# Supplement to “Uniform Estimation and Inference for Nonparametric Partitioning-Based M-Estimators”

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

September 8, 2024

**Abstract**

This supplement is self-contained. It presents more general theoretical results than those reported in the paper, as well as their proofs. In particular, a larger class of loss functions is allowed for, and weaker regularity conditions are employed when possible, which together enlarge the scope of our results. Furthermore, additional results not reported in the paper and their proofs are given. Some of the technical results presented may be of independent theoretical interest. Finally, some omitted details on the motivating examples are provided.

[*]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

## List of important statements

## List of lemmas

# A Introduction

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 $\big((y_i, \mathbf{x}_i)\big)_{1 \le i \le n}$ is a random sample, where $y_i \in \mathcal{Y} \subset \mathbb{R}$ is a scalar response variable, $\mathbf{x}_i$ is a $d$-dimensional covariate with values in $\mathcal{X}$. Let $\rho(\cdot, \cdot; \mathbf{q})$ be a loss function parametrized by $\mathbf{q} \in \mathcal{Q}$ (Borel-measurable in all three arguments), and let $\eta(\cdot)$ be a strictly monotonic continuously differentiable transformation function. (More detailed assumptions on $\rho(\cdot, \cdot; \mathbf{q})$ 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

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

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 $\mu_0$, and transformations thereof, employing the partitioning-based series $M$-estimator

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

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

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

is a dictionary of $K$ locally supported basis functions of order $m$ based on a quasi-uniform partition $\Delta = \{\delta_l : 1 \le l \le \bar{\kappa}\}$ 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 $\mu_0$ can be approximated by linear combinations of the basis (Assumption B.4 below and in the paper); the partition being quasi-uniform intuitively means that the largest size of a cell cannot get asymptotically bigger than the smallest one (Assumption B.3 below and in the paper). We consider large sample approximations where $d$ and $m$ are fixed constants, and $\bar{\kappa} \to \infty$ (and thus $K \to \infty$) as $n \to \infty$. Prior literature is discussed in the paper.

## A.1 Organization

Section 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.2 is weaker than the assumptions on the loss function described in Section 3 of the main paper. In particular, a complexity condition on $\{\psi(\cdot, \cdot; \mathbf{q})\}$ is absent in Section 3 because the main paper contains a stronger but simpler Assumption 3. Thus, the setup in this supplement is more general that in the main paper, and this fact is formally shown by Proposition I.1.

Section 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.

Section 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 important 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.

Section E presents our main Bahadur representation results, also categorized based on whether the objective function is convex or not. Theorem 1 in the paper corresponds to Theorem E.10 that allows for a possibly non-convex loss function. The other results in this section may be of independent theoretical and methodological interest as they require slightly different (weaker) assumptions for more special cases.

Section F develops strong approximation results using a generalized conditional Yurinskii's coupling approach. First, Section 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 [1], [4], [5], and references therein. Second, Section F.2 deploys those results to the setting of interest in our paper to verify our main result Theorem F.4, leading to Corollary F.5 that confirms Theorem 2 in the paper. That second subsection also includes other technical lemmas of potential theoretical interest (e.g., the construction of valid variance estimators).

Section G discusses results related to the implementation of uniform inference. In particular, it formally shows the validity of the plug-in approximation method and confidence bands described in the paper.

Section H discusses in detail the verification of our high-level assumptions for each of the four motivating examples in the paper.

Section I shows that the simplifying Assumption 3 imposed in the paper implies the more general assumptions imposed in this supplemental appendix.

Section J discusses other parameters of potential interest such as the level and marginal effect functions, which formalizes the claims made in Section 8 of the paper.

## A.2 Notation

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

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

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

We will denote $N(\mathcal{F}, \rho, \varepsilon)$ the $\varepsilon$-covering number of a class $\mathcal{F}$ with respect to a semi-metric $\rho$ 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 [11] 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

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

for all $0 < \varepsilon \le 1$, 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 $\xi$ is $\sigma^2$-*sub-Gaussian*, where $\sigma^2 > 0$, if

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

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

We will say that random element $\tilde{Z}$ (for example, a Gaussian process as a random element with values in a space of continuous vector-functions) is a copy of random element $Z$ if they have the same laws.

If we say the probability space is "rich enough", it means that, whenever the argument requires, we can find a random variable distributed uniformly on $[0, 1]$ independent of the data and such random variables previously used (informally, independent of everything we had before). This property is equivalent to having only one uniform random variable independent of the data (since it can be replicated: see Lemma 4.21 in [9]).

Finally, we will use the following notations:

$$\widehat{\boldsymbol{\beta}}(\mathbf{q}) := \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^K} \mathbb{E}_n[\rho(y_i, \eta(\boldsymbol{\beta}^\intercal \mathbf{p}(\mathbf{x}_i)); \mathbf{q})] \tag{A.5}$$

$$\widehat{\mu}(\mathbf{x}, \mathbf{q}) := \mathbf{p}(\mathbf{x})^\intercal \widehat{\boldsymbol{\beta}}(\mathbf{q}), \tag{A.6}$$

$$\bar{\boldsymbol{Q}}_{\mathbf{q}} := \mathbb{E}_n\Big[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\intercal \Psi_1(\mathbf{x}_i, \eta(\mu_0(\mathbf{x}_i, \mathbf{q})); \mathbf{q}) \eta^{(1)}(\mu_0(\mathbf{x}_i, \mathbf{q}))^2\Big] \tag{A.7}$$

$$\mathbf{Q}_{0,\mathbf{q}} := \mathbb{E}\Big[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\intercal \Psi_1(\mathbf{x}_i, \eta(\mu_0(\mathbf{x}_i, \mathbf{q})); \mathbf{q}) \eta^{(1)}(\mu_0(\mathbf{x}_i, \mathbf{q}))^2\Big] \tag{A.8}$$

$$\boldsymbol{\Sigma}_{0,\mathbf{q}} := \mathbb{E}\Big[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\intercal \psi(y_i, \eta(\mu_0(\mathbf{x}_i, \mathbf{q})); \mathbf{q})^2 \eta^{(1)}(\mu_0(\mathbf{x}_i, \mathbf{q}))^2\Big] \tag{A.9}$$

$$\bar{\boldsymbol{\Sigma}}_{\mathbf{q}} := \mathbb{E}_n\Big[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\intercal \sigma_{\mathbf{q}}^2(\mathbf{x}_i) \eta^{(1)}(\mu_0(\mathbf{x}_i, \mathbf{q}))^2\Big] \tag{A.10}$$

$$\boldsymbol{\Omega}_{\mathbf{v}}(\mathbf{x}, \mathbf{q}) := \mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\intercal \mathbf{Q}_{0,\mathbf{q}}^{-1} \boldsymbol{\Sigma}_{0,\mathbf{q}} \mathbf{Q}_{0,\mathbf{q}}^{-1} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) \tag{A.11}$$

$$\bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x}, \mathbf{q}) := \mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\intercal \bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} \bar{\boldsymbol{\Sigma}}_{\mathbf{q}} \bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}). \tag{A.12}$$

We usually prove a claim below its statement, and indicate the end of proof by the Q.E.D. symbol □ or with words.

Many mathematical notations are clickable and link to their definitions; for example, clicking on $\bar{\boldsymbol{\Sigma}}_{\mathbf{q}}$ should lead to (A.10).

# B 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.

For the following assumption and throughout the document, when speaking of the conditional distribution of $y_1$ given $\mathbf{x}_1$, 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.1** (Data generating process)**.**

*(i)* $\big((y_i, \mathbf{x}_i)\big)_{1 \le i \le n}$ *is a random sample satisfying* (A.1) *as described above. The random vector* $\mathbf{x}_i$ *has a Lebesgue density* $f_X(\cdot)$*, continuous and bounded away from zero on a compact support* $\mathcal{X} \subset \mathbb{R}^d$*, which is the closure of an open connected set.*

*(ii) There exists a conditional density of $y_i$ given $\mathbf{x}_i$, denoted by $f_{Y|X}(y|\mathbf{x})$, with respect to some (sigma-finite) measure $\mathfrak{M}$ on $\mathcal{Y}$. It satisfies that $\sup_{\mathbf{x}\in\mathcal{X}}\sup_{y\in\mathcal{Y}_{\mathbf{x}}} f_{Y|X}(y|\mathbf{x})<\infty$, where $\mathcal{Y}_{\mathbf{x}}$ is the support of the conditional density of $y_i$ given $\mathbf{x}_i=\mathbf{x}$.*

*(iii) $\mu_0(\cdot,\mathbf{q})$ is $m\geq 1$ times continuously differentiable. Moreover, $\mu_0(\cdot,\mathbf{q})$ and its derivatives of order no more than $m$ are bounded uniformly over $\mathbf{q}\in\mathcal{Q}$ and $\mathbf{x}\in\mathcal{X}$, and $\mu_0(\mathbf{x},\mathbf{q})$ is Lipschitz in parameter: $|\mu_0(\mathbf{x},\mathbf{q}_1)-\mu_0(\mathbf{x},\mathbf{q}_2)|\lesssim\|\mathbf{q}_1-\mathbf{q}_2\|$ with the constant in $\lesssim$ not depending on $\mathbf{x}\in\mathcal{X}$, $\mathbf{q}_1,\mathbf{q}_2\in\mathcal{Q}$.*

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

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

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

**Assumption B.2** (Loss function)**.**

*(i)* ***Piecewise Hölder weak derivative.*** *For each $\mathbf{q}\in\mathcal{Q}$ and $y\in\mathcal{Y}$, $\eta\mapsto\rho(y,\eta;\mathbf{q})$ is absolutely continuous on closed bounded intervals within $\mathcal{E}$, where $\mathcal{E}$ is an open connected subset of $\mathbb{R}$ not depending on $y$, and admits an a. e. derivative $\psi(y,\eta;\mathbf{q})$, Borel measurable as a function of $(y,\eta,\mathbf{q})$.*

*The function $\psi(y,\eta(\theta);\mathbf{q})$ can be decomposed into the product of two Borel measurable functions*

$$\psi(y,\eta(\theta);\mathbf{q})=\varphi(y,\eta(\theta);\mathbf{q})\varpi(\theta),\tag{B.2}$$

*in the following way. If $\mathfrak{M}$ in Assumption B.1(ii) is Lebesgue measure, for any $\mathbf{q}\in\mathcal{Q}$, $\mathbf{x}\in\mathcal{X}$ and a pair of points $\zeta_1,\zeta_2\in B_{\mathbf{q}}(\mathbf{x})$, we have that*

$$\begin{aligned}\sup_{y\notin[\eta(\zeta_1)\wedge\eta(\zeta_2),\eta(\zeta_1)\vee\eta(\zeta_2)]}|\varphi(y,\eta(\zeta_1);\mathbf{q})-\varphi(y,\eta(\zeta_2);\mathbf{q})|&\lesssim|\zeta_1-\zeta_2|^\alpha,\\ \sup_{y\in[\eta(\zeta_1)\wedge\eta(\zeta_2),\eta(\zeta_1)\vee\eta(\zeta_2)]}|\varphi(y,\eta(\zeta_1);\mathbf{q})-\varphi(y,\eta(\zeta_2);\mathbf{q})|&\lesssim 1;\end{aligned}\tag{B.3}$$

*otherwise (if $\mathfrak{M}$ is not Lebesgue), we have*

$$\sup_y|\varphi(y,\eta(\zeta_1);\mathbf{q})-\varphi(y,\eta(\zeta_2);\mathbf{q})|\lesssim|\zeta_1-\zeta_2|^\alpha,\tag{B.4}$$

*with constants in $\lesssim$ not depending on $\mathbf{q},\mathbf{x},\zeta_1,\zeta_2$.*

*$\varpi(\cdot)$ is continuously differentiable and either strictly positive or strictly negative. The real inverse link function $\eta(\cdot)\colon\mathbb{R}\to\mathcal{E}$ is strictly monotonic and two times continuously differentiable.*

*(ii)* ***Moments and envelope.*** *The following first-order optimality condition holds:*

$$\mathbb{E}[\psi(y_i,\eta(\mu_0(\mathbf{x}_i,\mathbf{q}));\mathbf{q})\mid\mathbf{x}_i]=0.$$

*The function*

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

*is bounded and bounded away from zero uniformly over $\mathbf{x}\in\mathcal{X}$, $\mathbf{q}\in\mathcal{Q}$, and Lipschitz in $\mathbf{q}$ uniformly in $\mathbf{x}$.*

*The family $\{\psi(y_i,\eta(\mu_0(\mathbf{x}_i,\mathbf{q}));\mathbf{q})\colon\mathbf{q}\in\mathcal{Q}\}$ has a positive measurable envelope $\overline{\psi}(\mathbf{x}_i,y_i)$ such that*

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

*(iii)* ***Conditional expectation of*** $\psi$***.*** $\Psi(\mathbf{x}, \eta; \mathbf{q}) := \mathbb{E}[\psi(y_i, \eta; \mathbf{q}) \mid \mathbf{x}_i = \mathbf{x}]$ *is twice continuously differentiable with respect to* $\eta$*. Moreover,*

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

*and*

$$\sup_{\mathbf{x}\in\mathcal{X},\mathbf{q}\in\mathcal{Q}} \sup_{\zeta\in B_{\mathbf{q}}(\mathbf{x})} |\Psi_1(\mathbf{x}, \eta(\zeta); \mathbf{q})| < \infty,$$

*where* $\Psi_1(\mathbf{x}, \eta; \mathbf{q}) := \frac{\partial}{\partial\eta}\Psi(\mathbf{x}, \eta; \mathbf{q})$*. Moreover,* $\sup_{\mathbf{x}\in\mathcal{X},\mathbf{q}\in\mathcal{Q}} \sup_{\zeta\in B_{\mathbf{q}}(\mathbf{x})}|\Psi_2(\mathbf{x}, \eta(\zeta); \mathbf{q})| < \infty$*, where* $\Psi_2(\mathbf{x}, \eta; \mathbf{q}) := \frac{\partial}{\partial\eta}\Psi_1(\mathbf{x}, \eta; \mathbf{q})$*.*

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

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

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

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

*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* $\delta$ *in no more than* $s$ *steps (following a continuous path).*

The following assumption is exactly the same as Assumption 2 in [3].

**Assumption B.3** (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* $\delta$*. Further, for* $h = 1/J = \max\{\mathrm{diam}(\delta) : \delta \in \Delta\}$*, assume* $h = o(1)$ *and* $\log(1/h) \lesssim \log(n)$*.*

The following assumption is exactly the same as Assumption 3 in [3].

**Assumption B.4** (Local basis)**.**

*(i) For each basis function $p_k$, $k = 1, \ldots, K$, the union of elements of $\Delta$ 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 $\mathbf{a} = (a_1, \ldots, a_K)^\intercal \in \mathbb{R}^K$*

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

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

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

*where $\mathrm{cl}\,\delta$ is the closure of $\delta$.*

Assumption B.4 implicitly relates the number of basis functions and the maximum mesh size: $K \asymp h^{-d} = J^d$.

The following assumption is similar but weaker than Assumption 4 in [3].

**Assumption B.5** (Approximation error)**.** *There exists a vector of coefficients $\boldsymbol{\beta}_0(\mathbf{q}) \in \mathbb{R}^K$ such that for all $\boldsymbol{\varsigma}$ satisfying $|\boldsymbol{\varsigma}| \leq \varsigma$ in Assumption B.4 we have for some positive constant $C_{\mathrm{appr}}$*

$$\sup_{\mathbf{q} \in \mathcal{Q}, \mathbf{x} \in \mathcal{X}} \left| \mu_0^{(\boldsymbol{\varsigma})}(\mathbf{x}, \mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q})^\intercal \mathbf{p}^{(\boldsymbol{\varsigma})}(\mathbf{x}; \Delta, m) \right| \leq C_{\mathrm{appr}} h^{m-|\boldsymbol{\varsigma}|}.$$

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

**Assumption B.6** (Estimator of the Gram matrix)**.** *$\widehat{\boldsymbol{Q}}_{\mathbf{q}}$ is an estimator of the matrix $\bar{\boldsymbol{Q}}_{\mathbf{q}}$ such that $\left\| \widehat{\boldsymbol{Q}}_{\mathbf{q}} - \bar{\boldsymbol{Q}}_{\mathbf{q}} \right\|_\infty \lesssim_{\mathbb{P}} h^d r_{\mathrm{Q}}$ and $\left\| \widehat{\boldsymbol{Q}}_{\mathbf{q}} \right\|_\infty \lesssim_{\mathbb{P}} h^{-d}$, where $r_{\mathrm{Q}} = o(1)$.*

**Assumption B.7** (Variance estimate)**.** *$\widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x}, \mathbf{q})$ is an estimator of the scalar function $\boldsymbol{\Omega}_{\mathbf{v}}(\mathbf{x}, \mathbf{q})$ such that*

$$\sup_{\mathbf{q} \in \mathcal{Q}} \sup_{\mathbf{x} \in \mathcal{X}} \left| \widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x}, \mathbf{q}) - \boldsymbol{\Omega}_{\mathbf{v}}(\mathbf{x}, \mathbf{q}) \right| \vee \sup_{\mathbf{q} \in \mathcal{Q}} \sup_{\mathbf{x} \in \mathcal{X}} \left| \widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x}, \mathbf{q}) - \bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x}, \mathbf{q}) \right| \lesssim_{\mathbb{P}} h^{-2|\mathbf{v}|-d} r_\Omega, \tag{B.5}$$

*where $r_\Omega = o(1)$.*

# 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 $\xi_1, \ldots, \xi_n$ be $\sigma^2$-sub-Gaussian random variables (not necessarily independent). Then*

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

*where $C_U$ is a universal constant.*

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

$$\mathbb{E}\Big[\max_{1\le i\le n} \xi_i^2\Big]^{1/2} \le \mathbb{E}\Big[\max_{1\le i\le n} \xi_i^p\Big]^{1/p} \le \Big(\sum_{i=1}^n \mathbb{E}[\xi_i^p]\Big)^{1/p} \lesssim n^{1/p}\cdot\sigma\cdot p^{1/p}\sqrt{p}$$
$$\lesssim \sigma n^{1/p}\sqrt{p} \qquad \text{using } p^{1/p} \le 2.$$

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

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

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

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

In other words, the sequence of random variables $\mathbb{P}\{|X_n| > \gamma_n r_n \mid \mathbf{D}_n\}$ 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, $\mathbf{D}_n$ a sequence of random vectors. The following are equivalent:*

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

*(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,

$$\mathbb{P}\{\mathbb{P}\{|X_n| > \varepsilon \mid \mathbf{D}_n\} > \gamma\} \le \frac{\mathbb{P}\{|X_n| > \varepsilon\}}{\gamma} \to 0,$$

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

**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*

$$\sup_{\mathbb{Q}} \log N\Big(\mathcal{F}\mathcal{G}, \|\cdot\|_{\mathbb{Q},2}, \varepsilon\|FG\|_{\mathbb{Q},2},\Big)$$
$$\le \sup_{\mathbb{Q}} \log N\Big(\mathcal{F}, \|\cdot\|_{\mathbb{Q},2}, \frac{\varepsilon\|F\|_{\mathbb{Q},2}}{2}\Big) + \sup_{\mathbb{Q}} \log N\Big(\mathcal{G}, \|\cdot\|_{\mathbb{Q},2}, \frac{\varepsilon\|G\|_{\mathbb{Q},2}}{2}\Big)$$

*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*

$$\sup_{\mathbb{Q}} \log N\Big(\mathcal{F}+\mathcal{G}, \|\cdot\|_{\mathbb{Q},2}, \varepsilon\|F+G\|_{\mathbb{Q},2}\Big)$$
$$\le \sup_{\mathbb{Q}} \log N\Big(\mathcal{F}, \|\cdot\|_{\mathbb{Q},2}, \frac{\varepsilon\|F\|_{\mathbb{Q},2}}{2}\Big) + \sup_{\mathbb{Q}} \log N\Big(\mathcal{G}, \|\cdot\|_{\mathbb{Q},2}, \frac{\varepsilon\|G\|_{\mathbb{Q},2}}{2}\Big)$$

*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, [11]. ☐

**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*

$$\mathbb{E}\left[\max_{1 \leq i \leq n} X_i\right] \leq \sigma\sqrt{2\log n},$$

$$\mathbb{E}\left[\max_{1 \leq i \leq n} |X_i|\right] \leq 2\sigma\sqrt{\log n}.$$

*If $\Sigma_1$ and $\Sigma_2$ are constant positive semi-definite $n \times n$ matrices and $N \sim \mathcal{N}(0, I_n)$, then*

$$\mathbb{E}\left[\left\|\Sigma_1^{1/2} N - \Sigma_2^{1/2} N\right\|_\infty\right] \leq 2\sqrt{\log n}\|\Sigma_1 - \Sigma_2\|_2^{1/2}.$$

*If further $\Sigma_1$ is positive definite, then*

$$\mathbb{E}\left[\left\|\Sigma_1^{1/2} N - \Sigma_2^{1/2} N\right\|_\infty\right] \leq \sqrt{\log n}\lambda_{\min}(\Sigma_1)^{-1/2}\|\Sigma_1 - \Sigma_2\|_2.$$

*Proof.* See Lemma SA31 in [4]. ☐

**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*

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

*Then we have*

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

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

*Proof.* See Lemmas SA24 and SA25 in [4]. ☐

**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*

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

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

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

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

*Proof.* This lemma is well-known. See, for example, [11]. □

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

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

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

*Proof.* See Lemma 13 in [1]. □

**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 $\varepsilon > 0$,*

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

*Proof.* See Corollary 2.1 in [7]. □

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.1(i) and B.1(ii) hold. We will suppress the dependence of $\rho(\cdot, \cdot)$ on $\mathbf{q}$ in this lemma because the result can be applied separately for each $\mathbf{q}$. 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 $\eta$ 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 $\mathbf{x} \in \mathcal{X}$ the expectation $\mathbb{E}[\psi(y_i, \zeta) \mid \mathbf{x}_i = \mathbf{x}]$ is negative for (real deterministic) $\zeta$ in a neighborhood of $a_l$, positive for $\zeta$ in a neighborhood of $a_r$, and continuous in $\zeta$ (in particular, it crosses zero).*

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

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

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

$$\mu_0(\mathbf{x}) \in \arg\min_{\zeta\in\mathbb{R}} \mathbb{E}[\rho(y_i, \eta(\zeta)) \mid \mathbf{x}_i = \mathbf{x}].$$

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

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

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

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

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 $(\eta(\zeta) - \eta(\mu_0(\mathbf{x})))\psi(y, \eta(\mu_0(\mathbf{x}))) \leq \rho(y, \eta(\zeta)) - \rho(y, \eta(\mu_0(\mathbf{x})))$, we have

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

so $\mu_0(\mathbf{x})$ 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, B.3, B.4, B.2 and B.5 hold. Then*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

is Lemma SA-2.1 in [3]. The properties (C.8) and (C.9) are not explicitly stated but follow from the proof.

In the general case, by Assumption B.2(iii) $\Psi_1(\mathbf{x}_i, \eta(\mu_0(\mathbf{x}_i, \mathbf{q})); \mathbf{q})\eta^{(1)}(\mu_0(\mathbf{x}_i, \mathbf{q}))^2$ is bounded and bounded away from zero uniformly over $i$, $n$ and $\mathbf{q}$, so (C.2) and (C.5) follow from the previous case. The additional $\Psi_1(\mathbf{x}_i, \eta(\mu_0(\mathbf{x}_i, \mathbf{q})); \mathbf{q})\eta^{(1)}(\mu_0(\mathbf{x}_i))^2$ 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

$$\left\|\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} - \mathbf{Q}_{0,\mathbf{q}}^{-1}\right\|_\infty \leq \left\|\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1}\right\|_\infty \cdot \left\|\bar{\boldsymbol{Q}}_{\mathbf{q}} - \mathbf{Q}_{0,\mathbf{q}}\right\|_\infty \cdot \left\|\mathbf{Q}_{0,\mathbf{q}}^{-1}\right\|_\infty,$$
$$\left\|\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} - \mathbf{Q}_{0,\mathbf{q}}^{-1}\right\| \leq \left\|\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1}\right\| \cdot \left\|\bar{\boldsymbol{Q}}_{\mathbf{q}} - \mathbf{Q}_{0,\mathbf{q}}\right\| \cdot \left\|\mathbf{Q}_{0,\mathbf{q}}^{-1}\right\|$$

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

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

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.2(iv), the class

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

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 $\left\{\eta^{(1)}\big(\mu_0(\mathbf{x}, \mathbf{q})\big)^2 \colon \mathbf{q} \in \mathcal{Q}\right\}$ because $\eta^{(1)}\big(\mu_0(\mathbf{x}, \mathbf{q})\big)^2$ is Lipschitz is $\mathbf{q}$, by Lemma C.4 the class

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

with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim 1$. The class $\{\mathbf{x} \mapsto p_k(\mathbf{x})p_l(\mathbf{x}) \colon 1 \leq k, l \leq K\}$ 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:

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

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

Applying Lemma C.6, we see that

$$\sup_{g \in \mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^n g(\mathbf{x}_i) - \mathbb{E}[g(\mathbf{x}_i)]\right| \lesssim_{\mathbb{P}} h^d \left(\frac{\log \frac{1}{h}}{nh^d}\right)^{1/2} + \frac{\log \frac{1}{h}}{n}$$

$$\lesssim h^d \left(\frac{\log \frac{1}{h}}{nh^d}\right)^{1/2} \qquad \text{since } \frac{\log \frac{1}{h}}{nh^d} = o(1).$$

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

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

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

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

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

Lemma C.11 is proven. □

The following Lemmas C.12 and C.13 are needed for the proof of the Bahadur representation theorems (Theorems E.1 and E.10), Corollary E.2 and a version of the consistency result (Lemma D.3).

**Lemma C.12** (Uniform convergence: variance)**.** *Suppose Assumptions B.1, B.3, B.4, B.2 and B.5 hold. If*

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

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

*then*

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

*Proof.* By Assumption B.4, $\sup_{\mathbf{x}\in\mathcal{X}}\left\|\mathbf{p}^{(\mathbf{v})}(\mathbf{x})\right\| \lesssim h^{-|\mathbf{v}|}$; by Lemma C.11, $\left\|\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1}\right\|_\infty \lesssim_{\mathbb{P}} h^{-d}$. Therefore, it is enough to show

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

Define the function class

$$\mathcal{G} := \left\{(\mathbf{x}, y) \mapsto p_l(\mathbf{x})\eta^{(1)}(\mu_0(\mathbf{x},\mathbf{q}))\psi(y,\eta(\mu_0(\mathbf{x},\mathbf{q}));\mathbf{q}) : 1 \leq l \leq K, \mathbf{q}\in\mathcal{Q}\right\}.$$

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

$$\mathcal{W}_1 := \{(\mathbf{x}, y) \mapsto p_l(\mathbf{x}) : 1 \leq l \leq K\},$$

$$\mathcal{W}_2 := \Big\{(\mathbf{x}, y) \mapsto \eta^{(1)}(\mu_0(\mathbf{x}, \mathbf{q})) : \mathbf{q} \in \mathcal{Q}\Big\},$$
$$\mathcal{W}_3 := \{(\mathbf{x}, y) \mapsto \psi(y, \eta(\mu_0(\mathbf{x}, \mathbf{q})); \mathbf{q}) : \mathbf{q} \in \mathcal{Q}\}.$$

$\mathcal{W}_1$ 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 $\mu_0(\mathbf{x}, \mathbf{q})$ is bounded uniformly over $\mathbf{x}, \mathbf{q}$ and Lipschitz in $\mathbf{q}$, $\eta^{(1)}$ on a fixed bounded interval is Lipschitz. $\mathcal{W}_3$ with envelope $\overline{\psi}(\mathbf{x}, y)$ satisfies the uniform entropy bound (A.3) with $A, V \lesssim 1$ by Assumption B.2(iv). By Lemma C.4, $\mathcal{G}$ with envelope $\overline{\psi}(\mathbf{x}, y)$ 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:

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

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

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

Finally, under (C.10)

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

and under (C.11)

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

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. □

The following lemma gives control on the projection approximation error.

**Lemma C.13** (Projection of approximation error)**.** *Suppose Assumptions B.1, B.3, B.4, B.2 and B.5 hold. If*

$$\frac{\left(\log \frac{1}{h}\right)^{\nu/(\nu-2)}}{n h^{\nu d/(\nu-2)}} = o(1), \text{ or} \tag{C.13}$$

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

*then*

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

*Proof.* Denote

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

By Assumption B.4, $\sup_{\mathbf{x}\in\mathcal{X}}\|\mathbf{p}^{(\mathbf{v})}(\mathbf{x})\|\lesssim h^{-|\mathbf{v}|}$; by Lemma C.11, $\|\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1}\|_\infty\lesssim_{\mathbb{P}} h^{-d}$. Therefore, it is enough to show

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

We will do this by showing the three bounds

$$\sup_{\mathbf{q}\in\mathcal{Q}}\|\mathbb{E}_n[\mathbf{p}(\mathbf{x}_i)A_{1,\mathbf{q}}(\mathbf{x}_i,y_i)]\|_\infty\lesssim_{\mathbb{P}} h^{\frac{d}{2}+m}\left(\frac{\log\frac{1}{h}}{n}\right)^{1/2},\tag{C.14}$$
$$\sup_{\mathbf{q}\in\mathcal{Q}}\|\mathbb{E}_n[\mathbb{E}[\mathbf{p}(\mathbf{x}_i)A_{2,\mathbf{q}}(\mathbf{x}_i,y_i)\mid\mathbf{x}_i]]\|_\infty\lesssim_{\mathbb{P}} h^{d+m},\tag{C.15}$$
$$\begin{aligned}&\sup_{\mathbf{q}\in\mathcal{Q}}\|\mathbb{E}_n[\mathbf{p}(\mathbf{x}_i)(A_{2,\mathbf{q}}(\mathbf{x}_i,y_i)-\mathbb{E}[A_{2,\mathbf{q}}(\mathbf{x}_i,y_i)\mid\mathbf{x}_i])]\|_\infty\\&\quad\lesssim_{\mathbb{P}} h^{\frac{d}{2}+(\alpha\wedge\frac{1}{2})m}\sqrt{\frac{\log\frac{1}{h}}{n}}+\frac{\log\frac{1}{h}}{n}\end{aligned}\tag{C.16}$$

To show (C.14), consider the class of functions

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

Note that $\sup_{\mathbf{q},\mathbf{x}}|\eta^{(1)}(\mu_0(\mathbf{x},\mathbf{q}))-\eta^{(1)}(\mathbf{p}(\mathbf{x})^{\mathsf{T}}\boldsymbol{\beta}_0(\mathbf{q}))|\lesssim h^m$ by Assumption B.5. (C.14) follows by the same concentration argument as in Lemma C.12.

To show (C.15), note that

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

where $\zeta$ is between $\eta(\mu_0(\mathbf{x}_i,\mathbf{q}))$ and $\eta(\mathbf{p}(\mathbf{x}_i)^\intercal\boldsymbol{\beta}_0(\mathbf{q}))$, $\tilde{\zeta}$ is between $\mu_0(\mathbf{x}_i,\mathbf{q})$ and $\mathbf{p}(\mathbf{x}_i)^\intercal\boldsymbol{\beta}_0(\mathbf{q})$. By Assumption B.2(iii) and B.5, it follows that a. s.

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

Since $\eta^{(1)}(\mathbf{p}(\mathbf{x}_i)^\intercal\boldsymbol{\beta}_0(\mathbf{q}))$ is bounded, (C.15) follows by applying Lemma C.6 to the class $\{\mathbf{x}\mapsto p_l(\mathbf{x}),1\le l\le K\}$.

It is left to show (C.16).

Consider the class of functions

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

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

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

By Assumption B.2, 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 $\{\eta^{(1)}(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}):\boldsymbol{\beta}\in\mathcal{B}_l\}$, 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 [11], $\{\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}:\boldsymbol{\beta}\in\mathcal{B}_l\}$ 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:

$$\sup_{g\in\mathcal{G}}\mathbb{E}_n[\mathbb{V}[g(\mathbf{x}_i,y_i)\mid\mathbf{x}_i]]\lesssim h^{d+(2\alpha\wedge 1)m}\quad\text{w. p. a. 1.}\tag{C.17}$$

By Assumption B.2(i), we have for any $\mathbf{x},y$

$$\begin{aligned}
&|\psi(y,\eta(\mu_0(\mathbf{x},\mathbf{q}));\mathbf{q})-\psi(y,\eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}_0(\mathbf{q}));\mathbf{q})|\\
&\quad=|\varphi(y,\eta(\mu_0(\mathbf{x},\mathbf{q}));\mathbf{q})\varpi(\mu_0(\mathbf{x},\mathbf{q}))-\varphi(y,\eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}_0(\mathbf{q}));\mathbf{q})\varpi(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}_0(\mathbf{q}))|\\
&\quad\le|\varphi(y,\eta(\mu_0(\mathbf{x},\mathbf{q}));\mathbf{q})|\cdot|\varpi(\mu_0(\mathbf{x},\mathbf{q}))-\varpi(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}_0(\mathbf{q}))|\\
&\qquad+|\varphi(y,\eta(\mu_0(\mathbf{x},\mathbf{q}));\mathbf{q})-\varphi(y,\eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}_0(\mathbf{q}));\mathbf{q})|\cdot|\varpi(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}_0(\mathbf{q}))|.
\end{aligned}$$

Recall that $\varpi(\cdot)$ in a fixed bounded interval is Lipschitz and its absolute value is bounded away from zero, which gives $|\varphi(y,\eta(\mu_0(\mathbf{x},\mathbf{q}));\mathbf{q})|\lesssim\overline{\psi}(\mathbf{x},y)$.

We will now proceed proving (C.17) under the assumption that $\mathfrak{M}$ is Lebesgue measure, so (B.3) holds; the argument under (B.4) is similar (and leads to an even stronger variance bound), so it is omitted. For $y$ outside the closed segment between $\eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}_0(\mathbf{q}))$ and $\eta(\mu_0(\mathbf{x},\mathbf{q}))$,

$$\begin{aligned}
&|\psi(y,\eta(\mu_0(\mathbf{x},\mathbf{q}));\mathbf{q})-\psi(y,\eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}_0(\mathbf{q}));\mathbf{q})|\\
&\quad\lesssim \big(\overline{\psi}(\mathbf{x},y)+1\big)\cdot|\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}_0(\mathbf{q})-\mu_0(\mathbf{x},\mathbf{q})|^\alpha\\
&\quad\lesssim \big(\overline{\psi}(\mathbf{x},y)+1\big)h^{\alpha m}, \qquad \text{(C.18)}
\end{aligned}$$

where in (C.18) we used Assumption B.5. For $y$ between $\eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}_0(\mathbf{q}))$ and $\eta(\mu_0(\mathbf{x},\mathbf{q}))$ inclusive,

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

where in (C.19) we again used Assumption B.5.

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 $\eta^{(1)}(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}_0(\mathbf{q}))$ is also bounded uniformly over $\mathbf{x}\in\mathcal{X}$, we have a. s.

$$\begin{aligned}
&\mathbb{E}\big[A_{2,\mathbf{q}}(\mathbf{x}_i,y_i)^2\;\big|\;\mathbf{x}_i\big]\\
&\quad=\mathbb{E}\big[A_{2,\mathbf{q}}(\mathbf{x}_i,y_i)^2\mathbb{1}\big\{y_i\notin\big[\eta(\mathbf{p}(\mathbf{x}_i)^\intercal\boldsymbol{\beta}_0(\mathbf{q})),\eta(\mu_0(\mathbf{x}_i,\mathbf{q}))\big]\big\}\;\big|\;\mathbf{x}_i\big]\\
&\qquad+\mathbb{E}\big[A_{2,\mathbf{q}}(\mathbf{x}_i,y_i)^2\mathbb{1}\big\{y_i\in\big[\eta(\mathbf{p}(\mathbf{x}_i)^\intercal\boldsymbol{\beta}_0(\mathbf{q})),\eta(\mu_0(\mathbf{x}_i,\mathbf{q}))\big]\big\}\;\big|\;\mathbf{x}_i\big]\\
&\quad\lesssim h^{2\alpha m}\mathbb{E}\Big[\Big(\overline{\psi}(\mathbf{x}_i,y_i)^2+1\Big)\mathbb{1}\big\{y_i\notin\big[\eta(\mathbf{p}(\mathbf{x}_i)^\intercal\boldsymbol{\beta}_0(\mathbf{q})),\eta(\mu_0(\mathbf{x}_i,\mathbf{q}))\big]\big\}\;\Big|\;\mathbf{x}_i\Big]\\
&\qquad+h^{2m}\mathbb{E}\Big[\overline{\psi}(\mathbf{x}_i,y_i)^2\mathbb{1}\big\{y_i\in\big[\eta(\mathbf{p}(\mathbf{x}_i)^\intercal\boldsymbol{\beta}_0(\mathbf{q})),\eta(\mu_0(\mathbf{x}_i,\mathbf{q}))\big]\big\}\;\Big|\;\mathbf{x}_i\Big]\\
&\qquad+\mathbb{P}\big\{y_i\in\big[\eta(\mathbf{p}(\mathbf{x}_i)^\intercal\boldsymbol{\beta}_0(\mathbf{q})),\eta(\mu_0(\mathbf{x}_i,\mathbf{q}))\big]\;\big|\;\mathbf{x}_i\big\}\\
&\quad\le h^{2\alpha m}\mathbb{E}\Big[\Big(\overline{\psi}(\mathbf{x}_i,y_i)^2+1\Big)\;\Big|\;\mathbf{x}_i\Big]+h^{2m}\mathbb{E}\Big[\overline{\psi}(\mathbf{x}_i,y_i)^2\;\Big|\;\mathbf{x}_i\Big]\\
&\qquad+\mathbb{P}\big\{y_i\in\big[\eta(\mathbf{p}(\mathbf{x}_i)^\intercal\boldsymbol{\beta}_0(\mathbf{q})),\eta(\mu_0(\mathbf{x}_i,\mathbf{q}))\big]\;\big|\;\mathbf{x}_i\big\}\\
&\quad\lesssim h^{2\alpha m}+h^{2m}+|\mathbf{p}(\mathbf{x}_i)^\intercal\boldsymbol{\beta}_0(\mathbf{q})-\mu_0(\mathbf{x}_i,\mathbf{q})| \qquad \text{(C.20)}\\
&\quad\lesssim h^{2\alpha m}+h^{2m}+h^m \qquad \text{(C.21)}\\
&\quad\lesssim h^{(2\alpha\wedge 1)m},
\end{aligned}$$

where in (C.20) we used that by Assumption B.1(ii) the conditional density of $y_i|\mathbf{x}_i$ is bounded and Assumption B.2(ii), in (C.21) we used Assumption B.5.

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

$$\begin{aligned}
&\mathbb{E}_n[\mathbb{V}[g(\mathbf{x}_i,y_i)\mid\mathbf{x}_i]]\le\mathbb{E}_n\big[\mathbb{E}\big[g(\mathbf{x}_i,y_i)^2\;\big|\;\mathbf{x}_i\big]\big]=\mathbb{E}_n\big[p_l(\mathbf{x}_i)^2\mathbb{E}\big[A_{2,\mathbf{q}}(\mathbf{x}_i,y_i)^2\;\big|\;\mathbf{x}_i\big]\big]\\
&\quad\lesssim h^{(2\alpha\wedge1)m}\mathbb{E}_n\big[p_l(\mathbf{x}_i)^2\big]\\
&\quad\le h^{(2\alpha\wedge1)m}\|\mathbb{E}_n[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\intercal]\|\\
&\quad\lesssim h^{d+(2\alpha\wedge1)m}\quad\text{w. p. a. 1}, \qquad \text{(C.22)}
\end{aligned}$$

where in (C.22) we used Lemma C.11. We have proven (C.17).

Applying Lemma C.6 conditionally on $\{\mathbf{x}_i\}_{i=1}^n$, on an event with probability approaching one, we get (C.16), and the proof of Lemma C.13 is finished. □

# 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.5 hold, $\rho(y, \eta(\theta); \mathbf{q})$ is convex with respect to $\theta$ with left or right derivative $\psi(y, \eta(\theta); \mathbf{q})\eta^{(1)}(\theta)$, and $m > d/2$. Furthermore, assume that one of the following two conditions holds:*

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

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

*Then*

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

$$\sup_{\mathbf{x}\in\mathcal{X}} \sup_{\mathbf{q}\in\mathcal{Q}} \left| \widehat{\boldsymbol{\beta}}(\mathbf{q})^\mathsf{T} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) - \mu_0^{(\mathbf{v})}(\mathbf{x}, \mathbf{q}) \right| = o_{\mathbb{P}}\left(h^{-|\mathbf{v}|}\right), \tag{D.4}$$

$$\sup_{\mathbf{q}\in\mathcal{Q}} \left( \int_{\mathcal{X}} \left( \widehat{\boldsymbol{\beta}}(\mathbf{q})^\mathsf{T} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) - \mu_0^{(\mathbf{v})}(\mathbf{x}, \mathbf{q}) \right)^2 f_X(\mathbf{x})\, \mathrm{d}\mathbf{x} \right)^{1/2} = o_{\mathbb{P}}\left(h^{d/2-|\mathbf{v}|}\right). \tag{D.5}$$

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

$$\begin{aligned}
&\left| \widehat{\boldsymbol{\beta}}(\mathbf{q})^\mathsf{T} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) - \mu_0^{(\mathbf{v})}(\mathbf{x}, \mathbf{q}) \right| \\
&\quad \le \left| \widehat{\boldsymbol{\beta}}(\mathbf{q})^\mathsf{T} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) - \boldsymbol{\beta}_0(\mathbf{q})^\mathsf{T} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) \right| + \left| \boldsymbol{\beta}_0(\mathbf{q})^\mathsf{T} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) - \mu_0^{(\mathbf{v})}(\mathbf{x}, \mathbf{q}) \right| \\
&\quad \lesssim \left\| \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) \right\|_\infty \cdot \left\| \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) \right\|_1 + \left| \boldsymbol{\beta}_0(\mathbf{q})^\mathsf{T} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) - \mu_0^{(\mathbf{v})}(\mathbf{x}, \mathbf{q}) \right| \\
&\quad \lesssim \left\| \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) \right\|_\infty \cdot h^{-|\mathbf{v}|} + \left| \boldsymbol{\beta}_0(\mathbf{q})^\mathsf{T} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) - \mu_0^{(\mathbf{v})}(\mathbf{x}, \mathbf{q}) \right| && \text{by Assumption B.4} \\
&\quad \lesssim \left\| \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) \right\|_\infty \cdot h^{-|\mathbf{v}|} + h^{m-|\mathbf{v}|} && \text{by Assumption B.5,}
\end{aligned}$$

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

$$\begin{aligned}
&\left\| \widehat{\boldsymbol{\beta}}(\mathbf{q})^\mathsf{T} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) - \mu_0^{(\mathbf{v})}(\mathbf{x}, \mathbf{q}) \right\|_{L_2(X)} \\
&\quad \le \left\| \widehat{\boldsymbol{\beta}}(\mathbf{q})^\mathsf{T} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) - \boldsymbol{\beta}_0(\mathbf{q})^\mathsf{T} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) \right\|_{L_2(X)} + \left\| \boldsymbol{\beta}_0(\mathbf{q})^\mathsf{T} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) - \mu_0^{(\mathbf{v})}(\mathbf{x}, q) \right\|_{L_2(X)} \\
&\quad \le \left\| \widehat{\boldsymbol{\beta}}(\mathbf{q})^\mathsf{T} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) - \boldsymbol{\beta}_0(\mathbf{q})^\mathsf{T} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) \right\|_{L_2(X)} + \sup_{\mathbf{x}\in\mathcal{X}} \left| \boldsymbol{\beta}_0(\mathbf{q})^\mathsf{T} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) - \mu_0^{(\mathbf{v})}(\mathbf{x}, \mathbf{q}) \right| \\
&\quad \lesssim \left\| \widehat{\boldsymbol{\beta}}(\mathbf{q})^\mathsf{T} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) - \boldsymbol{\beta}_0(\mathbf{q})^\mathsf{T} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) \right\|_{L_2(X)} + h^{m-|\mathbf{v}|} && \text{by Assumption B.5}
\end{aligned}$$

$$
\begin{aligned}
&= \left(\left(\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q})\right)^{\mathsf{T}} \mathbb{E}\left[\mathbf{p}^{(\mathbf{v})}(\mathbf{x}_i)\mathbf{p}^{(\mathbf{v})}(\mathbf{x}_i)^{\mathsf{T}}\right]\left(\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q})\right)\right)^{1/2} + h^{m-|\mathbf{v}|} \\
&\le \lambda_{\max}\left(\mathbb{E}\left[\mathbf{p}^{(\mathbf{v})}(\mathbf{x}_i)\mathbf{p}^{(\mathbf{v})}(\mathbf{x}_i)^{\mathsf{T}}\right]\right)^{1/2}\left\|\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q})\right\| + h^{m-|\mathbf{v}|} \\
&\le h^{d/2-|\mathbf{v}|} \cdot \left\|\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q})\right\| + h^{m-|\mathbf{v}|} = o_{\mathbb{P}}\left(h^{d/2-|\mathbf{v}|}\right) + h^{m-|\mathbf{v}|} = o_{\mathbb{P}}\left(h^{d/2-|\mathbf{v}|}\right).
\end{aligned}
$$

uniformly over $\mathbf{q} \in \mathcal{Q}$, where by $\|g(\mathbf{x})\|_{L_2(X)}$ we denote $\left(\int_{\mathcal{X}} g(\mathbf{x})^2 f_X(\mathbf{x})\,\mathrm{d}\mathbf{x}\right)^{1/2}$ for simplicity. In the last equality we used $m > d/2$ again. We also used that the largest eigenvalue of $\mathbb{E}\left[\mathbf{p}^{(\mathbf{v})}(\mathbf{x}_i)\mathbf{p}^{(\mathbf{v})}(\mathbf{x}_i)^{\mathsf{T}}\right]$ is bounded from above by $h^{d-2|\mathbf{v}|}$ up to a multiplicative coefficient, which is proven by the same argument as for $\mathbf{v} = 0$ in Lemma C.11 in combination with Assumption B.4.

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

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

Since $\mathbb{E}_n\left[\psi(y_i, \eta(\mathbf{p}(\mathbf{x}_i)^{\mathsf{T}}\boldsymbol{\beta}); \mathbf{q})\eta^{(1)}(\mathbf{p}(\mathbf{x}_i)^{\mathsf{T}}\boldsymbol{\beta})\mathbf{p}(\mathbf{x}_i)\right]$ is a subgradient of the convex (by Assumption B.2(i)) objective function $\mathbb{E}_n[\rho(y_i, \eta(\mathbf{p}(\mathbf{x}_i)^{\mathsf{T}}\boldsymbol{\beta}); \mathbf{q})]$ of $\boldsymbol{\beta}$, the strategy is to show that

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

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

To implement this, we will show

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

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

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

$$
\begin{aligned}
&\inf_{\mathbf{q},\boldsymbol{\alpha}} \mathbb{E}_n[\delta_{\mathbf{q},i}(\boldsymbol{\alpha})] \ge \\
&\quad \ge \inf_{\mathbf{q},\boldsymbol{\alpha}} \mathbb{E}_n[\mathbb{E}[\delta_{\mathbf{q},i}(\boldsymbol{\alpha}) \mid \{\mathbf{x}_k\}_{k=1}^n]] - \sup_{\mathbf{q},\boldsymbol{\alpha}} |\mathbb{E}_n[\delta_{\mathbf{q},i}(\boldsymbol{\alpha}) - \mathbb{E}[\delta_{\mathbf{q},i}(\boldsymbol{\alpha}) \mid \{\mathbf{x}_k\}_{k=1}^n]]| \\
&\quad \gtrsim \mathbb{E}_n[\boldsymbol{\alpha}^{\mathsf{T}}\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^{\mathsf{T}}\boldsymbol{\alpha}] + o_{\mathbb{P}}\left(h^d\right),
\end{aligned}
$$

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

We will now prove (D.7). By Assumption B.2(iii),

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

almost surely uniformly over $\mathbf{q}$, where $C$ is some positive constant not depending on $n$ or $i$, $\xi_{\mathbf{q},i}$ is between $\eta(\mathbf{p}(\mathbf{x}_i)^{\mathsf{T}}(\boldsymbol{\beta}_0(\mathbf{q}) + \gamma\boldsymbol{\alpha}))$ and $\eta(\mu_0(\mathbf{x}_i, \mathbf{q}))$, $\zeta_{\mathbf{q},i}$ between $\mathbf{p}(\mathbf{x}_i)^{\mathsf{T}}(\boldsymbol{\beta}_0(\mathbf{q}) + \gamma\boldsymbol{\alpha})$ and $\mu_0(\mathbf{x}_i, \mathbf{q})$. We used that $\Psi(\mathbf{x}_i, \eta(\mu_0(\mathbf{x}_i, \mathbf{q})); \mathbf{q}) = 0$, $\gamma$ is small enough, $\mathbf{p}(\mathbf{x})^{\mathsf{T}}\boldsymbol{\beta}_0(\mathbf{q})$ is (for large enough $n$)

uniformly close to $\mu_0(\mathbf{x}, \mathbf{q})$ by Assumption B.5 and $\eta(\cdot)$ is strictly monotonic by Assumption (B.2)(i) giving the positivity of the product $\eta^{(1)}(\zeta_{\mathbf{q},i})\eta^{(1)}(\mathbf{p}(\mathbf{x}_i)^\intercal(\boldsymbol{\beta}_0(\mathbf{q}) + \gamma\boldsymbol{\alpha}))$.

Again using the uniform approximation bound $\sup_{\mathbf{q}\in\mathcal{Q},\mathbf{x}\in\mathcal{X}}|\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}_0(\mathbf{q}) - \mu_0(\mathbf{x}, \mathbf{q})| \lesssim h^m$ by Assumption B.5, we obtain

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

since $\mathbb{E}_n[|\boldsymbol{\alpha}^\intercal\mathbf{p}(\mathbf{x}_i)|] \lesssim \mathbb{E}_n\Big[(\boldsymbol{\alpha}^\intercal\mathbf{p}(\mathbf{x}_i))^2\Big]^{1/2} \lesssim_{\mathbb{P}} h^{d/2}$ 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

$$\mathcal{G}_1 := \big\{(\mathbf{x}_i, y_i) \mapsto \delta_{\mathbf{q},i}(\boldsymbol{\alpha})\colon \boldsymbol{\alpha} \in \mathcal{S}^{K-1}, \mathbf{q} \in \mathcal{Q}\big\}.$$

By Assumption B.2(i), B.4(iii) and Assumption B.5, for $\gamma$ small enough

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

Recalling the envelope condition in Assumption B.2(ii) and that $\mu_0(\cdot, \mathbf{q})$ is bounded by Assumption B.1(iii), we see that $\sup_{g\in\mathcal{G}_1}|g| \lesssim 1 + \overline{\psi}(\mathbf{x}_i, y_i)$, which means that under (D.1)

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

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

$$\mathbb{E}\left[\max_{1\le i\le n}|\delta_{\mathbf{q},i}(\boldsymbol{\alpha})|^2 \,\middle|\, \{\mathbf{x}_k\}_{k=1}^n\right]^{1/2} \lesssim \sqrt{\log n} \quad \text{a. s.}$$

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

By Assumption B.2(iv), the class

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

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

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

has a constant envelope and is VC with index no more than $K + 2$ by Lemma 2.6.15 in [11], 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

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

and therefore of

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

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 $1 + \overline{\psi}(\mathbf{x}_i, y_i)$ 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). □

The following lemma considers the special case of unconnected basis functions.

**Lemma D.2** (Consistency: unconnected basis functions)**.** *Assume the following.*

*(i) Assumptions B.1 to B.5 hold.*

*(ii) $\rho(y, \eta(\theta); \mathbf{q})$ is convex with respect to $\theta$, and $\psi(y, \eta(\theta); \mathbf{q})\eta^{(1)}(\theta)$ is its left or right derivative.*

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

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

$$\frac{\left(\log \frac{1}{h}\right)^{\frac{\nu}{\nu-1}}}{nh^{\frac{\nu}{\nu-1}d}} = o(1), \text{ or}$$

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

*Then*

$$\sup_{\mathbf{q}\in\mathcal{Q}} \left\|\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q})\right\|_\infty = o_{\mathbb{P}}(1) \tag{D.10}$$

*and*

$$\sup_{\mathbf{x}\in\mathcal{X}} \sup_{\mathbf{q}\in\mathcal{Q}} \left|\widehat{\boldsymbol{\beta}}(\mathbf{q})^\intercal \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) - \mu_0^{(\mathbf{v})}(\mathbf{x}, \mathbf{q})\right| = o_{\mathbb{P}}\left(h^{-|\mathbf{v}|}\right). \tag{D.11}$$

*Proof.* As in Lemma D.1, (D.11) follows from (D.10).

For $l \in \{1, \dots, \bar{\kappa}\}$, the number $M_l$ of basis functions in $\mathbf{p}(\cdot)$ which are active on the $l$th cell of $\Delta$ is bounded by a constant. Denote the vector of such basis functions $\mathbf{p}_l := (p_{l,1}, \dots, p_{l,M_l})^\intercal$. Define the matrices $\mathbf{Q}_{0,\mathbf{q},l}$ and $\bar{\mathbf{Q}}_{\mathbf{q},l}$ as before with $\mathbf{p}$ replaced by $\mathbf{p}_l$ (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 $\mathbf{q} \in \mathcal{Q}$ and $l$

$$\begin{aligned} &h^d \lesssim \lambda_{\min}(\mathbf{Q}_{0,\mathbf{q},l}) \le \lambda_{\max}(\mathbf{Q}_{0,\mathbf{q},l}) \lesssim h^d, \\ &h^d \lesssim \lambda_{\min}(\bar{\mathbf{Q}}_{\mathbf{q},l}) \le \lambda_{\max}(\bar{\mathbf{Q}}_{\mathbf{q},l}) \lesssim h^d \quad \text{w. p. a. 1.} \end{aligned} \tag{D.12}$$

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

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

$$\delta_{\mathbf{q},i,l}(\boldsymbol{\alpha}_l) := \boldsymbol{\alpha}_l^\intercal \mathbf{p}_l(\mathbf{x}_i)\psi(y_i, \eta(\mathbf{p}_l(\mathbf{x}_i)^\intercal(\boldsymbol{\beta}_{0,\mathbf{q},l} + \gamma\boldsymbol{\alpha}_l)); \mathbf{q})\eta^{(1)}(\mathbf{p}_l(\mathbf{x}_i)^\intercal(\boldsymbol{\beta}_{0,\mathbf{q},l} + \gamma\boldsymbol{\alpha}_l)).$$

Proceeding in the same way as in Lemma D.1, we will show

$$\inf_{\mathbf{q},\boldsymbol{\alpha}_l,l} \mathbb{E}_n[\delta_{\mathbf{q},i,l}(\boldsymbol{\alpha}_l)] > 0 \quad \text{with probability approaching 1,} \tag{D.13}$$

which is again enough to prove the lemma by convexity. It will follow that with probability approaching one the minimizer $\widehat{\boldsymbol{\beta}}_{\mathbf{q},l}$ of $\mathbb{E}_n\left[\rho(y_i, \eta(\mathbf{p}_l(\mathbf{x}_i)^\intercal\boldsymbol{\beta}_l); \mathbf{q})\right]$ with respect to $\boldsymbol{\beta}_l$ has to lie inside the ball $\|\boldsymbol{\beta}_l-\boldsymbol{\beta}_{0,\mathbf{q},l}\|\le\gamma$, and in particular inside the cube $\|\boldsymbol{\beta}_l-\boldsymbol{\beta}_{0,\mathbf{q},l}\|_\infty\le\gamma$. But note that $\widehat{\boldsymbol{\beta}}(\mathbf{q})=\big(\widehat{\boldsymbol{\beta}}_{\mathbf{q},1}^{\mathsf{T}},\dots,\widehat{\boldsymbol{\beta}}_{\mathbf{q},\bar{k}}^{\mathsf{T}}\big)^{\mathsf{T}}$. So, with probability approaching one, for all $\mathbf{q}\in\mathcal{Q}$, we have $\big\|\widehat{\boldsymbol{\beta}}(\mathbf{q})-\boldsymbol{\beta}_0(\mathbf{q})\big\|_\infty\le\gamma$. Since $\gamma$ was arbitrary (small enough), it is equivalent to $\sup_{\mathbf{q}\in\mathcal{Q}}\big\|\widehat{\boldsymbol{\beta}}(\mathbf{q})-\boldsymbol{\beta}_0(\mathbf{q})\big\|_\infty=o_{\mathbb{P}}(1)$.

Equation (D.13) is proven analogously to the corresponding argument in Lemma D.1. The class of functions

$$\mathcal{G}_1:=\big\{(\mathbf{x}_i,y_i)\mapsto\delta_{\mathbf{q},i,l}(\boldsymbol{\alpha}_l)\colon\boldsymbol{\alpha}_l\in\mathcal{S}^{M_l-1},l\in\{1,\dots,\bar{\kappa}\},\mathbf{q}\in\mathcal{Q}\big\}$$

now satisfies the uniform entropy bound (A.3) with $A\lesssim\bar{\kappa}\asymp h^{-d}$ (since there are $\bar{\kappa}$ different values of $l$) and $V\lesssim 1$ (since the vectors $\boldsymbol{\alpha}_l$ are of bounded dimensions). The bound $\sup_{l,\boldsymbol{\alpha}_l}\mathbb{E}_n\big[|\boldsymbol{\alpha}_l^{\mathsf{T}}\mathbf{p}_l(\mathbf{x}_i)|\big]\lesssim h^d$ with probability approaching one can be proven without assuming $m>d/2$ by using

$$\sup_{l,\boldsymbol{\alpha}_l}\mathbb{E}_n\big[|\boldsymbol{\alpha}_l^{\mathsf{T}}\mathbf{p}_l(\mathbf{x}_i)|\big]\le\sup_{l,\boldsymbol{\alpha}_l}\|\boldsymbol{\alpha}_l\|_\infty\mathbb{E}_n\big[\|\mathbf{p}_l(\mathbf{x}_i)\|_1\big]\lesssim h^d\quad\text{w. p. a. 1},$$

since the dimension of $\mathbf{p}_l(\cdot)$ is uniformly bounded. □

Next, we state and prove another variant of the consistency result, only requiring

$$\frac{\big(\log\frac{1}{h}\big)^{\frac{\nu}{\nu-2}}}{nh^{\frac{\nu}{\nu-2}d}}=o(1)$$

instead of

$$\frac{\big(\log\frac{1}{h}\big)^{\frac{\nu}{\nu-1}}}{nh^{\frac{2\nu}{\nu-1}d}}=o(1).$$

The following lemma considers the special case of strongly convex and strongly smooth loss function.

**Lemma D.3** (Consistency: strongly convex and strongly smooth loss case)**.** *Assume the following conditions.*

*(i) Assumptions B.1, B.3, B.4, B.2 and B.5 hold.*

*(ii) The rate of convergence of $h$ to zero is restricted by*

$$\frac{\big(\log\frac{1}{h}\big)^{\frac{\nu}{\nu-2}}}{nh^{\frac{\nu}{\nu-2}d}}=o(1),\ \text{or}$$

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

*(iii) The function $\eta\mapsto\psi(y,\eta;\mathbf{q})$ is continuously differentiable on $\mathbb{R}$ (for all $y,\mathbf{q}$), and there exist fixed (not depending on $n$, $\mathbf{q}$ or $\theta$) numbers $\lambda,\Lambda$ such that*

$$0<\lambda\le\frac{\partial}{\partial\theta}\Big(\psi(y_i,\eta(\theta);\mathbf{q})\eta^{(1)}(\theta)\Big)\le\Lambda.$$

*Then*

$$\sup_{\mathbf{q}\in\mathcal{Q}}\Big\|\widehat{\boldsymbol{\beta}}(\mathbf{q})-\boldsymbol{\beta}_0(\mathbf{q})\Big\|_\infty=o_{\mathbb{P}}(1)\tag{D.14}$$

*and*

$$\sup_{\mathbf{x}\in\mathcal{X}}\sup_{\mathbf{q}\in\mathcal{Q}}\Big|\widehat{\boldsymbol{\beta}}(\mathbf{q})^{\mathsf{T}}\mathbf{p}^{(\mathbf{v})}(\mathbf{x})-\mu_0^{(\mathbf{v})}(\mathbf{x},\mathbf{q})\Big|=o_{\mathbb{P}}\Big(h^{-|\mathbf{v}|}\Big).\tag{D.15}$$

*Proof.* As in Lemma D.1, (D.15) follows from (D.14).

Denote for $\boldsymbol{\beta} \in \mathbb{R}^K$

$$G_n(\boldsymbol{\beta}) := \mathbb{E}_n\Big[\psi(y_i, \eta(\mathbf{p}(\mathbf{x}_i)^\intercal \boldsymbol{\beta}); \mathbf{q})\eta^{(1)}(\mathbf{p}(\mathbf{x}_i)^\intercal \boldsymbol{\beta})\mathbf{p}(\mathbf{x}_i)\Big],$$

which is the gradient of the convex (by Assumption B.2(i)) function $\mathbb{E}_n[\rho(y_i, \eta(\mathbf{p}(\mathbf{x}_i)^\intercal \boldsymbol{\beta}); \mathbf{q})]$ of $\boldsymbol{\beta}$. By definition of $\widehat{\boldsymbol{\beta}}(\mathbf{q})$ and differentiability, $G_n\Big(\widehat{\boldsymbol{\beta}}(\mathbf{q})\Big) = 0$. By the mean value theorem,

$$G_n(\boldsymbol{\beta}_0(\mathbf{q})) = G_n(\boldsymbol{\beta}_0(\mathbf{q})) - G_n\Big(\widehat{\boldsymbol{\beta}}(\mathbf{q})\Big) = \mathbb{E}_n[\mu_i \mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\intercal]\Big(\boldsymbol{\beta}_0(\mathbf{q}) - \widehat{\boldsymbol{\beta}}(\mathbf{q})\Big), \tag{D.16}$$

where

$$\mu_i := \frac{\partial}{\partial \theta}\Big(\psi(y_i, \eta(\theta); \mathbf{q})\eta^{(1)}(\theta)\Big)\Big|_{\theta = \widetilde{\theta}_i} \quad \text{for some } \widetilde{\theta}_i \text{ between } \mathbf{p}(\mathbf{x}_i)^\intercal \boldsymbol{\beta}_0(\mathbf{q}) \text{ and } \mathbf{p}(\mathbf{x}_i)^\intercal \widehat{\boldsymbol{\beta}}(\mathbf{q}).$$

By the assumption of the lemma, $0 < \lambda \leq \mu_i \leq \Lambda$. Therefore, for any vector $\boldsymbol{a} \in \mathbb{R}^K$

$$\begin{aligned}\lambda \cdot \boldsymbol{a}^\intercal \mathbb{E}_n[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\intercal]\boldsymbol{a} &\leq \boldsymbol{a}^\intercal \mathbb{E}_n[\mu_i \mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\intercal]\boldsymbol{a} \\ &= \mathbb{E}_n\Big[\mu_i (\mathbf{p}(\mathbf{x}_i)^\intercal \boldsymbol{a})^2\Big] \leq \Lambda \cdot \boldsymbol{a}^\intercal \mathbb{E}_n[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\intercal]\boldsymbol{a}.\end{aligned}$$

Moreover, the matrix

$$\mathbb{E}_n[\mu_i \mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\intercal]$$

has the same multi-banded structure as

$$\mathbb{E}_n[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\intercal].$$

That means that by the same argument as that in Lemma C.11 we have

$$\Big\|\mathbb{E}_n[\mu_i \mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\intercal]^{-1}\Big\|_\infty \lesssim_{\mathbb{P}} h^{-d}. \tag{D.17}$$

It is shown in the proofs of Lemma C.12 and Lemma C.13 that

$$\|G_n(\boldsymbol{\beta}_0(\mathbf{q}))\|_\infty \lesssim_{\mathbb{P}} \sqrt{\frac{h^d \log \frac{1}{h}}{n}} + h^{d+m} = o\Big(h^d\Big). \tag{D.18}$$

From (D.16)

$$\Big\|\boldsymbol{\beta}_0(\mathbf{q}) - \widehat{\boldsymbol{\beta}}(\mathbf{q})\Big\|_\infty \leq \Big\|\mathbb{E}_n[\mu_i \mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\intercal]^{-1}\Big\|_\infty \cdot \|G_n(\boldsymbol{\beta}_0(\mathbf{q}))\|_\infty,$$

which in combination with (D.17) and (D.18) gives

$$\Big\|\boldsymbol{\beta}_0(\mathbf{q}) - \widehat{\boldsymbol{\beta}}(\mathbf{q})\Big\|_\infty = o_{\mathbb{P}}(1)$$

uniformly over $\mathbf{q} \in \mathcal{Q}$. ∎

## D.2 Nonconvex case

Our next goal is to prove the consistency result Lemma D.5 for the nonconvex case. We will need the following lemma.

**Lemma D.4** (Preparation for consistency in the nonconvex case)**.** *Suppose Assumptions B.1, B.3 and B.4, Assumption B.2, Items (i) to (iii), and Assumption B.5 hold. Then the infinity norm of $\boldsymbol{\beta}_0(\mathbf{q})$ is bounded:*

$$\sup_{\mathbf{q}\in\mathcal{Q}} \|\boldsymbol{\beta}_0(\mathbf{q})\|_\infty \lesssim 1. \tag{D.19}$$

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

$$\sup_{\mathbf{q}\in\mathcal{Q}} \sup_{\|\boldsymbol{\beta}\|_\infty \le R} |\rho(y, \eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}); \mathbf{q}) - \rho(y, \eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}_0(\mathbf{q})); \mathbf{q})| \le C_1\big(1 + \overline{\psi}(\mathbf{x}, y)\big). \tag{D.20}$$

*Proof.* We prove (D.19) first. By Assumption B.5 ($\boldsymbol{\beta}_0(\mathbf{q})^\intercal \mathbf{p}(\mathbf{x})$ is close to $\mu_0(\mathbf{x}, \mathbf{q})$) and Assumption B.1(iii) ($\mu_0(\mathbf{x}, \mathbf{q})$ is uniformly bounded), $|\boldsymbol{\beta}_0(\mathbf{q})^\intercal \mathbf{p}(\mathbf{x})|$ is bounded uniformly over $\mathbf{q} \in \mathcal{Q}$ and $\mathbf{x} \in \mathcal{X}$. By Assumption B.4, we can bound the $k$th coordinate of $\boldsymbol{\beta}_0(\mathbf{q})$

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

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

Now we prove (D.20). Note that

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

By Assumption B.2(i), for any $\mathbf{x}$, $y$ and $t$ in the interval of integration

$$\begin{aligned}
&|\psi(y, \eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}_0(\mathbf{q}) + t); \mathbf{q}) - \psi(y, \eta(\mu_0(\mathbf{x}, \mathbf{q})); \mathbf{q})| \\
&\quad = |\varphi(y, \eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}_0(\mathbf{q}) + t); \mathbf{q}) \varpi(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}_0(\mathbf{q}) + t) - \varphi(y, \eta(\mu_0(\mathbf{x}, \mathbf{q})); \mathbf{q}) \varpi(\mu_0(\mathbf{x}, \mathbf{q}))| \\
&\quad \le |\varphi(y, \eta(\mu_0(\mathbf{x}, \mathbf{q})); \mathbf{q})| \cdot |\varpi(\mu_0(\mathbf{x}, \mathbf{q})) - \varpi(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}_0(\mathbf{q}) + t)| \\
&\qquad + |\varphi(y, \eta(\mu_0(\mathbf{x}, \mathbf{q})); \mathbf{q}) - \varphi(y, \eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}_0(\mathbf{q}) + t); \mathbf{q})| \cdot |\varpi(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}_0(\mathbf{q}) + t)|.
\end{aligned}$$

Recall that $\varpi(\cdot)$ in a fixed bounded interval is Lipschitz and its absolute value is bounded away from zero, which gives $|\varphi(y, \eta(\mu_0(\mathbf{x}, \mathbf{q})); \mathbf{q})| \lesssim \overline{\psi}(\mathbf{x}, y)$. So we have a bound

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

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

$$|\rho(y, \eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}); \mathbf{q}) - \rho(y, \eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}_0(\mathbf{q})); \mathbf{q})| \le C_2\big(1 + \overline{\psi}(\mathbf{x}, y)\big) \cdot |\mathbf{p}(\mathbf{x})^\intercal(\boldsymbol{\beta} - \boldsymbol{\beta}_0(\mathbf{q}))|$$

$$
\begin{aligned}
&\le C_2\big(1+\overline{\psi}(\mathbf{x},y)\big)\cdot\|\mathbf{p}(\mathbf{x})\|_1\cdot\|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\mathbf{q})\|_\infty\\
&\le C_2\big(1+\overline{\psi}(\mathbf{x},y)\big)\cdot\|\mathbf{p}(\mathbf{x})\|_1\cdot(\|\boldsymbol{\beta}\|_\infty+\|\boldsymbol{\beta}_0(\mathbf{q})\|_\infty)\\
&\le C_1\big(1+\overline{\psi}(\mathbf{x},y)\big),
\end{aligned}
$$

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, modulo the fact that the assumption on the form of the loss here is more general than Assumption 3 in the paper, as explained in Section 3.1 there.

**Lemma D.5** (Consistency in the nonconvex case)**.** *Assume the following conditions.*

*(i) Assumptions B.1, B.3 and B.4, Assumption B.2, Items (i) to (iii), and Assumption B.5 hold.*

*(ii)* $m > d/2$.

*(iii) The following rate condition holds:*

$$
\frac{\big(\log\frac{1}{h}\big)^{\frac{\nu}{\nu-1}}}{nh^{\frac{2\nu}{\nu-1}d}} = o(1), \text{ or} \tag{D.21}
$$

$$
\overline{\psi}(\mathbf{x}_i,y_i) \text{ is } \sigma^2\text{-sub-Gaussian conditionally on } \mathbf{x}_i \text{ and } \frac{\sqrt{\log n}\log(1/h)}{nh^{2d}} = o(1). \tag{D.22}
$$

*(iv)* $R>0$ *is a fixed number (not depending on* $n$*) such that* $\sup_{\mathbf{q}\in\mathcal{Q}}\|\boldsymbol{\beta}_0(\mathbf{q})\|_\infty \le R/2$ *(existing by Lemma D.4).*

*(v) There is a positive constant* $c$ *such that we have* $\inf \Psi_1(\mathbf{x},\zeta;\mathbf{q}) > c$*, where the infimum is over* $\mathbf{x}\in\mathcal{X}$*,* $\mathbf{q}\in\mathcal{Q}$*,* $\|\boldsymbol{\beta}\|_\infty\le R$*,* $\zeta$ *between* $\eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta})$ *and* $\eta(\mu_0(\mathbf{x},\mathbf{q}))$*.*

*(vi) The class of functions*

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

*with envelope* $C_1\big(1+\overline{\psi}(\mathbf{x},y)\big)$ *(by Lemma D.4) satisfies the uniform entropy bound (A.3) with* $A\lesssim 1$ *and* $V\lesssim K$*.*

*Then for*

$$
\widehat{\boldsymbol{\beta}}_{\mathrm{constr}}(\mathbf{q}) := \operatorname*{arg\,min}_{\|\boldsymbol{\beta}\|_\infty\le R}\mathbb{E}_n[\rho(y_i,\eta(\mathbf{p}(\mathbf{x}_i)^\intercal\boldsymbol{\beta});\mathbf{q})] \tag{D.23}
$$

*we have*

$$
\sup_{\mathbf{q}\in\mathcal{Q}}\Big\|\widehat{\boldsymbol{\beta}}_{\mathrm{constr}}(\mathbf{q})-\boldsymbol{\beta}_0(\mathbf{q})\Big\| = o_{\mathbb{P}}(1). \tag{D.24}
$$

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

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

Note that

$$
\begin{aligned}
\mathbb{E}[\delta_{\mathbf{q},i}(\boldsymbol{\beta})\mid\mathbf{x}_i] &= \int_0^{\mathbf{p}(\mathbf{x}_i)^\intercal(\boldsymbol{\beta}-\boldsymbol{\beta}_0(\mathbf{q}))}\Psi(\mathbf{x}_i,\eta(\mathbf{p}(\mathbf{x}_i)^\intercal\boldsymbol{\beta}_0(\mathbf{q})+t);\mathbf{q})\eta^{(1)}(\mathbf{p}(\mathbf{x}_i)^\intercal\boldsymbol{\beta}_0(\mathbf{q})+t)\,\mathrm{d}t\\
&\overset{\text{(a)}}{=} \int_0^{\mathbf{p}(\mathbf{x}_i)^\intercal(\boldsymbol{\beta}-\boldsymbol{\beta}_0(\mathbf{q}))}\Psi_1(\mathbf{x}_i,\xi_{\mathbf{q},i,t};\mathbf{q})
\end{aligned}
$$

$$
\begin{aligned}
&\times \eta^{(1)}(\zeta_{\mathbf{q},i,t})\eta^{(1)}(\mathbf{p}(\mathbf{x}_i)^\mathsf{T}\boldsymbol{\beta}_0(\mathbf{q})+t)\{\mathbf{p}(\mathbf{x}_i)^\mathsf{T}\boldsymbol{\beta}_0(\mathbf{q})-\mu_0(\mathbf{x}_i,\mathbf{q})+t\}\,\mathrm{d}t \\
&\overset{\text{(b)}}{\geq} C_4\{\mathbf{p}(\mathbf{x}_i)^\mathsf{T}(\boldsymbol{\beta}-\boldsymbol{\beta}_0(\mathbf{q}))\}^2 - C_3\sup_{\mathbf{q},\mathbf{x}}|\mathbf{p}(\mathbf{x})^\mathsf{T}\boldsymbol{\beta}_0(\mathbf{q})-\mu_0(\mathbf{x},\mathbf{q})|\cdot|\mathbf{p}(\mathbf{x}_i)^\mathsf{T}(\boldsymbol{\beta}-\boldsymbol{\beta}_0(\mathbf{q}))| \\
&\overset{\text{(c)}}{\geq} C_4\{\mathbf{p}(\mathbf{x}_i)^\mathsf{T}(\boldsymbol{\beta}-\boldsymbol{\beta}_0(\mathbf{q}))\}^2 - C_5h^m|\mathbf{p}(\mathbf{x}_i)^\mathsf{T}(\boldsymbol{\beta}-\boldsymbol{\beta}_0(\mathbf{q}))|,
\end{aligned}
$$

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

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

for some $\xi_{\mathbf{q},i,t}$ between $\eta(\mathbf{p}(\mathbf{x}_i)^\mathsf{T}\boldsymbol{\beta}_0(\mathbf{q})+t)$ and $\eta(\mu_0(\mathbf{x}_i,\mathbf{q}))$, and some $\zeta_{\mathbf{q},i,t}$ between $\mathbf{p}(\mathbf{x}_i)^\mathsf{T}\boldsymbol{\beta}_0(\mathbf{q})+t$ and $\mu_0(\mathbf{x}_i,\mathbf{q})$ by the mean-value theorem applied twice; in (b) we used Condition (v), Assumption B.2, in particular that $\eta(\cdot)$ is strictly monotonic giving the positivity of the product $\eta^{(1)}(\zeta_{\mathbf{q},i,t})\eta^{(1)}(\mathbf{p}(\mathbf{x}_i)^\mathsf{T}\boldsymbol{\beta}_0(\mathbf{q})+t)$; in (c) we used Assumption B.5. By Lyapunov's inequality, $\mathbb{E}_n[|\mathbf{p}(\mathbf{x}_i)^\mathsf{T}(\boldsymbol{\beta}-\boldsymbol{\beta}_0(\mathbf{q}))|] \leq \mathbb{E}_n\Big[(\mathbf{p}(\mathbf{x}_i)^\mathsf{T}(\boldsymbol{\beta}-\boldsymbol{\beta}_0(\mathbf{q})))^2\Big]^{1/2}$. We conclude

$$
\begin{aligned}
&\mathbb{E}_n[\mathbb{E}[\delta_{\mathbf{q},i}(\boldsymbol{\beta})\mid\mathbf{x}_i]] \\
&\quad\geq C_4(\boldsymbol{\beta}-\boldsymbol{\beta}_0(\mathbf{q}))^\mathsf{T}\mathbb{E}_n[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\mathsf{T}](\boldsymbol{\beta}-\boldsymbol{\beta}_0(\mathbf{q})) - C_5h^m\mathbb{E}_n\Big[(\mathbf{p}(\mathbf{x}_i)^\mathsf{T}(\boldsymbol{\beta}-\boldsymbol{\beta}_0(\mathbf{q})))^2\Big]^{1/2} \\
&\quad\overset{\text{(a)}}{\geq} C_6h^d\|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\mathbf{q})\|^2 - C_7h^{m+d/2}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0(\mathbf{q})\|
\end{aligned}
$$

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

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

because

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

Define the class of functions

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

It is a product of a subclass of the class

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

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

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

with envelope $C_1\big(1+\overline{\psi}(\mathbf{x},y)\big)$, satisfying the uniform entropy bound (A.3) with $A\lesssim 1$ and $V\lesssim K$ by the conditions of Lemma D.5.

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

Next, under (D.21)

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

with constants in $\lesssim$ depending on $R$ but not on $n$ or $\varepsilon$, and under (D.22) by Lemma C.1

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

Moreover,

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

where (a) is by Lemma C.11.

By Lemma C.6, we have

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

under (D.21) and (D.22) respectively (since $\varepsilon$ is fixed).

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

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

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 $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0(\mathbf{q})\| \le \varepsilon$ for all $\mathbf{q} \in \mathcal{Q}$ with probability approaching one. Since $\varepsilon$ was arbitrary smaller than $R/2$, it is equivalent to $\sup_{\mathbf{q}\in\mathcal{Q}} \left\| \widehat{\boldsymbol{\beta}}_{\mathrm{constr}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) \right\| = o_{\mathbb{P}}(1)$. □

The following lemma considers the special case of unconnected basis functions.

**Lemma D.6** (Consistency in the nonconvex case: unconnected basis)**.** *Assume that for all $k \in \{1, \ldots, K\}$ the $k$th basis function $p_k(\cdot)$ is only active on one of the cells of $\Delta$, and define $\mathbf{p}_l(\cdot)$, $M_l$, $\boldsymbol{\beta}_{0,\mathbf{q},l}$ as in the proof of Lemma D.2.*

*Assume the conditions of Lemma D.5 with Condition (ii) removed, Condition (iii) replaced by*

$$\frac{(\log(1/h))^{\frac{\nu}{\nu-1}}}{nh^{\frac{\nu}{\nu-1}d}} = o(1), \ \text{or} \tag{D.25}$$

$$\overline{\psi}(\mathbf{x}_i, y_i) \text{ is } \sigma^2\text{-sub-Gaussian conditionally on } \mathbf{x}_i \text{ and } \frac{\sqrt{\log n}\log(1/h)}{nh^d} = o(1), \tag{D.26}$$

*and Condition (vi) replaced by the following: the class*

$$\{(\mathbf{x}, y) \mapsto \rho(y, \eta(\mathbf{p}_l(\mathbf{x})^\mathsf{T}\boldsymbol{\beta}_l)) - \rho(y, \eta(\mathbf{p}_l(\mathbf{x})^\mathsf{T}\boldsymbol{\beta}_{0,\mathbf{q},l})) : \|\boldsymbol{\beta}_l\|_\infty \le R, \mathbf{q} \in \mathcal{Q}, l \in \{1, \ldots, \bar{\kappa}\}\}$$

*satisfies the uniform entropy bound (A.3) with $A \lesssim \bar{\kappa} \asymp h^{-d}$ and $V \lesssim 1$.*

*Then, for $\widehat{\boldsymbol{\beta}}_{\mathrm{constr}}$ defined as in (D.23), we have $\sup_{\mathbf{q}\in\mathcal{Q}} \left\| \widehat{\boldsymbol{\beta}}_{\mathrm{constr}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) \right\|_\infty = o_{\mathbb{P}}(1)$.*

*Proof.* Define matrices $\mathbf{Q}_{0,\mathbf{q},l}$ and $\bar{\mathbf{Q}}_{\mathbf{q},l}$ as in the proof of Lemma D.2, 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.12) holds.

For any $M_l$-dimensional vector $\boldsymbol{\beta}_l$ satisfying the constraint $\|\boldsymbol{\beta}_l\|_\infty \le R$, define

$$\begin{aligned}\delta_{\mathbf{q},i,l}(\boldsymbol{\beta}_l) &:= \rho(y_i, \eta(\mathbf{p}_l(\mathbf{x}_i)^\mathsf{T}\boldsymbol{\beta}_l); \mathbf{q}) - \rho(y_i, \eta(\mathbf{p}_l(\mathbf{x}_i)^\mathsf{T}\boldsymbol{\beta}_{0,\mathbf{q},l}); \mathbf{q}) \\ &= \int_0^{\mathbf{p}_l(\mathbf{x}_i)^\mathsf{T}(\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\mathbf{q},l})} \psi(y_i, \eta(\mathbf{p}_l(\mathbf{x}_i)^\mathsf{T}\boldsymbol{\beta}_{0,\mathbf{q},l} + t); \mathbf{q})\eta^{(1)}(\mathbf{p}_l(\mathbf{x}_i)^\mathsf{T}\boldsymbol{\beta}_{0,\mathbf{q},l} + t)\,\mathrm{d}t.\end{aligned}$$

By the same argument as in the proof of Lemma D.5, we have

$$\begin{aligned}&\mathbb{E}_n\left[\mathbb{E}[\delta_{\mathbf{q},i,l}(\boldsymbol{\beta}_l)|\mathbf{x}_i]\right] \\ &\quad\ge C_8(\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\mathbf{q},l})^\mathsf{T}\mathbb{E}_n[\mathbf{p}_l(\mathbf{x}_i)\mathbf{p}_l(\mathbf{x}_i)^\mathsf{T}](\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\mathbf{q},l}) - C_9 h^m \mathbb{E}_n\left[\left|\mathbf{p}_l(\mathbf{x}_i)^\mathsf{T}(\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\mathbf{q},l})\right|\right] \\ &\quad\overset{\text{(a)}}{\ge} C_8 h^d \|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\mathbf{q},l}\|^2 - C_{10} h^{m+d}\|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\mathbf{q},l}\|\end{aligned}$$

with probability approaching one for some positive constants $C_8$, $C_9$ and $C_{10}$ (depending on $R$, but not on $\mathbf{q}$ or $l$), where in (a) we used $\mathbb{E}_n\left[\left|\mathbf{p}_l(\mathbf{x}_i)^\mathsf{T}(\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\mathbf{q},l})\right|\right] \le \|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\mathbf{q},l}\|_\infty \mathbb{E}_n\left[\|\mathbf{p}_l(\mathbf{x}_i)\|_1\right] \lesssim \|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\mathbf{q},l}\|_\infty h^d \le \|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\mathbf{q},l}\|_\infty h^d$ with probability approaching one since the dimension of $\mathbf{p}_l(\cdot)$ is bounded.

Next, proceeding with the same concentration argument as in Lemma D.5, we will obtain that with probability approaching one for all $l \in \{1, \ldots, \bar{\kappa}\}$, $\mathbf{q} \in \mathcal{Q}$, $\|\boldsymbol{\beta}_l\|_\infty \le R$, $\|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\mathbf{q},l}\| > \varepsilon$,

$$\begin{aligned}\mathbb{E}_n[\delta_{\mathbf{q},i,l}(\boldsymbol{\beta}_l)] &\ge C_8 h^d \|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\mathbf{q},l}\|^2 - C_{10} h^{m+d}\|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\mathbf{q},l}\| + \|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\mathbf{q},l}\| \cdot o(h^d) \\ &= \|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\mathbf{q},l}\| \cdot \left\{C_8 h^d \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(\mathbf{q})\| - C_{10} h^{m+d} + o(h^d)\right\} \\ &> \|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\mathbf{q},l}\| \cdot \left\{C_8 \varepsilon h^d - C_{10} h^{m+d} + o(h^d)\right\}\end{aligned}$$

$$\overset{(a)}{=} \|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\mathbf{q},l}\| \cdot h^d\{C_8\varepsilon + o(1)\} > 0.$$

It follows that the constrained minimizer $\widehat{\boldsymbol{\beta}}_{\mathbf{q},\mathrm{constr},l}$ of $\mathbb{E}_n\big[\rho(y_i, \eta(\mathbf{p}_l(\mathbf{x}_i)^\intercal\boldsymbol{\beta}_l); \mathbf{q})\big]$ with respect to $\boldsymbol{\beta}_l$ under the constraint $\|\boldsymbol{\beta}_l\|_\infty \le R$ has to lie inside the ball $\|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\mathbf{q},l}\| \le \varepsilon$, and in particular inside the cube $\|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0,\mathbf{q},l}\|_\infty \le \varepsilon$. But this optimization can be solved separately for all $l$, i. e. $\widehat{\boldsymbol{\beta}}_{\mathrm{constr}}(\mathbf{q}) = \big(\widehat{\boldsymbol{\beta}}_{\mathbf{q},\mathrm{constr},1}, \dots, \widehat{\boldsymbol{\beta}}_{\mathbf{q},\mathrm{constr},\bar{k}}\big)^\intercal$. So, with probability approaching one, for all $\mathbf{q} \in \mathcal{Q}$, we have $\big\|\widehat{\boldsymbol{\beta}}_{\mathrm{constr}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q})\big\|_\infty \le \varepsilon$. Since $\varepsilon$ was arbitrary smaller than $R/2$, it is equivalent to $\sup_{\mathbf{q}\in\mathcal{Q}}\big\|\widehat{\boldsymbol{\beta}}_{\mathrm{constr}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q})\big\|_\infty = o_{\mathbb{P}}(1)$. □

# E Bahadur representation

The main purpose of this section is to prove two versions of the Bahadur representation theorem: Theorem E.1 and Theorem E.10. In the first variant, we study the case where $\theta \mapsto \rho(y, \eta(\theta); \mathbf{q})$ is convex, and in the second one we consider the general case allowing for a possibly nonconvex loss function. The reason we have two versions is that consistency results are different in the two cases (see Section D in the paper and Section D here), and consistency is a prerequisite for our Bahadur representation results. In addition, some of the technical assumptions on the form of the loss are different, though this does not affect the simplified setup considered in the paper. Since Theorem E.10 is essentially more general, Theorem 1 in the main paper corresponds to that version.

In addition to proving the theorems, we provide a result on the convergence rates, Corollary E.2, as an immediate corollary. We state it only for the setting of Theorem E.1 to avoid repetition but of course the analogous result holds for the general case, and the argument is the same. This proves the claims in Section 5.1 of the paper.

We remind the reader that in the supplement the assumption on the form of the loss is more general than Assumption 3 in the main paper, as discussed in Section 3.1 there.

## E.1 Convex case

**Theorem E.1** (Bahadur representation: convex case)**.**

*(a) Suppose Assumptions B.1 to B.5 hold, $\rho(y, \eta(\theta); \mathbf{q})$ is convex with respect to $\theta$ with left or right derivative $\psi(y, \eta(\theta); \mathbf{q})\eta^{(1)}(\theta)$, and $\widehat{\boldsymbol{\beta}}(\mathbf{q})$ is a consistent estimator of $\boldsymbol{\beta}_0(\mathbf{q})$, that is,*

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

*In addition, suppose*

$$\begin{aligned} &\frac{\log^{d+2} n}{nh^d} = o(1) \text{ and} \\ &\text{either } \frac{\big(h^{-1}\log n\big)^{\frac{\nu}{\nu-2}d}}{n} = o(1), \text{ or } \overline{\psi}(\mathbf{x}_i, y_i) \text{ is } \sigma^2\text{-sub-Gaussian conditionally on } \mathbf{x}_i. \end{aligned} \tag{E.1}$$

*Then*

$$\begin{aligned} \sup_{\mathbf{q}\in\mathcal{Q},\mathbf{x}\in\mathcal{X}}\Big|&\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\intercal\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \mu_0^{(\mathbf{v})}(\mathbf{x}, \mathbf{q}) \\ &+\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\intercal\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1}\mathbb{E}_n\Big[\mathbf{p}(\mathbf{x}_i)\eta^{(1)}(\mu_0(\mathbf{x}_i, \mathbf{q}))\psi(y_i, \eta(\mu_0(\mathbf{x}_i, \mathbf{q})); \mathbf{q})\Big]\Big| \\ &\lesssim_{\mathbb{P}} h^{-|\mathbf{v}|}r_{\mathrm{Bah}} \end{aligned} \tag{E.2}$$

*with*

$$r_{\mathrm{Bah}} := \left(\frac{\log^d n}{nh^d}\right)^{\frac{1}{2}+\left(\frac{\alpha}{2}\wedge\frac{1}{4}\right)} \log n + h^{\left(\alpha\wedge\frac{1}{2}\right)m}\left(\frac{\log^{d+1} n}{nh^d}\right)^{1/2} + h^m. \tag{E.3}$$

*(b) If, in addition to the previous conditions,* (B.4) *holds (without any restrictions on $y$), then*

$$\begin{aligned}\sup_{\mathbf{q}\in\mathcal{Q},\mathbf{x}\in\mathcal{X}}\Big|&\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\intercal\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \mu_0^{(\mathbf{v})}(\mathbf{x},\mathbf{q})\\ &+\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\intercal\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1}\mathbb{E}_n\Big[\mathbf{p}(\mathbf{x}_i)\eta^{(1)}(\mu_0(\mathbf{x}_i,\mathbf{q}))\psi(y_i,\eta(\mu_0(\mathbf{x}_i,\mathbf{q}));\mathbf{q})\Big]\Big|\\ &\lesssim_{\mathbb{P}} h^{-|\mathbf{v}|}\left\{\left(\frac{\log^d n}{nh^d}\right)^{\frac{1+\alpha}{2}}\log n + h^{\alpha m}\left(\frac{\log^{d+1} n}{nh^d}\right)^{1/2} + h^m\right\}.\end{aligned} \tag{E.4}$$

We prove this theorem below in this section. First, we state and prove the following corollary.

**Corollary E.2** (Rates of convergence)**.**
*(a) If the conditions of Theorem* E.1(a) *hold, then*

$$\sup_{\mathbf{q}\in\mathcal{Q},\mathbf{x}\in\mathcal{X}}\left|\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\intercal\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \mu_0^{(\mathbf{v})}(\mathbf{x},\mathbf{q})\right| \lesssim_{\mathbb{P}} h^{-|\mathbf{v}|}r_{\mathrm{uc}} := h^{-|\mathbf{v}|}\left[\left(\frac{\log^d n}{nh^d}\right)^{1/2}\log n + h^m\right] \tag{E.5}$$

*and, as a consequence,*

$$\sup_{\mathbf{q}\in\mathcal{Q}}\left(\int_{\mathcal{X}}\left|\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\intercal\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \mu_0^{(\mathbf{v})}(\mathbf{x},\mathbf{q})\right|^2 f_X(\mathbf{x})\,\mathrm{d}\mathbf{x}\right)^{1/2} \lesssim_{\mathbb{P}} h^{-|\mathbf{v}|}r_{\mathrm{uc}}. \tag{E.6}$$

*(b) If the conditions of Theorem* E.1(a) *hold and*

$$(\log n)^{(d+1)/(\alpha\wedge(1/2))+d} = O\Big(nh^d\Big), \quad h^{(\alpha\wedge(1/2))m}\log^{d/2} n = O(1), \tag{E.7}$$

*then*

$$\sup_{\mathbf{q}\in\mathcal{Q},\mathbf{x}\in\mathcal{X}}\left|\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\intercal\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \mu_0^{(\mathbf{v})}(\mathbf{x},\mathbf{q})\right| \lesssim_{\mathbb{P}} h^{-|\mathbf{v}|}\left[\left(\frac{\log n}{nh^d}\right)^{1/2} + h^m\right]. \tag{E.8}$$

*(c) Finally, if the conditions of Theorem* E.1(a) *hold and*

$$(\log n)^{(d+2)/(\alpha\wedge(1/2))+d} = o\Big(nh^d\Big), \quad h^{(\alpha\wedge(1/2))m}\log^{(d+1)/2} n = o(1), \tag{E.9}$$

*then*

$$\sup_{\mathbf{q}\in\mathcal{Q}}\left(\int_{\mathcal{X}}\left|\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\intercal\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \mu_0^{(\mathbf{v})}(\mathbf{x},\mathbf{q})\right|^2 f_X(\mathbf{x})\,\mathrm{d}\mathbf{x}\right)^{1/2} \lesssim_{\mathbb{P}} h^{-|\mathbf{v}|}\left[\frac{1}{\sqrt{nh^d}} + h^m\right]. \tag{E.10}$$

*Proof.* By Theorem E.1, Lemma C.12 and triangle inequality,

$$\begin{aligned}&\sup_{\mathbf{q}\in\mathcal{Q},\mathbf{x}\in\mathcal{X}}\left|\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\intercal\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \mu_0^{(\mathbf{v})}(\mathbf{x},\mathbf{q})\right|\\ &\quad\lesssim_{\mathbb{P}} h^{-|\mathbf{v}|}\left[\left(\frac{\log(1/h)}{nh^d}\right)^{\frac{1}{2}} + \left(\frac{\log^d n}{nh^d}\right)^{\frac{1}{2}+\frac{\alpha}{2}\wedge\frac{1}{4}}\log n + h^{\left(\alpha\wedge\frac{1}{2}\right)m}\left(\frac{\log^{d+1} n}{nh^d}\right)^{1/2} + h^m\right].\end{aligned}$$

Using $\log(1/h) \lesssim \log n$ and simplifying the right-hand size, we obtain (E.5), and (E.6) follows immediately since the density $f_X$ is bounded. Additional restrictions (E.7) allow us to get a slightly stronger result (E.8).

To prove (E.10), note that

$$\begin{aligned}
&\sup_{\mathbf{q}} \int_{\mathcal{X}} \left| \mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\mathsf{T} \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\mathsf{T} \boldsymbol{\beta}_0(\mathbf{q}) \right|^2 f_X(\mathbf{x})\, \mathrm{d}\mathbf{x} \\
&\quad = \sup_{\mathbf{q}} \left( \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) \right)^\mathsf{T} \mathbb{E}\left[ \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) \mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\mathsf{T} \right] \left( \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) \right) \overset{\text{(a)}}{\lesssim} h^{d-2|\mathbf{v}|} \left\| \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) \right\|^2,
\end{aligned}$$

where inequality (a) is true because the largest eigenvalue of $\mathbb{E}\left[\mathbf{p}^{(\mathbf{v})}(\mathbf{x}_i)\mathbf{p}^{(\mathbf{v})}(\mathbf{x}_i)^\mathsf{T}\right]$ is bounded from above by $h^{d-2|\mathbf{v}|}$ up to a multiplicative coefficient, which is proven by the same argument as for $\mathbf{v} = 0$ in Lemma C.11 in combination with Assumption B.4.

It is left to prove

$$\left\| \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) \right\| \lesssim_{\mathbb{P}} \frac{1}{h^d \sqrt{n}}. \tag{E.11}$$

By the triangle inequality,

$$\begin{aligned}
\left\| \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) \right\| &\le \left\| \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) + \bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} \mathbb{E}_n\left[ \mathbf{p}(\mathbf{x}_i) \eta^{(1)}(\mu_0(\mathbf{x}_i, \mathbf{q})) \psi(y_i, \eta(\mu_0(\mathbf{x}_i, \mathbf{q})); \mathbf{q}) \right] \right\| \\
&\quad + \left\| \bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} \mathbb{E}_n\left[ \mathbf{p}(\mathbf{x}_i) \eta^{(1)}(\mu_0(\mathbf{x}_i, \mathbf{q})) \psi(y_i, \eta(\mu_0(\mathbf{x}_i, \mathbf{q})); \mathbf{q}) \right] \right\|.
\end{aligned} \tag{E.12}$$

To bound the second term in (E.12) on the right-hand side, consider the expectation

$$\begin{aligned}
&\mathbb{E}\left[ \left\| \mathbb{E}_n\left[ \mathbf{p}(\mathbf{x}_i) \eta^{(1)}(\mu_0(\mathbf{x}_i, \mathbf{q})) \psi(y_i, \eta(\mu_0(\mathbf{x}_i, \mathbf{q})); \mathbf{q}) \right] \right\|^2 \right] \\
&\quad = \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}\left[ \eta^{(1)}(\mu_0(\mathbf{x}_i, \mathbf{q}))^2 \psi(y_i, \eta(\mu_0(\mathbf{x}_i, \mathbf{q})); \mathbf{q})^2 \|\mathbf{p}(\mathbf{x}_i)\|^2 \right] \\
&\quad = \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}\left[ \sigma_{\mathbf{q}}^2(\mathbf{x}_i) \eta^{(1)}(\mu_0(\mathbf{x}_i, \mathbf{q}))^2 \|\mathbf{p}(\mathbf{x}_i)\|^2 \right] \overset{\text{(a)}}{\lesssim} \frac{1}{n^2} \sum_{i=1} 1 = \frac{1}{n},
\end{aligned}$$

where in (a) we used uniform boundedness of $\sigma_{\mathbf{q}}^2(\mathbf{x})$ by Assumption B.2(ii), uniform boundedness of $\mu_0(\mathbf{x}, \mathbf{q})$ and $\|\mathbf{p}(\mathbf{x})\|$. By Markov's inequality and Lemma C.11, this immediately implies

$$\left\| \bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} \mathbb{E}_n\left[ \mathbf{p}(\mathbf{x}_i) \eta^{(1)}(\mu_0(\mathbf{x}_i, \mathbf{q})) \psi(y_i, \eta(\mu_0(\mathbf{x}_i, \mathbf{q})); \mathbf{q}) \right] \right\| \lesssim_{\mathbb{P}} \|\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1}\| \cdot \frac{1}{\sqrt{n}} \lesssim \frac{1}{h^d \sqrt{n}}.$$

Concerning the first term in (E.12), it is proven in Theorem E.1 that

$$\left\| \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) + \bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} \mathbb{E}_n\left[ \mathbf{p}(\mathbf{x}_i) \eta^{(1)}(\mu_0(\mathbf{x}_i, \mathbf{q})) \psi(y_i, \eta(\mu_0(\mathbf{x}_i, \mathbf{q})); \mathbf{q}) \right] \right\|_\infty \lesssim_{\mathbb{P}} r_{\text{Bah}}$$

for $r_{\text{Bah}}$ defined in (E.3), so that

$$\begin{aligned}
&\left\| \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) + \bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} \mathbb{E}_n\left[ \mathbf{p}(\mathbf{x}_i) \eta^{(1)}(\mu_0(\mathbf{x}_i, \mathbf{q})) \psi(y_i, \eta(\mu_0(\mathbf{x}_i, \mathbf{q})); \mathbf{q}) \right] \right\| \\
&\quad \lesssim_{\mathbb{P}} \sqrt{K} r_{\text{Bah}} \lesssim \frac{r_{\text{Bah}}}{h^{d/2}} \overset{\text{(a)}}{=} o\left( \frac{1}{h^d \sqrt{n}} \right),
\end{aligned}$$

where in equality (a) we used $r_{\text{Bah}} = o\left(1/\sqrt{nh^d}\right)$ under the assumptions. This concludes the proof of (E.11). $\square$

We will now prove Theorem E.1.

We only show the first part, since the argument for the second is very similar with minor changes in obvious places.

Denote $\mathbf{p}_i := \mathbf{p}(\mathbf{x}_i)$ for simplicity. Next, define

$$\begin{aligned}\mathcal{V} = \Big\{&\boldsymbol{v} \in \mathbb{R}^K : \exists\ d\text{-dimensional multi-index } \boldsymbol{k},\\ &|v_{\boldsymbol{\ell},l}| \le \varrho^{\|\boldsymbol{k}-\boldsymbol{\ell}\|_\infty}\varepsilon_n \text{ for } \|\boldsymbol{k}-\boldsymbol{\ell}\|_\infty \le M_n \text{ and } v_{\boldsymbol{\ell},l} = 0 \text{ otherwise}\Big\}\\ \mathcal{H}_l = \big\{&\mathbf{v} \in \mathbb{R}^K : \|\mathbf{v}\|_\infty \le r_{l,n}\big\} \text{ for } l = 1, 2,\end{aligned}$$

where $\varrho$ is the constant from Lemma C.11 and the $v_{\boldsymbol{\ell},l}$ notation is as in the bound (C.9), and putting $\beta := \alpha \wedge (1/2)$ to shorten the notations

$$\begin{aligned} r_{1,n} &:= \left[\left(\frac{\log^d n}{nh^d}\right)^{1/2} + h^m\right]\gamma_n^{1/(1+\beta)},\\ \mathfrak{r}_{2,n} &:= r_{\mathrm{Bah}}\gamma_n,\\ r_{2,n} &:= \mathfrak{z}\mathfrak{r}_{2,n},\\ \varepsilon_n &:= \mathfrak{z}'\mathfrak{r}_{2,n}\end{aligned} \tag{E.13}$$

for $\mathfrak{z}, \mathfrak{z}' > 0$ and $M_n = c_1 \log n$, and $\gamma_n \to \infty$ a positive sequence such that $r_{1,n} + \mathfrak{r}_{2,n} = o(1)$. In the last step of the proof, we will consider $\mathfrak{z} = 2^\ell, \ell = L, L+1, \dots, \bar{L}$ where $\bar{L}$ is the smallest integer such that $2^{\bar{L}}\mathfrak{r}_{2,n} \ge c$ for some sufficiently small $c > 0$, and $\mathfrak{z}' = 2^L$ for a large enough constant integer $L$.

Define for $\boldsymbol{\beta}_1 \in \mathcal{H}_1$, $\boldsymbol{\beta}_2 \in \mathcal{H}_2$ and $\boldsymbol{v} \in \mathcal{V}$

$$\begin{aligned}\delta_{\mathbf{q},i}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{v}) := \boldsymbol{v}^\intercal \mathbf{p}_i[&\psi(y_i, \eta(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{v})); \mathbf{q}) - \psi(y_i, \eta(\mathbf{p}_i^\intercal \boldsymbol{\beta}_0(\mathbf{q})); \mathbf{q})]\\ \times\, &\eta^{(1)}(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{v})).\end{aligned}$$

Note that $\delta_{\mathbf{q},i}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{v}) \ne 0$ only if $\boldsymbol{v}^\intercal \mathbf{p}_i \ne 0$. For each $\boldsymbol{v} \in \mathcal{V}$, let $\mathcal{J}_{\boldsymbol{v}} := \{j : v_j \ne 0\}$. By construction, the cardinality of $\mathcal{J}_{\boldsymbol{v}}$ is bounded by $(2M_n + 1)^d$. We have $\delta_{\mathbf{q},i}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{v}) \ne 0$ only if $p_j(\mathbf{x}_i) \ne 0$ for some $j \in \mathcal{J}_{\boldsymbol{v}}$, which happens only if $\mathbf{x}_i \in \mathcal{I}_{\boldsymbol{v}}$ where

$$\mathcal{I}_{\boldsymbol{v}} := \bigcup\{\delta \in \Delta\colon \delta \cap \operatorname{supp} p_j \ne \varnothing \text{ for some } j \in \mathcal{J}_{\boldsymbol{v}}\}.$$

$\mathcal{I}_{\boldsymbol{v}}$ includes at most $c_2 M_n^d$ cells. Moreover, at most $c_3 M_n^d$ basis functions in $\mathbf{p}$ have supports overlapping with $\mathcal{I}_{\boldsymbol{v}}$. Denote the set of indices of such basis functions by $\bar{\mathcal{J}}_{\boldsymbol{v}}$. Based on the above observations, we have $\delta_{\mathbf{q},i}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{v}) = \delta_{\mathbf{q},i}\big(\boldsymbol{\beta}_{1,\bar{\mathcal{J}}_{\boldsymbol{v}}}, \boldsymbol{\beta}_{2,\bar{\mathcal{J}}_{\boldsymbol{v}}}, \boldsymbol{v}\big)$, where

$$\begin{aligned}&\delta_{\mathbf{q},i}\big(\boldsymbol{\beta}_{1,\bar{\mathcal{J}}_{\boldsymbol{v}}}, \boldsymbol{\beta}_{2,\bar{\mathcal{J}}_{\boldsymbol{v}}}, \boldsymbol{v}\big)\\ &\quad:= \sum_{j \in \mathcal{J}_{\boldsymbol{v}}} p_{i,j} v_j \left[\psi\left(y_i, \eta\left(\sum_{l \in \bar{\mathcal{J}}_{\boldsymbol{v}}} p_{i,l}(\beta_{0,\mathbf{q},l} + \beta_{1,l} + \beta_{2,l}) - \sum_{j \in \mathcal{J}_{\boldsymbol{v}}} p_{i,j} v_j\right); \mathbf{q}\right)\right.\\ &\qquad\left. - \psi\left(y_i, \eta\left(\sum_{l \in \bar{\mathcal{J}}_{\boldsymbol{v}}} p_{i,l}\beta_{0,\mathbf{q},l}\right); \mathbf{q}\right)\right]\\ &\qquad \times \eta^{(1)}\left(\sum_{l \in \bar{\mathcal{J}}_{\boldsymbol{v}}} p_{i,l}(\beta_{0,\mathbf{q},l} + \beta_{1,l} + \beta_{2,l}) - \sum_{j \in \mathcal{J}_{\boldsymbol{v}}} p_{i,j} v_j\right) \mathbb{1}\{\mathbf{x}_i \in \mathcal{I}_{\boldsymbol{v}}\}.\end{aligned} \tag{E.14}$$

Accordingly, for $\widetilde{\boldsymbol{\beta}}_1 \in \mathbb{R}^{c_3 M_n^d}, \widetilde{\boldsymbol{\beta}}_2 \in \mathbb{R}^{c_3 M_n^d}$, define the following function class

$$\mathcal{G} := \left\{ (\mathbf{x}_i, y_i) \mapsto \delta_{\mathbf{q},i}\left(\widetilde{\boldsymbol{\beta}}_1, \widetilde{\boldsymbol{\beta}}_2, \boldsymbol{v}\right) : \mathbf{q} \in \mathcal{Q}, \boldsymbol{v} \in \mathcal{V}, \left\|\widetilde{\boldsymbol{\beta}}_1\right\|_\infty \leq r_{1,n}, \left\|\widetilde{\boldsymbol{\beta}}_2\right\|_\infty \leq r_{2,n} \right\}.$$

Also denote

$$\bar{\boldsymbol{\beta}}_{\mathbf{q}} := -\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} \mathbb{E}_n\Big[\mathbf{p}_i \eta^{(1)}(\mathbf{p}_i^\mathsf{T} \boldsymbol{\beta}_0(\mathbf{q})) \psi(y_i, \eta(\mathbf{p}_i^\mathsf{T} \boldsymbol{\beta}_0(\mathbf{q})); \mathbf{q})\Big].$$

By Lemmas C.12 and C.13, using

$$\left(\frac{\log^d n}{nh^d}\right)^{1/2} + h^m = o(r_{1,n}), \tag{E.15}$$

we have with probability approaching one

$$\sup_{\mathbf{q} \in \mathcal{Q}} \left\|\bar{\boldsymbol{\beta}}_{\mathbf{q}}\right\|_\infty \leq r_{1,n}. \tag{E.16}$$

**Step 1** We bound $\sup_{g \in \mathcal{G}} |\mathbb{E}_n[g(\mathbf{x}_i, y_i)] - \mathbb{E}[g(\mathbf{x}_i, y_i) \mathbf{x}_i]|$ in this step.

**Lemma E.3** (Bonding variance)**.** *There exists a constant $C_{11} > 0$ such that the probability of the event*

$$\mathcal{A}_1 := \left\{ \sup_{\delta \in \Delta} \mathbb{E}_n[\mathbb{1}(\mathbf{x}_i \in \delta)] \leq C_{11} h^d \right\}$$

*converges to one. On $\mathcal{A}_1$, class $\mathcal{G}$ satisfies the following variance bound:*

$$\sup_{g \in \mathcal{G}} \mathbb{E}_n[\mathbb{V}[g(\mathbf{x}_i, y_i) \mid \mathbf{x}_i]] \lesssim \varepsilon_n^2 h^d (r_{1,n} + r_{2,n} + \varepsilon_n)^{(2\alpha) \wedge 1}.$$

*Proof of Lemma E.3.* The first sentence follows from Assumption B.3 and the same concentration argument as in Lemma D.1.

We will now proceed under the assumption that $\mathfrak{M}$ is Lebesgue measure, so (B.3) holds; the argument under (B.4) 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.13, using $\left|\psi(y, \eta(p(\mathbf{x})^\mathsf{T} \boldsymbol{\beta}_0(\mathbf{q})); \mathbf{q})\right| \lesssim \overline{\psi}(\mathbf{x}, y) + 1$, for $y_i$ *outside* the closed segment

$$\text{between } \eta\left(\sum_{l \in \bar{\mathcal{J}}_v} p_{i,l} \beta_{0,\mathbf{q},l}\right) \text{ and } \eta\left(\sum_{l \in \bar{\mathcal{J}}_v} p_{i,l}(\beta_{0,\mathbf{q},l} + \beta_{1,l} + \beta_{2,l}) - \sum_{j \in \mathcal{J}_v} p_{i,j} v_j\right),$$

we have

$$\begin{aligned} \Bigg| \psi\left(y_i, \eta\left(\sum_{l \in \bar{\mathcal{J}}_v} p_{i,l}(\beta_{0,\mathbf{q},l} + \beta_{1,l} + \beta_{2,l}) - \sum_{j \in \mathcal{J}_v} p_{i,j} v_j\right); \mathbf{q}\right) & \\ - \psi\left(y_i, \eta\left(\sum_{l \in \bar{\mathcal{J}}_v} p_{i,l} \beta_{0,\mathbf{q},l}\right); \mathbf{q}\right) \Bigg| &\lesssim \big(\overline{\psi}(\mathbf{x}_i, y_i) + 1\big)(r_{1,n} + r_{2,n} + \varepsilon_n)^\alpha, \end{aligned}$$

and for $y_i$ *in* this segment we have

$$\Bigg| \psi\left(y_i, \eta\left(\sum_{l \in \bar{\mathcal{J}}_v} p_{i,l}(\beta_{0,\mathbf{q},l} + \beta_{1,l} + \beta_{2,l}) - \sum_{j \in \mathcal{J}_v} p_{i,j} v_j\right); \mathbf{q}\right)$$

$$- \psi\left(y_i, \eta\left(\sum_{l\in\bar{\mathcal{J}}_v} p_{i,l}\beta_{0,\mathbf{q},l}\right); \mathbf{q}\right)\Bigg| \lesssim \big(\overline{\psi}(\mathbf{x}_i, y_i)+1\big)(r_{1,n}+r_{2,n}+\varepsilon_n)+1$$

uniformly over $\mathbf{q}$.

By construction, for each $\boldsymbol{v}\in\mathcal{V}$, there exists some $\boldsymbol{k_v}$ such that $|v_{\boldsymbol{l},l}|\le \varrho^{\|\boldsymbol{l}-\boldsymbol{k_v}\|_\infty}\varepsilon_n$ if $\|\boldsymbol{l}-\boldsymbol{k_v}\|_\infty \le M_n$, and otherwise $v_{\boldsymbol{l},l}=0$. The above facts imply that for any $\mathbf{x}_i\in\delta\subset\mathcal{I}_{\boldsymbol{v}}$,

$$\mathbb{V}[\delta_{\mathbf{q},i}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{v})\mid \mathbf{x}_i] \lesssim \varepsilon_n^2(r_{1,n}+r_{2,n}+\varepsilon_n)^{(2\alpha)\wedge 1}\sum_{(\boldsymbol{l},l)\in\mathcal{L}_\delta}\varrho^{2\|\boldsymbol{l}-\boldsymbol{k_v}\|_\infty} \quad \text{for}$$
$$\mathcal{L}_\delta := \{(\boldsymbol{l},l)\colon \operatorname{supp} p_{\boldsymbol{l},l}\cap\delta\neq\varnothing\}.$$

In addition, since $\delta_{\mathbf{q},i}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{v})\neq 0$ only if $\mathbf{x}_i\in\mathcal{I}_{\boldsymbol{v}}$, for all $g\in\mathcal{G}$,

$$\begin{aligned}
&\mathbb{E}_n[\mathbb{V}[g(\mathbf{x}_i,y_i)\mid\mathbf{x}_i]] \lesssim \varepsilon_n^2(r_{1,n}+r_{2,n}+\varepsilon_n)^{(2\alpha)\wedge1}\sum_{\delta\subset\mathcal{I}_{\boldsymbol{v}}}\mathbb{E}_n[\mathbb{1}(\mathbf{x}_i\in\delta)]\sum_{(\boldsymbol{l},l)\in\mathcal{L}_\delta}\varrho^{2\|\boldsymbol{l}-\boldsymbol{k_v}\|_\infty}\\
&= \varepsilon_n^2(r_{1,n}+r_{2,n}+\varepsilon_n)^{(2\alpha)\wedge1}\sum_{\boldsymbol{l},l}\varrho^{2\|\boldsymbol{l}-\boldsymbol{k_v}\|_\infty}\sum_{\delta\in\mathcal{L}^\star_{\boldsymbol{l},l}}\mathbb{E}_n[\mathbb{1}(\mathbf{x}_i\in\delta)] \quad \text{for}\\
&\mathcal{L}^\star_{\boldsymbol{l},l} := \{\delta\subset\mathcal{I}_{\boldsymbol{v}}\colon \operatorname{supp} p_{\boldsymbol{l},l}\cap\delta\neq\varnothing\}.
\end{aligned}$$

Note that $\mathcal{L}^\star_{\boldsymbol{l},l}$ contains a bounded number of elements. Then on $\mathcal{A}_1$,

$$\begin{aligned}
&\sup_{g\in\mathcal{G}}\mathbb{E}_n[\mathbb{V}[g(\mathbf{x}_i,y_i)\mid\mathbf{x}_i]] \lesssim \varepsilon_n^2h^d(r_{1,n}+r_{2,n}+\varepsilon_n)^{(2\alpha)\wedge1}\sum_{\boldsymbol{l},l}\varrho^{2\|\boldsymbol{l}-\boldsymbol{k_v}\|_\infty}\\
&\lesssim \varepsilon_n^2h^d(r_{1,n}+r_{2,n}+\varepsilon_n)^{(2\alpha)\wedge1}\sum_{\boldsymbol{l}}\varrho^{2\|\boldsymbol{l}-\boldsymbol{k_v}\|_\infty} \qquad \text{since } l \text{ is bounded}\\
&\lesssim \varepsilon_n^2h^d(r_{1,n}+r_{2,n}+\varepsilon_n)^{(2\alpha)\wedge1}\sum_{\boldsymbol{t}\in\mathbb{Z}^d}\varrho^{2\|\boldsymbol{t}\|_\infty} \lesssim \varepsilon_n^2h^d(r_{1,n}+r_{2,n}+\varepsilon_n)^{(2\alpha)\wedge1},
\end{aligned}$$

concluding the proof of Lemma E.3. □

**Lemma E.4** (Complexity of class $\mathcal{G}$)**.** *Class $\mathcal{G}$ with envelope $\varepsilon_n$ 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.4.* First, indeed $\sup_{\mathbf{x},y}\sup_{g\in\mathcal{G}}|g(\mathbf{x},y)|\lesssim\varepsilon_n$.

Next, the class of functions $\mathcal{W}_1 := \{(\mathbf{x}_i,y_i)\mapsto \boldsymbol{v}^\intercal\mathbf{p}(\mathbf{x}_i)\colon \boldsymbol{v}\in\mathcal{V}\}$ is a union of $O\big(h^{-d}\big)$ classes $\mathcal{W}_{1,\boldsymbol{k}} := \{(\mathbf{x}_i,y_i)\mapsto \boldsymbol{v}^\intercal\mathbf{p}(\mathbf{x}_i)\colon \boldsymbol{v}\in\mathcal{V}_{\boldsymbol{k}}\}$, where

$$\mathcal{V}_{\boldsymbol{k}} = \Big\{\boldsymbol{v}\in\mathbb{R}^K : |v_{\boldsymbol{\ell},l}|\le\varrho^{\|\boldsymbol{k}-\boldsymbol{\ell}\|_\infty}\varepsilon_n \text{ for } \|\boldsymbol{k}-\boldsymbol{\ell}\|_\infty\le M_n \text{ and } v_{\boldsymbol{\ell},l}=0 \text{ otherwise}\Big\}.$$

Since $\mathcal{W}_{1,\boldsymbol{k}}$ is a subclass of a vector space of functions of dimension $O\big(\log^d n\big)$, by Lemma 2.6.15 in [11] it is VC with index $O\big(\log^d n\big)$. This implies that $\mathcal{W}_{1,\boldsymbol{k}}$ with envelope $O(\varepsilon_n)$ satisfies the uniform entropy bound (A.3) with $A\lesssim 1$ and $V\lesssim\log^d n$. Since there are $O\big(h^{-d}\big)$ such classes and $\log\frac{1}{h}\lesssim\log n$, using the chain

$$O\Big(h^{-d}\Big)\left(\frac{A}{\varepsilon}\right)^{O\left(\log^d n\right)} = e^{O(\log n)}\left(\frac{A}{\varepsilon}\right)^{O\left(\log^d n\right)} \le \left(\frac{A}{\varepsilon}\right)^{O(\log n)+O\left(\log^d n\right)} = \left(\frac{A}{\varepsilon}\right)^{O\left(\log^d n\right)} \tag{E.17}$$

(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.2(iv), the class of functions

$$\mathcal{W}_2 := \Big\{(\mathbf{x}_i, y_i) \mapsto [\psi(y_i, \eta(\mathbf{p}_i^\mathsf{T}(\boldsymbol{\beta}_0(\mathbf{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{v})); \mathbf{q}) - \psi(y_i, \eta(\mathbf{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\mathbf{q})); \mathbf{q})]\mathbb{1}\{\mathbf{x}_i \in \mathcal{I}_{\boldsymbol{v}}\}\colon \boldsymbol{\beta}_1 \in \mathcal{H}_1, \boldsymbol{\beta}_2 \in \mathcal{H}_2, \boldsymbol{v} \in \mathcal{V}, \mathbf{q} \in \mathcal{Q}\Big\}$$

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

The class of functions

$$\mathcal{W}_3 := \Big\{(\mathbf{x}_i, y_i) \mapsto \eta^{(1)}(\mathbf{p}(\mathbf{x}_i)^\mathsf{T}(\boldsymbol{\beta}_0(\mathbf{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{v}))\mathbb{1}\{\mathbf{x}_i \in \mathcal{I}_{\boldsymbol{v}}\}\colon \boldsymbol{\beta}_1 \in \mathcal{H}_1, \boldsymbol{\beta}_2 \in \mathcal{H}_2, \boldsymbol{v} \in \mathcal{V}, \mathbf{q} \in \mathcal{Q}\Big\}$$

is a subset of the union over $\delta \in \Delta$ of classes (for some fixed positive constants $c$ and $r$, $n$ large enough)

$$\mathcal{W}_{3,\delta} := \Big\{(\mathbf{x}_i, y_i) \mapsto \eta^{(1)}(\mathbf{p}(\mathbf{x}_i)^\mathsf{T}\boldsymbol{\beta})\mathbb{1}\{\mathbf{x}_i \in \mathcal{N}_{[c \log n]}(\delta)\}\colon \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(\mathbf{q})\|_\infty \leq r, \mathbf{q} \in \mathcal{Q}\Big\}.$$

Note that $\boldsymbol{\beta}$ can be assumed to lie in a fixed vector space $\mathcal{B}_\delta$ of dimension $\dim \mathcal{B}_\delta = O\big(\log^d n\big)$. Again applying Lemma 2.6.15 in [11] 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.4. $\square$

**Lemma E.5** (Uniform concentration in $\mathcal{G}$)**.** *On the event $\mathcal{A}_1$ defined in Lemma E.3 we have*

$$\begin{aligned}\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q},\boldsymbol{\beta}_1\in\mathcal{H}_1,\boldsymbol{\beta}_2\in\mathcal{H}_2,\boldsymbol{v}\in\mathcal{V}} |\mathbb{E}_n[\delta_{\mathbf{q},i}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{v}) - \mathbb{E}[\delta_{\mathbf{q},i}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{v})|\mathbf{x}_i]]| \,\middle|\, \{\mathbf{x}_i\}_{i=1}^n\right] \\ \lesssim \frac{h^{d/2}\varepsilon_n(r_{1,n} + r_{2,n} + \varepsilon_n)^{\alpha\wedge\frac{1}{2}}}{\sqrt{n}} \log^{(d+1)/2} n + \frac{\varepsilon_n \log^{d+1} n}{n}.\end{aligned} \tag{E.18}$$

*Proof of Lemma E.5.* This follows by applying Lemma C.6 conditionally on $\{\mathbf{x}_i\}_{i=1}^n$ on $\mathcal{A}_1$. $\square$

**Step 2**

**Lemma E.6.** *For $\widetilde{\boldsymbol{Q}}_\mathbf{q} := \mathbb{E}_n\Big[\mathbf{p}_i\mathbf{p}_i^\mathsf{T}\Psi_1(\mathbf{x}_i, \eta(\mathbf{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\mathbf{q})); \mathbf{q})\eta^{(1)}(\mathbf{p}_i^\mathsf{T}\boldsymbol{\beta}_0(\mathbf{q}))^2\Big]$, we have the following bound on the event $\mathcal{A}_1$:*

$$\sup_{\mathbf{q}\in\mathcal{Q},\boldsymbol{\beta}_1\in\mathcal{H}_1,\boldsymbol{\beta}_2\in\mathcal{H}_2,\boldsymbol{v}\in\mathcal{V}} \left|\boldsymbol{v}^\mathsf{T}\Big(\widetilde{\boldsymbol{Q}}_\mathbf{q} - \bar{\boldsymbol{Q}}_\mathbf{q}\Big)(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2)\right| \lesssim h^{m+d}\varepsilon_n(r_{1,n} + r_{2,n}). \tag{E.19}$$

*Proof of Lemma E.6.* By the same logic as in Lemma C.11, we have on $\mathcal{A}_1$

$$\left\|\bar{\boldsymbol{Q}}_\mathbf{q} - \widetilde{\boldsymbol{Q}}_\mathbf{q}\right\|_\infty \vee \left\|\bar{\boldsymbol{Q}}_\mathbf{q} - \widetilde{\boldsymbol{Q}}_\mathbf{q}\right\| \lesssim h^m h^d$$

uniformly over $\mathbf{q}$ with probability approaching one. This gives (E.19), proving Lemma E.6. $\square$

**Step 3**

**Lemma E.7.** *On $\mathcal{A}_1$, we have*

$$\sup_{\mathbf{q}\in\mathcal{Q},\boldsymbol{\beta}_1\in\mathcal{H}_1,\boldsymbol{\beta}_2\in\mathcal{H}_2,\boldsymbol{v}\in\mathcal{V}}\Big|\mathbb{E}_n[\mathbb{E}[\delta_{\mathbf{q},i}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{v})|\mathbf{x}_i]]-\boldsymbol{v}^\intercal\widetilde{\boldsymbol{Q}}_{\mathbf{q}}(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)\Big|$$
$$\lesssim \varepsilon_n^2 h^d+\varepsilon_n h^d(r_{1,n}+r_{2,n}+\varepsilon_n)^2.$$

*Proof of Lemma E.7.* First, on $\mathcal{A}_1$ the largest eigenvalue of $\widetilde{\boldsymbol{Q}}_{\mathbf{q}}$ is bounded by $h^d$ up to a constant factor (uniformly in $\mathbf{q}$):

$$\lambda_{\max}\big(\widetilde{\boldsymbol{Q}}_{\mathbf{q}}\big)=\sup_{\|\boldsymbol{\alpha}\|=1}\boldsymbol{\alpha}^\intercal\widetilde{\boldsymbol{Q}}_{\mathbf{q}}\boldsymbol{\alpha}=\sup_{\|\boldsymbol{\alpha}\|=1}\mathbb{E}_n\big[(\boldsymbol{\alpha}^\intercal\mathbf{p}_i)^2\Psi_1(\mathbf{x}_i,\eta(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}));\mathbf{q})\eta^{(1)}(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}))^2\big]$$
$$\leq \sup_{\|\boldsymbol{\alpha}\|=1}\mathbb{E}_n\big[(\boldsymbol{\alpha}^\intercal\mathbf{p}_i)^2\big|\Psi_1(\mathbf{x}_i,\eta(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}));\mathbf{q})\big|\eta^{(1)}(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}))^2\big]$$
$$\lesssim \sup_{\|\boldsymbol{\alpha}\|=1}\mathbb{E}_n\big[(\boldsymbol{\alpha}^\intercal\mathbf{p}_i)^2\big]$$

(because $\big|\Psi_1(\mathbf{x}_i,\eta(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}));\mathbf{q})\big|\eta^{(1)}(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}))^2\lesssim 1$ by Assumptions B.2 and B.5)

$$=\sup_{\|\boldsymbol{\alpha}\|=1}\sum_{l=1}^{\bar{\kappa}}\mathbb{E}_n\big[(\boldsymbol{\alpha}^\intercal\mathbf{p}_i)^2\mathbb{1}\{\mathbf{x}_i\in\delta_l\}\big]$$
$$\lesssim \sup_{\|\boldsymbol{\alpha}\|=1}\sum_{l=1}^{\bar{\kappa}}\bigg(\sum_{k=1}^{M_l}\alpha_{l,k}^2\bigg)\mathbb{E}_n\big[\mathbb{1}\{\mathbf{x}_i\in\delta_l\}\big]$$

(because $\sup_{\mathbf{x}\in\delta_l}(\boldsymbol{\alpha}^\intercal\mathbf{p}(\mathbf{x}))^2\lesssim\sum_{k=1}^{M_l}\alpha_{l,k}^2$, where $\{\alpha_{l,k}\}_{k=1}^{M_l}$ are the components of $\boldsymbol{\alpha}$ corresponding to the $M_l$ basis functions supported on $\delta_l$)

$$\leq \sup_{\delta\in\Delta}\mathbb{E}_n\big[\mathbb{1}\{\mathbf{x}_i\in\delta\}\big]\sup_{\|\boldsymbol{\alpha}\|=1}\sum_{l=1}^{\bar{\kappa}}\bigg(\sum_{k=1}^{M_l}\alpha_{l,k}^2\bigg)$$
$$\lesssim \sup_{\delta\in\Delta}\mathbb{E}_n\big[\mathbb{1}\{\mathbf{x}_i\in\delta\}\big]\sup_{\|\boldsymbol{\alpha}\|=1}\|\boldsymbol{\alpha}\|^2=\sup_{\delta\in\Delta}\mathbb{E}_n\big[\mathbb{1}\{\mathbf{x}_i\in\delta\}\big].$$

Next, by Taylor expansion,

$$\begin{aligned}
&\mathbb{E}[\delta_{\mathbf{q},i}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{v})|\mathbf{x}_i]\\
&\quad=\boldsymbol{v}^\intercal\mathbf{p}_i[\Psi(\mathbf{x}_i,\eta(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{v}));\mathbf{q})-\Psi(\mathbf{x}_i,\eta(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}));\mathbf{q})]\\
&\qquad\times\eta^{(1)}(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{v}))\\
&\quad=\boldsymbol{v}^\intercal\mathbf{p}_i\Big[\Psi_1(\mathbf{x}_i,\eta(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}));\mathbf{q})\Big\{\eta^{(1)}(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}))\mathbf{p}_i^\intercal(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{v})+\\
&\qquad\quad+\frac{1}{2}\eta^{(2)}(\xi_{\mathbf{q},i})(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{v}))^2\Big\}\\
&\qquad+\frac{1}{2}\Psi_2\Big(\mathbf{x}_i,\tilde{\xi}_{\mathbf{q},i};\mathbf{q}\Big)\{\eta(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{v}))-\eta(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}))\}^2\Big]\\
&\qquad\times\eta^{(1)}(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{v}))
\end{aligned}$$

for some $\xi_{\mathbf{q},i}$ between $\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q})$ and $\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{v})$, $\tilde{\xi}_{\mathbf{q},i}$ is between $\eta(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}))$ and $\eta(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{v}))$. This gives

$$\mathbb{E}_n[\mathbb{E}[\delta_{\mathbf{q},i}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{v})|\mathbf{x}_i]]=\boldsymbol{v}^\intercal\widetilde{\boldsymbol{Q}}_{\mathbf{q}}(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)-\boldsymbol{v}^\intercal\widetilde{\boldsymbol{Q}}_{\mathbf{q}}\boldsymbol{v}+\mathrm{I}+\mathrm{II}+\mathrm{III},$$

where for some $\check{\xi}_{\mathbf{q},i}$ between $\mathbf{p}_i^\intercal \boldsymbol{\beta}_0(\mathbf{q})$ and $\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{v})$

$$
\begin{aligned}
\mathrm{I} &:= \mathbb{E}_n\Big[\boldsymbol{v}^\intercal \mathbf{p}_i \Psi_1(\mathbf{x}_i, \eta(\mathbf{p}_i^\intercal \boldsymbol{\beta}_0(\mathbf{q})); \mathbf{q}) \eta^{(1)}(\mathbf{p}_i^\intercal \boldsymbol{\beta}_0(\mathbf{q})) \eta^{(2)}\big(\check{\xi}_{\mathbf{q},i}\big)(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{v}))^2\Big],\\
\mathrm{II} &:= \frac{1}{2}\mathbb{E}_n\Big[\boldsymbol{v}^\intercal \mathbf{p}_i \Psi_1(\mathbf{x}_i, \eta(\mathbf{p}_i^\intercal \boldsymbol{\beta}_0(\mathbf{q})); \mathbf{q}) \eta^{(2)}(\xi_{\mathbf{q},i})\\
&\quad \times \eta^{(1)}(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{v}))(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{v}))^2\Big],\\
\mathrm{III} &:= \frac{1}{2}\mathbb{E}_n\Big[\boldsymbol{v}^\intercal \mathbf{p}_i \Psi_2\Big(\mathbf{x}_i, \tilde{\xi}_{\mathbf{q},i}; \mathbf{q}\Big)\{\eta(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{v})) - \eta(\mathbf{p}_i^\intercal \boldsymbol{\beta}_0(\mathbf{q}))\}^2\\
&\quad \times \eta^{(1)}(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{v}))\Big]
\end{aligned}
$$

and

$$
\begin{aligned}
&\boldsymbol{v}^\intercal \widetilde{\boldsymbol{Q}}_{\mathbf{q}} \boldsymbol{v} \lesssim \varepsilon_n^2 h^d,\\
&\mathrm{I} \lesssim \varepsilon_n h^d (r_{1,n} + r_{2,n} + \varepsilon_n)^2,\\
&\mathrm{II} \lesssim \varepsilon_n h^d (r_{1,n} + r_{2,n} + \varepsilon_n)^2,\\
&\mathrm{III} \lesssim \varepsilon_n h^d (r_{1,n} + r_{2,n} + \varepsilon_n)^2
\end{aligned}
$$

on the event $\mathcal{A}_1$. □

**Step 4** We employ the following lemma.

**Lemma E.8.** *There exists an event $\mathcal{A}_2$ whose probability converges to one such that on $\mathcal{A}_2$*

$$
\begin{aligned}
\sup_{\mathbf{q}\in\mathcal{Q}, \boldsymbol{\beta}_1\in\mathcal{H}_1, \boldsymbol{\beta}_2\in\mathcal{H}_2, \boldsymbol{v}\in\mathcal{V}} &\Big|\mathbb{E}_n\Big[\boldsymbol{v}^\intercal \mathbf{p}_i \psi(y_i, \eta(\mathbf{p}_i^\intercal \boldsymbol{\beta}_0(\mathbf{q})); \mathbf{q}) \eta^{(1)}(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{v}))\Big]\\
&- \mathbb{E}_n\Big[\boldsymbol{v}^\intercal \mathbf{p}_i \psi(y_i, \eta(\mathbf{p}_i^\intercal \boldsymbol{\beta}_0(\mathbf{q})); \mathbf{q}) \eta^{(1)}(\mathbf{p}_i^\intercal \boldsymbol{\beta}_0(\mathbf{q}))\Big]\Big| \lesssim h^d r_{1,n}(r_{1,n} + r_{2,n} + \varepsilon_n)\varepsilon_n.
\end{aligned}
$$

*Proof of Lemma E.8.*

$$
\begin{aligned}
&\mathbb{E}_n\Big[\boldsymbol{v}^\intercal \mathbf{p}_i \psi(y_i, \eta(\mathbf{p}_i^\intercal \boldsymbol{\beta}_0(\mathbf{q})); \mathbf{q}) \eta^{(1)}(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{v}))\Big]\\
&\qquad - \mathbb{E}_n\Big[\boldsymbol{v}^\intercal \mathbf{p}_i \psi(y_i, \eta(\mathbf{p}_i^\intercal \boldsymbol{\beta}_0(\mathbf{q})); \mathbf{q}) \eta^{(1)}(\mathbf{p}_i^\intercal \boldsymbol{\beta}_0(\mathbf{q}))\Big]\\
&\quad = \boldsymbol{v}^\intercal \mathbb{E}_n\Big[\psi(y_i, \eta(\mathbf{p}_i^\intercal \boldsymbol{\beta}_0(\mathbf{q})); \mathbf{q}) \eta^{(2)}(\xi_{\mathbf{q},i}) \mathbf{p}_i \mathbf{p}_i^\intercal\Big](\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{v})\\
&\quad \lesssim h^d r_{1,n}(r_{1,n} + r_{2,n} + \varepsilon_n)\varepsilon_n,
\end{aligned}
$$

where $\xi_{\mathbf{q},i}$ is between $\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{v})$ and $\mathbf{p}_i^\intercal \boldsymbol{\beta}_0(\mathbf{q})$. The bound holds on the event

$$
\mathcal{A}_2 := \Big\{\sup\Big\|\mathbb{E}_n\Big[\psi(y_i, \eta(\mathbf{p}_i^\intercal \boldsymbol{\beta}_0(\mathbf{q})); \mathbf{q}) \eta^{(2)}(\xi_{\mathbf{q},i}) \mathbf{p}_i \mathbf{p}_i^\intercal\Big]\Big\|_\infty \le h^d r_{1,n}\Big\},
$$

where the supremum is over $\boldsymbol{\beta}_1 \in \mathcal{H}_1$, $\boldsymbol{\beta}_2 \in \mathcal{H}_2$, $\boldsymbol{v} \in \mathcal{V}$, $\mathbf{q} \in \mathcal{Q}$ and $\xi_{\mathbf{q},i}$ between $\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 - \boldsymbol{v})$ and $\mathbf{p}_i^\intercal \boldsymbol{\beta}_0(\mathbf{q})$. By the same argument as Lemmas C.12 and C.13, $\mathbb{P}\{\mathcal{A}_2\} \to 1$. □

**Step 5** Let $(\boldsymbol{k}, k)$ be such an index that $|\beta_{2,\boldsymbol{k},k}| = \|\boldsymbol{\beta}_2\|_\infty$. Let

$$\bar{\boldsymbol{v}}_{\mathbf{q}} = c_4 \varepsilon_n h^d \big[\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1}\big]_{\boldsymbol{k},k} \operatorname{sign}(\beta_{2,\boldsymbol{k},k})$$

for some $c_4 > 0$ chosen later, where $\big[\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1}\big]_{\boldsymbol{k},k}$ denotes the $(\boldsymbol{k}, k)$th row of $\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1}$. Take $\boldsymbol{v}_{\mathbf{q}} \in \mathbb{R}^K$ with components $v_{\mathbf{q},\boldsymbol{j},j} = \bar{v}_{\mathbf{q},\boldsymbol{j},j}$ for $\|\boldsymbol{j} - \boldsymbol{k}\|_\infty \le M_n$ and zero otherwise. Clearly, $\boldsymbol{v}_{\mathbf{q}} \in \mathcal{V}$ on an event $\mathcal{A}_3$ with $\mathbb{P}(\mathcal{A}_3) \to 1$. On $\mathcal{A}_1 \cap \mathcal{A}_3$,

$$\left|(\boldsymbol{v}_{\mathbf{q}} - \bar{\boldsymbol{v}}_{\mathbf{q}})^\intercal \bar{\boldsymbol{Q}}_{\mathbf{q}} \boldsymbol{\beta}_2\right| \lesssim \varepsilon_n h^d r_{2,n} n^{-c_5}$$

for some large $c_5 > 0$ if we let $c_1$ be sufficiently large.

**Step 6** For each $\mathbf{q}$, partition the whole parameter space into shells: $\mathcal{O}_{\mathbf{q}} := \bigcup_{\ell=-\infty}^{\bar{L}} \mathcal{O}_{\mathbf{q},\ell}$ where $\mathcal{O}_{\mathbf{q},\ell} := \big\{\boldsymbol{\beta} \in \mathbb{R}^K \colon 2^{\ell-1}\mathfrak{r}_{2,n} \le \big\|\boldsymbol{\beta} - \boldsymbol{\beta}_0(\mathbf{q}) - \bar{\boldsymbol{\beta}}_{\mathbf{q}}\big\|_\infty < 2^\ell \mathfrak{r}_{2,n}\big\}$ for the smallest $\bar{L}$ such that $2^{\bar{L}}\mathfrak{r}_{2,n} \ge c$. Since with probability approaching one both $\sup_{\mathbf{q}\in\mathcal{Q}}\Big\|\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q})\Big\|_\infty \le c/2$ and $\sup_{\mathbf{q}\in\mathcal{Q}}\big\|\bar{\boldsymbol{\beta}}_{\mathbf{q}}\big\|_\infty \le c/2$, we obtain that the probability of the event $\bigcap_{\mathbf{q}\in\mathcal{Q}}\Big\{\widehat{\boldsymbol{\beta}}(\mathbf{q}) \in \mathcal{O}_{\mathbf{q}}\Big\}$ approaches one. Define $\mathcal{A} := \mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3$.

Note that for any vector $\boldsymbol{v}$

$$\mathbb{E}_n\Big[\boldsymbol{v}^\intercal \mathbf{p}_i \psi\Big(y_i, \eta\Big(\mathbf{p}_i^\intercal\Big(\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{v}\Big)\Big); \mathbf{q}\Big)\eta^{(1)}\Big(\mathbf{p}_i^\intercal\Big(\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{v}\Big)\Big)\Big] \le 0, \tag{E.20}$$

because $g\Big(\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{v}\Big) := \mathbb{E}_n\Big[\psi\Big(y_i, \eta\Big(\mathbf{p}_i^\intercal\Big(\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{v}\Big)\Big); \mathbf{q}\Big)\eta^{(1)}\Big(\mathbf{p}_i^\intercal\Big(\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{v}\Big)\Big)\mathbf{p}_i\Big]$ is a subgradient of the function $f(\boldsymbol{\beta}) := \mathbb{E}_n[\rho(y_i, \eta(\mathbf{p}_i^\intercal\boldsymbol{\beta}); \mathbf{q})]$ at $\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{v}$, and $\widehat{\boldsymbol{\beta}}(\mathbf{q})$ is the minimizer of this function, giving $\boldsymbol{v}^\intercal g\Big(\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{v}\Big) \le f\Big(\widehat{\boldsymbol{\beta}}(\mathbf{q})\Big) - f\Big(\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{v}\Big) \le 0$.

Let $\ell \ge L$ for some constant $L$ chosen later, and put $\varepsilon_n := 2^L\mathfrak{r}_{2,n}$. Denote $\mathcal{H}_{2,\ell} := \big\{\boldsymbol{\beta} \in \mathbb{R}^K : \|\boldsymbol{\beta}\|_\infty \le 2^\ell \mathfrak{r}_{2,n}\big\}$, and note that if $\boldsymbol{\beta} \in \mathcal{O}_{\mathbf{q},\ell}$, then $\boldsymbol{\beta} - \boldsymbol{\beta}_0(\mathbf{q}) - \bar{\boldsymbol{\beta}}_{\mathbf{q}} \in \mathcal{H}_{2,\ell}$ by definition. Recall also that $\bar{\boldsymbol{\beta}}_{\mathbf{q}} \in \mathcal{H}_1$ on the event of probability approaching one by (E.16). Therefore, if $\widehat{\boldsymbol{\beta}}(\mathbf{q}) \in \mathcal{O}_{\mathbf{q},\ell}$, for $\boldsymbol{v}_{\mathbf{q}}$ defined in Step 5, on the event with probability approaching one

$$\begin{aligned}
0 &\ge \mathbb{E}_n\Big[\boldsymbol{v}_{\mathbf{q}}^\intercal \mathbf{p}_i \psi\Big(y_i, \eta\Big(\mathbf{p}_i^\intercal\Big(\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{v}_{\mathbf{q}}\Big)\Big); \mathbf{q}\Big)\eta^{(1)}\Big(\mathbf{p}_i^\intercal\Big(\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{v}_{\mathbf{q}}\Big)\Big)\Big] \qquad \text{(E.21)}\\
&= \frac{1}{\sqrt{n}}\mathbb{G}_n\Big[\delta_{\mathbf{q},i}\Big(\bar{\boldsymbol{\beta}}_{\mathbf{q}}, \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) - \bar{\boldsymbol{\beta}}_{\mathbf{q}}, \boldsymbol{v}_{\mathbf{q}}\Big)\Big]\\
&\quad + \mathbb{E}_n\Big[\mathbb{E}\Big[\delta_{\mathbf{q},i}\Big(\bar{\boldsymbol{\beta}}_{\mathbf{q}}, \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) - \bar{\boldsymbol{\beta}}_{\mathbf{q}}, \boldsymbol{v}_{\mathbf{q}}\Big) \,\Big|\, \mathbf{x}_i\Big]\Big]\\
&\quad + S_4 + \mathbb{E}_n\Big[\boldsymbol{v}_{\mathbf{q}}^\intercal \mathbf{p}_i \psi(y_i, \eta(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q})); \mathbf{q})\eta^{(1)}(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}))\Big]\\
&= \frac{1}{\sqrt{n}}\mathbb{G}_n\Big[\delta_{\mathbf{q},i}\Big(\bar{\boldsymbol{\beta}}_{\mathbf{q}}, \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) - \bar{\boldsymbol{\beta}}_{\mathbf{q}}, \boldsymbol{v}_{\mathbf{q}}\Big)\Big] + S_3 + S_4\\
&\quad + \boldsymbol{v}_{\mathbf{q}}^\intercal \widetilde{\boldsymbol{Q}}_{\mathbf{q}}\Big(\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q})\Big) - \boldsymbol{v}_{\mathbf{q}}^\intercal \bar{\boldsymbol{Q}}_{\mathbf{q}} \bar{\boldsymbol{\beta}}_{\mathbf{q}}\\
&= \frac{1}{\sqrt{n}}\mathbb{G}_n\Big[\delta_{\mathbf{q},i}\Big(\bar{\boldsymbol{\beta}}_{\mathbf{q}}, \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) - \bar{\boldsymbol{\beta}}_{\mathbf{q}}, \boldsymbol{v}_{\mathbf{q}}\Big)\Big] + S_3 + S_4\\
&\quad + \boldsymbol{v}_{\mathbf{q}}^\intercal\Big(\widetilde{\boldsymbol{Q}}_{\mathbf{q}} - \bar{\boldsymbol{Q}}_{\mathbf{q}}\Big)\Big(\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q})\Big) + \boldsymbol{v}_{\mathbf{q}}^\intercal \bar{\boldsymbol{Q}}_{\mathbf{q}}\Big(\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) - \bar{\boldsymbol{\beta}}_{\mathbf{q}}\Big)\\
&= \frac{1}{\sqrt{n}}\mathbb{G}_n\Big[\delta_{\mathbf{q},i}\Big(\bar{\boldsymbol{\beta}}_{\mathbf{q}}, \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) - \bar{\boldsymbol{\beta}}_{\mathbf{q}}, \boldsymbol{v}_{\mathbf{q}}\Big)\Big] + S_2 + S_3 + S_4
\end{aligned}$$

$$
\begin{aligned}
&+ (\boldsymbol{v}_{\mathbf{q}} - \bar{\boldsymbol{v}}_{\mathbf{q}})^{\mathsf{T}} \bar{\boldsymbol{Q}}_{\mathbf{q}} \Big( \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) - \bar{\boldsymbol{\beta}}_{\mathbf{q}} \Big) + \bar{\boldsymbol{v}}_{\mathbf{q}}^{\mathsf{T}} \bar{\boldsymbol{Q}}_{\mathbf{q}} \Big( \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) - \bar{\boldsymbol{\beta}}_{\mathbf{q}} \Big) \\
= &\frac{1}{\sqrt{n}} \mathbb{G}_n \Big[ \delta_{\mathbf{q},i} \Big( \bar{\boldsymbol{\beta}}_{\mathbf{q}}, \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) - \bar{\boldsymbol{\beta}}_{\mathbf{q}}, \boldsymbol{v}_{\mathbf{q}} \Big) \Big] + S_2 + S_3 + S_4 + S_5 \\
&+ c_4 \varepsilon_n h^d \Big\| \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) - \bar{\boldsymbol{\beta}}_{\mathbf{q}} \Big\|_\infty \\
= &\frac{1}{\sqrt{n}} \mathbb{G}_n \Big[ \delta_{\mathbf{q},i} \Big( \bar{\boldsymbol{\beta}}_{\mathbf{q}}, \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) - \bar{\boldsymbol{\beta}}_{\mathbf{q}}, \boldsymbol{v}_{\mathbf{q}} \Big) \Big] + S_2 + S_3 + S_4 + S_5 \\
&+ c_4 \varepsilon_n h^d \Big\| \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) - \bar{\boldsymbol{\beta}}_{\mathbf{q}} \Big\|_\infty,
\end{aligned}
$$

where

$$
\begin{aligned}
S_2 &:= \boldsymbol{v}_{\mathbf{q}}^{\mathsf{T}} \Big( \widetilde{\boldsymbol{Q}}_{\mathbf{q}} - \bar{\boldsymbol{Q}}_{\mathbf{q}} \Big) \Big( \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) \Big), \\
S_3 &:= \mathbb{E}_n \Big[ \mathbb{E} \Big[ \delta_{\mathbf{q},i} \Big( \bar{\boldsymbol{\beta}}_{\mathbf{q}}, \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) - \bar{\boldsymbol{\beta}}_{\mathbf{q}}, \boldsymbol{v}_{\mathbf{q}} \Big) \Bigm| \mathbf{x}_i \Big] \Big] - \boldsymbol{v}_{\mathbf{q}}^{\mathsf{T}} \widetilde{\boldsymbol{Q}}_{\mathbf{q}} \Big( \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) \Big), \\
S_4 &:= \mathbb{E}_n \Big[ \boldsymbol{v}_{\mathbf{q}}^{\mathsf{T}} \mathbf{p}_i \psi(y_i, \eta(\mathbf{p}_i^{\mathsf{T}} \boldsymbol{\beta}_0(\mathbf{q})); \mathbf{q}) \eta^{(1)} \Big( \mathbf{p}_i^{\mathsf{T}} \Big( \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{v}_{\mathbf{q}} \Big) \Big) \Big] \\
&\quad - \mathbb{E}_n \Big[ \boldsymbol{v}_{\mathbf{q}}^{\mathsf{T}} \mathbf{p}_i \psi(y_i, \eta(\mathbf{p}_i^{\mathsf{T}} \boldsymbol{\beta}_0(\mathbf{q})); \mathbf{q}) \eta^{(1)} (\mathbf{p}_i^{\mathsf{T}} \boldsymbol{\beta}_0(\mathbf{q})) \Big], \\
S_5 &:= (\boldsymbol{v}_{\mathbf{q}} - \bar{\boldsymbol{v}}_{\mathbf{q}})^{\mathsf{T}} \bar{\boldsymbol{Q}}_{\mathbf{q}} \Big( \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) - \bar{\boldsymbol{\beta}}_{\mathbf{q}} \Big).
\end{aligned}
$$

We now bound these terms uniformly over $\mathbf{q}$, for $\ell \geq L$, where $L$ is constant, and such that $2^\ell \mathfrak{r}_{2,n} = O(1)$. Using Lemma E.6 and

$$
h^m r_{1,n} = o(\mathfrak{r}_{2,n}), \tag{E.22}
$$

we bound

$$
S_2 = O\Big( h^{m+d} \varepsilon_n \Big( r_{1,n} + 2^\ell \mathfrak{r}_{2,n} \Big) \Big) = o\Big( h^d 2^\ell \mathfrak{r}_{2,n} \varepsilon_n \Big).
$$

Using Lemma E.7, we bound

$$
S_3 = O\Big( \varepsilon_n^2 h^d \Big) + O\Big( \varepsilon_n h^d \Big( r_{1,n} + 2^\ell \mathfrak{r}_{2,n} + \varepsilon_n \Big)^2 \Big)
$$

on $\mathcal{A}$. Using Lemma E.8 and

$$
r_{1,n}^2 = o(\mathfrak{r}_{2,n}), \tag{E.23}
$$

we bound

$$
S_4 = O\Big( h^d r_{1,n} (r_{1,n} + 2^\ell \mathfrak{r}_{2,n} + \varepsilon_n) \varepsilon_n \Big) = o\Big( h^d 2^\ell \mathfrak{r}_{2,n} \varepsilon_n \Big)
$$

on $\mathcal{A}$. Finally, using Step 5 and $n^{-c_5} = o(1)$, we bound

$$
S_5 = O\Big( \varepsilon_n h^d 2^\ell \mathfrak{r}_{2,n} n^{-c_5} \Big) = o\Big( h^d 2^\ell \mathfrak{r}_{2,n} \varepsilon_n \Big).
$$

If $\widehat{\boldsymbol{\beta}}(\mathbf{q}) \in \mathcal{O}_{\mathbf{q},\ell}$, then $c_4 \varepsilon_n h^d \Big\| \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) - \bar{\boldsymbol{\beta}}_{\mathbf{q}} \Big\|_\infty \geq c_4 \varepsilon_n h^d 2^{\ell-1} \mathfrak{r}_{2,n}$. Note that

$$
|S_3| \overset{\text{(a)}}{\leq} c_4^2 \tilde{C}_1 h^d \varepsilon_n^2 + \tilde{C}_2 \varepsilon_n h^d \cdot 2^{2\ell} \mathfrak{r}_{2,n}^2 + o\Big( h^d 2^\ell \mathfrak{r}_{2,n} \varepsilon_n \Big) \overset{\text{(b)}}{\leq} \frac{c_4}{8} \varepsilon_n h^d 2^\ell \mathfrak{r}_{2,n},
$$

where in (a) we again used $r_{1,n}^2 = o(\mathfrak{r}_{2,n})$ and (b) is true if we choose $c_4$ and $c$ small enough; $\tilde{C}_1$ and $\tilde{C}_2$ are constants allowing for a bound $|(\boldsymbol{v}_{\mathbf{q}}/c_4)^\intercal \widetilde{\boldsymbol{Q}}_{\mathbf{q}}(\boldsymbol{v}_{\mathbf{q}}/c_4)| \le \tilde{C}_1 h^d \varepsilon_n^2$, and such that $c_4$ can be chosen small independently of $\tilde{C}_1$ or $\tilde{C}_2$. Combining this with

$$|S_2| + |S_4| + |S_5| \le \frac{c_4}{8}\varepsilon_n h^d 2^\ell \mathfrak{r}_{2,n}$$

for large enough $n$, we obtain

$$S_2 + S_3 + S_4 + S_5 + c_4\varepsilon_n h^d \left\|\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) - \bar{\boldsymbol{\beta}}_{\mathbf{q}}\right\|_\infty \ge \frac{c_4}{4}\varepsilon_n h^d 2^\ell \mathfrak{r}_{2,n}.$$

Combining the considerations above, we conclude

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{\mathbf{q}\in\mathcal{Q}}\left\|\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \boldsymbol{\beta}_0(\mathbf{q}) - \bar{\boldsymbol{\beta}}_{\mathbf{q}}\right\|_\infty \ge 2^L \mathfrak{r}_{2,n} \;\middle|\; \mathbf{x}_i\right\} \\
&\quad\le \mathbb{P}\left\{\bigcup_{\ell=L}^{\bar{L}}\left\{\sup_{\mathbf{q}\in\mathcal{Q}}\sup_{\boldsymbol{\beta}_1\in\mathcal{H}_1}\sup_{\boldsymbol{\beta}_2\in\mathcal{H}_{2,\ell}}\sup_{\boldsymbol{v}\in\mathcal{V}}\frac{1}{\sqrt{n}}\left|\mathbb{G}_n[\delta_{\mathbf{q},i}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{v})]\right| \ge C_4 h^d 2^\ell \mathfrak{r}_{2,n}\varepsilon_n\right\} \cap \mathcal{A} \;\middle|\; \mathbf{x}_i\right\} \\
&\qquad + o_{\mathbb{P}}(1) \\
&\quad\le \sum_{\ell=L}^{\bar{L}}\left(C_6 h^d 2^\ell \mathfrak{r}_{2,n}\varepsilon_n\right)^{-1}\mathbb{1}(\mathcal{A}_1) \\
&\qquad \times \mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\sup_{\boldsymbol{\beta}_1\in\mathcal{H}_1}\sup_{\boldsymbol{\beta}_2\in\mathcal{H}_{2,\ell}}\sup_{\boldsymbol{v}\in\mathcal{V}}\frac{1}{\sqrt{n}}\left|\mathbb{G}_n[\delta_{\mathbf{q},i}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{v})]\right| \;\middle|\; \mathbf{x}_i\right] + o_{\mathbb{P}}(1).
\end{aligned}$$

By Step 1, the term before $o_{\mathbb{P}}(1)$ in the rightmost expression can be made arbitrarily small by choosing $L$ large enough, for $n$ sufficiently large, since

$$\frac{h^{d/2}\varepsilon_n (r_{1,n})^{\alpha\wedge\frac{1}{2}}}{\sqrt{n}}\log^{(d+1)/2} n \lesssim h^d \mathfrak{r}_{2,n}\varepsilon_n \quad\Leftrightarrow\quad (r_{1,n})^{\alpha\wedge\frac{1}{2}}\left(\frac{\log^{d+1} n}{nh^d}\right)^{1/2} \lesssim \mathfrak{r}_{2,n}, \tag{E.24}$$

$$\frac{h^{d/2}\varepsilon_n}{\sqrt{n}}\log^{(d+1)/2} n \lesssim h^d (\mathfrak{r}_{2,n})^{1-\alpha\wedge\frac{1}{2}}\varepsilon_n \quad\Leftrightarrow\quad \left(\frac{\log^{d+1} n}{nh^d}\right)^{1/2} \lesssim (\mathfrak{r}_{2,n})^{1-\alpha\wedge\frac{1}{2}}, \tag{E.25}$$

$$\frac{\varepsilon_n \log^{d+1} n}{n} \lesssim h^d \mathfrak{r}_{2,n}\varepsilon_n \quad\Leftrightarrow\quad \frac{\log^{d+1} n}{nh^d} \lesssim \mathfrak{r}_{2,n}. \tag{E.26}$$

Since $\gamma_n$ was an arbitrary positive diverging (slowly enough) sequence, we infer (E.2) from this, and the theorem is proven.

*Remark* E.9 (Rate restrictions)*.* The rates in the proof are determined by four restrictions: Eqs. (E.15) and (E.23) to (E.25). Equation (E.26) follows from Eqs. (E.15) and (E.24). Equation (E.22) follows from Eq. (E.15) and Eq. (E.23).

## E.2 General case

We will state another version of the Bahadur Representation theorem, where $\theta \mapsto \rho(y, \eta(\theta); \mathbf{q})$ is not assumed to be convex. The following technical changes in the main assumptions are required:

**Replacement of Assumption B.2(i)** Assumption B.2(i) holds with (B.3) replaced by

$$\begin{aligned}
&\sup_{\mathbf{x}} \sup_{\lambda\in[0,1]} \sup_{y\notin[\eta(\zeta_1)\wedge\eta(\zeta_2),\eta(\zeta_1)\vee\eta(\zeta_2)]} |\varphi(y,\eta(\zeta_1+\lambda(\zeta_2-\zeta_1));\mathbf{q})-\varphi(y,\eta(\zeta_2);\mathbf{q})| \lesssim |\zeta_1-\zeta_2|^{\alpha},\\
&\sup_{\mathbf{x}} \sup_{\lambda\in[0,1]} \sup_{y\in[\eta(\zeta_1)\wedge\eta(\zeta_2),\eta(\zeta_1)\vee\eta(\zeta_2)]} |\varphi(y,\eta(\zeta_1+\lambda(\zeta_2-\zeta_1));\mathbf{q})-\varphi(y,\eta(\zeta_2);\mathbf{q})| \lesssim 1,
\end{aligned} \tag{E.27}$$

and (B.4) replaced by

$$\sup_{\mathbf{x}} \sup_{\lambda\in[0,1]} \sup_{y} |\varphi(y,\eta(\zeta_1+\lambda(\zeta_2-\zeta_1));\mathbf{q})-\varphi(y,\eta(\zeta_2);\mathbf{q})| \lesssim |\zeta_1-\zeta_2|^{\alpha}. \tag{E.28}$$

**Addition to Assumption B.2(iv)** The following assumption is **added** to Assumption B.2(iv):
For any fixed $c>0$, $r>0$, and a positive sequence $\varepsilon_n\to 0$ the class of functions

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

with envelope $\varepsilon_n$ 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 $n$.

**Theorem E.10** (Bahadur representation: general case)**.**

*(a) Suppose Assumptions B.1 to B.5 hold with changes described immediately above, $\mathcal{B}\subseteq\mathbb{R}^K$ is such a set that $\{\mathbf{b}\in\mathbb{R}^K:\|\mathbf{b}-\boldsymbol{\beta}_0(\mathbf{q})\|_\infty\le c, \mathbf{q}\in\mathcal{Q}\}\subseteq\mathcal{B}$ for some positive $c$, and the constrained minimizer*

$$\widehat{\boldsymbol{\beta}}_{\mathrm{constr}}(\mathbf{q})\in\operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathcal{B}}\mathbb{E}_n[\rho(y_i,\eta(\mathbf{p}(\mathbf{x}_i)^{\mathsf{T}}\boldsymbol{\beta});\mathbf{q})]$$

*exists and is a consistent in $\infty$-norm estimator of $\boldsymbol{\beta}_0(\mathbf{q})$, namely:*

$$\sup_{\mathbf{q}\in\mathcal{Q}}\Big\|\widehat{\boldsymbol{\beta}}_{\mathrm{constr}}(\mathbf{q})-\boldsymbol{\beta}_0(\mathbf{q})\Big\|_\infty = o_{\mathbb{P}}(1). \tag{E.29}$$

*In addition, suppose (E.1).*

*Then (E.2) holds with $\widehat{\boldsymbol{\beta}}(\mathbf{q})$ replaced by $\widehat{\boldsymbol{\beta}}_{\mathrm{constr}}(\mathbf{q})$.*

*(b) If, in addition to conditions in (i), (E.28) holds (without any restrictions on $y$), then (E.4) holds with $\widehat{\boldsymbol{\beta}}(\mathbf{q})$ replaced by $\widehat{\boldsymbol{\beta}}_{\mathrm{constr}}(\mathbf{q})$.*

The argument for Theorem E.10 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 $\delta_{\mathbf{q},i}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{v})$ is replaced with

$$\begin{aligned}
&\delta_{\mathbf{q},i}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{v})\\
&\quad:=\rho(y_i,\eta(\mathbf{p}_i^{\mathsf{T}}(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2));\mathbf{q})-\rho(y_i,\eta(\mathbf{p}_i^{\mathsf{T}}(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{v}));\mathbf{q})\\
&\qquad-[\eta(\mathbf{p}_i^{\mathsf{T}}(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2))-\eta(\mathbf{p}_i^{\mathsf{T}}(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{v}))]\psi(y_i,\eta(\mathbf{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\mathbf{q}));\mathbf{q})\\
&\quad=\int_{-\mathbf{p}_i^{\mathsf{T}}\boldsymbol{v}}^{0}[\psi(y_i,\eta(\mathbf{p}_i^{\mathsf{T}}(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+t);\mathbf{q})-\psi(y_i,\eta(\mathbf{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\mathbf{q}));\mathbf{q})]
\end{aligned}$$

$$\times \eta^{(1)}(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + t)\,\mathrm{d}t,$$

and (E.14) is changed accordingly.

**Lemma E.11** (Bounding variance)**.** *On $\mathcal{A}_1$ as in Lemma E.3, class $\mathcal{G}$ satisfies the following variance bound:*

$$\sup_{g\in\mathcal{G}} \mathbb{E}_n[\mathbb{V}[g(\mathbf{x}_i, y_i) \mid \mathbf{x}_i]] \lesssim \varepsilon_n^2 h^d (r_{1,n} + r_{2,n} + \varepsilon_n)^{(2\alpha)\wedge 1}.$$

*Proof of Lemma E.11.* This is proven by the same argument as in the proof of Theorem E.1. □

**Lemma E.12** (Complexity of class $G$)**.** *Class $\mathcal{G}$ with envelope $\varepsilon_n$ multiplied by a large enough constant satisfies the uniform entropy bound (A.3) with $A \lesssim 1/\varepsilon_n$ and $V \lesssim \log^d n$.*

Note that $A$ is not constant in this statement but it will not matter since $\log(1/\varepsilon_n) \lesssim \log n$.

*Proof of Lemma E.12.* This is a directly assumed in the modified Assumption B.2 described in Section E.2. □

**Lemma E.13** (Uniform concentration in $\mathcal{G}$)**.** *On the event $\mathcal{A}_1$, (E.18) holds.*

*Proof of Lemma E.13.* This is proven by the same argument as in the proof of Theorem E.1. □

**Lemma E.14.** *For $\widetilde{\boldsymbol{Q}}_{\mathbf{q}} := \mathbb{E}_n\Big[\mathbf{p}_i\mathbf{p}_i^\intercal \Psi_1(\mathbf{x}_i, \eta(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q})); \mathbf{q})\eta^{(1)}(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}))^2\Big]$, (E.19) holds.*

*Proof of Lemma E.14.* This is proven by the same argument as in the proof of Theorem E.1. □

**Lemma E.15.** *On $\mathcal{A}_1$, we have*

$$\begin{aligned}
&\sup_{\mathbf{q}\in\mathcal{Q},\boldsymbol{\beta}_1\in\mathcal{H}_1,\boldsymbol{\beta}_2\in\mathcal{H}_2,\boldsymbol{v}\in\mathcal{V}} \Big|\mathbb{E}_n[\mathbb{E}[\delta_{\mathbf{q},i}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{v})|\mathbf{x}_i]] - \boldsymbol{v}^\intercal \widetilde{\boldsymbol{Q}}_{\mathbf{q}}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2)\Big| \\
&\quad \lesssim \varepsilon_n^2 h^d + \varepsilon_n h^d (r_{1,n} + r_{2,n} + \varepsilon_n)^2.
\end{aligned}$$

*Proof of Lemma E.15.* First, on $\mathcal{A}_1$ the largest eigenvalue of $\widetilde{\boldsymbol{Q}}_{\mathbf{q}}$ is bounded by $h^d$ up to a constant factor (uniformly in $\mathbf{q}$), which is proven in Lemma E.7.

Next, by the Taylor expansion,

$$\begin{aligned}
&\mathbb{E}[\delta_{\mathbf{q},i}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{v})|\mathbf{x}_i] \\
&\quad = \int_{-\mathbf{p}_i^\intercal \boldsymbol{v}}^{0} [\Psi(\mathbf{x}_i, \eta(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + t); \mathbf{q}) - \Psi(\mathbf{x}_i, \eta(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q})); \mathbf{q})] \\
&\qquad \times \eta^{(1)}(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + t)\,\mathrm{d}t \\
&\quad = \int_{-\mathbf{p}_i^\intercal \boldsymbol{v}}^{0} \Big[\Psi_1(\mathbf{x}_i, \eta(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q})); \mathbf{q})\Big\{\eta^{(1)}(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}))(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + t) \\
&\qquad\quad + \frac{1}{2}\eta^{(2)}(\xi_{\mathbf{q},i,t})(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + t)^2\Big\} \\
&\qquad + \frac{1}{2}\Psi_2\Big(\mathbf{x}_i, \tilde{\xi}_{\mathbf{q},i,t}; \mathbf{q}\Big)\{\eta(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + t) - \eta(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}))\}^2\Big] \\
&\qquad \times \eta^{(1)}(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q}) + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + t)\,\mathrm{d}t
\end{aligned}$$

for some $\xi_{\mathbf{q},i,t}$ between $\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q})$ and $\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+t$, $\tilde{\xi}_{\mathbf{q},i,t}$ between $\eta(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}))$ and $\eta(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+t)$. This gives

$$\mathbb{E}_n[\mathbb{E}[\delta_{\mathbf{q},i}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{v})|\mathbf{x}_i]]=\boldsymbol{v}^\intercal\widetilde{\boldsymbol{Q}}_{\mathbf{q}}(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)-\boldsymbol{v}^\intercal\widetilde{\boldsymbol{Q}}_{\mathbf{q}}\boldsymbol{v}+\mathrm{I}+\mathrm{II}+\mathrm{III},$$

where for some $\check{\xi}_{\mathbf{q},i,t}$ between $\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q})$ and $\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+t$ again

$$\begin{aligned}
\mathrm{I}&:=\mathbb{E}_n\left[\int_{-\mathbf{p}_i^\intercal\boldsymbol{v}}^0\Psi_1(\mathbf{x}_i,\eta(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}));\mathbf{q})\eta^{(1)}(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}))\eta^{(2)}\big(\check{\xi}_{\mathbf{q},i}\big)\big(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+t\big)^2\,\mathrm{d}t\right],\\
\mathrm{II}&:=\frac{1}{2}\mathbb{E}_n\Bigg[\int_{-\mathbf{p}_i^\intercal\boldsymbol{v}}^0\Psi_1(\mathbf{x}_i,\eta(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}));\mathbf{q})\eta^{(2)}(\xi_{\mathbf{q},i,t})\\
&\qquad\times\eta^{(1)}(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+t)\big(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+t\big)^2\Bigg]\,\mathrm{d}t,\\
\mathrm{III}&:=\frac{1}{2}\mathbb{E}_n\Bigg[\int_{-\mathbf{p}_i^\intercal\boldsymbol{v}}^0\Psi_2\Big(\mathbf{x}_i,\tilde{\xi}_{\mathbf{q},i,t};\mathbf{q}\Big)\{\eta(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+t)-\eta(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}))\}^2\\
&\qquad\times\eta^{(1)}(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)+t)\,\mathrm{d}t\Bigg]
\end{aligned}$$

and

$$\begin{aligned}
&\boldsymbol{v}^\intercal\widetilde{\boldsymbol{Q}}_{\mathbf{q}}\boldsymbol{v}\lesssim\varepsilon_n^2h^d,\\
&\mathrm{I}\lesssim\varepsilon_nh^d(r_{1,n}+r_{2,n}+\varepsilon_n)^2,\\
&\mathrm{II}\lesssim\varepsilon_nh^d(r_{1,n}+r_{2,n}+\varepsilon_n)^2,\\
&\mathrm{III}\lesssim\varepsilon_nh^d(r_{1,n}+r_{2,n}+\varepsilon_n)^2
\end{aligned}$$

on the event $\mathcal{A}_1$. □

**Lemma E.16.** *There exists an event $\mathcal{A}_2$ whose probability converges to one such that on $\mathcal{A}_2$*

$$\begin{aligned}
\sup_{\mathbf{q}\in\mathcal{Q},\boldsymbol{\beta}_1\in\mathcal{H}_1,\boldsymbol{\beta}_2\in\mathcal{H}_2,\boldsymbol{v}\in\mathcal{V}}&\Big|\mathbb{E}_n[\psi(y_i,\eta(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}));\mathbf{q})\\
&\times(\eta(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2))-\eta(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{v})))]\\
&-\mathbb{E}_n\Big[\boldsymbol{v}^\intercal\mathbf{p}_i\psi(y_i,\eta(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}));\mathbf{q})\eta^{(1)}(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}))\Big]\Big|\lesssim h^dr_{1,n}(r_{1,n}+r_{2,n}+\varepsilon_n)\varepsilon_n.
\end{aligned}$$

*Proof of Lemma E.16.* By the Taylor expansion,

$$\begin{aligned}
&|\mathbb{E}_n[\psi(y_i,\eta(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}));\mathbf{q})(\eta(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2))-\eta(\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{v})))]\\
&\qquad-\mathbb{E}_n\Big[\boldsymbol{v}^\intercal\mathbf{p}_i\psi(y_i,\eta(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}));\mathbf{q})\eta^{(1)}(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}))\Big]\Big|\\
&\quad=\left|\mathbb{E}_n\left[\mathbf{p}_i^\intercal\boldsymbol{v}\psi(y_i,\eta(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}));\mathbf{q})\left\{\eta^{(2)}(\xi_{\mathbf{q},i})\mathbf{p}_i^\intercal(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{v})+\frac{1}{2}\eta^{(2)}\Big(\tilde{\xi}_{\mathbf{q},i}\Big)\mathbf{p}_i^\intercal\boldsymbol{v}\right\}\right]\right|\\
&\quad\lesssim h^dr_{1,n}(r_{1,n}+r_{2,n}+\varepsilon_n)\varepsilon_n,
\end{aligned}$$

where $\xi_{\mathbf{q},i}$ is between $\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{v})$ and $\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q})$, $\tilde{\xi}_{\mathbf{q},i}$ between $\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)$ and $\mathbf{p}_i^\intercal(\boldsymbol{\beta}_0(\mathbf{q})+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2-\boldsymbol{v})$. The bound holds on the event $\mathcal{A}_2:=\mathcal{A}_2'\cap\mathcal{A}_2''$, where

$$\mathcal{A}_2':=\left\{\sup_{\boldsymbol{\beta}_1\in\mathcal{H}_1,\boldsymbol{\beta}_2\in\mathcal{H}_2,\boldsymbol{v}\in\mathcal{V},\mathbf{q}\in\mathcal{Q}}\left\|\mathbb{E}_n\Big[\psi(y_i,\eta(\mathbf{p}_i^\intercal\boldsymbol{\beta}_0(\mathbf{q}));\mathbf{q})\eta^{(2)}(\xi_{\mathbf{q},i})\mathbf{p}_i\mathbf{p}_i^\intercal\Big]\right\|_\infty\le h^dr_{1,n}\right\},$$

and $\mathcal{A}_2''$ is defined the same way as $\mathcal{A}_2'$ with $\xi_{\mathbf{q},i}$ replaced by $\tilde{\xi}_{\mathbf{q},i}$.

By the same argument as Lemma C.12, $\mathbb{P}\{\mathcal{A}_2\}\to1$. □

The last step in the proof is essentially the same as for Theorem E.1 with the following changes. First, $\widehat{\boldsymbol{\beta}}(\mathbf{q})$ is replaced with $\widehat{\boldsymbol{\beta}}_{\text{constr}}(\mathbf{q})$ everywhere in this step.

Instead of (E.20), we have for any vector $\boldsymbol{v}$

$$\mathbb{E}_n\Big[\rho\Big(y_i, \eta\Big(\mathbf{p}_i^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_{\text{constr}}(\mathbf{q})\Big); \mathbf{q}\Big) - \rho\Big(y_i, \eta\Big(\mathbf{p}_i^{\mathsf{T}}\Big(\widehat{\boldsymbol{\beta}}_{\text{constr}}(\mathbf{q}) - \boldsymbol{v}\Big)\Big); \mathbf{q}\Big)\Big] \leq 0$$

by the definition of $\widehat{\boldsymbol{\beta}}_{\text{constr}}(\mathbf{q})$ as long as $\|\boldsymbol{v}\|_\infty$ is small enough (so that $\widehat{\boldsymbol{\beta}}_{\text{constr}}(\mathbf{q}) - \boldsymbol{v}$ satisfies the constraints).

Also, (E.21) is replaced with

$$0 \geq \mathbb{E}_n\Big[\rho\Big(y_i, \eta\Big(\mathbf{p}_i^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_{\text{constr}}(\mathbf{q})\Big); \mathbf{q}\Big) - \rho\Big(y_i, \eta\Big(\mathbf{p}_i^{\mathsf{T}}\Big(\widehat{\boldsymbol{\beta}}_{\text{constr}}(\mathbf{q}) - \boldsymbol{v}_{\mathbf{q}}\Big)\Big); \mathbf{q}\Big)\Big],$$

and the definition of $S_4$ becomes

$$\begin{aligned} S_4 := \mathbb{E}_n\Big[\Big(\eta\Big(\mathbf{p}_i^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_{\text{constr}}(\mathbf{q})\Big) - \eta\Big(\mathbf{p}_i\Big(\widehat{\boldsymbol{\beta}}_{\text{constr}}(\mathbf{q}) - \boldsymbol{v}_{\mathbf{q}}\Big)\Big)\Big)\psi(y_i, \eta(\mathbf{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\mathbf{q})); \mathbf{q})\Big] \\ - \mathbb{E}_n\Big[\boldsymbol{v}_{\mathbf{q}}^{\mathsf{T}}\mathbf{p}_i\psi(y_i, \eta(\mathbf{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\mathbf{q})); \mathbf{q})\eta^{(1)}(\mathbf{p}_i^{\mathsf{T}}\boldsymbol{\beta}_0(\mathbf{q}))\Big]. \end{aligned}$$

# F Strong approximation

The first subsection collects general results that may be of independent theoretical interest, while the second section discusses our main results on strong approximations for partitioning-based M-estimation.

## 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 [10]. We use the statement from [6] 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 $\alpha$ and $\beta$*

$$\mathbb{E}^* \sup_{A \in \mathcal{B}}\{\mathbb{P}\{X \in A \mid \mathcal{J}\} - G(A^\alpha|\mathcal{J})\} \leq \beta,$$

*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 $\mathbb{P}\{d(X, Y) > \alpha\} \leq \beta$.*

The following theorem is a conditional version of Lemma 39 in [1]; its proof carefully leverages Theorem F.1. See also [5] 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 $\{\mathbf{g}_i\}_{i=1}^n$ with values in $\mathbb{R}^k$ be defined on the same probability space and be such that the all the $2n$ vectors $\{\boldsymbol{\xi}_i\}_{i=1}^n \cup \{\mathbf{g}_i\}_{i=1}^n$ are independent conditionally on $\boldsymbol{C}$, are mean zero conditionally on $\boldsymbol{C}$ and for each $i \in \{1, \ldots, n\}$ the distribution of $\mathbf{g}_i$ conditionally on $\boldsymbol{C}$ is $\mathcal{N}(0, \mathbb{V}[\boldsymbol{\xi}_i \mid \boldsymbol{C}])$. Assume that this probability space is rich enough: there exists a random variable $U$ that is independent of $\{\boldsymbol{\xi}_i\}_{i=1}^n \cup \{\mathbf{g}_i\}_{i=1}^n$. Denote*

$$\boldsymbol{S} := \boldsymbol{\xi}_1 + \ldots + \boldsymbol{\xi}_n, \quad \boldsymbol{T} := \mathbf{g}_1 + \ldots + \mathbf{g}_n,$$

*and let*

$$\beta := \sum_{i=1}^n \mathbb{E}\Big[\|\boldsymbol{\xi}_i\|^2 \|\boldsymbol{\xi}_i\|_d\Big] + \sum_{i=1}^n \mathbb{E}\Big[\|\mathbf{g}_i\|^2 \|\mathbf{g}_i\|_d\Big]$$

*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*

$$\mathbb{P}\big\{\big\|\boldsymbol{S} - \boldsymbol{T}'\big\|_d > 3\delta\big\} \le \min_{t \ge 0}\bigg(2\mathbb{P}\{\|\boldsymbol{Z}\|_d > t\} + \frac{\beta}{\delta^3} t^2\bigg),$$

*where $\boldsymbol{Z} \sim \mathcal{N}(0, I_k)$.*

*Proof.* By the conditional Strassen's theorem, it is enough to show

$$\mathbb{E}^* \sup_{A \in \mathcal{B}(\mathbb{R}^k)} \Big(\mathbb{P}\{\boldsymbol{S} \in A \mid \sigma(\boldsymbol{C})\} - \mathbb{P}_{\boldsymbol{T}|\sigma(\boldsymbol{C})}\Big(A^{3\delta}\Big)\Big) \le \min_{t \ge 0}\bigg(2\mathbb{P}\{\|\boldsymbol{Z}\|_d > t\} + \frac{\beta}{\delta^3} t^2\bigg),$$

or, equivalently, for any $t > 0$

$$\mathbb{E}^* \sup_{A \in \mathcal{B}(\mathbb{R}^k)} \Big(\mathbb{P}\{\boldsymbol{S} \in A \mid \boldsymbol{C}\} - \mathbb{P}\Big\{\boldsymbol{T} \in A^{3\delta} \,\Big|\, \boldsymbol{C}\Big\}\Big) \le 2\mathbb{P}\{\|\boldsymbol{Z}\|_d > t\} + \frac{\beta}{\delta^3} t^2. \tag{F.1}$$

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 [1], namely, it is such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^k$,

$$\begin{aligned}
&\left|f(\mathbf{x} + \mathbf{y}) - f(\mathbf{x}) - \mathbf{y}^\intercal \nabla f(\mathbf{x}) - \frac{1}{2}\mathbf{y}^\intercal \nabla^2 f(\mathbf{x}) \mathbf{y}\right| \le \frac{\|\mathbf{y}\|^2 \|\mathbf{y}\|_d}{\sigma^2 \delta}, \\
&(1 - \epsilon)1\{\mathbf{x} \in A\} \le f(\mathbf{x}) \le \epsilon + (1 - \epsilon)1\Big\{\mathbf{x} \in A^{3\delta}\Big\},
\end{aligned}$$

with $\sigma := \delta/t$ and $\epsilon := \mathbb{P}\{\|\boldsymbol{Z}\|_d > t\}$.

Then note that

$$\begin{aligned}
\mathbb{P}\{\boldsymbol{S} \in A \mid \boldsymbol{C}\} &= \mathbb{E}[\mathbb{1}\{\boldsymbol{S} \in A\} - f(\boldsymbol{S}) \mid \boldsymbol{C}] + \mathbb{E}[f(\boldsymbol{S}) - f(\boldsymbol{T}) \mid \boldsymbol{C}] + \mathbb{E}[f(\boldsymbol{T}) \mid \boldsymbol{C}] \\
&\le \epsilon \mathbb{E}[\mathbb{1}\{\boldsymbol{S} \in A\} \mid \boldsymbol{C}] + \mathbb{E}[f(\boldsymbol{S}) - f(\boldsymbol{T}) \mid \boldsymbol{C}] + \epsilon + (1 - \epsilon)\mathbb{E}\Big[\mathbb{1}\Big\{\boldsymbol{T} \in A^{3\delta}\Big\} \,\Big|\, \boldsymbol{C}\Big] \\
&\le 2\epsilon + \mathbb{E}\Big[\mathbb{1}\Big\{\boldsymbol{T} \in A^{3\delta}\Big\} \,\Big|\, \boldsymbol{C}\Big] + \mathbb{E}[f(\boldsymbol{S}) - f(\boldsymbol{T}) \mid \boldsymbol{C}].
\end{aligned} \tag{F.2}$$

Now we bound $\mathbb{E}[f(\boldsymbol{S}) - f(\boldsymbol{T}) \mid \boldsymbol{C}]$:

$$\mathbb{E}[f(\boldsymbol{S}) - f(\boldsymbol{T}) \mid \boldsymbol{C}] = \sum_{i=1}^n \mathbb{E}[f(\boldsymbol{X}_i + \boldsymbol{Y}_i) - f(\boldsymbol{X}_i + \boldsymbol{W}_i) \mid \boldsymbol{C}]$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{n} \mathbb{E}\left[ f(\boldsymbol{X}_i) + \boldsymbol{Y}_i^{\mathsf{T}} \nabla f(\boldsymbol{X}_i) + \frac{1}{2} \boldsymbol{Y}_i^{\mathsf{T}} \nabla^2 f(\boldsymbol{X}_i) \boldsymbol{Y}_i + \frac{\|\boldsymbol{Y}_i\|^2 \|\boldsymbol{Y}_i\|_d}{\sigma^2 \delta} \,\middle|\, \boldsymbol{C} \right] \\
&\quad - \sum_{i=1}^{n} \mathbb{E}\left[ f(\boldsymbol{X}_i) + \boldsymbol{W}_i^{\mathsf{T}} \nabla f(\boldsymbol{X}_i) + \frac{1}{2} \boldsymbol{W}_i^{\mathsf{T}} \nabla^2 f(\boldsymbol{X}_i) \boldsymbol{W}_i - \frac{\|\boldsymbol{W}_i\|^2 \|\boldsymbol{W}_i\|_d}{\sigma^2 \delta} \,\middle|\, \boldsymbol{C} \right] \\
&\overset{\text{(a)}}{=} \sum_{i=1}^{n} \mathbb{E}\left[ \frac{\|\boldsymbol{Y}_i\|^2 \|\boldsymbol{Y}_i\|_d + \|\boldsymbol{W}_i\|^2 \|\boldsymbol{W}_i\|_d}{\sigma^2 \delta} \,\middle|\, \boldsymbol{C} \right] \text{ a. s.}
\end{aligned}
$$

for $\boldsymbol{X}_i := \boldsymbol{\xi}_1 + \ldots + \boldsymbol{\xi}_{i-1} + \mathbf{g}_{i+1} + \ldots + \mathbf{g}_n$, $\boldsymbol{Y}_i := \boldsymbol{\xi}_i$, $\boldsymbol{W}_i := \mathbf{g}_i$, in (a) we used the conditional independence of the family $\{\boldsymbol{\xi}_i\}_{i=1}^n \cup \{\mathbf{g}_i\}_{i=1}^n$, that they are conditionally mean zero and the equality of the corresponding conditional second moments.

We conclude that almost surely

$$
\sup_{A \in \mathcal{B}(\mathbb{R}^k)} \Big( \mathbb{P}\{\boldsymbol{S} \in A \mid \boldsymbol{C}\} - \mathbb{P}\Big\{\boldsymbol{T} \in A^{3\delta} \,\Big|\, \boldsymbol{C}\Big\} \Big) \leq 2\epsilon + \sum_{i=1}^{n} \mathbb{E}\left[ \frac{\|\boldsymbol{\xi}_i\|^2 \|\boldsymbol{\xi}_i\|_d + \|\mathbf{g}_i\|^2 \|\mathbf{g}_i\|_d}{\sigma^2 \delta} \,\middle|\, \boldsymbol{C} \right].
$$

By the definition of outer expectation, this implies

$$
\mathbb{E}^* \sup_{A \in \mathcal{B}(\mathbb{R}^k)} \Big( \mathbb{P}\{\boldsymbol{S} \in A \mid \boldsymbol{C}\} - \mathbb{P}\Big\{\boldsymbol{T} \in A^{3\delta} \,\Big|\, \boldsymbol{C}\Big\} \Big) \leq 2\epsilon + \frac{\beta}{\sigma^2 \delta},
$$

which is (F.1). □

The following theorem generalizes Lemma 36 in [1], and also builds on the argument for Lemma SA27 in [4].

**Theorem F.3** (Yurinskii coupling: $K$-dimensional process). *Let $\{\mathbf{x}_i, y_i\}_{i=1}^n$ be a random sample, where $\mathbf{x}_i$ 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 $\sup_{\mathbf{q} \in \mathcal{Q}} |A_n(\mathbf{q}, \mathbf{x}_i, y_i)| \leq \bar{A}_n(\mathbf{x}_i, y_i)$, where $\bar{A}_n(\mathbf{x}_i, y_i)$ is a Borel measurable envelope, $\mathbb{E}[A_n(\mathbf{q}, \mathbf{x}_i, y_i) \mid \mathbf{x}_i] = 0$ for all $\mathbf{q} \in \mathcal{Q}$, $\sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}\big[|\bar{A}(\mathbf{x}_i, y_i)|^\nu \,\big|\, \mathbf{x}_i = \mathbf{x}\big] \leq \mu_n < \infty$ for some $\nu \geq 3$ with $\mu_n \gtrsim 1$ and $\log \mu_n \lesssim \log n$, which satisfies the Lipschitz condition*

$$
\sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}\Big[ |A_n(\mathbf{q}, \mathbf{x}_i, y_i) - A_n(\tilde{\mathbf{q}}, \mathbf{x}_i, y_i)|^2 \,\Big|\, \mathbf{x}_i = \mathbf{x} \Big] \lesssim \|\mathbf{q} - \tilde{\mathbf{q}}\|
$$

*for all $\mathbf{q}, \tilde{\mathbf{q}} \in \mathcal{Q}$. Also, the (regular) conditional variance $\mathbb{E}\Big[A_n(\mathbf{q}, \mathbf{x}_i, y_i)^2 \,\Big|\, \mathbf{x}_i = \mathbf{x}\Big]$ is continuous in $\mathbf{x} \in \mathcal{X}$. Moreover, assume that the class of functions $\{(\mathbf{x}, y) \mapsto A_n(\mathbf{q}, \mathbf{x}, y) \colon \mathbf{q} \in \mathcal{Q}\}$ is VC-subgraph with an index bounded from above by a constant not depending on $n$.*

*Let $\mathbf{b}(\cdot)$ 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 $\sup_{\mathbf{x} \in \mathcal{X}} \|\mathbf{b}(\mathbf{x})\| \leq \zeta_K$ and the probability of the event $\mathcal{A} := \Big\{ \sup_{\|\boldsymbol{\alpha}\|=1} \mathbb{E}_n\Big[(\boldsymbol{\alpha}^{\mathsf{T}} \mathbf{b}(\mathbf{x}_i))^2\Big] \leq C_{\mathrm{Gr}} \Big\}$ approaches one, where $C_{\mathrm{Gr}}$ is some positive constant. Assume $\zeta_K$ satisfies $1/\zeta_K \lesssim 1$, $|\log \zeta_K| \lesssim \log n$.*

*Let $r_{n,\mathrm{yur}} = r_{\mathrm{yur}} \to 0$ be a sequence of positive numbers satisfying*

$$
\left( \frac{\zeta_K^3 \mu_n^{3/\nu}}{\sqrt{n}} \right)^{\frac{1}{3 + 2d_{\mathcal{Q}}}} \sqrt{\log n} + \frac{\zeta_K \mu_n^{1/\nu}}{n^{1/2 - 1/\nu}} \log n = o(r_{\mathrm{yur}}). \tag{F.3}
$$

*Assume also that the probability space (for each $n$) is rich enough: there exists a family of independent random variables $\{U_1, U_2, U_3\}$ distributed uniformly on $[0,1]$ that is independent of $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$. Then, on the same probability space, there exists a $K$-dimensional process $\boldsymbol{Z}_n(\mathbf{q})$ on $\mathcal{Q}$ which is conditionally on $\{\mathbf{x}_i\}_{i=1}^n$ a mean-zero Gaussian process with the same conditional covariance structure as that of $\boldsymbol{G}_n(\mathbf{q}) := \mathbb{G}_n[A_n(\mathbf{q}, \mathbf{x}_i, y_i)\mathbf{b}(\mathbf{x}_i)]$, meaning that for all $\mathbf{q}, \tilde{\mathbf{q}} \in \mathcal{Q}$*

$$\mathbb{E}[\boldsymbol{G}_n(\mathbf{q})\boldsymbol{G}_n(\tilde{\mathbf{q}})^\intercal \mid \{\mathbf{x}_i\}_{i=1}^n] = \mathbb{E}[\boldsymbol{Z}_n(\mathbf{q})\boldsymbol{Z}_n(\tilde{\mathbf{q}})^\intercal \mid \{\mathbf{x}_i\}_{i=1}^n], \tag{F.4}$$

*such that*

$$\sup_{\mathbf{q}\in\mathcal{Q}} \|\boldsymbol{G}_n(\mathbf{q}) - \boldsymbol{Z}_n(\mathbf{q})\|_\infty = o_{\mathbb{P}}(r_{\mathrm{yur}}).$$

*Moreover, if $\bar{A}_n(\mathbf{q}, \mathbf{x}_i, y_i)$ is $\sigma^2$-sub-Gaussian, then (F.3) can be replaced with*

$$\left(\frac{\zeta_K^3}{\sqrt{n}}\right)^{\frac{1}{3+2d_{\mathcal{Q}}}} \sqrt{\log n} + \frac{\zeta_K}{\sqrt{n}} \log^{3/2} n = o(r_{\mathrm{yur}}).$$

*Proof.* Let $\mathcal{Q}_n^\delta := \left\{\mathbf{q}_1, \dots, \mathbf{q}_{|\mathcal{Q}_n^\delta|}\right\}$ be an internal $\delta_n$-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 $\delta_n$ 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).

The plan of attack is to

1. show that $\boldsymbol{G}_n(\mathbf{q})$ does not deviate too much in sup-norm from its projected version, i. e. bound the tails of $\sup_{\mathbf{q}\in\mathcal{Q}} \left\|\boldsymbol{G}_n(\mathbf{q}) - \boldsymbol{G}_n \circ \pi_n^\delta(\mathbf{q})\right\|_\infty$,
2. apply Yurinskii coupling to the finite-dimensional vector $\left(\boldsymbol{G}_n \circ \pi_n^\delta(\mathbf{q})\right)_{\mathbf{q}\in\mathcal{Q}_n^\delta}$ and obtain a conditionally on $\{\mathbf{x}_i\}_{i=1}^n$ Gaussian vector $\boldsymbol{Z}_n^\delta$ with the right structure that is close enough, i. e. with a bound on the tails of $\left\|\boldsymbol{G}_n \circ \pi_n^\delta - \boldsymbol{Z}_n^\delta\right\|_\infty$,
3. extend this conditionally Gaussian vector to a $K$-dimensional conditionally Gaussian process $\boldsymbol{Z}_n$,
4. and finally show that $\boldsymbol{Z}_n(\mathbf{q})$ does not deviate too much from its projected version, i. e. bound the tails of $\sup_{\mathbf{q}\in\mathcal{Q}} \left\|\boldsymbol{Z}_n(\mathbf{q}) - \boldsymbol{Z}_n \circ \pi_n^\delta(\mathbf{q})\right\|_\infty$.

If we complete these steps, it will prove the lemma by the triangle inequality.

**Discretization of $G_n$** Consider the class of functions

$$\mathcal{G}_n' := \{\mathcal{X} \times \mathcal{Y} \ni (\mathbf{x}, y) \mapsto A_n(\mathbf{q}, \mathbf{x}, y) b_l(\mathbf{x}) \colon 1 \le l \le K, \mathbf{q} \in \mathcal{Q}\}$$

with envelope $\zeta_K \bar{A}_n(X)$. Since $\{A_n(\mathbf{q}, \mathbf{x}, y)\}$ is a VC class with $O(1)$ index and envelope $\bar{A}_n(\mathbf{x}, y)$, $\mathcal{G}_n'$ satisfies the uniform entropy bound (A.3) with $A \lesssim K$ and $V \lesssim 1$.

Next, consider the class of functions

$$\mathcal{G}_n^\delta := \{\mathcal{X} \times \mathcal{Y} \ni (\mathbf{x}, y) \mapsto (A_n(\mathbf{q}, \mathbf{x}, y) - A_n(\tilde{\mathbf{q}}, \mathbf{x}, y)) b_l(\mathbf{x}) \colon 1 \le l \le K, \mathbf{q}, \tilde{\mathbf{q}} \in \mathcal{Q}, \|\mathbf{q} - \tilde{\mathbf{q}}\| \lesssim \delta_n\}$$

with envelope $(\mathbf{x}, y) \mapsto 2\zeta_K \bar{A}_n(\mathbf{x}, y)$. 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 $\{\mathbf{x}_i\}_{i=1}^n$ on $\mathcal{A}$ with $\|F\|_{\bar{\mathbb{P}},2} \le 2\zeta_K \mu_n^{1/\nu}$ since

$$\|\bar{A}_n(\mathbf{x}, y)\|_{\bar{\mathbb{P}},2}^2 = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[\bar{A}_n(\mathbf{x}_i, y_i)^2 \,\big|\, \mathbf{x}_i\big] \le \frac{1}{n}\sum_{i=1}^n \mu_n^{2/\nu} = \mu_n^{2/\nu},$$

$\|M\|_{\mathbb{P},2} \le 2\zeta_K(\mu_n n)^{1/\nu}$ since

$$\begin{aligned}
\mathbb{E}\left[\left(\max_{1\le i\le n} \bar{A}_n(\mathbf{x}_i, y_i)\right)^2 \,\middle|\, \{\mathbf{x}_i\}_{i=1}^n\right] &\le \mathbb{E}\left[\left(\max_{1\le i\le n} \bar{A}_n(\mathbf{x}_i, y_i)\right)^\nu \,\middle|\, \{\mathbf{x}_i\}_{i=1}^n\right]^{2/\nu} \\
&\le \mathbb{E}\left[\sum_{i=1}^n \bar{A}_n(\mathbf{x}_i, y_i)^\nu \,\middle|\, \{\mathbf{x}_i\}_{i=1}^n\right]^{2/\nu} \le (\mu_n n)^{2/\nu},
\end{aligned}$$

and $\sigma \lesssim \sqrt{\delta_n}$ since on $\mathcal{A}$

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[(A_n(\mathbf{q}, \mathbf{x}_i, y_i) - A_n(\tilde{\mathbf{q}}, \mathbf{x}_i, y_i))^2 b_l(\mathbf{x}_i)^2 \,\Big|\, \{\mathbf{x}_i\}_{i=1}^n\Big] \\
&\quad = \frac{1}{n}\sum_{i=1}^n b_l(\mathbf{x}_i)^2 \mathbb{E}\Big[(A_n(\mathbf{q}, \mathbf{x}_i, y_i) - A_n(\tilde{\mathbf{q}}, \mathbf{x}_i, y_i))^2 \,\Big|\, \{\mathbf{x}_i\}_{i=1}^n\Big] \\
&\quad \lesssim \|\mathbf{q} - \tilde{\mathbf{q}}\| \frac{1}{n}\sum_{i=1}^n b_l(\mathbf{x}_i)^2 \lesssim \|\mathbf{q} - \tilde{\mathbf{q}}\|.
\end{aligned}$$

This gives that on $\mathcal{A}$

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\|\mathbf{q}-\tilde{\mathbf{q}}\|\le\delta_n} \|\boldsymbol{G}_n(\mathbf{q}) - \boldsymbol{G}_n(\tilde{\mathbf{q}})\|_\infty \,\middle|\, \{\mathbf{x}_i\}_{i=1}^n\right] \\
&\quad \lesssim \sqrt{\delta_n \log \frac{K\zeta_K \mu_n^{1/\nu}}{\sqrt{\delta_n}}} + \frac{\zeta_K(\mu_n n)^{1/\nu}}{\sqrt{n}} \log \frac{K\zeta_K \mu_n^{1/\nu}}{\sqrt{\delta_n}}.
\end{aligned}$$

By Markov's inequality and since $\mathbb{P}\{\mathcal{A}\} \to 1$, for any sequence $t_n > 0$

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{\|\mathbf{q}-\tilde{\mathbf{q}}\|\le\delta_n} \|\boldsymbol{G}_n(\mathbf{q}) - \boldsymbol{G}_n(\tilde{\mathbf{q}})\|_\infty > t_n\right\} \\
&\quad \lesssim \frac{1}{t_n}\sqrt{\delta_n \log \frac{K\zeta_K \mu_n^{1/\nu}}{\sqrt{\delta_n}}} + \frac{\zeta_K(\mu_n n)^{1/\nu}}{t_n\sqrt{n}} \log \frac{K\zeta_K \mu_n^{1/\nu}}{\sqrt{\delta_n}} + o(1),
\end{aligned} \tag{F.5}$$

where the constant in $\lesssim$ does not depend on $n$.

**Coupling** Define a $K|\mathcal{Q}_n^\delta|$-dimensional vector $\boldsymbol{\xi}_i := (A_n(\mathbf{q}, \mathbf{x}_i, y_i) b_l(\mathbf{x}_i)/\sqrt{n})_{1\le l\le K, \mathbf{q}\in\mathcal{Q}_n^\delta}$, 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 $\mathbb{E}\Big[\|\boldsymbol{\xi}_i\|^2 \|\boldsymbol{\xi}_i\|_\infty \,\Big|\, \{\mathbf{x}_i\}_{i=1}^n\Big]$.

$$\sum_{i=1}^n \mathbb{E}\Big[\|\boldsymbol{\xi}_i\|^2 \|\boldsymbol{\xi}_i\|_\infty \,\Big|\, \{\mathbf{x}_i\}_{i=1}^n\Big]$$

$$
\begin{aligned}
&= \frac{1}{n^{3/2}} \sum_{i=1}^n \|\mathbf{b}(\mathbf{x}_i)\|^2 \|\mathbf{b}(\mathbf{x}_i)\|_\infty \mathbb{E}\left[\sum_{\mathbf{q}\in\mathcal{Q}_n^\delta} A_n(\mathbf{q}, \mathbf{x}_i, y_i)^2 \cdot \max_{\mathbf{q}\in\mathcal{Q}_n^\delta} |A_n(\mathbf{q}, \mathbf{x}_i, y_i)| \;\middle|\; \mathbf{x}_i\right] \\
&\le \frac{1}{n^{3/2}} |\mathcal{Q}_n^\delta| \sum_{i=1}^n \|\mathbf{b}(\mathbf{x}_i)\|^2 \|\mathbf{b}(\mathbf{x}_i)\|_\infty \mathbb{E}\big[\bar{A}_n(\mathbf{x}_i, y_i)^3 \;\big|\; \mathbf{x}_i\big] \\
&\le \frac{|\mathcal{Q}_n^\delta| \mu_n^{3/\nu}}{n^{3/2}} \sum_{i=1}^n \|\mathbf{b}(\mathbf{x}_i)\|^2 \|\mathbf{b}(\mathbf{x}_i)\|_\infty \lesssim \frac{|\mathcal{Q}_n^\delta| \mu_n^{3/\nu}}{n^{1/2}} \zeta_K^3.
\end{aligned}
$$

Secondly, for $i \in \{1, \dots, n\}$ let $\mathbf{g}_i \sim \mathcal{N}(0, \Sigma_i)$ be independent vectors, where $\Sigma_i = \mathbb{V}[\boldsymbol{\xi}_i \mid \mathbf{x}_i]$. Since there is an independent random variable $U_1$ distributed uniformly on $[0, 1]$, we can construct the family $\{\mathbf{g}_i\}$ on the same probability space. Then by Jensen's inequality for any $\lambda > 0$ we have

$$
\begin{aligned}
\mathbb{E}\Big[\|\mathbf{g}_i\|_\infty^2 \;\Big|\; \mathbf{x}_i\Big] &\le \frac{1}{\lambda} \log \mathbb{E}\Big[e^{\lambda \|\mathbf{g}_i\|_\infty^2} \;\Big|\; \mathbf{x}_i\Big] \le \frac{1}{\lambda} \log \sum_{t=1}^{K|\mathcal{Q}_n^\delta|} \mathbb{E}\Big[e^{\lambda (g_{it})^2} \;\Big|\; \mathbf{x}_i\Big] \\
&\le \frac{-\frac{1}{2}\log\Big(1 - \frac{2\lambda}{n}\zeta_K^2 \mu_n^{2/\nu}\Big) + \log K + \log\big|\mathcal{Q}_n^\delta\big|}{\lambda} \lesssim \frac{\zeta_K^2 \mu_n^{2/\nu}}{n}\Big(\log K + \log|\mathcal{Q}_n^\delta|\Big),
\end{aligned}
$$

where we used the moment-generating function of $\chi_1^2$: $\mathbb{E}\big[\exp\{\alpha\chi_1^2\}\big] = (1 - 2\alpha)^{-1/2}$ for $\alpha < 1/2$, the bound $\mathbb{V}[\xi_{it} \mid \mathbf{x}_i] \le \zeta_K^2 \mu_n^{2/\nu} \big/ n$, and put $\lambda := \Big(4\zeta_K^2 \mu_n^{2/\nu}/n\Big)^{-1}$. Also,

$$
\begin{aligned}
\mathbb{E}\Big[\|\mathbf{g}_i\|^4 \;\Big|\; \mathbf{x}_i\Big]^{1/2} &= \mathbb{E}\left[\left(\sum_{t=1}^{K|\mathcal{Q}_n^\delta|} g_{it}^2\right)^2 \;\middle|\; \mathbf{x}_i\right]^{1/2} \le \sum_{t=1}^{K|\mathcal{Q}_n^\delta|} \mathbb{E}\big[g_{it}^4 \mid \mathbf{x}_i\big]^{1/2} \\
&= \sum_{t=1}^{K|\mathcal{Q}_n^\delta|} \sqrt{3}\mathbb{E}\big[g_{it}^2 \mid \mathbf{x}_i\big] \lesssim \mathbb{E}\Big[\|\mathbf{g}_i\|^2 \;\Big|\; \mathbf{x}_i\Big] = \mathbb{E}\Big[\|\boldsymbol{\xi}_i\|^2 \;\Big|\; \mathbf{x}_i\Big],
\end{aligned}
$$

which gives

$$
\sum_{i=1}^n \mathbb{E}\Big[\|\mathbf{g}_i\|^4 \;\Big|\; \mathbf{x}_i\Big]^{1/2} \lesssim \sum_{i=1}^n \mathbb{E}\Big[\|\boldsymbol{\xi}_i\|^2 \;\Big|\; \mathbf{x}_i\Big] \le \mu_n^{2/\nu} |\mathcal{Q}_n^\delta| \mathbb{E}_n\big[\|\mathbf{b}(\mathbf{x}_i)\|^2\big] \lesssim \zeta_K^2 \mu_n^{2/\nu} |\mathcal{Q}_n^\delta|.
$$

Therefore, by the Cauchy-Schwarz inequality,

$$
\begin{aligned}
\sum_{i=1}^n \mathbb{E}\Big[\|\mathbf{g}_i\|^2 \|\mathbf{g}_i\|_\infty \;\Big|\; \mathbf{x}_i\Big] &\le \sum_{i=1}^n \mathbb{E}\Big[\|g_i\|^4 \;\Big|\; \mathbf{x}_i\Big]^{1/2} \mathbb{E}\Big[\|g_i\|_\infty^2 \;\Big|\; \mathbf{x}_i\Big]^{1/2} \\
&\lesssim \frac{\zeta_K^3 \mu_n^{3/\nu}}{\sqrt{n}} |\mathcal{Q}_n^\delta| \sqrt{\log(K|\mathcal{Q}_n^\delta|)}.
\end{aligned}
$$

Now, since there exists a random variable $U_2$ independent of $\{\boldsymbol{\xi}_i\}_{i=1}^n \cup \{\mathbf{g}_i\}_{i=1}^n$, applying Theorem F.2 with

$$
\beta := \sum_{i=1}^n \mathbb{E}\Big[\|\boldsymbol{\xi}_i\|^2 \|\boldsymbol{\xi}_i\|_\infty\Big] + \sum_{i=1}^n \mathbb{E}\Big[\|\mathbf{g}_i\|^2 \|\mathbf{g}_i\|_\infty\Big]
$$

$$\lesssim \frac{\zeta_K^3 \mu_n^{3/\nu}}{\sqrt{n}} |\mathcal{Q}_n^\delta| \sqrt{\log(K|\mathcal{Q}_n^\delta|)}$$
$$\lesssim \frac{\zeta_K^3 \mu_n^{3/\nu}}{n^{1/2} \delta_n^{d_\mathcal{Q}}} \sqrt{\log \frac{K}{\delta_n^{d_\mathcal{Q}}}}$$

gives that for any $t_n > 0$, on the same probability space there exists a vector $\boldsymbol{Z}_n^\delta \sim \mathcal{N}\big(0, \mathbb{V}[\boldsymbol{G}_n \circ \pi_n^\delta \mid \{\mathbf{x}_i\}_{i=1}^n]\big)$, generally different for different $t_n$, such that

$$\mathbb{P}\Big\{\Big\|\boldsymbol{G}_n \circ \pi_n^\delta - \boldsymbol{Z}_n^\delta\Big\|_\infty > 3t_n\Big\} \leq \min_{s \geq 0}\Big\{2\mathbb{P}\{\|\boldsymbol{N}\|_\infty > s\} + \frac{\beta}{t_n^3} s^2\Big\},$$

where $\boldsymbol{N}$ is a $K|\mathcal{Q}_n^\delta|$-dimensional standard Gaussian vector. By the union bound,

$$\mathbb{P}\{\|\boldsymbol{N}\|_\infty > s\} \leq 2K|\mathcal{Q}_n^\delta| e^{-s^2/2},$$

so by taking $s := C\sqrt{\log(K|\mathcal{Q}_n^\delta|)}$ for a positive constant $C$ not depending on $n$ (chosen later), we have

$$\mathbb{P}\Big\{\Big\|\boldsymbol{G}_n \circ \pi_n^\delta - \boldsymbol{Z}_n^\delta\Big\|_\infty > 3t_n\Big\} \lesssim \Big(K|\mathcal{Q}_n^\delta|\Big)^{1-C^2/2} + \frac{\beta}{t_n^3} C^2 \log\Big(K|\mathcal{Q}_n^\delta|\Big)$$
$$\lesssim \left(\frac{K}{\delta_n^{d_\mathcal{Q}}}\right)^{1-C^2/2} + \frac{\beta}{t_n^3} C^2 \log \frac{K}{\delta_n^{d_\mathcal{Q}}}$$
$$\lesssim \left(\frac{K}{\delta_n^{d_\mathcal{Q}}}\right)^{1-C^2/2} + C^2 \frac{\zeta_K^3 \mu_n^{3/\nu}}{t_n^3 \sqrt{n} \delta_n^{d_\mathcal{Q}}} \log^{3/2} \frac{K}{\delta_n^{d_\mathcal{Q}}},$$

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 $\boldsymbol{G}_n(\mathbf{q})$ given $\{\mathbf{x}_i\}_{i=1}^n = \boldsymbol{X}$. It follows from Kolmogorov's continuity criterion that this process can be defined on $\mathcal{C}(\mathcal{Q})^K$. The laws of such processes define a family of Gaussian probability measures on the Borel $\sigma$-algebra $\mathcal{B}\big(\mathcal{C}(\mathcal{Q})^K\big)$ of the space $\mathcal{C}(\mathcal{Q})^K$, denoted $\{\mathbb{P}_{\boldsymbol{X}}\}_{\boldsymbol{X} \in \mathcal{X}^n}$. In order to construct one process that is conditionally on $\{\mathbf{x}_i'\}_{i=1}^n$, a mean zero Gaussian process with the same conditional covariance structure as that of $\boldsymbol{G}_n(\mathbf{q})$, where $\{\mathbf{x}_i'\}_{i=1}^n$ is a copy of $\{\mathbf{x}_i\}_{i=1}^n$, we need to show that this family of measures is a probability kernel as a function $\mathcal{X}^n \times \mathcal{B}\big(\mathcal{C}(\mathcal{Q})^K\big) \to [0,1]$. This follows by a standard argument: we can take a $\pi$-system of sets, generating $\mathcal{B}\big(\mathcal{C}(\mathcal{Q})^K\big)$, of the form $B = \big\{\mathbf{f} \in \mathcal{C}(\mathcal{Q})^K : \mathbf{f}(\mathbf{q}_1) \in B_1, \ldots, \mathbf{f}(\mathbf{q}_m) \in B_m\big\}$, where $m \in \{1, 2, \ldots\}$, $\mathbf{q}_j \in \mathcal{Q}$, 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 $A \in \mathcal{B}\big(\mathcal{C}(\mathcal{Q})^K\big)$ such that $\boldsymbol{X} \mapsto \mathbb{P}_{\boldsymbol{X}}(A)$ is a Borel function form a $\lambda$-system. It is left to apply the monotone class theorem.

We have shown that there exists a law on $\mathcal{X}^n \times \mathcal{C}(\mathcal{Q})^K$ which is the joint law of $\{\mathbf{x}_i'\}_{i=1}^n$ and a conditionally on $\{\mathbf{x}_i'\}_{i=1}^n$ mean zero Gaussian process with the same conditional covariance structure as that of $\boldsymbol{G}_n(\mathbf{q})$. Projecting this $\mathcal{C}(\mathcal{Q})^K$-process on $\mathcal{Q}_n^\delta$ and adding the resulting (conditionally Gaussian) vector as the middle coordinate, we obtain a law $\mathbb{P}_{123}$ on the Polish space $\mathcal{X}^n \times \mathbb{R}^{K|\mathcal{Q}_n^\delta|} \times \mathcal{C}(\mathcal{Q})^K$ with projection $\mathbb{P}_{12}$ on the first two spaces, where $\mathbb{P}_{12}$ is the law of $\big(\{\mathbf{x}_i\}_{i=1}^n, \boldsymbol{Z}_n^\delta\big)$. Since there exists a random variable $U_3$ independent of $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \cup \{\boldsymbol{Z}_n^\delta\}$ with a continuous distribution function, we can apply Lemma 3.7.3 in [8] with $V$ denoting $\big(\{\mathbf{x}_i\}_{i=1}^n, \boldsymbol{Z}_n^\delta\big)$, and obtain that there exists a random element $W$ with values in $\mathcal{C}(\mathcal{Q})^K$ on the original probability space such that the law of $\big(\{\mathbf{x}_i\}_{i=1}^n, \boldsymbol{Z}_n^\delta, W\big)$ is exactly $\mathbb{P}_{123}$. This is what we are looking for: $W = \{\boldsymbol{Z}_n(\mathbf{q})\}_{\mathbf{q}\in\mathcal{Q}}$ 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 $t = (l, \mathbf{q}, \tilde{\mathbf{q}}) \in T$ with

$$T := \{(l, \mathbf{q}, \tilde{\mathbf{q}})\colon l \in \{1, 2, \dots, K\}, \mathbf{q}, \tilde{\mathbf{q}} \in \mathcal{Q}, \|\mathbf{q} - \tilde{\mathbf{q}}\| \le \delta_n\}$$

as $X_{n,t} := Z_{n,l}(\mathbf{q}) - Z_{n,l}(\tilde{\mathbf{q}})$. It is a separable (because each $Z_{n,l}(\cdot)$ has a. s. continuous trajectories) mean-zero Gaussian conditionally on $\{\mathbf{x}_i\}_{i=1}^n$ process with the index set $T$ considered a metric space: $\operatorname{dist}((l, \mathbf{q}, \tilde{\mathbf{q}}), (l', \mathbf{q}', \tilde{\mathbf{q}}')) = \|\mathbf{q} - \mathbf{q}'\| + \|\tilde{\mathbf{q}} - \tilde{\mathbf{q}}'\| + \mathbb{1}\{l \ne l'\}$.

We will apply Lemma C.7 to this process. Note that on the event $\mathcal{A}$

$$\begin{aligned}
\sigma(X_n)^2 &:= \sup_{t\in T} \mathbb{E}\big[X_{n,t}^2 \;\big|\; \{\mathbf{x}_i\}_{i=1}^n\big] = \sup_{\|\mathbf{q}-\tilde{\mathbf{q}}\|\le\delta_n} \max_l \mathbb{E}\Big[(Z_{n,l}(\mathbf{q}) - Z_{n,l}(\tilde{\mathbf{q}}))^2 \;\Big|\; \{\mathbf{x}_i\}_{i=1}^n\Big] \\
&= \sup_{\|\mathbf{q}-\tilde{\mathbf{q}}\|\le\delta_n} \max_l \mathbb{E}\Big[(G_{n,l}(\mathbf{q}) - G_{n,l}(\tilde{\mathbf{q}}))^2 \;\Big|\; \{\mathbf{x}_i\}_{i=1}^n\Big] \\
&= \sup_{\|\mathbf{q}-\tilde{\mathbf{q}}\|\le\delta_n} \max_l \frac{1}{n}\sum_{i=1}^n b_l(\mathbf{x}_i)^2 \mathbb{E}\Big[(A_n(\mathbf{q}, \mathbf{x}_i, y_i) - A_n(\tilde{\mathbf{q}}, \mathbf{x}_i, y_i))^2 \;\Big|\; \mathbf{x}_i\Big] \\
&\lesssim \sup_{\|\mathbf{q}-\tilde{\mathbf{q}}\|\le\delta_n} \max_l \frac{1}{n}\sum_{i=1} b_l(\mathbf{x}_i)^2 \|\mathbf{q} - \tilde{\mathbf{q}}\| \le \delta_n \max_l \mathbb{E}_n\big[b_l(\mathbf{x}_i)^2\big] \\
&\le \delta_n \underbrace{\sup_{\|\boldsymbol{\alpha}\|=1} \mathbb{E}_n\Big[(\boldsymbol{\alpha}^\intercal \mathbf{b}(\mathbf{x}_i))^2\Big]}_{\le C_{\mathrm{Gr}}} \lesssim \delta_n.
\end{aligned}$$

Next, we define and bound the semimetric $\rho(t, t')$:

$$\begin{aligned}
\rho(t, t')^2 &:= \mathbb{E}\Big[\big(X_{n,t} - X_{n,t'}\big)^2 \;\Big|\; \{\mathbf{x}_i\}_{i=1}^n\Big] \\
&= \mathbb{E}\Big[\big((Z_{n,l}(\mathbf{q}) - Z_{n,l}(\tilde{\mathbf{q}})) - \big(Z_{n,l'}(\mathbf{q}') - Z_{n,l'}(\tilde{\mathbf{q}}')\big)\big)^2 \;\Big|\; \{\mathbf{x}_i\}_{i=1}^n\Big] \\
&\lesssim \mathbb{E}\Big[\big(Z_{n,l}(\mathbf{q}) - Z_{n,l'}(\mathbf{q}')\big)^2 \;\Big|\; \{\mathbf{x}_i\}_{i=1}^n\Big] + \mathbb{E}\Big[\big(Z_{n,l}(\tilde{\mathbf{q}}) - Z_{n,l'}(\tilde{\mathbf{q}}')\big)^2 \;\Big|\; \{\mathbf{x}_i\}_{i=1}^n\Big].
\end{aligned}$$

The first term on the right is bounded the following way: if $l \ne l'$,

$$\begin{aligned}
&\mathbb{E}\Big[\big(Z_{n,l}(\mathbf{q}) - Z_{n,l'}(\mathbf{q}')\big)^2 \;\Big|\; \{\mathbf{x}_i\}_{i=1}^n\Big] = \mathbb{E}\Big[\big(G_{n,l}(\mathbf{q}) - G_{n,l'}(\mathbf{q}')\big)^2 \;\Big|\; \{\mathbf{x}_i\}_{i=1}^n\Big] \\
&\quad= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\big(A_n(\mathbf{q}, \mathbf{x}_i, y_i) b_l(\mathbf{x}_i) - A_n(\mathbf{q}', \mathbf{x}_i, y_i) b_{l'}(\mathbf{x}_i)\big)^2 \;\Big|\; \mathbf{x}_i\Big] \\
&\quad\lesssim \frac{1}{n}\sum_{i=1}^n b_l(\mathbf{x}_i)^2 \mathbb{E}\Big[\big(A_n(\mathbf{q}, \mathbf{x}_i, y_i) - A_n(\mathbf{q}', \mathbf{x}_i, y_i)\big)^2 \;\Big|\; \mathbf{x}_i\Big] \\
&\qquad+ \frac{1}{n}\sum_{i=1}^n \big(b_l(\mathbf{x}_i)^2 + b_{l'}(\mathbf{x}_i)^2\big) \mathbb{E}\big[A_n(\mathbf{q}', \mathbf{x}_i, y_i)^2 \;\big|\; \mathbf{x}_i\big] \\
&\quad\lesssim \|\mathbf{q} - \mathbf{q}'\| \mathbb{E}_n\big[b_l(\mathbf{x}_i)^2\big] + \mu_n^{2/\nu}\big(\mathbb{E}_n\big[b_l(\mathbf{x}_i)^2\big] + \mathbb{E}_n\big[b_{l'}(\mathbf{x}_i)^2\big]\big) \lesssim \|\mathbf{q} - \mathbf{q}'\| + \mu_n^{2/\nu}.
\end{aligned}$$

Similarly, if $l = l'$, $\mathbb{E}\Big[\big(Z_{n,l}(\mathbf{q}) - Z_{n,l'}(\mathbf{q}')\big)^2 \,\Big|\, \{\mathbf{x}_i\}_{i=1}^n\Big] \lesssim \|\mathbf{q} - \mathbf{q}'\|$.

The term $\mathbb{E}\Big[\big(Z_{n,l}(\tilde{\mathbf{q}}) - Z_{n,l'}(\tilde{\mathbf{q}}')\big)^2 \,\Big|\, \{\mathbf{x}_i\}_{i=1}^n\Big]$ is bounded the same way, and we conclude

$$\rho(t,t')^2 \lesssim \mu_n^{2/\nu}\big(\|\mathbf{q} - \mathbf{q}'\| + \|\tilde{\mathbf{q}} - \tilde{\mathbf{q}}'\| + \mathbb{1}\{l \neq l'\}\big) = \mu_n^{2/\nu}\mathrm{dist}\big((l, \mathbf{q}, \tilde{\mathbf{q}}), (l', \mathbf{q}', \tilde{\mathbf{q}}')\big).$$

In other words, we have proven that for some positive constant $C_{12}$ we have on $\mathcal{A}$

$$\rho(t,t')^2 \leq C_{12}\mu_n^{2/\nu}\mathrm{dist}\big((l, \mathbf{q}, \tilde{\mathbf{q}}), (l', \mathbf{q}', \tilde{\mathbf{q}}')\big).$$

This means that an $\Big(\varepsilon/\mu_n^{1/\nu}\sqrt{C_{12}}\Big)^2$-covering of $T$ with respect to $\mathrm{dist}(\cdot)$ induces an $\varepsilon$-covering of $T$ with respect to $\rho$, and hence

$$N(T, \rho, \varepsilon) \leq N\left(T, \mathrm{dist}(\cdot), \left(\frac{\varepsilon}{\mu_n^{1/\nu}\sqrt{C_{12}}}\right)^2\right). \tag{F.6}$$

Combining this with (F.6) we get

$$\log N(T, \rho, \varepsilon) \lesssim \log\Big(K\mu_n^{1/\nu}/\varepsilon\Big).$$

Applying Lemma C.7 gives that on the event $\mathcal{A}$

$$\begin{aligned}
\mathbb{E}\Big[\sup_{t \in T}|X_{n,t}| \,\Big|\, \{\mathbf{x}_i\}_{i=1}^n\Big] &\lesssim \sigma(X_n) + \int_0^{2\sigma(X_n)} \sqrt{\log N(T, \rho, \varepsilon)}\,\mathrm{d}\varepsilon \\
&\lesssim \sigma(X_n) + \sigma(X_n)\sqrt{\log\Big(K\mu_n^{1/\nu}/\sigma(X_n)\Big)} \overset{\text{(a)}}{\lesssim} \left(\delta_n \log\left(K\frac{\mu_n^{1/\nu}}{\delta_n}\right)\right)^{1/2},
\end{aligned}$$

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 $\mathcal{A}$

$$\mathbb{P}\left\{\sup_{\|\mathbf{q}-\tilde{\mathbf{q}}\| \leq \delta_n}\|\boldsymbol{Z}_n(\mathbf{q}) - \boldsymbol{Z}_n(\tilde{\mathbf{q}})\|_\infty > t_n \,\middle|\, \{\mathbf{x}_i\}_{i=1}^n\right\} \lesssim \frac{1}{t_n}\sqrt{\delta_n \log\frac{K\mu_n^{1/\nu}}{\delta_n}},$$

where the constant in $\lesssim$ does not depend on $n$. Since $\mathbb{P}\{\mathcal{A}\} \to 1$, this implies

$$\mathbb{P}\left\{\sup_{\|\mathbf{q}-\tilde{\mathbf{q}}\| \leq \delta_n}\|\boldsymbol{Z}_n(\mathbf{q}) - \boldsymbol{Z}_n(\tilde{\mathbf{q}})\|_\infty > t_n\right\} \lesssim \frac{1}{t_n}\sqrt{\delta_n \log\frac{K\mu_n^{1/\nu}}{\delta_n}} + o(1). \tag{F.7}$$

**Choosing $\delta_n$ and conclusion** Combining the bounds obtained above, for any given positive sequence $t_n$ and any constant $C > 0$ of our choice

$$\begin{aligned}
\mathbb{P}\left\{\sup_{\mathbf{q} \in \mathcal{Q}}\|\boldsymbol{G}_n(\mathbf{q}) - \boldsymbol{Z}_n(\mathbf{q})\|_\infty > t_n\right\} &\leq \mathbb{P}\left\{\sup_{\|\mathbf{q}-\tilde{\mathbf{q}}\| \leq \delta_n}\|\boldsymbol{G}_n(\mathbf{q}) - \boldsymbol{G}_n(\tilde{\mathbf{q}})\|_\infty > t_n/3\right\} \\
&+ \mathbb{P}\left\{\Big\|\boldsymbol{G}_n \circ \pi_n^\delta - \boldsymbol{Z}_n^\delta\Big\|_\infty > t_n/3\right\} + \mathbb{P}\left\{\sup_{\|\mathbf{q}-\tilde{\mathbf{q}}\| \leq \delta_n}\|Z_n(\mathbf{q}) - Z_n(\tilde{\mathbf{q}})\|_\infty > t_n/3\right\}
\end{aligned}$$

$$
\begin{aligned}
&\lesssim \frac{1}{t_n}\sqrt{\delta_n \log \frac{K\zeta_K \mu_n^{1/\nu}}{\sqrt{\delta_n}}} + \frac{\zeta_K(\mu_n n)^{1/\nu}}{t_n\sqrt{n}} \log \frac{K\zeta_K \mu_n^{1/\nu}}{\sqrt{\delta_n}} \\
&\quad + \left(\frac{K}{\delta_n^{d_{\mathcal{Q}}}}\right)^{1-C^2/2} + C^2 \frac{\zeta_K^3 \mu_n^{3/\nu}}{t_n^3\sqrt{n}\delta_n^{d_{\mathcal{Q}}}} \log^{3/2} \frac{K}{\delta_n^{d_{\mathcal{Q}}}} \\
&\quad + \frac{1}{t_n}\sqrt{\delta_n \log \frac{K\mu_n^{1/\nu}}{\delta_n}} + o(1),
\end{aligned}
$$

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 $\delta_n$. 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

$$
\left(\frac{\zeta_K^3 \mu_n^{3/\nu}}{\sqrt{n}}\right)^{\frac{1}{3+2d_{\mathcal{Q}}}} \sqrt{\log n} + \frac{\zeta_K \mu_n^{1/\nu}}{n^{1/2-1/\nu}} \log n = o(\ell_n r_{\mathrm{yur}}).
$$

Putting $t_n := \ell_n r_{\mathrm{yur}}$, we get

$$
\mathbb{P}\left\{\sup_{\mathbf{q}\in\mathcal{Q}} \|\boldsymbol{G}_n(\mathbf{q}) - \boldsymbol{Z}_n(\mathbf{q})\|_\infty > \ell_n r_{\mathrm{yur}}\right\} = o(1).
$$

Fix $\varepsilon > 0$. For $n$ large enough, $\ell_n < \varepsilon$. Then for these $n$ we have

$$
\mathbb{P}\left\{\sup_{\mathbf{q}\in\mathcal{Q}} \|\boldsymbol{G}_n(\mathbf{q}) - \boldsymbol{Z}_n(\mathbf{q})\|_\infty > \varepsilon r_{\mathrm{yur}}\right\} \le \mathbb{P}\left\{\sup_{\mathbf{q}\in\mathcal{Q}} \|\boldsymbol{G}_n(\mathbf{q}) - \boldsymbol{Z}_n(\mathbf{q})\|_\infty > \ell_n r_{\mathrm{yur}}\right\} = o(1).
$$

Theorem 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 Theorem F.3. To simplify exposition, the following notation will be helpful from this point onwards:

$$
\hat{\boldsymbol{\ell}}(\mathbf{x},\mathbf{q},\mathbf{v}) := -h^{d/2} \frac{\widehat{\boldsymbol{Q}}_{\mathbf{q}}^{-1}\mathbf{p}^{(\mathbf{v})}(\mathbf{x})}{\sqrt{\widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}}, \tag{F.8}
$$

$$
\bar{\boldsymbol{\ell}}(\mathbf{x},\mathbf{q},\mathbf{v}) := -h^{d/2} \frac{\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1}\mathbf{p}^{(\mathbf{v})}(\mathbf{x})}{\sqrt{\boldsymbol{\Omega}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}}, \tag{F.9}
$$

$$
\bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v}) := -h^{d/2} \frac{\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1}\mathbf{p}^{(\mathbf{v})}(\mathbf{x})}{\sqrt{\bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}}, \tag{F.10}
$$

$$
\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v}) := -h^{d/2} \frac{\mathbf{Q}_{0,\mathbf{q}}^{-1}\mathbf{p}^{(\mathbf{v})}(\mathbf{x})}{\sqrt{\boldsymbol{\Omega}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}}, \tag{F.11}
$$

$$\bar{t}(\mathbf{x},\mathbf{q},\mathbf{v}) := h^{-d/2}\bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\mathbb{G}_n\Big[\psi(y_i,\eta(\mu_0(\mathbf{x}_i,\mathbf{q}));\mathbf{q})\eta^{(1)}(\mu_0(\mathbf{x}_i,\mathbf{q}))\mathbf{p}(\mathbf{x}_i)\Big], \tag{F.12}$$

$$T(\mathbf{x},\mathbf{q},\mathbf{v}) := \frac{\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^{\mathsf{T}}\widehat{\boldsymbol{\beta}}(\mathbf{q}) - \mu_0^{(\mathbf{v})}(\mathbf{x},\mathbf{q})}{\sqrt{\widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})/n}}. \tag{F.13}$$

**Theorem F.4** (Strong approximation: Yurinskii)**.**

*(a) Suppose all the conditions of Theorem E.1(a) hold with $\nu \geq 3$, $1/\big(nh^{3d}\big) = o(1)$ and the following $L_2$-norm Hölder-type condition holds: for any $\mathbf{q},\tilde{\mathbf{q}} \in \mathcal{Q}$*

$$\begin{aligned}\mathbb{E}\Big[\Big|&\psi(y_i,\eta(\mu_0(\mathbf{x}_i,\mathbf{q}));\mathbf{q})\eta^{(1)}(\mu_0(\mathbf{x}_i,\mathbf{q}))\\ &-\psi(y_i,\eta(\mu_0(\mathbf{x}_i,\tilde{\mathbf{q}}));\tilde{\mathbf{q}})\eta^{(1)}(\mu_0(\mathbf{x}_i,\tilde{\mathbf{q}}))\Big|^2\mathbf{x}_i\Big]^{1/2} \lesssim \|\mathbf{q}-\tilde{\mathbf{q}}\|^{1/2} \quad a.s.\end{aligned} \tag{F.14}$$

*Assume that the probability space (for each $n$) is rich enough: there exists a family of independent random variables $\{U_1, U_2, U_3\}$ distributed uniformly on $[0,1]$ that is independent of $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$. Then there exists a $K$-dimensional process $\boldsymbol{Z}_n(\mathbf{q})$ on $\mathcal{Q}$ which is conditionally on $\{\mathbf{x}_i\}_{i=1}^n$ a mean-zero Gaussian process with a.s. continuous trajectories and the same conditional covariance structure as that of $\boldsymbol{G}_n(\mathbf{q}) := h^{-d/2}\mathbb{G}_n\big[\psi(y_i,\eta(\mu_0(\mathbf{x}_i,\mathbf{q}));\mathbf{q})\eta^{(1)}(\mu_0(\mathbf{x}_i,\mathbf{q}))\mathbf{p}(\mathbf{x}_i)\big]$, meaning*

$$\mathbb{E}[\boldsymbol{G}_n(\mathbf{q})\boldsymbol{G}_n(\tilde{\mathbf{q}})^{\mathsf{T}} \mid \{\mathbf{x}_i\}_{i=1}^n] = \mathbb{E}[\boldsymbol{Z}_n(\mathbf{q})\boldsymbol{Z}_n(\tilde{\mathbf{q}})^{\mathsf{T}} \mid \{\mathbf{x}_i\}_{i=1}^n],$$

*such that if $r_{\mathrm{str}} \to 0$ is a positive sequence of numbers satisfying*

$$\left(\frac{1}{nh^{3d}}\right)^{\frac{1}{6+4d_{\mathcal{Q}}}}\sqrt{\log n} + \frac{1}{h^{d/2}n^{1/2-1/\nu}}\log n = o(r_{\mathrm{str}}),$$

*then*

$$\sup_{\mathbf{x}\in\mathcal{X}}\sup_{\mathbf{q}\in\mathcal{Q}}\Big|\bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\boldsymbol{G}_n(\mathbf{q}) - \bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\boldsymbol{Z}_n(\mathbf{q})\Big| = o_{\mathbb{P}}(r_{\mathrm{str}}). \tag{F.15}$$

*(b) If, in addition to the previous conditions, $\overline{\psi}(\mathbf{x}_i, y_i)$ is $\sigma^2$-sub-Gaussian conditionally on $\mathbf{x}_i$, and $r_{\mathrm{str}}^{\mathrm{sub}} \to 0$ is a positive sequence of numbers satisfying*

$$\left(\frac{1}{nh^{3d}}\right)^{\frac{1}{6+4d_{\mathcal{Q}}}}\sqrt{\log n} + \frac{\log^{3/2} n}{\sqrt{nh^d}} = o(r_{\mathrm{str}}^{\mathrm{sub}}),$$

*then*

$$\sup_{\mathbf{x}\in\mathcal{X}}\sup_{\mathbf{q}\in\mathcal{Q}}\Big|\bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\boldsymbol{G}_n(\mathbf{q}) - \bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\boldsymbol{Z}_n(\mathbf{q})\Big| = o_{\mathbb{P}}(r_{\mathrm{str}}^{\mathrm{sub}}). \tag{F.16}$$

*Proof.* We will only show Assertion (a) since Assertion (b) is shown similarly.

It is enough to show

$$\sup_{\mathbf{q}}\|\boldsymbol{G}_n(\mathbf{q}) - \boldsymbol{Z}_n(\mathbf{q})\|_\infty = o_{\mathbb{P}}(r_{\mathrm{str}})$$

because $\sup_{\mathbf{q},\mathbf{x}}\big\|\bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v})\big\|_1 \lesssim_{\mathbb{P}} 1$ (which we will prove in Lemma F.7 below). To do this, apply Theorem F.3 with $\boldsymbol{G}_n(\cdot)$ as in the statement,

$$\begin{aligned}&A_n(\mathbf{q}, y) = A_n(\mathbf{q},\mathbf{x},y) := \psi(y,\eta(\mu_0(\mathbf{x},\mathbf{q}));\mathbf{q})\eta^{(1)}(\mu_0(\mathbf{x},\mathbf{q})),\\ &\mathbf{b}(\cdot) := h^{-d/2}\mathbf{p}(\cdot),\\ &\zeta_K \lesssim K^{1/2} \lesssim h^{-d/2},\\ &\mu_n \lesssim 1.\end{aligned}$$

□

Using this theorem, we will obtain the following result as a corollary. This corresponds to Theorem 2 in the main paper; the notation there is simplified for better readability, and the statement is different but the argument is the same. Specifically, $T(\mathbf{x}, q)$ in the main paper corresponds to $T(\mathbf{x}, \mathbf{q}, \mathbf{v})$ here and the approximating process $Z(\mathbf{x}, q)$ corresponds to $\bar{\bar{\boldsymbol{\ell}}}(\mathbf{x}, \mathbf{q}, \mathbf{v})^{\mathsf{T}} \boldsymbol{Z}_n(\mathbf{q})$ here. The reason for differing notation is that we require more precision here: given the form of Theorem F.3, it is more natural to use $K$-dimensional processes such as $\boldsymbol{Z}_n(\mathbf{q})$, also making for simpler presentation of precise results in Section G later.

**Corollary F.5** (Strong approximation of the t-process)**.** *In the setting of Theorem F.4(a), we have*

$$\sup_{\mathbf{q},\mathbf{x}} \left| T(\mathbf{x}, \mathbf{q}, \mathbf{v}) - \bar{\bar{\boldsymbol{\ell}}}(\mathbf{x}, \mathbf{q}, \mathbf{v})^{\mathsf{T}} \boldsymbol{Z}_n(\mathbf{q}) \right| \lesssim_{\mathbb{P}} \sqrt{n} h^{d/2} r_{\mathrm{uc}} r_{\Omega} + \sqrt{n h^d} r_{\mathrm{Bah}} + o(r_{\mathrm{str}}) = o(r_{\mathrm{str}} + r_{\mathrm{ho}}).$$

The rest of this subsection will be devoted to filling in the small detail in the proof of Theorem F.4 that we deferred, and proving Corollary F.5.

**Lemma F.6** (Asymptotic variance)**.** *Suppose Assumptions Assumptions B.1 to B.5 hold, and* $\frac{\log(1/h)}{nh^d} = o(1)$*. Then*

$$h^{-d-2|\mathbf{v}|} \lesssim \inf_{\mathbf{q} \in \mathcal{Q}} \inf_{\mathbf{x} \in \mathcal{X}} \boldsymbol{\Omega}_{\mathbf{v}}(\mathbf{x}, \mathbf{q}) \le \sup_{\mathbf{q} \in \mathcal{Q}} \sup_{\mathbf{x} \in \mathcal{X}} \boldsymbol{\Omega}_{\mathbf{v}}(\mathbf{x}, \mathbf{q}) \lesssim h^{-d-2|\mathbf{v}|}, \tag{F.17}$$

$$\sup_{\mathbf{q} \in \mathcal{Q}} \sup_{\mathbf{x} \in \mathcal{X}} \left| \bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x}, \mathbf{q}) - \boldsymbol{\Omega}_{\mathbf{v}}(\mathbf{x}, \mathbf{q}) \right| \lesssim_{\mathbb{P}} h^{-d-2|\mathbf{v}|} \sqrt{\frac{\log(1/h)}{nh^d}}, \tag{F.18}$$

$$h^{-d-2|\mathbf{v}|} \lesssim_{\mathbb{P}} \inf_{\mathbf{q} \in \mathcal{Q}} \inf_{\mathbf{x} \in \mathcal{X}} \bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x}, \mathbf{q}) \le \sup_{\mathbf{q} \in \mathcal{Q}} \sup_{\mathbf{x} \in \mathcal{X}} \bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x}, \mathbf{q}) \lesssim_{\mathbb{P}} h^{-d-2|\mathbf{v}|}. \tag{F.19}$$

*Proof.* For the lower bound in (F.17), we have

$$\begin{aligned} \boldsymbol{\Omega}_{\mathbf{v}}(\mathbf{x}, \mathbf{q}) &\ge \lambda_{\min}(\boldsymbol{\Sigma}_{0,\mathbf{q}}) \cdot \left\| \mathbf{Q}_{0,\mathbf{q}}^{-1} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) \right\|^2 \ge \lambda_{\min}(\boldsymbol{\Sigma}_{0,\mathbf{q}}) \cdot \|\mathbf{Q}_{0,\mathbf{q}}\|^{-2} \left\| \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) \right\|^2 \\ &\gtrsim \lambda_{\min}(\boldsymbol{\Sigma}_{0,\mathbf{q}}) \cdot h^{-2d} \cdot h^{-2|\mathbf{v}|} \gtrsim h^d \cdot h^{-2d} \cdot h^{-2|\mathbf{v}|} = h^{-d-2|\mathbf{v}|} \end{aligned} \tag{F.20}$$

by Assumptions B.4, B.2 ($\sigma_{\mathbf{q}}^2(\mathbf{x})$ and $\eta^{(1)}(\mu_0(\mathbf{x}, \mathbf{q}))^2$ are uniformly over $\mathbf{x}$ bounded away from zero) and Lemma C.11.

For the upper bound in (F.17), we have

$$\begin{aligned} \boldsymbol{\Omega}_{\mathbf{v}}(\mathbf{x}, \mathbf{q}) &\le \lambda_{\max}(\boldsymbol{\Sigma}_{0,\mathbf{q}}) \cdot \left\| \mathbf{Q}_{0,\mathbf{q}}^{-1} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) \right\|^2 \le \lambda_{\max}(\boldsymbol{\Sigma}_{0,\mathbf{q}}) \cdot \left[ \lambda_{\max}(\mathbf{Q}_{0,\mathbf{q}}^{-1}) \right]^2 \left\| \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) \right\|^2 \\ &\overset{\text{(a)}}{\lesssim} \lambda_{\max}(\boldsymbol{\Sigma}_{0,\mathbf{q}}) \cdot h^{-2d} \cdot h^{-2|\mathbf{v}|} \overset{\text{(b)}}{\lesssim} h^d \cdot h^{-2d} \cdot h^{-2|\mathbf{v}|} = h^{-d-2|\mathbf{v}|}, \end{aligned}$$

where (a) is by Assumption B.4 and Lemma C.11, (b) is by Assumption B.2 ($\sigma_{\mathbf{q}}^2(\mathbf{x})$ and $\eta^{(1)}(\mu_0(\mathbf{x}, \mathbf{q}))^2$ are uniformly over $\mathbf{x}$ bounded) and Lemma C.11.

We will now prove (F.18). We start by noticing

$$\begin{aligned} &\sup_{\mathbf{q},\mathbf{x}} \left| \mathbf{p}^{(\mathbf{v})}(\mathbf{x})^{\mathsf{T}} \mathbf{Q}_{0,\mathbf{q}}^{-1} \big( \bar{\boldsymbol{\Sigma}}_{\mathbf{q}} - \boldsymbol{\Sigma}_{0,\mathbf{q}} \big) \mathbf{Q}_{0,\mathbf{q}}^{-1} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) \right| \\ &\quad \lesssim_{\mathbb{P}} \sup_{\mathbf{q},\mathbf{x}} \left\| \bar{\boldsymbol{\Sigma}}_{\mathbf{q}} - \boldsymbol{\Sigma}_{0,\mathbf{q}} \right\| \cdot \left\| \mathbf{Q}_{0,\mathbf{q}}^{-1} \right\|^2 \cdot \left\| \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) \right\|^2 \\ &\quad \overset{\text{(a)}}{\lesssim_{\mathbb{P}}} h^{-2d} \cdot h^{-2|\mathbf{v}|} \sup_{\mathbf{q}} \left\| \bar{\boldsymbol{\Sigma}}_{\mathbf{q}} - \boldsymbol{\Sigma}_{0,\mathbf{q}} \right\| \overset{\text{(b)}}{\lesssim_{\mathbb{P}}} h^{-2d-2|\mathbf{v}|} h^d \sqrt{\frac{\log(1/h)}{nh^d}} = h^{-d-2|\mathbf{v}|} \sqrt{\frac{\log(1/h)}{nh^d}}, \end{aligned}$$

where in (a) we used $\sup_{\mathbf{q}}\left\|\mathbf{Q}_{0,\mathbf{q}}^{-1}\right\| \lesssim_{\mathbb{P}} h^{-d}$ by Lemma C.11 and $\sup_{\mathbf{x}}\|\mathbf{p}^{(\mathbf{v})}(\mathbf{x})\| \lesssim h^{-|\mathbf{v}|}$ by Assumption B.4, in (b) we used $\sup_{\mathbf{q}}\|\bar{\mathbf{\Sigma}}_{\mathbf{q}} - \mathbf{\Sigma}_{0,\mathbf{q}}\| \lesssim_{\mathbb{P}} h^d\sqrt{\frac{\log(1/h)}{nh^d}}$ which is proven by the same argument as $\sup_{\mathbf{q}}\|\bar{\boldsymbol{Q}}_{\mathbf{q}} - \mathbf{Q}_{0,\mathbf{q}}\| \lesssim_{\mathbb{P}} h^d\sqrt{\frac{\log(1/h)}{nh^d}}$ in Lemma C.11. Similarly,

$$
\begin{aligned}
&\sup_{\mathbf{q},\mathbf{x}}\left|\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^{\mathsf{T}}(\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} - \mathbf{Q}_{0,\mathbf{q}}^{-1})\bar{\mathbf{\Sigma}}_{\mathbf{q}}\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1}\mathbf{p}^{(\mathbf{v})}(\mathbf{x})\right| \\
&\quad\le \sup_{\mathbf{q},\mathbf{x}}\|\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} - \mathbf{Q}_{0,\mathbf{q}}^{-1}\| \cdot \|\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1}\| \cdot \|\bar{\mathbf{\Sigma}}_{\mathbf{q}}\| \cdot \|\mathbf{p}^{(\mathbf{v})}(\mathbf{x})\|^2 \\
&\quad\lesssim_{\mathbb{P}} h^{-d}\sqrt{\frac{\log(1/h)}{nh^d}} \cdot h^{-d} \cdot h^d \cdot h^{-2|\mathbf{v}|} = h^{-d-2|\mathbf{v}|}\sqrt{\frac{\log(1/h)}{nh^d}}.
\end{aligned}
$$

Finally,

$$
\sup_{\mathbf{q},\mathbf{x}}\left|\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^{\mathsf{T}}\mathbf{Q}_{0,\mathbf{q}}^{-1}\bar{\mathbf{\Sigma}}_{\mathbf{q}}(\mathbf{Q}_{0,\mathbf{q}}^{-1} - \bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1})\mathbf{p}^{(\mathbf{v})}(\mathbf{x})\right| \lesssim_{\mathbb{P}} h^{-d-2|\mathbf{v}|}\sqrt{\frac{\log(1/h)}{nh^d}}.
$$

Combining the bounds above gives $\sup_{\mathbf{q},\mathbf{x}}\left|\bar{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q}) - \mathbf{\Omega}_{\mathbf{v}}(\mathbf{x},\mathbf{q})\right| \lesssim_{\mathbb{P}} h^{-d-2|\mathbf{v}|}\sqrt{\frac{\log(1/h)}{nh^d}}$.

(F.19) is an immediate consequence of (F.17) and (F.18), since $\frac{\log(1/h)}{nh^d} = o(1)$. $\square$

**Lemma F.7** (Closeness of linear terms)**.** *Suppose all the conditions of Theorem E.1(a) hold. Then*

$$
\sup_{\mathbf{q},\mathbf{x}}\left|\widehat{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})^{-1/2} - \bar{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})^{-1/2}\right| \lesssim_{\mathbb{P}} h^{d/2+|\mathbf{v}|}r_{\Omega}, \tag{F.21}
$$

$$
\sup_{\mathbf{q},\mathbf{x}}\left|\bar{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})^{-1/2} - \mathbf{\Omega}_{\mathbf{v}}(\mathbf{x},\mathbf{q})^{-1/2}\right| \lesssim_{\mathbb{P}} h^{d/2+|\mathbf{v}|}\sqrt{\frac{\log(1/h)}{nh^d}}, \tag{F.22}
$$

$$
\sup_{\mathbf{q},\mathbf{x}}\|\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})\|_1 \lesssim 1, \tag{F.23}
$$

$$
\sup_{\mathbf{q},\mathbf{x}}\left\|\bar{\boldsymbol{\ell}}(\mathbf{x},\mathbf{q},\mathbf{v}) - \boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})\right\|_1 \lesssim_{\mathbb{P}} \sqrt{\frac{\log(1/h)}{nh^d}}, \tag{F.24}
$$

$$
\sup_{\mathbf{q},\mathbf{x}}\left\|\bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v}) - \boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})\right\|_1 \lesssim_{\mathbb{P}} \sqrt{\frac{\log(1/h)}{nh^d}}, \tag{F.25}
$$

$$
\sup_{x,\mathbf{x}}\left\|\bar{\boldsymbol{\ell}}(\mathbf{x},\mathbf{q},\mathbf{v})\right\|_1 \lesssim 1 \quad \textit{w. p. a. 1}, \tag{F.26}
$$

$$
\sup_{x,\mathbf{x}}\left\|\bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v})\right\|_1 \lesssim 1 \quad \textit{w. p. a. 1}. \tag{F.27}
$$

*If, in addition, Assumption B.6 holds, then*

$$
\sup_{\mathbf{q},\mathbf{x}}\left\|\hat{\boldsymbol{\ell}}(\mathbf{x},\mathbf{q},\mathbf{v}) - \bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v})\right\|_1 \lesssim_{\mathbb{P}} r_{\mathrm{Q}} + r_{\Omega}, \tag{F.28}
$$

$$
\sup_{x,\mathbf{x}}\left\|\hat{\boldsymbol{\ell}}(\mathbf{x},\mathbf{q},\mathbf{v})\right\|_1 \lesssim 1 \quad \textit{w. p. a. 1}. \tag{F.29}
$$

*Proof of* (F.21). This follows from the following chain of inequalities:

$$
\begin{aligned}
&\sup_{\mathbf{q},\mathbf{x}}\left|\widehat{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})^{-1/2} - \bar{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})^{-1/2}\right| \\
&\quad\le \sup_{\mathbf{q},\mathbf{x}}\frac{1}{\sqrt{\widehat{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})\bar{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}\left(\sqrt{\widehat{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})} + \sqrt{\bar{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}\right)}\sup_{\mathbf{q},\mathbf{x}}\left|\widehat{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q}) - \bar{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})\right|
\end{aligned}
$$

$$\overset{(a)}{\lesssim_{\mathbb{P}}} \left(h^{d+2|\mathbf{v}|}\right)^{3/2} \sup_{\mathbf{q},\mathbf{x}} \left|\widehat{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q}) - \bar{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})\right| \overset{(b)}{\lesssim_{\mathbb{P}}} h^{d/2+|\mathbf{v}|} r_{\Omega},$$

where in (a) we used that $\widehat{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})$ is close to $\bar{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})$ by Assumption B.7 and the lower bound on $\bar{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})$ from Lemma F.6, and in (b) we used Assumption B.7. □

*Proof of* (F.22). Recall that

$$\begin{aligned}
&\sup_{\mathbf{q},\mathbf{x}} \left|\bar{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q}) - \mathbf{\Omega}_{\mathbf{v}}(\mathbf{x},\mathbf{q})\right| \lesssim_{\mathbb{P}} h^{-d-2|\mathbf{v}|}\sqrt{\frac{\log(1/h)}{nh^d}}, \quad \text{and} \\
&\inf_{\mathbf{q},\mathbf{x}} \bar{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q}) \gtrsim_{\mathbb{P}} h^{-d-2|\mathbf{v}|}
\end{aligned}$$

by Lemma F.6. Therefore,

$$\begin{aligned}
&\sup_{\mathbf{q},\mathbf{x}} \left|\bar{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})^{-1/2} - \mathbf{\Omega}_{\mathbf{v}}(\mathbf{x},\mathbf{q})^{-1/2}\right| \\
&\quad \le \sup_{\mathbf{q},\mathbf{x}} \frac{1}{\sqrt{\bar{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})\mathbf{\Omega}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}\left(\sqrt{\bar{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})} + \sqrt{\mathbf{\Omega}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}\right)} \sup_{\mathbf{q},\mathbf{x}} \left|\bar{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q}) - \mathbf{\Omega}_{\mathbf{v}}(\mathbf{x},\mathbf{q})\right| \\
&\quad \lesssim_{\mathbb{P}} \left(h^{d+2|\mathbf{v}|}\right)^{3/2} \sup_{\mathbf{q},\mathbf{x}} \left|\bar{\mathbf{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q}) - \mathbf{\Omega}_{\mathbf{v}}(\mathbf{x},\mathbf{q})\right| \lesssim_{\mathbb{P}} h^{3d/2+3|\mathbf{v}|} h^{-d-2|\mathbf{v}|}\sqrt{\frac{\log(1/h)}{nh^d}} \\
&\quad = h^{d/2+|\mathbf{v}|}\sqrt{\frac{\log(1/h)}{nh^d}}.
\end{aligned}$$

□

*Proof of* (F.23). The result follows from

$$h^{d/2}\left\|\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1}\mathbf{p}^{(\mathbf{v})}(\mathbf{x})\right\|_1 \le h^{d/2}\left\|\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1}\right\|_1 \cdot \left\|\mathbf{p}^{(\mathbf{v})}(\mathbf{x})\right\|_1 \lesssim_{\mathbb{P}} h^{-d/2-|\mathbf{v}|}$$

and Lemma F.6. □

*Proof of* (F.24). The result follows from

$$\begin{aligned}
\sup_{\mathbf{q},\mathbf{x}} \left\|\bar{\boldsymbol{\ell}}(\mathbf{x},\mathbf{q},\mathbf{v}) - \boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})\right\|_1 &= \sup_{\mathbf{q},\mathbf{x}} \frac{h^{d/2}}{\sqrt{\mathbf{\Omega}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}} \left\|\left(\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} - \mathbf{Q}_{0,\mathbf{q}}^{-1}\right)\mathbf{p}^{(\mathbf{v})}(\mathbf{x})\right\|_1 \\
&\overset{(a)}{\lesssim} h^{d+|\mathbf{v}|} \sup_{\mathbf{q},\mathbf{x}} \left\|\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} - \mathbf{Q}_{0,\mathbf{q}}^{-1}\right\|_1 \left\|\mathbf{p}^{(\mathbf{v})}(\mathbf{x})\right\|_1 \\
&\overset{(b)}{\lesssim_{\mathbb{P}}} h^{d+|\mathbf{v}|} \cdot h^{-d}\sqrt{\frac{\log(1/h)}{nh^d}} \cdot h^{-|\mathbf{v}|} = \sqrt{\frac{\log(1/h)}{nh^d}},
\end{aligned}$$

where in (a) we used $\inf_{\mathbf{q},\mathbf{x}} \mathbf{\Omega}_{\mathbf{v}}(\mathbf{x},\mathbf{q}) \gtrsim h^{-d-2|\mathbf{v}|}$ by Lemma F.6, in (b) we used $\left\|\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} - \mathbf{Q}_{0,\mathbf{q}}^{-1}\right\|_1 \lesssim_{\mathbb{P}} h^{-d}\sqrt{\log(1/h)/(nh^d)}$ by Lemma C.11 and $\left\|\mathbf{p}^{(\mathbf{v})}(\mathbf{x})\right\|_1 \lesssim h^{-|\mathbf{v}|}$ by Assumption B.4. □

*Proof of* (F.25). We have that

$$\sup_{\mathbf{q}} \left\|\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1}\right\|_1 \lesssim_{\mathbb{P}} h^{-d}$$

by Lemma C.11, and

$$\sup_{\mathbf{x}} \left\| \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) \right\|_1 \lesssim h^{-|\mathbf{v}|}$$

by Assumption B.4. Combining this with (F.22) gives

$$\left\| \frac{h^{d/2} \bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} \mathbf{p}^{(\mathbf{v})}(\mathbf{x})}{\sqrt{\bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x}, \mathbf{q})}} - \frac{h^{d/2} \bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} \mathbf{p}^{(\mathbf{v})}(\mathbf{x})}{\sqrt{\boldsymbol{\Omega}_{\mathbf{v}}(\mathbf{x}, \mathbf{q})}} \right\|_1 \lesssim_{\mathbb{P}} \sqrt{\frac{\log(1/h)}{nh^d}}.$$

It is left to recall (F.24). □

*Proof of* (F.26). This follows from (F.24), (F.23) and $\sqrt{\log(1/h)/(nh^d)} = o(1)$. □

*Proof of* (F.27). This follows from (F.25), (F.23) and $\sqrt{\log(1/h)/(nh^d)} = o(1)$. □

*Proof of* (F.28). By the triangle inequality,

$$\sup_{\mathbf{q}, \mathbf{x}} \left\| \frac{\widehat{\boldsymbol{Q}}_{\mathbf{q}}^{-1} \mathbf{p}^{(\mathbf{v})}(\mathbf{x})}{\sqrt{\widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x}, \mathbf{q})}} - \frac{\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} \mathbf{p}^{(\mathbf{v})}(\mathbf{x})}{\sqrt{\bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x}, \mathbf{q})}} \right\|_1 \tag{F.30}$$

$$\lesssim \sup_{\mathbf{q}, \mathbf{x}} \left\| \frac{\left( \widehat{\boldsymbol{Q}}_{\mathbf{q}}^{-1} - \bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} \right) \mathbf{p}^{(\mathbf{v})}(\mathbf{x})}{\sqrt{\widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x}, \mathbf{q})}} \right\|_1 + \sup_{\mathbf{q}, \mathbf{x}} \left\| \bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) \left( \frac{1}{\sqrt{\widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x}, \mathbf{q})}} - \frac{1}{\sqrt{\bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x}, \mathbf{q})}} \right) \right\|_1. \tag{F.31}$$

To bound the first term in (F.31), recall that $\widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x}, \mathbf{q}) \gtrsim_{\mathbb{P}} h^{-d-2|\mathbf{v}|}$ by Lemma F.6 and Assumption B.7. Then

$$\begin{aligned}
\sup_{\mathbf{q}, \mathbf{x}} \left\| \frac{\left( \widehat{\boldsymbol{Q}}_{\mathbf{q}}^{-1} - \bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} \right) \mathbf{p}^{(\mathbf{v})}(\mathbf{x})}{\sqrt{\widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x}, \mathbf{q})}} \right\|_1 &\lesssim_{\mathbb{P}} h^{d/2+|\mathbf{v}|} \sup_{\mathbf{q}, \mathbf{x}} \left\| \left( \widehat{\boldsymbol{Q}}_{\mathbf{q}}^{-1} - \bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} \right) \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) \right\|_1 \\
&\lesssim h^{d/2+|\mathbf{v}|} \sup_{\mathbf{q}, \mathbf{x}} \left\| \widehat{\boldsymbol{Q}}_{\mathbf{q}}^{-1} - \bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} \right\|_1 \cdot \left\| \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) \right\|_1 \\
&\lesssim h^{d/2} \left\| \widehat{\boldsymbol{Q}}_{\mathbf{q}}^{-1} - \bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} \right\|_\infty \lesssim_{\mathbb{P}} h^{d/2} \cdot h^{-d} r_{\mathrm{Q}} = h^{-d/2} r_{\mathrm{Q}},
\end{aligned}$$

where in the last inequality we used $\left\| \widehat{\boldsymbol{Q}}_{\mathbf{q}} - \bar{\boldsymbol{Q}}_{\mathbf{q}} \right\|_\infty \lesssim_{\mathbb{P}} h^d r_{\mathrm{Q}}$ by assumption, and

$$\left\| \widehat{\boldsymbol{Q}}_{\mathbf{q}}^{-1} - \bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} \right\|_\infty \le \left\| \widehat{\boldsymbol{Q}}_{\mathbf{q}}^{-1} \right\|_\infty \cdot \left\| \widehat{\boldsymbol{Q}}_{\mathbf{q}} - \bar{\boldsymbol{Q}}_{\mathbf{q}} \right\|_\infty \cdot \| \bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} \|_\infty \lesssim_{\mathbb{P}} h^{-d} \left( h^d r_{\mathrm{Q}} \right) h^{-d} = h^{-d} r_{\mathrm{Q}}$$

also by assumption and Lemma C.11.

It is left to bound the second term in (F.31):

$$\begin{aligned}
&\sup_{\mathbf{q}, \mathbf{x}} \left\| \bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) \left( \widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x}, \mathbf{q})^{-1/2} - \bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x}, \mathbf{q})^{-1/2} \right) \right\|_1 \\
&\quad \le \sup_{\mathbf{q}} \| \bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1} \|_1 \sup_{\mathbf{x}} \left\| \mathbf{p}^{(\mathbf{v})}(\mathbf{x}) \right\|_1 \sup_{\mathbf{q}, \mathbf{x}} \left| \widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x}, \mathbf{q})^{-1/2} - \bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x}, \mathbf{q})^{-1/2} \right| \\
&\quad \overset{\text{(a)}}{\lesssim_{\mathbb{P}}} h^{-d} \cdot h^{-|\mathbf{v}|} \cdot \sup_{\mathbf{q}, \mathbf{x}} \left| \widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x}, \mathbf{q})^{-1/2} - \bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x}, \mathbf{q})^{-1/2} \right| \\
&\quad \overset{\text{(b)}}{\lesssim_{\mathbb{P}}} h^{-d-|\mathbf{v}|} h^{d/2+|\mathbf{v}|} r_\Omega = h^{-d/2} r_\Omega,
\end{aligned}$$

where (a) is by Lemma C.11 and Assumption B.4, (b) is by (F.21). □

*Proof of* (F.29). This follows from (F.28), (F.27) and $r_{\mathrm{uc}} + r_\Omega = o(1)$. $\square$

**Lemma F.8** (Hats off)**.** *Suppose all the conditions of Theorem E.1(a) hold. Define (and fix for all further arguments) $r_{\mathrm{ho}}$ as an arbitrary positive sequence satisfying*

$$\sqrt{n}h^{d/2} r_{\mathrm{uc}} r_\Omega + \sqrt{nh^d} r_{\mathrm{Bah}} = o(r_{\mathrm{ho}}) \tag{F.32}$$

*for $r_{\mathrm{Bah}}$ defined in (E.3).*

*Then*

$$\sup_{\mathbf{x}\in\mathcal{X}} \sup_{\mathbf{q}\in\mathcal{Q}} |T(\mathbf{x},\mathbf{q},\mathbf{v}) - \bar{t}(\mathbf{x},\mathbf{q},\mathbf{v})| \lesssim_{\mathbb{P}} \sqrt{n}h^{d/2} r_{\mathrm{uc}} r_\Omega + \sqrt{nh^d} r_{\mathrm{Bah}} = o(r_{\mathrm{ho}}). \tag{F.33}$$

*Proof.* First, note that by Corollary E.2, Lemma F.6 and Assumption B.7

$$\begin{aligned}
&\sup_{\mathbf{q},\mathbf{x}} \left| \frac{\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\intercal \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \mu_0^{(\mathbf{v})}(\mathbf{x},\mathbf{q})}{\sqrt{\widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})/n}} - \frac{\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\intercal \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \mu_0^{(\mathbf{v})}(\mathbf{x},\mathbf{q})}{\sqrt{\bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})/n}} \right| \\
&\le \sqrt{n} \sup_{\mathbf{q},\mathbf{x}} \left| \mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\intercal \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \mu_0^{(\mathbf{v})}(\mathbf{x},\mathbf{q}) \right| \sup_{\mathbf{q},\mathbf{x}} \frac{1}{\sqrt{\widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q}) \bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})} \left( \sqrt{\widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})} + \sqrt{\bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})} \right)} \\
&\quad \cdot \sup_{\mathbf{q},\mathbf{x}} \left| \widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q}) - \bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q}) \right| \\
&\lesssim_{\mathbb{P}} \sqrt{n} \cdot h^{-|\mathbf{v}|} r_{\mathrm{uc}} \cdot \left( h^{d+2|\mathbf{v}|} \right)^{3/2} \cdot h^{-2|\mathbf{v}|-d} r_\Omega \\
&= \sqrt{n} h^{d/2} r_{\mathrm{uc}} r_\Omega.
\end{aligned}$$

Now, by Theorem E.1 and $h^{-2|\mathbf{v}|-d} \lesssim_{\mathbb{P}} \inf_{\mathbf{q},\mathbf{x}} |\bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})|$ (by Assumption B.7 and Lemma F.6)

$$\sup_{\mathbf{q},\mathbf{x}} \left| \frac{\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\intercal \widehat{\boldsymbol{\beta}}(\mathbf{q}) - \mu_0^{(\mathbf{v})}(\mathbf{x},\mathbf{q})}{\sqrt{\bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})/n}} - \bar{t}(\mathbf{x},\mathbf{q},\mathbf{v}) \right| \lesssim_{\mathbb{P}} \sqrt{n} \left( h^{d+2|\mathbf{v}|} \right)^{1/2} h^{-|\mathbf{v}|} r_{\mathrm{Bah}} = \sqrt{nh^d} r_{\mathrm{Bah}},$$

where $r_{\mathrm{Bah}}$ is defined in (E.3).

It is left to apply the triangle inequality. $\square$

*Proof of Corollary F.5.* Combine Theorem F.4 and Lemma F.8. $\square$

# G Uniform inference

**Theorem G.1** (Plug-in approximation)**.** *Suppose that all of the following is true:*

*(i) All the conditions of Corollary E.2(a), Theorem F.4(a) hold.*

*(ii) Assumption B.6 holds with a rate restriction $(r_{\mathrm{Q}} + r_\Omega)\sqrt{\log(1/h)} = o(r_{\mathrm{str}})$.*

*(iii) On the same probability space, there is a $K$-dimensional Gaussian conditionally on $\{\mathbf{x}_i\}_{i=1}^n$ process $\widehat{\boldsymbol{Z}}_n^*(\mathbf{q})$ on $\mathcal{Q}$ with a. s. continuous sample paths, whose distribution is known (depends only on the data and known quantities), such that*

$$\sup_{\mathbf{q}\in\mathcal{Q}} \left\| \widehat{\boldsymbol{Z}}_n^*(\mathbf{q}) - \boldsymbol{Z}_n^*(\mathbf{q}) \right\|_\infty = o_{\mathbb{P}}(r_{\mathrm{str}}), \tag{G.1}$$

*where $\boldsymbol{Z}_n^*(\mathbf{q})$ is a copy of $\boldsymbol{Z}_n(\cdot)$ conditionally on $\{\mathbf{x}_i\}_{i=1}^n$ depending on the data only via $\{\mathbf{x}_i\}_{i=1}^n$.*

*Then*

$$h^{d/2}\sup_{\mathbf{q}\in\mathcal{Q}}\sup_{\mathbf{x}\in\mathcal{X}}\left|\frac{\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^{\mathsf{T}}\widehat{\boldsymbol{Q}}_{\mathbf{q}}^{-1}}{\sqrt{\widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}}\widehat{\boldsymbol{Z}}_n^*(\mathbf{q})-\frac{\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^{\mathsf{T}}\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1}}{\sqrt{\bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}}\boldsymbol{Z}_n^*(\mathbf{q})\right|=o_{\mathbb{P}}(r_{\mathrm{str}}),$$

*or, equivalently (by Lemma C.3), for any* $\varepsilon>0$

$$\mathbb{P}\left\{h^{d/2}\sup_{\mathbf{q}\in\mathcal{Q}}\sup_{\mathbf{x}\in\mathcal{X}}\left|\frac{\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^{\mathsf{T}}\widehat{\boldsymbol{Q}}_{\mathbf{q}}^{-1}}{\sqrt{\widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}}\widehat{\boldsymbol{Z}}_n^*(\mathbf{q})-\frac{\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^{\mathsf{T}}\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1}}{\sqrt{\bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}}\boldsymbol{Z}_n^*(\mathbf{q})\right|>\varepsilon r_{\mathrm{str}}\;\middle|\;\boldsymbol{D}_n\right\}=o_{\mathbb{P}}(1).$$

*If* $\overline{\psi}(\mathbf{x}_i,y_i)$ *is* $\sigma^2$*-sub-Gaussian conditionally on* $\mathbf{x}_i$*, then* $r_{\mathrm{str}}$ *in the conditions and conclusion of this theorem can be replaced with* $r_{\mathrm{str}}^{\mathrm{sub}}$*.*

For the proof, we only handle the case where $\nu$ is finite since the sub-Gaussian case is proven similarly. We build on the proof of Theorem 6.3 in [3].

**Lemma G.2** (Bounding the supremum of the Gaussian process)**.** *In the setting of Corollary E.2(a), we have* $\sup_{\mathbf{q}\in\mathcal{Q}}\|\boldsymbol{Z}_n(\mathbf{q})\|_\infty\lesssim_{\mathbb{P}}\sqrt{\log K}$*.*

*Proof.* Consider the stochastic process $X_n$ defined for $t=(l,\mathbf{q})\in T$ with

$$T:=\{(l,\mathbf{q}):l\in\{1,2,\dots,K\},\mathbf{q}\in\mathcal{Q}\}$$

as $X_{n,t}:=Z_{n,l}(\mathbf{q})$. It is a separable mean-zero Gaussian process with the index set $T$ considered a metric space: $\operatorname{dist}((l,\mathbf{q}),(l',\mathbf{q}'))=\|\mathbf{q}-\mathbf{q}'\|+\mathbb{1}\{l\neq l'\}$. Note that

$$\begin{aligned}\sigma(X_n)^2&:=\sup_{t\in T}\mathbb{E}\big[X_{n,t}^2\;\big|\;\{\mathbf{x}_i\}_{i=1}^n\big]=\sup_{\mathbf{q}\in\mathcal{Q}}\max_l\mathbb{E}\big[Z_{n,l}^2(\mathbf{q})\;\big|\;\{\mathbf{x}_i\}_{i=1}^n\big]\\&=h^{-d}\sup_{\mathbf{q}\in\mathcal{Q}}\max_l\frac{1}{n}\sum_{i=1}^n\mathbb{E}[\psi(y_i;\eta(\mu_0(\mathbf{x}_i,\mathbf{q})),;\mathbf{q})2\mid\mathbf{x}_i]\eta^{(1)}(\mu_0(\mathbf{x}_i,\mathbf{q}))^2p_l(\mathbf{x}_i)^2\\&\lesssim h^{-d}\mathbb{E}_n\big[p_l(\mathbf{x}_i)^2\big]\leq h^{-d}\sup_{\boldsymbol{\alpha}\in\mathcal{S}^{K-1}}\mathbb{E}_n[\boldsymbol{\alpha}^{\mathsf{T}}\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^{\mathsf{T}}\boldsymbol{\alpha}]\lesssim1\quad\text{w. p. a. 1.}\end{aligned}$$

Next, we will bound

$$\rho(t,t')^2:=\mathbb{E}\Big[\big(X_{n,t}-X_{n,t'}\big)^2\;\Big|\;\{\mathbf{x}_i\}_{i=1}^n\Big]=\mathbb{E}\Big[\big(Z_{n,l}(\mathbf{q})-Z_{n,l'}(\mathbf{q}')\big)^2\;\Big|\;\{\mathbf{x}_i\}_{i=1}^n\Big].$$

If $l\neq l'$,

$$\begin{aligned}&\mathbb{E}\Big[\big(Z_{n,l}(\mathbf{q})-Z_{n,l'}(\mathbf{q}')\big)^2\;\Big|\;\{\mathbf{x}_i\}_{i=1}^n\Big]\\&\quad=\frac{1}{n}\sum_{i=1}^n h^{-d}\mathbb{E}\Big[\Big(\psi(y_i,\eta(\mu_0(\mathbf{x}_i,\mathbf{q}));\mathbf{q})\eta^{(1)}(\mu_0(\mathbf{x}_i,\mathbf{q}))p_l(\mathbf{x}_i)\\&\qquad-\psi\big(y_i,\eta\big(\mu_0(\mathbf{x}_i,\mathbf{q}')\big);\mathbf{q}'\big)\eta^{(1)}\big(\mu_0(\mathbf{x}_i,\mathbf{q}')\big)p_{l'}(\mathbf{x}_i)\Big)^2\{\mathbf{x}_i\}_{i=1}^n\Big]\\&\quad=h^{-d}\mathbb{E}_n\Big[\mathbb{E}\Big[\big(A_n(\mathbf{q},y_i,\mathbf{x}_i)-A_n(\mathbf{q}',y_i,\mathbf{x}_i)\big)^2\;\Big|\;\{\mathbf{x}_i\}_{i=1}^n\Big]p_l(\mathbf{x}_i)^2\Big]\\&\qquad+h^{-d}\mathbb{E}_n\big[\mathbb{E}\big[A_n(\mathbf{q}',y_i,\mathbf{x}_i)^2\;\big|\;\{\mathbf{x}_i\}_{i=1}^n\big]\big(p_l(\mathbf{x}_i)^2+p_{l'}(\mathbf{x}_i)^2\big)\big]\\&\quad\lesssim h^{-d}\|\mathbf{q}-\mathbf{q}'\|\mathbb{E}_n\big[p_l(\mathbf{x}_i)^2\big]+h^{-d}\mathbb{E}_n\big[p_l(\mathbf{x}_i)^2+p_{l'}(\mathbf{x}_i)^2\big]\end{aligned}$$

$$\lesssim \|\mathbf{q}-\mathbf{q}'\| + 1 \quad \text{w. p. a. } 1,$$

where we denoted $A_n(\mathbf{q}, y_i, \mathbf{x}_i) := \psi(y_i, \eta(\mu_0(\mathbf{x}_i, \mathbf{q})); \mathbf{q})\eta^{(1)}(\mu_0(\mathbf{x}_i, \mathbf{q}))$ to simplify notation. Similarly, if $l = l'$,

$$\mathbb{E}\Big[\big(Z_{n,l}(\mathbf{q}) - Z_{n,l'}(\mathbf{q}')\big)^2 \,\Big|\, \{\mathbf{x}_i\}_{i=1}^n\Big] \lesssim \|\mathbf{q}-\mathbf{q}'\| \quad \text{w. p. a. } 1.$$

We conclude

$$\rho(t,t')^2 \lesssim \operatorname{dist}\big((l,\mathbf{q}),(l',\mathbf{q}')\big).$$

This means that an $\varepsilon^2$-covering of $T$ with respect to $\operatorname{dist}(\cdot)$ induces an $\varepsilon$-covering of $T$ with respect to $\rho$, and hence

$$N(T,\rho,\varepsilon) \le N\big(T, \operatorname{dist}(\cdot), \varepsilon^2\big). \tag{G.2}$$

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.2) we get

$$\log N(T,\rho,\varepsilon) \lesssim \log(K/\varepsilon).$$

Now we apply Lemma C.7:

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in T}|X_{n,t}| \,\middle|\, \{\mathbf{x}_i\}_{i=1}^n\right] &\lesssim \sigma(X_n) + \int_0^{2\sigma(X_n)} \sqrt{\log N(T,\rho,\varepsilon)}\,\mathrm{d}\varepsilon \\ &\lesssim \sigma(X_n) + \sigma(X_n)\sqrt{\log(K/\sigma(X_n))} \lesssim_1 \sqrt{\log K} \quad \text{w. p. a. } 1,\end{aligned}$$

where in $\lesssim_1$ we used our bound $\sigma(X_n) \lesssim 1$ above. Rewriting, we obtain

$$\mathbb{E}\left[\sup_{\mathbf{q}\in\mathcal{Q}}\|\boldsymbol{Z}_n(\mathbf{q})\|_\infty \,\middle|\, \{\mathbf{x}_i\}_{i=1}^n\right] \lesssim \sqrt{\log K} \quad \text{w. p. a. } 1. \tag{G.3}$$

By Markov's inequality this gives $\sup_{\mathbf{q}\in\mathcal{Q}}\|\boldsymbol{Z}_n(\mathbf{q})\|_\infty \lesssim_{\mathbb{P}} \sqrt{\log K}$. □

*Concluding the proof of Theorem G.1.* By the triangle inequality, it is enough to show

$$h^{d/2}\sup_{\mathbf{q},\mathbf{x}}\left|\frac{\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\mathsf{T}\widehat{\boldsymbol{Q}}_{\mathbf{q}}^{-1}}{\sqrt{\widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}}\widehat{\boldsymbol{Z}}_n^*(\mathbf{q}) - \frac{\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\mathsf{T}\widehat{\boldsymbol{Q}}_{\mathbf{q}}^{-1}}{\sqrt{\widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}}\boldsymbol{Z}_n^*(\mathbf{q})\right| = o_{\mathbb{P}}(r_{\mathrm{str}}), \tag{G.4}$$

$$h^{d/2}\sup_{\mathbf{q},\mathbf{x}}\left|\frac{\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\mathsf{T}\widehat{\boldsymbol{Q}}_{\mathbf{q}}^{-1}}{\sqrt{\widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}}\boldsymbol{Z}_n^*(\mathbf{q}) - \frac{\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\mathsf{T}\bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1}}{\sqrt{\bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}}\boldsymbol{Z}_n^*(\mathbf{q})\right| = o_{\mathbb{P}}(r_{\mathrm{str}}), \tag{G.5}$$

To prove (G.4), combine $\sup_{\mathbf{q},\mathbf{x}}\left\|\hat{\boldsymbol{\ell}}(\mathbf{x},\mathbf{q},\mathbf{v})\right\|_1 \lesssim_{\mathbb{P}} 1$ by Lemma F.7 with (G.1).

To prove (G.5), combine Lemma G.2 with $\sup_{\mathbf{q},\mathbf{x}}\left\|\hat{\boldsymbol{\ell}}(\mathbf{x},\mathbf{q},\mathbf{v}) - \bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v})\right\|_1 \lesssim_{\mathbb{P}} r_{\mathrm{Q}} + r_\Omega$ by Lemma F.7, and use Condition (ii) of this theorem. □

**Theorem G.3** (Confidence bands)**.** *Suppose that the following is true:*

*(i) All conditions of Theorem F.4(a) hold.*

*(ii) Assumption B.6 holds with a rate restriction* $(r_{\mathrm{Q}} + r_\Omega)\sqrt{\log(1/h)} = o(r_{\mathrm{str}})$.

*(iii)* $(r_{\mathrm{str}} + r_{\mathrm{ho}})\sqrt{\log(1/h)} = o(1)$.

*Then* $\mathbb{P}\big\{\sup_{\mathbf{q},\mathbf{x}}|T(\mathbf{x},\mathbf{q},\mathbf{v})| < k^*(1-\alpha)\big\} = 1-\alpha+o(1)$, *where* $k^*(\eta)$ *is the conditional* $\eta$*-quantile of* $\sup_{\mathbf{q},\mathbf{x}}\big|\hat{\boldsymbol{\ell}}(\mathbf{x},\mathbf{q},\mathbf{v})^\mathsf{T}\widehat{\boldsymbol{Z}}_n^*(\mathbf{q})\big|$ *given the data.*

To prove this result we first introduce some auxiliary lemmas. To simplify the exposition, let

$$V := \sup_{\mathbf{q},\mathbf{x}} |T(\mathbf{x},\mathbf{q},\mathbf{v})|, \tag{G.6}$$

$$V^* := \sup_{\mathbf{q},\mathbf{x}} \left| h^{d/2} \frac{\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\mathsf{T} \widehat{\boldsymbol{Q}}_{\mathbf{q}}^{-1}}{\sqrt{\widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}} \widehat{\boldsymbol{Z}}_n^*(\mathbf{q}) \right|, \tag{G.7}$$

$$\tilde{V} := \sup_{\mathbf{q},\mathbf{x}} \left| h^{d/2} \frac{\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\mathsf{T} \bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1}}{\sqrt{\bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}} \boldsymbol{Z}_n(\mathbf{q}) \right|, \tag{G.8}$$

$$\tilde{V}^* := \sup_{\mathbf{q},\mathbf{x}} \left| h^{d/2} \frac{\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\mathsf{T} \bar{\boldsymbol{Q}}_{\mathbf{q}}^{-1}}{\sqrt{\bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}} \boldsymbol{Z}_n^*(\mathbf{q}) \right|. \tag{G.9}$$

Let $\tilde{k}^*(\eta)$ denote the conditional $\eta$-quantile of $\tilde{V}^*$ given $\{\mathbf{x}_i\}_{i=1}^n$.

**Lemma G.4** (Closeness rates)**.** *Random variables $V$, $\tilde{V}$, $V^*$, $\tilde{V}^*$ satisfy the following:*

*(a)* $\left|V - \tilde{V}\right| = o_{\mathbb{P}}(r_{\mathrm{str}} + r_{\mathrm{ho}})$*;*

*(b)* $\left|V^* - \tilde{V}^*\right| = o_{\mathbb{P}}(r_{\mathrm{str}})$*;*

*(c)* $\tilde{V}^*$ *depends on the data only via* $\{\mathbf{x}_i\}_{i=1}^n$*.*

*Proof.* By Lemma F.8 and Theorem F.4, we have

$$\left| V - \sup_{\mathbf{q},\mathbf{x}} \left| \bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v})^\mathsf{T} \boldsymbol{Z}_n(\mathbf{q}) \right| \right| = o_{\mathbb{P}}(r_{\mathrm{str}} + r_{\mathrm{ho}}),$$

which is Assertion (a).

By Theorem G.1, we have

$$\left| V^* - \sup_{\mathbf{q},\mathbf{x}} \left| \bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v})^\mathsf{T} \boldsymbol{Z}_n^*(\mathbf{q}) \right| \right| = o_{\mathbb{P}}(r_{\mathrm{str}}),$$

which is Assertion (b).

Assertion (c) follows from the definition of the process $\boldsymbol{Z}_n^*(\cdot)$ and the fact that $\bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v})$ only depends on the data via $\{\mathbf{x}_i\}_{i=1}^n$. □

**Lemma G.5** (First sequence)**.** *There exists a sequence of positive numbers $\nu_{n,1} \to 0$ such that w. p. a. 1*

$$k^*(1-\alpha) \le \tilde{k}^*(1-\alpha+\nu_{n,1}) + r_{\mathrm{str}} \quad \text{and} \quad k^*(1-\alpha) \ge \tilde{k}^*(1-\alpha-\nu_{n,1}) - r_{\mathrm{str}}.$$

*Proof.* This follows from $|V^* - \tilde{V}^*| = o_{\mathbb{P}}(r_{\mathrm{str}})$ by Lemma G.4, directly applying Lemma C.8. □

**Lemma G.6** (Second sequence)**.** *There exists a constant $C_{\tilde{V}^*} > 0$ such that w. p. a. 1*

$$\sup_{u\in\mathbb{R}} \mathbb{P}\left\{ \left|\tilde{V}^* - u\right| \le 2r_{\mathrm{str}} + r_{\mathrm{ho}} \,\middle|\, \boldsymbol{D}_n \right\} \le C_{\tilde{V}^*}(r_{\mathrm{str}} + r_{\mathrm{ho}})\sqrt{\log(1/h)} =: \nu_{n,2},$$

*Moreover, for the sequence $\nu_{n,2} \to 0$ just defined, the following holds w. p. a. 1:*

$$\tilde{k}^*(1-\alpha-\nu_{n,1}) - \tilde{k}^*(1-\alpha-\nu_{n,1}-\nu_{n,2}) \ge 2r_{\mathrm{str}} + r_{\mathrm{ho}}, \tag{G.10}$$

$$\tilde{k}^*(1-\alpha+\nu_{n,1}+\nu_{n,2}) - \tilde{k}^*(1-\alpha+\nu_{n,1}) \ge 2r_{\mathrm{str}} + r_{\mathrm{ho}}. \tag{G.11}$$

*Proof.* By Lemma C.9, using that $\bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v})^\mathsf{T}\boldsymbol{Z}_n^*(\mathbf{q})$ is a separable mean-zero Gaussian conditionally on $\{\mathbf{x}_i\}_{i=1}^n$ process on $\mathcal{Q}\times\mathcal{X}$ with $\mathbb{E}\big[\big(\bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v})^\mathsf{T}\boldsymbol{Z}_n^*(\mathbf{q})\big)^2 \bigm| \{\mathbf{x}_i\}_{i=1}^n\big] = 1$, we have w. p. a. 1

$$\begin{aligned}
&\sup_{u\in\mathbb{R}}\mathbb{P}\Big\{\Big|\tilde{V}^* - u\Big| \le 2r_{\mathrm{str}} + r_{\mathrm{ho}} \Bigm| \boldsymbol{D}_n\Big\}\\
&\quad\le 4(2r_{\mathrm{str}} + r_{\mathrm{ho}})\left(\mathbb{E}\left[\sup_{\mathbf{q},\mathbf{x}}\left|\bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v})^\mathsf{T}\boldsymbol{Z}_n^*(\mathbf{q})\right| \Bigm| \{\mathbf{x}_i\}_{i=1}^n\right] + 1\right)\\
&\quad\le 4(2r_{\mathrm{str}} + r_{\mathrm{ho}})\left(\sup_{\mathbf{q},\mathbf{x}}\left\|\bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v})\right\|_1\mathbb{E}\left[\sup_{\mathbf{q}}\|\boldsymbol{Z}_n^*(\mathbf{q})\|_\infty \Bigm| \{\mathbf{x}_i\}_{i=1}^n\right] + 1\right)\\
&\quad\overset{(a)}{\lesssim} 4(2r_{\mathrm{str}} + r_{\mathrm{ho}})\left(\mathbb{E}\left[\sup_{\mathbf{q}}\|\boldsymbol{Z}_n^*(\mathbf{q})\|_\infty \Bigm| \{\mathbf{x}_i\}_{i=1}^n\right] + 1\right)\\
&\quad\overset{(b)}{\lesssim} (r_{\mathrm{str}} + r_{\mathrm{ho}})\sqrt{\log(1/h)},
\end{aligned}$$

where in (a) we used $\left\|\bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v})\right\|_1 \lesssim 1$ w. p. a. 1 by Lemma F.7; (b) is by Lemma G.2.

We will now prove (G.10). Note that $\sup_{u\in\mathbb{R}}\mathbb{P}\{|\tilde{V}^* - u| \le 2r_{\mathrm{str}} + r_{\mathrm{ho}} \bigm| \boldsymbol{D}_n\} \le \nu_{n,2}$ w. p. a. 1 by Lemma G.6 implies that w. p. a. 1

$$\sup_{u\in\mathbb{R}}\mathbb{P}\Big\{u < \tilde{V}^* \le u + 2r_{\mathrm{str}} + r_{\mathrm{ho}} \Bigm| \boldsymbol{D}_n\Big\} \le \nu_{n,2}.$$

This can be rewritten as

$$\sup_{u\in\mathbb{R}}\Big\{\mathbb{P}\Big\{\tilde{V}^* \le u + 2r_{\mathrm{str}} + r_{\mathrm{ho}} \Bigm| \boldsymbol{D}_n\Big\} - \mathbb{P}\Big\{\tilde{V}^* \le u \Bigm| \boldsymbol{D}_n\Big\}\Big\} \le \nu_{n,2}.$$

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 $\mathbb{P}\Big\{\tilde{V}^* \le \tilde{k}^*(1-\alpha-\nu_{n,1}-\nu_{n,2}) \Bigm| \boldsymbol{D}_n\Big\} \ge 1-\alpha-\nu_{n,1}-\nu_{n,2}$, this gives w. p. a. 1

$$\begin{aligned}
&\mathbb{P}\Big\{\tilde{V}^* \le \tilde{k}^*(1-\alpha-\nu_{n,1}-\nu_{n,2}) + 2r_{\mathrm{str}} + r_{\mathrm{ho}} \Bigm| \boldsymbol{D}_n\Big\}\\
&\qquad - (1-\alpha-\nu_{n,1}-\nu_{n,2}) \le \nu_{n,2}
\end{aligned}$$

or

$$\mathbb{P}\Big\{\tilde{V}^* \le \tilde{k}^*(1-\alpha-\nu_{n,1}-\nu_{n,2}) + 2r_{\mathrm{str}} + r_{\mathrm{ho}} \Bigm| \boldsymbol{D}_n\Big\} \le 1-\alpha-\nu_{n,1}.$$

By monotonicity of a (conditional) distribution function, this means that w. p. a. 1

$$\tilde{k}^*(1-\alpha-\nu_{n,1}-\nu_{n,2}) + 2r_{\mathrm{str}} + r_{\mathrm{ho}} \le \tilde{k}^*(1-\alpha-\nu_{n,1}).$$

This proves the inequality (G.10). The inequality (G.11) is proven similarly. □

*Concluding the proof of Theorem G.3.* Note that

$$\begin{aligned}
\mathbb{P}\{V > k^*(1-\alpha)\} &\overset{(a)}{\le} \mathbb{P}\Big\{V > \tilde{k}^*(1-\alpha-\nu_{n,1}) - r_{\mathrm{str}}\Big\} + o(1)\\
&\overset{(b)}{\le} \mathbb{P}\Big\{\tilde{V} > \tilde{k}^*(1-\alpha-\nu_{n,1}) - (2r_{\mathrm{str}} + r_{\mathrm{ho}})\Big\} + o(1)\\
&\overset{(c)}{\le} \mathbb{P}\Big\{\tilde{V} > \tilde{k}^*(1-\alpha-\nu_{n,1}-\nu_{n,2})\Big\} + o(1)
\end{aligned}$$

$$\stackrel{(d)}{=} \mathbb{E}\Big[\mathbb{P}\Big\{\tilde{V} > \tilde{k}^*(1-\alpha-\nu_{n,1}-\nu_{n,2}) \,\Big|\, \{\mathbf{x}_i\}_{i=1}^n\Big\}\Big] + o(1)$$
$$\stackrel{(e)}{\le} \alpha + \nu_{n,1} + \nu_{n,2} + o(1) = \alpha + o(1),$$

where (a) is by Lemma G.5, (b) is by Assertion (a) in Lemma G.4, (c) is by Lemma G.6, (d) is by the law of iterated expectations, (e) is by the definition of a conditional quantile and using that $\tilde{V}$ has the same conditional distribution as $\tilde{V}^*$.

Similarly,

$$\mathbb{P}\{V > k^*(1-\alpha)\} \stackrel{(a)}{\ge} \mathbb{P}\Big\{V > \tilde{k}^*(1-\alpha+\nu_{n,1}) + r_{\mathrm{str}}\Big\} + o(1)$$
$$\stackrel{(b)}{\ge} \mathbb{P}\Big\{\tilde{V} > \tilde{k}^*(1-\alpha+\nu_{n,1}) + (2r_{\mathrm{str}} + r_{\mathrm{ho}})\Big\} + o(1)$$
$$\stackrel{(c)}{\ge} \mathbb{P}\Big\{\tilde{V} > \tilde{k}^*(1-\alpha+\nu_{n,1}+\nu_{n,2})\Big\} + o(1)$$
$$\stackrel{(d)}{=} \mathbb{E}\Big[\mathbb{P}\Big\{\tilde{V} > \tilde{k}^*(1-\alpha+\nu_{n,1}+\nu_{n,2}) \,\Big|\, \{\mathbf{x}_i\}_{i=1}^n\Big\}\Big] + o(1)$$
$$\stackrel{(e)}{=} \alpha - \nu_{n,1} - \nu_{n,2} + o(1) = \alpha + o(1),$$

where (a) is by Lemma G.5, (b) is by Assertion (a) in Lemma G.4, (c) is by Lemma G.6, (d) is by the law of iterated expectations, in (e) we used that the distribution function of $\tilde{V}$ conditional on $\{\mathbf{x}_i\}_{i=1}^n$ is continuous w. p. a. 1, because by the same anti-concentration argument as in the proof of Lemma G.6 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 $\varepsilon > 0$

$$\sup_{u\in\mathbb{R}} \mathbb{P}\Big\{|\tilde{V} - u| \le \varepsilon \,\Big|\, \{\mathbf{x}_i\}_{i=1}^n\Big\} \le C\varepsilon\sqrt{\log(1/h)}.$$

In particular, on $\mathcal{A}_n$ all jumps of the distribution function of $\tilde{V}$ conditional on $\{\mathbf{x}_i\}_{i=1}^n$ are bounded by $C\varepsilon\sqrt{\log(1/h)}$, which implies that the distribution function is continuous on $\mathcal{A}_n$, since $\varepsilon$ is arbitrary. Theorem G.3 is proven. □

**Theorem G.7** (K-S Distance)**.** *Suppose the following holds:*

*(i) All the conditions of Theorem G.1 hold, and* $(r_{\mathrm{str}} + r_{\mathrm{ho}})\sqrt{\log(1/h)} = o(1)$.

*(ii) There exists a $K$-dimensional mean-zero unconditionally Gaussian process $\widetilde{\boldsymbol{Z}}_n(\mathbf{q})$ on $\mathcal{Q}$ with a. s. continuous sample paths such that*

$$\mathbb{E}\Big[\Big(\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}_n(\mathbf{q})\Big)^2 \,\Big|\, \{\mathbf{x}_i\}_{i=1}^n\Big] = 1$$

*and*

$$\mathbb{E}\Big[\sup_{\mathbf{q}\in\mathcal{Q}}\Big\|\boldsymbol{Z}_n(\mathbf{q}) - \widetilde{\boldsymbol{Z}}_n(\mathbf{q})\Big\|_\infty \,\Big|\, \{\mathbf{x}_i\}_{i=1}^n\Big] = o_{\mathbb{P}}(r_{\mathrm{str}}). \tag{G.12}$$

*(iii) On the same probability space there exists a copy, conditional on $\{\mathbf{x}_i\}_{i=1}^n$, of $\widetilde{\boldsymbol{Z}}_n(\cdot)$, denoted $\widetilde{\boldsymbol{Z}}^*_n(\cdot)$, independent of the data and such that*

$$\mathbb{E}\Big[\sup_{\mathbf{q}\in\mathcal{Q}}\Big\|\boldsymbol{Z}^*_n(\mathbf{q}) - \widetilde{\boldsymbol{Z}}^*_n(\mathbf{q})\Big\|_\infty \,\Big|\, \{\mathbf{x}_i\}_{i=1}^n\Big] = o_{\mathbb{P}}(r_{\mathrm{str}}). \tag{G.13}$$

*Then*

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\left\{\sup_{\mathbf{q}\in\mathcal{Q}}\sup_{\mathbf{x}\in\mathcal{X}}|T(\mathbf{x},\mathbf{q},\mathbf{v})|\leq u\right\}-\mathbb{P}\left\{\sup_{\mathbf{q}\in\mathcal{Q}}\sup_{\mathbf{x}\in\mathcal{X}}\left|\frac{h^{d/2}\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^{\mathsf{T}}\widehat{\boldsymbol{Q}}_{\mathbf{q}}^{-1}}{\sqrt{\widehat{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}}\widehat{\boldsymbol{Z}}_n^*(\mathbf{q})\right|\leq u\;\middle|\;\boldsymbol{D}_n\right\}\right|=o_{\mathbb{P}}(1).$$

*Proof.* Since $\widetilde{\boldsymbol{Z}}_n^*(\cdot)$ is a copy of $\widetilde{\boldsymbol{Z}}_n(\cdot)$ conditional on $\{\mathbf{x}_i\}_{i=1}^n$, in particular they have the same unconditional laws, giving

$$\mathbb{P}\Big\{\sup_{\mathbf{q},\mathbf{x}}\Big|\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}_n(\mathbf{q})\Big|\leq u\Big\}=\mathbb{P}\Big\{\sup_{\mathbf{q},\mathbf{x}}\Big|\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}_n^*(\mathbf{q})\Big|\leq u\Big\}.$$

This means that, by the triangle inequality, it is sufficient to prove the following bounds:

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\left\{\sup_{\mathbf{q},\mathbf{x}}|T(\mathbf{x},\mathbf{q},\mathbf{v})|\leq u\right\}-\mathbb{P}\left\{\sup_{\mathbf{q},\mathbf{x}}\left|\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}_n(\mathbf{q})\right|\leq u\right\}\right|=o(1), \tag{G.14}$$

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\left\{\sup_{\mathbf{q},\mathbf{x}}\left|\hat{\boldsymbol{\ell}}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widehat{\boldsymbol{Z}}_n^*(\mathbf{q})\right|\leq u\;\middle|\;\boldsymbol{D}_n\right\}-\mathbb{P}\left\{\sup_{\mathbf{q},\mathbf{x}}\left|\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}_n^*(\mathbf{q})\right|\leq u\right\}\right|=o_{\mathbb{P}}(1). \tag{G.15}$$

We will now prove (G.14). Note that for any random variables $\xi$ and $\eta$ and any $s>0$,

$$\sup_{u\in\mathbb{R}}|\mathbb{P}\{\xi\leq u\}-\mathbb{P}\{\eta\leq u\}|\leq\sup_{u\in\mathbb{R}}\mathbb{P}\{|\eta-u|\leq s\}+\mathbb{P}\{|\xi-\eta|>s\}, \tag{G.16}$$

which follows from the two bounds

$$\begin{aligned}
\mathbb{P}\{\xi\leq u\}&\leq\mathbb{P}\{\xi\leq u\text{ and }|\xi-\eta|\leq s\}+\mathbb{P}\{|\xi-\eta|>s\}\\
&\leq\mathbb{P}\{\eta\leq u+s\}+\mathbb{P}\{|\xi-\eta|>s\}\\
&\leq\mathbb{P}\{\eta\leq u\}+\mathbb{P}\{u<\eta\leq u+s\}+\mathbb{P}\{|\xi-\eta|>s\};\\
\mathbb{P}\{\xi>u\}&\leq\mathbb{P}\{\xi>u\text{ and }|\xi-\eta|\leq s\}+\mathbb{P}\{|\xi-\eta|>s\}\\
&\leq\mathbb{P}\{\eta>u-s\}+\mathbb{P}\{|\xi-\eta|>s\}\\
&\leq\mathbb{P}\{\eta>u\}+\mathbb{P}\{u-s<\eta\leq u\}+\mathbb{P}\{|\xi-\eta|>s\}.
\end{aligned}$$

By $\sup_{\mathbf{q},\mathbf{x}}\|\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})\|_1\lesssim 1$ (from Lemma F.7) and (G.12), a simple application of Lemma C.2 gives

$$h^{d/2}\sup_{\mathbf{q},\mathbf{x}}\left|\frac{\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^{\mathsf{T}}\mathbf{Q}_{0,\mathbf{q}}^{-1}}{\sqrt{\boldsymbol{\Omega}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}}\boldsymbol{Z}_n(\mathbf{q})-\frac{\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^{\mathsf{T}}\mathbf{Q}_{0,\mathbf{q}}^{-1}}{\sqrt{\boldsymbol{\Omega}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}}\widetilde{\boldsymbol{Z}}_n(\mathbf{q})\right|=o_{\mathbb{P}}(r_{\mathrm{str}}). \tag{G.17}$$

By (F.25) in Lemma F.7, we have $\sup_{\mathbf{q},\mathbf{x}}\|\bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v})-\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})\|_1\lesssim_{\mathbb{P}}\sqrt{\frac{\log(1/h)}{nh^d}}$, which by Lemma G.2 gives $\sup_{\mathbf{q},\mathbf{x}}|\bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\boldsymbol{Z}_n(\mathbf{q})-\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\boldsymbol{Z}_n(\mathbf{q})|\lesssim_{\mathbb{P}}\frac{\log(1/h)}{\sqrt{nh^d}}=o(r_{\mathrm{str}})$. Combining it with Theorem F.4, Lemma F.8 and (G.17),

$$\mathbb{P}\left\{\left|\sup_{\mathbf{q},\mathbf{x}}|T(\mathbf{x},\mathbf{q},\mathbf{v})|-\sup_{\mathbf{q},\mathbf{x}}\left|\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}_n(\mathbf{q})\right|\right|>r_{\mathrm{str}}+r_{\mathrm{ho}}\right\}=o(1).$$

Then we can apply (G.16) with $\xi=\sup_{\mathbf{q},\mathbf{x}}|T(\mathbf{x},\mathbf{q},\mathbf{v})|$, $\eta=\sup_{\mathbf{q},\mathbf{x}}\left|\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}_n(\mathbf{q})\right|$ and $s=r_{\mathrm{str}}+r_{\mathrm{ho}}$, and get

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\left\{\sup_{\mathbf{q},\mathbf{x}}|T(\mathbf{x},\mathbf{q},\mathbf{v})|\leq u\right\}-\mathbb{P}\left\{\sup_{\mathbf{q},\mathbf{x}}\left|\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}_n(\mathbf{q})\right|\leq u\right\}\right|$$

$$\leq \sup_{u\in\mathbb{R}} \mathbb{P}\left\{ \left| \sup_{\mathbf{q},\mathbf{x}} \left| \boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}} \widetilde{\boldsymbol{Z}}_n(\mathbf{q}) \right| - u \right| \leq r_{\mathrm{str}} + r_{\mathrm{ho}} \right\} + o(1).$$

For (G.14), it is left to show that

$$\sup_{u\in\mathbb{R}} \mathbb{P}\left\{ \left| \sup_{\mathbf{q},\mathbf{x}} \left| \boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}} \widetilde{\boldsymbol{Z}}_n(\mathbf{q}) \right| - u \right| \leq r_{\mathrm{str}} + r_{\mathrm{ho}} \right\} = o(1). \tag{G.18}$$

We apply the Gaussian anti-concentration result given in Corollary 2.1 in [7]:

$$\begin{aligned}
&\sup_{u\in\mathbb{R}} \mathbb{P}\left\{ \left| \sup_{\mathbf{q},\mathbf{x}} \left| \boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}} \widetilde{\boldsymbol{Z}}_n(\mathbf{q}) \right| - u \right| \leq r_{\mathrm{str}} + r_{\mathrm{ho}} \right\} \\
&\quad \leq 4(r_{\mathrm{str}} + r_{\mathrm{ho}}) \left( \mathbb{E}\left[ \sup_{\mathbf{q},\mathbf{x}} \left| \boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}} \widetilde{\boldsymbol{Z}}_n(\mathbf{q}) \right| \right] + 1 \right) \\
&\quad \leq 4(r_{\mathrm{str}} + r_{\mathrm{ho}}) \left( \mathbb{E}\left[ \sup_{\mathbf{q},\mathbf{x}} \|\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})\|_1 \left\| \widetilde{\boldsymbol{Z}}_n(\mathbf{q}) \right\|_\infty \right] + 1 \right) \\
&\quad \overset{\text{(a)}}{\lesssim} (r_{\mathrm{str}} + r_{\mathrm{ho}}) \left( \mathbb{E}\left[ \sup_{\mathbf{q}} \left\| \widetilde{\boldsymbol{Z}}_n(\mathbf{q}) \right\|_\infty \right] + 1 \right) \\
&\quad \overset{\text{(b)}}{\lesssim} (r_{\mathrm{str}} + r_{\mathrm{ho}}) \sqrt{\log(1/h)},
\end{aligned} \tag{G.19}$$

where in (a) we used $\|\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})\|_1 \lesssim 1$, see (F.23); (b) is by

$$\begin{aligned}
\mathbb{E}\left[ \sup_{\mathbf{q}} \left\| \widetilde{\boldsymbol{Z}}_n(\mathbf{q}) \right\|_\infty \right] &= \mathbb{E}\left[ \sup_{\mathbf{q}} \left\| \widetilde{\boldsymbol{Z}}_n(\mathbf{q}) \right\|_\infty \;\middle|\; \{\mathbf{x}_i\}_{i=1}^n \right] \\
&\leq \mathbb{E}\left[ \sup_{\mathbf{q}} \left\| \widetilde{\boldsymbol{Z}}_n(\mathbf{q}) - \boldsymbol{Z}_n(\mathbf{q}) \right\|_\infty \;\middle|\; \{\mathbf{x}_i\}_{i=1}^n \right] + \mathbb{E}\left[ \sup_{\mathbf{q}} \|\boldsymbol{Z}_n(\mathbf{q})\|_\infty \;\middle|\; \{\mathbf{x}_i\}_{i=1}^n \right] \\
&\overset{\text{(c)}}{=} o_{\mathbb{P}}(r_{\mathrm{str}}) + O_{\mathbb{P}}\left( \sqrt{\log(1/h)} \right) \overset{\text{(d)}}{=} O_{\mathbb{P}}\left( \sqrt{\log(1/h)} \right)
\end{aligned}$$

using (G.12), Lemma G.2 for (c), $r_{\mathrm{str}} \lesssim \sqrt{\log(1/h)}$ for (d), and noting that

$$\mathbb{E}\left[ \sup_{\mathbf{q}} \left\| \widetilde{\boldsymbol{Z}}_n(\mathbf{q}) \right\|_\infty \right] \lesssim_{\mathbb{P}} \sqrt{\log(1/h)}$$

is equivalent to

$$\mathbb{E}\left[ \sup_{\mathbf{q}} \left\| \widetilde{\boldsymbol{Z}}_n(\mathbf{q}) \right\|_\infty \right] \lesssim \sqrt{\log(1/h)}.$$

Since $(r_{\mathrm{str}} + r_{\mathrm{ho}})\sqrt{\log(1/h)} = o(1)$, the right-hand side in (G.19) is $o(1)$, proving (G.18), which was sufficient for (G.14).

We will now prove (G.15). By $\sup_{\mathbf{q},\mathbf{x}} \|\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})\|_1 \lesssim 1$ (from Lemma F.7) and (G.13), a simple application of Lemma C.2 gives

$$h^{d/2} \sup_{\mathbf{q},\mathbf{x}} \left| \frac{\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^{\mathsf{T}} \mathbf{Q}_{0,\mathbf{q}}^{-1}}{\sqrt{\boldsymbol{\Omega}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}} \boldsymbol{Z}_n^*(\mathbf{q}) - \frac{\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^{\mathsf{T}} \mathbf{Q}_{0,\mathbf{q}}^{-1}}{\sqrt{\boldsymbol{\Omega}_{\mathbf{v}}(\mathbf{x},\mathbf{q})}} \widetilde{\boldsymbol{Z}}_n^*(\mathbf{q}) \right| = o_{\mathbb{P}}(r_{\mathrm{str}}). \tag{G.20}$$

By Theorem G.1, we have $\sup_{\mathbf{q},\mathbf{x}}\left|\hat{\boldsymbol{\ell}}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widehat{\boldsymbol{Z}}_n^*(\mathbf{q}) - \bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v})\boldsymbol{Z}_n^*(\mathbf{q})\right| = o_{\mathbb{P}}(r_{\mathrm{str}})$. By (F.25) in Lemma F.7, we have $\sup_{\mathbf{q},\mathbf{x}}\|\bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v}) - \boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})\|_1 \lesssim_{\mathbb{P}} \sqrt{\frac{\log(1/h)}{nh^d}}$, which by Lemma G.2 gives $\sup_{\mathbf{q},\mathbf{x}}|\bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\boldsymbol{Z}_n^*(\mathbf{q}) - \boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\boldsymbol{Z}_n^*(\mathbf{q})| \lesssim_{\mathbb{P}} \frac{\log(1/h)}{\sqrt{nh^d}} = o(r_{\mathrm{str}})$. By the triangle inequality,

$$\sup_{\mathbf{q},\mathbf{x}}|\hat{\boldsymbol{\ell}}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widehat{\boldsymbol{Z}}_n^*(\mathbf{q}) - \boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\boldsymbol{Z}_n^*(\mathbf{q})| = o_{\mathbb{P}}(r_{\mathrm{str}}).$$

Combining with (G.20) and applying the triangle inequality again, we obtain

$$\mathbb{P}\left\{\left|\sup_{\mathbf{q},\mathbf{x}}\left|\hat{\boldsymbol{\ell}}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widehat{\boldsymbol{Z}}_n^*(\mathbf{q})\right| - \sup_{\mathbf{q},\mathbf{x}}\left|\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}_n^*(\mathbf{q})\right|\right| > r_{\mathrm{str}}\right\} = o(1),$$

implying by Markov's inequality that for any $\varepsilon > 0$,

$$\mathbb{P}\left\{\mathbb{P}\left\{\left|\sup_{\mathbf{q},\mathbf{x}}\left|\hat{\boldsymbol{\ell}}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widehat{\boldsymbol{Z}}_n^*(\mathbf{q})\right| - \sup_{\mathbf{q},\mathbf{x}}\left|\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}_n^*(\mathbf{q})\right|\right| > r_{\mathrm{str}} \,\middle|\, \boldsymbol{D}_n\right\} > \varepsilon\right\} = o(1),$$

that is,

$$\mathbb{P}\left\{\left|\sup_{\mathbf{q},\mathbf{x}}\left|\hat{\boldsymbol{\ell}}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widehat{\boldsymbol{Z}}_n^*(\mathbf{q})\right| - \sup_{\mathbf{q},\mathbf{x}}\left|\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}_n^*(\mathbf{q})\right|\right| > r_{\mathrm{str}} \,\middle|\, \boldsymbol{D}_n\right\} = o_{\mathbb{P}}(1).$$

Then we can apply (G.16) with $\mathbb{P}\{\cdot \mid \boldsymbol{D}_n\}$ instead of $\mathbb{P}\{\cdot\}$, $\xi = \sup_{\mathbf{q},\mathbf{x}}\left|\hat{\boldsymbol{\ell}}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widehat{\boldsymbol{Z}}_n^*(\mathbf{q})\right|$, $\eta = \sup_{\mathbf{q},\mathbf{x}}\left|\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}_n^*(\mathbf{q})\right|$ and $s = r_{\mathrm{str}}$, and get

$$\begin{aligned}
&\sup_{u\in\mathbb{R}}\left|\mathbb{P}\left\{\sup_{\mathbf{q},\mathbf{x}}\left|\hat{\boldsymbol{\ell}}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widehat{\boldsymbol{Z}}_n^*(\mathbf{q})\right| \le u \,\middle|\, \boldsymbol{D}_n\right\} - \mathbb{P}\left\{\sup_{\mathbf{q},\mathbf{x}}\left|\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}_n^*(\mathbf{q})\right| \le u\right\}\right| \\
&\quad\le \sup_{u\in\mathbb{R}}\mathbb{P}\left\{\left|\sup_{\mathbf{q},\mathbf{x}}\left|\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}_n^*(\mathbf{q})\right| - u\right| \le r_{\mathrm{str}}\right\} + o_{\mathbb{P}}(1) \\
&\quad= \sup_{u\in\mathbb{R}}\mathbb{P}\left\{\left|\sup_{\mathbf{q},\mathbf{x}}\left|\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}_n(\mathbf{q})\right| - u\right| \le r_{\mathrm{str}}\right\} + o_{\mathbb{P}}(1),
\end{aligned}$$

where we used that $\widetilde{\boldsymbol{Z}}_n^*(\mathbf{q})$ is independent of the data allowing us to remove the conditioning on $\boldsymbol{D}_n$, and again that $\widetilde{\boldsymbol{Z}}_n(\cdot)$ and $\widetilde{\boldsymbol{Z}}_n^*(\cdot)$ have the same laws.

It is left to use that

$$\begin{aligned}
&\sup_{u\in\mathbb{R}}\mathbb{P}\left\{\left|\sup_{\mathbf{q},\mathbf{x}}\left|\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}_n(\mathbf{q})\right| - u\right| \le r_{\mathrm{str}}\right\} \\
&\quad\le \sup_{u\in\mathbb{R}}\mathbb{P}\left\{\left|\sup_{\mathbf{q},\mathbf{x}}\left|\boldsymbol{\ell}(\mathbf{x},\mathbf{q},\mathbf{v})^{\mathsf{T}}\widetilde{\boldsymbol{Z}}_n(\mathbf{q})\right| - u\right| \le r_{\mathrm{str}} + r_{\mathrm{ho}}\right\} \overset{\text{(a)}}{=} o(1),
\end{aligned}$$

where (a) is by (G.18).

Theorem G.7 is proven. □

# 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 [1] is a special case of our setting.

**Proposition H.1** (Verification of Assumption B.2 for quantile regression)**.** *Consider the model in Eq. (A.1) with $\mathcal{Q} = [\varepsilon_0, 1 - \varepsilon_0]$ for some positive $\varepsilon_0 < 1/2$ and $\rho(y, \eta; q) = (q - \mathbb{1}\{y < \eta\})(y - \eta)$. Suppose the following conditions.*

*(i) Assumptions B.1, B.3, B.4 hold.*

*(ii) $\eta \mapsto F_{Y|X}(\eta|\mathbf{x})$ is twice continuously differentiable with first derivative $f_{Y|X}(\eta|\mathbf{x})$ (in particular, $\mathfrak{M}$ is the Lebesgue measure).*

*(iii) $\mathbb{E}[|y_1|] < \infty$.*

*(iv) The real inverse link function $\eta(\cdot)\colon \mathbb{R} \to \mathcal{E}$ is strictly monotonic and two times continuously differentiable, where $\mathcal{E}$ is an open connected subset of $\mathbb{R}$ containing the $q$-quantile of $y_1$ conditionally on $\mathbf{x}_1 = \mathbf{x}$ for all $(\mathbf{x}, q)$.*

*(v) $f_{Y|X}\big(\eta\big(\mu_0(\mathbf{x}, q)\big)|\mathbf{x}\big)$ is bounded away from zero uniformly over $q \in \mathcal{Q}$, $\mathbf{x} \in \mathcal{X}$; the derivative of $y \mapsto f_{Y|X}(y|\mathbf{x})$ is continuous and bounded in absolute value from above uniformly over $y \in \mathcal{Y}_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}$.*

*(vi) $\eta\big(\mu_0(\mathbf{x}, q)\big)$ is the $q$-quantile of the conditional distribution of $y_1$ given $\mathbf{x}_1 = \mathbf{x}$, where $\mathbf{x} \in \mathcal{X}$.*

*Then Assumption B.2, its stronger version described in Section E.2, and (F.14) are also true.*

*Remark.* Taking $\eta\big(\mu_0(\mathbf{x}, q)\big)$ 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 $\mu_0(\mathbf{x}, q)$ is Lipschitz in parameter (as we do in Assumption B.1(iii)). Since

$$\frac{\partial}{\partial q}\mu_0(\mathbf{x}, q) = \frac{1}{f_{Y|X}(\mu_0(\mathbf{x}, q)|\mathbf{x})},$$

the Lipschitz property follows from $f_{Y|X}(\mu_0(\mathbf{x}, q)|\mathbf{x})$ being bounded from below uniformly over $\mathbf{x} \in \mathcal{X}$, $q \in \mathcal{Q}$.

*Proof.* We will verify the assumptions one by one.

**Verifying Assumption B.2(i)** Clearly, $\rho(y, \eta; q)$ is convex in $\eta$ and the a. e. derivative in $\eta$ is $\psi(y, \eta; q) \equiv \mathbb{1}\{y - \eta < 0\} - q$ is piecewise constant with only one jump (therefore piecewise Hölder with $\alpha = 1$). $\eta$ is strictly monotonic and three times continuously differentiable by assumption.

**Verifying Assumption B.2(ii)** Since $\eta(\mu_0(\cdot, q))$ is the conditional $q$-quantile, we have

$$\mathbb{E}[\psi(y_i, \eta(\mu_0(\mathbf{x}_i, q)); q) \mid \mathbf{x}_i] = \mathbb{E}[\mathbb{1}\{y_i < \eta(\mu_0(\mathbf{x}_i, q))\} - q \mid \mathbf{x}_i] = q - q = 0$$

and

$$\begin{aligned}\sigma_q^2(\mathbf{x}) &= \mathbb{E}\Big[\psi(y_i, \eta(\mu_0(\mathbf{x}_i, q)); q)^2 \,\Big|\, \mathbf{x}_i = \mathbf{x}\Big] \\ &= \mathbb{E}\Big[(\mathbb{1}\{y_i < \eta(\mu_0(\mathbf{x}_i, q))\} - q)^2 \,\Big|\, \mathbf{x}_i = \mathbf{x}\Big] = q - 2q^2 + q^2 = q(1 - q)\end{aligned}$$

is constant in $\mathbf{x}$ (in particular continuous in $\mathbf{x}$) and bounded away from zero since both $q$ and $1 - q$ are bounded away from zero. Since $q(1 - q)$ is smooth, $\sigma_q^2(\mathbf{x})$ is Lipshitz in $q$. The family $\{\psi(y_i, \eta(\mu_0(\mathbf{x}_i, q)); q)\colon q \in \mathcal{Q}\}$ has a positive measurable envelope 1 which has uniformly bounded conditional moments of any order.

**Verifying Assumption B.2(iii)** The conditional expectation

$$\Psi(x,\eta;q) = \mathbb{E}[\mathbb{1}\{y<\eta\} - q \mid \mathbf{x}_i = \mathbf{x}] = \int_{-\infty}^{\eta} f_{Y|X}(y|\mathbf{x})\,\mathrm{d}y - q$$

is twice continuously differentiable in $\eta$ (the integral on the right is a, possibly improper, Riemann integral) and its second derivative $f'_{Y|X}(\eta|\mathbf{x})$ is continuous and bounded in absolute value. By the mean value theorem, this means that $f_{Y|X}(\eta(\mu_0(\mathbf{x},q))|\mathbf{x})$ being bounded away from zero implies $f_{Y|X}(\eta(\zeta)|\mathbf{x})$ is bounded away from zero for $\zeta$ sufficiently close to $\mu_0(\mathbf{x},q)$. The bound on $|\Psi_1(\mathbf{x},\eta(\zeta);q)|$ from above in such a neighborhood (and in fact everywhere) is automatic since $f_{Y|X}(y|\mathbf{x})$ is bounded from above (uniformly over $y\in\mathcal{Y}_{\mathbf{x}}$, $\mathbf{x}\in\mathcal{X}$).

**Verifying Assumption B.2(iv)** This verification proceeds similarly to Lemmas 25–28 in [1].

The class of functions

$$\mathcal{W}_1 := \big\{(\mathbf{x},y)\mapsto \mathbb{1}\{y<\eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta})\} \colon \boldsymbol{\beta}\in\mathbb{R}^K\big\}$$

is VC with index $O(K)$ by Lemmas 2.6.15 and 2.6.18 in [11] (since $\eta(\cdot)$ is monotone). The class of functions

$$\mathcal{G}_2 = \{(\mathbf{x},y)\mapsto \mathbb{1}\{y<\eta(\mu_0(\mathbf{x},q))\}\colon q\in\mathcal{Q}\}$$

is VC with index 2 since $\eta(\mu_0(\mathbf{x},q))$ is increasing in $q$ for any $\mathbf{x}\in\mathcal{X}$, giving that the class of sets $\{(\mathbf{x},y) : y<\eta(\mu_0(\mathbf{x},q))\}$ 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

$$\mathcal{G}_1 = \{(\mathbf{x},y)\mapsto \mathbb{1}\{y<\eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta})\} - \mathbb{1}\{y<\eta(\mu_0(\mathbf{x},q))\} : \|\boldsymbol{\beta}-\boldsymbol{\beta}_0(q)\|_\infty \le r, q\in\mathcal{Q}\}$$

with envelope 2 satisfies the uniform entropy bound (A.3) with $A\lesssim 1$ and $V\lesssim K$ because it is a subclass of $\mathcal{W}_1-\mathcal{G}_2$.

For a fixed vector space $\mathcal{B}$ of dimension $\dim\mathcal{B}$,

$$\mathcal{W}_{\mathcal{B}} := \{(\mathbf{x},y)\mapsto \mathbb{1}\{y<\eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta})\} : \boldsymbol{\beta}\in\mathcal{B}\}$$

is VC with index $O(\dim\mathcal{B})$ by Lemmas 2.6.15 and 2.6.18 in [11] (since $\eta(\cdot)$ is monotone). Therefore, for any fixed $c>0$, $\delta\in\Delta$, the class

$$\mathcal{W}_{2,\delta} := \big\{(\mathbf{x},y)\mapsto \mathbb{1}\{y<\eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta})\}\mathbb{1}\big\{\mathbf{x}\in\mathcal{N}_{[c\log n]}(\delta)\big\} : \boldsymbol{\beta}\in\mathbb{R}^K\big\}$$

is VC with index $O\big(\log^d n\big)$ by Lemma 2.6.18 in [11] 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 $\mathbb{1}\big\{\mathbf{x}\in\mathcal{N}_{[c\log n]}(\delta)\big\}$. 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 $O\big(h^{-d}\big)$ such classes

$$\mathcal{W}_2 := \big\{(\mathbf{x},y)\mapsto \mathbb{1}\{y<\eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta})\}\mathbb{1}\big\{\mathbf{x}\in\mathcal{N}_{[c\log n]}(\delta)\big\} : \boldsymbol{\beta}\in\mathbb{R}^K, \delta\in\Delta\big\}$$

with envelope 1 satisfies the uniform entropy bound (A.3) with $A\lesssim 1$ and $V\lesssim\log^d n$, see (E.17). By Lemma C.4, the same is true of

$$\mathcal{G}_3 = \{(\mathbf{x},y)\mapsto$$

$$
\begin{aligned}
&[\mathbb{1}\{y < \eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta})\} - \mathbb{1}\{y < \eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}_0(q))\}]\mathbb{1}\big\{\mathbf{x} \in \mathcal{N}_{[c\log n]}(\delta)\big\} :\\
&\boldsymbol{\beta} \in \mathbb{R}^K, \delta \in \Delta, q \in \mathcal{Q}\big\}
\end{aligned}
$$

with envelope 2 because it is a subclass of $\mathcal{W}_2 - \mathcal{W}_2$.

The class of functions

$$
\mathcal{G}_4 = \big\{\mathbf{x} \mapsto f_{Y|X}(\eta(\mu_0(\mathbf{x}, q))|\mathbf{x}) : q \in \mathcal{Q}\big\}
$$

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 $\xi_{\mathbf{x},q_1,q_2}$ between $\eta(\mu_0(\mathbf{x}, q_1))$ and $\eta(\mu_0(\mathbf{x}, q_2))$

$$
\begin{aligned}
&\big|f_{Y|X}(\eta(\mu_0(\mathbf{x}, q_1))|\mathbf{x}) - f_{Y|X}(\eta(\mu_0(\mathbf{x}, q_2))|\mathbf{x})\big|\\
&\quad = \Big|f'_{Y|X}(\xi_{\mathbf{x},q_1,q_2})\Big| \cdot |\eta(\mu_0(\mathbf{x}, q_1)) - \eta(\mu_0(\mathbf{x}, q_2))|\\
&\quad \lesssim |\eta(\mu_0(\mathbf{x}, q_1)) - \eta(\mu_0(\mathbf{x}, q_2))| && \text{since } f'_{Y|X} \text{ is uniformly bounded}\\
&\quad \lesssim |\mu_0(\mathbf{x}, q_1) - \mu_0(\mathbf{x}, q_2)| && \text{since } \eta(\cdot) \text{ is Lipschitz}\\
&\quad \lesssim |q_1 - q_2| && \text{since } \mu_0(\mathbf{x}, q) \text{ is Lipschitz in } q.
\end{aligned}
$$

with constants in $\lesssim$ not depending on $\mathbf{x}, q_1, q_2$ or $n$ (this is also proven in Lemma 20 in [1]). 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

$$
\big\{(\mathbf{x}, y) \mapsto p_l(\mathbf{x})\mathbb{1}\{y < \eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta})\} : \boldsymbol{\beta} \in \mathbb{R}^K\big\}
$$

is VC 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 $p_l(\mathbf{x})$. 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

$$
\begin{aligned}
\mathcal{G}_5 = \Big\{&(\mathbf{x}, y) \mapsto\\
&p_l(\mathbf{x})[\mathbb{1}\{y < \eta(\mu_0(\mathbf{x}, q))\} - \mathbb{1}\{y < \eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}_0(q))\}] : q \in \mathcal{Q}\Big\}.
\end{aligned}
$$

**Verifying the addition to Assumption B.2(iv) described in Section E.2** The functions in the class have the form

$$
\begin{aligned}
&\rho(y, \eta(\mathbf{p}(\mathbf{x})^\intercal(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta})); q)\mathbb{1}\big(\mathbf{x} \in \mathcal{N}_{[c\log n]}(\delta)\big)\\
&\qquad - \rho(y, \eta(\mathbf{p}(\mathbf{x})^\intercal(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta} - \boldsymbol{v})); q)\mathbb{1}\big(\mathbf{x} \in \mathcal{N}_{[c\log n]}(\delta)\big)\\
&\qquad - [\eta(\mathbf{p}(\mathbf{x})^\intercal(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta})) - \eta(\mathbf{p}(\mathbf{x})^\intercal(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta} - \boldsymbol{v}))]\psi(y, \eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}_0(q)); q)\\
&\qquad \times \mathbb{1}\big(\mathbf{x} \in \mathcal{N}_{[c\log n]}(\delta)\big)\\
&\quad =: T_1 + T_2 + T_3 + T_4,
\end{aligned}
$$

where

$$
\begin{aligned}
T_1 &:= y[\mathbb{1}\{y < \eta(\mathbf{p}(\mathbf{x})^\intercal(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta} - \boldsymbol{v}))\} - \mathbb{1}\{y < \eta(\mathbf{p}(\mathbf{x})^\intercal(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta}))\}]\\
&\quad \times \mathbb{1}\big(\mathbf{x} \in \mathcal{N}_{[c\log n]}(\delta)\big),\\
T_2 &:= \eta(\mathbf{p}(\mathbf{x})^\intercal(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta}))\mathbb{1}\{y < \eta(\mathbf{p}(\mathbf{x})^\intercal(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta}))\}\mathbb{1}\big(\mathbf{x} \in \mathcal{N}_{[c\log n]}(\delta)\big),
\end{aligned}
$$

$$T_3 := -\eta(\mathbf{p}(\mathbf{x})^\intercal(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta} - \boldsymbol{v}))\mathbb{1}\{y < \eta(\mathbf{p}(\mathbf{x})^\intercal(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta} - \boldsymbol{v}))\}\mathbb{1}\big(\mathbf{x} \in \mathcal{N}_{[c\log n]}(\delta)\big),$$
$$T_4 := -\mathbb{1}\{y < \eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}_0(q))\}[\eta(\mathbf{p}(\mathbf{x})^\intercal(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta})) - \eta(\mathbf{p}(\mathbf{x})^\intercal(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta} - \boldsymbol{v}))] \times \mathbb{1}\big(\mathbf{x} \in \mathcal{N}_{[c\log n]}(\delta)\big).$$

Note that for $T_1$ to be nonzero, $y$ has to lie in a fixed interval (not depending on $n$), say $\Big[-\tilde{R}, \tilde{R}\Big]$. The class of functions

$$\Big\{(\mathbf{x}, y) \mapsto y\mathbb{1}\Big\{|y| \leq \tilde{R}\Big\}\mathbb{1}\{y < \eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta})\} : \boldsymbol{\beta} \in \mathcal{B}\Big\},$$

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 [11] (since $\eta(\cdot)$ is monotone and $y \mapsto y\mathbb{1}\Big\{|y| \leq \tilde{R}\Big\}$ is one fixed function). The class $\{(\mathbf{x}, y) \mapsto T_1, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \leq r, \|\boldsymbol{v}\|_\infty \leq \varepsilon_n, q \in \mathcal{Q}\}$ with $\delta$ 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 $O\big(h^{-d}\big)$ such classes

$$\{(\mathbf{x}, y) \mapsto T_1, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \leq r, \|\boldsymbol{v}\|_\infty \leq \varepsilon_n, \delta \in \Delta, q \in \mathcal{Q}\}$$

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.17).

Similarly, the classes

$$\{(\mathbf{x}, y) \mapsto T_2, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \leq r, \|\boldsymbol{v}\|_\infty \leq \varepsilon_n, \delta \in \Delta, q \in \mathcal{Q}\},$$
$$\{(\mathbf{x}, y) \mapsto T_3, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \leq r, \|\boldsymbol{v}\|_\infty \leq \varepsilon_n, \delta \in \Delta, q \in \mathcal{Q}\},$$
$$\{(\mathbf{x}, y) \mapsto T_4, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \leq r, \|\boldsymbol{v}\|_\infty \leq \varepsilon_n, \delta \in \Delta, q \in \mathcal{Q}\}$$

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 $\infty$-norm (like $\boldsymbol{\beta}$). Applying Lemma C.4 one more time, we have that there exist some constants $C_{13} \geq e$, $C_{14} \geq 1$ and $C_{15} > 0$ such that

$$\sup_{\mathbb{Q}} N\Big(\mathcal{G}, \|\cdot\|_{\mathbb{Q},2}, \varepsilon C_{15}\Big) \leq \left(\frac{C_{13}}{\varepsilon}\right)^{C_{14}\log^d n} \tag{H.1}$$

for all $0 < \varepsilon \leq 1$, where the supremum is taken over all finite discrete probability measures $\mathbb{Q}$ and $\mathcal{G}$ is the class defined in the modified Assumption B.2 in Section E.2. 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 $C_{16}\varepsilon_n$, where $C_{16}$ is a large enough constant.

For large enough $n$ we can replace $\varepsilon$ with $C_{16}\varepsilon\varepsilon_n/C_{15}$ in (H.1), giving

$$\sup_{\mathbb{Q}} N\Big(\mathcal{G}, \|\cdot\|_{\mathbb{Q},2}, C_{16}\varepsilon\varepsilon_n\Big) \leq \left(\frac{C_{13}C_{15}}{C_{16}\varepsilon\varepsilon_n}\right)^{C_{14}\log^d n}$$

for all $0 < \varepsilon \leq 1$. For large enough $n$, $\frac{C_{13}C_{15}}{C_{16}\varepsilon_n} \geq e$. The verification is complete.

**Verifying** (F.14) In this case, $\psi(y, \eta; q) = \mathbb{1}\{y < \eta\} - q$ and $\eta(\mu_0(\mathbf{x}_i, q))$ is the $q$-quantile of $y_i$ conditional on $\mathbf{x}_i$. Without loss of generality, we will assume that $\eta(\cdot)$ is strictly increasing and $q \leq \tilde{q}$ (the other cases are symmetric).

$$\mathbb{E}\left[\left|\psi(y_i, \eta(\mu_0(\mathbf{x}_i, q)); q)\eta^{(1)}(\mu_0(\mathbf{x}_i, q)) - \psi(y_i, \eta(\mu_0(\mathbf{x}_i, \tilde{q})); \tilde{q})\eta^{(1)}(\mu_0(\mathbf{x}_i, \tilde{q}))\right|^2 \,\middle|\, \mathbf{x}_i\right]$$

$$
\begin{aligned}
&= \mathbb{E}\Big[\psi(y_i, \eta(\mu_0(\mathbf{x}_i, q)); q)^2 \,\Big|\, \mathbf{x}_i\Big] \eta^{(1)}(\mu_0(\mathbf{x}_i, q))^2 \\
&\quad + \mathbb{E}\Big[\psi(y_i, \eta(\mu_0(\mathbf{x}_i, \tilde{q})); \tilde{q})^2 \,\Big|\, \mathbf{x}_i\Big] \eta^{(1)}(\mu_0(\mathbf{x}_i, \tilde{q}))^2 \\
&\quad - 2\mathbb{E}[\psi(y_i, \eta(\mu_0(\mathbf{x}_i, q)); q)\psi(y_i, \eta(\mu_0(\mathbf{x}_i, \tilde{q})); \tilde{q}) \mid \mathbf{x}_i] \eta^{(1)}(\mu_0(\mathbf{x}_i, q))\eta^{(1)}(\mu_0(\mathbf{x}_i, \tilde{q})) \\
&= (q - q^2)\eta^{(1)}(\mu_0(\mathbf{x}_i, q))^2 + (\tilde{q} - \tilde{q}^2)\eta^{(1)}(\mu_0(\mathbf{x}_i, \tilde{q}))^2 \\
&\quad - 2(q - q\tilde{q})\eta^{(1)}(\mu_0(\mathbf{x}_i, q))\eta^{(1)}(\mu_0(\mathbf{x}_i, \tilde{q})) \\
&= q\big|\eta^{(1)}(\mu_0(\mathbf{x}_i, q)) - \eta^{(1)}(\mu_0(\mathbf{x}_i, \tilde{q}))\big|^2 + (\tilde{q} - q)\eta^{(1)}(\mu_0(\mathbf{x}_i, \tilde{q}))^2 \\
&\quad - \big|q\eta^{(1)}(\mu_0(\mathbf{x}_i, q)) - \tilde{q}\eta^{(1)}(\mu_0(\mathbf{x}_i, \tilde{q}))\big|^2 \\
&\le q\big|\eta^{(1)}(\mu_0(\mathbf{x}_i, q)) - \eta^{(1)}(\mu_0(\mathbf{x}_i, \tilde{q}))\big|^2 + (\tilde{q} - q)\eta^{(1)}(\mu_0(\mathbf{x}_i, \tilde{q}))^2 \\
&\overset{\text{(a)}}{\lesssim} q(\tilde{q} - q)^2 + \tilde{q} - q \lesssim \tilde{q} - q,
\end{aligned}
$$

where in (a) we used that $\eta^{(1)}(\cdot)$ on a fixed compact is Lipschitz and $\mu_0(\mathbf{x}, q)$ is Lipschitz in $q$ uniformly over $\mathbf{x}$, as well as boundedness of $\mu_0(\mathbf{x}, q)$ uniformly over $q$ and $\mathbf{x}$.

This concludes the proof of Proposition H.1. □

**Proposition H.2** (Verification of the conditions of Lemma D.5)**.** *Suppose all conditions of Proposition H.1 hold. In addition, suppose there is a positive constant $C_{17}$ such that we have $\inf f_{Y|X}(y|\mathbf{x}) > C_{17}$, where the infimum is over $\mathbf{x} \in \mathcal{X}$, $q \in \mathcal{Q}$, $\|\boldsymbol{\beta}\|_\infty \le R$ for $R$ described in Lemma D.5, $y$ between $\eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta})$ and $\eta(\mu_0(\mathbf{x}, q))$. Then conditions in Conditions (v) and (vi) of Lemma D.5 also hold.*

*Proof.* We only need to verify Lemma D.5(vi) since Lemma D.5(v) is directly assumed in this lemma ($\Psi_1(\mathbf{x}, \eta; q) = f_{Y|X}(\eta|\mathbf{x})$ in this case).

In this verification, we will use $\theta_1 := \mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}$, $\theta_2 := \mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}_0(q)$ 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

$$
\{(\mathbf{x}, y) \mapsto y[\mathbb{1}\{y < \eta(\theta_2)\} - \mathbb{1}\{y < \eta(\theta_1)\}] : \|\boldsymbol{\beta}\|_\infty \le R, q \in \mathcal{Q}\}
$$

with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim K$.

The class $\{(\mathbf{x}, y) \mapsto q : q \in \mathcal{Q}\}$ is of course VC with a constant index (as a subclass of the class of constant functions), and the class $\{(\mathbf{x}, y) \mapsto \eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathbb{R}^K\}$ is VC with index $O(K)$ because the space of functions $\mathbf{p}(\mathbf{x})\boldsymbol{\beta}$ is a linear space with $O(K)$ dimension, and $\eta(\cdot)$ is monotone. Applying Lemma C.4, we see that the class $\{(\mathbf{x}, y) \mapsto q[\eta(\theta_1) - \eta(\theta_2)] : \|\boldsymbol{\beta}\|_\infty \le R, q \in \mathcal{Q}\}$ with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim K$.

The class $\{(\mathbf{x}, y) \mapsto \mathbb{1}\{y < \eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta})\} : \boldsymbol{\beta} \in \mathbb{R}^K\}$ is VC with index $O(K)$ because the space of functions $\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}$ is a linear space with $O(K)$ dimension, and $\eta(\cdot)$ is monotone. Again applying Lemma C.4, we see that the class $\{(\mathbf{x}, y) \mapsto \eta(\theta_1)\mathbb{1}\{y < \eta(\theta_1)\} : \|\boldsymbol{\beta}\|_\infty \le R\}$ 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 $\{(\mathbf{x}, y) \mapsto \eta(\theta_2)\mathbb{1}\{y < \eta(\theta_2)\} : q \in \mathcal{Q}\}$.

It is left to apply Lemma C.4 again, concluding that the class described in Lemma D.5(vi) with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim K$. □

**Proposition H.3** (Verification of plug-in approximation assumptions for quantile regression)**.** *Let the setting be as in Proposition H.1 with identity link function $\eta(a) \equiv a$.*

*(a) Suppose that Conditions (i) and (ii) of Theorem G.1 are true. Then so is Condition (iii).*

*(b) Suppose that Condition (i) of Theorem G.7 is true. Then so are Conditions (ii) and (iii).*

*Proof.* The argument for Assertion (a) can be found in [1]; we include it here for completeness. $\boldsymbol{Z}_n(q)$ is a mean-zero Gaussian conditionally on $\{\mathbf{x}_i\}_{i=1}^n$ process with a. s. continuous sample paths with covariance structure

$$\begin{aligned}
&\mathbb{E}[\boldsymbol{Z}_n(q)\boldsymbol{Z}_n(\tilde{q})^\mathsf{T} \mid \{\mathbf{x}_i\}_{i=1}^n] \\
&\quad = h^{-d}\mathbb{E}_n\Big[S_{q,\tilde{q}}(\mathbf{x}_i)\eta^{(1)}(\mu_0(\mathbf{x}_i, q))\eta^{(1)}(\mu_0(\mathbf{x}_i, \tilde{q}))\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\mathsf{T}\Big] \\
&\quad = (q \wedge \tilde{q} - q\tilde{q})h^{-d}\mathbb{E}_n[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\mathsf{T}],
\end{aligned}$$

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

This means that $\mathbf{B}_K(q) := h^{d/2}\mathbb{E}_n[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\mathsf{T}]^{-1/2}\boldsymbol{Z}_n(q)$ is, conditionally on $\{\mathbf{x}_i\}_{i=1}^n$, a $K$-dimensional Brownian bridge on $\mathcal{Q}$ (i. e., a $K$-dimensional vector of independent 1-dimensional Brownian bridges on $\mathcal{Q}$). Simulating any Brownian bridge $\mathbf{B}_K^*(\cdot)$ on $\mathcal{Q}$ with a. s. continuous sample paths independent of the data, we note that $h^{-d/2}\mathbb{E}_n[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\mathsf{T}]^{1/2}\mathbf{B}_K^*(\cdot)$ is a feasible copy (conditional on $\{\mathbf{x}_i\}_{i=1}^n$) of $\boldsymbol{Z}_n(\cdot)$, depending on the data only via $\{\mathbf{x}_i\}_{i=1}^n$, so the supremum in (G.1) is zero.

We will now prove Assertion (b). Take the unconditionally Gaussian process

$$\widetilde{\boldsymbol{Z}}_n(\cdot) := h^{-d/2}\mathbb{E}[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\mathsf{T}]^{1/2}\mathbf{B}_K(\cdot).$$

Then

$$\Big\|\widetilde{\boldsymbol{Z}}_n(q) - \boldsymbol{Z}_n(q)\Big\|_\infty = h^{-d/2}\Big\|\Big(\mathbb{E}_n[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\mathsf{T}]^{1/2} - \mathbb{E}[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\mathsf{T}]^{1/2}\Big)\mathbf{B}_K(q)\Big\|_\infty$$

By the same argument as for Lemma G.2 (applying the Gaussian maximal inequality), we can note

$$\begin{aligned}
&\mathbb{E}\Big[\Big\|\widetilde{\boldsymbol{Z}}_n(q) - \boldsymbol{Z}_n(q)\Big\|_\infty \Bigm| \{\mathbf{x}_i\}_{i=1}^n\Big] \\
&\quad \lesssim h^{-d/2}\Big\|\mathbb{E}_n[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\mathsf{T}]^{1/2} - \mathbb{E}[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\mathsf{T}]^{1/2}\Big\|\sqrt{\log K},
\end{aligned}$$

(where the constant in $\lesssim$ is the same everywhere except a negligible set). By Theorem X.3.8 in [2] and Lemma C.11

$$h^{-d/2}\Big\|\mathbb{E}_n[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\mathsf{T}]^{1/2} - \mathbb{E}[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\mathsf{T}]^{1/2}\Big\| \lesssim_\mathbb{P} \sqrt{\frac{\log(1/h)}{nh^d}},$$

so we can conclude

$$\mathbb{E}\Big[\Big\|\widetilde{\boldsymbol{Z}}_n(q) - \boldsymbol{Z}_n(q)\Big\|_\infty \Bigm| \{\mathbf{x}_i\}_{i=1}^n\Big] \lesssim_\mathbb{P} \frac{\log(1/h)}{\sqrt{nh^d}} \lesssim r_{\mathrm{str}},$$

verifying Condition (ii) of Theorem G.7.

Now take

$$\widetilde{\boldsymbol{Z}}_n^*(\cdot) := h^{-d/2}\mathbb{E}[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\mathsf{T}]^{1/2}\mathbf{B}_K^*(\cdot),$$

which is independent of the data. By the same argument (replacing $\widetilde{\boldsymbol{Z}}_n(\cdot)$ with $\widetilde{\boldsymbol{Z}}_n^*(\cdot)$ and $\boldsymbol{Z}_n(\cdot)$ with $\boldsymbol{Z}_n^*(\cdot)$), we have

$$\mathbb{E}\Big[\Big\|\widetilde{\boldsymbol{Z}}_n^*(q) - \boldsymbol{Z}_n^*(q)\Big\|_\infty \Bigm| \{\mathbf{x}_i\}_{i=1}^n\Big] \lesssim_\mathbb{P} \frac{\log(1/h)}{\sqrt{nh^d}} \lesssim r_{\mathrm{str}},$$

verifying Condition (iii) of Theorem G.7. □

## H.2 Distribution regression

**Proposition H.4** (Verification of assumptions for distribution regression)**.** *Consider the model* (A.1) *with a strictly monotonic two times continuously differentiable link function* $\eta(\cdot)\colon \mathbb{R} \to (0,1)$, $\mathcal{Q} = [-A, A]$ *for some* $A > 0$, *and* $\rho(y, \eta; q) = (\mathbb{1}\{y \le q\} - \eta)^2$. *Suppose the following conditions.*

*(i) Assumptions* B.1, B.3 *and* B.4 *hold.*

*(ii) The conditional distribution function* $\mathcal{X} \ni \mathbf{x} \mapsto F_{Y|X}(q|\mathbf{x})$, *as a function of* $\mathbf{x}$, *is a continuous function whose values lie in a compact subset of* $(0,1)$, *where this compact does not depend on* $q \in \mathcal{Q}$.

*(iii) As a function of* $q \in \mathcal{Q}$, $F_{Y|X}(q|\mathbf{x})$ *is continuously differentiable and* $f_{Y|X}(q|\mathbf{x})$ *is its derivative (in particular,* $\mathfrak{M}$ *is Lebesgue measure).*

*(iv)* $\eta(\mu_0(\mathbf{x}, q)) = F_{Y|X}(q|\mathbf{x}) = \mathbb{P}\{y_1 \le q \mid \mathbf{x}_1 = \mathbf{x}\}$.

*Then Assumption* B.2, *its modified version described in Section* E.2, (F.14), *and conditions in Conditions* (v) *and* (vi) *of Lemma* D.5 *are also true.*

*Remark.* Taking $\eta(\mu_0(\mathbf{x}, q))$ 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

$$
\begin{aligned}
&\mathbb{E}\big[\big(\mathbb{1}\{y_1 \le q\} - \eta(\mu(\mathbf{x}_1))\big)^2\big] \\
&\quad = \mathbb{E}\big[\big(\mathbb{1}\{y_1 \le q\} - F_{Y|X}(q|\mathbf{x}_1) + F_{Y|X}(q|\mathbf{x}_1) - \eta(\mu(\mathbf{x}_1))\big)^2\big] \\
&\quad = \mathbb{E}\big[\big(\mathbb{1}\{y_1 \le q\} - F_{Y|X}(q|\mathbf{x}_1)\big)^2\big] \\
&\qquad + 2\mathbb{E}\big[\big(\mathbb{1}\{y_1 \le q\} - F_{Y|X}(q|\mathbf{x}_1)\big)\big(F_{Y|X}(q|\mathbf{x}_1) - \eta(\mu(\mathbf{x}_1))\big)\big] \\
&\qquad + \mathbb{E}\big[\big(F_{Y|X}(q|\mathbf{x}_1) - \eta(\mu(\mathbf{x}_1))\big)^2\big].
\end{aligned}
$$

Since the cross term is zero (proven by conditioning on $\mathbf{x}_1$), this means

$$
\begin{aligned}
&\mathbb{E}\big[\big(\mathbb{1}\{y_1 \le q\} - \eta(\mu(\mathbf{x}_1))\big)^2\big] \\
&\quad = \mathbb{E}\big[\big(\mathbb{1}\{y_1 \le q\} - F_{Y|X}(q|\mathbf{x}_1)\big)^2\big] + \mathbb{E}\big[\big(F_{Y|X}(q|\mathbf{x}_1) - \eta(\mu(\mathbf{x}_1))\big)^2\big] \\
&\quad \ge \mathbb{E}\big[\big(\mathbb{1}\{y_1 \le q\} - F_{Y|X}(q|\mathbf{x}_1)\big)^2\big].
\end{aligned}
$$

Pointwise in $q \in \mathcal{Q}$, equality holds if and only if $\eta(\mu(\mathbf{x}_1)) = F_{Y|X}(q|\mathbf{x}_1)$ almost surely.

*Remark.* In this case,

$$
\begin{aligned}
&\mathbb{E}[\boldsymbol{Z}_n(q)\boldsymbol{Z}_n(\tilde{q})^\intercal \mid \{\mathbf{x}_i\}_{i=1}^n] \\
&\quad = 4h^{-d}\mathbb{E}_n\Big[F_{Y|X}(q \wedge \tilde{q}|\mathbf{x}_i)\big(1 - F_{Y|X}(q \vee \tilde{q}|\mathbf{x}_i)\big)\eta^{(1)}(\mu_0(\mathbf{x}_i, q))\eta^{(1)}(\mu_0(\mathbf{x}_i, \tilde{q}))\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\intercal\Big].
\end{aligned}
$$

This covariance structure is not known, but it can be estimated by

$$
4h^{-d}\mathbb{E}_n\Big[\eta(\widehat{\mu}_{0,q\wedge\tilde{q}}(\mathbf{x}_i))\big(1 - \eta(\widehat{\mu}_{0,q\vee\tilde{q}}(\mathbf{x}_i))\big)\eta^{(1)}(\widehat{\mu}_{0,q}(\mathbf{x}_i))\eta^{(1)}(\widehat{\mu}_{0,\tilde{q}}(\mathbf{x}_i))\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\intercal\Big] \tag{H.2}
$$

Condition (iii) in Theorem G.1 requires constructing a process $\widehat{\boldsymbol{Z}}_n^*(q)$ whose distribution is known. This is done by the discretization argument as in the proof of Theorem F.3:

- prove the existence of $\boldsymbol{Z}_n^*(q)$ by applying Lemma 3.7.3 in [8] with $V$ denoting $\boldsymbol{D}_n$;
- project $\boldsymbol{Z}_n^*(q)$ on a discrete $\delta$-net $\tilde{\mathcal{Q}}^\delta$ of $\mathcal{Q}$ and show that the process does not deviate too much from its projection, denoted $\boldsymbol{Z}_n^*(q|_{\tilde{\mathcal{Q}}^\delta})$;

- approximate the projection vector $\boldsymbol{Z}_n^*(q|_{\tilde{\mathcal{Q}}^\delta})$ by a different conditionally Gaussian vector with a known covariance matrix, using (H.2) (for $q$ and $\tilde{q}$ in the covering);
- use the same argument as in the proof of Theorem F.3 to extend the resulting conditionally Gaussian vector to a conditionally Gaussian process $\widehat{\boldsymbol{Z}}_n^*(q)$;
- finally, show that the resulting process does not deviate too much from its projection onto the covering (its "discretization").

*Proof.* We will verify the assumptions one by one.

**Verifying Assumption B.2(i)** $\rho(y, \eta; q)$ is indeed absolutely continuous, and even infinitely smooth, with respect to $\eta \in \mathcal{E} := (-\infty, 0)$. Its first derivative is

$$\psi(y, \eta; q) = 2(\eta - \mathbb{1}\{y \le q\}).$$

Since the derivative of $\eta(\cdot)$ is bounded on a compact interval, the function $\psi(y, \eta(\theta); q)$ is Lipschitz in $\theta$ on a compact interval, so we can take $\psi(y, \eta(\theta); q) = \varphi(y, \eta(\theta); q)$, and $\varpi(\cdot)$ is the identity function.

**Verifying Assumption B.2(ii)** The first-order optimality condition

$$\mathbb{E}[2(\eta(\mu_0(\mathbf{x}_i, q)) - \mathbb{1}\{y_i \le q\}) \mid \mathbf{x}_i] = 2\big(F_{Y|X}(q|\mathbf{x}_i) - \mathbb{E}[\mathbb{1}\{y_i \le q\} \mid \mathbf{x}_i]\big) = 0$$

indeed holds. The conditional variance

$$\sigma_q^2(\mathbf{x}) := 4\mathbb{E}\Big[(\eta(\mu_0(\mathbf{x}_i, q)) - \mathbb{1}\{y_i \le q\})^2 \,\Big|\, \mathbf{x}_i = \mathbf{x}\Big] = 4F_{Y|X}(q|\mathbf{x})\big(1 - F_{Y|X}(q|\mathbf{x})\big)$$

is continuous and bounded away from zero by the assumptions ($F_{Y|X}(q|\mathbf{x})$ cannot achieve 0 or 1).

The family $\big\{2\big(F_{Y|X}(q|\mathbf{x}) - \mathbb{1}\{y \le q\}\big) : q \in \mathcal{Q}\big\}$ is bounded in absolute value by $\overline{\psi}(\mathbf{x}, y) \equiv 2$.

**Verifying Assumption B.2(iii)** The conditional expectation

$$\Psi(\mathbf{x}, \eta; q) := \mathbb{E}[2(\eta - \mathbb{1}\{y_i \le q\}) \mid \mathbf{x}_i = \mathbf{x}] = 2\eta - 2F_{Y|X}(q|\mathbf{x})$$

is linear, and in particular infinitely smooth, in $\eta$. Its first partial derivative

$$\Psi_1(\mathbf{x}, \eta; q) = \frac{\partial}{\partial \eta}\Psi(\mathbf{x}, \eta; q) = 2$$

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.2(iv)** The class of functions

$$\mathcal{G}_{11} := \big\{(\mathbf{x}, y) \mapsto F_{Y|X}(q|\mathbf{x}) : q \in \mathcal{Q}\big\}$$

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

$$\mathcal{G}_{12} := \big\{(\mathbf{x}, y) \mapsto \mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^K\big\}$$

is VC with index not exceeding $K+2$ by Lemma 2.6.15 in [11]. Since in a fixed bounded interval $\eta(\cdot)$ is Lipschitz, the class

$$\mathcal{G}_{13} := \{(\mathbf{x}, y) \mapsto \eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}) : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r\}$$

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

$$\mathcal{G}_1 := \big\{(\mathbf{x}, y) \mapsto 2\big(\eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}) - F_{Y|X}(q|\mathbf{x})\big) : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, q \in \mathcal{Q}\big\},$$

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 $\{(\mathbf{x}, y) : y \le q\}$ 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

$$\mathcal{G}_{21} := \{(\mathbf{x}, y) \mapsto \mathbb{1}\{y \le q\} : q \in \mathcal{Q}\},$$

giving by Lemma C.4 that

$$\mathcal{G}_2 := \big\{(\mathbf{x}, y) \mapsto 2\big(F_{Y|X}(q|\mathbf{x}) - \mathbb{1}\{y \le q\}\big) : q \in \mathcal{Q}\big\},$$

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 $\dim \mathcal{B}$,

$$\mathcal{W}_{\mathcal{B}} := \{(\mathbf{x}, y) \mapsto \mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta} : \boldsymbol{\beta} \in \mathcal{B}\}$$

is VC-subgraph with index $O(\dim \mathcal{B})$ by Lemmas 2.6.15 and 2.6.18 in [11]. 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

$$\mathcal{W}_{3,\delta} := \big\{(\mathbf{x}, y) \mapsto \eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}) \mathbb{1}\big\{\mathbf{x} \in \mathcal{N}_{[c \log n]}(\delta)\big\} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, q \in \mathcal{Q}\big\}$$

with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim \log^d n$. Then the union of $O(h^{-d})$ such classes

$$\mathcal{W}_3 := \big\{(\mathbf{x}, y) \mapsto \eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}) \mathbb{1}\big\{\mathbf{x} \in \mathcal{N}_{[c \log n]}(\delta)\big\} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, q \in \mathcal{Q}, \delta \in \Delta\big\}$$

satisfies the uniform entropy bound (A.3) with $A \lesssim 1$ and $V \lesssim \log^d n$ (see (E.17)). The same is true of

$$\begin{aligned}\mathcal{G}_3 := \big\{&(\mathbf{x}, y) \mapsto 2(\eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}) - \eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}_0(q))) \mathbb{1}\big\{\mathbf{x} \in \mathcal{N}_{[c \log n]}(\delta)\big\} :\\ &\|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, q \in \mathcal{Q}, \delta \in \Delta\big\}\end{aligned}$$

because it is a subclass of $2\mathcal{W}_3 - 2\mathcal{W}_3$.

The class

$$\mathcal{G}_4 := \{\mathbf{x} \mapsto 2\}$$

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 $l \in \{1, \ldots, K\}$ the class

$$\mathcal{G}_{51} := \{(\mathbf{x}, y) \mapsto p_l(\mathbf{x})\eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}_0(q)) : q \in \mathcal{Q}\}$$

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

$$\mathcal{G}_5 := \big\{(\mathbf{x}, y) \mapsto 2p_l(\mathbf{x})\big(F_{Y|X}(q|\mathbf{x}) - \eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}_0(q))\big) : q \in \mathcal{Q}\big\}$$

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

**Verifying modified Assumption B.2** The replacement of Assumption B.2(i) holds trivially: see the argument in Section H.2.

The addition to Assumption B.2(iv) holds as well because the class of functions described there

$$\begin{aligned}
\big\{(\mathbf{x}, y) \mapsto\ & [\eta(\mathbf{p}(\mathbf{x})^\intercal(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta})) - \eta(\mathbf{p}(\mathbf{x})^\intercal(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta} - \boldsymbol{v}))]\\
&\times [\eta(\mathbf{p}(\mathbf{x})^\intercal(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta})) + \eta(\mathbf{p}(\mathbf{x})^\intercal(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta} - \boldsymbol{v})) - 2\eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}_0(q))]\\
&\times \mathbb{1}\big(\mathbf{x} \in \mathcal{N}_{[c \log n]}(\delta)\big) :\\
&\|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, \|\boldsymbol{v}\|_\infty \le \varepsilon_n, \delta \in \Delta, q \in \mathcal{Q}\big\}
\end{aligned}$$

is contained in the product of two classes

$$\begin{aligned}
\mathcal{V}_1 := \big\{&(\mathbf{x}, y) \mapsto [\eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}) - \eta(\mathbf{p}(\mathbf{x})^\intercal(\boldsymbol{\beta} - \boldsymbol{v}))]\mathbb{1}\big(\mathbf{x} \in \mathcal{N}_{[c \log n]}(\delta)\big) :\\
&\|\boldsymbol{\beta}\|_\infty \le \tilde{r}, \|\boldsymbol{v}\|_\infty \le \varepsilon_n, \delta \in \Delta, q \in \mathcal{Q}\big\},\\
\mathcal{V}_2 := \big\{&(\mathbf{x}, y) \mapsto [\eta(\mathbf{p}(\mathbf{x})^\intercal(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta})) + \eta(\mathbf{p}(\mathbf{x})^\intercal(\boldsymbol{\beta}_0(q) + \boldsymbol{\beta} - \boldsymbol{v})) - 2\eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}_0(q))]\\
&\times \mathbb{1}\big(\mathbf{x} \in \mathcal{N}_{[c \log n]}(\delta)\big) : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, \|\boldsymbol{v}\|_\infty \le \varepsilon_n, \delta \in \Delta, q \in \mathcal{Q}\big\}
\end{aligned}$$

for some fixed $\tilde{r} > 0$. $\mathcal{V}_1$ with envelope $\varepsilon_n$ multiplied by a large enough constant (since $\eta$ 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 $\{\eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta})\}$ and $\{\eta(\mathbf{p}(\mathbf{x})^\intercal(\boldsymbol{\beta} - \boldsymbol{v}))\}$ with constant envelopes, using Lemma C.4 and then replacing $\varepsilon$ in the uniform entropy bound by $\varepsilon \cdot \varepsilon_n$). $\mathcal{V}_2$ 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.2(iv).

**Verifying Condition (v) in Lemma D.5** This condition holds trivially because $\Psi_1(\cdot, \cdot; q)$ is a positive constant.

**Verifying Condition (vi) in Lemma D.5** The class of functions described in this condition is

$$\begin{aligned}
\{(\mathbf{x}, y) \mapsto\ & (\eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}_0(q)) - \eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta})) \cdot (2\mathbb{1}\{y \le q\} - \eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}) - \eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}_0(q))) :\\
&\|\boldsymbol{\beta}\|_\infty \le R, q \in \mathcal{Q}\}.
\end{aligned}$$

The assertion follows by Lemma C.4 since

$$\{(\mathbf{x}, y) \mapsto \mathbb{1}\{y \le q\} : q \in \mathcal{Q}\}$$

is a VC-subgraph class with a constant index and

$$\{(\mathbf{x}, y) \mapsto \eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}) : \|\boldsymbol{\beta}\|_\infty \le R\}$$

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.14) Without loss of generality, assume $q \le \tilde{q}$ (the other case is symmetric).

$$
\begin{aligned}
&\mathbb{E}\Big[\Big|\psi(y_i,\eta(\mu_0(\mathbf{x}_i,q));q)\eta^{(1)}(\mu_0(\mathbf{x}_i,q)) - \psi(y_i,\eta(\mu_0(\mathbf{x}_i,\tilde{q}));\tilde{q})\eta^{(1)}(\mu_0(\mathbf{x}_i,\tilde{q}))\Big|^2 \,\Big|\, \mathbf{x}_i\Big] \\
&\quad = \mathbb{E}\Big[\psi(y_i,\eta(\mu_0(\mathbf{x}_i,q));q)^2 \,\Big|\, \mathbf{x}_i\Big]\eta^{(1)}(\mu_0(\mathbf{x}_i,q))^2 \\
&\qquad + \mathbb{E}\Big[\psi(y_i,\eta(\mu_0(\mathbf{x}_i,\tilde{q}));\tilde{q})^2 \,\Big|\, \mathbf{x}_i\Big]\eta^{(1)}(\mu_0(\mathbf{x}_i,\tilde{q}))^2 \\
&\qquad - 2\mathbb{E}[\psi(y_i,\eta(\mu_0(\mathbf{x}_i,q));q)\psi(y_i,\eta(\mu_0(\mathbf{x}_i,\tilde{q}));\tilde{q}) \mid \mathbf{x}_i]\eta^{(1)}(\mu_0(\mathbf{x}_i,q))\eta^{(1)}(\mu_0(\mathbf{x}_i,\tilde{q})) \\
&\quad = 4\big[F_{Y|X}(q|\mathbf{x}_i) - F_{Y|X}(q|\mathbf{x}_i)^2\big]\eta^{(1)}(\mu_0(\mathbf{x}_i,q))^2 \\
&\qquad + 4\big[F_{Y|X}(\tilde{q}|\mathbf{x}_i) - F_{Y|X}(\tilde{q}|\mathbf{x}_i)^2\big]\eta^{(1)}(\mu_0(\mathbf{x}_i,\tilde{q}))^2 \\
&\qquad - 8\big[F_{Y|X}(q|\mathbf{x}_i) - F_{Y|X}(q|\mathbf{x}_i)F_{Y|X}(\tilde{q}|\mathbf{x}_i)\big]\eta^{(1)}(\mu_0(\mathbf{x}_i,q))\eta^{(1)}(\mu_0(\mathbf{x}_i,\tilde{q})) \\
&\quad = 4F_{Y|X}(q|\mathbf{x}_i)\big|\eta^{(1)}(\mu_0(\mathbf{x}_i,q)) - \eta^{(1)}(\mu_0(\mathbf{x}_i,\tilde{q}))\big|^2 \\
&\qquad + 4\big[F_{Y|X}(\tilde{q}|\mathbf{x}_i) - F_{Y|X}(q|\mathbf{x}_i)\big]\eta^{(1)}(\mu_0(\mathbf{x}_i,\tilde{q}))^2 \\
&\qquad - 4\big|F_{Y|X}(q|\mathbf{x}_i)\eta^{(1)}(\mu_0(\mathbf{x}_i,q)) - F_{Y|X}(\tilde{q}|\mathbf{x}_i)\eta^{(1)}(\mu_0(\mathbf{x}_i,\tilde{q}))\big|^2 \\
&\quad \le 4F_{Y|X}(q|\mathbf{x}_i)\big|\eta^{(1)}(\mu_0(\mathbf{x}_i,q)) - \eta^{(1)}(\mu_0(\mathbf{x}_i,\tilde{q}))\big|^2 \\
&\qquad + 4\big[F_{Y|X}(\tilde{q}|\mathbf{x}_i) - F_{Y|X}(q|\mathbf{x}_i)\big]\eta^{(1)}(\mu_0(\mathbf{x}_i,\tilde{q}))^2 \\
&\quad \overset{\text{(a)}}{\lesssim} (\tilde{q}-q)^2 + (\tilde{q}-q) \lesssim \tilde{q}-q,
\end{aligned}
$$

where in (a) we used that $\eta^{(1)}(\cdot)$ and $\eta(\cdot)$ on a fixed compact are Lipschitz and $\mu_0(\mathbf{x},q)$ is Lipschitz in $q$ uniformly over $\mathbf{x}$ (therefore, $\eta(\mu_0(\mathbf{x},q)) = F_{Y|X}(q|\mathbf{x})$ is also Lipschitz in $q$ uniformly over $\mathbf{x}$), as well as boundedness of $\mu_0(\mathbf{x},q)$ uniformly over $q$ and $\mathbf{x}$.

Proposition H.4 is proven. □

## H.3 $L_p$ regression

**Proposition H.5** (Verification of Assumption B.2 for $L_p$ regression)**.** *Suppose Assumptions B.1, B.3, B.4 hold with $\mathcal{Q}$ a singleton, $\rho(y,\eta) = |y-\eta|^p$, where $1 < p \le 2$, $\mu_0(\cdot)$ 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 two times continuously differentiable, where $\mathcal{E}$ is 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 $\int_{\mathbb{R}} \psi(y,a_l) f_{Y|X}(y|\mathbf{x})\,\mathrm{d}y < 0$ if $a_l$ is finite, and $\int_{\mathbb{R}} \psi(y,a_r) f_{Y|X}(y|\mathbf{x})\,\mathrm{d}y > 0$ if $a_r$ is finite.*

*Also assume that $|y_1|^{\nu(p-1)}$ is integrable for some $\nu > 2$ and $\mathbf{x} \mapsto f_{Y|X}(y|\mathbf{x})$ is continuous for any $y \in \mathcal{Y}$. Moreover, for any $\mathbf{x}$ the conditional density $f_{Y|X}(y|\mathbf{x})$ is a member of the Schwartz space $\mathcal{S}(\mathbb{R})$; the function $\int_{\mathbb{R}} |\eta(\zeta) - y|^{p-1} \operatorname{sign}(\eta(\zeta)-y) f'_{Y|X}(y|\mathbf{x})\,\mathrm{d}y$ is bounded and bounded away from zero uniformly over $\mathbf{x} \in \mathcal{X}$, $\zeta \in B(\mathbf{x})$ with $B(\mathbf{x})$ defined in* (B.1)*; the function $\int_{\mathbb{R}} |\eta(\zeta) - y|^{p-1} \operatorname{sign}(\eta(\zeta)-y) f''_{Y|X}(y|\mathbf{x})\,\mathrm{d}y$ is bounded in absolute value uniformly over $\mathbf{x} \in \mathcal{X}$, $\zeta \in B(\mathbf{x})$. Then Assumption B.2, its stronger modified version described in Section E.2 and* (F.14) *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

$$\zeta \mapsto \int_{\mathbb{R}} \psi(y,\zeta) f_{Y|X}(y|\mathbf{x})\,\mathrm{d}y$$

is continuous is proven below in the verification of Assumption B.2(i). To ensure that it crosses zero if $a_l$ or $a_r$ is not finite, we show

$$\int_{\mathbb{R}} \psi(y,\zeta) f_{Y|X}(y|\mathbf{x})\,\mathrm{d}y \to -\infty \text{ as } \zeta \to -\infty, \tag{H.3}$$

$$\int_{\mathbb{R}} \psi(y,\zeta) f_{Y|X}(y|\mathbf{x})\,\mathrm{d}y \to +\infty \text{ as } \zeta \to +\infty. \tag{H.4}$$

To prove (H.3), recall that $\psi(y,\zeta) = p|y-\zeta|^{p-1}\operatorname{sign}(\zeta - y)$ and therefore

$$\begin{aligned}
\int_{\mathbb{R}} \psi(y,\zeta) f_{Y|X}(y|\mathbf{x})\,\mathrm{d}y &= -p\int_{\zeta}^{+\infty} (y-\zeta)^{p-1} f_{Y|X}(y|\mathbf{x})\,\mathrm{d}y + p\int_{-\infty}^{\zeta} (\zeta - y)^{p-1} f_{Y|X}(y|\mathbf{x})\,\mathrm{d}y \\
&= -p(-\zeta)^{p-1} \underbrace{\int_{\mathbb{R}} \left(1 + \frac{y}{-\zeta}\right)^{p-1} \mathbb{1}\{y > \zeta\} f_{Y|X}(y|\mathbf{x})\,\mathrm{d}y}_{\to 1} \\
&\quad + p \underbrace{\int_{\mathbb{R}} (\zeta - y)^{p-1} \mathbb{1}\{y \le \zeta\} f_{Y|X}(y|\mathbf{x})\,\mathrm{d}y}_{\to 0} \to -\infty,
\end{aligned}$$

where we used dominated convergence because for $-\zeta > 1$ we have $1 + \frac{y}{-\zeta} \le 1 + |y|$ in the first integral and $\zeta - y \le -y = |y|$ in the second integral. (H.4) is proven similarly.

**Verifying Assumption B.2(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 $\eta$ on bounded intervals.

The function $x \mapsto |x|^{\alpha}\operatorname{sign}(x)$ for $\alpha \in (0,1]$ is $\alpha$-Hölder for $x \in \mathbb{R}$ (with constant 2). Therefore, putting $\alpha := p-1$, for any pair of reals $\zeta_1$ and $\zeta_2$ in a fixed bounded interval we have

$$\begin{aligned}
&\sup_{\mathbf{x}} \sup_{\lambda \in [0,1]} \sup_{y} |\psi(y, \eta(\zeta_1 + \lambda(\zeta_2 - \zeta_1))) - \psi(y, \eta(\zeta_2))| \\
&\le 2p|\eta(\zeta_1 + \lambda(\zeta_2 - \zeta_1)) - \eta(\zeta_2)|^{p-1} \overset{\text{(a)}}{\lesssim} |\zeta_1 - \zeta_2|^{p-1},
\end{aligned}$$

where in (a) we used that the link function $\eta(\cdot)$ in a fixed bounded interval is Lipschitz.

We can take $\varphi(y,\eta) \equiv \psi(y,\eta)$ and $\varpi(\theta) \equiv 1$.

**Verifying Assumption B.2(ii)** The first-order optimality condition is true because $\mu_0(\cdot)$ is defined this way in (C.1).

The function

$$\sigma^2(\mathbf{x}) := \mathbb{E}\Big[\psi(y_i, \mu_0(\mathbf{x}))^2 \,\Big|\, \mathbf{x}_i = \mathbf{x}\Big] = p^2 \int_{\mathbb{R}} |y - \mu_0(\mathbf{x})|^{2p-2} f_{Y|X}(y|\mathbf{x})\,\mathrm{d}y$$

is continuous because $\mathbf{x} \mapsto |y - \mu_0(\mathbf{x})|^{2p-2} f_{Y|X}(y|\mathbf{x})$ 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, $\sigma^2(\mathbf{x})$ is bounded away from zero because it is non-zero since $y_1$ has a conditional density.

The family of functions $\Big\{p|y - \eta(\mu_0(\mathbf{x}))|^{p-1}\operatorname{sign}(\eta(\mu_0(\mathbf{x})) - y)\Big\}$ only contains one element. Note that $\mathbf{x} \mapsto |y - \eta(\mu_0(\mathbf{x}))|^{\nu(p-1)} f_{Y|X}(y|\mathbf{x})$ is continuous and dominated by $(|y|^{\nu(p-1)} + C)C'$ for large enough constants $C$ and $C'$. Therefore, the function

$$\mathbf{x} \mapsto \int_{\mathbb{R}} |y - \eta(\mu_0(\mathbf{x}))|^{\nu(p-1)} f_{Y|X}(y|\mathbf{x})\,\mathrm{d}y$$

is also continuous. As a continuous function on a compact set, it is bounded.

**Verifying Assumption B.2(iii)** For each $\mathbf{x}$, the function

$$\eta \mapsto \int_{\mathbb{R}} |\eta - y|^{p-1} \operatorname{sign}(\eta - y) f_{Y|X}(y|\mathbf{x}) \, \mathrm{d}y$$

is a convolution of two (locally integrable) functions $|y|^{p-1} \operatorname{sign}(y)$ and $f_{Y|X}(y|\mathbf{x})$. The first one grows no faster than a polynomial, and therefore defines a tempered distribution, i.e. it can be considered a generalized function in $\mathcal{S}'(\mathbb{R})$. Since $f_{Y|X}(\cdot|\mathbf{x})$ lies in the Schwartz space $\mathcal{S}(\mathbb{R})$, it is well-known that this convolution is infinitely differentiable, and its derivative of any order is a convolution of $|y|^{p-1} \operatorname{sign}(y)$ and the corresponding derivative of $f_{Y|X}(y|\mathbf{x})$. The other conditions in Assumption B.2(iii) are directly assumed in the statement of Proposition H.5.

**Verifying Assumption B.2(iv)** 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 $\left\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\|_\infty \le r$. Note that

$$\begin{aligned}
&\left|[\psi(y, \eta(\mathbf{p}(\mathbf{x})^\mathsf{T}\boldsymbol{\beta})) - \psi(y, \eta(\mu_0(\mathbf{x})))] - \left[\psi\Big(y, \eta\Big(\mathbf{p}(\mathbf{x})^\mathsf{T}\widetilde{\boldsymbol{\beta}}\Big)\Big) - \psi(y, \eta(\mu_0(\mathbf{x})))\right]\right| \\
&\quad \le 2p\left|\eta(\mathbf{p}(\mathbf{x})^\mathsf{T}\boldsymbol{\beta}) - \eta\Big(\mathbf{p}(\mathbf{x})^\mathsf{T}\widetilde{\boldsymbol{\beta}}\Big)\right|^{p-1} \lesssim \left\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right\|_\infty^{p-1}.
\end{aligned}$$

The result for $\mathcal{G}_1$ follows.

Similarly,

$$\begin{aligned}
&\left|[\psi(y, \eta(\mathbf{p}(\mathbf{x})^\mathsf{T}\boldsymbol{\beta})) - \psi(y, \eta(\mathbf{p}(\mathbf{x})^\mathsf{T}\boldsymbol{\beta}_0))] - \left[\psi\Big(y, \eta\Big(\mathbf{p}(\mathbf{x})^\mathsf{T}\widetilde{\boldsymbol{\beta}}\Big)\Big) - \psi(y, \eta(\mathbf{p}(\mathbf{x})^\mathsf{T}\boldsymbol{\beta}_0))\right]\right| \\
&\quad \le 2p\left|\eta(\mathbf{p}(\mathbf{x})^\mathsf{T}\boldsymbol{\beta}) - \eta\Big(\mathbf{p}(\mathbf{x})^\mathsf{T}\widetilde{\boldsymbol{\beta}}\Big)\right|^{p-1} \lesssim \left\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right\|_\infty^{p-1}.
\end{aligned} \tag{H.5}$$

For a fixed cell $\delta \in \Delta$ the class of functions of the form

$$[\psi(y, \eta(\mathbf{p}(\mathbf{x})^\mathsf{T}\boldsymbol{\beta}); q) - \psi(y, \eta(\mathbf{p}(\mathbf{x})^\mathsf{T}\boldsymbol{\beta}_0); q)]\mathbb{1}\big(\mathbf{x} \in \mathcal{N}_{[c \log n]}(\delta)\big)$$

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.17) similarly to the proof of Proposition H.7).

**Verifying the addition to Assumption B.2(iv) described in Section E.2** Fix $\delta \in \Delta$. Let $\boldsymbol{\beta}$ and $\widetilde{\boldsymbol{\beta}}$ be such that $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_\infty \le r$ and $\left\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\|_\infty \le r$; 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

$$\begin{aligned}
g(t) &:= [\psi(y, \eta(\mathbf{p}(\mathbf{x})^\mathsf{T}(\boldsymbol{\beta}_0 + \boldsymbol{\beta}) + t)) - \psi(y, \eta(\mathbf{p}(\mathbf{x})^\mathsf{T}\boldsymbol{\beta}_0))] \\
&\quad \times \eta^{(1)}(\mathbf{p}(\mathbf{x})^\mathsf{T}(\boldsymbol{\beta}_0 + \boldsymbol{\beta}) + t), \\
\tilde{g}(t) &:= \left[\psi\Big(y, \eta\Big(\mathbf{p}(\mathbf{x})^\mathsf{T}\Big(\boldsymbol{\beta}_0 + \widetilde{\boldsymbol{\beta}}\Big) + t\Big)\Big) - \psi(y, \eta(\mathbf{p}(\mathbf{x})^\mathsf{T}\boldsymbol{\beta}_0))\right] \\
&\quad \times \eta^{(1)}\Big(\mathbf{p}(\mathbf{x})^\mathsf{T}\Big(\boldsymbol{\beta}_0 + \widetilde{\boldsymbol{\beta}}\Big) + t\Big).
\end{aligned}$$

Note that

$$\int_{-\mathbf{p}(\mathbf{x})^\mathsf{T}\boldsymbol{v}}^0 g(t)\,\mathrm{d}t - \int_{-\mathbf{p}(\mathbf{x})^\mathsf{T}\widetilde{\boldsymbol{v}}}^0 \tilde{g}(t)\,\mathrm{d}t = \int_{-\mathbf{p}(\mathbf{x})^\mathsf{T}\boldsymbol{v}}^0 (g(t) - \tilde{g}(t))\,\mathrm{d}t + \int_{-\mathbf{p}(\mathbf{x})^\mathsf{T}\boldsymbol{v}}^{-\mathbf{p}(\mathbf{x})^\mathsf{T}\widetilde{\boldsymbol{v}}} \tilde{g}(t)\,\mathrm{d}t.$$

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 $\mathbf{x}$ in these integrals

$$|g(t)-\tilde{g}(t)| \lesssim \left\|\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right\|_\infty^{p-1} + \left\|\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right\|_\infty,$$

and $|\tilde{g}(t)|$ is bounded. This gives

$$\left|\int_{-\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{v}}^{0} g(t)\,\mathrm{d}t - \int_{-\mathbf{p}(\mathbf{x})^\intercal \widetilde{\boldsymbol{v}}}^{0} \tilde{g}(t)\,\mathrm{d}t\right| \lesssim \varepsilon_n\left(\left\|\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right\|_\infty^{p-1} + \left\|\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right\|_\infty\right) + \|\boldsymbol{v}-\widetilde{\boldsymbol{v}}\|_\infty$$

It means that taking an $\varepsilon$-net (for $\varepsilon$ smaller than 1) in the space of $\boldsymbol{\beta}$ and an $\varepsilon_n\varepsilon$-net in the space of $\boldsymbol{v}$ induces an $C_{18}\varepsilon_n\varepsilon$-net in the space of functions

$$\left\{(\mathbf{x},y)\mapsto \int_{-\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{v}}^{0} g(t)\,\mathrm{d}t \cdot \mathbb{1}\big(\mathbf{x}\in\mathcal{N}_{[c\log n]}(\delta)\big) : \|\boldsymbol{\beta}-\boldsymbol{\beta}_0(q)\|_\infty \le r, \|\boldsymbol{v}\|_\infty \le \varepsilon_n\right\}$$

in terms of the sup-norm, where $C_{18}$ is some constant. Possibly increasing $C_{18}$, we can conclude that this class with envelope $C_{18}$ 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.17), the same can be said about the union of $O\big(h^{-d}\big)$ such classes (corresponding to different $\delta$). The verification is concluded.

**Verifying** (F.14) This is obvious because $\mathcal{Q}$ is a singleton.

**Verifying Lemma D.5(vi)** It follows from the proof of Lemma D.4 that

$$\begin{aligned}\left|\rho(y,\eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta})) - \rho\Big(y,\eta(\mathbf{p}(\mathbf{x})^\intercal\widetilde{\boldsymbol{\beta}})\Big)\right| &\lesssim \big(1+\overline{\psi}(\mathbf{x},y)\big)\left|\mathbf{p}(\mathbf{x})^\intercal(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}})\right| \\ &\lesssim \big(1+\overline{\psi}(\mathbf{x},y)\big)\left\|\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right\|_\infty.\end{aligned}$$

The required uniform entropy bound follows immediately from this.

Proposition H.5 is proven. □

**Proposition H.6** (Verification of strong approximation assumptions)**.** *Suppose the setting of Proposition H.5 applies, and the probability space is rich enough.*

*(a) If Conditions (i) and (ii) of Theorem G.1 are true, then so is Condition (iii), as long as there is an estimator $\widehat{\boldsymbol{\Sigma}}$ (known measurable function of the data) such that*

$$\left\|\widehat{\boldsymbol{\Sigma}}-\bar{\boldsymbol{\Sigma}}\right\| = o\left(\frac{h^d r_{\mathrm{str}}}{\sqrt{\log K}}\right).$$

*(b) If Condition (i) of Theorem G.7 is true, then so are Conditions (ii) and (iii).*

*Proof.* Since $\mathcal{Q}$ is a singleton, this is proven by exactly the same argument as in Lemma H.10. □

## H.4 Logistic regression

**Proposition H.7** (Verification of Assumption B.2 and others for logistic regression)**.** *Suppose Assumptions B.1, B.3, B.4 hold with $\mathcal{Q}$ a singleton, $\mathcal{Y}=\{0,1\}$, $\eta(\theta)=1/\big(1+e^{-\theta}\big)$, $\mathfrak{M}$ is the counting measure on $\{0,1\}$, and $\rho(y,\eta)=-y\log(\eta)-(1-y)\log(1-\eta)$. Assume also $\pi(\mathbf{x}) := \mathbb{P}\{y_i=1 \mid \mathbf{x}_i=\mathbf{x}\}$ is continuous and $\pi(\mathbf{x})$ lies in the interval $(0,1)$ for any $\mathbf{x}\in\mathcal{X}$. Then Assumption B.2, its stronger modified version described in Section E.2, and* (F.14) *are also true. Moreover, the conditions in Conditions (v) and (vi) of Lemma D.5 are true.*

We will prove Proposition H.7 now. Since $\mathcal{Q}$ is a singleton, we will omit the index $q$ in notations.

**Verifying Assumption B.2(i) and its stronger version described in Section E.2** 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

$$\frac{\partial}{\partial\theta}\rho(y,\eta(\theta)) = \psi(y,\eta(\theta))\eta^{(1)}(\theta) = (1-y)\eta(\theta) - y(1-\eta(\theta)) = \eta(\theta) - y,$$
$$\frac{\partial^2}{\partial\theta^2}\rho(y,\eta(\theta)) = \eta^{(1)}(\theta) = \eta(\theta)(1-\eta(\theta)).$$

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 $\theta$ for any $y$.

The following decomposition holds: $\psi(y,\eta(\theta)) = \varphi(y,\eta(\theta))\varpi(\theta)$, where $\varphi(y,\eta) = \eta - y$, $\varpi(\theta) = 1/\eta^{(1)}(\theta)$. Uniformly over $\zeta_1$ and $\zeta_2$ in a fixed bounded interval, we have

$$\sup_y |\varphi(y,\eta(\zeta_1)) - \varphi(y,\eta(\zeta_2))| = |\eta(\zeta_1) - \eta(\zeta_2)| \overset{\text{(a)}}{\leq} |\zeta_1 - \zeta_2|,$$

where in (a) we used that the derivative of $\eta(\cdot)$ does not exceed 1. So in this case the Hölder parameter $\alpha = 1$.

$\varpi(\cdot)$ is infinitely smooth and strictly positive on $\mathbb{R}$. The logistic function $\eta(\cdot)$ is strictly monotonic and infinitely smooth on $\mathbb{R}$.

**Verifying Assumption B.2(ii)** We have $\mathbb{E}[\psi(y_i,\eta(\mu_0(\mathbf{x}_i))) \mid \mathbf{x}_i] = 0$ since

$$\eta(\mu_0(\mathbf{x}_i)) = \mathbb{E}[y_i \mid \mathbf{x}_i] = \mathbb{P}\{y_i = 1 \mid \mathbf{x}_i\} =: \pi(\mathbf{x}_i).$$

Next,

$$\sigma^2(\mathbf{x}) = \mathbb{E}\Big[\psi(y_i,\eta(\mu_0(\mathbf{x}_i)))^2 \,\Big|\, \mathbf{x}_i = \mathbf{x}\Big] = \frac{\mathbb{E}\Big[(y_i - \pi(\mathbf{x}_i))^2 \,\Big|\, \mathbf{x}_i = \mathbf{x}\Big]}{\eta^{(1)}(\mu_0(\mathbf{x}))^2}$$
$$= \frac{\pi(\mathbf{x})(1-\pi(\mathbf{x}))}{\pi(\mathbf{x})^2(1-\pi(\mathbf{x}))^2} = \frac{1}{\pi(\mathbf{x})(1-\pi(\mathbf{x}))}$$

is continuous and bounded away from zero (is not less than 4).

Since $\mathbf{x}_i$ lies in a compact set, $\psi(y,\eta(\mu_0(\mathbf{x}_i)))$ is bounded, so it has moments of any order.

**Verifying Assumption B.2(iii)** In this case

$$\Psi(\mathbf{x};\eta) = \frac{\eta - \mathbb{E}[y_i \mid \mathbf{x}_i = \mathbf{x}]}{\eta(1-\eta)} = \frac{\eta - \pi(\mathbf{x})}{\eta(1-\eta)}.$$

This function is infinitely smooth in $\eta$ for $\eta \in (0,1)$. Its first derivative

$$\Psi_1(\mathbf{x},\eta) = \frac{\partial}{\partial\eta}\Psi(\mathbf{x};\eta) = \frac{\eta^2 - 2\eta\pi(\mathbf{x}) + \pi(\mathbf{x})}{\eta^2(1-\eta)^2}.$$

Therefore,

$$\Psi_1(\mathbf{x},\eta(\zeta))\eta^{(1)}(\zeta)^2 = \eta(\zeta)^2 - 2\eta(\zeta)\pi(\mathbf{x}) + \pi(\mathbf{x}). \tag{H.6}$$

If $|\zeta - \mu_0(\mathbf{x})| \le r$, then $|\eta(\zeta) - \eta(\mu_0(\mathbf{x}))| = |\eta(\zeta) - \pi(\mathbf{x})| \le r$ (since the derivative of $\eta(\cdot)$ does not exceed 1). Since $0 < \min_{\mathbf{x}\in\mathcal{X}} \pi(\mathbf{x})(1-\pi(\mathbf{x})) \le \max_{\mathbf{x}\in\mathcal{X}} \pi(\mathbf{x})(1-\pi(\mathbf{x})) < 1$, for small enough $r$ the right-hand side of (H.6) for such $\zeta$ is also bounded away from zero and one.

Finally,

$$\Psi_2(\mathbf{x},\eta) = \frac{\partial}{\partial\eta}\Psi_1(\mathbf{x},\eta) = \frac{2\big(\eta^3 - 3\pi(\mathbf{x})\eta^2 + 3\pi(\mathbf{x})\eta - \pi(\mathbf{x})\big)}{\eta^3(1-\eta)^3}.$$

Again since $0 < \min_{\mathbf{x}\in\mathcal{X}} \pi(\mathbf{x})(1-\pi(\mathbf{x})) \le \max_{\mathbf{x}\in\mathcal{X}} \pi(\mathbf{x})(1-\pi(\mathbf{x})) < 1$, for such $\zeta$ that $|\zeta - \mu_0(\mathbf{x})| \le r$ and $r$ small enough, the product $\eta(\zeta)(1-\eta(\zeta))$ is bounded away from zero. So for such $\zeta$, $|\Psi_2(\mathbf{x},\eta(\zeta))|$ is uniformly bounded.

**Lemma H.8** (Class $G_1$)**.** *The class*

$$\mathcal{G}_1 = \left\{\mathcal{X}\times\mathbb{R} \ni (\mathbf{x},y) \mapsto \frac{\eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}) - y}{\eta^{(1)}(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta})} - \frac{\pi(\mathbf{x}) - y}{\pi(\mathbf{x})(1-\pi(\mathbf{x}))} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le r\right\}$$

*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.8.* The class

$$\mathcal{G}_{11} = \big\{(\mathbf{x},y) \mapsto \mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^K\big\}$$

is VC with index not exceeding $K+2$ by Lemma 2.6.15 in [11]. Since in a fixed bounded interval $\eta(\cdot)$ is Lipschitz and $(\mathbf{x},y) \mapsto y$ is one fixed function, the class

$$\mathcal{G}_{12} = \{(\mathbf{x},y) \mapsto \eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}) - y : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_\infty \le r\}$$

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

$$\mathcal{G}_{13} = \Big\{(\mathbf{x},y) \mapsto \eta^{(1)}(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta})^{-1} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_\infty \le r\Big\},$$

where we used again that under these constraints $\eta^{(1)}(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta})$ is bounded away from zero. This implies by Lemma C.4 that it is true of $\mathcal{G}_{12}\cdot\mathcal{G}_{13} - \psi(y,\eta(\mathbf{x}))$ (since $\psi(y,\eta(\mathbf{x}))$ is one fixed function), which is what we need. ☐

**Lemma H.9** (Class $\mathcal{G}_3$)**.** *The class*

$$\begin{aligned}\mathcal{G}_3 := \Big\{&\mathcal{X}\times\mathbb{R} \ni (\mathbf{x},y) \mapsto \\ &\left[\frac{\eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}) - y}{\eta^{(1)}(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta})} - \frac{\eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}_0) - y}{\eta^{(1)}(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}_0)}\right]\mathbb{1}\big(\mathbf{x} \in \mathcal{N}_{[c\log n]}(\delta)\big) : \\ &\|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r, \delta \in \Delta\Big\}\end{aligned}$$

*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.9.* For a fixed vector space $\mathcal{B}$ of dimension $\dim \mathcal{B}$,

$$\mathcal{W}_{\mathcal{B}} := \left\{(\mathbf{x}, y) \mapsto \frac{\eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}) - y}{\eta^{(1)}(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta})} : \boldsymbol{\beta} \in \mathcal{B}, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r\right\}$$

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.8. Therefore, for any fixed $c > 0$, $\delta \in \Delta$, the class

$$\mathcal{W}_{3,\delta} := \left\{(\mathbf{x}, y) \mapsto \frac{\eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}) - y}{\eta^{(1)}(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta})} \mathbb{1}\big(\mathbf{x} \in \mathcal{N}_{[c \log n]}(\delta)\big) : \boldsymbol{\beta} \in \mathbb{R}^K, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r\right\}$$

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 $\mathbb{1}\big(\mathbf{x} \in \mathcal{N}_{[c \log n]}(\delta)\big)$. Subtracting a fixed bounded function does not change this fact, so the same is true of

$$\begin{aligned}\mathcal{G}_{3,\delta} := \Big\{&(\mathbf{x}, y) \mapsto \\ &\left[\frac{\eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}) - y}{\eta^{(1)}(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta})} - \frac{\eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}_0) - y}{\eta^{(1)}(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}_0)}\right] \mathbb{1}\big(\mathbf{x} \in \mathcal{N}_{[c \log n]}(\delta)\big) : \\ &\boldsymbol{\beta} \in \mathbb{R}^K, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0(q)\|_\infty \le r\Big\}\end{aligned}$$

Since there are $O\big(h^{-d}\big)$ such classes and $\log(1/h) \lesssim \log n$, using the chain (E.17) we obtain that $\mathcal{G}_3$ satisfies the uniform entropy bound (A.3) with with $A \lesssim 1$ and $V \lesssim \log^d n$. $\square$

**Verifying Assumption B.2(iv)** 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.8 and Lemma H.9.

**Verifying the addition to Assumption B.2(iv) described in Section E.2** This is verified (in a more general setting) in Section I.

**Verifying (F.14)** This is obvious since $\mathcal{Q}$ is a singleton.

**Verifying the condition in Lemma D.5(v)** Recall that in this case

$$\Psi_1(\mathbf{x}, \eta) = \frac{\eta^2 - 2\eta\pi(\mathbf{x}) + \pi(\mathbf{x})}{\eta^2(1 - \eta)^2}.$$

The numerator is always positive since $0 < \pi(\mathbf{x}) < 1$, and the denominator is also positive since $\eta \in (0, 1)$. Since $\Psi_1(\mathbf{x}, \eta)$ 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 $(0, 1)$, $\Psi_1(\mathbf{x}, \eta)$ is bounded away from zero uniformly over $\mathbf{x} \in \mathcal{X}$ and $\eta$ lying in this compact subset.

**Verifying the condition in Lemma D.5(vi)** In this verification, we will use $\theta_1 := \mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}$, $\tilde{\theta}_1 := \mathbf{p}(\mathbf{x})^\intercal \widetilde{\boldsymbol{\beta}}$ and $\theta_2 := \mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}_0$ to simplify notations. Note that for $\theta$ lying in a fixed compact, both functions $\log \eta(\theta)$ and $\log(1 - \eta(\theta))$ are Lipschitz in $\theta$, so if $\|\boldsymbol{\beta}\|_\infty \le R$ and $\|\tilde{\boldsymbol{\beta}}\|_\infty \le R$, we have

$$|\rho(y, \eta(\theta_1)) - \rho(y, \eta(\theta_2)) - \rho\Big(y, \eta(\tilde{\theta}_1)\Big) + \rho(y, \eta(\theta_2))| \lesssim |\theta_1 - \tilde{\theta}_1| \lesssim |\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}|,$$

where the constants in $\lesssim$ are allowed to depend on $R$ but not on $\boldsymbol{\beta}, \tilde{\boldsymbol{\beta}}$, $\mathbf{x}$ or $y$ (we used that $y$ and $1-y$ are bounded by 1). The result follows.

This concludes the proof of Proposition H.7.

**Lemma H.10** (Verification of strong approximation assumptions)**.** *Suppose the setting of Proposition H.7 applies, and the probability space is rich enough.*

*(a) If Conditions (i) and (ii) of Theorem G.1 are true, then so is Condition (iii), as long as there is an estimator $\widehat{\mathbf{\Sigma}}$ (known measurable function of the data) such that*

$$\|\widehat{\mathbf{\Sigma}}-\bar{\mathbf{\Sigma}}\| = o\biggl(\frac{h^d r_{\mathrm{str}}}{\sqrt{\log K}}\biggr).$$

*(b) If Condition (i) of Theorem G.7 is true, then so are Conditions (ii) and (iii).*

*Proof.* (a) In this case, $\boldsymbol{Z}_n$ is a $K$-dimensional vector, which is conditionally on $\{\mathbf{x}_i\}_{i=1}^n$ mean-zero and Gaussian with covariance matrix

$$\mathbb{E}[\boldsymbol{Z}_n\boldsymbol{Z}_n{}^\mathsf{T} \mid \{\mathbf{x}_i\}_{i=1}^n] = h^{-d}\mathbb{E}_n\Bigl[\sigma^2(\mathbf{x}_i)\eta^{(1)}(\mu_0(\mathbf{x}_i))^2\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\mathsf{T}\Bigr] = h^{-d}\bar{\mathbf{\Sigma}}.$$

Let $N_K$ be a standard $K$-dimensional Gaussian vector independent of the data. Clearly, we can put $\boldsymbol{Z}_n^* = h^{-d/2}\bar{\mathbf{\Sigma}}^{1/2}N_K$.

If we find an estimator $\widehat{\mathbf{\Sigma}}$ of $\bar{\mathbf{\Sigma}}$, we can put $\widehat{\boldsymbol{Z}}_n^* = h^{-d/2}\widehat{\mathbf{\Sigma}}^{1/2}N_K$, giving by Lemma C.5

$$\begin{aligned}&\mathbb{E}\Bigl[\bigl\|\widehat{\boldsymbol{Z}}_n^* - \boldsymbol{Z}_n^*\bigr\|_\infty \Bigm| \boldsymbol{D}_n\Bigr]\\ &\quad= \mathbb{E}\Bigl[h^{-d/2}\bigl\|\widehat{\mathbf{\Sigma}}^{1/2}N_K - \bar{\mathbf{\Sigma}}_{1/2}N_K\bigr\|_\infty \Bigm| \boldsymbol{D}_n\Bigr] \le h^{-d/2}\lambda_{\min}(\bar{\mathbf{\Sigma}})^{-1/2}\|\widehat{\mathbf{\Sigma}}-\bar{\mathbf{\Sigma}}\|\sqrt{\log K}.\end{aligned}$$

Using $\lambda_{\min}(\bar{\mathbf{\Sigma}}) \gtrsim_{\mathbb{P}} h^d$, which is proven by the same argument as in Lemma C.11, and Markov's inequality, we see that (G.1) is true as long as

$$h^{-d}\|\widehat{\mathbf{\Sigma}}-\bar{\mathbf{\Sigma}}\|\sqrt{\log K} = o_{\mathbb{P}}(r_{\mathrm{str}}).$$

(b) Since $\boldsymbol{Z}_n$ is a conditionally on $\{\mathbf{x}_i\}_{i=1}^n$ Gaussian vector with a conditional covariance matrix $h^{-d}\bar{\mathbf{\Sigma}}$, we get that the vector

$$\boldsymbol{\xi}_n = h^{d/2}\bar{\mathbf{\Sigma}}_{-1/2}\boldsymbol{Z}_n$$

is conditionally on $\{\mathbf{x}_i\}_{i=1}^n$ standard Gaussian (and, in particular, unconditionally standard Gaussian). Put $\widetilde{\boldsymbol{Z}}_n = h^{-d/2}\mathbf{\Sigma}_0^{1/2}\boldsymbol{\xi}_n$. Again using Lemma C.5, we get

$$\mathbb{E}\Bigl[\bigl\|\widetilde{\boldsymbol{Z}}_n - \boldsymbol{Z}_n\bigr\|_\infty \Bigm| \{\mathbf{x}_i\}_{i=1}^n\Bigr] \le h^{-d/2}\lambda_{\min}(\mathbf{\Sigma}_0)^{-1/2}\|\bar{\mathbf{\Sigma}}-\mathbf{\Sigma}_0\|\sqrt{\log K} \overset{\text{(a)}}{\lesssim_{\mathbb{P}}} \frac{\log(1/h)}{\sqrt{nh^d}} \lesssim r_{\mathrm{str}},$$

where in (a) we used $\lambda_{\min}(\mathbf{\Sigma}_0) \gtrsim h^d$ and $\|\bar{\mathbf{\Sigma}}-\mathbf{\Sigma}_0\| \lesssim_{\mathbb{P}} h^d\sqrt{\frac{\log(1/h)}{nh^d}}$, which is proven by the same argument as in Lemma C.11.

Condition (iii) is obvious: just take $\widetilde{\boldsymbol{Z}}_n^* = h^{-d/2}\mathbf{\Sigma}_{0,1/2}N_K$, where $N_K$ is defined above in this proof. ∎

# I 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 in Assumption B.5(iv).

**Proposition I.1** (A simpler generalization of all examples)**.** *Assume the following conditions.*

*(i) The function $\rho(y,\eta;q)$ is of the form*

$$\rho(y,\eta;q) = \mathcal{T}(y,\eta) + (f_1(y) + D_1\eta)\mathbb{1}\{y \le \eta\} + (f_2(y) + D_2\eta)\mathbb{1}\{y \le q\} + (f_3(y) + D_3\eta)q,$$

*where $f_j$ are fixed continuous functions of $y$, $D_j$ are universal constants and $\mathcal{T}(y;\eta) : \mathcal{Y} \times \mathcal{E} \to \mathbb{R}$ is a measurable function not depending on $q$, differentiable in $\eta$ for any fixed $y$ with a derivative $\tau(y,\eta) := \frac{\partial}{\partial\eta}\mathcal{T}(y,\eta)$.*

*(ii) $\mathbb{E}[\tau(y_i,\eta) \mid \mathbf{x}_i = \mathbf{x}]$ is also differentiable in $\eta$ for any $\mathbf{x}$. The functions*

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

*are continuous in their arguments, and $\alpha$-Hölder continuous in $\eta$ 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 this compact $\mathcal{K}$, but not on the other argument, i.e. $y$ and $\mathbf{x}$ respectively), $\alpha \in (0,1]$.*

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

*(iv) Assumptions B.1, B.3 and B.4 and Assumption B.2 Items (i) to (iii) hold with $\overline{\psi}(\mathbf{x},y) = \overline{\tau}(\mathbf{x},y) + |D_1| + |D_2| + |D_3|\max_{q\in\mathcal{Q}}|q|$, where $\overline{\tau}(\mathbf{x},y)$ is a measurable envelope of $\Big\{(\mathbf{x},y) \mapsto \tau(y,\eta(\mu_0(\mathbf{x},q)))\Big\}$.*

*(v) $q \mapsto \mu_0(\mathbf{x},q)$ is nondecreasing.*

*Then Assumption B.2(iv) and the addition to it described in Section E.2 also hold.*

We will prove this now.

In this case

$$\psi(y,\eta;q) = \tau(y,\eta) + D_1\mathbb{1}\{y \le \eta\} + D_2\mathbb{1}\{y \le q\} + D_3 q.$$

**Lemma I.2** (Class $G_1$)**.** *The class $\mathcal{G}_1$ described in Assumption B.2(iv) with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$, $V \lesssim K$.*

*Proof of Lemma I.2.* 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

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

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 $\alpha$-Hölder continuity of $\tau(y,\cdot)$ on compacta

$$\left|\tau(y,\eta(\mathbf{p}(\mathbf{x})^{\mathsf{T}}\boldsymbol{\beta})) - \tau(y;\eta(\mu_0(\mathbf{x},q))) - \tau\Big(y,\eta(\mathbf{p}(\mathbf{x})^{\mathsf{T}}\widetilde{\boldsymbol{\beta}})\Big) + \tau(y,\eta(\mu_0(\mathbf{x},\tilde{q})))\right|$$

$$
\begin{aligned}
&\lesssim \left|\eta(\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}) - \eta(\mathbf{p}(\mathbf{x})^\intercal \widetilde{\boldsymbol{\beta}})\right|^\alpha + |\eta(\mu_0(\mathbf{x}, q)) - \eta(\mu_0(\mathbf{x}, \tilde{q}))|^\alpha \\
&\lesssim \left\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right\|_\infty^\alpha + |q - \tilde{q}|^\alpha,
\end{aligned}
$$

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 $q \mapsto \mu_0(\mathbf{x}, q)$ is Lipschitz (uniformly in $\mathbf{x}$). Again by $\alpha$-Hölder continuity for any fixed $r > 0$ the class

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

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 I.2 by Lemma C.4. ☐

**Lemma I.3** (Class $G_2$)**.** *The class $\mathcal{G}_2$ described in Assumption B.2(iv) with envelope $\overline{\psi}(\mathbf{x}, y)$ satisfies the uniform entropy bound (A.3) with $A \lesssim 1$, $V \lesssim 1$.*

*Proof of Lemma I.3.* The class

$$
\left\{\mathcal{X} \times \mathcal{Y} \ni (\mathbf{x}, y) \mapsto \tau(y, \eta(\mu_0(\mathbf{x}, q))) \colon q \in \mathcal{Q}\right\}
$$

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

The class

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

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

The class

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

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.4).

The class

$$
\left\{\mathcal{X} \times \mathcal{Y} \ni (\mathbf{x}, y) \mapsto q \colon q \in \mathcal{Q}\right\}
$$

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 $\mathbf{x}, y$).

It is left to apply Lemma C.4, concluding the proof of Lemma I.3. ☐

**Lemma I.4** (Class $G_3$)**.** *The class $\mathcal{G}_3$ described in Assumption B.2(iv) 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 I.4.* Fix $\delta \in \Delta$ and some large enough $R > 0$. Note that if $\mathbb{1}\{\mathbf{x} \in \mathcal{N}_{[c \log n]}(\delta)\}$ and $\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}$ 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 $\varepsilon$, the class of vectors $\{\boldsymbol{\beta} \in \mathcal{B}_\delta, \|\boldsymbol{\beta}\|_\infty < R\}$ has an infinity-norm $\varepsilon$-net $\bar{\mathcal{B}}_\delta^\varepsilon$ such that

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

where $C$ is some positive constant.

By $\alpha$-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

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

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 $O\bigl(h^{-d}\bigr)$ such classes

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

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.17).

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

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

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

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

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 $O\bigl(h^{-d}\bigr)$ such classes

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

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.17).

It is left to apply Lemma C.4 once again, concluding the proof of Lemma I.4. □

**Lemma I.5** (Class $G_4$)**.** *The class $\mathcal{G}_4$ described in Assumption B.2(iv) with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A\lesssim 1$, $V\lesssim 1$.*

*Proof of Lemma I.5.* In this case

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

and

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

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

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

where in (a) we used that $\eta(\cdot)$ on compacta is Lipschitz and $q \mapsto \mu_0(\mathbf{x}, q)$ is uniformly Lipschitz in $q$. The result of Lemma I.5 follows. □

**Lemma I.6** (Class $G_5$)**.** *The class $\mathcal{G}_5$ described in Assumption B.2(iv) with a large enough constant envelope satisfies the uniform entropy bound (A.3) with $A \lesssim 1$, $V \lesssim 1$.*

*Proof of Lemma I.6.* Take $R > 0$ fixed and large enough so that $\|\boldsymbol{\beta}_{0,q}\|_\infty \leq R$ for all $q$ and $n$. Note that for $p_l(\mathbf{x})$ and $\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}$ 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 $\{\boldsymbol{\beta} \in \mathcal{B}_l, \|\boldsymbol{\beta}\|_\infty \leq R\}$ has an infinity-norm $\varepsilon$-net $\bar{\mathcal{B}}_l^\varepsilon$ such that

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

where $C$ is some positive constant.

By $\alpha$-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

$$\mathcal{G}_5' := \{(\mathbf{x}, y) \mapsto p_l(\mathbf{x})(\tau(y, \eta(\mu_0(\mathbf{x}, q))) - \tau(y, \eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta}_0(q)))) : q \in \mathcal{Q}\}$$

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 I.3, the class

$$\left\{\mathcal{X} \times \mathcal{Y} \ni (\mathbf{x}, y) \mapsto \mathbb{1}\{y \leq \eta(\mu_0(\mathbf{x}, q))\} \colon q \in \mathcal{Q}\right\}$$

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

$$\left\{\mathcal{X} \times \mathcal{Y} \ni (\mathbf{x}, y) \mapsto p_l(\mathbf{x})\mathbb{1}\{y \leq \eta(\mathbf{p}(\mathbf{x})^\intercal\boldsymbol{\beta})\} : \boldsymbol{\beta} \in \mathbb{R}^K\right\}$$

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 I.6. □

**Verifying the addition to Assumption B.2(iv) described in Section E.2** 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 $\alpha$-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

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

has a covering number bound

$$\log N(\mathcal{G}', \text{sup-norm}, \varepsilon) \lesssim \log^d n \log\left(\frac{C_1}{\varepsilon}\right), \tag{I.1}$$

for all small enough positive $\varepsilon$, 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 $O\big(\log^d n\big)$ if $\mathbb{1}\big\{\mathbf{x} \in \mathcal{N}_{[c \log n]}(\delta)\big\}$ is not zero. Applying the mean-value theorem to

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

and $\alpha$-Hölder continuity again, we see that class $\mathcal{G}'$ has an envelope which is $\varepsilon_n$ multiplied by a large enough constant $C_2$. Replacing $\varepsilon$ with $C_2\varepsilon\varepsilon_n$ (for large enough $n$ this is small enough) in (I.1), we get a covering number bound

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

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 $O\big(h^{-d}\big)$ such classes

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

also with envelope $\varepsilon_n$ 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.17).

Next,

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

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

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

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 Section E.2.

It is left to apply Lemma C.4.

The proof of Proposition I.1 is finished.

# J Other parameters of interest

This section formalizes the discussion in Section 8 of the paper. The following theorem is now a simple corollary of the previous results presented in this supplemental appendix.

**Theorem J.1** (Other parameters of interest)**.**

*(a) Suppose all the conditions of Theorem E.1(a) hold with* $\mathbf{v} = \mathbf{0}$*, and all the conditions of Theorem F.4(a) hold. Then*

$$
\begin{aligned}
&\sup_{\mathbf{q}\in\mathcal{Q}}\sup_{\mathbf{x}\in\mathcal{X}}\left|\frac{\eta\big(\widehat{\mu}(\mathbf{x},\mathbf{q})\big) - \eta\big(\mu_0(\mathbf{x},\mathbf{q})\big)}{\big|\eta^{(1)}(\mu_0(\mathbf{x},\mathbf{q}))\big|\sqrt{\bar{\boldsymbol{\Omega}}_{\mathbf{0}}(\mathbf{x},\mathbf{q})/n}} - \operatorname{sign}\big\{\eta^{(1)}(\mu_0(\mathbf{x},\mathbf{q}))\big\}\bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{0})^{\mathsf{T}}\boldsymbol{Z}_n(\mathbf{q})\right| \\
&\quad = o_{\mathbb{P}}(r_{\mathrm{str}}) + O_{\mathbb{P}}\Big(\sqrt{nh^d}\big(r_{\mathrm{uc}}^2 + r_{\mathrm{Bah}}\big)\Big)
\end{aligned}
$$

*with* $\boldsymbol{Z}_n(\mathbf{q})$ *defined in Theorem F.4.*

*(b) Fix any* $k \in \{1, \dots, d\}$*. Suppose all the conditions of Theorem E.1(a) hold with* $\mathbf{v} = \mathbf{e}_k$*, where* $\mathbf{e}_k = (0, \dots, 0, 1, 0, \dots, 0)^{\mathsf{T}} \in \mathbb{R}^d$ *with 1 at the kth place, and all the conditions of Theorem F.4(a) hold. Then*

$$
\begin{aligned}
\sup_{\mathbf{q},\mathbf{x}}\Bigg|&\frac{\eta^{(1)}\big(\widehat{\mu}(\mathbf{x},\mathbf{q})\big)\widehat{\mu}^{(\mathbf{e}_k)}(\mathbf{x},\mathbf{q}) - \eta^{(1)}\big(\mu_0(\mathbf{x},\mathbf{q})\big)\mu_0^{(\mathbf{e}_k)}(\mathbf{x},\mathbf{q})}{\big|\eta^{(1)}(\mu_0(\mathbf{x},\mathbf{q}))\big|\sqrt{\bar{\boldsymbol{\Omega}}_{\mathbf{e}_k}(\mathbf{x},\mathbf{q})}} \\
&- \operatorname{sign}\big\{\eta^{(1)}(\mu_0(\mathbf{x},\mathbf{q}))\big\}\bar{\bar{\boldsymbol{\ell}}}(\mathbf{x},\mathbf{q},\mathbf{e}_k)^{\mathsf{T}}\boldsymbol{Z}_n(\mathbf{q})\Bigg| \\
&= o_{\mathbb{P}}(r_{\mathrm{str}}) + O_{\mathbb{P}}\Big(\sqrt{nh^d}\big(r_{\mathrm{uc}}^2 + r_{\mathrm{Bah}} + hr_{\mathrm{uc}}\big)\Big).
\end{aligned}
$$

(c) *If $\overline{\psi}(\mathbf{x}_i, y_i)$ is $\sigma^2$-sub-Gaussian conditionally on $\mathbf{x}_i$, then in Assertions (a) and (b) $r_{\mathrm{str}}$ can be replaced with $r_{\mathrm{str}}^{\mathrm{sub}}$.*

*Proof.* (a) We have uniformly over $\mathbf{q}$ and $\mathbf{x}$

$$
\begin{aligned}
&\eta(\widehat{\mu}(\mathbf{x},\mathbf{q})) - \eta(\mu_0(\mathbf{x},\mathbf{q})) \\
&\overset{(a)}{=} \eta^{(1)}(\mu_0(\mathbf{x},\mathbf{q}))(\widehat{\mu}(\mathbf{x},\mathbf{q}) - \mu_0(\mathbf{x},\mathbf{q})) + \frac{\eta^{(2)}(\xi)}{2}(\widehat{\mu}(\mathbf{x},\mathbf{q}) - \mu_0(\mathbf{x},\mathbf{q}))^2 \\
&\overset{(b)}{=} \eta^{(1)}(\mu_0(\mathbf{x},\mathbf{q}))(\widehat{\mu}(\mathbf{x},\mathbf{q}) - \mu_0(\mathbf{x},\mathbf{q})) + O_{\mathbb{P}}\big(r_{\mathrm{uc}}^2\big) \\
&\overset{(c)}{=} \eta^{(1)}(\mu_0(\mathbf{x},\mathbf{q}))\bar{t}(\mathbf{x},\mathbf{q},\mathbf{0})\sqrt{\bar{\boldsymbol{\Omega}}_{\mathbf{0}}(\mathbf{x},\mathbf{q})/n} + O_{\mathbb{P}}\big(r_{\mathrm{uc}}^2 + r_{\mathrm{Bah}}\big) \\
&\overset{(d)}{=} \eta^{(1)}(\mu_0(\mathbf{x},\mathbf{q}))\sqrt{\bar{\boldsymbol{\Omega}}_{\mathbf{0}}(\mathbf{x},\mathbf{q})/n}\Big(\bar{t}(\mathbf{x},\mathbf{q},\mathbf{0}) + O_{\mathbb{P}}\Big(\sqrt{nh^d}\big(r_{\mathrm{uc}}^2 + r_{\mathrm{Bah}}\big)\Big)\Big).
\end{aligned}
$$

Here, (a) is by Taylor expansion, with some $\xi = \xi_{\mathbf{q},\mathbf{x}}$ between $\widehat{\mu}(\mathbf{x},\mathbf{q})$ and $\mu_0(\mathbf{x},\mathbf{q})$. In (b), we used consistency (giving that $\eta^{(2)}(\xi)$ does not exceed a fixed constant not depending on $\mathbf{q}$ or $\mathbf{x}$) and Corollary E.2. (c) is by Theorem E.1(a) and since $\eta^{(1)}(\mu_0(\mathbf{x},\mathbf{q}))$ is uniformly bounded. (d) is by $h^{-2|\mathbf{v}|-d} \lesssim_{\mathbb{P}} \inf_{\mathbf{q},\mathbf{x}} |\bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x},\mathbf{q})|$ (by Lemma F.6) and since $\big|\eta^{(1)}(\mu_0(\mathbf{x},\mathbf{q}))\big|$ is bounded away from zero by Assumption B.2(iii).

Rewriting, we obtain

$$
\sup_{\mathbf{q},\mathbf{x}}\left|\frac{\eta(\widehat{\mu}(\mathbf{x},\mathbf{q})) - \eta(\mu_0(\mathbf{x},\mathbf{q}))}{\eta^{(1)}(\mu_0(\mathbf{x},\mathbf{q}))\sqrt{\bar{\boldsymbol{\Omega}}_{\mathbf{0}}(\mathbf{x},\mathbf{q})/n}} - \bar{t}(\mathbf{x},\mathbf{q},\mathbf{0})\right| \lesssim_{\mathbb{P}} \sqrt{nh^d}\big(r_{\mathrm{uc}}^2 + r_{\mathrm{Bah}}\big).
$$

It is left to combine this with Theorem F.4 and use the triangle inequality.

(b) We have uniformly over $\mathbf{q}$ and $\mathbf{x}$

$$
\begin{aligned}
&\eta^{(1)}(\widehat{\mu}(\mathbf{x},\mathbf{q}))\widehat{\mu}^{(\mathbf{e}_k)}(\mathbf{x},\mathbf{q}) - \eta^{(1)}(\mu_0(\mathbf{x},\mathbf{q}))\mu_0^{(\mathbf{e}_k)}(\mathbf{x},\mathbf{q}) \\
&\overset{(a)}{=} \eta^{(1)}(\mu_0(\mathbf{x},\mathbf{q}))\Big(\widehat{\mu}^{(\mathbf{e}_k)}(\mathbf{x},\mathbf{q}) - \mu_0^{(\mathbf{e}_k)}(\mathbf{x},\mathbf{q})\Big) \\
&\quad + \Big(\eta^{(1)}(\widehat{\mu}(\mathbf{x},\mathbf{q})) - \eta^{(1)}(\mu_0(\mathbf{x},\mathbf{q}))\Big)\widehat{\mu}^{(\mathbf{e}_k)}(\mathbf{x},\mathbf{q}) \\
&\overset{(b)}{=} \eta^{(1)}(\mu_0(\mathbf{x},\mathbf{q}))\Big(\widehat{\mu}^{(\mathbf{e}_k)}(\mathbf{x},\mathbf{q}) - \mu_0^{(\mathbf{e}_k)}(\mathbf{x},\mathbf{q})\Big) + \eta^{(2)}(\zeta)(\widehat{\mu}(\mathbf{x},\mathbf{q}) - \mu_0(\mathbf{x},\mathbf{q}))\widehat{\mu}^{(\mathbf{e}_k)}(\mathbf{x},\mathbf{q}) \\
&\overset{(c)}{=} \eta^{(1)}(\mu_0(\mathbf{x},\mathbf{q}))\bar{t}(\mathbf{x},\mathbf{q},\mathbf{e}_k)\sqrt{\bar{\boldsymbol{\Omega}}_{\mathbf{e}_k}(\mathbf{x},\mathbf{q})/n} \\
&\quad + \eta^{(2)}(\zeta)(\widehat{\mu}(\mathbf{x},\mathbf{q}) - \mu_0(\mathbf{x},\mathbf{q}))\widehat{\mu}^{(\mathbf{e}_k)}(\mathbf{x},\mathbf{q}) + O_{\mathbb{P}}\big(h^{-1}r_{\mathrm{Bah}}\big) \\
&\overset{(d)}{=} \eta^{(1)}(\mu_0(\mathbf{x},\mathbf{q}))\bar{t}(\mathbf{x},\mathbf{q},\mathbf{e}_k)\sqrt{\bar{\boldsymbol{\Omega}}_{\mathbf{e}_k}(\mathbf{x},\mathbf{q})/n} \\
&\quad + \eta^{(2)}(\zeta)(\widehat{\mu}(\mathbf{x},\mathbf{q}) - \mu_0(\mathbf{x},\mathbf{q}))\Big(\widehat{\mu}^{(\mathbf{e}_k)}(\mathbf{x},\mathbf{q}) - \mu_0^{(\mathbf{e}_k)}(\mathbf{x},\mathbf{q})\Big) \\
&\quad + \eta^{(2)}(\zeta)(\widehat{\mu}(\mathbf{x},\mathbf{q}) - \mu_0(\mathbf{x},\mathbf{q}))\mu_0^{(\mathbf{e}_k)}(\mathbf{x},\mathbf{q}) + O_{\mathbb{P}}\big(h^{-1}r_{\mathrm{Bah}}\big) \\
&\overset{(e)}{=} \eta^{(1)}(\mu_0(\mathbf{x},\mathbf{q}))\bar{t}(\mathbf{x},\mathbf{q},\mathbf{e}_k)\sqrt{\bar{\boldsymbol{\Omega}}_{\mathbf{e}_k}(\mathbf{x},\mathbf{q})/n} + O_{\mathbb{P}}\big(h^{-1}(r_{\mathrm{uc}}^2 + r_{\mathrm{Bah}}) + r_{\mathrm{uc}}\big) \\
&\overset{(f)}{=} \eta^{(1)}(\mu_0(\mathbf{x},\mathbf{q}))\sqrt{\bar{\boldsymbol{\Omega}}_{\mathbf{e}_k}(\mathbf{x},\mathbf{q})/n}\Big[\bar{t}(\mathbf{x},\mathbf{q},\mathbf{e}_k) + O_{\mathbb{P}}\Big(h\sqrt{nh^d}\big(h^{-1}(r_{\mathrm{uc}}^2 + r_{\mathrm{Bah}}) + r_{\mathrm{uc}}\big)\Big)\Big].
\end{aligned}
$$

Here, (a) is just rewriting. In (b) we used the mean-value theorem, with some $\zeta = \zeta_{\mathbf{q},\mathbf{x}}$ between $\widehat{\mu}(\mathbf{x},\mathbf{q})$ and $\mu_0(\mathbf{x},\mathbf{q})$. In (c) we used Theorem E.1(a) with $\mathbf{v} = \mathbf{e}_k$ and that $\eta^{(1)}(\mu_0(\mathbf{x},\mathbf{q}))$ is bounded uniformly over $\mathbf{q},\mathbf{x}$. (d) is just rewriting. In (e) we used consistency (giving that $\eta^{(2)}(\zeta)$ does not exceed a fixed constant not depending on $\mathbf{q}$ or $\mathbf{x}$), Corollary E.2 with $\mathbf{v} = \mathbf{0}$ and $\mathbf{v} = \mathbf{e}_k$, and uniform boundedness of first partial derivatives of $\mu_0(\cdot, \mathbf{q})$. (f) is by $h^{-2|\mathbf{v}|-d} \lesssim_{\mathbb{P}} \inf_{\mathbf{q},\mathbf{x}} |\bar{\boldsymbol{\Omega}}_{\mathbf{v}}(\mathbf{x}, \mathbf{q})|$ (by Lemma F.6) and since $\left|\eta^{(1)}(\mu_0(\mathbf{x}, \mathbf{q}))\right|$ is bounded away from zero by Assumption B.2(iii).

Rewriting, we obtain

$$\begin{aligned}
&\sup_{\mathbf{q},\mathbf{x}} \left| \frac{\eta^{(1)}(\widehat{\mu}(\mathbf{x}, \mathbf{q}))\widehat{\mu}^{(\mathbf{e}_k)}(\mathbf{x}, \mathbf{q}) - \eta^{(1)}(\mu_0(\mathbf{x}, \mathbf{q}))\mu_0^{(\mathbf{e}_k)}(\mathbf{x}, \mathbf{q})}{\eta^{(1)}(\mu_0(\mathbf{x}, \mathbf{q}))\sqrt{\bar{\boldsymbol{\Omega}}_{\mathbf{e}_k}(\mathbf{x}, \mathbf{q})}} - \bar{t}(\mathbf{x}, \mathbf{q}, \mathbf{e}_k) \right| \\
&\quad \lesssim_{\mathbb{P}} \sqrt{nh^d}\big(r_{\mathrm{uc}}^2 + r_{\mathrm{Bah}} + hr_{\mathrm{uc}}\big).
\end{aligned}$$

It is left to combine this with Theorem F.4 and use the triangle inequality.

(c) The argument is the same as for Parts (i) and (ii) with $r_{\mathrm{str}}$ replaced by $r_{\mathrm{str}}^{\mathrm{sub}}$. □

# References

[1] 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.

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

[3] 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.

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

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

[6] 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.

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

[8] Dudley, R. M. (2014). *Uniform central limit theorems*, *142*: Cambridge university press.

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

[10] 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.

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