{E}\big[\mathbb{1}\{\boldsymbol{S}\in A\} - f(\boldsymbol{S})\,\big|\,\boldsymbol{C}\big] + \mathbb{E}\big[f(\boldsymbol{S}) - f(\boldsymbol{T})\,\big|\,\boldsymbol{C}\big] + \mathbb{E}[f(\boldsymbol{T})\,|\,\boldsymbol{C}]\\
&\le \epsilon\mathbb{E}\big[\mathbb{1}\{\boldsymbol{S}\in A\}\,\big|\,\boldsymbol{C}\big] + \mathbb{E}\big[f(\boldsymbol{S}) - f(\boldsymbol{T})\,\big|\,\boldsymbol{C}\big] + \epsilon + (1-\epsilon)\mathbb{E}\big[\mathbb{1}\{\boldsymbol{T}\in A^{3\delta}\}\,\big|\,\boldsymbol{C}\big]\\
&\le 2\epsilon + \mathbb{E}\big[\mathbb{1}\{\boldsymbol{T}\in A^{3\delta}\}\,\big|\,\boldsymbol{C}\big] + \mathbb{E}\big[f(\boldsymbol{S}) - f(\boldsymbol{T})\,\big|\,\boldsymbol{C}\big].
\end{aligned} \tag{F.2}$$

Now we bound $\mathbb{E}[f(\boldsymbol{S}) - f(\boldsymbol{T})\,|\,\boldsymbol{C}]$:

$$\begin{aligned}
\mathbb{E}[f(\boldsymbol{S}) - f(\boldsymbol{T})\,|\,\boldsymbol{C}] &= \sum_{i=1}^{n}\mathbb{E}[f(\boldsymbol{X}_i+\boldsymbol{Y}_i) - f(\boldsymbol{X}_i+\boldsymbol{W}_i)\,|\,\boldsymbol{C}]\\
&\le \sum_{i=1}^{n}\mathbb{E}\bigg[f(\boldsymbol{X}_i) + \boldsymbol{Y}_i^\mathsf{T}\nabla f(\boldsymbol{X}_i) + \frac{1}{2}\boldsymbol{Y}_i^\mathsf{T}\nabla^2 f(\boldsymbol{X}_i)\boldsymbol{Y}_i + \frac{\|\boldsymbol{Y}_i\|^2\|\boldsymbol{Y}_i\|_d}{\sigma^2\delta}\,\bigg|\,\boldsymbol{C}\bigg]\\
&\quad - \sum_{i=1}^{n}\mathbb{E}\bigg[f(\boldsymbol{X}_i) + \boldsymbol{W}_i^\mathsf{T}\nabla f(\boldsymbol{X}_i) + \frac{1}{2}\boldsymbol{W}_i^\mathsf{T}\nabla^2 f(\boldsymbol{X}_i)\boldsymbol{W}_i - \frac{\|\boldsymbol{W}_i\|^2\|\boldsymbol{W}_i\|_d}{\sigma^2\delta}\,\bigg|\,\boldsymbol{C}\bigg]\\
&\overset{\text{(a)}}{=} \sum_{i=1}^{n}\mathbb{E}[\frac{\|\boldsymbol{Y}_i\|^2\|\boldsymbol{Y}_i\|_d + \|\boldsymbol{W}_i\|^2\|\boldsymbol{W}_i\|_d}{\sigma^2\delta}\,|\,\boldsymbol{C}]\ \text{a.s.}
\end{aligned}$$

for $\boldsymbol{X}_i := \boldsymbol{\xi}_1 + \ldots + \boldsymbol{\xi}_{i-1} + \boldsymbol{g}_{i+1} + \ldots + \boldsymbol{g}_n$, $\boldsymbol{Y}_i := \boldsymbol{\xi}_i$, $\boldsymbol{W}_i := \boldsymbol{g}_i$, in (a) we used the conditional independence of the family $\{\boldsymbol{\xi}_i\}_{i=1}^n \cup \{\boldsymbol{g}_i\}_{i=1}^n$, that they are conditionally mean zero and the equality of the corresponding conditional second moments.

We conclude that almost surely

$$\sup_{A\in\mathcal{B}(\mathbb{R}^k)}\big(\mathbb{P}\{\boldsymbol{S}\in A\,|\,\boldsymbol{C}\}-\mathbb{P}\{\boldsymbol{T}\in A^{3\delta}\,|\,\boldsymbol{C}\}\big)\le 2\epsilon+\sum_{i=1}^n\mathbb{E}\bigg[\frac{\|\boldsymbol{\xi}_i\|^2\|\boldsymbol{\xi}_i\|_d+\|\boldsymbol{g}_i\|^2\|\boldsymbol{g}_i\|_d}{\sigma^2\delta}\,\bigg|\,\boldsymbol{C}\bigg].$$

By the definition of outer expectation, this implies

$$\mathbb{E}^*\Big[\sup_{A\in\mathcal{B}(\mathbb{R}^k)}\big(\mathbb{P}\{\boldsymbol{S}\in A\,|\,\boldsymbol{C}\}-\mathbb{P}\{\boldsymbol{T}\in A^{3\delta}\,|\,\boldsymbol{C}\}\big)\Big]\le 2\epsilon+\frac{\beta}{\sigma^2\delta},$$

which is (F.1). □

The following lemma generalizes Lemma 36 in [3], and also builds on the argument for Lemma SA27 in [12].

Lemma F.3 (Yurinskii coupling: K-dimensional process)**.** *Let $\{\boldsymbol{x}_i,y_i\}_{i=1}^n$ be a random sample, where $\boldsymbol{x}_i$ has compact support $\mathcal{X}\subset\mathbb{R}^d$, $y_i\in\mathcal{Y}\subset\mathbb{R}$ is a scalar. Also let $\mathcal{Q}\subseteq\mathbb{R}^{d_\mathcal{Q}}$ be a fixed compact set.*

Let $A_n\colon\mathcal{Q}\times\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ be a Borel measurable function satisfying $\sup_{\boldsymbol{q}\in\mathcal{Q}}|A_n(\boldsymbol{q},\boldsymbol{x}_i,y_i)|\le\bar{A}_n(\boldsymbol{x}_i,y_i)$, where $\bar{A}_n(\boldsymbol{x}_i,y_i)$ is a Borel measurable envelope, $\mathbb{E}[A_n(\boldsymbol{q},\boldsymbol{x}_i,y_i)\,|\,\boldsymbol{x}_i]=0$ for all $\boldsymbol{q}\in\mathcal{Q}$, $\sup_{\boldsymbol{x}\in\mathcal{X}}\mathbb{E}[|\bar{A}(\boldsymbol{x}_i,y_i)|^\nu\,|\,\boldsymbol{x}_i=\boldsymbol{x}]\le\mu_n<\infty$ for some $\nu\ge3$ with $\mu_n\gtrsim1$ and $\log\mu_n\lesssim\log n$, which satisfies the Lipschitz condition

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\mathbb{E}\big[|A_n(\boldsymbol{q},\boldsymbol{x}_i,y_i)-A_n(\tilde{\boldsymbol{q}},\boldsymbol{x}_i,y_i)|^2\,\big|\,\boldsymbol{x}_i=\boldsymbol{x}\big]\lesssim\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|$$

for all $\boldsymbol{q},\tilde{\boldsymbol{q}}\in\mathcal{Q}$. Also, the (regular) conditional variance $\mathbb{E}[A_n(\boldsymbol{q},\boldsymbol{x}_i,y_i)^2\,|\,\boldsymbol{x}_i=\boldsymbol{x}]$ is continuous in $\boldsymbol{x}\in\mathcal{X}$. Moreover, assume that the class of functions $\{(\boldsymbol{x},y)\mapsto A_n(\boldsymbol{q},\boldsymbol{x},y)\colon\boldsymbol{q}\in\mathcal{Q}\}$ is VC-subgraph with an index bounded from above by a constant not depending on n.

Let $\boldsymbol{b}(\cdot)$ be a Borel measurable function $\mathcal{X}\to\mathbb{R}^K$ (where $K=K_n$ is some sequence of positive integers tending to infinity and satisfying $\log K\lesssim\log n$) such that $\sup_{\boldsymbol{x}\in\mathcal{X}}\|\boldsymbol{b}(\boldsymbol{x})\|\le\zeta_K$ and the probability of the event $\mathcal{A}:=\{\sup_{\|\boldsymbol{\alpha}\|=1}\mathbb{E}_n\big[(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{b}(\boldsymbol{x}_i))^2\big]\le C_{\mathrm{Gr}}\}$ approaches one, where C_{Gr} is some positive constant. Assume ζ_K satisfies $1/\zeta_K\lesssim1$, $|\log\zeta_K|\lesssim\log n$.

Let $r_{n,\mathrm{yur}}=r_{\mathtt{YUR}}\to0$ be a sequence of positive numbers satisfying

$$\bigg(\frac{\zeta_K^3\mu_n^{3/\nu}}{\sqrt{n}}\bigg)^{\frac{1}{3+2d_\mathcal{Q}}}\sqrt{\log n}+\frac{\zeta_K\mu_n^{1/\nu}}{n^{1/2-1/\nu}}\log n=o(r_{\mathtt{YUR}}).\tag{F.3}$$

Assume also that the probability space is rich enough, and denote

$$\boldsymbol{G}_n(\boldsymbol{q}):=\sqrt{n}\mathbb{E}_n[A_n(\boldsymbol{q},\boldsymbol{x}_i,y_i)\boldsymbol{b}(\boldsymbol{x}_i)].$$

Then, on the same probability space, there exists a K-dimensional, conditionally on $\mathbf{X}_n$ mean-zero Gaussian, process $\boldsymbol{Z}_n(\boldsymbol{q})$ on $\mathcal{Q}$ with a. s. continuous trajectories, satisfying

$$\mathbb{E}\big[\boldsymbol{G}_n(\boldsymbol{q})\boldsymbol{G}_n(\tilde{\boldsymbol{q}})^\mathsf{T}\,\big|\,\mathbf{X}_n\big]=\mathbb{E}\big[\boldsymbol{Z}_n(\boldsymbol{q})\boldsymbol{Z}_n(\tilde{\boldsymbol{q}})^\mathsf{T}\,\big|\,\mathbf{X}_n\big],\quad \boldsymbol{q},\tilde{\boldsymbol{q}}\in\mathcal{Q};$$
$$\sup_{\boldsymbol{q}\in\mathcal{Q}}\|\boldsymbol{G}_n(\boldsymbol{q})-\boldsymbol{Z}_n(\boldsymbol{q})\|_\infty=o_\mathbb{P}(r_{\mathtt{YUR}}).$$

Moreover, if $\bar{A}_n(\boldsymbol{q},\boldsymbol{x}_i,y_i)$ is σ^2-sub-Gaussian, then (F.3) can be replaced with

$$\bigg(\frac{\zeta_K^3}{\sqrt{n}}\bigg)^{\frac{1}{3+2d_\mathcal{Q}}}\sqrt{\log n}+\frac{\zeta_K}{\sqrt{n}}\log^{3/2}n=o(r_{\mathtt{YUR}}).$$

Proof. Let $\mathcal{Q}_n^\delta := \{\boldsymbol{q}_1, \dots, \boldsymbol{q}_{|\mathcal{Q}_n^\delta|}\}$ be an internal δ_n-covering of $\mathcal{Q}$ with respect to the 2-norm $\|\cdot\|$ of cardinality $|\mathcal{Q}_n^\delta| \lesssim 1/\delta_n^{d_\mathcal{Q}}$, where δ_n is chosen later. Denote $\pi_n^\delta \colon \mathcal{Q} \to \mathcal{Q}$ a sequence of projections associated with this covering: it maps each point in $\mathcal{Q}$ to the center of the ball containing this point (if such a ball is not unique, choose one by an arbitrary rule).

Strategy The plan of attack is to

1. show that $\boldsymbol{G}_n(\boldsymbol{q})$ does not deviate too much in sup-norm from its projected version, i. e. bound the tails of $\sup_{\boldsymbol{q}\in\mathcal{Q}}\|\boldsymbol{G}_n(\boldsymbol{q}) - \boldsymbol{G}_n \circ \pi_n^\delta(\boldsymbol{q})\|_\infty$,
2. apply Yurinskii coupling to the finite-dimensional vector $\big(\boldsymbol{G}_n \circ \pi_n^\delta(\boldsymbol{q})\big)_{\boldsymbol{q}\in\mathcal{Q}_n^\delta}$ and obtain a conditionally on $\mathbf{X}_n$ Gaussian vector $\boldsymbol{Z}_n^\delta$ with the right structure that is close enough, i. e. with a bound on the tails of $\|\boldsymbol{G}_n \circ \pi_n^\delta - \boldsymbol{Z}_n^\delta\|_\infty$,
3. extend this conditionally Gaussian vector to a K-dimensional conditionally Gaussian process $\boldsymbol{Z}_n$,
4. and finally show that $\boldsymbol{Z}_n(\boldsymbol{q})$ does not deviate too much from its projected version, i. e. bound the tails of $\sup_{\boldsymbol{q}\in\mathcal{Q}}\|\boldsymbol{Z}_n(\boldsymbol{q}) - \boldsymbol{Z}_n \circ \pi_n^\delta(\boldsymbol{q})\|_\infty$.

If we complete these steps, it will prove the theorem by the triangle inequality.

Discretization of G_n Consider the class of functions

$$\mathcal{G}_n' := \{\mathcal{X} \times \mathcal{Y} \ni (\boldsymbol{x}, y) \mapsto A_n(\boldsymbol{q}, \boldsymbol{x}, y) b_l(\boldsymbol{x}) \colon 1 \le l \le K, \boldsymbol{q} \in \mathcal{Q}\}$$

with envelope $\zeta_K \bar{A}_n(X)$. Since $\{A_n(\boldsymbol{q}, \boldsymbol{x}, y)\}$ is a VC class with $O(1)$ index and envelope $\bar{A}_n(\boldsymbol{x}, y)$, $\mathcal{G}_n'$ satisfies the uniform entropy bound (A.3) with $A \lesssim K$ and $V \lesssim 1$.

Next, consider the class of functions

$$\mathcal{G}_n^\delta := \{\mathcal{X} \times \mathcal{Y} \ni (\boldsymbol{x}, y) \mapsto \big(A_n(\boldsymbol{q}, \boldsymbol{x}, y) - A_n(\tilde{\boldsymbol{q}}, \boldsymbol{x}, y)\big) b_l(\boldsymbol{x}) \colon 1 \le l \le K, \boldsymbol{q}, \tilde{\boldsymbol{q}} \in \mathcal{Q}, \|\boldsymbol{q} - \tilde{\boldsymbol{q}}\| \lesssim \delta_n\}$$

with envelope $(\boldsymbol{x}, y) \mapsto 2\zeta_K \bar{A}_n(\boldsymbol{x}, y)$. Using Lemma C.4, we get that this class satisfies the uniform entropy bound (A.3) with $A \lesssim K$ and $V \lesssim 1$.

Now we apply Lemma C.6 conditionally on $\mathbf{X}_n$ on $\mathcal{A}$ with $\|F\|_{\bar{\mathbb{P}},2} \le 2\zeta_K \mu_n^{1/\nu}$ since

$$\|\bar{A}_n(\boldsymbol{x}, y)\|_{\bar{\mathbb{P}},2}^2 = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\bar{A}_n(\boldsymbol{x}_i, y_i)^2 \,|\, \boldsymbol{x}_i] \le \frac{1}{n}\sum_{i=1}^n \mu_n^{2/\nu} = \mu_n^{2/\nu},$$

$\|M\|_{\mathbb{P},2} \le 2\zeta_K (\mu_n n)^{1/\nu}$ since

$$\begin{aligned}\mathbb{E}\Big[\Big(\max_{1\le i\le n} \bar{A}_n(\boldsymbol{x}_i, y_i)\Big)^2 \,\Big|\, \mathbf{X}_n\Big] &\le \mathbb{E}\Big[\Big(\max_{1\le i\le n} \bar{A}_n(\boldsymbol{x}_i, y_i)\Big)^\nu \,\Big|\, \mathbf{X}_n\Big]^{2/\nu} \\ &\le \mathbb{E}\Big[\sum_{i=1}^n \bar{A}_n(\boldsymbol{x}_i, y_i)^\nu \,\Big|\, \mathbf{X}_n\Big]^{2/\nu} \le (\mu_n n)^{2/\nu},\end{aligned}$$

and $\sigma \lesssim \sqrt{\delta_n}$ since on $\mathcal{A}$

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[\big(A_n(\boldsymbol{q},\boldsymbol{x}_i,y_i) - A_n(\tilde{\boldsymbol{q}},\boldsymbol{x}_i,y_i)\big)^2 b_l(\boldsymbol{x}_i)^2 \,\big|\, \mathbf{X}_n\big] \\
&\quad = \frac{1}{n}\sum_{i=1}^n b_l(\boldsymbol{x}_i)^2 \mathbb{E}\big[\big(A_n(\boldsymbol{q},\boldsymbol{x}_i,y_i) - A_n(\tilde{\boldsymbol{q}},\boldsymbol{x}_i,y_i)\big)^2 \,\big|\, \mathbf{X}_n\big] \\
&\quad \lesssim \|\boldsymbol{q}-\tilde{\boldsymbol{q}}\| \frac{1}{n}\sum_{i=1}^n b_l(\boldsymbol{x}_i)^2 \lesssim \|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|.
\end{aligned}$$

This gives that on $\mathcal{A}$

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|\le\delta_n} \|\boldsymbol{G}_n(\boldsymbol{q}) - \boldsymbol{G}_n(\tilde{\boldsymbol{q}})\|_\infty \,\Big|\, \mathbf{X}_n\Big] \\
&\quad \lesssim \sqrt{\delta_n \log \frac{K\zeta_K \mu_n^{1/\nu}}{\sqrt{\delta_n}}} + \frac{\zeta_K (\mu_n n)^{1/\nu}}{\sqrt{n}} \log \frac{K\zeta_K\mu_n^{1/\nu}}{\sqrt{\delta_n}}.
\end{aligned}$$

By Markov's inequality and since $\mathbb{P}\{\mathcal{A}\} \to 1$, for any sequence $t_n > 0$

$$\begin{aligned}
&\mathbb{P}\Big\{\sup_{\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|\le\delta_n} \|\boldsymbol{G}_n(\boldsymbol{q}) - \boldsymbol{G}_n(\tilde{\boldsymbol{q}})\|_\infty > t_n\Big\} \\
&\quad \lesssim \frac{1}{t_n}\sqrt{\delta_n \log \frac{K\zeta_K \mu_n^{1/\nu}}{\sqrt{\delta_n}}} + \frac{\zeta_K (\mu_n n)^{1/\nu}}{t_n\sqrt{n}} \log \frac{K\zeta_K\mu_n^{1/\nu}}{\sqrt{\delta_n}} + o(1),
\end{aligned} \tag{F.4}$$

where the constant in $\lesssim$ does not depend on n.

Coupling Define a $K|\mathcal{Q}_n^\delta|$-dimensional vector $\boldsymbol{\xi}_i := \big(A_n(\boldsymbol{q},\boldsymbol{x}_i,y_i) b_l(\boldsymbol{x}_i)/\sqrt{n}\big)_{1\le l\le K, \boldsymbol{q}\in\mathcal{Q}_n^\delta}$, so that we have $\boldsymbol{G}_n \circ \pi_n^\delta = \sum_{i=1}^n \boldsymbol{\xi}_i$. We make some preparations before applying Theorem F.2.

Firstly, we bound $\mathbb{E}\big[\|\boldsymbol{\xi}_i\|^2\|\boldsymbol{\xi}_i\|_\infty \,\big|\, \mathbf{X}_n\big]$.

$$\begin{aligned}
&\sum_{i=1}^n \mathbb{E}\big[\|\boldsymbol{\xi}_i\|^2\|\boldsymbol{\xi}_i\|_\infty \,\big|\, \mathbf{X}_n\big] \\
&\quad = \frac{1}{n^{3/2}}\sum_{i=1}^n \|\boldsymbol{b}(\boldsymbol{x}_i)\|^2 \|\boldsymbol{b}(\boldsymbol{x}_i)\|_\infty \mathbb{E}\Big[\sum_{\boldsymbol{q}\in\mathcal{Q}_n^\delta} A_n(\boldsymbol{q},\boldsymbol{x}_i,y_i)^2 \cdot \max_{\boldsymbol{q}\in\mathcal{Q}_n^\delta} |A_n(\boldsymbol{q},\boldsymbol{x}_i,y_i)| \,\Big|\, \boldsymbol{x}_i\Big] \\
&\quad \le \frac{1}{n^{3/2}} |\mathcal{Q}_n^\delta| \sum_{i=1}^n \|\boldsymbol{b}(\boldsymbol{x}_i)\|^2 \|\boldsymbol{b}(\boldsymbol{x}_i)\|_\infty \mathbb{E}\big[\bar{A}_n(\boldsymbol{x}_i,y_i)^3 \,\big|\, \boldsymbol{x}_i\big] \\
&\quad \le \frac{|\mathcal{Q}_n^\delta| \mu_n^{3/\nu}}{n^{3/2}} \sum_{i=1}^n \|\boldsymbol{b}(\boldsymbol{x}_i)\|^2 \|\boldsymbol{b}(\boldsymbol{x}_i)\|_\infty \lesssim \frac{|\mathcal{Q}_n^\delta|\mu_n^{3/\nu}}{n^{1/2}} \zeta_K^3.
\end{aligned}$$

Secondly, for $i \in \{1,\dots,n\}$ let $\boldsymbol{g}_i \sim \mathcal{N}(0,\Sigma_i)$ be independent vectors, where $\Sigma_i = \mathbb{V}[\boldsymbol{\xi}_i \,|\, \boldsymbol{x}_i]$. Since there is an independent random variable U_1 distributed uniformly on $[0,1]$, we can construct the family $\{\boldsymbol{g}_i\}$ on the same probability space. Then by Jensen's inequality for any $\lambda > 0$ we have

$$\begin{aligned}
\mathbb{E}\big[\|\boldsymbol{g}_i\|_\infty^2 \,\big|\, \boldsymbol{x}_i\big] &\le \frac{1}{\lambda}\log \mathbb{E}\big[e^{\lambda\|\boldsymbol{g}_i\|_\infty^2} \,\big|\, \boldsymbol{x}_i\big] \le \frac{1}{\lambda}\log \sum_{t=1}^{K|\mathcal{Q}_n^\delta|} \mathbb{E}\big[e^{\lambda(g_{it})^2} \,\big|\, \boldsymbol{x}_i\big] \\
&\le \frac{-\frac{1}{2}\log\Big(1 - \frac{2\lambda}{n}\zeta_K^2\mu_n^{2/\nu}\Big) + \log K + \log|\mathcal{Q}_n^\delta|}{\lambda} \lesssim \frac{\zeta_K^2\mu_n^{2/\nu}}{n}\big(\log K + \log|\mathcal{Q}_n^\delta|\big),
\end{aligned}$$

where we used the moment-generating function of χ_1^2: $\mathbb{E}[\exp\{\alpha\chi_1^2\}] = (1-2\alpha)^{-1/2}$ for $\alpha < 1/2$, the bound $\mathbb{V}[\xi_{it} \,|\, \boldsymbol{x}_i] \le \zeta_K^2\mu_n^{2/\nu}\big/n$, and put $\lambda := \big(4\zeta_K^2\mu_n^{2/\nu}/n\big)^{-1}$. Also,

$$\begin{aligned}
\mathbb{E}\big[\|\boldsymbol{g}_i\|^4 \,\big|\, \boldsymbol{x}_i\big]^{1/2} &= \mathbb{E}\bigg[\bigg(\sum_{t=1}^{K|\mathcal{Q}_n^\delta|} g_{it}^2\bigg)^2 \,\bigg|\, \boldsymbol{x}_i\bigg]^{1/2} \le \sum_{t=1}^{K|\mathcal{Q}_n^\delta|} \mathbb{E}[g_{it}^4 \,|\, \boldsymbol{x}_i]^{1/2} \\
&= \sum_{t=1}^{K|\mathcal{Q}_n^\delta|} \sqrt{3}\mathbb{E}[g_{it}^2 \,|\, \boldsymbol{x}_i] \lesssim \mathbb{E}\big[\|\boldsymbol{g}_i\|^2 \,\big|\, \boldsymbol{x}_i\big] = \mathbb{E}[\|\boldsymbol{\xi}_i\|^2 \,|\, \boldsymbol{x}_i],
\end{aligned}$$

which gives

$$\sum_{i=1}^n \mathbb{E}\big[\|\boldsymbol{g}_i\|^4 \,\big|\, \boldsymbol{x}_i\big]^{1/2} \lesssim \sum_{i=1}^n \mathbb{E}\big[\|\boldsymbol{\xi}_i\|^2 \,\big|\, \boldsymbol{x}_i\big] \le \mu_n^{2/\nu}|\mathcal{Q}_n^\delta|\mathbb{E}_n\big[\|\boldsymbol{b}(\boldsymbol{x}_i)\|^2\big] \lesssim \zeta_K^2\mu_n^{2/\nu}|\mathcal{Q}_n^\delta|.$$

Therefore, by the Cauchy-Schwarz inequality,

$$\begin{aligned}
\sum_{i=1}^n \mathbb{E}\big[\|\boldsymbol{g}_i\|^2\|\boldsymbol{g}_i\|_\infty \,\big|\, \boldsymbol{x}_i\big] &\le \sum_{i=1}^n \mathbb{E}\big[\|g_i\|^4 \,\big|\, \boldsymbol{x}_i\big]^{1/2}\mathbb{E}\big[\|g_i\|_\infty^2 \,\big|\, \boldsymbol{x}_i\big]^{1/2} \\
&\lesssim \frac{\zeta_K^3\mu_n^{3/\nu}}{\sqrt{n}}|\mathcal{Q}_n^\delta|\sqrt{\log\big(K|\mathcal{Q}_n^\delta|\big)}.
\end{aligned}$$

Now, since there exists a random variable U_2 independent of $\{\boldsymbol{\xi}_i\}_{i=1}^n \cup \{\boldsymbol{g}_i\}_{i=1}^n$, applying Theorem F.2 with

$$\begin{aligned}
\beta &:= \sum_{i=1}^n \mathbb{E}\big[\|\boldsymbol{\xi}_i\|^2\|\boldsymbol{\xi}_i\|_\infty\big] + \sum_{i=1}^n \mathbb{E}\big[\|\boldsymbol{g}_i\|^2\|\boldsymbol{g}_i\|_\infty\big] \\
&\lesssim \frac{\zeta_K^3\mu_n^{3/\nu}}{\sqrt{n}}|\mathcal{Q}_n^\delta|\sqrt{\log\big(K|\mathcal{Q}_n^\delta|\big)} \\
&\lesssim \frac{\zeta_K^3\mu_n^{3/\nu}}{n^{1/2}\delta_n^{d_\mathcal{Q}}}\sqrt{\log\frac{K}{\delta_n^{d_\mathcal{Q}}}}
\end{aligned}$$

gives that for any $t_n > 0$, on the same probability space there exists a vector $\boldsymbol{Z}_n^\delta \sim \mathcal{N}\big(0, \mathbb{V}[\boldsymbol{G}_n \circ \pi_n^\delta \,|\, \mathbf{X}_n]\big)$, generally different for different t_n, such that

$$\mathbb{P}\big\{\big\|\boldsymbol{G}_n \circ \pi_n^\delta - \boldsymbol{Z}_n^\delta\big\|_\infty > 3t_n\big\} \le \min_{s\ge 0}\big\{2\mathbb{P}\{\|\boldsymbol{N}\|_\infty > s\} + \frac{\beta}{t_n^3}s^2\big\},$$

where $\boldsymbol{N}$ is a $K|\mathcal{Q}_n^\delta|$-dimensional standard Gaussian vector. By the union bound,

$$\mathbb{P}\{\|\boldsymbol{N}\|_\infty > s\} \le 2K|\mathcal{Q}_n^\delta|e^{-s^2/2},$$

so by taking $s := C\sqrt{\log\bigl(K|\mathcal{Q}_n^\delta|\bigr)}$ for a positive constant C not depending on n (chosen later), we have

$$\begin{aligned}\mathbb{P}\{\|\boldsymbol{G}_n \circ \pi_n^\delta - \boldsymbol{Z}_n^\delta\|_\infty > 3t_n\} &\lesssim \bigl(K|\mathcal{Q}_n^\delta|\bigr)^{1-C^2/2} + \frac{\beta}{t_n^3} C^2 \log\bigl(K|\mathcal{Q}_n^\delta|\bigr)\\ &\lesssim \left(\frac{K}{\delta_n^{d_{\mathcal{Q}}}}\right)^{1-C^2/2} + \frac{\beta}{t_n^3} C^2 \log\frac{K}{\delta_n^{d_{\mathcal{Q}}}}\\ &\lesssim \left(\frac{K}{\delta_n^{d_{\mathcal{Q}}}}\right)^{1-C^2/2} + C^2 \frac{\zeta_K^3 \mu_n^{3/\nu}}{t_n^3\sqrt{n}\delta_n^{d_{\mathcal{Q}}}} \log^{3/2}\frac{K}{\delta_n^{d_{\mathcal{Q}}}},\end{aligned}$$

where the constant in $\lesssim$ does not depend on n.

Embedding a conditionally Gaussian vector into a conditionally Gaussian process For a fixed vector in $\boldsymbol{X} \in \mathcal{X}^n$, by standard existence results for Gaussian processes, there exists a mean-zero K-dimensional Gaussian process whose covariance structure is the same as that of $\boldsymbol{G}_n(\boldsymbol{q})$ given $\mathbf{X}_n = \boldsymbol{X}$. It follows from Kolmogorov's continuity criterion that this process can be defined on $\mathcal{C}(\mathcal{Q}, \mathbb{R}^K)$. The laws of such processes define a family of Gaussian probability measures on the Borel σ-algebra $\mathcal{B}(\mathcal{C}(\mathcal{Q}, \mathbb{R}^K))$ of the space $\mathcal{C}(\mathcal{Q}, \mathbb{R}^K)$, denoted $\{\mathbb{P}_{\boldsymbol{X}}\}_{\boldsymbol{X}\in\mathcal{X}^n}$. In order to construct one process that is conditionally on $\mathbf{X}_n'$, a mean zero Gaussian process with the same conditional covariance structure as that of $\boldsymbol{G}_n(\boldsymbol{q})$, where $\mathbf{X}_n' \overset{d}{=} \mathbf{X}_n$, we need to show that this family of measures is a probability kernel as a function $\mathcal{X}^n \times \mathcal{B}(\mathcal{C}(\mathcal{Q}, \mathbb{R}^K)) \to [0,1]$. This follows by a standard argument: we can take a π-system of sets, generating $\mathcal{B}(\mathcal{C}(\mathcal{Q}, \mathbb{R}^K))$, of the form $B = \{\boldsymbol{f} \in \mathcal{C}(\mathcal{Q}, \mathbb{R}^K) : \boldsymbol{f}(\boldsymbol{q}_1) \in B_1, \dots, \boldsymbol{f}(\boldsymbol{q}_m) \in B_m\}$, where $m \in \{1, 2, \dots\}$, $\boldsymbol{q}_j \in \mathcal{Q}$, and each B_j is a parallelepiped in $\mathbb{R}^K$ with edges parallel to the coordinate axes, and notice that for such sets $\boldsymbol{X} \mapsto \mathbb{P}_{\boldsymbol{X}}(B)$ is a Borel function (since a mean-zero Gaussian vector is a linear transformation of a standard Gaussian vector). The sets $A \in \mathcal{B}(\mathcal{C}(\mathcal{Q}, \mathbb{R}^K))$ such that $\boldsymbol{X} \mapsto \mathbb{P}_{\boldsymbol{X}}(A)$ is a Borel function form a λ-system. It is left to apply the monotone class theorem.

We have shown that there exists a law on $\mathcal{X}^n \times \mathcal{C}(\mathcal{Q}, \mathbb{R}^K)$ which is the joint law of $\bigl(\mathbf{X}_n', \{\boldsymbol{Z}_n'(\boldsymbol{q})\}_{\boldsymbol{q}\in\mathcal{Q}}\bigr)$, where $\{\boldsymbol{Z}_n'(\boldsymbol{q})\}_{\boldsymbol{q}\in\mathcal{Q}}$ is a conditionally on $\mathbf{X}_n'$ mean zero Gaussian process with the same conditional covariance structure as that of $\boldsymbol{G}_n'(\boldsymbol{q})$ (where $\boldsymbol{G}_n'(\boldsymbol{q})$ is the same function of $\mathbf{X}_n$ as $\boldsymbol{G}_n(\boldsymbol{q})$ except $\mathbf{X}_n$ is replaced by $\mathbf{X}_n'$). Projecting this $\mathcal{C}(\mathcal{Q}, \mathbb{R}^K)$-process onto $\mathcal{Q}_n^\delta$, we obtain a vector $\boldsymbol{Z}_n^{\delta\prime}$ such that

$$\bigl(\mathbf{X}_n', \boldsymbol{Z}_n^{\delta\prime}\bigr) \overset{d}{=} \bigl(\mathbf{X}_n, \boldsymbol{Z}_n^\delta\bigr).$$

Since $\mathcal{C}(\mathcal{Q}, \mathbb{R}^K)$ is Polish and there exists a uniformly distributed random variable U_3 independent of $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \cup \{\boldsymbol{Z}_n^\delta\}$, we can apply Theorem 8.17 (transfer) in [28], and obtain that there exists (on the same probability space) a random element $\{\boldsymbol{Z}_n(\boldsymbol{q})\}_{\boldsymbol{q}\in\mathcal{Q}} \in \mathcal{C}(\mathcal{Q}, \mathbb{R}^K)$ such that

$$\bigl(\mathbf{X}_n, \boldsymbol{Z}_n^\delta, \{\boldsymbol{Z}_n(\boldsymbol{q})\}_{\boldsymbol{q}\in\mathcal{Q}}\bigr) = \bigl(\mathbf{X}_n', \boldsymbol{Z}_n^{\delta\prime}, \{\boldsymbol{Z}_n'(\boldsymbol{q})\}_{\boldsymbol{q}\in\mathcal{Q}}\bigr).$$

In particular, $\{\boldsymbol{Z}_n(\boldsymbol{q})\}_{\boldsymbol{q}\in\mathcal{Q}}$ is the conditionally Gaussian process whose projection on $\mathcal{Q}_n^\delta$ is the vector $\boldsymbol{Z}_n^\delta$ a. s.

Discretization of Z_n Consider the stochastic process X_n defined for $t = (l, \boldsymbol{q}, \tilde{\boldsymbol{q}}) \in T$ with

$$T := \{(l, \boldsymbol{q}, \tilde{\boldsymbol{q}}) \colon l \in \{1, 2, \dots, K\}, \boldsymbol{q}, \tilde{\boldsymbol{q}} \in \mathcal{Q}, \|\boldsymbol{q} - \tilde{\boldsymbol{q}}\| \le \delta_n\}$$

as $X_{n,t} := Z_{n,l}(\boldsymbol{q}) - Z_{n,l}(\tilde{\boldsymbol{q}})$. It is a separable (because each $Z_{n,l}(\cdot)$ has a. s. continuous trajectories) mean-zero Gaussian conditionally on $\mathbf{X}_n$ process with the index set T considered a metric space: $\operatorname{dist}\big((l,\boldsymbol{q},\tilde{\boldsymbol{q}}),(l',\boldsymbol{q}',\tilde{\boldsymbol{q}}')\big) = \|\boldsymbol{q}-\boldsymbol{q}'\| + \|\tilde{\boldsymbol{q}}-\tilde{\boldsymbol{q}}'\| + \mathbb{1}\{l \neq l'\}$.

We will apply Lemma C.7 to this process. Note that on the event $\mathcal{A}$

$$
\begin{aligned}
\sigma(X_n)^2 &:= \sup_{t\in T}\mathbb{E}\big[X_{n,t}^2 \,\big|\, \mathbf{X}_n\big] = \sup_{\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|\le\delta_n}\max_l \mathbb{E}\big[\big(Z_{n,l}(\boldsymbol{q}) - Z_{n,l}(\tilde{\boldsymbol{q}})\big)^2 \,\big|\, \mathbf{X}_n\big]\\
&= \sup_{\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|\le\delta_n}\max_l \mathbb{E}\big[\big(G_{n,l}(\boldsymbol{q}) - G_{n,l}(\tilde{\boldsymbol{q}})\big)^2 \,\big|\, \mathbf{X}_n\big]\\
&= \sup_{\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|\le\delta_n}\max_l \frac{1}{n}\sum_{i=1}^n b_l(\boldsymbol{x}_i)^2\mathbb{E}\big[\big(A_n(\boldsymbol{q},\boldsymbol{x}_i,y_i) - A_n(\tilde{\boldsymbol{q}},\boldsymbol{x}_i,y_i)\big)^2 \,\big|\, \boldsymbol{x}_i\big]\\
&\lesssim \sup_{\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|\le\delta_n}\max_l \frac{1}{n}\sum_{i=1} b_l(\boldsymbol{x}_i)^2\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\| \le \delta_n \max_l \mathbb{E}_n[b_l(\boldsymbol{x}_i)^2]\\
&\le \delta_n \underbrace{\sup_{\|\boldsymbol{\alpha}\|=1}\mathbb{E}_n\big[\big(\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{b}(\boldsymbol{x}_i)\big)^2\big]}_{\le C_{\mathrm{Gr}}} \lesssim \delta_n.
\end{aligned}
$$

Next, we define and bound the semimetric $\rho(t,t')$:

$$
\begin{aligned}
\rho(t,t')^2 &:= \mathbb{E}\big[(X_{n,t} - X_{n,t'})^2 \,\big|\, \mathbf{X}_n\big]\\
&= \mathbb{E}\big[\big(\big(Z_{n,l}(\boldsymbol{q}) - Z_{n,l}(\tilde{\boldsymbol{q}})\big) - \big(Z_{n,l'}(\boldsymbol{q}') - Z_{n,l'}(\tilde{\boldsymbol{q}}')\big)\big)^2 \,\big|\, \mathbf{X}_n\big]\\
&\lesssim \mathbb{E}\big[\big(Z_{n,l}(\boldsymbol{q}) - Z_{n,l'}(\boldsymbol{q}')\big)^2 \,\big|\, \mathbf{X}_n\big] + \mathbb{E}\big[\big(Z_{n,l}(\tilde{\boldsymbol{q}}) - Z_{n,l'}(\tilde{\boldsymbol{q}}')\big)^2 \,\big|\, \mathbf{X}_n\big].
\end{aligned}
$$

The first term on the right is bounded the following way: if $l \neq l'$,

$$
\begin{aligned}
&\mathbb{E}\big[\big(Z_{n,l}(\boldsymbol{q}) - Z_{n,l'}(\boldsymbol{q}')\big)^2 \,\big|\, \mathbf{X}_n\big] = \mathbb{E}\big[\big(G_{n,l}(\boldsymbol{q}) - G_{n,l'}(\boldsymbol{q}')\big)^2 \,\big|\, \mathbf{X}_n\big]\\
&\quad= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[\big(A_n(\boldsymbol{q},\boldsymbol{x}_i,y_i)b_l(\boldsymbol{x}_i) - A_n(\boldsymbol{q}',\boldsymbol{x}_i,y_i)b_{l'}(\boldsymbol{x}_i)\big)^2 \,\big|\, \boldsymbol{x}_i\big]\\
&\quad\lesssim \frac{1}{n}\sum_{i=1}^n b_l(\boldsymbol{x}_i)^2\mathbb{E}\big[\big(A_n(\boldsymbol{q},\boldsymbol{x}_i,y_i) - A_n(\boldsymbol{q}',\boldsymbol{x}_i,y_i)\big)^2 \,\big|\, \boldsymbol{x}_i\big]\\
&\qquad+ \frac{1}{n}\sum_{i=1}^n \big(b_l(\boldsymbol{x}_i)^2 + b_{l'}(\boldsymbol{x}_i)^2\big)\mathbb{E}\big[A_n(\boldsymbol{q}',\boldsymbol{x}_i,y_i)^2 \,\big|\, \boldsymbol{x}_i\big]\\
&\quad\lesssim \|\boldsymbol{q}-\boldsymbol{q}'\|\mathbb{E}_n\big[b_l(\boldsymbol{x}_i)^2\big] + \mu_n^{2/\nu}\big(\mathbb{E}_n[b_l(\boldsymbol{x}_i)^2] + \mathbb{E}_n[b_{l'}(\boldsymbol{x}_i)^2]\big) \lesssim \|\boldsymbol{q}-\boldsymbol{q}'\| + \mu_n^{2/\nu}.
\end{aligned}
$$

Similarly, if $l = l'$, $\mathbb{E}\big[\big(Z_{n,l}(\boldsymbol{q}) - Z_{n,l'}(\boldsymbol{q}')\big)^2 \,\big|\, \mathbf{X}_n\big] \lesssim \|\boldsymbol{q}-\boldsymbol{q}'\|$.

The term $\mathbb{E}\big[\big(Z_{n,l}(\tilde{\boldsymbol{q}}) - Z_{n,l'}(\tilde{\boldsymbol{q}}')\big)^2 \,\big|\, \mathbf{X}_n\big]$ is bounded the same way, and we conclude

$$
\rho(t,t')^2 \lesssim \mu_n^{2/\nu}\big(\|\boldsymbol{q}-\boldsymbol{q}'\| + \|\tilde{\boldsymbol{q}}-\tilde{\boldsymbol{q}}'\| + \mathbb{1}\{l\neq l'\}\big) = \mu_n^{2/\nu}\operatorname{dist}\big((l,\boldsymbol{q},\tilde{\boldsymbol{q}}),(l',\boldsymbol{q}',\tilde{\boldsymbol{q}}')\big).
$$

In other words, we have proven that for some positive constant C_{13} we have on $\mathcal{A}$

$$
\rho(t,t')^2 \le C_{13}\mu_n^{2/\nu}\operatorname{dist}\big((l,\boldsymbol{q},\tilde{\boldsymbol{q}}),(l',\boldsymbol{q}',\tilde{\boldsymbol{q}}')\big).
$$

This means that an $\big(\varepsilon/\mu_n^{1/\nu}\sqrt{C_{13}}\big)^2$-covering of T with respect to $\operatorname{dist}(\cdot)$ induces an ε-covering of T with respect to ρ, and hence

$$
N(T,\rho,\varepsilon) \le N\Bigg(T, \operatorname{dist}(\cdot), \bigg(\frac{\varepsilon}{\mu_n^{1/\nu}\sqrt{C_{13}}}\bigg)^2\Bigg). \tag{F.5}
$$

Therefore,

$$\log N(T,\rho,\varepsilon) \lesssim \log(K\mu_n^{1/\nu}/\varepsilon).$$

Applying Lemma C.7 gives that on the event $\mathcal{A}$

$$\begin{aligned}\mathbb{E}\Big[\sup_{t\in T}|X_{n,t}|\,\Big|\,\mathbf{X}_n\Big] &\lesssim \sigma(X_n) + \int_0^{2\sigma(X_n)} \sqrt{\log N(T,\rho,\varepsilon)}\,\mathrm{d}\varepsilon \\ &\lesssim \sigma(X_n) + \sigma(X_n)\sqrt{\log\big(K\mu_n^{1/\nu}/\sigma(X_n)\big)} \overset{\text{(a)}}{\lesssim} \left(\delta_n \log\left(K\frac{\mu_n^{1/\nu}}{\delta_n}\right)\right)^{1/2},\end{aligned}$$

where in (a) we used our bound $\sigma(X_n) \lesssim \sqrt{\delta_n}$ above and that $x \mapsto x\log\frac{1}{x}$ is increasing for sufficiently small x. Rewriting and applying Markov's inequality, we obtain that on $\mathcal{A}$

$$\mathbb{P}\left\{\sup_{\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|\le\delta_n} \|\boldsymbol{Z}_n(\boldsymbol{q}) - \boldsymbol{Z}_n(\tilde{\boldsymbol{q}})\|_\infty > t_n \,\middle|\, \mathbf{X}_n\right\} \lesssim \frac{1}{t_n}\sqrt{\delta_n \log\frac{K\mu_n^{1/\nu}}{\delta_n}},$$

where the constant in $\lesssim$ does not depend on n. Since $\mathbb{P}\{\mathcal{A}\}\to 1$, this implies

$$\mathbb{P}\left\{\sup_{\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|\le\delta_n} \|\boldsymbol{Z}_n(\boldsymbol{q}) - \boldsymbol{Z}_n(\tilde{\boldsymbol{q}})\|_\infty > t_n\right\} \lesssim \frac{1}{t_n}\sqrt{\delta_n \log\frac{K\mu_n^{1/\nu}}{\delta_n}} + o(1). \tag{F.6}$$

Choosing δ_n and conclusion Combining the bounds obtained above, for any given positive sequence t_n and any constant $C>0$ of our choice

$$\begin{aligned}\mathbb{P}\left\{\sup_{\boldsymbol{q}\in\mathcal{Q}}\|\boldsymbol{G}_n(\boldsymbol{q}) - \boldsymbol{Z}_n(\boldsymbol{q})\|_\infty > t_n\right\} &\le \mathbb{P}\left\{\sup_{\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|\le\delta_n}\|\boldsymbol{G}_n(\boldsymbol{q}) - \boldsymbol{G}_n(\tilde{\boldsymbol{q}})\|_\infty > t_n/3\right\} \\ &\quad + \mathbb{P}\left\{\big\|\boldsymbol{G}_n\circ\pi_n^\delta - \boldsymbol{Z}_n^\delta\big\|_\infty > t_n/3\right\} + \mathbb{P}\left\{\sup_{\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|\le\delta_n}\|Z_n(\boldsymbol{q}) - Z_n(\tilde{\boldsymbol{q}})\|_\infty > t_n/3\right\} \\ &\lesssim \frac{1}{t_n}\sqrt{\delta_n\log\frac{K\zeta_K\mu_n^{1/\nu}}{\sqrt{\delta_n}}} + \frac{\zeta_K(\mu_n n)^{1/\nu}}{t_n\sqrt{n}}\log\frac{K\zeta_K\mu_n^{1/\nu}}{\sqrt{\delta_n}} \\ &\quad + \left(\frac{K}{\delta_n^{d_\mathcal{Q}}}\right)^{1-C^2/2} + C^2\frac{\zeta_K^3\mu_n^{3/\nu}}{t_n^3\sqrt{n}\delta_n^{d_\mathcal{Q}}}\log^{3/2}\frac{K}{\delta_n^{d_\mathcal{Q}}} \\ &\quad + \frac{1}{t_n}\sqrt{\delta_n\log\frac{K\mu_n^{1/\nu}}{\delta_n}} + o(1),\end{aligned}$$

where the constant in $\lesssim$ does not depend on n.

Take, for example, $C=2$ (so that $1-C^2/2$ is negative). Now we approximately (assuming that each $\log(\cdot)$ on the right is $O(\log n)$ and ignoring constant coefficients) optimize this over δ_n. This gives

$$\delta_n := \left(\frac{\zeta_K^3\mu_n^{3/\nu}}{t_n^2\sqrt{n}}\log n\right)^{\frac{2}{1+2d_\mathcal{Q}}}.$$

Let $\ell_n\to 0$ be a positive sequence satisfying

$$\left(\frac{\zeta_K^3\mu_n^{3/\nu}}{\sqrt{n}}\right)^{\frac{1}{3+2d_\mathcal{Q}}}\sqrt{\log n} + \frac{\zeta_K\mu_n^{1/\nu}}{n^{1/2-1/\nu}}\log n = o(\ell_n r_{\mathtt{YUR}}).$$

Clearly, whenever (F.3) holds, such a sequence exists. Putting $t_n := \ell_n r_{\text{yur}}$, we get

$$\mathbb{P}\Big\{\sup_{\boldsymbol{q}\in\mathcal{Q}}\|\boldsymbol{G}_n(\boldsymbol{q})-\boldsymbol{Z}_n(\boldsymbol{q})\|_\infty > \ell_n r_{\text{yur}}\Big\} = o(1).$$

Fix $\varepsilon > 0$. For n large enough, $\ell_n < \varepsilon$. Then for these n we have

$$\mathbb{P}\Big\{\sup_{\boldsymbol{q}\in\mathcal{Q}}\|\boldsymbol{G}_n(\boldsymbol{q})-\boldsymbol{Z}_n(\boldsymbol{q})\|_\infty > \varepsilon r_{\text{yur}}\Big\} \le \mathbb{P}\Big\{\sup_{\boldsymbol{q}\in\mathcal{Q}}\|\boldsymbol{G}_n(\boldsymbol{q})-\boldsymbol{Z}_n(\boldsymbol{q})\|_\infty > \ell_n r_{\text{yur}}\Big\} = o(1).$$

Lemma F.3 is proven. □

F.2 Main Result

We begin by presenting our main strong approximation result, which is a special case of more technical and lengthy Lemma F.3. To simplify exposition, the following notation will be helpful from this point onwards:

$$\hat{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) := -h^{d/2}\frac{\widehat{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})}{\sqrt{\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})}}, \tag{F.7}$$

$$\bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) := -h^{d/2}\frac{\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})}{\sqrt{\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})}}, \tag{F.8}$$

$$\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) := -h^{d/2}\frac{\mathbf{Q}_{0,\boldsymbol{q}}^{-1}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})}{\sqrt{\Omega_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})}}, \tag{F.9}$$

$$t_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) := h^{-d/2}\bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\sqrt{n}\mathbb{E}_n\big[\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}));\boldsymbol{q})\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))\boldsymbol{p}(\boldsymbol{x}_i)\big], \tag{F.10}$$

$$T_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) := \frac{\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}}\widehat{\boldsymbol{\beta}}(\boldsymbol{q})-\mu_0^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})}{\sqrt{\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})/n}}, \tag{F.11}$$

where $\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})$ is some feasible estimator of variance.

Theorem F.4 (Strong approximation)**.**

(a) Suppose Assumptions B.1 to B.6 hold with $\nu \ge 3$. Furthermore, assume the following conditions hold:

$$\sup_{\boldsymbol{q}\in\mathcal{Q}}\sup_{\boldsymbol{x}\in\mathcal{X}}|\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})-\mu_0^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})| \lesssim_{\mathbb{P}} h^{-|\boldsymbol{v}|}r_{\mathtt{UC}}; \tag{F.12}$$

$$\sup_{\boldsymbol{q}\in\mathcal{Q}}\sup_{\boldsymbol{x}\in\mathcal{X}}|\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})-\mu_0^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})-\mathsf{L}^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})| \lesssim_{\mathbb{P}} h^{-|\boldsymbol{v}|}r_{\mathtt{BR}} \text{ with } \frac{\log n}{nh^d} \lesssim r_{\mathtt{BR}}; \tag{F.13}$$

$$\sup_{\boldsymbol{q}\in\mathcal{Q}}\sup_{\boldsymbol{x}\in\mathcal{X}}\big|\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})-\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\big| \lesssim_{\mathbb{P}} h^{-2|\boldsymbol{v}|-d}r_{\mathtt{VC}} \text{ with } r_{\mathtt{VC}} = o(1); \tag{F.14}$$

$$\begin{aligned}\mathbb{E}[(\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}));\boldsymbol{q})\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))-\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\tilde{\boldsymbol{q}}));\tilde{\boldsymbol{q}})\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\tilde{\boldsymbol{q}})))^2\mid\boldsymbol{x}_i]&\\ \lesssim \|\boldsymbol{q}-\tilde{\boldsymbol{q}}\| \text{ for all } \boldsymbol{q},\tilde{\boldsymbol{q}}\in\mathcal{Q}.&\end{aligned} \tag{F.15}$$

Let $r_{\mathtt{SA}}$ is any positive sequence of numbers converging to zero[1] *such that*

$$\Big(\frac{1}{nh^{3d}}\Big)^{\frac{1}{6+4d_{\mathcal{Q}}}}\sqrt{\log n}+\frac{1}{h^{d/2}n^{1/2-1/\nu}}\log n = o(r_{\mathtt{SA}}). \tag{F.16}$$

[1] In particular, the left-hand side of (F.16) is automatically assumed to be $o(1)$.

Denote

$$\boldsymbol{G}(\boldsymbol{q}) := h^{-d/2}\sqrt{n}\mathbb{E}_n\big[\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q})); \boldsymbol{q})\eta^{(1)}(\mu_0(\boldsymbol{x}_i, \boldsymbol{q}))\boldsymbol{p}(\boldsymbol{x}_i)\big].$$

Then (provided the probability space is rich enough) there exists a K-dimensional, conditionally on $\mathbf{X}_n$ mean-zero Gaussian, process $\boldsymbol{Z}(\boldsymbol{q})$ on $\mathcal{Q}$ with a. s. continuous trajectories, satisfying

$$\mathbb{E}[\boldsymbol{G}(\boldsymbol{q})\boldsymbol{G}(\tilde{\boldsymbol{q}})^{\mathsf{T}} \mid \mathbf{X}_n] = \mathbb{E}[\boldsymbol{Z}(\boldsymbol{q})\boldsymbol{Z}(\tilde{\boldsymbol{q}})^{\mathsf{T}} \mid \mathbf{X}_n]; \tag{F.17}$$

$$\sup_{\boldsymbol{q}}\|\boldsymbol{G}(\boldsymbol{q}) - \boldsymbol{Z}(\boldsymbol{q})\|_\infty = o_{\mathbb{P}}(r_{\mathtt{SA}}); \tag{F.18}$$

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\mathcal{Q}}\big|\bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q})^{\mathsf{T}}\boldsymbol{G}(\boldsymbol{q}) - \bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q})^{\mathsf{T}}\boldsymbol{Z}(\boldsymbol{q})\big| = o_{\mathbb{P}}(r_{\mathtt{SA}}), \tag{F.19}$$

$$\sup_{\boldsymbol{q},\boldsymbol{x}}\big|T_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q}) - \bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q})^{\mathsf{T}}\boldsymbol{Z}(\boldsymbol{q})\big| \lesssim_{\mathbb{P}} \sqrt{nh^d}(r_{\mathtt{UC}}r_{\mathtt{VC}} + r_{\mathtt{BR}}) + o(r_{\mathtt{SA}}). \tag{F.20}$$

(b) If, in addition to the previous conditions, $\overline{\psi}(\boldsymbol{x}_i, y_i)$ is σ^2-sub-Gaussian conditionally on $\boldsymbol{x}_i$, and $r_{\mathtt{SA}}^{\mathrm{sub}} \to 0$ is a positive sequence of numbers satisfying

$$\left(\frac{1}{nh^{3d}}\right)^{\frac{1}{6+4d_{\mathcal{Q}}}}\sqrt{\log n} + \frac{\log^{3/2} n}{\sqrt{nh^d}} = o(r_{\mathtt{SA}}^{\mathrm{sub}}),$$

then

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\mathcal{Q}}\big|\bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q})^{\mathsf{T}}\boldsymbol{G}(\boldsymbol{q}) - \bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q})^{\mathsf{T}}\boldsymbol{Z}(\boldsymbol{q})\big| = o_{\mathbb{P}}(r_{\mathtt{SA}}^{\mathrm{sub}}). \tag{F.21}$$

Proof. Note that (F.18) implies (F.19) because $\sup_{\boldsymbol{q},\boldsymbol{x}}\big\|\bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q})\big\|_1 \lesssim_{\mathbb{P}} 1$ by Lemma F.7 below. To prove (F.18), apply Lemma F.3 with $\boldsymbol{G}(\cdot)$ as in the statement,

$$\begin{aligned}
&A_n(\boldsymbol{q}, y) = A_n(\boldsymbol{q}, \boldsymbol{x}, y) := \psi(y, \eta(\mu_0(\boldsymbol{x}, \boldsymbol{q})); \boldsymbol{q})\eta^{(1)}(\mu_0(\boldsymbol{x}, \boldsymbol{q})),\\
&\boldsymbol{b}(\cdot) := h^{-d/2}\boldsymbol{p}(\cdot),\\
&\zeta_K \lesssim K^{1/2} \lesssim h^{-d/2},\\
&\mu_n \lesssim 1.
\end{aligned}$$

Finally, (F.20) follows from combining (F.19) with Lemma F.8. □

Remark F.5. This theorem contains Theorem 2 in the main paper; the notation there is simplified for better readability. Specifically, $T(\boldsymbol{x}, q)$ in the main paper corresponds to $T_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q})$ here and the approximating process $Z(\boldsymbol{x}, q)$ corresponds to $\bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q})^{\mathsf{T}}\boldsymbol{Z}(\boldsymbol{q})$ here. The reason for differing notation is that we require more precision here: given the form of Lemma F.3, it is more natural to use K-dimensional processes such as $\boldsymbol{Z}(\boldsymbol{q})$, also making for simpler presentation of precise results in Appendix G later.

The rest of this subsection will be devoted to proving the lemmas used in the proof of Theorem F.4. First, we prove bounds on the asymptotic variance and its approximation. Similar bounds were proven in [9, 11].

Lemma F.6 (Asymptotic variance)**.** *Suppose Assumptions Assumptions B.1 to B.6 hold, and $\frac{\log(1/h)}{nh^d} = o(1)$. Then*

$$h^{-d-2|\boldsymbol{v}|} \lesssim \inf_{\boldsymbol{q}\in\mathcal{Q}}\inf_{\boldsymbol{x}\in\mathcal{X}}\Omega_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q}) \le \sup_{\boldsymbol{q}\in\mathcal{Q}}\sup_{\boldsymbol{x}\in\mathcal{X}}\Omega_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q}) \lesssim h^{-d-2|\boldsymbol{v}|}, \tag{F.22}$$

$$\sup_{\boldsymbol{q}\in\mathcal{Q}}\sup_{\boldsymbol{x}\in\mathcal{X}}\big|\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q}) - \Omega_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q})\big| \lesssim_{\mathbb{P}} h^{-d-2|\boldsymbol{v}|}\sqrt{\frac{\log(1/h)}{nh^d}}, \tag{F.23}$$

$$h^{-d-2|\boldsymbol{v}|} \lesssim_{\mathbb{P}} \inf_{\boldsymbol{q}\in\mathcal{Q}}\inf_{\boldsymbol{x}\in\mathcal{X}}\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q}) \le \sup_{\boldsymbol{q}\in\mathcal{Q}}\sup_{\boldsymbol{x}\in\mathcal{X}}\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q}) \lesssim_{\mathbb{P}} h^{-d-2|\boldsymbol{v}|}. \tag{F.24}$$

Proof. For the lower bound in (F.22), we have

$$\begin{aligned}\Omega_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) &\geq \lambda_{\min}(\boldsymbol{\Sigma}_{0,\boldsymbol{q}})\cdot\|\mathbf{Q}_{0,\boldsymbol{q}}^{-1}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\|^2 \geq \lambda_{\min}(\boldsymbol{\Sigma}_{0,\boldsymbol{q}})\cdot\|\mathbf{Q}_{0,\boldsymbol{q}}\|^{-2}\|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\|^2\\ &\gtrsim \lambda_{\min}(\boldsymbol{\Sigma}_{0,\boldsymbol{q}})\cdot h^{-2d}\cdot h^{-2|\boldsymbol{v}|} \gtrsim h^d\cdot h^{-2d}\cdot h^{-2|\boldsymbol{v}|} = h^{-d-2|\boldsymbol{v}|}\end{aligned} \tag{F.25}$$

by Assumptions B.2, B.4 ($\sigma_{\boldsymbol{q}}^2(\boldsymbol{x})$ and $\eta^{(1)}\big(\mu_0(\boldsymbol{x},\boldsymbol{q})\big)^2$ are uniformly over $\boldsymbol{x}$ bounded away from zero) and Lemma C.11.

For the upper bound in (F.22), we have

$$\begin{aligned}\Omega_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) &\leq \lambda_{\max}(\boldsymbol{\Sigma}_{0,\boldsymbol{q}})\cdot\big\|\mathbf{Q}_{0,\boldsymbol{q}}^{-1}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\big\|^2 \leq \lambda_{\max}(\boldsymbol{\Sigma}_{0,\boldsymbol{q}})\cdot\big[\lambda_{\max}(\mathbf{Q}_{0,\boldsymbol{q}}^{-1})\big]^2\big\|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\big\|^2\\ &\overset{\text{(a)}}{\lesssim} \lambda_{\max}(\boldsymbol{\Sigma}_{0,\boldsymbol{q}})\cdot h^{-2d}\cdot h^{-2|\boldsymbol{v}|} \overset{\text{(b)}}{\lesssim} h^d\cdot h^{-2d}\cdot h^{-2|\boldsymbol{v}|} = h^{-d-2|\boldsymbol{v}|},\end{aligned}$$

where (a) is by Assumption B.2 and Lemma C.11, (b) is by Assumption B.4 ($\sigma_{\boldsymbol{q}}^2(\boldsymbol{x})$ and $\eta^{(1)}\big(\mu_0(\boldsymbol{x},\boldsymbol{q})\big)^2$ are uniformly over $\boldsymbol{x}$ bounded) and Lemma C.11.

We will now prove (F.23). We start by noticing

$$\begin{aligned}&\sup_{\boldsymbol{q},\boldsymbol{x}}\big|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}}\mathbf{Q}_{0,\boldsymbol{q}}^{-1}(\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q}}-\boldsymbol{\Sigma}_{0,\boldsymbol{q}})\mathbf{Q}_{0,\boldsymbol{q}}^{-1}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\big|\\ &\quad\lesssim_{\mathbb{P}} \sup_{\boldsymbol{q},\boldsymbol{x}}\|\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q}}-\boldsymbol{\Sigma}_{0,\boldsymbol{q}}\|\cdot\|\mathbf{Q}_{0,\boldsymbol{q}}^{-1}\|^2\cdot\|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\|^2\\ &\quad\overset{\text{(a)}}{\lesssim_{\mathbb{P}}} h^{-2d}\cdot h^{-2|\boldsymbol{v}|}\sup_{\boldsymbol{q}}\|\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q}}-\boldsymbol{\Sigma}_{0,\boldsymbol{q}}\| \overset{\text{(b)}}{\lesssim_{\mathbb{P}}} h^{-2d-2|\boldsymbol{v}|}h^d\sqrt{\frac{\log(1/h)}{nh^d}} = h^{-d-2|\boldsymbol{v}|}\sqrt{\frac{\log(1/h)}{nh^d}},\end{aligned}$$

where in (a) we used $\sup_{\boldsymbol{q}}\|\mathbf{Q}_{0,\boldsymbol{q}}^{-1}\| \lesssim_{\mathbb{P}} h^{-d}$ by Lemma C.11 and $\sup_{\boldsymbol{x}}\|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\| \lesssim h^{-|\boldsymbol{v}|}$ by Assumption B.2, in (b) we used $\sup_{\boldsymbol{q}}\|\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q}}-\boldsymbol{\Sigma}_{0,\boldsymbol{q}}\| \lesssim_{\mathbb{P}} h^d\sqrt{\frac{\log(1/h)}{nh^d}}$ which is proven by the same argument as $\sup_{\boldsymbol{q}}\|\bar{\mathbf{Q}}_{\boldsymbol{q}}-\mathbf{Q}_{0,\boldsymbol{q}}\| \lesssim_{\mathbb{P}} h^d\sqrt{\frac{\log(1/h)}{nh^d}}$ in Lemma C.11. Similarly,

$$\begin{aligned}&\sup_{\boldsymbol{q},\boldsymbol{x}}\big|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}}(\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}-\mathbf{Q}_{0,\boldsymbol{q}}^{-1})\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q}}\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\big|\\ &\quad\leq \sup_{\boldsymbol{q},\boldsymbol{x}}\|\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}-\mathbf{Q}_{0,\boldsymbol{q}}^{-1}\|\cdot\|\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\|\cdot\|\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q}}\|\cdot\|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\|^2\\ &\quad\lesssim_{\mathbb{P}} h^{-d}\sqrt{\frac{\log(1/h)}{nh^d}}\cdot h^{-d}\cdot h^d\cdot h^{-2|\boldsymbol{v}|} = h^{-d-2|\boldsymbol{v}|}\sqrt{\frac{\log(1/h)}{nh^d}}.\end{aligned}$$

Finally,

$$\sup_{\boldsymbol{q},\boldsymbol{x}}\big|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}}\mathbf{Q}_{0,\boldsymbol{q}}^{-1}\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q}}(\mathbf{Q}_{0,\boldsymbol{q}}^{-1}-\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1})\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\big| \lesssim_{\mathbb{P}} h^{-d-2|\boldsymbol{v}|}\sqrt{\frac{\log(1/h)}{nh^d}}.$$

Combining the bounds above gives $\sup_{\boldsymbol{q},\boldsymbol{x}}\big|\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})-\Omega_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\big| \lesssim_{\mathbb{P}} h^{-d-2|\boldsymbol{v}|}\sqrt{\frac{\log(1/h)}{nh^d}}$.

(F.24) is an immediate consequence of (F.22) and (F.23), since $\frac{\log(1/h)}{nh^d} = o(1)$. $\square$

Lemma F.7 (Closeness of linear terms)**.** *Suppose Assumptions B.1 to B.6,* (F.14) *hold, and*

$\frac{\log(1/h)}{nh^d} = o(1)$. *Then*

$$\sup_{\boldsymbol{q},\boldsymbol{x}} \left| \widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{-1/2} - \bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{-1/2} \right| \lesssim_{\mathbb{P}} h^{d/2+|\boldsymbol{v}|} r_{\mathtt{VC}}, \tag{F.26}$$

$$\sup_{\boldsymbol{q},\boldsymbol{x}} \left| \bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{-1/2} - \Omega_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{-1/2} \right| \lesssim_{\mathbb{P}} h^{d/2+|\boldsymbol{v}|} \sqrt{\frac{\log(1/h)}{nh^d}}, \tag{F.27}$$

$$\sup_{\boldsymbol{q},\boldsymbol{x}} \|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\|_1 \lesssim 1, \tag{F.28}$$

$$\sup_{\boldsymbol{q},\boldsymbol{x}} \|\bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) - \boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\|_1 \lesssim_{\mathbb{P}} \sqrt{\frac{\log(1/h)}{nh^d}}, \tag{F.29}$$

$$\sup_{\boldsymbol{q},\boldsymbol{x}} \|\bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\|_1 \lesssim 1 \quad \textit{w. p. a. 1.} \tag{F.30}$$

If, in addition, Assumption B.7 holds, then

$$\sup_{\boldsymbol{q},\boldsymbol{x}} \|\hat{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) - \bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\|_1 \lesssim_{\mathbb{P}} r_{\mathtt{Q}} + r_{\mathtt{VC}}, \tag{F.31}$$

$$\sup_{\boldsymbol{q},\boldsymbol{x}} \|\hat{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\|_1 \lesssim 1 \quad \textit{w. p. a. 1.} \tag{F.32}$$

Proof. Equation (F.26) follows from the following chain of inequalities:

$$\begin{aligned}
&\sup_{\boldsymbol{q},\boldsymbol{x}} \left| \widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{-1/2} - \bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{-1/2} \right| \\
&\quad \le \sup_{\boldsymbol{q},\boldsymbol{x}} \left[\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) \bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\right]^{-1/2} \left[\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{1/2} + \bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{1/2}\right]^{-1} \sup_{\boldsymbol{q},\boldsymbol{x}} \left| \widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) - \bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) \right| \\
&\quad \overset{\text{(a)}}{\lesssim_{\mathbb{P}}} (h^{d+2|\boldsymbol{v}|})^{3/2} \sup_{\boldsymbol{q},\boldsymbol{x}} \left| \widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) - \bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) \right| \overset{\text{(b)}}{\lesssim_{\mathbb{P}}} h^{d/2+|\boldsymbol{v}|} r_{\mathtt{VC}},
\end{aligned}$$

where in (a) we used that $\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})$ is close to $\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})$ by (F.14) and the lower bound on $\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})$ from Lemma F.6, and in (b) we used (F.14).

To prove (F.27), recall that

$$\begin{aligned}
&\sup_{\boldsymbol{q},\boldsymbol{x}} \left| \bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) - \Omega_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) \right| \lesssim_{\mathbb{P}} h^{-d-2|\boldsymbol{v}|} \sqrt{\frac{\log(1/h)}{nh^d}}, \quad \text{and} \\
&\inf_{\boldsymbol{q},\boldsymbol{x}} \bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) \gtrsim_{\mathbb{P}} h^{-d-2|\boldsymbol{v}|}
\end{aligned}$$

by Lemma F.6. Therefore,

$$\begin{aligned}
&\sup_{\boldsymbol{q},\boldsymbol{x}} \left| \bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{-1/2} - \Omega_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{-1/2} \right| \\
&\quad \le \sup_{\boldsymbol{q},\boldsymbol{x}} \left[\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) \Omega_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\right]^{-1/2} \left[\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{1/2} + \Omega_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{1/2}\right]^{-1} \sup_{\boldsymbol{q},\boldsymbol{x}} \left| \bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) - \Omega_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) \right| \\
&\quad \lesssim_{\mathbb{P}} (h^{d+2|\boldsymbol{v}|})^{3/2} \sup_{\boldsymbol{q},\boldsymbol{x}} \left| \bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) - \Omega_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) \right| \lesssim_{\mathbb{P}} h^{3d/2+3|\boldsymbol{v}|} h^{-d-2|\boldsymbol{v}|} \sqrt{\frac{\log(1/h)}{nh^d}} \\
&\quad = h^{d/2+|\boldsymbol{v}|} \sqrt{\frac{\log(1/h)}{nh^d}}.
\end{aligned}$$

Equation (F.28) follows from

$$h^{d/2}\|\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\|_1 \le h^{d/2}\|\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\|_1 \cdot \|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\|_1 \lesssim_{\mathbb{P}} h^{-d/2-|\boldsymbol{v}|}$$

and Lemma F.6.

We will now prove (F.29). We have that

$$\sup_{\boldsymbol{q}}\|\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\|_1 \lesssim_{\mathbb{P}} h^{-d}$$

by Lemma C.11, and

$$\sup_{\boldsymbol{x}}\|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\|_1 \lesssim h^{-|\boldsymbol{v}|}$$

by Assumption B.2. Combining this with (F.27) gives

$$\left\|\frac{h^{d/2}\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})}{\sqrt{\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})}} - \frac{h^{d/2}\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})}{\sqrt{\Omega_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})}}\right\|_1 \lesssim_{\mathbb{P}} \sqrt{\frac{\log(1/h)}{nh^d}}.$$

It is left to bound

$$\begin{aligned}
&\sup_{\boldsymbol{q},\boldsymbol{x}} \frac{h^{d/2}}{\sqrt{\Omega_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})}}\big\|(\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1} - \mathbf{Q}_{0,\boldsymbol{q}}^{-1})\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\big\|_1 \\
&\overset{\text{(a)}}{\lesssim} h^{d+|\boldsymbol{v}|}\sup_{\boldsymbol{q},\boldsymbol{x}}\|\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1} - \mathbf{Q}_{0,\boldsymbol{q}}^{-1}\|_1\|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\|_1 \\
&\overset{\text{(b)}}{\lesssim_{\mathbb{P}}} h^{d+|\boldsymbol{v}|}\cdot h^{-d}\sqrt{\frac{\log(1/h)}{nh^d}}\cdot h^{-|\boldsymbol{v}|} = \sqrt{\frac{\log(1/h)}{nh^d}},
\end{aligned}$$

where in (a) we used $\inf_{\boldsymbol{q},\boldsymbol{x}}\Omega_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) \gtrsim h^{-d-2|\boldsymbol{v}|}$ by Lemma F.6, in (b) we used $\|\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1} - \mathbf{Q}_{0,\boldsymbol{q}}^{-1}\|_1 \lesssim_{\mathbb{P}} h^{-d}\sqrt{\log(1/h)/(nh^d)}$ by Lemma C.11 and $\|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\|_1 \lesssim h^{-|\boldsymbol{v}|}$ by Assumption B.2.

Equation (F.30) follows from (F.29), (F.28) and $\log(1/h)/(nh^d) = o(1)$.

We will now prove (F.31). By the triangle inequality,

$$\sup_{\boldsymbol{q},\boldsymbol{x}}\left\|\frac{\widehat{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})}{\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{1/2}} - \frac{\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})}{\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{1/2}}\right\|_1 \tag{F.33}$$

$$\lesssim \sup_{\boldsymbol{q},\boldsymbol{x}}\left\|\frac{(\widehat{\mathbf{Q}}_{\boldsymbol{q}}^{-1} - \bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1})\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})}{\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{1/2}}\right\|_1 + \sup_{\boldsymbol{q},\boldsymbol{x}}\big\|\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\big(\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{-1/2} - \bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{-1/2}\big)\big\|_1. \tag{F.34}$$

To bound the first term in (F.34), recall that $\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) \gtrsim_{\mathbb{P}} h^{-d-2|\boldsymbol{v}|}$ by Lemma F.6 and (F.14). Then

$$\begin{aligned}
\sup_{\boldsymbol{q},\boldsymbol{x}}\left\|\frac{(\widehat{\mathbf{Q}}_{\boldsymbol{q}}^{-1} - \bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1})\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})}{\sqrt{\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})}}\right\|_1 &\lesssim_{\mathbb{P}} h^{d/2+|\boldsymbol{v}|}\sup_{\boldsymbol{q},\boldsymbol{x}}\big\|(\widehat{\mathbf{Q}}_{\boldsymbol{q}}^{-1} - \bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1})\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\big\|_1 \\
&\lesssim h^{d/2+|\boldsymbol{v}|}\sup_{\boldsymbol{q},\boldsymbol{x}}\|\widehat{\mathbf{Q}}_{\boldsymbol{q}}^{-1} - \bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\|_1\cdot\|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\|_1 \\
&\lesssim h^{d/2}\|\widehat{\mathbf{Q}}_{\boldsymbol{q}}^{-1} - \bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\|_\infty \lesssim_{\mathbb{P}} h^{d/2}\cdot h^{-d}r_{\mathtt{Q}} = h^{-d/2}r_{\mathtt{Q}},
\end{aligned}$$

where in the last inequality we used $\|\widehat{\mathbf{Q}}_{\boldsymbol{q}} - \bar{\mathbf{Q}}_{\boldsymbol{q}}\|_\infty \lesssim_{\mathbb{P}} h^d r_{\mathtt{Q}}$ by assumption, and

$$\|\widehat{\mathbf{Q}}_{\boldsymbol{q}}^{-1} - \bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\|_\infty \le \|\widehat{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\|_\infty \cdot \|\widehat{\mathbf{Q}}_{\boldsymbol{q}} - \bar{\mathbf{Q}}_{\boldsymbol{q}}\|_\infty \cdot \|\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\|_\infty \lesssim_{\mathbb{P}} h^{-d}(h^d r_{\mathtt{Q}})h^{-d} = h^{-d} r_{\mathtt{Q}}$$

also by assumption and Lemma C.11.

It is left to bound the second term in (F.34):

$$\begin{aligned}
&\sup_{\boldsymbol{q},\boldsymbol{x}}\Big\|\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\big(\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{-1/2} - \bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{-1/2}\big)\Big\|_1 \\
&\quad\le \sup_{\boldsymbol{q}}\|\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\|_1 \sup_{\boldsymbol{x}}\|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\|_1 \sup_{\boldsymbol{q},\boldsymbol{x}}\big|\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{-1/2} - \bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{-1/2}\big| \\
&\quad\overset{\text{(a)}}{\lesssim_{\mathbb{P}}} h^{-d}\cdot h^{-|\boldsymbol{v}|}\cdot\sup_{\boldsymbol{q},\boldsymbol{x}}\big|\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{-1/2} - \bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{-1/2}\big| \\
&\quad\overset{\text{(b)}}{\lesssim_{\mathbb{P}}} h^{-d-|\boldsymbol{v}|}h^{d/2+|\boldsymbol{v}|}r_{\mathtt{VC}} = h^{-d/2}r_{\mathtt{VC}},
\end{aligned}$$

where (a) is by Lemma C.11 and Assumption B.2, (b) is by (F.26). Equation (F.31) is proven.

Equation (F.32) follows from (F.31), (F.30) and $r_{\mathtt{Q}} + r_{\mathtt{VC}} = o(1)$. □

Lemma F.8 (Hats off)**.** *Suppose Assumptions B.1 to B.6, (F.12), (F.13) and (F.14) hold,* $\nu \ge 3$, *and* $\frac{\log^3 n}{nh^{3d}} = o(1)$. *Define (and fix for all further arguments)* $r_{\mathtt{HO}}$ *as an arbitrary positive sequence satisfying*

$$\sqrt{nh^d}(r_{\mathtt{UC}}r_{\mathtt{VC}} + r_{\mathtt{BR}}) = o(r_{\mathtt{HO}}). \tag{F.35}$$

Then

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\mathcal{Q}}|T_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) - t_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})| \lesssim_{\mathbb{P}} \sqrt{nh^d}(r_{\mathtt{UC}}r_{\mathtt{VC}} + r_{\mathtt{BR}}) = o(r_{\mathtt{HO}}). \tag{F.36}$$

Proof. First, note that by Lemma F.6, (F.12) and (F.14)

$$\begin{aligned}
&\sup_{\boldsymbol{q},\boldsymbol{x}}\left|\frac{\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q}) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})}{\sqrt{\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})/n}} - \frac{\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q}) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})}{\sqrt{\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})/n}}\right| \\
&\le \sqrt{n}\sup_{\boldsymbol{q},\boldsymbol{x}}\big|\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q}) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})\big| \sup_{\boldsymbol{q},\boldsymbol{x}}\frac{1}{\sqrt{\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})}\Big(\sqrt{\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})} + \sqrt{\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})}\Big)} \\
&\qquad\cdot\sup_{\boldsymbol{q},\boldsymbol{x}}\big|\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) - \bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\big| \\
&\lesssim_{\mathbb{P}} \sqrt{n}\cdot h^{-|\boldsymbol{v}|}r_{\mathtt{UC}}\cdot(h^{d+2|\boldsymbol{v}|})^{3/2}\cdot h^{-2|\boldsymbol{v}|-d}r_{\mathtt{VC}} \\
&= \sqrt{n}h^{d/2}r_{\mathtt{UC}}r_{\mathtt{VC}}.
\end{aligned}$$

By (F.13), Assumption B.2 and Lemma C.11, Eq. (C.12) in Lemma C.12,

$$\begin{aligned}
&\sup_{\boldsymbol{q},\boldsymbol{x}}\Big|\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q}) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q}) + \boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}}\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}));\boldsymbol{q})]\Big| \\
&\quad\le \sup_{\boldsymbol{q},\boldsymbol{x}}|\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q}) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q}) - \mathsf{L}^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})| \\
&\qquad+\sup_{\boldsymbol{q},\boldsymbol{x}}\|\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})\|_1\big\|\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1} - \mathbf{Q}_{0,\boldsymbol{q}}^{-1}\big\|_\infty\big\|\mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}));\boldsymbol{q})]\big\|_\infty \\
&\quad\lesssim h^{-|\boldsymbol{v}|}r_{\mathtt{BR}} + h^{-d-|\boldsymbol{v}|}\sqrt{\frac{\log(1/h)}{nh^d}}\sqrt{\frac{h^d\log(1/h)}{n}} \\
&\quad= h^{-|\boldsymbol{v}|}r_{\mathtt{BR}} + h^{-|\boldsymbol{v}|}\frac{\log(1/h)}{nh^d} \overset{\text{(a)}}{\lesssim} h^{-|\boldsymbol{v}|}r_{\mathtt{BR}},
\end{aligned} \tag{F.37}$$

where in (a) we used $\frac{\log(1/h)}{nh^d} \lesssim \frac{\log n}{nh^d} \lesssim r_{\mathtt{BR}}$. Combining this with $h^{-2|\boldsymbol{v}|-d} \lesssim_{\mathbb{P}} \inf_{\boldsymbol{q},\boldsymbol{x}} |\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})|$ (by Lemma F.6)

$$\sup_{\boldsymbol{q},\boldsymbol{x}} \left| \frac{\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^\mathsf{T} \widehat{\boldsymbol{\beta}}(\boldsymbol{q}) - \mu_0^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q})}{\sqrt{\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})/n}} - t_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) \right| \lesssim_{\mathbb{P}} \sqrt{n}\big(h^{d+2|\boldsymbol{v}|}\big)^{1/2} h^{-|\boldsymbol{v}|} r_{\mathtt{BR}} = \sqrt{nh^d}\, r_{\mathtt{BR}}. \tag{F.38}$$

It is left to apply the triangle inequality. □

G Uniform inference

G.1 Plug-in approximation

Recall that the conditional covariance structure of $\boldsymbol{Z}(\cdot)$ is

$$\mathbb{E}\big[\boldsymbol{Z}(\boldsymbol{q})\boldsymbol{Z}(\tilde{\boldsymbol{q}})^\mathsf{T} \,\big|\, \mathbf{X}_n\big] = h^{-d} \bar{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$$

with

$$\begin{aligned} \bar{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}} &:= \mathbb{E}_n\big[S_{\boldsymbol{q},\tilde{\boldsymbol{q}}}(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\tilde{\boldsymbol{q}}))\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\big] \in \mathcal{C}(\mathcal{Q}\times\mathcal{Q}, \mathbb{R}^{K\times K}), \\ S_{\boldsymbol{q},\tilde{\boldsymbol{q}}}(\boldsymbol{x}) &:= \mathbb{E}\big[\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q})); \boldsymbol{q})\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i,\tilde{\boldsymbol{q}})); \tilde{\boldsymbol{q}}) \,\big|\, \boldsymbol{x}_i = \boldsymbol{x}\big]. \end{aligned} \tag{G.1}$$

Theorem G.1 (Plug-in approximation)**.** *Suppose that all of the following is true:*

(i) All the conditions of Theorem F.4(a) and Assumption B.7 hold.

(ii) The function $\boldsymbol{x} \mapsto S_{\boldsymbol{q},\tilde{\boldsymbol{q}}}(\boldsymbol{x})$ defined in (G.1) is continuous for all $\boldsymbol{q}, \tilde{\boldsymbol{q}}$, and

$$\sup_{\boldsymbol{x},\boldsymbol{q},\boldsymbol{q}_1\neq\boldsymbol{q}_2} \frac{|S_{\boldsymbol{q},\boldsymbol{q}_1}(\boldsymbol{x}) - S_{\boldsymbol{q},\boldsymbol{q}_2}(\boldsymbol{x})|}{\|\boldsymbol{q}_1 - \boldsymbol{q}_2\|} \lesssim 1. \tag{G.2}$$

(iii) There is an estimator $\widehat{S}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}(\boldsymbol{x})$ which is a known (not depending on $\mathbb{P}_{\mathbf{D}_n}$) measurable function of $(\mathbf{D}_n, \boldsymbol{q}, \tilde{\boldsymbol{q}}, \boldsymbol{x})$, satisfying the bound

$$\sup_{\boldsymbol{q},\tilde{\boldsymbol{q}},\boldsymbol{x}} \big|\widehat{S}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}(\boldsymbol{x}) - S_{\boldsymbol{q},\tilde{\boldsymbol{q}}}(\boldsymbol{x})\big| \lesssim_{\mathbb{P}} r_{\mathtt{S}} \quad \text{with } r_{\mathtt{S}} = o(1). \tag{G.3}$$

Then, on the same probability space, there is a mean-zero Gaussian conditionally on $\mathbf{D}_n$ process $\widehat{\boldsymbol{Z}}^*(\boldsymbol{q})$ in $\mathcal{C}(\mathcal{Q}, \mathbb{R}^K)$, whose distribution is known (does not depend on $\mathbb{P}_{\mathbf{D}_n}$), such that

$$\mathbb{E}\Big[\sup_{\boldsymbol{q}\in\mathcal{Q}} \big\|\boldsymbol{Z}^*(\boldsymbol{q}) - \widehat{\boldsymbol{Z}}^*(\boldsymbol{q})\big\|_\infty \,\Big|\, \mathbf{D}_n\Big] \lesssim_{\mathbb{P}} (r_{\mathtt{UC}} + r_{\mathtt{S}})^{1/(2d_{\mathcal{Q}}+2)} \sqrt{\log n}, \tag{G.4}$$

where $\boldsymbol{Z}^*(\boldsymbol{q})$ is a process in $\mathcal{C}(\mathcal{Q}, \mathbb{R}^K)$ satisfying

$$(\mathbf{X}_n, \boldsymbol{Z}(\cdot)) \overset{d}{=} (\mathbf{X}_n, \boldsymbol{Z}^*(\cdot)), \quad \boldsymbol{Z}^*(\cdot) \underset{\mathbf{X}_n}{\perp\!\!\!\perp} \boldsymbol{y}_n.$$

Further,

$$h^{d/2} \sup_{\boldsymbol{q}\in\mathcal{Q}} \sup_{\boldsymbol{x}\in\mathcal{X}} \left| \frac{\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^\mathsf{T} \widehat{\mathbf{Q}}_{\boldsymbol{q}}^{-1}}{\sqrt{\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})}} \widehat{\boldsymbol{Z}}^*(\boldsymbol{q}) - \frac{\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^\mathsf{T} \bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}}{\sqrt{\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})}} \boldsymbol{Z}^*(\boldsymbol{q}) \right| \lesssim_{\mathbb{P}} r_{\mathtt{PI}}, \tag{G.5}$$

$$\text{where} \quad r_{\mathtt{PI}} := \big[(r_{\mathtt{UC}} + r_{\mathtt{S}})^{1/(2d_{\mathcal{Q}}+2)} + r_{\mathtt{Q}} + r_{\mathtt{VC}}\big] \sqrt{\log n}. \tag{G.6}$$

Equivalently (see Lemma C.3), fixing an arbitrary positive sequence $R_{\mathtt{PI}}$ such that $r_{\mathtt{PI}} = o(R_{\mathtt{PI}})$, we have

$$\mathbb{P}\left\{ h^{d/2} \sup_{\boldsymbol{q}\in\mathcal{Q}} \sup_{\boldsymbol{x}\in\mathcal{X}} \left| \frac{\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^\mathsf{T} \widehat{\mathbf{Q}}_{\boldsymbol{q}}^{-1}}{\sqrt{\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})}} \widehat{\boldsymbol{Z}}^*(\boldsymbol{q}) - \frac{\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^\mathsf{T} \bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}}{\sqrt{\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})}} \boldsymbol{Z}^*(\boldsymbol{q}) \right| > R_{\mathtt{PI}} \,\middle|\, \boldsymbol{D}_n \right\} = o_{\mathbb{P}}(1).$$

Remark G.2. One-dimensional processes $Z^\star(\boldsymbol{x}, \boldsymbol{q})$ and $\widehat{Z}(\boldsymbol{x}, \boldsymbol{q})$ in the main paper are defined as follows:

$$Z^\star(\boldsymbol{x}, \boldsymbol{q}) := \bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q})^\mathsf{T} \boldsymbol{Z}^*(\boldsymbol{q}) = -h^{d/2} \frac{\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^\mathsf{T} \bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}}{\sqrt{\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q})}} \boldsymbol{Z}^*(\boldsymbol{q}),$$

$$\widehat{Z}(\boldsymbol{x}, \boldsymbol{q}) := \hat{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q})^\mathsf{T} \widehat{\boldsymbol{Z}}^*(\boldsymbol{q}) = -h^{d/2} \frac{\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^\mathsf{T} \widehat{\mathbf{Q}}_{\boldsymbol{q}}^{-1}}{\sqrt{\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q})}} \widehat{\boldsymbol{Z}}^*(\boldsymbol{q}).$$

The rest of Appendix G.1 is devoted to proving Theorem G.1.

G.1.1 Strategy

First, $\boldsymbol{Z}^*(\cdot)$ exists by Theorem 8.17 (transfer) in [28] with $(\xi, \eta, \zeta) = (\mathbf{X}_n, \boldsymbol{Z}(\cdot), \boldsymbol{y}_n)$, $\tilde{\xi} = \xi$.

The rest of the proof will be broken up in two steps.

First, we will prove the existence of $\widehat{\boldsymbol{Z}}^*(\boldsymbol{q})$. A natural approach is to approximate the covariance structure $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$ using plug-in, giving an estimated covariance structure $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$, discretize $\mathcal{Q}$ by a suitable δ_n-covering $\mathcal{Q}^{\delta_n}$ and argue that since the discrete versions $\bar{\boldsymbol{\Sigma}}^{\delta_n}$ and $\widehat{\boldsymbol{\Sigma}}^{\delta_n}$ of $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$ are close, the Gaussian vector $\boldsymbol{Z}(\boldsymbol{q}|_{\mathcal{Q}^{\delta_n}}) \in \mathbb{R}^{K|\mathcal{Q}^{\delta_n}|}$ is close to a "feasible" Gaussian vector with covariance $\widehat{\boldsymbol{\Sigma}}^{\delta_n}$. Then we can embed this vector into a conditionally Gaussian process $\widehat{\boldsymbol{Z}}^*(\boldsymbol{q})$. One technical caveat is that a simple plug-in method can lead to $\widehat{\boldsymbol{\Sigma}}^{\delta_n}$ not being positive-semidefinite, or to $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$ not having a suitable Lipschitz property in $\boldsymbol{q}$ and $\tilde{\boldsymbol{q}}$ (necessary for controlling the discretization error between a continuous process and a discrete vector). For this reason, we just project $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$ on the space of such structures that the wanted properties hold (program (G.9)).

The second step is proving (G.4) (given $\widehat{\boldsymbol{Z}}^*(\boldsymbol{q})$ and $\boldsymbol{Z}^*(\boldsymbol{q})$ that are close) by bounding $\sup_{\boldsymbol{q}\in\mathcal{Q}} \boldsymbol{Z}^*(\boldsymbol{q})$ using the maximal inequality in Lemma C.7, and applying Lemma F.7.

G.1.2 Existence of the feasible process

As the first step, we will prove that under Conditions (i) to (iii), the process $\widehat{\boldsymbol{Z}}^*(\boldsymbol{q})$ as described in Theorem G.1 exists.

Estimating the covariance structure By Lemma C.11, for a large enough constant C_{14} the probability of the event

$$\mathcal{A} := \{\lambda_{\max}(\mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}]) \le C_{14} h^d\} \tag{G.7}$$

converges to one. Note that $S_{\boldsymbol{q},\tilde{\boldsymbol{q}}}(\boldsymbol{x})$ is bounded uniformly in $\boldsymbol{x}, \boldsymbol{q}, \tilde{\boldsymbol{q}}$ because it is continuous in these arguments and $\mathcal{Q} \times \mathcal{Q} \times \mathcal{X}$ is compact. Combining with Eq. (G.2), we see that there is a large enough C_{15} such that on $\mathcal{A}$

$$\sup_{\boldsymbol{q},\tilde{\boldsymbol{q}}} \|\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}\| \le C_{15} h^d$$

and

$$\sup_{\boldsymbol{q},\boldsymbol{q}_1 \ne \boldsymbol{q}_2} \frac{|\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q},\boldsymbol{q}_1} - \bar{\boldsymbol{\Sigma}}_{\boldsymbol{q},\boldsymbol{q}_2}|}{\|\boldsymbol{q}_1 - \boldsymbol{q}_2\|} \le C_{15} h^d.$$

We first approximate $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$ by

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}} := \mathbb{E}_n\big[\widehat{S}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}(\boldsymbol{x}_i)\eta^{(1)}(\widehat{\mu}(\boldsymbol{x}_i, \boldsymbol{q}))\eta^{(1)}(\widehat{\mu}(\boldsymbol{x}_i, \tilde{\boldsymbol{q}}))\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\big]. \tag{G.8}$$

The latter is known (a function of $\mathbf{D}_n$ not depending on $\mathbb{P}_{\mathbf{D}_n}$), but may not be positive semidefinite or Lipschitz in the arguments. In order to recover these properties, we solve the following program:

$$
\begin{aligned}
\text{minimize} \quad & \sup_{\boldsymbol{q},\tilde{\boldsymbol{q}}} \|\mathbf{M}_{\boldsymbol{q},\tilde{\boldsymbol{q}}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}\| \\
\text{over} \quad & \mathbf{M}_{\boldsymbol{q},\tilde{\boldsymbol{q}}} \in \mathcal{C}(\mathcal{Q} \times \mathcal{Q}, \mathbb{R}^{K\times K}), \\
& \sup_{\boldsymbol{q},\tilde{\boldsymbol{q}}} \|\mathbf{M}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}\| \le C_{15} h^d, \\
& \mathbf{M}_{\boldsymbol{q},\tilde{\boldsymbol{q}}} = \mathbf{M}_{\tilde{\boldsymbol{q}},\boldsymbol{q}} = \mathbf{M}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}^{\mathsf{T}} = \mathbf{M}_{\tilde{\boldsymbol{q}},\boldsymbol{q}}^{\mathsf{T}}, \quad \boldsymbol{q}, \tilde{\boldsymbol{q}} \in \mathcal{Q}, \\
& \big(\mathbf{M}_{\boldsymbol{q}_k,\boldsymbol{q}_m}\big)_{k,m=1}^{M} \in \mathbb{R}^{MK\times MK} \text{ is symmetric positive semidefinite for any } (\boldsymbol{q}_k)_{k=1}^{M}, \\
& \|\mathbf{M}_{\boldsymbol{q},\boldsymbol{q}_1} - \mathbf{M}_{\boldsymbol{q},\boldsymbol{q}_2}\| \le C_{15} h^d \|\boldsymbol{q}_1 - \boldsymbol{q}_2\|, \quad \boldsymbol{q}, \boldsymbol{q}_1, \boldsymbol{q}_2 \in \mathcal{Q}.
\end{aligned} \tag{G.9}
$$

Lemma G.3 (Solution to the program)**.** *There exists an exact minimizer* $\widehat{\boldsymbol{\Sigma}}^{+}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$ *of* (G.9) *which is a (known) measurable function of* $\mathbf{D}_n$ *with values in* $\mathcal{C}(\mathcal{Q} \times \mathcal{Q}, \mathbb{R}^{K\times K})$.

Proof. The space $\mathcal{C}(\mathcal{Q} \times \mathcal{Q}, \mathbb{R}^{K\times K})$ with the sup-norm

$$
\sup_{\boldsymbol{q},\tilde{\boldsymbol{q}}} \|\mathbf{M}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}\|
$$

is Polish, and the feasible set is non-empty and compact by the multi-dimensional Arzelà–Ascoli theorem. The function

$$
\big(\mathcal{X}^n \times \mathcal{Y}, \mathcal{C}(\mathcal{Q} \times \mathcal{Q}, \mathbb{R}^{K\times K})\big) \ni \big(\mathbf{D}_n, \mathbf{M}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}\big) \mapsto \sup_{\boldsymbol{q},\tilde{\boldsymbol{q}}} \|\mathbf{M}_{\boldsymbol{q},\tilde{\boldsymbol{q}}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}\| \in \mathbb{R}
$$

is a Carathéodory function, that is, measurable in $\mathbf{D}_n$ and continuous in $\mathbf{M}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$. By the Measurable Maximum Theorem 18.19 in [2] (here minimum), the result follows. □

Now we prove that $\widehat{\boldsymbol{\Sigma}}^{+}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$ is sufficiently close to $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$.

Lemma G.4 (Consistency of $\widehat{\boldsymbol{\Sigma}}^{+}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$)**.** *The function* $\widehat{\boldsymbol{\Sigma}}^{+}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$ *defined in Lemma G.3 satisfies*

$$
\sup_{\boldsymbol{q},\tilde{\boldsymbol{q}}} \big\|\widehat{\boldsymbol{\Sigma}}^{+}_{\boldsymbol{q},\tilde{\boldsymbol{q}}} - \bar{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}\big\| \lesssim_{\mathbb{P}} h^d (r_{\mathtt{UC}} + r_{\mathtt{S}}).
$$

Proof. From Eqs. (F.12) and (G.3) and Lemma C.11 we have

$$
\sup_{\boldsymbol{q},\tilde{\boldsymbol{q}}} \big\|\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}\big\| \lesssim_{\mathbb{P}} h^d (r_{\mathtt{UC}} + r_{\mathtt{S}}). \tag{G.10}
$$

But $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$ lies in the feasibility set of (G.9) on $\mathcal{A}$, which implies that

$$
\sup_{\boldsymbol{q},\tilde{\boldsymbol{q}}} \big\|\widehat{\boldsymbol{\Sigma}}^{+}_{\boldsymbol{q},\tilde{\boldsymbol{q}}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}\big\| \le \sup_{\boldsymbol{q},\tilde{\boldsymbol{q}}} \big\|\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}\big\| \quad \text{on } \mathcal{A}.
$$

Since $\mathbb{P}(\mathcal{A}) = 1 - o(1)$, we have

$$
\sup_{\boldsymbol{q},\tilde{\boldsymbol{q}}} \big\|\widehat{\boldsymbol{\Sigma}}^{+}_{\boldsymbol{q},\tilde{\boldsymbol{q}}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}\big\| \lesssim_{\mathbb{P}} h^d (r_{\mathtt{UC}} + r_{\mathtt{S}}),
$$

and by the triangle inequality the result follows. □

Discretization and controlling the deviations of the infeasible process Having estimated the covariance structure, we proceed with the same steps as in Lemma F.3. First, we discretize $\mathcal{Q}$. Let $\mathcal{Q}^{\delta_n} := \{\boldsymbol{q}_1, \dots, \boldsymbol{q}_{|\mathcal{Q}^{\delta_n}|}\}$ be an internal δ_n-covering[2] of $\mathcal{Q}$ with respect to the 2-norm $\|\cdot\|$ of cardinality $|\mathcal{Q}^{\delta_n}| \lesssim 1/\delta_n^{d_{\mathcal{Q}}}$, where δ_n is chosen later. As already proven in Lemma F.3, there is a bound on the deviations of $\boldsymbol{Z}^*(\cdot)$:

$$\mathbb{E}\Big[\sup_{\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|\le\delta_n}\|\boldsymbol{Z}^*(\boldsymbol{q}) - \boldsymbol{Z}^*(\tilde{\boldsymbol{q}})\|_\infty \,\Big|\, \mathbf{X}_n\Big] = \mathbb{E}\Big[\sup_{\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|\le\delta_n}\|\boldsymbol{Z}^*(\boldsymbol{q}) - \boldsymbol{Z}^*(\tilde{\boldsymbol{q}})\|_\infty \,\Big|\, \mathbf{D}_n\Big] \lesssim \sqrt{\delta_n \log\frac{K}{\delta_n}} \quad \text{on } \mathcal{A},$$

where in the equality we used that $\boldsymbol{Z}^*(\cdot)$ is independent of $\boldsymbol{y}_n$ conditionally on $\mathbf{X}_n$.

Closeness of the discrete vectors Consider now the conditionally Gaussian vector $\boldsymbol{Z}^*(\boldsymbol{q}|_{\mathcal{Q}^{\delta_n}}) \in \mathbb{R}^{K|\mathcal{Q}^{\delta_n}|}$ and denote

$$\bar{\boldsymbol{\Sigma}}^{\delta_n} := h^d \mathbb{V}[\boldsymbol{Z}^*(\boldsymbol{q}|_{\mathcal{Q}^{\delta_n}}) \,|\, \mathbf{X}_n] = (\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q}_k,\boldsymbol{q}_m})_{k,m=1}^{|\mathcal{Q}^{\delta_n}|} \in \mathbb{R}^{K|\mathcal{Q}^{\delta_n}|\times K|\mathcal{Q}^{\delta_n}|},$$

so that

$$\boldsymbol{Z}^*(\boldsymbol{q}|_{\mathcal{Q}^{\delta_n}}) = h^{-d/2}(\bar{\boldsymbol{\Sigma}}^{\delta_n})^{1/2}\boldsymbol{\xi}_n, \quad \text{where } \mathbb{P}_{\boldsymbol{\xi}_n|\mathbf{X}_n} = \mathcal{N}\big(0, \mathbf{I}_{\mathbb{R}^{K|\mathcal{Q}^{\delta_n}|}}\big),$$

and $\boldsymbol{\xi}_n$ is independent of $\boldsymbol{y}_n$ given $\mathbf{X}_n$. (Since $\mathbb{P}_{\boldsymbol{\xi}_n|\mathbf{X}_n}$ does not depend on $\mathbf{X}_n$, this vector is in fact independent of $\mathbf{D}_n$.) Discretizing $\widehat{\boldsymbol{\Sigma}}^+_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$ in the same way, we can put

$$\widehat{\boldsymbol{\Sigma}}^{\delta_n} := \big(\widehat{\boldsymbol{\Sigma}}^+_{\boldsymbol{q}_k,\boldsymbol{q}_m}\big)_{k,m=1}^{|\mathcal{Q}^{\delta_n}|} \in \mathbb{R}^{K|\mathcal{Q}^{\delta_n}|\times K|\mathcal{Q}^{\delta_n}|}.$$

Since the functions $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$ and $\widehat{\boldsymbol{\Sigma}}^+_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$ are close, the two matrices are close as well, which we will make precise in the following lemma.

Lemma G.5 (The matrices $\bar{\boldsymbol{\Sigma}}^{\delta_n}$ and $\widehat{\boldsymbol{\Sigma}}^{\delta_n}$ are close)**.** *We have*

$$\big\|\bar{\boldsymbol{\Sigma}}^{\delta_n} - \widehat{\boldsymbol{\Sigma}}^{\delta_n}\big\| \lesssim_{\mathbb{P}} |\mathcal{Q}^{\delta_n}| h^d (r_{\mathtt{UC}} + r_{\mathtt{S}}).$$

Proof. We can bound the Frobenius norm using the the bound on each element of the matrix:

$$\big\|\bar{\boldsymbol{\Sigma}}^{\delta_n} - \widehat{\boldsymbol{\Sigma}}^{\delta_n}\big\| \le \big\|\bar{\boldsymbol{\Sigma}}^{\delta_n} - \widehat{\boldsymbol{\Sigma}}^{\delta_n}\big\|_F \le |\mathcal{Q}^{\delta_n}| \sup_{\boldsymbol{q},\tilde{\boldsymbol{q}}}\big\|\widehat{\boldsymbol{\Sigma}}^+_{\boldsymbol{q},\tilde{\boldsymbol{q}}} - \bar{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}\big\|.$$

It is left to combine with Lemma G.4. $\square$

Applying Lemma C.5, we then have

$$\begin{aligned} h^{-d/2}\mathbb{E}\big[\big\|(\bar{\boldsymbol{\Sigma}}^{\delta_n})^{1/2}\boldsymbol{\xi}_n - (\widehat{\boldsymbol{\Sigma}}^{\delta_n})^{1/2}\boldsymbol{\xi}_n\big\|_\infty \,\big|\, \mathbf{D}_n\big] &\le 2h^{-d/2}\sqrt{\log(K|\mathcal{Q}^{\delta_n}|)}\|\bar{\boldsymbol{\Sigma}}^{\delta_n} - \widehat{\boldsymbol{\Sigma}}^{\delta_n}\|^{1/2} \\ &\lesssim_{\mathbb{P}} |\mathcal{Q}^{\delta_n}|^{1/2}(r_{\mathtt{UC}} + r_{\mathtt{S}})^{1/2}\sqrt{\log(K|\mathcal{Q}^{\delta_n}|)}. \end{aligned} \tag{G.11}$$

[2] This means $\mathcal{Q}_n^\delta \subset \mathcal{Q}$ and balls with radius δ_n centered at these points cover $\mathcal{Q}$.

Embedding the conditionally Gaussian vector $(\widehat{\mathbf{\Sigma}}^{\delta_n})^{1/2}\boldsymbol{\xi}_n$ into a conditionally Gaussian process By the same argument as in Lemma F.3, there exists a law on $(\mathcal{X}\times\mathcal{Y})^n\times\mathcal{C}(\mathcal{Q},\mathbb{R}^K)$ which is the joint law of $\mathbf{D}'_n \overset{d}{=} \mathbf{D}_n$ and a conditionally on $\mathbf{D}'_n$ mean-zero Gaussian process $\widehat{\boldsymbol{Z}}_n^{*\prime}(\boldsymbol{q})$ with the conditional covariance structure

$$\mathbb{E}[\widehat{\boldsymbol{Z}}_n^{*\prime}(\boldsymbol{q})\widehat{\boldsymbol{Z}}_n^{*\prime}(\tilde{\boldsymbol{q}})^{\mathsf{T}}\,|\,\mathbf{D}'_n] = h^{-d}\widehat{\mathbf{\Sigma}}'^{+}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$$

(where $\widehat{\mathbf{\Sigma}}'^{+}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$ is the same function of the data as $\widehat{\mathbf{\Sigma}}^{+}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$ with $\mathbf{D}_n$ replaced by $\mathbf{D}'_n$). Continuity follows by the Kolmogorov-Chentsov theorem:

$$\mathbb{E}\big[\big\|\widehat{\boldsymbol{Z}}_n^{*\prime}(\boldsymbol{q}) - \widehat{\boldsymbol{Z}}_n^{*\prime}(\tilde{\boldsymbol{q}})\big\|^a\,\big|\,\mathbf{D}'_n\big] \le C_n\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|^{d_{\mathcal{Q}}+b}, \tag{G.12}$$

for $a,b>0$ chosen as follows. The vector $\widehat{\boldsymbol{Z}}_n^{*\prime}(\boldsymbol{q}) - \widehat{\boldsymbol{Z}}_n^{*\prime}(\tilde{\boldsymbol{q}})$ is conditionally on $\mathbf{D}'_n$ Gaussian with covariance

$$h^{-d}\big(\widehat{\mathbf{\Sigma}}'^{+}_{\boldsymbol{q},\boldsymbol{q}} - 2\widehat{\mathbf{\Sigma}}'^{+}_{\boldsymbol{q},\tilde{\boldsymbol{q}}} + \widehat{\mathbf{\Sigma}}'^{+}_{\tilde{\boldsymbol{q}},\tilde{\boldsymbol{q}}}\big).$$

Hence, we have for any $m>0$

$$\begin{aligned}
&\mathbb{E}\big[\big\|\widehat{\boldsymbol{Z}}_n^{*\prime}(\boldsymbol{q}) - \widehat{\boldsymbol{Z}}_n^{*\prime}(\tilde{\boldsymbol{q}})\big\|^{2m}\,\big|\,\mathbf{D}'_n\big]\\
&\quad\le h^{-dm}\big\|\widehat{\mathbf{\Sigma}}'^{+}_{\boldsymbol{q},\boldsymbol{q}} - 2\widehat{\mathbf{\Sigma}}'^{+}_{\boldsymbol{q},\tilde{\boldsymbol{q}}} + \widehat{\mathbf{\Sigma}}'^{+}_{\tilde{\boldsymbol{q}},\tilde{\boldsymbol{q}}}\big\|^m\mathbb{E}[\|\boldsymbol{\xi}_K\|^{2m}]\\
&\quad\le (2C_{15})^m\mathbb{E}[\|\boldsymbol{\xi}_K\|^{2m}]\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|^m,
\end{aligned}$$

where $\boldsymbol{\xi}_K\sim\mathcal{N}(0,\mathbf{I}_K)$. So we can take, for example, $m=d_{\mathcal{Q}}+1$, $a=2m$, $b=1$ in (G.12).

Projecting this $\mathcal{C}(\mathcal{Q},\mathbb{R}^K)$-process onto $\mathcal{Q}^{\delta_n}$, we obtain a vector $\widehat{\boldsymbol{Z}}_n^{*\prime}\big(\boldsymbol{q}|_{\mathcal{Q}^{\delta_n}}\big)\in\mathbb{R}^{K|\mathcal{Q}^{\delta_n}|}$ such that

$$\big(\mathbf{D}'_n, \widehat{\boldsymbol{Z}}_n^{*\prime}\big(\boldsymbol{q}|_{\mathcal{Q}^{\delta_n}}\big)\big) \overset{d}{=} \big(\mathbf{D}_n, h^{-d/2}(\widehat{\mathbf{\Sigma}}^{\delta_n})^{1/2}\boldsymbol{\xi}_n\big).$$

Since $\mathcal{C}(\mathcal{Q},\mathbb{R}^K)$ is Polish, by Theorem 8.17 (transfer) in [28] on our probability space there exists a random element $\widehat{\boldsymbol{Z}}^*(\boldsymbol{q})\in\mathcal{C}(\mathcal{Q},\mathbb{R}^K)$ such that

$$\big(\mathbf{D}'_n, \widehat{\boldsymbol{Z}}_n^{*\prime}\big(\boldsymbol{q}|_{\mathcal{Q}^{\delta_n}}\big), \widehat{\boldsymbol{Z}}_n^{*\prime}(\boldsymbol{q})\big) \overset{d}{=} \big(\mathbf{D}_n, h^{-d/2}(\widehat{\mathbf{\Sigma}}^{\delta_n})^{1/2}\boldsymbol{\xi}_n, \widehat{\boldsymbol{Z}}^*(\boldsymbol{q})\big).$$

In particular, almost surely $h^{-d/2}(\widehat{\mathbf{\Sigma}}^{\delta_n})^{1/2}\boldsymbol{\xi}_n$ is the projection of $\widehat{\boldsymbol{Z}}^*$ on $\mathcal{Q}^{\delta_n}$.

Controlling the deviations of $\widehat{\boldsymbol{Z}}^*(\boldsymbol{q})$ Consider the stochastic process X_n with index set

$$T := \{(l,\boldsymbol{q},\tilde{\boldsymbol{q}})\colon l\in\{1,2,\dots,K\}, \boldsymbol{q},\tilde{\boldsymbol{q}}\in\mathcal{Q}, \|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|\le\delta_n\}$$

whose value at $t=(l,\boldsymbol{q},\tilde{\boldsymbol{q}})$ is $X_{n,t} := \widehat{Z}_l^*(\boldsymbol{q}) - \widehat{Z}_l^*(\tilde{\boldsymbol{q}})$. It is a separable (because each $\widehat{Z}_l^*(\cdot)$ has continuous trajectories) mean-zero Gaussian conditionally on $\mathbf{D}_n$ process with the index set T considered a metric space: $\operatorname{dist}\big((l,\boldsymbol{q},\tilde{\boldsymbol{q}}),(l',\boldsymbol{q}',\tilde{\boldsymbol{q}}')\big) = \|\boldsymbol{q}-\boldsymbol{q}'\| + \|\tilde{\boldsymbol{q}}-\tilde{\boldsymbol{q}}'\| + \mathbb{1}\{l\ne l'\}$.

The quantity $\sigma(X_n)^2$ is defined and bounded as follows:

$$\sigma(X_n)^2 := \sup_{t\in T}\mathbb{E}[X_{n,t}^2\,|\,\mathbf{D}_n] = \sup_{\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|\le\delta_n}\max_l\mathbb{E}\big[\big(\widehat{Z}_l^*(\boldsymbol{q}) - \widehat{Z}_l^*(\tilde{\boldsymbol{q}})\big)^2\,\big|\,\mathbf{D}_n\big] \lesssim \delta_n,$$

because $\widehat{Z}_l^*(\boldsymbol{q}) - \widehat{Z}_l^*(\tilde{\boldsymbol{q}})$ is conditionally on $\mathbf{D}_n$ mean-zero normal variable with variance

$$\begin{aligned}
&h^{-d}\big(\big(\widehat{\Sigma}^{+}_{\boldsymbol{q},\boldsymbol{q}}\big)_{ll} - 2\big(\widehat{\Sigma}^{+}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}\big)_{ll} + \big(\widehat{\Sigma}^{+}_{\tilde{\boldsymbol{q}},\tilde{\boldsymbol{q}}}\big)_{ll}\big)\\
&\quad\le h^{-d}\big|\big(\widehat{\Sigma}^{+}_{\boldsymbol{q},\boldsymbol{q}}\big)_{ll} - \big(\widehat{\Sigma}^{+}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}\big)_{ll}\big| + h^{-d}\big|\big(\widehat{\Sigma}^{+}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}\big)_{ll} - \big(\widehat{\Sigma}^{+}_{\tilde{\boldsymbol{q}},\tilde{\boldsymbol{q}}}\big)_{ll}\big|\\
&\quad\le h^{-d}\big\|\widehat{\mathbf{\Sigma}}^{+}_{\boldsymbol{q},\boldsymbol{q}} - \widehat{\mathbf{\Sigma}}^{+}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}\big\| + h^{-d}\big\|\widehat{\mathbf{\Sigma}}^{+}_{\boldsymbol{q},\tilde{\boldsymbol{q}}} - \widehat{\mathbf{\Sigma}}^{+}_{\tilde{\boldsymbol{q}},\tilde{\boldsymbol{q}}}\big\|\\
&\quad\le 2C_{15}\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|.
\end{aligned}$$

Next, we define and bound the semimetric $\rho(t,t')$:

$$\begin{aligned}\rho(t,t')^2 &:= \mathbb{E}[(X_{n,t}-X_{n,t'})^2 \,|\, \mathbf{D}_n] \\ &= \mathbb{E}\big[\big(\big(\widehat{Z}_l^*(\boldsymbol{q})-\widehat{Z}_l^*(\tilde{\boldsymbol{q}})\big)-\big(\widehat{Z}_{l'}^*(\boldsymbol{q}')-\widehat{Z}_{l'}^*(\tilde{\boldsymbol{q}}')\big)\big)^2 \,\big|\, \mathbf{D}_n\big] \\ &\lesssim \mathbb{E}\big[\big(\widehat{Z}_l^*(\boldsymbol{q})-\widehat{Z}_{l'}^*(\boldsymbol{q}')\big)^2 \,\big|\, \mathbf{D}_n\big]-\mathbb{E}\big[\big(\widehat{Z}_l^*(\tilde{\boldsymbol{q}})-\widehat{Z}_{l'}^*(\tilde{\boldsymbol{q}}')\big)^2 \,\big|\, \mathbf{D}_n\big].\end{aligned}$$

The first term on the right is bounded the following way: if $l \neq l'$,

$$\begin{aligned}&\mathbb{E}\big[\big(\widehat{Z}_l^*(\boldsymbol{q})-\widehat{Z}_{l'}^*(\boldsymbol{q}')\big)^2 \,\big|\, \mathbf{D}_n\big] \\ &\quad\lesssim \mathbb{E}\big[\big(\widehat{Z}_l^*(\boldsymbol{q})-\widehat{Z}_l^*(\boldsymbol{q}')\big)^2 \,\big|\, \mathbf{D}_n\big]+\mathbb{E}\big[\widehat{Z}_l^*(\boldsymbol{q}')^2 \,\big|\, \mathbf{D}_n\big]+\mathbb{E}\big[\widehat{Z}_{l'}^*(\boldsymbol{q}')^2 \,\big|\, \mathbf{D}_n\big] \\ &\quad\lesssim \|\boldsymbol{q}-\boldsymbol{q}'\|+1,\end{aligned}$$

and if $l = l'$,

$$\mathbb{E}\big[\big(\widehat{Z}_l^*(\boldsymbol{q})-\widehat{Z}_{l'}^*(\boldsymbol{q}')\big)^2 \,\big|\, \mathbf{D}_n\big] \lesssim \|\boldsymbol{q}-\boldsymbol{q}'\|.$$

The term $\mathbb{E}\big[\big(\widehat{Z}_l^*(\tilde{\boldsymbol{q}})-\widehat{Z}_{l'}^*(\tilde{\boldsymbol{q}}')\big)^2 \,\big|\, \mathbf{D}_n\big]$ is bounded the same way, and we conclude that

$$\rho(t,t')^2 \leq C_{16}\big(\|\boldsymbol{q}-\boldsymbol{q}'\|+\|\tilde{\boldsymbol{q}}-\tilde{\boldsymbol{q}}'\|+\mathbb{1}\{l \neq l'\}\big) = C_{16}\mathrm{dist}\big((l,\boldsymbol{q},\tilde{\boldsymbol{q}}),(l',\boldsymbol{q}',\tilde{\boldsymbol{q}}')\big).$$

for some positive constant C_{16}. This means that, for any $\varepsilon > 0$, an $C_{16}^{-1}\varepsilon^2$-covering of T with respect to $\mathrm{dist}(\cdot)$ induces an ε-covering of T with respect to ρ, and hence

$$N(T,\rho,\varepsilon) \leq N\big(T,\mathrm{dist}(\cdot),C_{16}^{-1}\varepsilon^2\big).$$

But the right-hand side is clearly $O(K\varepsilon^{-2})$ because T is a subset of $\{1,\dots,K\}\times\mathcal{Q}\times\mathcal{Q}$, so for $\varepsilon \in (0,1)$

$$\log N(T,\rho,\varepsilon) \lesssim \log\frac{K}{\varepsilon}.$$

Applying Lemma C.7 gives

$$\begin{aligned}&\mathbb{E}\Big[\sup_{t\in T}|X_{n,t}| \,\Big|\, \mathbf{D}_n\Big] = \mathbb{E}\Big[\sup_{\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|\leq\delta_n}\big\|\widehat{\boldsymbol{Z}}^*(\boldsymbol{q})-\widehat{\boldsymbol{Z}}^*(\tilde{\boldsymbol{q}})\big\|_\infty \,\Big|\, \mathbf{D}_n\Big] \lesssim \sigma(X_n)+\int_0^{2\sigma(X_n)}\sqrt{\log N(T,\rho,\varepsilon)}\,\mathrm{d}\varepsilon \\ &\lesssim \sigma(X_n)+\int_0^{2\sigma(X_n)}\sqrt{\log(K/\varepsilon)}\,\mathrm{d}\varepsilon \lesssim \sigma(X_n)+\int_0^{2\sigma(X_n)}\big(\sqrt{\log K}+\sqrt{\log(\varepsilon^{-1})}\big)\,\mathrm{d}\varepsilon \\ &\lesssim \sigma(X_n)\sqrt{\log K}+\frac{1}{\sqrt{\log(\sigma(X_n)^{-1})}}\int_0^{2\sigma(X_n)}\log(\varepsilon^{-1})\,\mathrm{d}\varepsilon \\ &\lesssim \sigma(X_n)\sqrt{\log K}+\frac{1+\log(\sigma(X_n)^{-1})}{\sqrt{\log(\sigma(X_n)^{-1})}}\sigma(X_n) \lesssim \sigma(X_n)\sqrt{\log(K\sigma(X_n)^{-1})} \lesssim \sqrt{\delta_n\log\frac{K}{\delta_n}},\end{aligned}$$

where in the last inequality we used the bound $\sigma(X_n) \lesssim \sqrt{\delta_n}$ from above and that $x \mapsto x\log(1/x)$ is increasing for sufficiently small x.

Choosing δ_n and conclusion Combining the bounds obtained above, we can write

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{\boldsymbol{q}\in\mathcal{Q}}\big\|\boldsymbol{Z}^*(\boldsymbol{q})-\widehat{\boldsymbol{Z}}^*(\boldsymbol{q})\big\|_\infty \,\Big|\, \mathbf{D}_n\Big]\\
&\quad\le \mathbb{E}\Big[\sup_{\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|\le\delta_n}\big\|\boldsymbol{Z}^*(\boldsymbol{q})-\boldsymbol{Z}^*(\tilde{\boldsymbol{q}})\big\|_\infty \,\Big|\, \mathbf{D}_n\Big] + h^{-d/2}\mathbb{E}\big[\big\|(\bar{\boldsymbol{\Sigma}}^{\delta_n})^{1/2}\boldsymbol{\xi}_n-(\widehat{\boldsymbol{\Sigma}}^{\delta_n})^{1/2}\boldsymbol{\xi}_n\big\|_\infty \,\big|\, \mathbf{D}_n\big]\\
&\qquad + \mathbb{E}\Big[\sup_{\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|\le\delta_n}\big\|\widehat{\boldsymbol{Z}}^*(\boldsymbol{q})-\widehat{\boldsymbol{Z}}^*(\tilde{\boldsymbol{q}})\big\|_\infty \,\Big|\, \mathbf{D}_n\Big]\\
&\quad\lesssim_{\mathbb{P}} \left(\sqrt{\delta_n}+\frac{(r_{\mathtt{UC}}+r_{\mathtt{S}})^{1/2}}{\delta_n^{d_{\mathcal{Q}}/2}}\right)\sqrt{\log\frac{K}{\delta_n}}.
\end{aligned}$$

Choose the approximately optimal

$$\delta_n := (r_{\mathtt{UC}}+r_{\mathtt{S}})^{1/(d_{\mathcal{Q}}+1)},$$

giving

$$\mathbb{E}\Big[\sup_{\boldsymbol{q}\in\mathcal{Q}}\big\|\boldsymbol{Z}^*(\boldsymbol{q})-\widehat{\boldsymbol{Z}}^*(\boldsymbol{q})\big\|_\infty \,\Big|\, \mathbf{D}_n\Big] \lesssim_{\mathbb{P}} (r_{\mathtt{UC}}+r_{\mathtt{S}})^{1/(2d_{\mathcal{Q}}+2)}\sqrt{\log n}.$$

G.1.3 Proof of (G.5)

Having proven the existence of $\widehat{\boldsymbol{Z}}^*(\cdot)$, we will now prove that under Conditions (i) to (iii), Eq. (G.5) holds. We build on the proof of Theorem 6.3 in [11].

Lemma G.6 (Bounding the supremum of the Gaussian process)**.** *In the setting of Theorem G.1, we have* $\sup_{\boldsymbol{q}\in\mathcal{Q}}\|\boldsymbol{Z}(\boldsymbol{q})\|_\infty \lesssim_{\mathbb{P}} \sqrt{\log K}$.

Proof. Consider the stochastic process X_n defined for $t=(l,\boldsymbol{q})\in T$ with

$$T := \{(l,\boldsymbol{q}) : l\in\{1,2,\ldots,K\}, \boldsymbol{q}\in\mathcal{Q}\}$$

as $X_{n,t} := Z_{n,l}(\boldsymbol{q})$. It is a separable mean-zero Gaussian process with the index set T considered a metric space: $\operatorname{dist}\big((l,\boldsymbol{q}),(l',\boldsymbol{q}')\big) = \|\boldsymbol{q}-\boldsymbol{q}'\| + \mathbb{1}\{l\neq l'\}$. Note that

$$\begin{aligned}
\sigma(X_n)^2 &:= \sup_{t\in T}\mathbb{E}\big[X_{n,t}^2 \,\big|\, \mathbf{X}_n\big] = \sup_{\boldsymbol{q}\in\mathcal{Q}}\max_l \mathbb{E}\big[Z_{n,l}^2(\boldsymbol{q}) \,\big|\, \mathbf{X}_n\big]\\
&= h^{-d}\sup_{\boldsymbol{q}\in\mathcal{Q}}\max_l \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}));\boldsymbol{q})^2 \,\big|\, \boldsymbol{x}_i\big]\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))^2 p_l(\boldsymbol{x}_i)^2\\
&\lesssim h^{-d}\mathbb{E}_n[p_l(\boldsymbol{x}_i)^2] \le h^{-d}\sup_{\boldsymbol{\alpha}\in\mathcal{S}^{K-1}}\mathbb{E}_n\big[\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\boldsymbol{\alpha}\big] \lesssim 1 \quad \text{w. p. a. 1.}
\end{aligned}$$

Next, we will bound

$$\rho(t,t')^2 := \mathbb{E}\big[(X_{n,t}-X_{n,t'})^2 \,\big|\, \mathbf{X}_n\big] = \mathbb{E}\big[(Z_{n,l}(\boldsymbol{q})-Z_{n,l'}(\boldsymbol{q}'))^2 \,\big|\, \mathbf{X}_n\big].$$

If $l \neq l'$,

$$
\begin{aligned}
&\mathbb{E}\big[(Z_{n,l}(\boldsymbol{q}) - Z_{n,l'}(\boldsymbol{q}'))^2 \,\big|\, \mathbf{X}_n\big] \\
&\quad = \frac{1}{n}\sum_{i=1}^{n} h^{-d}\mathbb{E}\Big[\Big(\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q})); \boldsymbol{q})\eta^{(1)}(\mu_0(\boldsymbol{x}_i, \boldsymbol{q}))p_l(\boldsymbol{x}_i) \\
&\qquad - \psi(y_i, \eta\big(\mu_0(\boldsymbol{x}_i, \boldsymbol{q}')\big); \boldsymbol{q}')\eta^{(1)}\big(\mu_0(\boldsymbol{x}_i, \boldsymbol{q}')\big)p_{l'}(\boldsymbol{x}_i)\Big)^2 \mathbf{X}_n\Big] \\
&\quad = h^{-d}\mathbb{E}_n\big[\mathbb{E}\big[(A_n(\boldsymbol{q}, y_i, \boldsymbol{x}_i) - A_n(\boldsymbol{q}', y_i, \boldsymbol{x}_i))^2 \,\big|\, \mathbf{X}_n\big] p_l(\boldsymbol{x}_i)^2\big] \\
&\qquad + h^{-d}\mathbb{E}_n\big[\mathbb{E}\big[A_n(\boldsymbol{q}', y_i, \boldsymbol{x}_i)^2 \,\big|\, \mathbf{X}_n\big]\big(p_l(\boldsymbol{x}_i)^2 + p_{l'}(\boldsymbol{x}_i)^2\big)\big] \\
&\quad \lesssim h^{-d}\|\boldsymbol{q} - \boldsymbol{q}'\|\mathbb{E}_n\big[p_l(\boldsymbol{x}_i)^2\big] + h^{-d}\mathbb{E}_n\big[p_l(\boldsymbol{x}_i)^2 + p_{l'}(\boldsymbol{x}_i)^2\big] \\
&\quad \lesssim \|\boldsymbol{q} - \boldsymbol{q}'\| + 1 \quad \text{w. p. a. 1,}
\end{aligned}
$$

where we denoted $A_n(\boldsymbol{q}, y_i, \boldsymbol{x}_i) := \psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, \boldsymbol{q})); \boldsymbol{q})\eta^{(1)}(\mu_0(\boldsymbol{x}_i, \boldsymbol{q}))$ to simplify notation. Similarly, if $l = l'$,

$$
\mathbb{E}\big[(Z_{n,l}(\boldsymbol{q}) - Z_{n,l'}(\boldsymbol{q}'))^2 \,\big|\, \mathbf{X}_n\big] \lesssim \|\boldsymbol{q} - \boldsymbol{q}'\| \quad \text{w. p. a. 1.}
$$

We conclude

$$
\rho(t, t')^2 \lesssim \operatorname{dist}\big((l, \boldsymbol{q}), (l', \boldsymbol{q}')\big).
$$

This means that an ε^2-covering of T with respect to $\operatorname{dist}(\cdot)$ induces an ε-covering of T with respect to ρ, and hence

$$
N(T, \rho, \varepsilon) \leq N(T, \operatorname{dist}(\cdot), \varepsilon^2). \tag{G.13}
$$

On the other hand, since $\mathcal{Q}$ does not depend on n, clearly for (sufficiently small) $\tilde{\varepsilon} > 0$, $N(T, \operatorname{dist}(\cdot), \tilde{\varepsilon}) \lesssim K(C_1/\tilde{\varepsilon})^{C_2}$, where C_1 and C_2 are both constants. Combining this with (G.13) we get

$$
\log N(T, \rho, \varepsilon) \lesssim \log(K/\varepsilon).
$$

Now we apply Lemma C.7:

$$
\begin{aligned}
\mathbb{E}\Big[\sup_{t \in T}|X_{n,t}| \,\Big|\, \mathbf{X}_n\Big] &\lesssim \sigma(X_n) + \int_0^{2\sigma(X_n)} \sqrt{\log N(T, \rho, \varepsilon)}\,\mathrm{d}\varepsilon \\
&\lesssim \sigma(X_n) + \sigma(X_n)\sqrt{\log(K/\sigma(X_n))} \lesssim_1 \sqrt{\log K} \quad \text{w. p. a. 1,}
\end{aligned}
$$

where in $\lesssim_1$ we used our bound $\sigma(X_n) \lesssim 1$ above. Rewriting, we obtain

$$
\mathbb{E}\Big[\sup_{\boldsymbol{q} \in \mathcal{Q}}\|\boldsymbol{Z}(\boldsymbol{q})\|_\infty \,\Big|\, \mathbf{X}_n\Big] \lesssim \sqrt{\log K} \quad \text{w. p. a. 1.} \tag{G.14}
$$

By Markov's inequality this gives $\sup_{\boldsymbol{q} \in \mathcal{Q}}\|\boldsymbol{Z}(\boldsymbol{q})\|_\infty \lesssim_{\mathbb{P}} \sqrt{\log K}$. □

By the triangle inequality, Eq. (G.5) will follow from the two bounds

$$
h^{d/2}\sup_{\boldsymbol{q}, \boldsymbol{x}}\left|\frac{\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^\mathsf{T}\widehat{\mathbf{Q}}_{\boldsymbol{q}}^{-1}}{\sqrt{\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q})}}\widehat{\boldsymbol{Z}}^*(\boldsymbol{q}) - \frac{\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^\mathsf{T}\widehat{\mathbf{Q}}_{\boldsymbol{q}}^{-1}}{\sqrt{\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q})}}\boldsymbol{Z}^*(\boldsymbol{q})\right| \lesssim_{\mathbb{P}} (r_{\mathtt{UC}} + r_{\mathtt{S}})^{1/(2d_{\mathcal{Q}}+2)}\sqrt{\log n}, \tag{G.15}
$$

$$
h^{d/2}\sup_{\boldsymbol{q}, \boldsymbol{x}}\left|\frac{\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^\mathsf{T}\widehat{\mathbf{Q}}_{\boldsymbol{q}}^{-1}}{\sqrt{\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q})}}\boldsymbol{Z}^*(\boldsymbol{q}) - \frac{\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^\mathsf{T}\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}}{\sqrt{\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{q})}}\boldsymbol{Z}^*(\boldsymbol{q})\right| \lesssim_{\mathbb{P}} (r_{\mathtt{Q}} + r_{\mathtt{VC}})\sqrt{\log n}. \tag{G.16}
$$

To prove (G.15), combine $\sup_{\boldsymbol{q},\boldsymbol{x}}\big\|\hat{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\big\|_1 \lesssim_{\mathbb{P}} 1$ by Lemma F.7 with (G.4).

To prove (G.16), combine Lemma G.6 with $\sup_{\boldsymbol{q},\boldsymbol{x}}\big\|\hat{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) - \bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\big\|_1 \lesssim_{\mathbb{P}} r_{\mathtt{Q}} + r_{\mathtt{VC}}$ by Lemma F.7.

This completes the proof of Theorem G.1.

G.2 Confidence bands

As a corollary of Theorem G.1, we can obtain the following result, whose proof is based on [3].

Theorem G.7 (Confidence bands)**.** *Suppose all conditions of Theorem G.1 are true, and* $(R_{\mathtt{PI}} + r_{\mathtt{SA}} + r_{\mathtt{HO}})\sqrt{\log n} = o(1)$*. Then* $\mathbb{P}\big\{\sup_{\boldsymbol{q},\boldsymbol{x}} |T_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})| < k^*(1-\alpha)\big\} = 1-\alpha+o(1)$*, where* $k^*(\eta)$ *is the conditional* η*-quantile of* $\sup_{\boldsymbol{q},\boldsymbol{x}}\big|\hat{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widehat{\boldsymbol{Z}}^*(\boldsymbol{q})\big|$ *given the data. Equivalently,* $\mathbb{P}\big\{\mu_0^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q}) \in \mathsf{CB}_{1-\alpha}(\boldsymbol{x},\boldsymbol{q},\boldsymbol{v})\big\} = 1-\alpha+o(1)$ *with*

$$\mathsf{CB}_{1-\alpha}(\boldsymbol{x},\boldsymbol{q},\boldsymbol{v}) := \Big(\widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q}) - k^*(1-\alpha)\sqrt{\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})/n}, \widehat{\mu}^{(\boldsymbol{v})}(\boldsymbol{x},\boldsymbol{q}) + k^*(1-\alpha)\sqrt{\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})/n}\Big). \quad \text{(G.17)}$$

To prove this result we first introduce some auxiliary lemmas. To simplify the exposition, let

$$V := \sup_{\boldsymbol{q},\boldsymbol{x}} |T_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})|, \quad \text{(G.18)}$$

$$V^* := \sup_{\boldsymbol{q},\boldsymbol{x}}\left|h^{d/2}\frac{\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}}\widehat{\mathbf{Q}}_{\boldsymbol{q}}^{-1}}{\sqrt{\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})}}\widehat{\boldsymbol{Z}}^*(\boldsymbol{q})\right|, \quad \text{(G.19)}$$

$$\widetilde{V} := \sup_{\boldsymbol{q},\boldsymbol{x}}\left|h^{d/2}\frac{\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}}\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}}{\sqrt{\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})}}\boldsymbol{Z}(\boldsymbol{q})\right|, \quad \text{(G.20)}$$

$$\widetilde{V}^* := \sup_{\boldsymbol{q},\boldsymbol{x}}\left|h^{d/2}\frac{\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}}\bar{\mathbf{Q}}_{\boldsymbol{q}}^{-1}}{\sqrt{\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})}}\boldsymbol{Z}^*(\boldsymbol{q})\right|. \quad \text{(G.21)}$$

Let $\tilde{k}^*(\eta)$ denote the conditional η-quantile of $\widetilde{V}^*$ given $\mathbf{X}_n$.

Lemma G.8 (Closeness rates)**.** *Random variables* V, $\widetilde{V}$, V^*, $\widetilde{V}^*$ *satisfy the following:*

(a) $|V - \widetilde{V}| = o_{\mathbb{P}}(r_{\mathtt{SA}} + r_{\mathtt{HO}})$*;*

(b) $|V^* - \widetilde{V}^*| = o_{\mathbb{P}}(R_{\mathtt{PI}})$*, where* $R_{\mathtt{PI}}$ *is defined in Theorem G.1;*

(c) $\widetilde{V}^* \underset{\mathbf{X}_n}{\perp\!\!\!\perp} \boldsymbol{y}_n$.

Proof. By Theorem F.4, we have

$$\left|V - \sup_{\boldsymbol{q},\boldsymbol{x}}\left|\bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\boldsymbol{Z}(\boldsymbol{q})\right|\right| = o_{\mathbb{P}}(r_{\mathtt{SA}} + r_{\mathtt{HO}}),$$

which is Assertion (a).

By Theorem G.1, we have

$$\left|V^* - \sup_{\boldsymbol{q},\boldsymbol{x}}\left|\bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\boldsymbol{Z}^*(\boldsymbol{q})\right|\right| = o_{\mathbb{P}}(R_{\mathtt{PI}}),$$

which is Assertion (b).

Assertion (c) follows from the definition of the process $\boldsymbol{Z}^*(\cdot)$ and the fact that $\bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})$ only depends on the data via $\mathbf{X}_n$. □

Lemma G.9 (First sequence)**.** *There exists a sequence of positive numbers $\nu_{n,1} \to 0$ such that w. p. a. 1*

$$k^*(1-\alpha) \le \tilde{k}^*(1-\alpha+\nu_{n,1}) + R_{\texttt{PI}} \quad \text{and} \quad k^*(1-\alpha) \ge \tilde{k}^*(1-\alpha-\nu_{n,1}) - R_{\texttt{PI}}.$$

Proof. This follows from $|V^* - \widetilde{V}^*| = o_{\mathbb{P}}(R_{\texttt{PI}})$ by Lemma G.8, directly applying Lemma C.8. □

Lemma G.10 (Second sequence)**.** *There exists a constant $C_{\widetilde{V}^*} > 0$ such that with $1-o(1)$ probability*

$$\sup_{u\in\mathbb{R}} \mathbb{P}\big\{\big|\widetilde{V}^* - u\big| \le R_{\texttt{PI}} + r_{\texttt{SA}} + r_{\texttt{HO}} \,\big|\, \boldsymbol{D}_n\big\} \le C_{\widetilde{V}^*}(R_{\texttt{PI}} + r_{\texttt{SA}} + r_{\texttt{HO}})\sqrt{\log(1/h)} =: \nu_{n,2},$$

Moreover, for the sequence $\nu_{n,2} \to 0$ just defined, the following holds w. p. a. 1:

$$\tilde{k}^*(1-\alpha-\nu_{n,1}) - \tilde{k}^*(1-\alpha-\nu_{n,1}-\nu_{n,2}) \ge R_{\texttt{PI}} + r_{\texttt{SA}} + r_{\texttt{HO}}, \tag{G.22}$$

$$\tilde{k}^*(1-\alpha+\nu_{n,1}+\nu_{n,2}) - \tilde{k}^*(1-\alpha+\nu_{n,1}) \ge R_{\texttt{PI}} + r_{\texttt{SA}} + r_{\texttt{HO}}. \tag{G.23}$$

Proof. By Lemma C.9, using that $\bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\boldsymbol{Z}^*(\boldsymbol{q})$ is a separable mean-zero Gaussian conditionally on $\mathbf{X}_n$ process on $\mathcal{Q}\times\mathcal{X}$ with $\mathbb{E}\big[\big(\bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\boldsymbol{Z}^*(\boldsymbol{q})\big)^2 \,\big|\, \mathbf{X}_n\big] = 1$, we have with $1-o(1)$ probability

$$\begin{aligned}
&\sup_{u\in\mathbb{R}} \mathbb{P}\big\{\big|\widetilde{V}^* - u\big| \le R_{\texttt{PI}} + r_{\texttt{SA}} + r_{\texttt{HO}} \,\big|\, \boldsymbol{D}_n\big\} \\
&\quad = \sup_{u\in\mathbb{R}} \mathbb{P}\big\{\big|\widetilde{V}^* - u\big| \le R_{\texttt{PI}} + r_{\texttt{SA}} + r_{\texttt{HO}} \,\big|\, \mathbf{X}_n\big\} \\
&\quad \le 4(R_{\texttt{PI}} + r_{\texttt{SA}} + r_{\texttt{HO}})\bigg(\mathbb{E}\Big[\sup_{\boldsymbol{q},\boldsymbol{x}}\big|\bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\boldsymbol{Z}^*(\boldsymbol{q})\big| \,\Big|\, \mathbf{X}_n\Big] + 1\bigg) \\
&\quad \le 4(R_{\texttt{PI}} + r_{\texttt{SA}} + r_{\texttt{HO}})\bigg(\sup_{\boldsymbol{q},\boldsymbol{x}}\big\|\bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\big\|_1 \mathbb{E}\Big[\sup_{\boldsymbol{q}}\big\|\boldsymbol{Z}^*(\boldsymbol{q})\big\|_\infty \,\Big|\, \mathbf{X}_n\Big] + 1\bigg) \\
&\quad \overset{\text{(a)}}{\lesssim} (R_{\texttt{PI}} + r_{\texttt{SA}} + r_{\texttt{HO}})\bigg(\mathbb{E}\Big[\sup_{\boldsymbol{q}}\big\|\boldsymbol{Z}^*(\boldsymbol{q})\big\|_\infty \,\Big|\, \mathbf{X}_n\Big] + 1\bigg) \\
&\quad \overset{\text{(b)}}{\lesssim} (R_{\texttt{PI}} + r_{\texttt{SA}} + r_{\texttt{HO}})\sqrt{\log(1/h)},
\end{aligned}$$

where in (a) we used $\big\|\bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\big\|_1 \lesssim 1$ w. p. a. 1 by Lemma F.7; (b) is by Lemma G.6.

We will now prove (G.22). Note that $\sup_{u\in\mathbb{R}} \mathbb{P}\big\{|\widetilde{V}^* - u| \le R_{\texttt{PI}} + r_{\texttt{SA}} + r_{\texttt{HO}} \,\big|\, \boldsymbol{D}_n\big\} \le \nu_{n,2}$ w. p. a. 1 implies that w. p. a. 1

$$\sup_{u\in\mathbb{R}} \mathbb{P}\Big\{u < \widetilde{V}^* \le u + R_{\texttt{PI}} + r_{\texttt{SA}} + r_{\texttt{HO}} \,\Big|\, \boldsymbol{D}_n\Big\} \le \nu_{n,2},$$

that is,

$$\sup_{u\in\mathbb{R}}\Big\{\mathbb{P}\big\{\widetilde{V}^* \le u + R_{\texttt{PI}} + r_{\texttt{SA}} + r_{\texttt{HO}} \,\big|\, \boldsymbol{D}_n\big\} - \mathbb{P}\big\{\widetilde{V}^* \le u \,\big|\, \boldsymbol{D}_n\big\}\Big\} \le \nu_{n,2}.$$

Since this is true for any u, we can in particular replace u with a random variable $\tilde{k}^*(1-\alpha-\nu_{n,1}-\nu_{n,2})$. Using $\mathbb{P}\big\{\widetilde{V}^* \le \tilde{k}^*(1-\alpha-\nu_{n,1}-\nu_{n,2}) \,\big|\, \boldsymbol{D}_n\big\} \ge 1-\alpha-\nu_{n,1}-\nu_{n,2}$, this gives w. p. a. 1

$$\mathbb{P}\big\{\widetilde{V}^* \le \tilde{k}^*(1-\alpha-\nu_{n,1}-\nu_{n,2}) + R_{\texttt{PI}} + r_{\texttt{SA}} + r_{\texttt{HO}} \,\big|\, \boldsymbol{D}_n\big\} - (1-\alpha-\nu_{n,1}-\nu_{n,2}) \le \nu_{n,2}$$

or

$$\mathbb{P}\big\{\widetilde{V}^* \le \tilde{k}^*(1-\alpha-\nu_{n,1}-\nu_{n,2}) + R_{\texttt{PI}} + r_{\texttt{SA}} + r_{\texttt{HO}} \,\big|\, \boldsymbol{D}_n\big\} \le 1-\alpha-\nu_{n,1}.$$

By monotonicity of a (conditional) distribution function, this means that w. p. a. 1

$$\tilde{k}^*(1-\alpha-\nu_{n,1}-\nu_{n,2}) + R_{\mathtt{PI}} + r_{\mathtt{SA}} + r_{\mathtt{HO}} \le \tilde{k}^*(1-\alpha-\nu_{n,1}).$$

This proves the inequality (G.22). The inequality (G.23) is proven similarly. □

Concluding the proof of Theorem G.7. Note that

$$\begin{aligned}
\mathbb{P}\{V > k^*(1-\alpha)\} &\overset{\text{(a)}}{\le} \mathbb{P}\Big\{V > \tilde{k}^*(1-\alpha-\nu_{n,1}) - R_{\mathtt{PI}}\Big\} \\
&\overset{\text{(b)}}{\le} \mathbb{P}\Big\{\widetilde{V} > \tilde{k}^*(1-\alpha-\nu_{n,1}) - (R_{\mathtt{PI}} + r_{\mathtt{SA}} + r_{\mathtt{HO}})\Big\} + o(1) \\
&\overset{\text{(c)}}{\le} \mathbb{P}\Big\{\widetilde{V} > \tilde{k}^*(1-\alpha-\nu_{n,1}-\nu_{n,2})\Big\} + o(1) \\
&\overset{\text{(d)}}{=} \mathbb{E}\Big[\mathbb{P}\Big\{\widetilde{V} > \tilde{k}^*(1-\alpha-\nu_{n,1}-\nu_{n,2}) \,\Big|\, \mathbf{X}_n\Big\}\Big] + o(1) \\
&\overset{\text{(e)}}{\le} \alpha + \nu_{n,1} + \nu_{n,2} + o(1) = \alpha + o(1),
\end{aligned}$$

where (a) is by Lemma G.9, (b) is by Assertion (a) in Lemma G.8, (c) is by Lemma G.10, (d) is by the law of iterated expectations, (e) is by the definition of a conditional quantile and using that $\widetilde{V}$ has the same conditional distribution as $\widetilde{V}^*$.

Similarly,

$$\begin{aligned}
\mathbb{P}\big\{V > k^*(1-\alpha)\big\} &\overset{\text{(a)}}{\ge} \mathbb{P}\big\{V > \tilde{k}^*(1-\alpha+\nu_{n,1}) + R_{\mathtt{PI}}\big\} + o(1) \\
&\overset{\text{(b)}}{\ge} \mathbb{P}\big\{\widetilde{V} > \tilde{k}^*(1-\alpha+\nu_{n,1}) + (R_{\mathtt{PI}} + r_{\mathtt{SA}} + r_{\mathtt{HO}})\big\} + o(1) \\
&\overset{\text{(c)}}{\ge} \mathbb{P}\big\{\widetilde{V} > \tilde{k}^*(1-\alpha+\nu_{n,1}+\nu_{n,2})\big\} + o(1) \\
&\overset{\text{(d)}}{=} \mathbb{E}\big[\mathbb{P}\big\{\widetilde{V} > \tilde{k}^*(1-\alpha+\nu_{n,1}+\nu_{n,2}) \,\big|\, \mathbf{X}_n\big\}\big] + o(1) \\
&\overset{\text{(e)}}{=} \alpha - \nu_{n,1} - \nu_{n,2} + o(1) = \alpha + o(1),
\end{aligned}$$

where (a) is by Lemma G.9, (b) is by Assertion (a) in Lemma G.8, (c) is by Lemma G.10, (d) is by the law of iterated expectations. In (e) we used that the distribution function of $\widetilde{V}$ conditional on $\mathbf{X}_n$ is continuous w. p. a. 1, because by the same anti-concentration argument as in the proof of Lemma G.10 there is a positive constant C such that on an event $\mathcal{A}_n$ satisfying $\mathbb{P}\{\mathcal{A}_n\} \to 1$ we have for any $\varepsilon > 0$

$$\sup_{u\in\mathbb{R}} \mathbb{P}\big\{|\widetilde{V} - u| \le \varepsilon \,\big|\, \mathbf{X}_n\big\} \le C\varepsilon\sqrt{\log(1/h)}.$$

In particular, on $\mathcal{A}_n$ all jumps of the distribution function of $\widetilde{V}$ conditional on $\mathbf{X}_n$ are bounded by $C\varepsilon\sqrt{\log(1/h)}$, which implies that the distribution function is continuous on $\mathcal{A}_n$, since ε is arbitrary. Theorem G.7 is proven. □

The following theorem uses a Gaussian anti-concentration result from [17] to convert $o_{\mathbb{P}}(\cdot)$ bounds obtained above to the bound on the Kolmogorov-Smirnov distance (sup-norm distance of distribution functions), similarly to [11].

G.3 Kolmogorov-Smirnov distance bound

Theorem G.11 (Kolmogorov-Smirnov Distance)**.** *Suppose all conditions of Theorem G.1 are true, and* $(R_{\mathtt{PI}} + r_{\mathtt{SA}} + r_{\mathtt{HO}})\sqrt{\log n} = o(1)$*. Then*

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\left\{\sup_{\boldsymbol{q}\in\mathcal{Q}}\sup_{\boldsymbol{x}\in\mathcal{X}}|T_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})| \le u\right\} - \mathbb{P}\left\{\sup_{\boldsymbol{q}\in\mathcal{Q}}\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\frac{h^{d/2}\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}}\widehat{\mathbf{Q}}_{\boldsymbol{q}}^{-1}}{\sqrt{\widehat{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})}}\widehat{\boldsymbol{Z}}^*(\boldsymbol{q})\right| \le u \,\middle|\, \boldsymbol{D}_n\right\}\right| = o_{\mathbb{P}}(1).$$

Proof. We will rely on the following lemma which is proven in Appendix G.3.1 by the same discretization argument as in Theorem G.1.

Lemma G.12. *On the same probability space, there exists a mean-zero unconditionally Gaussian process* $\widetilde{\boldsymbol{Z}}(\cdot)$ *in* $\mathcal{C}(\mathcal{Q},\mathbb{R}^K)$ *such that*

$$\widetilde{\boldsymbol{Z}}(\cdot) \perp\!\!\!\perp \mathbf{X}_n, \tag{G.24}$$

$$\mathbb{E}\left[\left(\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}(\boldsymbol{q})\right)^2\right] = 1, \tag{G.25}$$

$$\mathbb{E}\left[\sup_{\boldsymbol{q}\in\mathcal{Q}}\left\|\boldsymbol{Z}(\boldsymbol{q}) - \widetilde{\boldsymbol{Z}}(\boldsymbol{q})\right\|_\infty \,\middle|\, \mathbf{X}_n\right] \lesssim_{\mathbb{P}} \left(\frac{\log n}{nh^d}\right)^{1/(4d_{\mathcal{Q}}+4)}\sqrt{\log n}. \tag{G.26}$$

Also, there exists a process $\widetilde{\boldsymbol{Z}}^*(\cdot)$ *in* $\mathcal{C}(\mathcal{Q},\mathbb{R}^K)$*, such that*

$$\widetilde{\boldsymbol{Z}}^*(\cdot) \perp\!\!\!\perp \mathbf{D}_n, \tag{G.27}$$

$$\left(\mathbf{X}_n, \boldsymbol{Z}(\cdot), \widetilde{\boldsymbol{Z}}(\cdot)\right) \overset{d}{=} \left(\mathbf{X}_n, \boldsymbol{Z}^*(\cdot), \widetilde{\boldsymbol{Z}}^*(\cdot)\right), \tag{G.28}$$

and in particular

$$\mathbb{E}\left[\sup_{\boldsymbol{q}\in\mathcal{Q}}\left\|\boldsymbol{Z}^*(\boldsymbol{q}) - \widetilde{\boldsymbol{Z}}^*(\boldsymbol{q})\right\|_\infty \,\middle|\, \mathbf{X}_n\right] \lesssim_{\mathbb{P}} \left(\frac{\log n}{nh^d}\right)^{1/(4d_{\mathcal{Q}}+4)}\sqrt{\log n}. \tag{G.29}$$

Since $\widetilde{\boldsymbol{Z}}^*(\cdot) \overset{d}{=} \widetilde{\boldsymbol{Z}}(\cdot)$,

$$\mathbb{P}\left\{\sup_{\boldsymbol{q},\boldsymbol{x}}\left|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}(\boldsymbol{q})\right| \le u\right\} = \mathbb{P}\left\{\sup_{\boldsymbol{q},\boldsymbol{x}}\left|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}^*(\boldsymbol{q})\right| \le u\right\}.$$

This means that, by the triangle inequality, it is sufficient to prove the following bounds:

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\left\{\sup_{\boldsymbol{q},\boldsymbol{x}}\left|T_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\right| \le u\right\} - \mathbb{P}\left\{\sup_{\boldsymbol{q},\boldsymbol{x}}\left|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}(\boldsymbol{q})\right| \le u\right\}\right| = o(1), \tag{G.30}$$

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\left\{\sup_{\boldsymbol{q},\boldsymbol{x}}\left|\hat{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widehat{\boldsymbol{Z}}^*(\boldsymbol{q})\right| \le u \,\middle|\, \boldsymbol{D}_n\right\} - \mathbb{P}\left\{\sup_{\boldsymbol{q},\boldsymbol{x}}\left|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}^*(\boldsymbol{q})\right| \le u\right\}\right| = o_{\mathbb{P}}(1). \tag{G.31}$$

We will now prove (G.30). Note that for any random variables ξ and η and any $s > 0$,

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\{\xi \le u\} - \mathbb{P}\{\eta \le u\}\right| \le \sup_{u\in\mathbb{R}}\mathbb{P}\{|\eta - u| \le s\} + \mathbb{P}\{|\xi - \eta| > s\}, \tag{G.32}$$

which follows from the two bounds

$$\begin{aligned}
\mathbb{P}\{\xi \le u\} &\le \mathbb{P}\{\xi \le u \text{ and } |\xi - \eta| \le s\} + \mathbb{P}\{|\xi - \eta| > s\} \\
&\le \mathbb{P}\{\eta \le u + s\} + \mathbb{P}\{|\xi - \eta| > s\} \\
&\le \mathbb{P}\{\eta \le u\} + \mathbb{P}\{u < \eta \le u + s\} + \mathbb{P}\{|\xi - \eta| > s\}; \\
\mathbb{P}\{\xi > u\} &\le \mathbb{P}\{\xi > u \text{ and } |\xi - \eta| \le s\} + \mathbb{P}\{|\xi - \eta| > s\} \\
&\le \mathbb{P}\{\eta > u - s\} + \mathbb{P}\{|\xi - \eta| > s\} \\
&\le \mathbb{P}\{\eta > u\} + \mathbb{P}\{u - s < \eta \le u\} + \mathbb{P}\{|\xi - \eta| > s\}.
\end{aligned}$$

Strategy To obtain (G.30), we will apply (G.32) with $\xi = \sup_{\boldsymbol{q},\boldsymbol{x}}|T_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})|$, $\eta = \sup_{\boldsymbol{q},\boldsymbol{x}}\big|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}(\boldsymbol{q})\big|$ and $s = r_{\mathtt{SA}} + r_{\mathtt{HO}}$, use that the term $\mathbb{P}\{|\xi-\eta| > s\}$ is $o(1)$ because $\big|T_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\big|$ and $\big|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}(\boldsymbol{q})\big|$ are close, and apply Gaussian anti-concentration Lemma C.9 to show that $\sup_{u\in\mathbb{R}}\mathbb{P}\{|\eta-u|\le s\}$ is also $o(1)$. The argument for (G.31) will be similar.

Proof of (G.30) By $\sup_{\boldsymbol{q},\boldsymbol{x}}\|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\|_1 \lesssim 1$ (from Lemma F.7) and (G.26), a simple application of Lemma C.2 gives

$$h^{d/2}\sup_{\boldsymbol{q},\boldsymbol{x}}\left|\frac{\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}}\mathbf{Q}_{0,\boldsymbol{q}}^{-1}}{\sqrt{\Omega_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})}}\boldsymbol{Z}(\boldsymbol{q}) - \frac{\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}}\mathbf{Q}_{0,\boldsymbol{q}}^{-1}}{\sqrt{\Omega_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})}}\widetilde{\boldsymbol{Z}}(\boldsymbol{q})\right| = o_{\mathbb{P}}(r_{\mathtt{SA}}). \tag{G.33}$$

By (F.29) in Lemma F.7, we have $\sup_{\boldsymbol{q},\boldsymbol{x}}\|\bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) - \boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\|_1 \lesssim_{\mathbb{P}} \sqrt{\frac{\log(1/h)}{nh^d}}$, which by Lemma G.6 gives $\sup_{\boldsymbol{q},\boldsymbol{x}}|\bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\boldsymbol{Z}(\boldsymbol{q}) - \boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\boldsymbol{Z}(\boldsymbol{q})| \lesssim_{\mathbb{P}} \frac{\log(1/h)}{\sqrt{nh^d}} = o(r_{\mathtt{SA}})$. Combining it with Theorem F.4, Lemma F.8 and (G.33),

$$\mathbb{P}\Big\{\Big|\sup_{\boldsymbol{q},\boldsymbol{x}}\big|T_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\big| - \sup_{\boldsymbol{q},\boldsymbol{x}}\big|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}(\boldsymbol{q})\big|\Big| > r_{\mathtt{SA}} + r_{\mathtt{HO}}\Big\} = o(1).$$

Then we can apply (G.32) with $\xi = \sup_{\boldsymbol{q},\boldsymbol{x}}|T_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})|$, $\eta = \sup_{\boldsymbol{q},\boldsymbol{x}}\big|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}(\boldsymbol{q})\big|$ and $s = r_{\mathtt{SA}} + r_{\mathtt{HO}}$, and get

$$\begin{aligned}
&\sup_{u\in\mathbb{R}}\Big|\mathbb{P}\Big\{\sup_{\boldsymbol{q},\boldsymbol{x}}|T_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})| \le u\Big\} - \mathbb{P}\Big\{\sup_{\boldsymbol{q},\boldsymbol{x}}\big|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}(\boldsymbol{q})\big| \le u\Big\}\Big| \\
&\quad\le \sup_{u\in\mathbb{R}}\mathbb{P}\Big\{\Big|\sup_{\boldsymbol{q},\boldsymbol{x}}\Big|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}(\boldsymbol{q})\Big| - u\Big| \le r_{\mathtt{SA}} + r_{\mathtt{HO}}\Big\} + o(1).
\end{aligned}$$

For (G.30), it is left to show that

$$\sup_{u\in\mathbb{R}}\mathbb{P}\Big\{\Big|\sup_{\boldsymbol{q},\boldsymbol{x}}\big|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}(\boldsymbol{q})\big| - u\Big| \le r_{\mathtt{SA}} + r_{\mathtt{HO}}\Big\} = o(1).$$

We will show a stronger version

$$\sup_{u\in\mathbb{R}}\mathbb{P}\Big\{\Big|\sup_{\boldsymbol{q},\boldsymbol{x}}\big|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}(\boldsymbol{q})\big| - u\Big| \le R_{\mathtt{PI}} + r_{\mathtt{SA}} + r_{\mathtt{HO}}\Big\} = o(1). \tag{G.34}$$

We apply Lemma C.9:

$$\begin{aligned}
&\sup_{u\in\mathbb{R}}\mathbb{P}\Big\{\Big|\sup_{\boldsymbol{q},\boldsymbol{x}}\big|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}(\boldsymbol{q})\big| - u\Big| \le R_{\mathtt{PI}} + r_{\mathtt{SA}} + r_{\mathtt{HO}}\Big\} \\
&\quad\le 4(R_{\mathtt{PI}} + r_{\mathtt{SA}} + r_{\mathtt{HO}})\Big(\mathbb{E}\Big[\sup_{\boldsymbol{q},\boldsymbol{x}}\big|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}(\boldsymbol{q})\big|\Big] + 1\Big) \\
&\quad\le 4(R_{\mathtt{PI}} + r_{\mathtt{SA}} + r_{\mathtt{HO}})\Big(\mathbb{E}\Big[\sup_{\boldsymbol{q},\boldsymbol{x}}\|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\|_1\big\|\widetilde{\boldsymbol{Z}}(\boldsymbol{q})\big\|_\infty\Big] + 1\Big) \\
&\quad\overset{\text{(a)}}{\lesssim} (R_{\mathtt{PI}} + r_{\mathtt{SA}} + r_{\mathtt{HO}})\Big(\mathbb{E}\Big[\sup_{\boldsymbol{q}}\big\|\widetilde{\boldsymbol{Z}}(\boldsymbol{q})\big\|_\infty\Big] + 1\Big) \\
&\quad\overset{\text{(b)}}{\lesssim} (R_{\mathtt{PI}} + r_{\mathtt{SA}} + r_{\mathtt{HO}})\sqrt{\log(1/h)},
\end{aligned} \tag{G.35}$$

where in (a) we used $\|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\|_1 \lesssim 1$, see (F.28); (b) is by

$$\begin{aligned}\mathbb{E}\Big[\sup_{\boldsymbol{q}}\big\|\widetilde{\boldsymbol{Z}}(\boldsymbol{q})\big\|_\infty\Big] &= \mathbb{E}\Big[\sup_{\boldsymbol{q}}\big\|\widetilde{\boldsymbol{Z}}(\boldsymbol{q})\big\|_\infty \,\Big|\, \mathbf{X}_n\Big] \\ &\le \mathbb{E}\Big[\sup_{\boldsymbol{q}}\big\|\widetilde{\boldsymbol{Z}}(\boldsymbol{q}) - \boldsymbol{Z}(\boldsymbol{q})\big\|_\infty \,\Big|\, \mathbf{X}_n\Big] + \mathbb{E}\Big[\sup_{\boldsymbol{q}}\big\|\boldsymbol{Z}(\boldsymbol{q})\big\|_\infty \,\Big|\, \mathbf{X}_n\Big] \\ &\overset{\text{(c)}}{=} o_{\mathbb{P}}(r_{\mathtt{SA}}) + O_{\mathbb{P}}\big(\sqrt{\log(1/h)}\big) \overset{\text{(d)}}{=} O_{\mathbb{P}}\big(\sqrt{\log(1/h)}\big)\end{aligned}$$

using (G.26) and Lemma G.6 for (c), $r_{\mathtt{SA}} \lesssim 1 \lesssim \sqrt{\log(1/h)}$ for (d), and noting that

$$\mathbb{E}\Big[\sup_{\boldsymbol{q}}\big\|\widetilde{\boldsymbol{Z}}(\boldsymbol{q})\big\|_\infty\Big] \lesssim_{\mathbb{P}} \sqrt{\log(1/h)}$$

is equivalent to

$$\mathbb{E}\Big[\sup_{\boldsymbol{q}}\big\|\widetilde{\boldsymbol{Z}}(\boldsymbol{q})\big\|_\infty\Big] \lesssim \sqrt{\log(1/h)}.$$

Since $(R_{\mathtt{PI}} + r_{\mathtt{SA}} + r_{\mathtt{HO}})\sqrt{\log(1/h)} = o(1)$, the right-hand side in (G.35) is $o(1)$, proving (G.34), which was sufficient for (G.30).

Proof of (G.31) By $\sup_{\boldsymbol{q},\boldsymbol{x}}\|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\|_1 \lesssim 1$ (from Lemma F.7) and (G.29), a simple application of Lemma C.2 gives

$$h^{d/2}\sup_{\boldsymbol{q},\boldsymbol{x}}\left|\frac{\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}}\mathbf{Q}_{0,\boldsymbol{q}}^{-1}}{\sqrt{\Omega_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})}}\boldsymbol{Z}^*(\boldsymbol{q}) - \frac{\boldsymbol{p}^{(\boldsymbol{v})}(\boldsymbol{x})^{\mathsf{T}}\mathbf{Q}_{0,\boldsymbol{q}}^{-1}}{\sqrt{\Omega_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})}}\widetilde{\boldsymbol{Z}}^*(\boldsymbol{q})\right| = o_{\mathbb{P}}(r_{\mathtt{SA}}). \tag{G.36}$$

By Theorem G.1, we have $\sup_{\boldsymbol{q},\boldsymbol{x}}\big|\hat{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widehat{\boldsymbol{Z}}^*(\boldsymbol{q}) - \bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\boldsymbol{Z}^*(\boldsymbol{q})\big| = o_{\mathbb{P}}(R_{\mathtt{PI}})$. By (F.29) in Lemma F.7, we have $\sup_{\boldsymbol{q},\boldsymbol{x}}\|\bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q}) - \boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})\|_1 \lesssim_{\mathbb{P}} \sqrt{\frac{\log(1/h)}{nh^d}}$, which by Lemma G.6 gives $\sup_{\boldsymbol{q},\boldsymbol{x}}|\bar{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\boldsymbol{Z}^*(\boldsymbol{q}) - \boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\boldsymbol{Z}^*(\boldsymbol{q})| \lesssim_{\mathbb{P}} \frac{\log(1/h)}{\sqrt{nh^d}} = o(r_{\mathtt{SA}})$. By the triangle inequality,

$$\sup_{\boldsymbol{q},\boldsymbol{x}}|\hat{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widehat{\boldsymbol{Z}}^*(\boldsymbol{q}) - \boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\boldsymbol{Z}^*(\boldsymbol{q})| = o_{\mathbb{P}}(R_{\mathtt{PI}} + r_{\mathtt{SA}}).$$

Combining with (G.36) and applying the triangle inequality again, we obtain

$$\mathbb{P}\Big\{\Big|\sup_{\boldsymbol{q},\boldsymbol{x}}\big|\hat{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widehat{\boldsymbol{Z}}^*(\boldsymbol{q})\big| - \sup_{\boldsymbol{q},\boldsymbol{x}}\big|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}^*(\boldsymbol{q})\big|\Big| > R_{\mathtt{PI}} + r_{\mathtt{SA}}\Big\} = o(1),$$

implying by Markov's inequality that for any $\varepsilon > 0$,

$$\mathbb{P}\Big\{\mathbb{P}\Big\{\Big|\sup_{\boldsymbol{q},\boldsymbol{x}}\big|\hat{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widehat{\boldsymbol{Z}}^*(\boldsymbol{q})\big| - \sup_{\boldsymbol{q},\boldsymbol{x}}\big|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}^*(\boldsymbol{q})\big|\Big| > R_{\mathtt{PI}} + r_{\mathtt{SA}} \,\Big|\, \boldsymbol{D}_n\Big\} > \varepsilon\Big\} = o(1),$$

that is,

$$\mathbb{P}\Big\{\Big|\sup_{\boldsymbol{q},\boldsymbol{x}}\big|\hat{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widehat{\boldsymbol{Z}}^*(\boldsymbol{q})\big| - \sup_{\boldsymbol{q},\boldsymbol{x}}\big|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}^*(\boldsymbol{q})\big|\Big| > R_{\mathtt{PI}} + r_{\mathtt{SA}} \,\Big|\, \boldsymbol{D}_n\Big\} = o_{\mathbb{P}}(1).$$

Then we can apply (G.32) with $\mathbb{P}\{\cdot \mid \boldsymbol{D}_n\}$ instead of $\mathbb{P}\{\cdot\}$, $\xi = \sup_{\boldsymbol{q},\boldsymbol{x}}\big|\hat{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widehat{\boldsymbol{Z}}^*(\boldsymbol{q})\big|$, $\eta = \sup_{\boldsymbol{q},\boldsymbol{x}}\big|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}^*(\boldsymbol{q})\big|$ and $s = R_{\mathtt{PI}} + r_{\mathtt{SA}}$, and get

$$
\begin{aligned}
&\sup_{u\in\mathbb{R}}\Big|\mathbb{P}\Big\{\sup_{\boldsymbol{q},\boldsymbol{x}}\big|\hat{\boldsymbol{\ell}}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widehat{\boldsymbol{Z}}^*(\boldsymbol{q})\big| \le u \,\Big|\, \boldsymbol{D}_n\Big\} - \mathbb{P}\Big\{\sup_{\boldsymbol{q},\boldsymbol{x}}\big|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}^*(\boldsymbol{q})\big| \le u\Big\}\Big| \\
&\quad \le \sup_{u\in\mathbb{R}}\mathbb{P}\Big\{\Big|\sup_{\boldsymbol{q},\boldsymbol{x}}\big|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}^*(\boldsymbol{q})\big| - u\Big| \le R_{\mathtt{PI}} + r_{\mathtt{SA}}\Big\} + o_{\mathbb{P}}(1) \\
&\quad = \sup_{u\in\mathbb{R}}\mathbb{P}\Big\{\Big|\sup_{\boldsymbol{q},\boldsymbol{x}}\Big|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}(\boldsymbol{q})\Big| - u\Big| \le R_{\mathtt{PI}} + r_{\mathtt{SA}}\Big\} + o_{\mathbb{P}}(1),
\end{aligned}
$$

where we used that $\widetilde{\boldsymbol{Z}}^*(\boldsymbol{q})$ is independent of the data allowing us to remove the conditioning on $\boldsymbol{D}_n$, and again that $\widetilde{\boldsymbol{Z}}(\cdot)$ and $\widetilde{\boldsymbol{Z}}^*(\cdot)$ have the same laws.

It is left to use that

$$
\begin{aligned}
&\sup_{u\in\mathbb{R}}\mathbb{P}\Big\{\Big|\sup_{\boldsymbol{q},\boldsymbol{x}}\big|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}(\boldsymbol{q})\big| - u\Big| \le R_{\mathtt{PI}} + r_{\mathtt{SA}}\Big\} \\
&\quad \le \sup_{u\in\mathbb{R}}\mathbb{P}\Big\{\Big|\sup_{\boldsymbol{q},\boldsymbol{x}}\big|\boldsymbol{\ell}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}(\boldsymbol{q})\big| - u\Big| \le R_{\mathtt{PI}} + r_{\mathtt{SA}} + r_{\mathtt{HO}}\Big\} \stackrel{\text{(a)}}{=} o(1),
\end{aligned}
$$

where (a) is by (G.34).

Theorem G.11 is proven. ☐

G.3.1 Proof of Lemma G.12: construction of $\widetilde{Z}(\cdot)$ and $\widetilde{Z}^*(\cdot)$

Recall that the conditional covariance structure of $\boldsymbol{Z}(\cdot)$ is

$$
\mathbb{E}\big[\boldsymbol{Z}(\boldsymbol{q})\boldsymbol{Z}(\tilde{\boldsymbol{q}})^{\mathsf{T}} \,\big|\, \mathbf{X}_n\big] = h^{-d}\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}
$$

with $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$ defined in (G.1). The matrix $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$ is approximated by the non-random matrix $\boldsymbol{\Sigma}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$ defined by

$$
\boldsymbol{\Sigma}_{\boldsymbol{q},\tilde{\boldsymbol{q}}} := \mathbb{E}\big[S_{\boldsymbol{q},\tilde{\boldsymbol{q}}}(\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\boldsymbol{q}))\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\tilde{\boldsymbol{q}}))\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\big].
$$

By the same argument as in Lemma C.11, we have

$$
\sup_{\boldsymbol{q},\tilde{\boldsymbol{q}}}\|\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q},\tilde{\boldsymbol{q}}} - \boldsymbol{\Sigma}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}\| \lesssim_{\mathbb{P}} h^d\Big(\frac{\log(1/h)}{nh^d}\Big)^{1/2} =: h^d r_{\boldsymbol{\Sigma}}. \tag{G.37}
$$

Discretization and controlling the deviations of $\boldsymbol{Z}(\cdot)$ Then we proceed with the same discretization as in Theorem G.1. Let $\mathcal{Q}^{\delta_n} := \{\boldsymbol{q}_1,\ldots,\boldsymbol{q}_{|\mathcal{Q}^{\delta_n}|}\}$ be an internal δ_n-covering of $\mathcal{Q}$ with respect to the 2-norm $\|\cdot\|$ of cardinality $|\mathcal{Q}^{\delta_n}| \lesssim 1/\delta_n^{d_{\mathcal{Q}}}$, where δ_n is chosen later. As already proven in Lemma F.3, there is a bound on the deviations of $\boldsymbol{Z}(\cdot)$:

$$
\mathbb{E}\Big[\sup_{\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|\le\delta_n}\|\boldsymbol{Z}(\boldsymbol{q}) - \boldsymbol{Z}(\tilde{\boldsymbol{q}})\|_\infty \,\Big|\, \mathbf{X}_n\Big] \lesssim \sqrt{\delta_n\log\frac{K}{\delta_n}} \quad \text{on } \mathcal{A},
$$

where $\mathcal{A}$ is the event from (G.7).

Closeness of the discrete vectors Consider now the conditionally Gaussian vector $\boldsymbol{Z}(\boldsymbol{q}|_{\mathcal{Q}^{\delta_n}}) \in \mathbb{R}^{K|\mathcal{Q}^{\delta_n}|}$ and denote

$$\bar{\boldsymbol{\Sigma}}^{\delta_n} := h^d \mathbb{V}[\boldsymbol{Z}(\boldsymbol{q}|_{\mathcal{Q}^{\delta_n}}) \,|\, \mathbf{X}_n] = (\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q}_k, \boldsymbol{q}_m})_{k,m=1}^{|\mathcal{Q}^{\delta_n}|} \in \mathbb{R}^{K|\mathcal{Q}^{\delta_n}| \times K|\mathcal{Q}^{\delta_n}|},$$

so that

$$\boldsymbol{Z}(\boldsymbol{q}|_{\mathcal{Q}^{\delta_n}}) = h^{-d/2} (\bar{\boldsymbol{\Sigma}}^{\delta_n})^{1/2} \boldsymbol{\xi}_n, \quad \text{where } \mathbb{P}_{\boldsymbol{\xi}_n | \mathbf{X}_n} = \mathcal{N}\big(0, \mathbf{I}_{\mathbb{R}^{K|\mathcal{Q}^{\delta_n}|}}\big).$$

(Since $\mathbb{P}_{\boldsymbol{\xi}_n | \mathbf{X}_n}$ does not depend on $\mathbf{X}_n$, the vector $\boldsymbol{\xi}_n$ is independent of $\mathbf{X}_n$.) Discretizing $\boldsymbol{\Sigma}_{\boldsymbol{q}, \tilde{\boldsymbol{q}}}$ in the same way, we can put

$$\boldsymbol{\Sigma}^{\delta_n} := \big(\boldsymbol{\Sigma}_{\boldsymbol{q}_k, \boldsymbol{q}_m}\big)_{k,m=1}^{|\mathcal{Q}^{\delta_n}|} \in \mathbb{R}^{K|\mathcal{Q}^{\delta_n}| \times K|\mathcal{Q}^{\delta_n}|}.$$

Since the functions $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q}, \tilde{\boldsymbol{q}}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{q}, \tilde{\boldsymbol{q}}}$ are close, the two matrices are close as well, which we will make precise in the following lemma.

Lemma G.13 (The matrices $\bar{\boldsymbol{\Sigma}}^{\delta_n}$ and $\boldsymbol{\Sigma}^{\delta_n}$ are close)**.** *We have*

$$\big\|\bar{\boldsymbol{\Sigma}}^{\delta_n} - \boldsymbol{\Sigma}^{\delta_n}\big\| \lesssim_{\mathbb{P}} |\mathcal{Q}^{\delta_n}| h^d r_{\boldsymbol{\Sigma}}.$$

Proof. We can bound the Frobenius norm using the the bound on each element of the matrix:

$$\big\|\bar{\boldsymbol{\Sigma}}^{\delta_n} - \boldsymbol{\Sigma}^{\delta_n}\big\| \leq \big\|\bar{\boldsymbol{\Sigma}}^{\delta_n} - \boldsymbol{\Sigma}^{\delta_n}\big\|_F \leq |\mathcal{Q}^{\delta_n}| \sup_{\boldsymbol{q}, \tilde{\boldsymbol{q}}} \big\|\bar{\boldsymbol{\Sigma}}_{\boldsymbol{q}, \tilde{\boldsymbol{q}}} - \boldsymbol{\Sigma}_{\boldsymbol{q}, \tilde{\boldsymbol{q}}}\big\|.$$

It is left to combine with Eq. (G.37). □

Applying Lemma C.5, we then have

$$\begin{aligned} h^{-d/2} \mathbb{E}\big[\big\|(\bar{\boldsymbol{\Sigma}}^{\delta_n})^{1/2} \boldsymbol{\xi}_n - (\boldsymbol{\Sigma}^{\delta_n})^{1/2} \boldsymbol{\xi}_n\big\|_\infty \,|\, \mathbf{X}_n\big] &\leq 2 h^{-d/2} \sqrt{\log(K|\mathcal{Q}^{\delta_n}|)} \|\bar{\boldsymbol{\Sigma}}^{\delta_n} - \boldsymbol{\Sigma}^{\delta_n}\|^{1/2} \\ &\lesssim_{\mathbb{P}} |\mathcal{Q}^{\delta_n}|^{1/2} {r_{\boldsymbol{\Sigma}}}^{1/2} \sqrt{\log(K|\mathcal{Q}^{\delta_n}|)}. \end{aligned} \tag{G.38}$$

Embedding the conditionally Gaussian vector into a conditionally Gaussian process By the same argument as in Lemma F.3, there exists a law on $\mathcal{X}^n \times \mathcal{C}(\mathcal{Q}, \mathbb{R}^K)$ which is the joint law of $\mathbf{X}_n' \overset{d}{=} \mathbf{X}_n$ and a conditionally on $\mathbf{X}_n'$ mean-zero Gaussian process $\widetilde{\boldsymbol{Z}}_n'(\boldsymbol{q})$ with the conditional covariance structure

$$\mathbb{E}\big[\widetilde{\boldsymbol{Z}}_n'(\boldsymbol{q}) \widetilde{\boldsymbol{Z}}_n'(\tilde{\boldsymbol{q}})^\mathsf{T} \,\big|\, \mathbf{X}_n'\big] = h^{-d} \boldsymbol{\Sigma}_{\boldsymbol{q}, \tilde{\boldsymbol{q}}}.$$

Continuity follows by the Kolmogorov-Chentsov theorem:

$$\mathbb{E}\big[\big\|\widetilde{\boldsymbol{Z}}_n'(\boldsymbol{q}) - \widetilde{\boldsymbol{Z}}_n'(\tilde{\boldsymbol{q}})\big\|^a \,\big|\, \mathbf{X}_n'\big] \leq C_n \|\boldsymbol{q} - \tilde{\boldsymbol{q}}\|^{d_{\mathcal{Q}} + b}, \tag{G.39}$$

for $a, b > 0$ chosen as follows. The vector $\widetilde{\boldsymbol{Z}}_n'(\boldsymbol{q}) - \widetilde{\boldsymbol{Z}}_n'(\tilde{\boldsymbol{q}})$ is conditionally on $\mathbf{X}_n'$ Gaussian with covariance

$$h^{-d}\big(\boldsymbol{\Sigma}_{\boldsymbol{q}, \boldsymbol{q}} - 2\boldsymbol{\Sigma}_{\boldsymbol{q}, \tilde{\boldsymbol{q}}} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{q}}}\big).$$

Hence, we have for any $m > 0$

$$\begin{aligned} &\mathbb{E}\big[\big\|\widetilde{\boldsymbol{Z}}_n'(\boldsymbol{q}) - \widetilde{\boldsymbol{Z}}_n'(\tilde{\boldsymbol{q}})\big\|^{2m} \,\big|\, \mathbf{X}_n'\big] \\ &\quad \leq h^{-dm} \big\|\boldsymbol{\Sigma}_{\boldsymbol{q}, \boldsymbol{q}} - 2\boldsymbol{\Sigma}_{\boldsymbol{q}, \tilde{\boldsymbol{q}}} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{q}}}\big\|^m \mathbb{E}[\|\boldsymbol{\xi}_K\|^{2m}] \\ &\quad \leq C_n \|\boldsymbol{q} - \tilde{\boldsymbol{q}}\|^m, \end{aligned}$$

where $\boldsymbol{\xi}_K \sim \mathcal{N}(0, \mathbf{I}_K)$, and we used that $\boldsymbol{\Sigma}_{\boldsymbol{q},\tilde{\boldsymbol{q}}}$ is Lipschitz in $\tilde{\boldsymbol{q}}$ a. s. (with the constant allowed to depend on n). So we can take, for example, $m = d_{\mathcal{Q}} + 1$, $a = 2m$, $b = 1$ in (G.39).

Projecting this $\mathcal{C}(\mathcal{Q}, \mathbb{R}^K)$-process onto $\mathcal{Q}^{\delta_n}$, we obtain a vector $\widetilde{\boldsymbol{Z}}'_n(\boldsymbol{q}|_{\mathcal{Q}^{\delta_n}}) \in \mathbb{R}^{K|\mathcal{Q}^{\delta_n}|}$ such that

$$\big(\mathbf{X}'_n, \widetilde{\boldsymbol{Z}}'_n(\boldsymbol{q}|_{\mathcal{Q}^{\delta_n}})\big) \overset{d}{=} \big(\mathbf{X}_n, h^{-d/2}(\boldsymbol{\Sigma}^{\delta_n})^{1/2}\boldsymbol{\xi}_n\big).$$

Since $\mathcal{C}(\mathcal{Q}, \mathbb{R}^K)$ is Polish, by Theorem 8.17 (transfer) in [28] on our probability space there exists a random element $\widetilde{\boldsymbol{Z}}(\boldsymbol{q}) \in \mathcal{C}(\mathcal{Q}, \mathbb{R}^K)$ such that

$$\big(\mathbf{X}'_n, \widetilde{\boldsymbol{Z}}'_n(\boldsymbol{q}|_{\mathcal{Q}^{\delta_n}}), \widetilde{\boldsymbol{Z}}'_n(\boldsymbol{q})\big) \overset{d}{=} \big(\mathbf{X}_n, h^{-d/2}(\boldsymbol{\Sigma}^{\delta_n})^{1/2}\boldsymbol{\xi}_n, \widetilde{\boldsymbol{Z}}(\boldsymbol{q})\big),$$

In particular, almost surely $h^{-d/2}(\boldsymbol{\Sigma}^{\delta_n})^{1/2}\boldsymbol{\xi}_n$ is the projection of $\widetilde{\boldsymbol{Z}}(\cdot)$ on $\mathcal{Q}^{\delta_n}$. Note that $\mathbb{P}_{\widetilde{\boldsymbol{Z}}(\boldsymbol{q})|\mathbf{X}_n}$ does not depend on $\mathbf{X}_n$, which means $\widetilde{\boldsymbol{Z}}(\boldsymbol{q})$ is independent of $\mathbf{X}_n$. In addition, note that (G.25) holds.

Controlling the deviations of $\widetilde{\boldsymbol{Z}}(\boldsymbol{q})$ It is proven by the same argument as in Lemma F.3 and Theorem G.1 that

$$\mathbb{E}\Big[\sup_{\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|\le\delta_n} \|\widetilde{\boldsymbol{Z}}(\boldsymbol{q}) - \widetilde{\boldsymbol{Z}}(\tilde{\boldsymbol{q}})\|_\infty \,\Big|\, \mathbf{X}_n\Big] = \mathbb{E}\Big[\sup_{\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|\le\delta_n} \|\widetilde{\boldsymbol{Z}}(\boldsymbol{q}) - \widetilde{\boldsymbol{Z}}(\tilde{\boldsymbol{q}})\|_\infty\Big] \lesssim \sqrt{\delta_n \log\frac{K}{\delta_n}}.$$

Choosing δ_n and conclusion Combining the bounds obtained above, we can write

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{\boldsymbol{q}\in\mathcal{Q}}\big\|\boldsymbol{Z}(\boldsymbol{q}) - \widetilde{\boldsymbol{Z}}(\boldsymbol{q})\big\|_\infty \,\Big|\, \mathbf{X}_n\Big] \\
&\quad\le \mathbb{E}\Big[\sup_{\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|\le\delta_n}\big\|\boldsymbol{Z}(\boldsymbol{q}) - \boldsymbol{Z}(\tilde{\boldsymbol{q}})\big\|_\infty \,\Big|\, \mathbf{X}_n\Big] + h^{-d/2}\mathbb{E}\big[\big\|(\bar{\boldsymbol{\Sigma}}^{\delta_n})^{1/2}\boldsymbol{\xi}_n - (\boldsymbol{\Sigma}^{\delta_n})^{1/2}\boldsymbol{\xi}_n\big\|_\infty \,\big|\, \mathbf{X}_n\big] \\
&\qquad + \mathbb{E}\Big[\sup_{\|\boldsymbol{q}-\tilde{\boldsymbol{q}}\|\le\delta_n}\big\|\widetilde{\boldsymbol{Z}}(\boldsymbol{q}) - \widetilde{\boldsymbol{Z}}(\tilde{\boldsymbol{q}})\big\|_\infty \,\Big|\, \mathbf{X}_n\Big] \\
&\quad\lesssim_{\mathbb{P}} \left(\sqrt{\delta_n} + \frac{{r_{\boldsymbol{\Sigma}}}^{1/2}}{\delta_n^{d_{\mathcal{Q}}/2}}\right)\sqrt{\log\frac{K}{\delta_n}}.
\end{aligned}$$

Choose the approximately optimal

$$\delta_n := (r_{\boldsymbol{\Sigma}})^{1/(d_{\mathcal{Q}}+1)},$$

giving

$$\mathbb{E}\Big[\sup_{\boldsymbol{q}\in\mathcal{Q}}\big\|\boldsymbol{Z}(\boldsymbol{q}) - \widetilde{\boldsymbol{Z}}(\boldsymbol{q})\big\|_\infty \,\Big|\, \mathbf{X}_n\Big] \lesssim_{\mathbb{P}} (r_{\boldsymbol{\Sigma}})^{1/(2d_{\mathcal{Q}}+2)}\sqrt{\log n}.$$

Constructing $\widetilde{\boldsymbol{Z}}^*(\cdot)$ By Theorem 8.17 (transfer) in [28], on our probability space there is a random element $\widetilde{\boldsymbol{Z}}^*(\cdot)$ with values in $\mathcal{C}(\mathcal{Q}, \mathbb{R}^K)$ such that

$$\big(\mathbf{X}_n, \boldsymbol{Z}(\cdot), \widetilde{\boldsymbol{Z}}(\cdot)\big) \overset{d}{=} \big(\mathbf{X}_n, \boldsymbol{Z}^*(\cdot), \widetilde{\boldsymbol{Z}}^*(\cdot)\big), \tag{G.40}$$

and $\widetilde{\boldsymbol{Z}}^*(\cdot)$ is conditionally on $(\mathbf{X}_n, \boldsymbol{Z}^*(\cdot))$ independent of $\boldsymbol{y}_n$. Since also $\boldsymbol{Z}^*(\cdot)$ conditionally on $\mathbf{X}_n$ independent of $\boldsymbol{y}_n$, we have by Theorem 8.12 (chain rule) in [28] that $(\boldsymbol{Z}^*(\cdot), \widetilde{\boldsymbol{Z}}^*(\cdot))$ is conditionally on $\mathbf{X}_n$ independent of $\boldsymbol{y}_n$; in particular, $\widetilde{\boldsymbol{Z}}^*(\cdot)$ is conditionally on $\mathbf{X}_n$ independent of $\boldsymbol{y}_n$. But by (G.40), $\widetilde{\boldsymbol{Z}}^*(\cdot)$ is independent of $\mathbf{X}_n$. Again by the chain rule, $\widetilde{\boldsymbol{Z}}^*(\cdot)$ is independent of $(\mathbf{X}_n, \boldsymbol{y}_n) = \mathbf{D}_n$.

H Examples

This section discusses in detail the four motivating examples introduced in the paper.

H.1 Quantile regression

The next result verifies that the quantile regression case under the same conditions on $f_{Y|X}$ as Condition S.2 in [3] is a special case of our setting.

Proposition H.1 (Verification of Assumption B.4 for quantile regression)**.** *Suppose Assumptions B.1 to B.3 hold with $\mathcal{Q} = [\varepsilon_0, 1-\varepsilon_0]$ for some $\varepsilon_0 \in (0, 1/2)$, the loss is given by $\rho(y, \eta; q) = (q - \mathbb{1}\{y < \eta\})(y - \eta)$, and $\mathbb{E}[|y_1|] < \infty$. Assume further that $\eta(\cdot)\colon \mathbb{R} \to \mathcal{E}$ is strictly monotonic and twice continuously differentiable with $\mathcal{E}$ an open connected subset of $\mathbb{R}$ containing the conditional q-quantile of $y_1|\boldsymbol{x}_1 = \boldsymbol{x}$, given by $\eta\big(\mu_0(\boldsymbol{x}, q)\big)$ for all $(\boldsymbol{x}, q)$; $y \mapsto F_{Y|X}(y|\boldsymbol{x})$ is twice continuously differentiable with first derivative $f_{Y|X}(y|\boldsymbol{x})$ (in particular, $\mathfrak{M}$ is the Lebesgue measure); $f_{Y|X}\big(\eta\big(\mu_0(\boldsymbol{x}, q)\big)|\boldsymbol{x}\big)$ is bounded away from zero uniformly over $q \in \mathcal{Q}$, $\boldsymbol{x} \in \mathcal{X}$, and the derivative of $y \mapsto f_{Y|X}(y|\boldsymbol{x})$ is continuous and bounded in absolute value from above uniformly over $y \in \mathcal{Y}_{\boldsymbol{x}}, \boldsymbol{x} \in \mathcal{X}$. Then Assumptions B.4, B.5 and B.8 and Eq. (F.15) are also true.*

Remark. Taking $\eta\big(\mu_0(\boldsymbol{x}, q)\big)$ to be the conditional q-quantile does not violate (A.1) by Lemma C.10.

Remark. In the setting of Proposition H.1, it is not necessary to assume that $\mu_0(\boldsymbol{x}, q)$ is Lipschitz in parameter (as we do in Assumption B.3(iv)). Since

$$\frac{\partial}{\partial q}\mu_0(\boldsymbol{x}, q) = \frac{1}{f_{Y|X}(\mu_0(\boldsymbol{x}, q)|\boldsymbol{x})},$$

the Lipschitz property follows from $f_{Y|X}(\mu_0(\boldsymbol{x}, q)|\boldsymbol{x})$ being bounded from below uniformly over $\boldsymbol{x} \in \mathcal{X}$, $q \in \mathcal{Q}$.

Proof. We will verify the assumptions one by one.

Verifying Assumption B.4(i) It is easy to see that the a. e. derivative $\eta \mapsto \psi(y, \eta; q) \equiv \mathbb{1}\{y - \eta < 0\} - q$ of $\eta \mapsto \rho(y, \eta; q)$ is Lebesgue integrable and satisfies

$$\int_a^b \psi(y, \eta; q)\,\mathrm{d}\eta = \rho(y, b; q) - \rho(y, a; q)$$

for any $[a, b] \subset \mathcal{E}$.

Verifying Assumption B.4(ii) Since $\eta(\mu_0(\cdot, q))$ is the conditional q-quantile, we have

$$\mathbb{E}[\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, q)); q) \,|\, \boldsymbol{x}_i] = \mathbb{E}[\mathbb{1}\{y_i < \eta(\mu_0(\boldsymbol{x}_i, q))\} - q \,|\, \boldsymbol{x}_i] = q - q = 0$$

and

$$\begin{aligned}\sigma_q^2(\boldsymbol{x}) &= \mathbb{E}[\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, q)); q)^2 \,|\, \boldsymbol{x}_i = \boldsymbol{x}] \\ &= \mathbb{E}[(\mathbb{1}\{y_i < \eta(\mu_0(\boldsymbol{x}_i, q))\} - q)^2 \,|\, \boldsymbol{x}_i = \boldsymbol{x}] = q - 2q^2 + q^2 = q(1-q)\end{aligned}$$

is constant in $\boldsymbol{x}$ (in particular continuous in $\boldsymbol{x}$) and bounded away from zero since both q and $1 - q$ are bounded away from zero. Since $q(1-q)$ is smooth, $\sigma_q^2(\boldsymbol{x})$ is Lipshitz in q. The family $\{\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, q)); q)\colon q \in \mathcal{Q}\}$ has a positive measurable envelope 1 which has uniformly bounded conditional moments of any order.

Verifying Assumption B.4(iii) Clearly, $\rho(y, \eta; q)$ is convex in η and the a. e. derivative in η is $\psi(y, \eta; q) \equiv \mathbb{1}\{y - \eta < 0\} - q$ is piecewise constant with only one jump (therefore piecewise Hölder with $\alpha = 1$). The link function $\eta(\cdot)$ is strictly monotonic and twice continuously differentiable by assumption.

Verifying Assumption B.4(iv) The conditional expectation

$$\Psi(x, \eta; q) = \mathbb{E}[\mathbb{1}\{y < \eta\} - q \mid \boldsymbol{x}_i = \boldsymbol{x}] = \int_{-\infty}^{\eta} f_{Y|X}(y|\boldsymbol{x})\,\mathrm{d}y - q$$

is twice continuously differentiable in η (the integral on the right is a Riemann integral, possibly improper) and its second derivative $f'_{Y|X}(\eta|\boldsymbol{x})$ is continuous and bounded in absolute value. By the mean value theorem, this means that $f_{Y|X}(\eta(\mu_0(\boldsymbol{x}, q))|\boldsymbol{x})$ being bounded away from zero implies $f_{Y|X}(\eta(\zeta)|\boldsymbol{x})$ is bounded away from zero for ζ sufficiently close to $\mu_0(\boldsymbol{x}, q)$. The bound on $|\Psi_1(\boldsymbol{x}, \eta(\zeta); q)|$ from above in such a neighborhood (and in fact everywhere) is automatic since $f_{Y|X}(y|\boldsymbol{x})$ is bounded from above (uniformly over $y \in \mathcal{Y}_{\boldsymbol{x}}$, $\boldsymbol{x} \in \mathcal{X}$).

Verifying Assumption B.5 This verification proceeds similarly to Lemmas 25–28 in [3].

The class of functions

$$\mathcal{W}_1 := \left\{(\boldsymbol{x}, y) \mapsto \mathbb{1}\{y < \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})\} \colon \boldsymbol{\beta} \in \mathbb{R}^K\right\}$$

is VC-subgraph with index $O(K)$ by Lemmas 2.6.15 and 2.6.18 in [41] (since $\eta(\cdot)$ is monotone). The class of functions

$$\mathcal{G}_2 = \left\{(\boldsymbol{x}, y) \mapsto \mathbb{1}\{y < \eta(\mu_0(\boldsymbol{x}, q))\} \colon q \in \mathcal{Q}\right\}$$

is VC-subgraph with index 2 since $\eta(\mu_0(\boldsymbol{x}, q))$ is increasing in q for any $\boldsymbol{x} \in \mathcal{X}$, giving that the class of sets $\{(\boldsymbol{x}, y) : y < \eta(\mu_0(\boldsymbol{x}, q))\}$ with $q \in \mathcal{Q}$ is linearly ordered by inclusion. The VC property of $\mathcal{W}_1$ with envelope 1 implies that it satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim K$. The VC property of $\mathcal{G}_2$ with envelope 1 implies that it satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim 1$. By Lemma C.4, for any fixed $r > 0$ the class

$$\mathcal{G}_1 = \left\{(\boldsymbol{x}, y) \mapsto \mathbb{1}\{y < \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})\} - \mathbb{1}\{y < \eta(\mu_0(\boldsymbol{x}, q))\} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \leq r, q \in \mathcal{Q}\right\}$$

with envelope 2 satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim K$ because it is a subclass of $\mathcal{W}_1 - \mathcal{G}_2$.

For a fixed vector space $\mathcal{B}$ of dimension $\dim \mathcal{B}$,

$$\mathcal{W}_{\mathcal{B}} := \{(\boldsymbol{x}, y) \mapsto \mathbb{1}\{y < \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})\} : \boldsymbol{\beta} \in \mathcal{B}\}$$

is VC with index $O(\dim \mathcal{B})$ by Lemmas 2.6.15 and 2.6.18 in [41] (since $\eta(\cdot)$ is monotone). Therefore, for any fixed $c > 0$, $\delta \in \Delta$, the class

$$\mathcal{W}_{2,\delta} := \left\{(\boldsymbol{x}, y) \mapsto \mathbb{1}\{y < \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})\}\mathbb{1}\{\boldsymbol{x} \in \mathcal{N}_{[c \log n]}(\delta)\} : \boldsymbol{\beta} \in \mathbb{R}^K\right\}$$

is VC-subgraph with index $O(\log^d n)$ by Lemma 2.6.18 in [41] because it is contained in the product of $\mathcal{W}_{\mathcal{B}_{2,\delta}}$ for some vector space $\mathcal{B}_{2,\delta}$ of dimension $\dim \mathcal{B}_{2,\delta} \lesssim \log^d n$ and a fixed function $\mathbb{1}\{\boldsymbol{x} \in \mathcal{N}_{[c \log n]}(\delta)\}$. This means $\mathcal{W}_{2,\delta}$ with envelope 1 satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim \log^d n$. Then the union of $O(h^{-d})$ such classes

$$\mathcal{W}_2 := \left\{(\boldsymbol{x}, y) \mapsto \mathbb{1}\{y < \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})\}\mathbb{1}\{\boldsymbol{x} \in \mathcal{N}_{[c \log n]}(\delta)\} : \boldsymbol{\beta} \in \mathbb{R}^K, \delta \in \Delta\right\}$$

with envelope 1 satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim \log^d n$, see (E.34). By Lemma C.4, the same is true of

$$\begin{aligned}\mathcal{G}_3 = \big\{&(\boldsymbol{x}, y) \mapsto \\ &\big[\mathbb{1}\{y < \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})\} - \mathbb{1}\{y < \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(q))\}\big]\mathbb{1}\{\boldsymbol{x} \in \mathcal{N}_{[c\log n]}(\delta)\} : \\ &\boldsymbol{\beta} \in \mathbb{R}^K, \delta \in \Delta, q \in \mathcal{Q}\big\}\end{aligned}$$

with envelope 2 because it is a subclass of $\mathcal{W}_2 - \mathcal{W}_2$.

The class of functions

$$\mathcal{G}_4 = \{\boldsymbol{x} \mapsto f_{Y|X}(\eta(\mu_0(\boldsymbol{x}, q))|\boldsymbol{x}) : q \in \mathcal{Q}\}$$

has a bounded envelope by the assumptions of the lemma. Moreover, $\mathcal{G}_4$ has the following property: for any $q_1, q_2 \in \mathcal{Q}$ we have for some $\xi_{\boldsymbol{x},q_1,q_2}$ between $\eta(\mu_0(\boldsymbol{x}, q_1))$ and $\eta(\mu_0(\boldsymbol{x}, q_2))$

$$\begin{aligned}&\big|f_{Y|X}(\eta(\mu_0(\boldsymbol{x}, q_1))|\boldsymbol{x}) - f_{Y|X}(\eta(\mu_0(\boldsymbol{x}, q_2))|\boldsymbol{x})\big| \\ &\quad = \big|f'_{Y|X}(\xi_{\boldsymbol{x},q_1,q_2})\big| \cdot |\eta(\mu_0(\boldsymbol{x}, q_1)) - \eta(\mu_0(\boldsymbol{x}, q_2))| \\ &\quad \lesssim \big|\eta(\mu_0(\boldsymbol{x}, q_1)) - \eta(\mu_0(\boldsymbol{x}, q_2))\big| && \text{since } f'_{Y|X} \text{ is uniformly bounded} \\ &\quad \lesssim \big|\mu_0(\boldsymbol{x}, q_1) - \mu_0(\boldsymbol{x}, q_2)\big| && \text{since } \eta(\cdot) \text{ is Lipschitz} \\ &\quad \lesssim \big|q_1 - q_2\big| && \text{since } \mu_0(\boldsymbol{x}, q) \text{ is Lipschitz in } q.\end{aligned}$$

with constants in $\lesssim$ not depending on $\boldsymbol{x}, q_1, q_2$ or n (this is also proven in Lemma 20 in [3]). Since $\mathcal{Q}$ is a fixed one-dimensional segment, this implies that $\mathcal{G}_4$ satisfies the uniform entropy bound (A.3) with $A, V \lesssim 1$.

For a fixed l, the class of functions

$$\{(\boldsymbol{x}, y) \mapsto p_l(\boldsymbol{x})\mathbb{1}\{y < \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})\} : \boldsymbol{\beta} \in \mathbb{R}^K\}$$

is VC-subgraph with index $O(1)$ because it is contained in the product of $\mathcal{W}_{\mathcal{B}_{5,l}}$ for some vector space $\mathcal{B}_{5,l}$ of dimension $\dim \mathcal{B}_{5,l} \lesssim 1$, and a fixed function $p_l(\boldsymbol{x})$. Then this class with envelope $O(1)$ satisfies the uniform entropy bound (A.3) with $A, V \lesssim 1$. Since, as we have shown above, the same is true of $\mathcal{G}_2$, by Lemma C.4, it is also true of

$$\begin{aligned}\mathcal{G}_5 = \big\{&(\boldsymbol{x}, y) \mapsto \\ &p_l(\boldsymbol{x})\big[\mathbb{1}\{y < \eta(\mu_0(\boldsymbol{x}, q))\} - \mathbb{1}\{y < \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(q))\}\big] : q \in \mathcal{Q}\big\}.\end{aligned}$$

Verifying Assumption B.8 The functions in the class have the form

$$\begin{aligned}&\rho(y, \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta})); q)\mathbb{1}(\boldsymbol{x} \in \mathcal{N}_{[c\log n]}(\delta)) \\ &\quad - \rho(y, \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta} - \boldsymbol{v})); q)\mathbb{1}(\boldsymbol{x} \in \mathcal{N}_{[c\log n]}(\delta)) \\ &\quad - \big[\eta\big(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta})\big) - \eta\big(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta} - \boldsymbol{v})\big)\big]\psi(y, \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(q)); q) \\ &\quad \times \mathbb{1}(\boldsymbol{x} \in \mathcal{N}_{[c\log n]}(\delta)) \\ &=: T_1 + T_2 + T_3 + T_4,\end{aligned}$$

where

$$
\begin{aligned}
T_1 &:= y\big[\mathbb{1}\big\{y < \eta\big(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta}-\boldsymbol{v})\big)\big\} - \mathbb{1}\big\{y < \eta\big(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta})\big)\big\}\big] \\
&\quad \times \mathbb{1}\{\boldsymbol{x} \in \mathcal{N}_{[c\log n]}(\delta)\}, \\
T_2 &:= \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta}))\mathbb{1}\big\{y < \eta\big(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta})\big)\big\}\mathbb{1}\{\boldsymbol{x} \in \mathcal{N}_{[c\log n]}(\delta)\}, \\
T_3 &:= -\eta\big(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta}-\boldsymbol{v})\big)\mathbb{1}\big\{y < \eta\big(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta}-\boldsymbol{v})\big)\big\}\mathbb{1}\{\boldsymbol{x} \in \mathcal{N}_{[c\log n]}(\delta)\}, \\
T_4 &:= -\mathbb{1}\big\{y < \eta\big(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(q)\big)\big\}\big[\eta\big(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta})\big) - \eta\big(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta}-\boldsymbol{v})\big)\big] \\
&\quad \times \mathbb{1}\{\boldsymbol{x} \in \mathcal{N}_{[c\log n]}(\delta)\}.
\end{aligned}
$$

Note that for T_1 to be nonzero, y has to lie in a fixed interval (not depending on n), say $[-\tilde{R}, \tilde{R}]$. The class of functions

$$
\{(\boldsymbol{x}, y) \mapsto y\mathbb{1}\{|y| \le \tilde{R}\}\mathbb{1}\{y < \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})\} : \boldsymbol{\beta} \in \mathcal{B}\},
$$

where $\mathcal{B}$ is any linear subspace of $\mathbb{R}^K$, is VC-subgraph with index $O(\dim \mathcal{B})$ by Lemmas 2.6.15 and 2.6.18 in [41] (since $\eta(\cdot)$ is monotone and $y \mapsto y\mathbb{1}\{|y| \le \tilde{R}\}$ is one fixed function). The class $\{(\boldsymbol{x}, y) \mapsto T_1, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, \|\boldsymbol{v}\|_\infty \le \varepsilon_n, q \in \mathcal{Q}\}$ with δ fixed is a subclass of the difference of two such classes, so by Lemma C.4 this class with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim \log^d n$. Therefore, the union of $O(h^{-d})$ such classes

$$
\{(\boldsymbol{x}, y) \mapsto T_1, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, \|\boldsymbol{v}\|_\infty \le \varepsilon_n, \delta \in \Delta, q \in \mathcal{Q}\}
$$

with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim \log^d n$, see (E.34).

Similarly, the classes

$$
\begin{aligned}
&\{(\boldsymbol{x}, y) \mapsto T_2, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, \|\boldsymbol{v}\|_\infty \le \varepsilon_n, \delta \in \Delta, q \in \mathcal{Q}\}, \\
&\{(\boldsymbol{x}, y) \mapsto T_3, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, \|\boldsymbol{v}\|_\infty \le \varepsilon_n, \delta \in \Delta, q \in \mathcal{Q}\}, \\
&\{(\boldsymbol{x}, y) \mapsto T_4, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, \|\boldsymbol{v}\|_\infty \le \varepsilon_n, \delta \in \Delta, q \in \mathcal{Q}\}
\end{aligned}
$$

with large enough constant envelopes also satisfy the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim \log^d n$. We used that $\varepsilon_n \to 0$ giving that $\boldsymbol{v}$ is bounded in ∞-norm (like $\boldsymbol{\beta}$). Applying Lemma C.4 one more time, we have that there exist some constants $C_{17} \ge e$, $C_{18} \ge 1$ and $C_{19} > 0$ such that

$$
\sup_{\mathbb{Q}} N(\mathcal{G}, \|\cdot\|_{\mathbb{Q},2}, \varepsilon C_{19}) \le \left(\frac{C_{17}}{\varepsilon}\right)^{C_{18}\log^d n} \tag{H.1}
$$

for all $0 < \varepsilon \le 1$, where the supremum is taken over all finite discrete probability measures $\mathbb{Q}$ and $\mathcal{G}$ is the class defined in Assumption B.8. Note that the integral representation of $\mathcal{G}$ makes it clear that this class not only has a large enough constant envelope, but is also bounded by $C_{20}\varepsilon_n$, where C_{20} is a large enough constant.

For large enough n we can replace ε with $C_{20}\varepsilon\varepsilon_n/C_{19}$ in (H.1), giving

$$
\sup_{\mathbb{Q}} N(\mathcal{G}, \|\cdot\|_{\mathbb{Q},2}, C_{20}\varepsilon\varepsilon_n) \le \left(\frac{C_{17}C_{19}}{C_{20}\varepsilon\varepsilon_n}\right)^{C_{18}\log^d n}
$$

for all $0 < \varepsilon \le 1$. For large enough n, $C_{17}C_{19}/(C_{20}\varepsilon_n) \ge e$. The verification is complete.

Verifying (F.15) In this case, $\psi(y,\eta;q) = \mathbb{1}\{y < \eta\} - q$ and $\eta(\mu_0(\boldsymbol{x}_i, q))$ is the q-quantile of y_i conditional on $\boldsymbol{x}_i$. Without loss of generality, we will assume that $\eta(\cdot)$ is strictly increasing and $q \le \tilde{q}$ (the other cases are symmetric).

$$
\begin{aligned}
&\mathbb{E}\big[\big|\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, q)); q)\eta^{(1)}(\mu_0(\boldsymbol{x}_i, q)) - \psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, \tilde{q})); \tilde{q})\eta^{(1)}(\mu_0(\boldsymbol{x}_i, \tilde{q}))\big|^2 \,\big|\, \boldsymbol{x}_i\big] \\
&\quad = \mathbb{E}\big[\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, q)); q)^2 \,\big|\, \boldsymbol{x}_i\big]\eta^{(1)}(\mu_0(\boldsymbol{x}_i, q))^2 \\
&\qquad + \mathbb{E}\big[\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, \tilde{q})); \tilde{q})^2 \,\big|\, \boldsymbol{x}_i\big]\eta^{(1)}(\mu_0(\boldsymbol{x}_i, \tilde{q}))^2 \\
&\qquad - 2\mathbb{E}\big[\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, q)); q)\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i, \tilde{q})); \tilde{q}) \,\big|\, \boldsymbol{x}_i\big]\eta^{(1)}(\mu_0(\boldsymbol{x}_i, q))\eta^{(1)}(\mu_0(\boldsymbol{x}_i, \tilde{q})) \\
&\quad = (q - q^2)\eta^{(1)}(\mu_0(\boldsymbol{x}_i, q))^2 + (\tilde{q} - \tilde{q}^2)\eta^{(1)}(\mu_0(\boldsymbol{x}_i, \tilde{q}))^2 \\
&\qquad - 2(q - q\tilde{q})\eta^{(1)}(\mu_0(\boldsymbol{x}_i, q))\eta^{(1)}(\mu_0(\boldsymbol{x}_i, \tilde{q})) \\
&\quad = q\big|\eta^{(1)}(\mu_0(\boldsymbol{x}_i, q)) - \eta^{(1)}(\mu_0(\boldsymbol{x}_i, \tilde{q}))\big|^2 + (\tilde{q} - q)\eta^{(1)}(\mu_0(\boldsymbol{x}_i, \tilde{q}))^2 \\
&\qquad - \big|q\eta^{(1)}(\mu_0(\boldsymbol{x}_i, q)) - \tilde{q}\eta^{(1)}(\mu_0(\boldsymbol{x}_i, \tilde{q}))\big|^2 \\
&\quad \le q\big|\eta^{(1)}(\mu_0(\boldsymbol{x}_i, q)) - \eta^{(1)}(\mu_0(\boldsymbol{x}_i, \tilde{q}))\big|^2 + (\tilde{q} - q)\eta^{(1)}(\mu_0(\boldsymbol{x}_i, \tilde{q}))^2 \\
&\quad \overset{\text{(a)}}{\lesssim} q(\tilde{q} - q)^2 + \tilde{q} - q \lesssim \tilde{q} - q,
\end{aligned}
$$

where in (a) we used that $\eta^{(1)}(\cdot)$ on a fixed compact is Lipschitz and $\mu_0(\boldsymbol{x}, q)$ is Lipschitz in q uniformly over $\boldsymbol{x}$, as well as boundedness of $\mu_0(\boldsymbol{x}, q)$ uniformly over q and $\boldsymbol{x}$.

This concludes the proof of Proposition H.1. □

Proposition H.2 (Verification of the conditions of Lemma D.3)**.** *Suppose all conditions of Proposition H.1 hold. In addition, suppose there is a positive constant C_{21} such that we have $\inf f_{Y|X}(y|\boldsymbol{x}) > C_{21}$, where the infimum is over $\boldsymbol{x} \in \mathcal{X}$, $q \in \mathcal{Q}$, $\|\boldsymbol{\beta}\|_\infty \le R$ for R described in Lemma D.3, y between $\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})$ and $\eta(\mu_0(\boldsymbol{x}, q))$. Then conditions in Conditions (v) and (vi) of Lemma D.3 also hold.*

Proof. We only need to verify Lemma D.3(vi) since Lemma D.3(v) is directly assumed in this lemma ($\Psi_1(\boldsymbol{x}, \eta; q) = f_{Y|X}(\eta|\boldsymbol{x})$ in this case).

In this verification, we will use $\theta_1 := \boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}$, $\theta_2 := \boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(q)$ to simplify notations. Rewrite

$$
\begin{aligned}
&\rho(y, \eta(\theta_1); q) - \rho(y, \eta(\theta_2); q) \\
&\quad = y[\mathbb{1}\{y < \eta(\theta_2)\} - \mathbb{1}\{y < \eta(\theta_1)\}] + q[\eta(\theta_1) - \eta(\theta_2)] \\
&\qquad + \eta(\theta_1)\mathbb{1}\{y < \eta(\theta_1)\} - \eta(\theta_2)\mathbb{1}\{y < \eta(\theta_2)\}.
\end{aligned}
$$

By the same argument as in the proof of Proposition H.1, the class

$$
\{(\boldsymbol{x}, y) \mapsto y[\mathbb{1}\{y < \eta(\theta_2)\} - \mathbb{1}\{y < \eta(\theta_1)\}] : \|\boldsymbol{\beta}\|_\infty \le R, q \in \mathcal{Q}\}
$$

with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim K$.

The class $\{(\boldsymbol{x}, y) \mapsto q : q \in \mathcal{Q}\}$ is of course VC with a constant index (as a subclass of the class of constant functions), and the class $\{(\boldsymbol{x}, y) \mapsto \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathbb{R}^K\}$ is VC with index $O(K)$ because the space of functions $\boldsymbol{p}(\boldsymbol{x})\boldsymbol{\beta}$ is a linear space with $O(K)$ dimension, and $\eta(\cdot)$ is monotone. Applying Lemma C.4, we see that the class $\{(\boldsymbol{x}, y) \mapsto q[\eta(\theta_1) - \eta(\theta_2)] : \|\boldsymbol{\beta}\|_\infty \le R, q \in \mathcal{Q}\}$ with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim K$.

The class $\{(\boldsymbol{x}, y) \mapsto \mathbb{1}\{y < \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})\} : \boldsymbol{\beta} \in \mathbb{R}^K\}$ is VC with index $O(K)$ because the space of functions $\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}$ is a linear space with $O(K)$ dimension, and $\eta(\cdot)$ is monotone. Again applying

Lemma C.4, we see that the class $\{(\boldsymbol{x}, y) \mapsto \eta(\theta_1)\mathbb{1}\{y < \eta(\theta_1)\} : \|\boldsymbol{\beta}\|_\infty \le R\}$ with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim K$. The same is true of its subclass $\{(\boldsymbol{x}, y) \mapsto \eta(\theta_2)\mathbb{1}\{y < \eta(\theta_2)\} : q \in \mathcal{Q}\}$.

It is left to apply Lemma C.4 again, concluding that the class described in Lemma D.3(vi) with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim K$. □

H.2 Distribution regression

Proposition H.3 (Verification of assumptions for distribution regression)**.** *Let $\mathcal{Q} = [-A, A]$ for some $A > 0$. Suppose Assumption B.3 holds with the loss function $\rho(y, \eta; q) = (\mathbb{1}\{y \le q\} - \eta)^2$, the link function $\eta(\cdot)\colon \mathbb{R} \to (0, 1)$ is strictly monotonic and twice continuously differentiable, $q \mapsto F_{Y|X}(q|\boldsymbol{x})$ is continuously differentiable with derivative $f_{Y|X}(q|\boldsymbol{x})$ (in particular, $\mathfrak{M}$ is Lebesgue measure), $\boldsymbol{x} \mapsto F_{Y|X}(q|\boldsymbol{x})$ is a continuous function, and $F_{Y|X}(q|\boldsymbol{x}) = \eta(\mu_0(\boldsymbol{x}, q))$ lies in a compact subset of $(0, 1)$ for all $q \in \mathcal{Q}, \boldsymbol{x} \in \mathcal{X}$ (this subset does not depend on q or $\boldsymbol{x}$). Then Assumptions B.4, B.5 and B.8, Eq. (F.15), and Conditions (v) and (vi) of Lemma D.3 are also true.*

Remark. Taking $\eta(\mu_0(\boldsymbol{x}, q))$ to be the conditional distribution function does not violate Eq. (A.1) by the following standard argument. For any Borel function $\mu(\cdot)\colon \mathcal{X} \to \mathbb{R}$ one has

$$
\begin{aligned}
&\mathbb{E}\bigl[\bigl(\mathbb{1}\{y_1 \le q\} - \eta(\mu(\boldsymbol{x}_1))\bigr)^2\bigr] \\
&\quad = \mathbb{E}\bigl[\bigl(\mathbb{1}\{y_1 \le q\} - F_{Y|X}(q|\boldsymbol{x}_1) + F_{Y|X}(q|\boldsymbol{x}_1) - \eta(\mu(\boldsymbol{x}_1))\bigr)^2\bigr] \\
&\quad = \mathbb{E}\bigl[\bigl(\mathbb{1}\{y_1 \le q\} - F_{Y|X}(q|\boldsymbol{x}_1)\bigr)^2\bigr] \\
&\qquad + 2\mathbb{E}\bigl[\bigl(\mathbb{1}\{y_1 \le q\} - F_{Y|X}(q|\boldsymbol{x}_1)\bigr)\bigl(F_{Y|X}(q|\boldsymbol{x}_1) - \eta(\mu(\boldsymbol{x}_1))\bigr)\bigr] \\
&\qquad + \mathbb{E}\bigl[\bigl(F_{Y|X}(q|\boldsymbol{x}_1) - \eta(\mu(\boldsymbol{x}_1))\bigr)^2\bigr].
\end{aligned}
$$

Since the cross term is zero (proven by conditioning on $\boldsymbol{x}_1$), this means

$$
\begin{aligned}
&\mathbb{E}\bigl[\bigl(\mathbb{1}\{y_1 \le q\} - \eta(\mu(\boldsymbol{x}_1))\bigr)^2\bigr] \\
&\quad = \mathbb{E}\bigl[\bigl(\mathbb{1}\{y_1 \le q\} - F_{Y|X}(q|\boldsymbol{x}_1)\bigr)^2\bigr] + \mathbb{E}\bigl[\bigl(F_{Y|X}(q|\boldsymbol{x}_1) - \eta(\mu(\boldsymbol{x}_1))\bigr)^2\bigr] \\
&\quad \ge \mathbb{E}\bigl[\bigl(\mathbb{1}\{y_1 \le q\} - F_{Y|X}(q|\boldsymbol{x}_1)\bigr)^2\bigr].
\end{aligned}
$$

Pointwise in $q \in \mathcal{Q}$, equality holds if and only if $\eta(\mu(\boldsymbol{x}_1)) = F_{Y|X}(q|\boldsymbol{x}_1)$ almost surely.

Remark H.4. In this case,

$$
\mathbb{E}\bigl[\boldsymbol{Z}(q)\boldsymbol{Z}(\tilde{q})^{\mathsf{T}} \bigm| \mathbf{X}_n\bigr] = h^{-d}\mathbb{E}_n\bigl[S_{q,\tilde{q}}(\boldsymbol{x})\eta^{(1)}(\mu_0(\boldsymbol{x}_i, q))\eta^{(1)}(\mu_0(\boldsymbol{x}_i, \tilde{q}))\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\bigr]
$$

with

$$
S_{q,\tilde{q}}(\boldsymbol{x}) = 4F_{Y|X}(q \wedge \tilde{q}|\boldsymbol{x}_i)\bigl(1 - F_{Y|X}(q \vee \tilde{q}|\boldsymbol{x}_i)\bigr).
$$

This covariance structure is not known, but it can be estimated by

$$
h^{-d}\mathbb{E}_n\bigl[\widehat{S}_{q,\tilde{q}}(\boldsymbol{x})\eta^{(1)}(\widehat{\mu}(\boldsymbol{x}_i, q))\eta^{(1)}(\widehat{\mu}(\boldsymbol{x}_i, \tilde{q}))\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^{\mathsf{T}}\bigr]. \tag{H.2}
$$

with

$$
\widehat{S}_{q,\tilde{q}}(\boldsymbol{x}) = 4\eta(\widehat{\mu}(\boldsymbol{x}_i, q \wedge \tilde{q}))\bigl(1 - \eta(\widehat{\mu}(\boldsymbol{x}_i, q \vee \tilde{q}))\bigr).
$$

Proof. We will verify the assumptions one by one.

Verifying Assumption B.4(i) It is easy to see that the a. e. derivative $\eta \mapsto \psi(y, \eta; q) \equiv 2(\eta - \mathbb{1}\{y \leq q\})$ of $\eta \mapsto \rho(y, \eta; q)$ is Lebesgue integrable and satisfies

$$\int_a^b \psi(y, \eta; q)\,\mathrm{d}\eta = \rho(y, b; q) - \rho(y, a; q)$$

for any $[a, b] \subset (0, 1)$.

Verifying Assumption B.4(ii) The first-order optimality condition

$$\mathbb{E}\big[2(\eta(\mu_0(\boldsymbol{x}_i, q)) - \mathbb{1}\{y_i \leq q\}) \,\big|\, \boldsymbol{x}_i\big] = 2\big(F_{Y|X}(q|\boldsymbol{x}_i) - \mathbb{E}[\mathbb{1}\{y_i \leq q\} \,|\, \boldsymbol{x}_i]\big) = 0$$

indeed holds. The conditional variance

$$\sigma_q^2(\boldsymbol{x}) := 4\mathbb{E}[(\eta(\mu_0(\boldsymbol{x}_i, q)) - \mathbb{1}\{y_i \leq q\})^2 \,|\, \boldsymbol{x}_i = \boldsymbol{x}] = 4F_{Y|X}(q|\boldsymbol{x})(1 - F_{Y|X}(q|\boldsymbol{x}))$$

is continuous and bounded away from zero by the assumptions ($F_{Y|X}(q|\boldsymbol{x})$ cannot achieve 0 or 1).

The family $\{2(F_{Y|X}(q|\boldsymbol{x}) - \mathbb{1}\{y \leq q\}) : q \in \mathcal{Q}\}$ is bounded in absolute value by $\overline{\psi}(\boldsymbol{x}, y) \equiv 2$.

Verifying Assumption B.4(iii) The loss function $\rho(y, \eta; q)$ is infinitely smooth with respect to η. Its first derivative is $\psi(y, \eta; q) = 2(\eta - \mathbb{1}\{y \leq q\})$. Since the derivative of $\eta(\cdot)$ is bounded on a compact interval, the function $\psi(y, \eta(\theta); q)$ is Lipschitz in θ on a compact interval.

Verifying Assumption B.4(iv) The conditional expectation

$$\Psi(\boldsymbol{x}, \eta; q) := \mathbb{E}[2(\eta - \mathbb{1}\{y_i \leq q\}) \,|\, \boldsymbol{x}_i = \boldsymbol{x}] = 2\eta - 2F_{Y|X}(q|\boldsymbol{x})$$

is linear, and in particular infinitely smooth, in η. Its first partial derivative

$$\Psi_1(\boldsymbol{x}, \eta; q) = \frac{\partial}{\partial \eta}\Psi(\boldsymbol{x}, \eta; q) = 2$$

is a nonzero constant, so it is bounded and bounded away from zero everywhere. The second partial derivative is zero, and so it is also bounded.

Verifying Assumption B.5 The class of functions

$$\mathcal{G}_{11} := \{(\boldsymbol{x}, y) \mapsto F_{Y|X}(q|\boldsymbol{x}) : q \in \mathcal{Q}\}$$

with a constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim 1$ because it is VC-subgraph with index 1 since the subgraphs are linearly ordered by inclusion (by monotonicity in q).

The class

$$\mathcal{G}_{12} := \{(\boldsymbol{x}, y) \mapsto \boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^K\}$$

is VC with index not exceeding $K + 2$ by Lemma 2.6.15 in [41]. Since in a fixed bounded interval $\eta(\cdot)$ is Lipschitz, the class

$$\mathcal{G}_{13} := \{(\boldsymbol{x}, y) \mapsto \eta(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}) : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \leq r\}$$

with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim K$. By Lemma C.4, the class

$$\mathcal{G}_1 := \big\{(\boldsymbol{x}, y) \mapsto 2\big(\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}) - F_{Y|X}(q|\boldsymbol{x})\big) : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, q \in \mathcal{Q}\big\},$$

with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim K$.

The class of sets $\{(\boldsymbol{x}, y) : y \le q\}$ with $q \in \mathcal{Q}$ is linearly ordered by inclusion, so it is VC with a constant index and so is the class of functions

$$\mathcal{G}_{21} := \{(\boldsymbol{x}, y) \mapsto \mathbb{1}\{y \le q\} : q \in \mathcal{Q}\},$$

giving by Lemma C.4 that

$$\mathcal{G}_2 := \{(\boldsymbol{x}, y) \mapsto 2(F_{Y|X}(q|\boldsymbol{x}) - \mathbb{1}\{y \le q\}) : q \in \mathcal{Q}\},$$

which is a subclass of $2(\mathcal{G}_{11} - \mathcal{G}_{21})$, satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim 1$.

For a fixed vector space $\mathcal{B}$ of dimension $\dim \mathcal{B}$,

$$\mathcal{W}_\mathcal{B} := \{(\boldsymbol{x}, y) \mapsto \boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathcal{B}\}$$

is VC-subgraph with index $O(\dim \mathcal{B})$ by Lemmas 2.6.15 and 2.6.18 in [41]. Therefore, again using that $\eta(\cdot)$ is Lipschitz in a compact interval, by Lemma C.4 we have that for any fixed $c > 0$, $\delta \in \Delta$, the class

$$\mathcal{W}_{3,\delta} := \{(\boldsymbol{x}, y) \mapsto \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})\mathbb{1}\{\boldsymbol{x} \in \mathcal{N}_{[c \log n]}(\delta)\} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, q \in \mathcal{Q}\}$$

with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim \log^d n$. Then the union of $O(h^{-d})$ such classes

$$\mathcal{W}_3 := \{(\boldsymbol{x}, y) \mapsto \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})\mathbb{1}\{\boldsymbol{x} \in \mathcal{N}_{[c \log n]}(\delta)\} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, q \in \mathcal{Q}, \delta \in \Delta\}$$

satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim \log^d n$ (see (E.34)). The same is true of

$$\begin{aligned}\mathcal{G}_3 := \big\{&(\boldsymbol{x}, y) \mapsto 2\big(\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}) - \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(q))\big)\mathbb{1}\{\boldsymbol{x} \in \mathcal{N}_{[c \log n]}(\delta)\} :\\ &\|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, q \in \mathcal{Q}, \delta \in \Delta\big\}\end{aligned}$$

because it is a subclass of $2\mathcal{W}_3 - 2\mathcal{W}_3$.

The class

$$\mathcal{G}_4 := \{\boldsymbol{x} \mapsto 2\}$$

consists of just one bounded function, so clearly it satisfies the uniform entropy bound (A.3) with envelope 2, $A \lesssim 1$, $V \lesssim 1$.

Finally, for any fixed $l \in \{1, \dots, K\}$ the class

$$\mathcal{G}_{51} := \{(\boldsymbol{x}, y) \mapsto p_l(\boldsymbol{x})\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(q)) : q \in \mathcal{Q}\}$$

satisfies the uniform entropy bound (A.3) with a large enough constant envelope, $A \lesssim 1$ and $V \lesssim 1$ because $\eta(\cdot)$ is Lipschitz and $\mathcal{G}_{51}$ is contained in a fixed function multiplied by $\eta(\mathcal{W}_\mathcal{B})$ for a linear space $\mathcal{B}$ of a constant dimension, and by Lemma C.4

$$\mathcal{G}_5 := \{(\boldsymbol{x}, y) \mapsto 2p_l(\boldsymbol{x})(F_{Y|X}(q|\boldsymbol{x}) - \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(q))) : q \in \mathcal{Q}\}$$

with a constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim 1$.

Verifying Assumption B.8 Assumption B.8 holds because the class of functions

$$\begin{aligned}
&\big\{(\boldsymbol{x}, y) \mapsto \big[\eta\big(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta})\big) - \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta}-\boldsymbol{v}))\big] \\
&\quad \times \big[\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta})) + \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta}-\boldsymbol{v})) - 2\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(q))\big] \\
&\quad \times \mathbb{1}(\boldsymbol{x} \in \mathcal{N}_{[c\log n]}(\delta)) : \\
&\quad \|\boldsymbol{\beta}-\boldsymbol{\beta}_0(q)\|_\infty \le r, \|\boldsymbol{v}\|_\infty \le \varepsilon_n, \delta \in \Delta, q \in \mathcal{Q}\big\}
\end{aligned}$$

is contained in the product of two classes

$$\begin{aligned}
\mathcal{V}_1 := \big\{&(\boldsymbol{x}, y) \mapsto \big[\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}) - \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}-\boldsymbol{v}))\big]\mathbb{1}(\boldsymbol{x} \in \mathcal{N}_{[c\log n]}(\delta)) : \\
&\|\boldsymbol{\beta}\|_\infty \le \tilde{r}, \|\boldsymbol{v}\|_\infty \le \varepsilon_n, \delta \in \Delta, q \in \mathcal{Q}\big\},
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{V}_2 := \big\{&(\boldsymbol{x}, y) \mapsto \big[\eta\big(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta})\big) + \eta\big(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0(q)+\boldsymbol{\beta}-\boldsymbol{v})\big) - 2\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(q))\big] \\
&\times \mathbb{1}(\boldsymbol{x} \in \mathcal{N}_{[c\log n]}(\delta)) : \|\boldsymbol{\beta}-\boldsymbol{\beta}_0(q)\|_\infty \le r, \|\boldsymbol{v}\|_\infty \le \varepsilon_n, \delta \in \Delta, q \in \mathcal{Q}\big\}
\end{aligned}$$

for some fixed $\tilde{r} > 0$. Class $\mathcal{V}_1$ with envelope ε_n multiplied by a large enough constant (since η is Lipschitz) satisfies the uniform entropy bound (A.3) with $A \lesssim 1/\varepsilon_n$ and $V \lesssim \log^d n$ (this can be shown by further breaking down $\mathcal{V}_1$ into classes $\{\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})\}$ and $\{\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}-\boldsymbol{v}))\}$ with constant envelopes, using Lemma C.4 and then replacing ε in the uniform entropy bound by $\varepsilon \cdot \varepsilon_n$). Class $\mathcal{V}_2$ with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim \log^d n$ because it is true for each of the three additive terms it can be broken down into. We omit the details since they are the same as in the verification of Assumption B.5.

Verifying Condition (v) in Lemma D.3 This condition holds trivially because $\Psi_1(\cdot, \cdot; q)$ is a positive constant.

Verifying Condition (vi) in Lemma D.3 The class of functions described in this condition is

$$\begin{aligned}
&\big\{(\boldsymbol{x}, y) \mapsto \big(\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(q)) - \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})\big) \cdot \big(2\mathbb{1}\{y \le q\} - \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}) - \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0(q))\big) : \\
&\|\boldsymbol{\beta}\|_\infty \le R, q \in \mathcal{Q}\big\}.
\end{aligned}$$

The assertion follows by Lemma C.4 since

$$\{(\boldsymbol{x}, y) \mapsto \mathbb{1}\{y \le q\} : q \in \mathcal{Q}\}$$

is a VC-subgraph class with a constant index and

$$\{(\boldsymbol{x}, y) \mapsto \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}) : \|\boldsymbol{\beta}\|_\infty \le R\}$$

with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim K$ because $\eta(\cdot)$ is Lipschitz.

Verifying (F.15) Without loss of generality, assume $q \le \tilde q$ (the other case is symmetric).

$$\begin{aligned}
&\mathbb{E}\big[|\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,q));q)\eta^{(1)}(\mu_0(\boldsymbol{x}_i,q)) - \psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\tilde q));\tilde q)\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\tilde q))|^2 \,\big|\, \boldsymbol{x}_i\big] \\
&\quad= \mathbb{E}\big[\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,q));q)^2 \,\big|\, \boldsymbol{x}_i\big]\eta^{(1)}(\mu_0(\boldsymbol{x}_i,q))^2 \\
&\qquad+ \mathbb{E}\big[\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\tilde q));\tilde q)^2 \,\big|\, \boldsymbol{x}_i\big]\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\tilde q))^2 \\
&\qquad- 2\mathbb{E}\big[\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,q));q)\psi(y_i,\eta(\mu_0(\boldsymbol{x}_i,\tilde q));\tilde q) \,\big|\, \boldsymbol{x}_i\big]\eta^{(1)}(\mu_0(\boldsymbol{x}_i,q))\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\tilde q)) \\
&\quad= 4\big[F_{Y|X}(q|\boldsymbol{x}_i) - F_{Y|X}(q|\boldsymbol{x}_i)^2\big]\eta^{(1)}(\mu_0(\boldsymbol{x}_i,q))^2 \\
&\qquad+ 4\big[F_{Y|X}(\tilde q|\boldsymbol{x}_i) - F_{Y|X}(\tilde q|\boldsymbol{x}_i)^2\big]\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\tilde q))^2 \\
&\qquad- 8\big[F_{Y|X}(q|\boldsymbol{x}_i) - F_{Y|X}(q|\boldsymbol{x}_i)F_{Y|X}(\tilde q|\boldsymbol{x}_i)\big]\eta^{(1)}(\mu_0(\boldsymbol{x}_i,q))\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\tilde q)) \\
&\quad= 4F_{Y|X}(q|\boldsymbol{x}_i)\big|\eta^{(1)}(\mu_0(\boldsymbol{x}_i,q)) - \eta^{(1)}(\mu_0(\boldsymbol{x}_i,\tilde q))\big|^2 \\
&\qquad+ 4\big[F_{Y|X}(\tilde q|\boldsymbol{x}_i) - F_{Y|X}(q|\boldsymbol{x}_i)\big]\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\tilde q))^2 \\
&\qquad- 4\big|F_{Y|X}(q|\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i,q)) - F_{Y|X}(\tilde q|\boldsymbol{x}_i)\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\tilde q))\big|^2 \\
&\quad\le 4F_{Y|X}(q|\boldsymbol{x}_i)\big|\eta^{(1)}(\mu_0(\boldsymbol{x}_i,q)) - \eta^{(1)}(\mu_0(\boldsymbol{x}_i,\tilde q))\big|^2 \\
&\qquad+ 4\big[F_{Y|X}(\tilde q|\boldsymbol{x}_i) - F_{Y|X}(q|\boldsymbol{x}_i)\big]\eta^{(1)}(\mu_0(\boldsymbol{x}_i,\tilde q))^2 \\
&\quad\overset{(a)}{\lesssim} (\tilde q - q)^2 + (\tilde q - q) \lesssim \tilde q - q,
\end{aligned}$$

where in (a) we used that $\eta^{(1)}(\cdot)$ and $\eta(\cdot)$ on a fixed compact are Lipschitz and $\mu_0(\boldsymbol{x},q)$ is Lipschitz in q uniformly over $\boldsymbol{x}$ (therefore, $\eta(\mu_0(\boldsymbol{x},q)) = F_{Y|X}(q|\boldsymbol{x})$ is also Lipschitz in q uniformly over $\boldsymbol{x}$), as well as boundedness of $\mu_0(\boldsymbol{x},q)$ uniformly over q and $\boldsymbol{x}$.

Proposition H.3 is proven. □

H.3 L_p regression

Proposition H.5 (Verification of Assumption B.4 for L_p regression)**.** *Suppose Assumptions B.1 to B.3 hold with $\mathcal{Q}$ a singleton, loss function $\rho(y,\eta) = |y-\eta|^p$, $p \in (1,2]$, $\mu_0(\cdot)$ as defined in (C.1), $\mathfrak{M}$ the Lebesgue measure. Assume the real inverse link function $\eta(\cdot)\colon \mathbb{R} \to \mathcal{E}$ is strictly monotonic and twice continuously differentiable with $\mathcal{E}$ an open connected subset of $\mathbb{R}$. Denoting by a_l and a_r the left and the right ends of $\mathcal{E}$ respectively (possibly $\pm\infty$), assume that $\int_{\mathbb{R}} \psi(y,a_l) f_{Y|X}(y|\boldsymbol{x})\,\mathrm{d}y < 0$ if a_l is finite, and $\int_{\mathbb{R}} \psi(y,a_r) f_{Y|X}(y|\boldsymbol{x})\,\mathrm{d}y > 0$ if a_r is finite. Also assume that $\mathbb{E}[|y_1|^{\nu(p-1)}] < \infty$ for some $\nu > 2$, and that $\boldsymbol{x} \mapsto f_{Y|X}(y|\boldsymbol{x})$ is continuous for any $y \in \mathcal{Y}$. In addition, assume that $\eta \mapsto \int_{\mathbb{R}} |\eta(\zeta) - y|^{p-1}\operatorname{sign}(\eta(\zeta)-y) f_{Y|X}(y|\boldsymbol{x})\,\mathrm{d}y$ is twice continuously differentiable with derivatives $\frac{\mathrm{d}^j}{\mathrm{d}\eta^j}\int_{\mathbb{R}} |\eta(\zeta) - y|^{p-1}\operatorname{sign}(\eta(\zeta)-y) f_{Y|X}(y|\boldsymbol{x})\,\mathrm{d}y = \int_{\mathbb{R}} |\eta(\zeta) - y|^{p-1}\operatorname{sign}(\eta(\zeta)-y)\frac{\partial^j}{\partial y^j} f_{Y|X}(y|\boldsymbol{x})\,\mathrm{d}y$ for $j \in \{1,2\}$. Moreover, the function $\int_{\mathbb{R}} |\eta(\zeta) - y|^{p-1}\operatorname{sign}(\eta(\zeta)-y)\frac{\partial}{\partial y} f_{Y|X}(y|\boldsymbol{x})\,\mathrm{d}y$ is bounded and bounded away from zero uniformly over $\boldsymbol{x} \in \mathcal{X}$, $\zeta \in B(\boldsymbol{x})$ with $B(\boldsymbol{x})$ defined in* (B.1)*, and the function $\int_{\mathbb{R}} |\eta(\zeta) - y|^{p-1}\operatorname{sign}(\eta(\zeta)-y)\frac{\partial^2}{\partial y^2} f_{Y|X}(y|\boldsymbol{x})\,\mathrm{d}y$ is bounded in absolute value uniformly over $\boldsymbol{x} \in \mathcal{X}$, $\zeta \in B(\boldsymbol{x})$. Then Assumptions B.4, B.5 and B.8, Eq.* (F.15) *and Condition (vi) of Lemma D.3 are also true.*

Proof. Since $\mathcal{Q}$ is a singleton, we will omit the index q in notations.

Verifying the assumptions of Lemma C.10 The fact that

$$\zeta \mapsto \int_{\mathbb{R}} \psi(y,\zeta) f_{Y|X}(y|\boldsymbol{x})\,\mathrm{d}y$$

is continuous is proven below in the verification of Assumption B.4(iii). To ensure that it crosses zero if a_l or a_r is not finite, we show

$$\int_{\mathbb{R}} \psi(y,\zeta) f_{Y|X}(y|\boldsymbol{x})\,\mathrm{d}y \to -\infty \text{ as } \zeta \to -\infty, \tag{H.3}$$

$$\int_{\mathbb{R}} \psi(y,\zeta) f_{Y|X}(y|\boldsymbol{x})\,\mathrm{d}y \to +\infty \text{ as } \zeta \to +\infty. \tag{H.4}$$

To prove (H.3), recall that $\psi(y,\zeta) = p|y-\zeta|^{p-1}\operatorname{sign}(\zeta - y)$ and therefore

$$\begin{aligned}
\int_{\mathbb{R}} \psi(y,\zeta) f_{Y|X}(y|\boldsymbol{x})\,\mathrm{d}y &= -p\int_{\zeta}^{+\infty} (y-\zeta)^{p-1} f_{Y|X}(y|\boldsymbol{x})\,\mathrm{d}y + p\int_{-\infty}^{\zeta} (\zeta - y)^{p-1} f_{Y|X}(y|\boldsymbol{x})\,\mathrm{d}y \\
&= -p(-\zeta)^{p-1} \underbrace{\int_{\mathbb{R}} \left(1 + \frac{y}{-\zeta}\right)^{p-1} \mathbb{1}\{y > \zeta\} f_{Y|X}(y|\boldsymbol{x})\,\mathrm{d}y}_{\to 1} \\
&\quad + p\underbrace{\int_{\mathbb{R}} (\zeta - y)^{p-1} \mathbb{1}\{y \le \zeta\} f_{Y|X}(y|\boldsymbol{x})\,\mathrm{d}y}_{\to 0} \to -\infty,
\end{aligned}$$

where we used dominated convergence because for $-\zeta > 1$ we have $1 + y/(-\zeta) \le 1 + |y|$ in the first integral and $\zeta - y \le -y = |y|$ in the second integral. Equation (H.4) is proven similarly.

Verifying Assumption B.4(i) The function $\rho(y,\eta)$ is continuously differentiable with respect to $\eta \in \mathbb{R}$, and its first derivative is the continuous function $\psi(y,\eta) = p|y-\eta|^{p-1}\operatorname{sign}(\eta - y)$, therefore $\rho(y,\eta)$ for any fixed y is absolutely continuous with respect to η on bounded intervals.

Verifying Assumption B.4(ii) The first-order optimality condition is true because $\mu_0(\cdot)$ is defined this way in (C.1).

The function

$$\sigma^2(\boldsymbol{x}) := \mathbb{E}\big[\psi(y_i, \mu_0(\boldsymbol{x}))^2 \,\big|\, \boldsymbol{x}_i = \boldsymbol{x}\big] = p^2 \int_{\mathbb{R}} |y - \mu_0(\boldsymbol{x})|^{2p-2} f_{Y|X}(y|\boldsymbol{x})\,\mathrm{d}y$$

is continuous because $\boldsymbol{x} \mapsto |y - \mu_0(\boldsymbol{x})|^{2p-2} f_{Y|X}(y|\boldsymbol{x})$ is continuous and dominated by $(|y|^{2p-2} + C)C'$ for large enough constants C and C'. As a continuous function on a compact set, $\sigma^2(\boldsymbol{x})$ is bounded away from zero because it is non-zero since y_1 has a conditional density.

The family of functions $\{p|y - \eta(\mu_0(\boldsymbol{x}))|^{p-1}\operatorname{sign}(\eta(\mu_0(\boldsymbol{x})) - y)\}$ only contains one element. Note that $\boldsymbol{x} \mapsto |y - \eta(\mu_0(\boldsymbol{x}))|^{\nu(p-1)} f_{Y|X}(y|\boldsymbol{x})$ is continuous and dominated by $(|y|^{\nu(p-1)} + C)C'$ for large enough constants C and C'. Therefore, the function

$$\boldsymbol{x} \mapsto \int_{\mathbb{R}} |y - \eta(\mu_0(\boldsymbol{x}))|^{\nu(p-1)} f_{Y|X}(y|\boldsymbol{x})\,\mathrm{d}y$$

is also continuous. As a continuous function on a compact set, it is bounded.

Verifying Assumption B.4(iii) The function $x \mapsto |x|^{\alpha}\operatorname{sign}(x)$ for $\alpha \in (0,1]$ is α-Hölder for $x \in \mathbb{R}$ (with constant 2). Therefore, putting $\alpha := p - 1$, for any pair of reals ζ_1 and ζ_2 in a fixed bounded interval we have

$$\begin{aligned}
&\sup_{\boldsymbol{x}} \sup_{\lambda \in [0,1]} \sup_{y} |\psi(y, \eta(\zeta_1 + \lambda(\zeta_2 - \zeta_1))) - \psi(y, \eta(\zeta_2))| \\
&\quad \le 2p|\eta(\zeta_1 + \lambda(\zeta_2 - \zeta_1)) - \eta(\zeta_2)|^{p-1} \overset{\text{(a)}}{\lesssim} |\zeta_1 - \zeta_2|^{p-1},
\end{aligned}$$

where in (a) we used that the link function $\eta(\cdot)$ in a fixed bounded interval is Lipschitz.

Verifying Assumption B.4(iv) The conditions of Assumption B.4(iv) are directly assumed in the statement of Proposition H.5.

Verifying Assumption B.5 Since $\mathcal{G}_2$, $\mathcal{G}_4$, $\mathcal{G}_5$ are just singletons (and the existence of corresponding envelopes holds trivially), it is enough to consider $\mathcal{G}_1$ and $\mathcal{G}_3$.

Assume that $\boldsymbol{\beta}$ and $\widetilde{\boldsymbol{\beta}}$ are such that $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_\infty \le r$ and $\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_\infty \le r$. Note that

$$\begin{aligned}&\big|\big[\psi(y,\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}))-\psi(y,\eta(\mu_0(\boldsymbol{x})))\big]-\big[\psi(y,\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\widetilde{\boldsymbol{\beta}}))-\psi(y,\eta(\mu_0(\boldsymbol{x})))\big]\big|\\&\quad\le 2p\big|\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})-\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\widetilde{\boldsymbol{\beta}})\big|^{p-1}\lesssim\|\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\|_\infty^{p-1}.\end{aligned}$$

The result for $\mathcal{G}_1$ follows.

Similarly,

$$\begin{aligned}&\big|\big[\psi(y,\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}))-\psi(y,\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0))\big]-\big[\psi(y,\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\widetilde{\boldsymbol{\beta}}))-\psi(y,\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0))\big]\big|\\&\quad\le 2p\big|\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})-\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\widetilde{\boldsymbol{\beta}})\big|^{p-1}\lesssim\|\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\|_\infty^{p-1}.\end{aligned}\tag{H.5}$$

For a fixed cell $\delta\in\Delta$ the class of functions of the form

$$\big[\psi(y,\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta});q)-\psi(y,\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0);q)\big]\mathbb{1}\{\boldsymbol{x}\in\mathcal{N}_{[c\log n]}(\delta)\}$$

can be parametrized by $\boldsymbol{\beta}$ lying in a fixed vector space $\mathcal{B}_\delta$ of dimension $O(\log^d n)$. The result now follows from the bound (H.5) (by using (E.34) similarly to the proof of Proposition H.6).

Verifying the addition to Assumption B.8 Fix $\delta\in\Delta$. Let $\boldsymbol{\beta}$ and $\widetilde{\boldsymbol{\beta}}$ be such that $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_\infty\le r$ and $\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_\infty\le r$; let $\boldsymbol{v}$ and $\widetilde{\boldsymbol{v}}$ be such that $\|\boldsymbol{v}\|_\infty\le\varepsilon_n$ and $\|\widetilde{\boldsymbol{v}}\|_\infty\le\varepsilon_n$. To declutter notation, put

$$\begin{aligned}g(t)&:=\big[\psi(y,\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0+\boldsymbol{\beta})+t))-\psi(y,\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0))\big]\\&\qquad\times\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0+\boldsymbol{\beta})+t),\\\tilde{g}(t)&:=\big[\psi(y,\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0+\widetilde{\boldsymbol{\beta}})+t))-\psi(y,\eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0))\big]\\&\qquad\times\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta}_0+\widetilde{\boldsymbol{\beta}})+t).\end{aligned}$$

Note that

$$\int_{-\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{v}}^0 g(t)\,\mathrm{d}t-\int_{-\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\widetilde{\boldsymbol{v}}}^0\tilde{g}(t)\,\mathrm{d}t=\int_{-\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{v}}^0\big(g(t)-\tilde{g}(t)\big)\,\mathrm{d}t+\int_{-\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{v}}^{-\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\widetilde{\boldsymbol{v}}}\tilde{g}(t)\,\mathrm{d}t.$$

Since $\psi(y,\cdot)$ is $(p-1)$-Hölder continuous in the second argument and functions $\eta(\cdot)$, $\eta^{(1)}(\cdot)$ in a fixed bounded interval are Lipschitz, we get that uniformly over t and $\boldsymbol{x}$ in these integrals

$$\big|g(t)-\tilde{g}(t)\big|\lesssim\|\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\|_\infty^{p-1}+\big\|\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\big\|_\infty,$$

and $\big|\tilde{g}(t)\big|$ is bounded. This gives

$$\left|\int_{-\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{v}}^0 g(t)\,\mathrm{d}t-\int_{-\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\widetilde{\boldsymbol{v}}}^0\tilde{g}(t)\,\mathrm{d}t\right|\lesssim\varepsilon_n\big(\|\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\|_\infty^{p-1}+\big\|\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\big\|_\infty\big)+\big\|\boldsymbol{v}-\widetilde{\boldsymbol{v}}\big\|_\infty$$

It means that taking an ε-net (for ε smaller than 1) in the space of $\boldsymbol{\beta}$ and an $\varepsilon_n\varepsilon$-net in the space of $\boldsymbol{v}$ induces an $C_{22}\varepsilon_n\varepsilon$-net in the space of functions

$$\left\{(\boldsymbol{x}, y) \mapsto \int_{-\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{v}}^{0} g(t)\,\mathrm{d}t \cdot \mathbb{1}\{\boldsymbol{x} \in \mathcal{N}_{[c\log n]}(\delta)\} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, \|\boldsymbol{v}\|_\infty \le \varepsilon_n\right\}$$

in terms of the sup-norm, where C_{22} is some constant. Possibly increasing C_{22}, we can conclude that this class with envelope C_{22} satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim \log^d n$ (where we used that $\boldsymbol{\beta}$ and $\boldsymbol{v}$ can be assumed to lie in a vector space of dimension $O(\log^d n)$). By (E.34), the same can be said about the union of $O(h^{-d})$ such classes (corresponding to different δ). The verification is concluded.

Verifying (F.15) This is obvious because $\mathcal{Q}$ is a singleton.

Verifying Lemma D.3(vi) It follows from the proof of Lemma D.2 that

$$\begin{aligned}\big|\rho(y, \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta})) - \rho(y, \eta(\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\widetilde{\boldsymbol{\beta}}))\big| &\lesssim \big(1 + \overline{\psi}(\boldsymbol{x}, y)\big)\big|\boldsymbol{p}(\boldsymbol{x})^\mathsf{T}(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}})\big| \\ &\lesssim \big(1 + \overline{\psi}(\boldsymbol{x}, y)\big)\big\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\big\|_\infty.\end{aligned}$$

The required uniform entropy bound follows immediately from this.

Proposition H.5 is proven. □

H.4 Logistic regression

Proposition H.6 (Verification of Assumption B.4 and others for logistic regression)**.** *Suppose Assumptions B.1 to B.3 hold with $\mathcal{Q}$ a singleton, $\mathcal{Y} = \{0, 1\}$, $\eta(\theta) = 1/(1+e^{-\theta})$, $\mathfrak{M}$ is the counting measure on $\{0, 1\}$, and $\rho(y, \eta) = -y\log(\eta) - (1-y)\log(1-\eta)$. Assume also $\pi(\boldsymbol{x}) := \mathbb{P}\{y_1 = 1 \,|\, \boldsymbol{x}_1 = \boldsymbol{x}\}$ is continuous and $\pi(\boldsymbol{x})$ lies in the interval $(0, 1)$ for any $\boldsymbol{x} \in \mathcal{X}$. Then Assumptions B.4, B.5 and B.8, (F.15), Conditions (v) and (vi) of Lemma D.3 are true.*

We will prove Proposition H.6 now. Since $\mathcal{Q}$ is a singleton, we will omit the index q in notations.

Verifying Assumption B.4(i) The function $\rho(y, \eta)$ is infinitely smooth with respect to $\eta \in (0, 1)$, and its first derivative is $\psi(y, \eta) = (1-y)/(1-\eta) - y/\eta$, therefore $\rho(y, \eta)$ for any fixed y is absolutely continuous with respect to η on bounded intervals.

Verifying Assumption B.4(ii) We have $\mathbb{E}[\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i))) \,|\, \boldsymbol{x}_i] = 0$ since

$$\eta(\mu_0(\boldsymbol{x}_i)) = \mathbb{E}[y_i \,|\, \boldsymbol{x}_i] = \mathbb{P}\{y_i = 1 \,|\, \boldsymbol{x}_i\} = \pi(\boldsymbol{x}_i).$$

Next,

$$\begin{aligned}\sigma^2(\boldsymbol{x}) &= \mathbb{E}\big[\psi(y_i, \eta(\mu_0(\boldsymbol{x}_i)))^2 \,\big|\, \boldsymbol{x}_i = \boldsymbol{x}\big] = \frac{\mathbb{E}\big[(y_i - \pi(\boldsymbol{x}_i))^2 \,\big|\, \boldsymbol{x}_i = \boldsymbol{x}\big]}{\eta^{(1)}(\mu_0(\boldsymbol{x}))^2} \\ &= \frac{\pi(\boldsymbol{x})(1-\pi(\boldsymbol{x}))}{\pi(\boldsymbol{x})^2(1-\pi(\boldsymbol{x}))^2} = \frac{1}{\pi(\boldsymbol{x})(1-\pi(\boldsymbol{x}))}\end{aligned}$$

is continuous and bounded away from zero (is not less than 4).

Since $\boldsymbol{x}_i$ lies in a compact set, $\psi(y, \eta(\mu_0(\boldsymbol{x}_i)))$ is bounded, so it has moments of any order.

Verifying Assumption B.4(iii) The function $\rho(y,\eta)$ is infinitely smooth with respect to $\eta \in (0,1)$, and its first derivative is $\psi(y,\eta) = (1-y)/(1-\eta) - y/\eta$. Using the famous expression for the derivative of the logistic function $\eta^{(1)}(\theta) = \eta(\theta)(1-\eta(\theta))$, we get

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\rho(y,\eta(\theta)) = \psi(y,\eta(\theta))\eta^{(1)}(\theta) = (1-y)\eta(\theta) - y(1-\eta(\theta)) = \eta(\theta) - y,\\
&\frac{\partial^2}{\partial\theta^2}\rho(y,\eta(\theta)) = \eta^{(1)}(\theta) = \eta(\theta)(1-\eta(\theta)).
\end{aligned}$$

Since the logistic link maps to $(0,1)$, the second derivative is positive (and does not depend on y). Therefore, $\rho(y,\eta(\theta))$ is convex with respect to θ for any y.

Uniformly over ζ_1 and ζ_2 in a fixed bounded interval, we have

$$\sup_y \big|\psi(y,\eta(\zeta_1)) - \psi(y,\eta(\zeta_2))\big| \lesssim \big|\eta(\zeta_1) - \eta(\zeta_2)\big| \overset{\text{(a)}}{\leq} |\zeta_1 - \zeta_2|,$$

where in (a) we used that the derivative of $\eta(\cdot)$ does not exceed 1. So in this case the Hölder parameter $\alpha = 1$.

The logistic function $\eta(\cdot)$ is strictly monotonic and infinitely smooth on $\mathbb{R}$.

Verifying Assumption B.4(iv) In this case

$$\Psi(\boldsymbol{x};\eta) = \frac{\eta - \mathbb{E}[y_i \,|\, \boldsymbol{x}_i = \boldsymbol{x}]}{\eta(1-\eta)} = \frac{\eta - \pi(\boldsymbol{x})}{\eta(1-\eta)}.$$

This function is infinitely smooth in η for $\eta \in (0,1)$. Its first derivative

$$\Psi_1(\boldsymbol{x},\eta) = \frac{\partial}{\partial\eta}\Psi(\boldsymbol{x};\eta) = \frac{\eta^2 - 2\eta\pi(\boldsymbol{x}) + \pi(\boldsymbol{x})}{\eta^2(1-\eta)^2}.$$

Therefore,

$$\Psi_1(\boldsymbol{x},\eta(\zeta))\eta^{(1)}(\zeta)^2 = \eta(\zeta)^2 - 2\eta(\zeta)\pi(\boldsymbol{x}) + \pi(\boldsymbol{x}). \tag{H.6}$$

If $|\zeta - \mu_0(\boldsymbol{x})| \leq r$, then $\big|\eta(\zeta) - \eta(\mu_0(\boldsymbol{x}))\big| = \big|\eta(\zeta) - \pi(\boldsymbol{x})\big| \leq r$ (since the derivative of $\eta(\cdot)$ does not exceed 1). Since $0 < \min_{\boldsymbol{x}\in\mathcal{X}} \pi(\boldsymbol{x})(1-\pi(\boldsymbol{x})) \leq \max_{\boldsymbol{x}\in\mathcal{X}} \pi(\boldsymbol{x})(1-\pi(\boldsymbol{x})) < 1$, for small enough r the right-hand side of (H.6) for such ζ is also bounded away from zero and one.

Finally,

$$\Psi_2(\boldsymbol{x},\eta) = \frac{\partial}{\partial\eta}\Psi_1(\boldsymbol{x},\eta) = \frac{2\big(\eta^3 - 3\pi(\boldsymbol{x})\eta^2 + 3\pi(\boldsymbol{x})\eta - \pi(\boldsymbol{x})\big)}{\eta^3(1-\eta)^3}.$$

Again since $0 < \min_{\boldsymbol{x}\in\mathcal{X}} \pi(\boldsymbol{x})(1-\pi(\boldsymbol{x})) \leq \max_{\boldsymbol{x}\in\mathcal{X}} \pi(\boldsymbol{x})(1-\pi(\boldsymbol{x})) < 1$, for such ζ that $|\zeta - \mu_0(\boldsymbol{x})| \leq r$ and r small enough, the product $\eta(\zeta)(1-\eta(\zeta))$ is bounded away from zero. So for such ζ, $|\Psi_2(\boldsymbol{x},\eta(\zeta))|$ is uniformly bounded.

Lemma H.7 (Class G_1)**.** *The class*

$$\mathcal{G}_1 = \left\{\mathcal{X}\times\mathbb{R} \ni (\boldsymbol{x},y) \mapsto \frac{\eta(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta}) - y}{\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta})} - \frac{\pi(\boldsymbol{x}) - y}{\pi(\boldsymbol{x})\big(1-\pi(\boldsymbol{x})\big)} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq r\right\}$$

with a large enough constant envelope satisfies the uniform entropy bound (A.3) *with $A \lesssim 1$ and $V \lesssim K$.*

Proof of Lemma H.7. The class

$$\mathcal{G}_{11} = \left\{(\boldsymbol{x}, y) \mapsto \boldsymbol{p}(\boldsymbol{x})^\top \boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^K\right\}$$

is VC with index not exceeding $K + 2$ by Lemma 2.6.15 in [41]. Since in a fixed bounded interval $\eta(\cdot)$ is Lipschitz and $(\boldsymbol{x}, y) \mapsto y$ is one fixed function, the class

$$\mathcal{G}_{12} = \left\{(\boldsymbol{x}, y) \mapsto \eta(\boldsymbol{p}(\boldsymbol{x})^\top \boldsymbol{\beta}) - y : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_\infty \leq r\right\}$$

with a large enough constant envelope (recall that $\mathcal{Y}$ is a bounded set) satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim K$. Since $1/\eta^{(1)}(\cdot)$ in a fixed bounded interval is Lipschitz, the same is true of

$$\mathcal{G}_{13} = \left\{(\boldsymbol{x}, y) \mapsto \eta^{(1)}(\boldsymbol{p}(\boldsymbol{x})^\top \boldsymbol{\beta})^{-1} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_\infty \leq r\right\},$$

where we used again that under these constraints $\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x})^\top \boldsymbol{\beta})$ is bounded away from zero. This implies by Lemma C.4 that it is true of $\mathcal{G}_{12} \cdot \mathcal{G}_{13} - \psi(y, \eta(\boldsymbol{x}))$ (since $\psi(y, \eta(\boldsymbol{x}))$ is one fixed function), which is what we need. □

Lemma H.8 (Class $\mathcal{G}_3$)**.** *The class*

$$\begin{aligned}\mathcal{G}_3 := \Big\{&\mathcal{X} \times \mathbb{R} \ni (\boldsymbol{x}, y) \mapsto \\ &\left[\frac{\eta(\boldsymbol{p}(\boldsymbol{x})^\top \boldsymbol{\beta}) - y}{\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x})^\top \boldsymbol{\beta})} - \frac{\eta(\boldsymbol{p}(\boldsymbol{x})^\top \boldsymbol{\beta}_0) - y}{\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x})^\top \boldsymbol{\beta}_0)}\right] \mathbb{1}\{\boldsymbol{x} \in \mathcal{N}_{[c \log n]}(\delta)\} : \\ &\|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \leq r, \delta \in \Delta\Big\}\end{aligned}$$

with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim \log^d n$.

Proof of Lemma H.8. For a fixed vector space $\mathcal{B}$ of dimension $\dim \mathcal{B}$,

$$\mathcal{W}_{\mathcal{B}} := \left\{(\boldsymbol{x}, y) \mapsto \frac{\eta(\boldsymbol{p}(\boldsymbol{x})^\top \boldsymbol{\beta}) - y}{\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x})^\top \boldsymbol{\beta})} : \boldsymbol{\beta} \in \mathcal{B}, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \leq r\right\}$$

with a large enough constant envelope (recall that $\mathcal{Y}$ is a bounded set) satisfies the uniform entropy bound with $A \lesssim 1$ and $V \lesssim \dim \mathcal{B}$ by the same argument as in the proof of Lemma H.7. Therefore, for any fixed $c > 0$, $\delta \in \Delta$, the class

$$\mathcal{W}_{3,\delta} := \left\{(\boldsymbol{x}, y) \mapsto \frac{\eta(\boldsymbol{p}(\boldsymbol{x})^\top \boldsymbol{\beta}) - y}{\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x})^\top \boldsymbol{\beta})} \mathbb{1}\{\boldsymbol{x} \in \mathcal{N}_{[c \log n]}(\delta)\} : \boldsymbol{\beta} \in \mathbb{R}^K, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \leq r\right\}$$

with a large enough constant envelope also satisfies the uniform entropy bound with $A \lesssim 1$ and $V \lesssim \log^d n$ because it is contained in the product of $\mathcal{W}_{\mathcal{B}_{3,\delta}}$ for some vector space $\mathcal{B}_{3,\delta}$ of dimension $\dim \mathcal{B}_{3,\delta} \lesssim \log^d n$ and a fixed function $\mathbb{1}\{\boldsymbol{x} \in \mathcal{N}_{[c \log n]}(\delta)\}$. Subtracting a fixed bounded function does not change this fact, so the same is true of

$$\begin{aligned}\mathcal{G}_{3,\delta} := \Big\{&(\boldsymbol{x}, y) \mapsto \\ &\left[\frac{\eta(\boldsymbol{p}(\boldsymbol{x})^\top \boldsymbol{\beta}) - y}{\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x})^\top \boldsymbol{\beta})} - \frac{\eta(\boldsymbol{p}(\boldsymbol{x})^\top \boldsymbol{\beta}_0) - y}{\eta^{(1)}(\boldsymbol{p}(\boldsymbol{x})^\top \boldsymbol{\beta}_0)}\right] \mathbb{1}(\boldsymbol{x} \in \mathcal{N}_{[c \log n]}(\delta)) : \\ &\boldsymbol{\beta} \in \mathbb{R}^K, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \leq r\Big\}\end{aligned}$$

Since there are $O(h^{-d})$ such classes and $\log(1/h) \lesssim \log n$, using the chain (E.34) we obtain that $\mathcal{G}_3$ satisfies the uniform entropy bound (A.3) with with $A \lesssim 1$ and $V \lesssim \log^d n$. □

Verifying Assumption B.5 Classes $\mathcal{G}_2$, $\mathcal{G}_4$, $\mathcal{G}_5$ are just singletons (and the existence of corresponding envelopes holds trivially). Classes $\mathcal{G}_1$ and $\mathcal{G}_3$ are tackled in Lemma H.7 and Lemma H.8.

Verifying Assumption B.8 This is verified (in a more general setting) in Appendix B.1.2.

Verifying (F.15) This is obvious since $\mathcal{Q}$ is a singleton.

Verifying the condition in Lemma D.3(v) Recall that in this case

$$\Psi_1(\boldsymbol{x}, \eta) = \frac{\eta^2 - 2\eta\pi(\boldsymbol{x}) + \pi(\boldsymbol{x})}{\eta^2(1-\eta)^2}.$$

The numerator is always positive since $0 < \pi(\boldsymbol{x}) < 1$, and the denominator is also positive since $\eta \in (0,1)$. Since $\Psi_1(\boldsymbol{x}, \eta)$ is continuous in both arguments and the image of a compact set under a continuous mapping is compact, we see that for any fixed compact subset of $(0,1)$, $\Psi_1(\boldsymbol{x}, \eta)$ is bounded away from zero uniformly over $\boldsymbol{x} \in \mathcal{X}$ and η lying in this compact subset.

Verifying the condition in Lemma D.3(vi) In this verification, we will use $\theta_1 := \boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}$, $\tilde{\theta}_1 := \boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\widetilde{\boldsymbol{\beta}}$ and $\theta_2 := \boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\beta}_0$ to simplify notations. Note that for θ lying in a fixed compact, both functions $\log \eta(\theta)$ and $\log(1-\eta(\theta))$ are Lipschitz in θ, so if $\|\boldsymbol{\beta}\|_\infty \le R$ and $\|\widetilde{\boldsymbol{\beta}}\|_\infty \le R$, we have

$$|\rho(y, \eta(\theta_1)) - \rho(y, \eta(\theta_2)) - \rho(y, \eta(\tilde{\theta}_1)) + \rho(y, \eta(\theta_2))| \lesssim |\theta_1 - \tilde{\theta}_1| \lesssim |\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}|,$$

where the constants in $\lesssim$ are allowed to depend on R but not on $\boldsymbol{\beta}, \widetilde{\boldsymbol{\beta}}$, $\boldsymbol{x}$ or y (we used that y and $1-y$ are bounded by 1). The result follows.

This concludes the proof of Proposition H.6.

H.5 Examples of basis functions

Fix $0 \le s < m$. Take any segment $[a,b]$ and partition it into J sub-segments by taking $J-1$ knots $a < \tau_1 < \ldots \tau_{J-1} < b$. Consider a non-increasing tuple of real numbers $(t_k)_{k=1}^{2m+(m-s)(J-1)}$ defined as follows:

$$\begin{aligned}
&t_1 = \ldots = t_m = a,\\
&t_{m+1} = \ldots = t_{2m-s} = \tau_1,\\
&\ldots\\
&t_{m+(m-s)(J-2)+1} = \ldots = t_{m+(m-s)(J-1)} = \tau_{J-1},\\
&t_{m+(m-s)(J-1)+1} = \ldots = t_{2m+(m-s)(J-1)} = b,
\end{aligned}$$

and put for $l \in \{1, 2, \ldots, m + (m-s)(J-1)\}$

$$p_l(x) = (-1)^m (t_{l+m} - t_l)[t_l, \ldots, t_{l+m}](x-t)_+^{m-1}, \quad x \in [a,b),$$
$$p_{m+(m-s)(J-1)}(b) = \lim_{x \uparrow b} p_{m+(m-s)(J-1)}(x),$$

where $[t_l, \ldots, t_{l+m}]g(t,x)$ denotes the divided difference [37, Definition 2.49] of $t \mapsto g(t,x)$ over points $t_l, \ldots, t_{l+m}$, and $\alpha_+ := \alpha \vee 0$. We will call the basis $\{p_l(\cdot)\}_{l=1}^{m+(m-s)(J-1)}$ the (normalized)

B-spline basis of order m with knots $\tau_1 < \ldots < \tau_{J-1}$, each of multiplicity $m-s$. The functions $\{p_l(\cdot)\}_{l=1}^{m+(m-s)(J-1)}$ form a basis in the function space

$$\big\{f(\cdot) \in \mathcal{C}^{s-1}[a,b] : f(\cdot) \text{ is a polynomial of degree } m-1 \text{ on each } [\tau_j, \tau_{j+1}],\ j \in \{0,\ldots,J-1\}\big\},$$

where we put $\tau_0 = a$, $\tau_J = b$ for simplicity ([37, Corollary 4.10])[3]; moreover, each $p_l(\cdot)$ is strictly positive on (t_l, t_{l+m}), zero on the complement of $[t_l, t_{l+m}]$, and does not exceed 1 ([37, Theorem 4.9]).

This definition can be extended to dimension d by considering $\mathcal{X} = \bigotimes_{\ell=1}^d [a_\ell, b_\ell]$ and the corresponding tensor products

$$p_{l_1,\ldots,l_d}(\boldsymbol{x}) = p_{l_1}(x_1)\ldots p_{l_d}(x_d)$$

for all tuples $\boldsymbol{l} = (l_1, \ldots, l_d)$ such that $l_\ell \in \{1, 2, \ldots, m+(m-s)(J_\ell - 1)\}$. The d partitions of $[a_\ell, b_\ell]$ induce a tensor-product partition Δ of $\mathcal{X}$ (with $\kappa = J_1 \ldots J_d$ cells), and the basis $\boldsymbol{p}(\boldsymbol{x})$ (arranged in a lexicographic order of $\boldsymbol{l}$) is called the *tensor-product B-spline basis* of order m associated with the tensor-product partition Δ ([37, Definition 12.3]).

We refer to [37] for details.

Proposition H.9 (Verifying Assumptions B.2 and B.6)**.** *Suppose Assumptions B.1 and B.3 hold with the tensor-product partition Δ as described above. Let $\boldsymbol{p}(\boldsymbol{x})$ be a tensor-product B-spline basis of order m associated with Δ. Then $\boldsymbol{p}(\boldsymbol{x})$ satisfies Assumptions B.2 and B.6.*

Proof. It follows from the more general argument for Lemma SA-6.1 in [11]. □

Taking $s = m-1$ recovers standard tensor-product B-splines with simple (multiplicity 1) partition knots, whereas $s = 0$ corresponds to piecewise polynomials, where, in particular, each basis function is only supported on one cell (making Lemmas D.4 and D.5 applicable). Additional examples are provided in [11, Section SA-6].

I Other parameters of interest

This section formalizes the discussion in Section 9 of the paper. The following theorem is now a simple corollary of the previous results presented in this supplemental appendix.

Theorem I.1 (Other parameters of interest)**.**

(a) *Suppose all the conditions of Theorem F.4(a) hold with $\boldsymbol{v} = \boldsymbol{0}$ and $r_{\mathtt{UC}} = o(1)$. Then*

$$\begin{aligned}&\sup_{\boldsymbol{q}\in\mathcal{Q}}\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\frac{\eta\big(\widehat{\mu}(\boldsymbol{x},\boldsymbol{q})\big)-\eta\big(\mu_0(\boldsymbol{x},\boldsymbol{q})\big)}{\big|\eta^{(1)}(\mu_0(\boldsymbol{x},\boldsymbol{q}))\big|\sqrt{\bar{\Omega}_{\boldsymbol{0}}(\boldsymbol{x},\boldsymbol{q})/n}} - \operatorname{sign}\big\{\eta^{(1)}(\mu_0(\boldsymbol{x},\boldsymbol{q}))\big\}\bar{\boldsymbol{\ell}}_{\boldsymbol{0}}(\boldsymbol{x},\boldsymbol{q})^{\mathsf{T}}\boldsymbol{Z}(\boldsymbol{q})\right| \\ &\quad = o_{\mathbb{P}}(r_{\mathtt{SA}}) + O_{\mathbb{P}}\Big(\sqrt{nh^d}\big(r_{\mathtt{UC}}^2 + r_{\mathtt{BR}}\big)\Big)\end{aligned}$$

with $\boldsymbol{Z}(\boldsymbol{q})$ defined in Theorem F.4.

[3] The symbol $\mathcal{C}^{-1}[a,b]$ here corresponds to the family of all functions $[a,b] \to \mathbb{R}$ (with no smoothness restrictions).

(b) Fix any $k \in \{1, \ldots, d\}$. Suppose all the conditions of Theorem F.4(a) hold with $\boldsymbol{v} = \boldsymbol{e}_k$, where $\boldsymbol{e}_k = (0, \ldots, 0, 1, 0, \ldots, 0)^\mathsf{T} \in \mathbb{R}^d$ with 1 at the kth place. Then

$$\sup_{\boldsymbol{q},\boldsymbol{x}} \left| \frac{\eta^{(1)}\big(\widehat{\mu}(\boldsymbol{x},\boldsymbol{q})\big)\widehat{\mu}^{(\boldsymbol{e}_k)}(\boldsymbol{x},\boldsymbol{q}) - \eta^{(1)}\big(\mu_0(\boldsymbol{x},\boldsymbol{q})\big)\mu_0^{(\boldsymbol{e}_k)}(\boldsymbol{x},\boldsymbol{q})}{\big|\eta^{(1)}(\mu_0(\boldsymbol{x},\boldsymbol{q}))\big|\sqrt{\bar{\Omega}_{\boldsymbol{e}_k}(\boldsymbol{x},\boldsymbol{q})}} - \operatorname{sign}\big\{\eta^{(1)}(\mu_0(\boldsymbol{x},\boldsymbol{q}))\big\}\bar{\boldsymbol{\ell}}_{\boldsymbol{e}_k}(\boldsymbol{x},\boldsymbol{q})^\mathsf{T}\boldsymbol{Z}(\boldsymbol{q}) \right|$$
$$= o_{\mathbb{P}}(r_{\mathtt{SA}}) + O_{\mathbb{P}}\Big(\sqrt{nh^d}\big(r_{\mathtt{UC}}^2 + r_{\mathtt{BR}} + hr_{\mathtt{UC}}\big)\Big).$$

(c) If $\overline{\psi}(\boldsymbol{x}_i, y_i)$ is σ^2-sub-Gaussian conditionally on $\boldsymbol{x}_i$, then in Assertions (a) and (b) $r_{\mathtt{SA}}$ can be replaced with $r_{\mathtt{SA}}^{\mathrm{sub}}$.

Proof. (a) We have uniformly over $\boldsymbol{q}$ and $\boldsymbol{x}$

$$\begin{aligned}
&\eta(\widehat{\mu}(\boldsymbol{x},\boldsymbol{q})) - \eta(\mu_0(\boldsymbol{x},\boldsymbol{q})) \\
&\overset{\text{(a)}}{=} \eta^{(1)}(\mu_0(\boldsymbol{x},\boldsymbol{q}))(\widehat{\mu}(\boldsymbol{x},\boldsymbol{q}) - \mu_0(\boldsymbol{x},\boldsymbol{q})) + \frac{\eta^{(2)}(\xi)}{2}(\widehat{\mu}(\boldsymbol{x},\boldsymbol{q}) - \mu_0(\boldsymbol{x},\boldsymbol{q}))^2 \\
&\overset{\text{(b)}}{=} \eta^{(1)}(\mu_0(\boldsymbol{x},\boldsymbol{q}))(\widehat{\mu}(\boldsymbol{x},\boldsymbol{q}) - \mu_0(\boldsymbol{x},\boldsymbol{q})) + O_{\mathbb{P}}(r_{\mathtt{UC}}^2) \\
&\overset{\text{(c)}}{=} \eta^{(1)}(\mu_0(\boldsymbol{x},\boldsymbol{q}))t_{\boldsymbol{0}}(\boldsymbol{x},\boldsymbol{q})\sqrt{\bar{\Omega}_{\boldsymbol{0}}(\boldsymbol{x},\boldsymbol{q})/n} + O_{\mathbb{P}}(r_{\mathtt{UC}}^2 + r_{\mathtt{BR}}) \\
&\overset{\text{(d)}}{=} \eta^{(1)}(\mu_0(\boldsymbol{x},\boldsymbol{q}))\sqrt{\bar{\Omega}_{\boldsymbol{0}}(\boldsymbol{x},\boldsymbol{q})/n}\Big[t_{\boldsymbol{0}}(\boldsymbol{x},\boldsymbol{q}) + O_{\mathbb{P}}\big(\sqrt{nh^d}(r_{\mathtt{UC}}^2 + r_{\mathtt{BR}})\big)\Big].
\end{aligned}$$

Here, (a) is by Taylor expansion, with some $\xi = \xi_{\boldsymbol{q},\boldsymbol{x}}$ between $\widehat{\mu}(\boldsymbol{x},\boldsymbol{q})$ and $\mu_0(\boldsymbol{x},\boldsymbol{q})$. In (b), we used $r_{\mathtt{UC}} = o(1)$ (giving that $\eta^{(2)}(\xi)$ does not exceed a fixed constant not depending on $\boldsymbol{q}$ or $\boldsymbol{x}$) and (F.12). (c) is by (F.37) and since $\eta^{(1)}(\mu_0(\boldsymbol{x},\boldsymbol{q}))$ is uniformly bounded. (d) is by $h^{-2|\boldsymbol{v}|-d} \lesssim_{\mathbb{P}} \inf_{\boldsymbol{q},\boldsymbol{x}} |\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})|$ (by Lemma F.6) and since $|\eta^{(1)}(\mu_0(\boldsymbol{x},\boldsymbol{q}))|$ is bounded away from zero by Assumption B.4(iv).

Rewriting, we obtain

$$\sup_{\boldsymbol{q},\boldsymbol{x}} \left| \frac{\eta(\widehat{\mu}(\boldsymbol{x},\boldsymbol{q})) - \eta(\mu_0(\boldsymbol{x},\boldsymbol{q}))}{\eta^{(1)}(\mu_0(\boldsymbol{x},\boldsymbol{q}))\sqrt{\bar{\Omega}_{\boldsymbol{0}}(\boldsymbol{x},\boldsymbol{q})/n}} - t_{\boldsymbol{0}}(\boldsymbol{x},\boldsymbol{q}) \right| \lesssim_{\mathbb{P}} \sqrt{nh^d}\big(r_{\mathtt{UC}}^2 + r_{\mathtt{BR}}\big).$$

It is left to combine this with (F.19) and use the triangle inequality.

(b) We have uniformly over $\boldsymbol{q}$ and $\boldsymbol{x}$

$$
\begin{aligned}
&\eta^{(1)}(\widehat{\mu}(\boldsymbol{x},\boldsymbol{q}))\widehat{\mu}^{(\boldsymbol{e}_k)}(\boldsymbol{x},\boldsymbol{q}) - \eta^{(1)}(\mu_0(\boldsymbol{x},\boldsymbol{q}))\mu_0^{(\boldsymbol{e}_k)}(\boldsymbol{x},\boldsymbol{q}) \\
&\overset{\text{(a)}}{=} \eta^{(1)}(\mu_0(\boldsymbol{x},\boldsymbol{q}))\big(\widehat{\mu}^{(\boldsymbol{e}_k)}(\boldsymbol{x},\boldsymbol{q}) - \mu_0^{(\boldsymbol{e}_k)}(\boldsymbol{x},\boldsymbol{q})\big) \\
&\quad + \big(\eta^{(1)}(\widehat{\mu}(\boldsymbol{x},\boldsymbol{q})) - \eta^{(1)}(\mu_0(\boldsymbol{x},\boldsymbol{q}))\big)\widehat{\mu}^{(\boldsymbol{e}_k)}(\boldsymbol{x},\boldsymbol{q}) \\
&\overset{\text{(b)}}{=} \eta^{(1)}(\mu_0(\boldsymbol{x},\boldsymbol{q}))\big(\widehat{\mu}^{(\boldsymbol{e}_k)}(\boldsymbol{x},\boldsymbol{q}) - \mu_0^{(\boldsymbol{e}_k)}(\boldsymbol{x},\boldsymbol{q})\big) + \eta^{(2)}(\zeta)(\widehat{\mu}(\boldsymbol{x},\boldsymbol{q}) - \mu_0(\boldsymbol{x},\boldsymbol{q}))\widehat{\mu}^{(\boldsymbol{e}_k)}(\boldsymbol{x},\boldsymbol{q}) \\
&\overset{\text{(c)}}{=} \eta^{(1)}(\mu_0(\boldsymbol{x},\boldsymbol{q}))t_{\boldsymbol{e}_k}(\boldsymbol{x},\boldsymbol{q})\sqrt{\bar{\Omega}_{\boldsymbol{e}_k}(\boldsymbol{x},\boldsymbol{q})/n} \\
&\quad + \eta^{(2)}(\zeta)\big(\widehat{\mu}(\boldsymbol{x},\boldsymbol{q}) - \mu_0(\boldsymbol{x},\boldsymbol{q})\big)\widehat{\mu}^{(\boldsymbol{e}_k)}(\boldsymbol{x},\boldsymbol{q}) + O_{\mathbb{P}}(h^{-1}r_{\mathtt{BR}}) \\
&\overset{\text{(d)}}{=} \eta^{(1)}(\mu_0(\boldsymbol{x},\boldsymbol{q}))t_{\boldsymbol{e}_k}(\boldsymbol{x},\boldsymbol{q})\sqrt{\bar{\Omega}_{\boldsymbol{e}_k}(\boldsymbol{x},\boldsymbol{q})/n} \\
&\quad + \eta^{(2)}(\zeta)\big(\widehat{\mu}(\boldsymbol{x},\boldsymbol{q}) - \mu_0(\boldsymbol{x},\boldsymbol{q})\big)\big(\widehat{\mu}^{(\boldsymbol{e}_k)}(\boldsymbol{x},\boldsymbol{q}) - \mu_0^{(\boldsymbol{e}_k)}(\boldsymbol{x},\boldsymbol{q})\big) \\
&\quad + \eta^{(2)}(\zeta)\big(\widehat{\mu}(\boldsymbol{x},\boldsymbol{q}) - \mu_0(\boldsymbol{x},\boldsymbol{q})\big)\mu_0^{(\boldsymbol{e}_k)}(\boldsymbol{x},\boldsymbol{q}) + O_{\mathbb{P}}(h^{-1}r_{\mathtt{BR}}) \\
&\overset{\text{(e)}}{=} \eta^{(1)}(\mu_0(\boldsymbol{x},\boldsymbol{q}))t_{\boldsymbol{e}_k}(\boldsymbol{x},\boldsymbol{q})\sqrt{\bar{\Omega}_{\boldsymbol{e}_k}(\boldsymbol{x},\boldsymbol{q})/n} + O_{\mathbb{P}}\big(h^{-1}(r_{\mathtt{UC}}^2 + r_{\mathtt{BR}}) + r_{\mathtt{UC}}\big) \\
&\overset{\text{(f)}}{=} \eta^{(1)}(\mu_0(\boldsymbol{x},\boldsymbol{q}))\sqrt{\bar{\Omega}_{\boldsymbol{e}_k}(\boldsymbol{x},\boldsymbol{q})/n}\Big[t_{\boldsymbol{e}_k}(\boldsymbol{x},\boldsymbol{q}) + O_{\mathbb{P}}\Big(h\sqrt{nh^d}\big(h^{-1}(r_{\mathtt{UC}}^2 + r_{\mathtt{BR}}) + r_{\mathtt{UC}}\big)\Big)\Big].
\end{aligned}
$$

Here, (a) is just rewriting. In (b) we used the mean-value theorem, with some $\zeta = \zeta_{\boldsymbol{q},\boldsymbol{x}}$ between $\widehat{\mu}(\boldsymbol{x},\boldsymbol{q})$ and $\mu_0(\boldsymbol{x},\boldsymbol{q})$. In (c) we used (F.37) with $\boldsymbol{v} = \boldsymbol{e}_k$ and that $\eta^{(1)}(\mu_0(\boldsymbol{x},\boldsymbol{q}))$ is bounded uniformly over $\boldsymbol{q},\boldsymbol{x}$. (d) is just rewriting. In (e) we used $r_{\mathtt{UC}} = o(1)$ (giving that $\eta^{(2)}(\zeta)$ does not exceed a fixed constant not depending on $\boldsymbol{q}$ or $\boldsymbol{x}$), (F.12) with $\boldsymbol{v} = \boldsymbol{0}$ and $\boldsymbol{v} = \boldsymbol{e}_k$, and uniform boundedness of first partial derivatives of $\mu_0(\cdot,\boldsymbol{q})$. (f) is by $h^{-2|\boldsymbol{v}|-d} \lesssim_{\mathbb{P}} \inf_{\boldsymbol{q},\boldsymbol{x}} |\bar{\Omega}_{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{q})|$ (by Lemma F.6) and since $\big|\eta^{(1)}(\mu_0(\boldsymbol{x},\boldsymbol{q}))\big|$ is bounded away from zero by Assumption B.4(iv).

Rewriting, we obtain

$$
\begin{aligned}
\sup_{\boldsymbol{q},\boldsymbol{x}}&\left|\frac{\eta^{(1)}(\widehat{\mu}(\boldsymbol{x},\boldsymbol{q}))\widehat{\mu}^{(\boldsymbol{e}_k)}(\boldsymbol{x},\boldsymbol{q}) - \eta^{(1)}(\mu_0(\boldsymbol{x},\boldsymbol{q}))\mu_0^{(\boldsymbol{e}_k)}(\boldsymbol{x},\boldsymbol{q})}{\eta^{(1)}(\mu_0(\boldsymbol{x},\boldsymbol{q}))\sqrt{\bar{\Omega}_{\boldsymbol{e}_k}(\boldsymbol{x},\boldsymbol{q})}} - t_{\boldsymbol{e}_k}(\boldsymbol{x},\boldsymbol{q})\right| \\
&\lesssim_{\mathbb{P}} \sqrt{nh^d}\big(r_{\mathtt{UC}}^2 + r_{\mathtt{BR}} + hr_{\mathtt{UC}}\big).
\end{aligned}
$$

It is left to combine this with (F.19) and use the triangle inequality.

(c) The argument is the same as for Parts (i) and (ii) with $r_{\mathtt{SA}}$ replaced by $r_{\mathtt{SA}}^{\mathrm{sub}}$. □

J Simulation Evidence

We conducted a small simulation experiment to demonstrate the finite sample properties of the partitioning-based M-estimation methodology. We studied pointwise and uniform (over $q \in \mathcal{Q}$, $\boldsymbol{x} \in \mathcal{X}$) estimation and inference for the conditional distribution function (Example 2 in the paper, and discussed in Section H.2 above).

We considered four data generating processes. Model 1 through 3 were chosen to satisfy $y_i = m(\boldsymbol{x}_i) + \epsilon_i$, where $\boldsymbol{x}_i \sim \text{Uniform}(\mathcal{X})$ with $\mathcal{X} = [0,1]^d$, $\epsilon_i \sim \text{Normal}(0,1)$, $\boldsymbol{x}_i \perp\!\!\!\perp \epsilon_i$, $m(\boldsymbol{x})$ is defined in Table 4. For Model 4, we took the treatment effect model from [21] (with a slight change so that the support of $\boldsymbol{x}_i$ is also $[0,1]$):

$$
y_i = T_i\left[\mathbf{1}_{V_i \le 1-x_i}\frac{V_i^2}{1-x_i} + \mathbf{1}_{V_i > 1-x_i}V_i\right] + (1-T_i)\left[\mathbf{1}_{U_i \le x_i}\frac{U_i^2}{x_i} + \mathbf{1}_{U_i > x_i}U_i\right],
$$

where $T_i = \mathbf{1}_{U_i^T < x_i}$, and $x_i, U_i^T, U_i, V_i \sim \text{Uniform}[0,1]$ are independent. The estimand is $\mu_0(\boldsymbol{x}, q)$, where

$$\eta(\mu_0(\boldsymbol{x}, q)) = F(q|\boldsymbol{x}) = \mathbb{P}\{y_i \leq q \,|\, \boldsymbol{x}_i = \boldsymbol{x}\}$$

with $q \in \mathcal{Q} = [-0.2, 0.2]$ for Models 1 through 3 and $\mathcal{Q} = [0.2, 0.8]$ for Model 4; the link function $\eta(\cdot)$ is the complementary log-log link: $\eta(t) = 1 - e^{-e^t}$.

Table 4: Definition of the $m(\boldsymbol{x})$ function.

	$d = 1$	$d = 2$
Model 1	$\sin(2\pi x)/2$	$(\sin(2\pi x_1) + \sin(2\pi x_2))/4$
Model 2	$\sin(5x)\sin(10x)/2$	$\sin(5x_1)\sin(10x_2)/2$
Model 3	$(1 - (4x-2)^2)^2/4$	$(1 - (4x_1 - 2)^2)^2 \sin(5x_2)/4$

We use the tensor-product B-spline basis $\boldsymbol{p}(\boldsymbol{x})$ on $[0,1]^d$ as defined in Appendix H.5 with equally spaced knots. (Recall that the number of knots on each segment is $J-1$, and thus $\kappa = J^d$.) The estimator $\widehat{\mu}(\boldsymbol{x}, q) = \boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\widehat{\boldsymbol{\beta}}(q)$ is constructed by solving the optimization problem

$$\widehat{\boldsymbol{\beta}}(q) \in \operatorname*{arg\,min}_{\boldsymbol{b} \in \mathbb{R}^K} \sum_{i=1}^n \rho(y_i, \eta(\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\boldsymbol{b}); q)$$

for each q of interest separately (with $\rho(\cdot)$ defined as in Proposition H.3), using the L-BFGS-B algorithm initialized at $\mathbf{0}$. The matrices $\bar{\mathbf{Q}}_q$ and $\bar{\boldsymbol{\Sigma}}_q$ were estimated using, respectively,

$$\widehat{\mathbf{Q}}_{\boldsymbol{q}} = 2\mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\eta^{(1)}(\widehat{\mu}(\boldsymbol{x}_i, q))^2]$$

and

$$\widehat{\boldsymbol{\Sigma}}_q = 4\mathbb{E}_n[\boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^\mathsf{T}\widehat{F}_{Y|X}(q|\boldsymbol{x})(1 - \widehat{F}_{Y|X}(q|\boldsymbol{x}))\eta^{(1)}(\widehat{\mu}(\boldsymbol{x}_i, \boldsymbol{q}))^2],$$

with $\widehat{F}_{Y|X}(q|\boldsymbol{x}) = \eta(\widehat{\mu}(\boldsymbol{x}, q))$. Thus, the variance estimator is $\widehat{\Omega}(\boldsymbol{x}, q) = \boldsymbol{p}(\boldsymbol{x})^\mathsf{T}\widehat{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{q}}\widehat{\mathbf{Q}}_{\boldsymbol{q}}^{-1}\boldsymbol{p}(\boldsymbol{x})$.

The simulation experiment considered two sample sizes ($n = 5\,000$ and $n = 50\,000$), two dimensions ($d = 1$ and $d = 2$), and $1\,000$ replications. The pointwise estimation and inference results are presented in Tables 5 and 7 ($n = 5\,000$) and Tables 9 and 11 ($n = 50\,000$). For each model, we consider two choices of J and three evaluation points for $(q, \boldsymbol{x})$. The root mean squared error (RMSE) for point estimators, the coverage rates and the average widths of pointwise 95% nominal confidence intervals (CIs) are reported. The uniform estimation and inference results are presented in Tables 6 and 8 ($n = 5\,000$) and Tables 10 and 12 ($n = 50\,000$). We consider two choices of J, and use a discrete grid of 8 equally spaced points in place of the continuous segment $\mathcal{Q}$ and 10 (for $d = 1$) or $5 \times 5 = 25$ (for $d = 2$) points in place of $\mathcal{X}$. The uniform 95% nominal confidence bands (CBs) are constructed using the plug-in approach as in Appendix G.1, and the covariance structure of the discrete analogue of $\widehat{\boldsymbol{Z}}(q)$ is obtained as in Remark H.4. The maximum estimation error, the uniform coverage rates and the average widths of confidence bands are reported.

Table 5: Pointwise simulation results for points $\{(q^{(k)}, x^{(k)})\}_{k=1}^3$, averaged across 1 000 replications with $n = 5\,000$, $d = 1$, where $(q^{(1)}, q^{(2)}, q^{(3)}) = (-0.2, 0, 0.2)$ for Model 1 through 3, $(q^{(1)}, q^{(2)}, q^{(3)}) = (0.45, 0.6, 0.75)$ for Model 4; $(x^{(1)}, x^{(2)}, x^{(3)}) = (0.3, 0.1, 0.2)$.

		RMSE			Coverage			CI width		
Model	J	Point 1	Point 2	Point 3	Point 1	Point 2	Point 3	Point 1	Point 2	Point 3
1	4	0.088	0.074	0.074	0.952	0.953	0.944	0.336	0.289	0.287
	6	0.110	0.077	0.088	0.946	0.951	0.944	0.420	0.303	0.339
2	4	0.069	0.071	0.070	0.953	0.947	0.946	0.270	0.277	0.274
	6	0.088	0.076	0.085	0.950	0.942	0.948	0.343	0.291	0.321
3	4	0.073	0.099	0.068	0.953	0.943	0.945	0.279	0.370	0.258
	6	0.084	0.110	0.072	0.952	0.956	0.954	0.337	0.435	0.292
4	4	0.065	0.058	0.060	0.952	0.961	0.942	0.252	0.237	0.225
	6	0.078	0.063	0.068	0.963	0.953	0.952	0.308	0.246	0.263

Table 6: Uniform simulation results, averaged across 1 000 replications with $n = 5\,000$, $d = 1$.

Model	J	$\sup_{q,\boldsymbol{x}} \lvert\widehat{\mu}(\boldsymbol{x}, q) - \mu_0(\boldsymbol{x}, q)\rvert$	Uniform coverage	Av. CB width
1	4	0.151	0.950	0.399
	6	0.181	0.948	0.475
2	4	0.146	0.939	0.381
	6	0.170	0.940	0.453
3	4	0.187	0.922	0.441
	6	0.223	0.952	0.550
4	4	0.173	0.953	0.415
	6	0.201	0.947	0.494

Table 7: Pointwise simulation results for points $\{(q^{(k)}, \boldsymbol{x}^{(k)})\}_{k=1}^{3}$, averaged across 1 000 replications with $n = 5\,000$, $d = 2$, where $(q^{(1)}, q^{(2)}, q^{(3)}) = (-0.2, 0, 0.2)$, $\boldsymbol{x}^{(1)} = (0.3, 0.1)$, $\boldsymbol{x}^{(2)} = (0.1, 0.4)$ and $\boldsymbol{x}^{(3)} = (0.2, 0.2)$.

		RMSE			Coverage			CI width		
Model	J	Point 1	Point 2	Point 3	Point 1	Point 2	Point 3	Point 1	Point 2	Point 3
1	4	0.272	0.194	0.243	0.957	0.956	0.954	1.052	0.768	0.949
	6	0.455	0.256	0.366	0.944	0.957	0.935	1.417	1.027	1.328
2	4	0.273	0.182	0.237	0.957	0.934	0.941	1.059	0.662	0.904
	6	0.425	0.231	0.318	0.950	0.939	0.955	1.422	0.855	1.248
3	4	0.244	0.260	0.219	0.926	0.936	0.950	0.896	0.944	0.839
	6	0.314	0.784	0.290	0.940	0.932	0.950	1.154	1.558	1.129

Table 8: Uniform simulation results, averaged across 1 000 replications with $n = 5\,000$, $d = 2$.

Model	J	$\sup_{q,\boldsymbol{x}} \lvert \widehat{\mu}(\boldsymbol{x}, q) - \mu_0(\boldsymbol{x}, q) \rvert$	Uniform coverage	Av. CB width
1	4	0.487	0.940	1.163
	6	0.730	0.950	1.593
2	4	0.491	0.935	1.173
	6	0.732	0.944	1.608
3	4	0.478	0.950	1.172
	6	0.806	0.921	1.656

Table 9: Pointwise simulation results for points $\{(q^{(k)}, x^{(k)})\}_{k=1}^{3}$, averaged across 1 000 replications with $n = 50\,000$, $d = 1$, where $(q^{(1)}, q^{(2)}, q^{(3)}) = (-0.2, 0, 0.2)$ for Model 1 through 3, $(q^{(1)}, q^{(2)}, q^{(3)}) = (0.45, 0.6, 0.75)$ for Model 4; $(x^{(1)}, x^{(2)}, x^{(3)}) = (0.3, 0.1, 0.2)$.

		RMSE			Coverage			CI width		
Model	J	Point 1	Point 2	Point 3	Point 1	Point 2	Point 3	Point 1	Point 2	Point 3
1	4	0.027	0.024	0.023	0.947	0.950	0.956	0.106	0.091	0.091
	6	0.034	0.025	0.027	0.949	0.950	0.957	0.132	0.096	0.107
2	4	0.026	0.023	0.022	0.906	0.940	0.943	0.085	0.088	0.086
	6	0.028	0.023	0.026	0.942	0.963	0.947	0.108	0.092	0.101
3	4	0.023	0.035	0.021	0.946	0.904	0.960	0.088	0.117	0.081
	6	0.028	0.035	0.024	0.946	0.942	0.959	0.107	0.136	0.092
4	4	0.020	0.019	0.019	0.955	0.947	0.943	0.080	0.075	0.071
	6	0.024	0.019	0.022	0.954	0.951	0.950	0.097	0.078	0.083

Table 10: Uniform simulation results, averaged across 1 000 replications with $n = 50\,000$, $d = 1$.

Model	J	$\sup_{q,x} \lvert\widehat{\mu}(x,q) - \mu_0(x,q)\rvert$	Uniform coverage	Av. CB width
1	4	0.047	0.943	0.126
	6	0.057	0.953	0.150
2	4	0.054	0.697	0.120
	6	0.053	0.944	0.143
3	4	0.066	0.871	0.139
	6	0.070	0.941	0.173
4	4	0.055	0.938	0.131
	6	0.063	0.943	0.156

Table 11: Pointwise simulation results for points $\{(q^{(k)}, \boldsymbol{x}^{(k)})\}_{k=1}^{3}$, averaged across 1 000 replications with $n = 50\,000$, $d = 2$, where $(q^{(1)}, q^{(2)}, q^{(3)}) = (-0.2, 0, 0.2)$, $\boldsymbol{x}^{(1)} = (0.3, 0.1)$, $\boldsymbol{x}^{(2)} = (0.1, 0.4)$ and $\boldsymbol{x}^{(3)} = (0.2, 0.2)$.

		RMSE			Coverage			CI width		
Model	J	Point 1	Point 2	Point 3	Point 1	Point 2	Point 3	Point 1	Point 2	Point 3
1	4	0.085	0.064	0.075	0.947	0.940	0.952	0.326	0.239	0.295
	6	0.108	0.084	0.102	0.952	0.941	0.951	0.423	0.313	0.407
2	4	0.083	0.053	0.074	0.952	0.945	0.943	0.328	0.206	0.281
	6	0.107	0.067	0.095	0.951	0.949	0.959	0.424	0.263	0.385
3	4	0.072	0.076	0.068	0.953	0.941	0.944	0.278	0.291	0.262
	6	0.090	0.107	0.088	0.952	0.948	0.952	0.352	0.419	0.350

Table 12: Uniform simulation results, averaged across 1 000 replications with $n = 50\,000$, $d = 2$.

Model	J	$\sup_{q,\boldsymbol{x}} \lvert\widehat{\mu}(\boldsymbol{x}, q) - \mu_0(\boldsymbol{x}, q)\rvert$	Uniform coverage	Av. CB width
1	4	0.146	0.953	0.363
	6	0.193	0.960	0.484
2	4	0.146	0.958	0.366
	6	0.198	0.944	0.489
3	4	0.145	0.930	0.365
	6	0.198	0.950	0.494