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

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

September 8, 2024

**Abstract**

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

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

[*] Department of Operations Research and Financial Engineering, Princeton University.

[†] School of Economics and Management, Tsinghua University.

[‡] Department of Operations Research and Financial Engineering, Princeton University.

# Contents

# 1 Introduction

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

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

where the minimization is over the space of measurable functions from $\mathcal{X}$ to $\mathbb{R}$ (in particular, assume that the minimum is achieved, which is true in most cases). This setup covers many settings of interest in nonparametric statistics, econometrics, and data science, including generalized linear models, robust nonlinear regression, and generalized conditional quantile regression. In practice, the parameter of interest may be $\mu_0$ itself, or otherwise specific transformations thereof such as $\eta(\mu_0(\mathbf{x}_i))$ or its partial derivatives.

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

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

where $\mathbf{x} \mapsto \mathbf{p}(\mathbf{x}) = \mathbf{p}(\mathbf{x}; \Delta, m) = \big(p_1(\mathbf{x}; \Delta, m), \ldots, p_K(\mathbf{x}; \Delta, m)\big)^\mathsf{T}$ 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 $\mathcal{X}$, and $\mathcal{B} \subseteq \mathbb{R}^K$ is the feasible set of the optimization problem. The $m$ parameter controls how well $\mu_0$ can be approximated by linear combinations of the basis (Assumption 6); the partition being quasi-uniform intuitively means that the largest size of a cell cannot get asymptotically bigger than the smallest one (Assumption 4). We consider large sample approximations where $d$ and $m$ are fixed constants, and $\bar{\kappa} \to \infty$ (and thus $K \to \infty$) as $n \to \infty$. As a consequence, appropriate choices of $\Delta$ and $\mathbf{p}(\cdot)$ will enable valid nonparametric approximations of $\mu_0$, and transformations thereof. See, e.g., [19] for a textbook introduction to partitioning-based methods, and Section 1.1 for prior literature. Typical basis functions covered by our conditions include piecewise polynomials, splines, and compactly supported wavelets, as well as those generated by certain decision tree methods. Section 3 gives the precise conditions on $\mathbf{p}(\cdot)$ and $\Delta$ (Assumptions 4 and 5).

Provided that the vector-valued functional coefficient estimator $\widehat{\boldsymbol{\beta}}(q)$ satisfies the uniform consistency requirement $\sup_{q \in \mathcal{Q}} \big\|\widehat{\boldsymbol{\beta}}(q) - \boldsymbol{\beta}_0(q)\big\|_\infty = o_\mathbb{P}(1)$, where $\|\cdot\|_\infty$ denotes the $\ell^\infty$-norm and $\boldsymbol{\beta}_0\colon \mathcal{Q} \to \mathbb{R}^K$ denotes coefficients such that $\boldsymbol{\beta}_0(q)^\mathsf{T}\mathbf{p}$ approximates $\mu_0$ well enough uniformly over $\mathcal{X} \times \mathcal{Q}$ (Assumption 6), we present three main theoretical results for the partitioning-based series $M$-estimator in (1.2):

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

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

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

These results contribute to the literature (Section 1.1) by offering estimation and inference methods that are uniformly valid over both $\mathcal{X}$ and $\mathcal{Q}$, while allowing for a large class of possibly non-smooth loss functions, and under improved side conditions on the tuning parameter $K$ which is the number of elements in the approximation basis $\mathbf{p}(\cdot)$. The closest antecedent to our work is [2], which considers exclusively nonparametric conditional quantile series regression estimation and inference uniformly over $\mathcal{X} \times \mathcal{Q}$ with $\eta(u) = u$, and under the side condition $K^4/n \to 0$, up to polylog$(n)$ terms, among other requirements. In contrast, for the special case of nonparametric quantile regression, this paper allows for a non-identity (inverse) link function $\eta(\cdot)$, and establishes convergence rates under the weaker conditions $K/n \to 0$ for piecewise polynomials and $K^2/n \to 0$ for connected basis, while for uniform inference it requires the weaker condition $K^3/n \to 0$, in all cases up to polylog$(n)$ terms. We also weaken other assumptions imposed in prior work, as discussed precisely throughout the manuscript. See Example 1 in Sections 2 and 7.1, and Sections 5.2 and 6.1.

More broadly, our paper allows for a large class of possibly non-smooth loss functions, beyond the check function for quantile regression, and characterizes precisely how their degree of smoothness affects the order of the remainder in the uniform Bahadur representation for $\widehat{\mu}$, its convergence rates, and the validity of the associated uniform inference procedures. As a consequence, our general theory gives uniform estimation and inference results for new nonparametric estimators of interest in statistics and data science. We demonstrate the broad applicability of our theoretical results with the following new applications:

- *Generalized Conditional Distribution Regression.* [14] studies this problem in a parametric setting ($K$ fixed) for counterfactual analysis and causal inference. We give uniform estimation and inference for nonparametric partitioning-based conditional distribution regression. In particular, our uniform (over $\mathcal{Q}$) results are useful for constructing nonparametric inference procedures in treatment effect and policy evaluation settings. See Example 2 in Sections 2 and 7.2.

- *Generalized $L_p$ Regression Estimation.* [26] studies $L_p$ regression estimation with identity transformation $\eta(\cdot)$ in a parametric setting ($K$ fixed). We study uniform estimation and inference for nonparametric partitioning-based generalized $L_p$ regression with $p \in [1, 2]$, covering the full interpolation between nonparametric generalized median regression ($p = 1$) and nonparametric nonlinear least squares regression ($p = 2$), while also allowing for a non-identity $\eta(\cdot)$. See Example 3 in Sections 2 and 7.3. These results are useful in nonparametric robust statistics settings.

- *Other Generalized, Nonlinear, Robust Regression Methods.* Our results also cover other application of interest such as nonparametric partitioning-based (quasi-) maximum likelihood logistic regression, Poisson regression, censored and truncated regression, and Tukey and Huber regression, just to mention a few examples. In particular, Example 4 in Sections 2 and 7.4 considers the case of nonparametric partitioning-based logistic regression because of its importance in classification and machine learning settings.

To our knowledge, uniform nonparametric partitioning-based estimation and inference for the examples mentioned so far (and many others) have not been studied in the literature before at the level of generality we achieve. The only exception is nonparametric quantile regression [2] with identity (inverse) link $\eta(\cdot)$, for which our general theoretical results substantially improve upon the side rate conditions and other assumptions imposed previously.

A challenge in our analysis is that many examples of interest have non-convex objective functions in (1.2), thereby requiring a more careful analysis when it comes to the construction of the partitioning-based estimator $\widehat{\mu}(\mathbf{x}, q)$. To address this challenge, we separate our theoretical work into two parts. On the one hand, as already mentioned, our main uniform estimation and inference results are obtained under the high-level consistency assumption $\sup_{q\in\mathcal{Q}} \big\|\widehat{\boldsymbol{\beta}}(q) - \boldsymbol{\beta}_0(q)\big\|_\infty = o_{\mathbb{P}}(1)$, which is agnostic about the shape features of the objective function and other optimization-related aspects underlying the nonlinear partitioning-based estimator. On the other hand, we provide primitive conditions to verify the high-level consistency condition depending on whether the loss function $\theta \mapsto \rho(y, \eta(\theta); q)$ is convex or not. More precisely, when the loss function is convex, we verify the high-level consistency condition with $\mathcal{B} = \mathbb{R}^K$, leading to unconstrained optimization in (1.2). When the loss function is non-convex, we verify the high-level consistency condition with $\mathcal{B} = \{\mathbf{b} \in \mathbb{R}^K : \|\mathbf{b}\|_\infty \leq R\}$ for some large enough fixed constant $R > 0$, leading to constrained optimization in (1.2). Such a "box" constraint is arguably a mild assumption in practice, and may be justified in theory under different regularity conditions. We also illustrate these high-level consistency results in the context of our four motivating examples in Sections 2 and 7.

## 1.1 Prior Literature

Our paper contributes to the literature on nonparametric estimation [19, 18], focusing in particular on series (or sieve) approximation methods.

The nonparametric series estimation literature is mature and well-developed for the very special case of a square loss function $\rho(y, \eta(\theta); q) = (y - \theta)^2$ with identity transformation $\eta$, which is not a function of $q \in \mathcal{Q}$. See, for example, [38], [22], [9], [3], [12], [10], [8] for pointwise and uniform over $\mathcal{X}$ estimation and inference results at different levels of generality, and with increasingly weaker technical conditions. The results in this strand of the literature explicitly exploit the special structure of the square loss function and identity transformation function, which leads to a closed-form solution of the estimator in (1.2), and hence are often obtained under minimal assumptions and technical regularity conditions. To be more precise, up to polylog$(n)$ terms and some regularity conditions, the minimal requirement $K/n \to \infty$ has been shown to be sufficient for optimal convergence rates for any $d \geq 1$, and for strong approximations uniformly over $\mathcal{X}$ when $d = 1$. Furthermore, valid strong approximations uniformly over $\mathcal{X}$ have been established for $d > 1$ under the requirement $K^3/n \to \infty$, up to polylog$(n)$ terms and mild regularity conditions.

Despite aiming for generality (i.e., allowing for a large class of loss functions with different levels of smoothness and a non-identity transformation function), this paper establishes uniform over *both* $\mathcal{X}$ and $\mathcal{Q}$ estimation and inference results under the weak assumption $K^2/n \to \infty$ for convergence rates, and under the same condition $K^3/n \to \infty$ for strong approximations as in the special case of square loss and identity transformation functions. Beyond the case of square loss function, [2] is the closest prior work we are aware of, which focuses specifically on quantile regression with identity transformation function, and imposes stronger assumptions than those herein. More broadly, we are not aware of other prior results from the nonparametric sieve M-estimation literature at the level of generality and under the weak conditions considered in this paper.

Our contributions can also be compared to recent work on nonparametric M-estimation employing other smoothing techniques. For example, [25] considers local polynomial methods, and [32] considers smoothing spline methods. In Sections 5.2 and 6.1 we discuss precisely how our results are either on par with or improve upon those prior contributions from the broader nonparametric literature.

## 1.2 Notation

We employ standard notation in probability, statistics and empirical process theory [4, 17, 23, 35]. For any vector $\mathbf{a} = (a_1, \cdots, a_M) \in \mathbb{R}^M$, we write $\|\mathbf{a}\| = (\sum_{j=1}^M a_j^2)^{1/2}$ and $\|\mathbf{a}\|_\infty = \max_{1\leq j\leq M} |a_j|$. For any real function $f$ depending on $d$ variables $(t_1, \ldots, t_d)$ and any vector $\mathbf{v} = (v_1, \cdots, v_d)$ of nonnegative integers, denote $f^{(\mathbf{v})} = \frac{\partial^{|\mathbf{v}|}}{\partial t_1^{v_1} \ldots \partial t_d^{v_d}} f$ where $|\mathbf{v}| = \sum_{k=1}^{d} v_j$. For functions that depend on $(\mathbf{x}, q)$, the multi-index derivative notation is taken with respect to the first argument $\mathbf{x}$, unless otherwise noted. We say a function $f$ is $\alpha$-Hölder on a set $\mathcal{I}$ if for some constant $C > 0$ and $\alpha > 0$, $|f(\mathbf{x}_1) - f(\mathbf{x}_2)| \leq C\|\mathbf{x}_1 - \mathbf{x}_2\|^{\alpha}$ for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{I}$. For any two numbers $a$ and $b$, $a \vee b = \max\{a, b\}$, and $a \wedge b = \min\{a, b\}$. Let $\mathbb{E}_n[g(x_i)] = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$ and $\mathbb{G}_n[g(x_i)] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(g(x_i) - \mathbb{E}[g(x_i)])$. For sequences, $a_n = O(b_n)$ or $a_n \lesssim b_n$ denotes $\limsup_n |a_n/b_n|$ is finite, $a_n = O_{\mathbb{P}}(b_n)$ denotes $\limsup_{\epsilon\to\infty} \limsup_{n\to\infty} \mathbb{P}[|a_n/b_n| \geq \epsilon] = 0$, $a_n = o(b_n)$ denotes $a_n/b_n \to 0$, and $a_n = o_{\mathbb{P}}(b_n)$ denotes $a_n/b_n \to_{\mathbb{P}} 0$, where $\to_{\mathbb{P}}$ is convergence in probability. Limits are taken as $n \to \infty$, and the dependence on $n$ is often suppressed, e. g. $K = K_n$. Also, we say a random variable $\xi$ is sub-Gaussian conditional on $\boldsymbol{X}$ if for some constant $\sigma^2 > 0$, $\mathbb{P}(|\xi| \geq t|\boldsymbol{X} = \mathbf{x}) \leq 2\exp(-t^2/\sigma^2)$ for all $t \geq 0$ and $\mathbf{x} \in \mathcal{X}$.

### 1.3 Organization

Section 2 presents our four motivating examples, which motivates our general theoretical work and demonstrates its applicability. Section 3 presents the slightly simplified high-level technical assumptions used throughout the paper; their most general form is given in the supplemental appendix. Section 4 gives sufficient conditions for the uniform consistency requirement $\sup_{q\in\mathcal{Q}} \big\|\widehat{\boldsymbol{\beta}}(q) - \boldsymbol{\beta}_0(q)\big\|_\infty = o_{\mathbb{P}}(1)$, depending on whether the objective function is convex or not. Sections 5 and 6 present our main general theoretical results: Bahadur representation, rates of convergence, strong approximation, and uniform inference. Section 7 demonstrates how our general sufficient conditions are verified for each of our motivating examples. Section 8 discusses how our results can be extended to cover other parameters of interest. Finally, Section 9 concludes.

The supplemental appendix collects all the technical proofs, and also presents other theoretical results that may be of independent interest. In particular, (i) we allow for $\mathcal{Q}$ to be a set of vectors rather than scalars, which can be useful in other examples beyond those studied in this paper; (ii) we consider more complex (VC-type) classes of loss and transformation functions, thereby covering a broader class of settings than those studied herein, but at the cost of additional, cumbersome notation and technicalities; and (iii) we present new strong approximation results for a class of $K$-dimensional linear stochastic processes indexed by $\mathcal{X} \times \mathcal{Q}$ under standard complexity and smoothness conditions, leveraging a conditional Strassen's Theorem [31, 13] and generalizing prior Yurinskii's coupling results in the literature [36, 2].

## 2 Motivating Examples

We discuss four examples of interest covered by our theoretical results. Section 7 demonstrates how our high-level assumptions, introduced in Section 3, are verified for these examples in order to obtain uniform estimation and inference results; the supplemental appendix collects omitted details.

Our first example generalizes the work of [2], who studies the large sample properties of nonparametric conditional quantile series regression with $\eta(u) = u$. We allow for non-identity transformation under substantially weaker technical conditions.

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

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

*where $q \in \mathcal{Q} = [\varepsilon_0, 1 - \varepsilon_0]$ denotes the quantile position, with $\varepsilon_0 > 0$. Then, $\eta(\mu_0(\mathbf{x}, q))$ is the $q$-th conditional quantile function of $Y$ given $\boldsymbol{X} = \mathbf{x}$, and the partitioning-based quantile regression* *estimator is $\eta(\widehat{\mu}(\mathbf{x},q))$ as defined in (1.2). In the classical case, $\eta(\cdot)$ is the identity function, but our theory accommodates other transformations. Interest lies on the quantile process estimator $(\eta(\widehat{\mu}(\mathbf{x},q)) : (\mathbf{x},q) \in \mathcal{X} \times \mathcal{Q})$, which can be used to characterize heterogeneous effects of covariates on the outcome distribution and to conduct specification testing. See Section 7.1 for our main results, and for Section 8 for results on transformations.* ▲

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

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

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

*where we can, for example, use the complementary log-log link $\eta(a) = 1 - \exp(-\exp(a))$. Estimand of interest are $\eta(\mu_0(\mathbf{x}))$, which corresponds to the conditional distribution function of $Y$ given $\boldsymbol{X} = \mathbf{x}$ (i.e., $F_{Y|\boldsymbol{X}}(q|\mathbf{x}) = \mathbb{E}[\mathbb{1}(Y \leq q)|\boldsymbol{X} = \mathbf{x}]$), and derivatives thereof. Uniform estimation and inference results based on $(\eta(\widehat{\mu}(\mathbf{x},q)) : (\mathbf{x},q) \in \mathcal{X} \times \mathcal{Q})$ are useful for a variety of purposes, including heterogeneous treatment effect estimation and specification testing. See Section 7.2 for our main results, and Section 8 for transformations of $\widehat{\mu}(\cdot,\cdot)$.* ▲

The next example considers a novel class of nonparametric partitioning-based estimators in the context of robust $L_p$ regression. For comparison and background, see [26] for the latest results on parametric ($K$ fixed) $L_p$ regression estimation.

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

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

*for a fixed $p > 0$. In particular, $p = 2$ leads to nonlinear least squares, and $p = 1$ leads to generalized least absolute deviations, when a non-identity transformation function $\eta$ is used. The estimand of interest is usually the transformed regression function $\eta(\mu_0(\mathbf{x}))$, which needs to be interpreted in context. Our general theory applies to any choice $p \in [1,2]$, delivering uniform estimation and inference methods based on $(\widehat{\mu}(\mathbf{x}) : \mathbf{x} \in \mathcal{X})$, and transformations thereof. See Section 7.3 for our main results, and Section 8 for transformations.* ▲

The final example considers (nonparametric) Generalized Linear Models [30]. For specificity, we focus on (quasi-)maximum likelihood logistic regression, but our results cover many other examples within this class such as regression models with limited dependent variables (e.g., Poisson, fractional, censored and truncation regression).

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

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

*with $\mathcal{Y} = \{0,1\}$. The estimand $\eta(\mu_0(\mathbf{x}))$ characterizes the conditional probability of $Y = 1$ given $\boldsymbol{X} = \mathbf{x}$. See Section 7.4 for uniform estimation and inference methods based on $(\widehat{\mu}(\mathbf{x}) : \mathbf{x} \in \mathcal{X})$, and Section 8 for transformations thereof. Furthermore, our results also cover other related quasi-maximum likelihood (and non-linear least squares) problems such as fractional regression where $\mathcal{Y} = [0,1]$.* ▲

The four examples introduced so far cover the main different settings of interest from a technical perspective. To be precise, in Example 1 uniformity over $\mathcal{X} \times \mathcal{Q}$ is of interest, and the loss function is non-smooth as a function of $\mathbf{x} \in \mathcal{X}$ but smooth as a function $q \in \mathcal{Q}$. Example 2 is the "dual" of Example 1 in the sense that uniformity over $\mathcal{X} \times \mathcal{Q}$ is also of interest, but now the loss function is smooth as a function of $\mathbf{x} \in \mathcal{X}$ and non-smooth as a function $q \in \mathcal{Q}$. In Example 3 only uniformity over $\mathcal{X}$ is of interest because $q \in \mathcal{Q}$ is not present in the loss function, but its smoothness depends on $p \in [0, 1]$; the a. e. derivative of $\eta \mapsto \rho(y, \eta)$ ranges from discontinuous ($p = 1$), to Hölder continuous ($p \in (1, 2)$), to linear ($p = 2$). Likewise, Example 4 only involves uniformity over $\mathcal{X}$ because $q \in \mathcal{Q}$ is not present in the loss function, but now the loss function is smooth and well-behaved; this last example serves as a benchmark for our theoretical development, in addition to being of broad practical interest. All of the examples above have a convex loss function when $\eta(u) = u$, but can be non-convex when $\eta(\cdot)$ is not the identity function: our theoretical work will either explicitly take into account the presence of non-convexity, or circumvent this challenge altogether.

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

# 3 Assumptions

Our theoretical work proceeds under five general assumptions. The first three assumptions concern the data generating process and the loss function, the next two assumptions concern the partitioning-based estimation method, and the last assumption links the statistical model and partition-based approximation.

## 3.1 Statistical Model

Our first assumption imposes basic regularity on the model.

**Assumption 1** (Data Generating Process)**.**

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

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

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

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

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

Assumption 1 imposes standard conditions from the nonparametric regression literature, including basic support and smoothness restrictions. Minimal additional regularity is imposed to accommodate uniformity over $q \in \mathcal{Q}$, and different types of conditional distributions of $Y|\boldsymbol{X}$ (e. g., absolutely continuous, discrete or mixed) are allowed.

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

**Assumption 2** (Loss Function)**.**

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

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

(iii) *$\psi(y, \eta(\theta); q) = \varphi(y, \eta(\theta); q)\varpi(\theta)$, where $\eta(\cdot)$ is strictly monotonic and twice continuously differentiable, and $\varpi(\cdot)$ is continuously differentiable and strictly positive or negative. Furthermore, for some fixed constant $\alpha \in (0, 1]$, for any $(\mathbf{x}, q) \in \mathcal{X} \times \mathcal{Q}$, and a pair of points $\zeta_1, \zeta_2 \in B_q(\mathbf{x})$, $\varphi(\cdot)$ satisfies the following (constants hidden in $\lesssim$ do not depend on $\mathbf{x}$, $q$, $\zeta_1$, $\zeta_2$):*

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

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

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

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

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

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

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

*and*

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

Assumption 2 is carefully crafted to accommodate all the examples discussed in Section 2, and many others. Part (i) allows for different degrees of smoothness in the loss function, assuming only absolute continuity (with respect to $\eta$). Part (ii) formalizes the idea that $\mu_0(\mathbf{x}, q)$ may not be a unique (global) minimizer in (1.1), and consequently it is only required to be a root of the (conditional) first-order condition; the rest of the assumptions in that part are mild regularity conditions. In some applications, $\mu_0(\mathbf{x}; q)$ can be the unique minimizer; see, for example, [28], [29], and references therein.

Part (iii) of Assumption 2 imposes additional structure on the a. e. first derivative of the loss function, allowing for all types of outcome data (discrete, mixed, and continuous) and rescalings emerging in some of the motivating examples. Importantly, this part characterizes precisely the role of (Hölder) smoothness, which is controlled by the parameter $\alpha \in (0, 1]$. We illustrate the full power of this general assumption in Section 7, where $\alpha = 1$ in Examples 1 and 2, $\alpha = p - 1$ in Example 3 when $p > 1$, and $\alpha = 1$ in Example 4. Furthermore, the special multiplicative structure of $\psi(\cdot)$ plays a key role in Example 4, as shown in Section 7.4 and in the supplemental appendix. Finally, part (iv) of Assumption 2 collects mild regularity conditions on the smoothed-out a. e. derivative of the loss function.

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

**Assumption 3** (Simplified Setup)**.**

(i) *$q \mapsto \mu_0(\mathbf{x}, q)$ is non-decreasing, and $\rho(y, \eta; q) = \sum_{j=1}^{4} \omega_j \rho_j(y, \eta; q)$, where $(\omega_1, \omega_2, \omega_3, \omega_4)$ are constants, and the functions $(\rho_1(\cdot), \rho_2(\cdot), \rho_3(\cdot), \rho_4(\cdot))$ are of the following types.*

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

*where $f_j$ are fixed continuous functions, $D_j$ are universal constants, and $\eta \mapsto \mathcal{T}(y; \eta)$ is differentiable.*

(ii) *If $\omega_1 \neq 0$, then the following conditions hold:*

(a) *$\eta \mapsto \mathbb{E}[\tau(y_i, \eta)|\mathbf{x}_i = \mathbf{x}]$ is differentiable, where $\tau(y, \eta) = \frac{\partial}{\partial \eta}\mathcal{T}(y, \eta)$.*

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

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

(iii) *If $\omega_2 \neq 0$, then $F_{Y|X}$ is differentiable with a Lebesgue density $f_{Y|X}$, and $f_{Y|X}$ is continuous in both arguments and $y \mapsto f_{Y|X}(y|\mathbf{x})$ is continuously differentiable.*

This third assumption imposes an additional weak monotonicity condition on $\mu_0(\mathbf{x}, q)$ as a function of $q$, which is compatible with all the examples in Section 2. Finally, the key restriction emerging from Assumption 3 is on the structure of the loss function, which allows for linear combinations of smooth loss functions of $y$ and $\eta$, and non-smooth loss functions involving indicator functions of either $y$ and $\eta$, or $y$ and $q$. These restrictions are still general enough to cover all examples discussed before: the loss function in Example 1 is a combination of Type I and Type III functions with $f_1$ and $f_3$ being linear functions of $y$; the loss function in Example 2 is a combination of Type II and Type IV functions; the loss function in Example 3 is of Type IV for $p > 1$ and corresponds to median regression for $p = 1$; and the loss function in Example 4 is usually a Type IV function. See Section 7 and the supplemental appendix for details.

### 3.2 Partitioning-Based Methodology

The next two assumptions concern the regularities of the partition and the local basis constructed on it, which are the core ingredients for the partition-based M-estimator in (1.2). The conditions assumed are the same as those imposed for the special case of least squares partitioning-based series methods [22, 10].

**Assumption 4** (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 $h = \max\{\mathrm{diam}(\delta) : \delta \in \Delta\}$, assume $h = o(1)$ and $\log(1/h) \lesssim \log n$.*

Assumption 4 requires the partition $\Delta$ be quasi-uniform: the elements in the partition $\Delta$ do not differ too much in size asymptotically. As a consequence, we can use the maximum diameter $h$ as a universal measure of mesh sizes.

The next assumption requires the basis be "locally supported", non-collinear, and bounded in a proper sense. A function $p(\cdot)$ on $\mathcal{X}$ is *active* on $\delta \in \Delta$ if it is not identically zero on $\delta$.

**Assumption 5** (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 $\mathbf{a} = (a_1, \ldots, a_K)^\mathsf{T} \in \mathbb{R}^K$,*

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

(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 $\delta$.*

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

Assumptions 4 and 5 implicitly relate the number of approximating series terms, the number of cells in $\Delta$, and the maximum mesh size: $K \asymp \bar{\kappa} \asymp h^{-d}$. These conditions are easily verified for local nonparametric methods such as piecewise polynomial regression, splines, and compactly supported wavelets: see [22], [9], [3], [10], [8], and references therein. Furthermore, the conditions can be used to justify employing estimators based on tree methodology: $\boldsymbol{X}$-adaptive tree constructions [16] and certain other recursive adaptive partitioning methods [37] may be accommodated under additional restrictions (e.g., via sample splitting).

### 3.3 Uniform Approximation

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

**Assumption 6** (Approximation Error)**.** *There exists a vector of coefficients $\boldsymbol{\beta}_0(q) \in \mathbb{R}^K$ such that for all $\boldsymbol{\varsigma}$ satisfying $|\boldsymbol{\varsigma}| \leq \varsigma$ in Assumption 5,*

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

The vector $\boldsymbol{\beta}_0(q)$ can be viewed as a pseudo-true value, and does not have to be unique. The existence of such $\boldsymbol{\beta}_0(q)$ can be established using approximation theory or related methods, and necessarily depends on the specific underlying structure of the statistical model (determining $\mu_0(\mathbf{x}, q)$) and the partitioning-based method (determining $\mathbf{p}^{(\boldsymbol{\varsigma})}(\mathbf{x}; \Delta, m)$). For more disucssion, see [22], [9], [3], [10], [8], and references therein.

## 4 Consistency

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

### 4.1 Convex Loss Function

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

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

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

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

*Then*

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

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

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

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

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

These restrictions are comparable to, or improve upon, the existing results in the literature. For example, for quantile regression with tensor-product of $B$-splines, Corollary 1 of [2] implies that the $L_2$-consistency (4.3) can be obtained under $1/(nh^{2d}) \asymp K^2/n = o(1)$, and Corollary 2 therein implies that the uniform consistency (4.2) can be obtained under $1/(nh^{4d}) \asymp K^4/n = o(1)$. In contrast, noting the generalized residual from quantile regression has a sub-Gaussian envelope, we only need $1/(nh^{2d}) \asymp K^2/n = o(1)$ to establish both kinds of consistency, substantially improving the requirements for uniform consistency.

Moreover, when an unconnected basis is used or the loss function is strongly convex and smooth (e.g., the square loss for mean regression), the weakest possible restriction suffices: $1/(nh^d) \asymp K/n = o(1)$, up to polylog$(n)$ terms. See Section 4.3 for details. This is on par with the existing best results for series-based least squares regression [3, 10] in the literature, but our results can also cover other nonlinear cases such as piecewise-polynomial-based quantile, nonlinear, or robust regression. Whether it is possible to establish consistency under the weakest condition $1/(nh^d) \asymp K/n = o(1)$ for general partitioning-based $M$-estimators remains an open question.

Another important feature of Lemma 1 is that *no* constraints are imposed on the coefficients in the optimization procedure, which allows the estimation space to be, for example, piecewise polynomials. In contrast, many studies of series (or sieve) methods restrict the functions in the estimation space to satisfy certain smoothness conditions, e.g., Lipschitz continuity, to derive the uniform consistency [e.g., 15], which may contradict common practical implementations.

## 4.2 Non-Convex Loss Function

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

A proof of consistency for the unconstrained estimator in (1.2) with non-convex loss function is not available, but we are able to establish consistency of a minimally regularized $M$-estimator.

Specifically, we add a *fixed* "box" constraint: for some fixed constant $R > 0$,

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

In the supplemental appendix we show that the pseudo-true coefficients $\boldsymbol{\beta}_0(q)$ from Assumption 6 are bounded in sup-norm by a universal constant: $\sup_{q\in\mathcal{Q}} \|\boldsymbol{\beta}_0(q)\|_\infty \lesssim 1$ (because $\mathbf{p}(\mathbf{x})^\intercal \boldsymbol{\beta}_0(q)$ have to be close to $\mu_0(\mathbf{x}, q)$ which is uniformly bounded). Therefore, we can always choose a sufficiently large constant $R$ in the optimization procedure, making the box constraint set contain $\boldsymbol{\beta}_0(q)$ as an interior point. The following lemma, proven in the supplemental appendix (Lemma D.5), establishes consistency of the constrained estimator.

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

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

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

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

Two additional restrictions are imposed in this lemma. The first one, $R \geq 2\sup_{q\in\mathcal{Q}} \|\boldsymbol{\beta}_0(q)\|_\infty$, can be theoretically justified by Lemma D.4 in the supplemental appendix, and in practice a large enough $R$ is recommended. The other restriction concerns a lower bound for $\Psi_1$, which implies that the (population) loss function is strongly convex in a neighborhood of the true value $\eta(\mu_0(\mathbf{x}, q))$, making the (constrained) minimizer well defined. (Note that this condition does *not* break because of the shape of $\eta(\cdot)$, in contrast with the convexity of $\rho(y, \eta(\theta); q)$ in $\theta$.) The other conditions in this lemma are the same as those in the convex case, and thus all improvements discussed before also apply to this second consistency result.

### 4.3 Weaker Conditions for Special Cases

In the supplemental appendix we provide additional consistency results for two special cases of theoretical and practical relevance:

- the loss function is strongly convex and smooth (i.e., the second "derivative" of $\rho(y, \eta(\cdot); q)$ is bounded and bounded away from zero), or

- an "unconnected" basis (i.e., each basis function is supported on a single cell of $\Delta$) is employed.

The first case covers, for example, the usual least squares regression, and the second case covers any $M$-estimation problem based on, for example, piecewise polynomials. Notably, in these two scenarios the key consistency result $\|\widehat{\boldsymbol{\beta}}(q) - \boldsymbol{\beta}(q)\|_\infty = o_{\mathbb{P}}(1)$ can be established for any $m$ and $d$, so the prior requirement $m > d/2$ is unnecessary, under the seemingly minimal rate restrictions $1/(nh^d) \asymp K/n = o(1)$ in the sub-Gaussian case, and $1/(nh^{\frac{\nu}{\nu-1}d}) \asymp K^{\frac{\nu}{\nu-1}}/n = o(1)$ in the bounded $\nu$-th moment case, up to polylog$(n)$ terms. See Section D in the supplemental appendix for more details.

# 5 Bahadur Representation and Convergence Rates

This section presents our first main result for uniform estimation and inference: a novel uniform Bahadur representation for partitioning-based $M$-estimators. The Bahadur representation is

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

where

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

The following theorem takes the sup-norm consistency of the coefficient estimators $\widehat{\boldsymbol{\beta}}(q)$ as a high-level assumption, and thus avoids imposing any of the specific sufficient conditions discussed in Section 4. The proof is provided in the supplemental appendix (Theorem E.10).

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

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

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

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

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

*Then*

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

*If, in addition,* $\sup_{y\in\mathcal{Y},q\in\mathcal{Q}}|\varphi(y,\eta(\zeta_1);q)-\varphi(y,\eta(\zeta_2);q)|\lesssim|\zeta_1-\zeta_2|^\alpha$ *without any restrictions on* $y$ *in Assumption 2(iii), then*

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

The Bahadur representation (5.1) applies to the case where the "derivative" $\psi(\cdot,\cdot;q)$ of the loss function may be discontinuous. One typical example is quantile regression (Example 1), where the "derivative" $\psi(y,\eta;q)=\mathbb{1}(y-\eta<0)-q$, as a function of $(y-\eta)$, is piecewise constant with a jump at zero. In this case we can let $\alpha=1$, and (5.1) implies that the order of the remainder in the Bahadur representation for partitioning-based quantile regression is $O(h^{-|\mathbf{v}|}(nh^d)^{-3/4}+h^{m-|\mathbf{v}|})$, up to polylog$(n)$ terms. Another example is $L_p$ regression with $p\in(1,2)$, Example 3, where the derivative of the loss function is given by $\psi(y,\eta)\equiv\psi(y-\eta)=p|y-\eta|^{p-1}\mathrm{sgn}(\eta-y)$ with $\mathrm{sgn}(\cdot)$ denoting the sign function. As a function of $(y-\eta)$, $\psi(\cdot)$ is $\alpha$-Hölder on $[0,\infty)$ or $(-\infty,0]$ for all $\alpha\in(0,p-1]$ but not for $\alpha>p-1$. Thus, (5.1) applies with the order of the remainder depending on $p$, which is the same as that for quantile regression when $p\ge3/2$.

On the other hand, the Bahadur representation (5.2) applies to the case where the "derivative" of the loss is a continuous function of $(y-\eta)$. Nonlinear least squares regression (Example 2 and quasi-maximum likelihood estimation of generalized linear models (Example 4) fall into this category with the Hölder parameter $\alpha=1$. In such cases, (5.2) implies that the order of the remainder in the Bahadur representation is $O(h^{-|\mathbf{v}|}(nh^d)^{-1}+h^{m-|\mathbf{v}|})$, up to polylog$(n)$ terms, which is a tighter upper bound than that implied by (5.1). See Section 7 and the supplemental appendix for more details.

In both cases, the remainder of the Bahadur representation consists of two terms. The last term $h^{m-|\mathbf{v}|}$ corresponds to the error from approximating the function $\mu_0$ using the partitioning basis (c.f., Assumption 6), whereas the first term arises from the (potential) nonlinearity underlying the $M$-estimation, and reflects explicitly the role of non-smoothness of the loss function. Specifically, when the "derivative" of the loss function has discontinuity points, the order of the remainder in (5.1) is greater than that in the continuous case (5.2); with a smaller Hölder parameter $\alpha$, the order of the remainder in both cases could increase.

## 5.1 Rates of Convergence

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

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

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

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

*Then*

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

By setting $h \asymp \left(\frac{\log n}{n}\right)^{\frac{1}{2m+d}}$, Corollary 1 implies that the partitioning-based $M$-estimator can achieve the uniform convergence rate $\left(\frac{\log n}{n}\right)^{\frac{m}{2m+d}}$. This matches the optimal rate of convergence in sup-norm for nonparametric estimators of the conditional mean [34] and conditional quantiles [11]. In this sense, the rate of convergence in Corollary 1 is optimal and cannot be further improved at our level of generality.

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

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

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

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

*Then*

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

By setting $h \asymp n^{-\frac{1}{2m+d}}$, Corollary 2 implies that the partitioning-based $M$-estimator can also achieve the $L_2$ convergence rate $n^{-\frac{m}{2m+d}}$, uniformly over $\mathcal{Q}$, thereby matching the optimal rate of convergence in $L_2$-norm for nonparametric estimators of conditional means [33] and conditional quantiles [11].

The convergence rates in (5.3) and (5.4) capture two contributions: the first term reflects the variance of the estimator, while the second term arises from the error of approximating the unknown $\mu_0$ by the partitioning basis. In the case of Corollary 2, it is possible to further leverage Theorem 1 to obtain a precise first-order asymptotic approximation for the integrated mean square error of the partitioning-based $M$-estimator, uniformly over $\mathcal{Q}$, which in turn could be used to develop plug-in asymptotically optimal rules for selecting $K \asymp h^{-d}$. See, for example, Theorem 4.2 in [10] for a similar result in the special case of square loss and identity transformation functions. We do not pursue this result here for brevity.

## 5.2 Comparison with Existing Results

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

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

### Mean Regression

The usual mean regression is a special case of our general setup where $\rho(\cdot)$ is the square loss, $\eta(\cdot)$ is the identity link, and $\mathcal{Q}$ is a singleton. Bahadur representations for this special case were established by [3] and [10]. Since the derivative of the square loss for mean regression is linear, the first term in (5.1) or (5.2) does not show up in the uniform linearization of least squares series estimators. See, for example, Lemma SA-4.2 of [10]; note that $R_{1n,q}$ defined therein has been implicitly included in the leading variance term in (5.2) above. Theorem 1 substantially extends these prior results to other nonlinear settings, under minimal additional conditions.

Finally, Corollaries 1 and 2 demonstrate the convergence rate optimality of general partitioning-based series $M$-estimation, recovering in particular known results for mean regression [3, 10] under essentially the same minimal conditions.

### Quantile Regression

Theorem 1 improves upon prior theoretical results for nonparametric series quantile regression estimators. The most recent advance in this literature is due to [2], which establishes a uniform linear approximation for general series-based quantile regression estimators. In comparison, we exploit the "local support" feature of the partitioning basis, and make improvements in (at least) four aspects. To summarize these improvements without additional cumbersome notation, we set $\mathbf{v} = \mathbf{0}$ and ignore the smoothing bias $h^m$ in the Bahadur approximation remainders.

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

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

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

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

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

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

### Other Nonparametric Smoothing Methods

[25] establishes a similar Bahadur representation for kernel-based $M$-estimators using weakly stationary time series data. They consider a special case of our setup in Assumption 2: their loss function class $\mathcal{Q}$ is a singleton, $\eta$ is an identity function, and the "derivative" of the loss can be written as $\psi(y, \eta) \equiv \psi(y - \eta)$ and is assumed to be piecewise Lipschitz continuous. In a comparable cross-sectional context with $\alpha = 1$ and $\mathbf{v} = \mathbf{0}$, the order of the remainder in our Bahadur representation (5.1) is $O((nh^d)^{-3/4})$, up to polylog$(n)$ terms, and thus Theorem 1 matches their approximation error up to a minor difference in $\log n$ terms. Taking $nh^d$ to be the effective sample size, the approximation rate can not be further improved at this level of generality, and hence Theorem 1 establishes that the partitioning-based series $M$-estimator in (1.2) can achieve the same

best Bahadur approximation as local polynomial kernel methods, up to polylog$(n)$ terms.

Furthermore, compared to [25] or other similar contributions in the literature, Theorem 1 exhibits (at least) two novel features. First, the Bahadur representations (5.1) and (5.2) hold uniformly not only over the evaluation point $x \in \mathcal{X}$, but also over the loss function index $q \in \mathcal{Q}$, which may be important, for example, to study simultaneous quantile regression where the *entire* conditional quantile process may be of interest. Second, Theorem 1 also covers the more general setup where the "derivative" function may exhibit different degrees of smoothness, reflected by discontinuity points and/or the Hölder parameter $\alpha$, or admits a more complex structure so that $\psi(y, \eta; q)$ cannot be written as $\psi(y - \eta; q)$. Thus, we cover more examples such as distribution regression (Example 2) and $L_p$ regression with $p \in (1, 2)$ (Example 3). Finally, [25] does not discuss convergence rates as we do in Corollaries 1 and 2.

In the context of nonparametric penalized smoothing spline methods, [32] also establishes a uniform Bahadur representation (and other results) that can be compared to Theorem 1. However, their paper imposes more stringent assumptions and hence cover a smaller class of settings: using our notation, they assume that (i) $\mathcal{Q}$ is a singleton so their uniformity is only over $\mathcal{X}$; (ii) $d = 1$ so they consider only scalar covariate $\mathbf{x}_i$; and (iii) $\rho(\cdot, \eta(\cdot))$ is smooth so they rule out many important examples such as quantile regression, and Tukey and Huber regression. Furthermore, their results do not take explicit advantage of specific moment and boundedness conditions, or the structure of the nonparametric estimator, and instead impose the generic side condition $nh^2 \to \infty$, which is comparable to our condition $K^2/n \to \infty$, up to polylog$(n)$ terms. Most importantly, in the closest comparable case ($d = 1$, $\alpha = 1$, and $\mathbf{v} = \mathbf{0}$), and only focusing on the variance component for simplicity, the order of the remainder in their uniform Bahadur representation (a combination of Theorem 3.4 and Lemma 3.1 in [32]) is $O((nh)^{-1}h^{-(6m-1)/(4m)})$, while (5.2) in Theorem 1 gives the optimal result $O((nh)^{-1})$, thereby demonstrating a substantial improvement over their result. Finally, as for convergence rates, Proposition 3.3 in [32] and our Corollary 2 are essentially equivalent, both delivering optimal mean square convergence. They do not explicitly discuss uniform convergence rates as we do in Corollary 1.

# 6 Strong Approximation and Uniform Inference

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

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

where

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

Accordingly, we define a feasible variance estimator as

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

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

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

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

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

Employing Theorem 1, or more precise arguments under slightly weaker conditions, it is easy to show that $T(\mathbf{x},q)$ converges in distribution to $\mathsf{N}(0,1)$ for each $(\mathbf{x},q)\in\mathcal{X}\times\mathcal{Q}$. However, the stochastic process $(T(\mathbf{x},q):(\mathbf{x},q)\in\mathcal{X}\times\mathcal{Q})$ is generally not asymptotically tight and, therefore, does not converge weakly in $\ell^\infty(\mathcal{X}\times\mathcal{Q})$, where $\ell^\infty(\mathcal{X}\times\mathcal{Q})$ denotes the set of all (uniformly) bounded real functions on $\mathcal{X}\times\mathcal{Q}$ equipped with uniform norm [35]. Nevertheless, we can carefully construct a Gaussian process, in a possibly enlarged probability space, that approximates the entire process $T(\cdot)$ sufficiently fast, which then can be used to characterize the finite sample distribution of the function $M$-estimator $\widehat{\mu}^{(\mathbf{v})}(\cdot)$.

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

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

It follows that, conditional on $\{\mathbf{x}_i\}_{i=1}^n$, the randomness of $t(\mathbf{x},q)$ comes exclusively from the $K$-dimensional vector $\mathbb{G}_n[\mathbf{p}(\mathbf{x}_i)\eta^{(1)}(\mu_0(\mathbf{x}_i,q))\psi(y_i,\eta(\mu_0(\mathbf{x}_i,q));q)]$. Thus, our proof strategy is to further "discretize" this vector with respect to $q\in\mathcal{Q}$, and then apply Yurinskii' coupling [36] to construct a (conditional) Gaussian process that is close to the original $t$-statistic process $T(\mathbf{x},q)$ uniformly over both $\mathbf{x}\in\mathcal{X}$ and $q\in\mathcal{Q}$. Our construction leverages a conditional Strassen's theorem [13, Theorem B.2] to generalize prior coupling results [2, Lemma 36]. See Section F in the supplemental appendix for details.

Our strong approximation approach is formalized in the next theorem. We employ high-level conditions to ease the exposition, but those conditions can be verified using Corollaries 1 and 2, and Theorem 1, as well as using the more general results in the supplemental appendix. Let $r_{\mathtt{UC}}$, $r_{\mathtt{BR}}$, $r_{\mathtt{VC}}$, and $r_{\mathtt{SA}}$ be positive non-random sequences as $n\to\infty$. The proof is available in the supplemental appendix (Corollary F.5).

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

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

(ii) $\sup_{q\in\mathcal{Q}}\sup_{\mathbf{x}\in\mathcal{X}}\big|\widehat{\mu}(\mathbf{x},q)-\mu_0(\mathbf{x},q)-\mathsf{L}(\mathbf{x},q)\big|\lesssim_{\mathbb{P}} r_{\mathtt{BR}}$.

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

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

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

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

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

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

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

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

The speed of strong approximation in Theorem 2 is determined by four factors: the uniform convergence rate $r_{\mathtt{UC}}$, the order of the remainder in the Bahadur representation $r_{\mathtt{BR}}$, the convergence rate $r_{\mathtt{VC}}$ of the variance estimator $\widehat{\Omega}_{\mathbf{v}}$, and the strong approximation rate $r_{\mathtt{SA}}$. Therefore, our strong approximation results are established at a high-level of generality, building on our prior theoretical results: Corollary 1 for $r_{\mathtt{UC}}$, and Theorem 1 for $r_{\mathtt{BR}}$, while $r_{\mathtt{VC}}$ is a high-level condition that needs to be verified on a case-by-case basis (see Sections 6.2 and 7 for more discussion.

With respect to strong approximation rate, Theorem 2 lays down two versions of lower bounds on $r_{\mathtt{SA}}$, depending on the tail behavior of the generalized residuals. Such restrictions may not be optimal, but are still weak enough to cover almost all partition size choices commonly used in practice. In particular, the restriction on $r_{\mathtt{SA}}$ in Theorem 2 allows for the MSE-optimal choice $h \asymp n^{-\frac{1}{2m+d}}$ in all cases except the unidimensional Haar basis approximation ($m = d = 1$); there is also room for undersmoothing in order to make the smoothing bias negligible in all cases but $m = d = 1$. The strong approximation for one dimensional partitioning-based series estimators in the special case of square loss and identity transformation functions was studied in [10, 8] via a different coupling strategy, which delivered tighter approximation results allowing for an MSE-optimal choice of $h$. We conjecture those techniques could be adapted to cover the case $m = d = 1$ for general partitioning-based $M$-estimator in (1.2), but we do not pursue this line of research here because it would require a different theoretical treatment.

### 6.1 Comparison with Existing Results

Theorem 2 is the first to establish strong approximation results for general partitioning-based $M$-estimators at the level of generality considered in this paper. In the prior literature, similar results are usually available only in specific scenarios such as least squares regression or quantile regression. To be more precise, in the least squares context ($\mathcal{Q}$ is a singleton), [10] establishes uniform inference theory for univariate regression ($d = 1$) and multivariate regression ($d > 1$) separately via different strong approximation methods. In particular, when $d > 1$, the same Yurinskii coupling technique is employed to obtain strong approximation for $t$-statistic processes, leading to similar rate restrictions on $h$. Theorem 2 is a substantial generalization of results therein, not just covering other loss functions, but also providing distributional approximation uniformly over loss function index $q \in \mathcal{Q}$.

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

Finally, [32] establishes uniform inference results for nonparametric penalized smoothing spline M-estimators. As mentioned before, their work is more specialized because they assume that $\mathcal{Q}$ is a singleton, $d = 1$, and $\rho(\cdot, \eta(\cdot))$ is smooth. Furthermore, their approach to constructing valid confidence bands and related uniform inference methods relies on approximating the suprema of the stochastic process directly via extreme value theory [32, Theorem 5.1], which leads to substantially slower approximation rates and requires stronger assumption and side rate conditions; see [3] and [10] for more discussion in the context of nonparametric least squares series estimation. In contrast, Theorem 2, and our related uniform inference methods, provide a pre-asymptotic approximation with better finite sample properties, faster approximation rates, and weaker regularity conditions.

### 6.2 Implementation and Feasible Uniform Inference

The approximation process $(Z(\mathbf{x}, q) : (\mathbf{x}, q) \in \mathcal{X} \times \mathcal{Q})$ in Theorem 2 is still infeasible since its covariance structure contains unknowns. In practice, two general strategies can be used: *Plug-in method* or *Bootstrap method*. We discuss in detail the former approach, which we later employ in Section 7 to develop feasible uniform inference in the context of our motivating examples. We also briefly review the latter approach at the end of this section, but leave its theoretical analysis for future research.

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

The covariance structure of the process $Z(\mathbf{x}, q)$ in Theorem 2 is

$$\mathbb{E}[Z(\mathbf{x}, q)Z(\tilde{\mathbf{x}}, \tilde{q})|\boldsymbol{X}_n] = \ell(\mathbf{x}, q)^\mathsf{T}\bar{\mathbf{\Sigma}}_{q,\tilde{q}}\ell(\tilde{\mathbf{x}}, \tilde{q}), \qquad \ell(\mathbf{x}, q)^\mathsf{T} = \frac{\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\mathsf{T}\bar{\boldsymbol{Q}}_q^{-1}}{\sqrt{\bar{\Omega}_{\mathbf{v}}(\mathbf{x}, q)}} \tag{6.1}$$

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

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

with $S_{q,\tilde{q}}(\mathbf{x}) = \mathbb{E}\big[\psi(y_i, \eta(\mu_0(\mathbf{x}_i, q)); q)\psi(y_i, \eta(\mu_0(\mathbf{x}_i, \tilde{q})); \tilde{q})|\mathbf{x}_i = \mathbf{x}\big]$. Given context-specific estimates $\widehat{\boldsymbol{Q}}_q$ and $\mathbf{x} \mapsto \widehat{S}_{q,\tilde{q}}(\mathbf{x})$, we can put

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

and $\widehat{\Omega}_{\mathbf{v}}(\mathbf{x}, q) = \mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\mathsf{T}\widehat{\boldsymbol{Q}}_q^{-1}\widehat{\mathbf{\Sigma}}_{q,q}\widehat{\boldsymbol{Q}}_q^{-1}\mathbf{p}^{(\mathbf{v})}(\mathbf{x})$ as above. The discussion about finding these prerequisite estimates $\widehat{\boldsymbol{Q}}_q$ and $\mathbf{x} \mapsto \widehat{S}_{q,\tilde{q}}(\mathbf{x})$ is deferred to Section 7. Then, a feasible Gaussian approximation $\widehat{Z}(\mathbf{x}, q)$ can be constructed as a Gaussian process conditional on the data $\boldsymbol{D}_n = ((y_1, \mathbf{x}_1), \cdots, (y_n, \mathbf{x}_n))$ with conditional covariance structure

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

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

Once we have a feasible process $\widehat{Z}(\mathbf{x}, q)$ that is "close" to a copy of $Z(\mathbf{x}, q)$ uniformly over $\mathcal{X} \times \mathcal{Q}$ conditional on the data, then $\widehat{Z}(\mathbf{x}, q)$ can be used to make inference on the entire function

$\mu_0(\mathbf{x}, q)$, and functionals thereof. For example, our strong approximation results can be converted to convergence of Kolmogorov distance between the distributions of $\sup_{q\in\mathcal{Q}}\sup_{\mathbf{x}\in\mathcal{X}}|T(\mathbf{x}, q)|$ and its feasible Gaussian approximation $\sup_{q\in\mathcal{Q}}\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{Z}(\mathbf{x}, q)|$. Under some technical assumptions (see Theorem G.7 for a formal statement),

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\Big(\sup_{q\in\mathcal{Q}}\sup_{\mathbf{x}\in\mathcal{X}}|T(\mathbf{x}, q)| \leq u\Big) - \mathbb{P}\Big(\sup_{q\in\mathcal{Q}}\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{Z}(\mathbf{x}, q)| \leq u\Big)\right| = o_{\mathbb{P}}(1).$$

In turn, this result leads to the asymptotically valid uniform confidence band for $\mu_0^{(\mathbf{v})}$ given by

$$\mathsf{CB}_{1-\alpha}(\mathbf{x}, q) = \Big[\widehat{\mu}^{(\mathbf{v})}(\mathbf{x}, q) \pm \mathfrak{c}_{1-\alpha}\sqrt{\widehat{\Omega}_{\mathbf{v}}(\mathbf{x}, q)}\Big] \tag{6.2}$$

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

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

provided the smoothing (or misspecification) bias relative to the standard error of the estimator is small, which could be achieved by undersmoothing, bias correction [20], simply ignoring the bias [21], robust bias correction [6, 7], or the Lepskii's method [27, 5], among other possibilities. Thus, under regularity conditions, it can be shown that the confidence band (6.2) covers $\mu_0^{(\mathbf{v})}$ with probability approximately $1 - \alpha$ in large samples, that is,

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

We illustrate these ideas, and general blueprint, in the next section when developing estimation and inference methods for our four motivating examples introduced in Section 2. See also the supplemental appendix.

An alternative strategy to approximate the distribution of the stochastic process $Z(\cdot, \cdot)$ in Theorem 2 is bootstrapping. Various bootstrap-based methods are available to simulate a feasible process $\widehat{Z}(\mathbf{x}, q)$ conditional on the data, which might help avoid the estimation of a potentially complex covariance structure. The bootstrap design has to ensure $\widehat{Z}(\mathbf{x}, q)$ can be approximated by the same process $Z(\mathbf{x}, q)$. To illustrate the main idea consider the weighted bootstrap method, which relies on a set of weights $(e_1, \cdots, e_n)$ that are i.i.d. draws from a distribution with mean 1 and variance 1, and are independent of the data $\boldsymbol{D}_n$. For each draw of such weights, define a weighted bootstrap draw of the partitioning-based $M$-estimator by

$$\widehat{\mu}^*(\mathbf{x}, q) = \mathbf{p}(\mathbf{x})^\intercal\widehat{\boldsymbol{\beta}}^*(q), \qquad \widehat{\boldsymbol{\beta}}^*(q) \in \arg\min_{\mathbf{b}\in\mathcal{B}}\sum_{i=1}^n e_i\rho(y_i, \eta(\mathbf{p}(\mathbf{x}_i^\intercal\mathbf{b})); q).$$

Analogues of Theorems 1 and 2, and other theoretical results, can be established for the bootstrap process $(\widehat{\mu}^*(\mathbf{x}, q) - \widehat{\mu}(\mathbf{x}, q) : (\mathbf{x}, q) \in \mathcal{X}\times\mathcal{Q})$ or a Studentized version thereof, leveraging our general technical results in the supplemental appendix. Consequently, conditional on the data $\boldsymbol{D}_n$, and under minimal additional regularity conditions, the bootstrap process will be close in distribution to same (conditional) Gaussian process $Z(\mathbf{x}, q)$, and therefore it can be used to approximate the desired critical value for valid uniform inference. We do not discuss this approach further to avoid additional technicalities and repetitions.

# 7 Verification of Assumptions in Examples

We focus attention on two issues for our four motivating examples (Section 2): (i) how the high-level conditions imposed previously, including Assumptions 2 and 3, and Condition (iv) in Theorem 2, can be verified under intuitive primitive assumptions; and (ii) how to implement uniform inference based on our theory following the blueprint outlined in Section 6.2.

## 7.1 Example 1: Generalized Conditional Quantile Regression

This example considers generalized conditional quantile regression with a possibly non-identity link: $\rho(y, \eta; q) = (q - \mathbb{1}(y < \eta))(y - \eta)$, where $q \in \mathcal{Q}$ denotes the quantile position. Thus, let $\eta(\mu_0(\mathbf{x}, q))$ be the conditional $q$-quantile of $Y$ given $\boldsymbol{X} = \mathbf{x}$; we verify in the supplemental appendix that such $\mu_0$ solves (1.1). For this example, the following simple proposition, proven in the supplemental appendix (Proposition H.1), gives sufficient conditions to verify the general Assumptions 2 and 3, and Condition (iv) in Theorem 2.

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

The additional conditions in this proposition are primitive and easy-to-interpret, only restricting the conditional density of $Y$ given $\boldsymbol{X}$ to be bounded and smooth in a mild sense. Our assumptions are on par with or are weaker than those imposed in [2], despite the high level of generality of our theoretical results.

We can implement uniform inference following the plug-in (or bootstrap-based) method described in Section 6.2. For the plug-in approach, note that $S_{q,\tilde{q}}(\mathbf{x}) = q \wedge \tilde{q} - q\tilde{q}$ is known and constant in $\mathbf{x}$, so a natural plug-in estimator of $\bar{\boldsymbol{\Sigma}}_{q,\tilde{q}}$ is

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

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

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

depends on the unknown conditional density $f_{Y|X}$, and a plug-in estimator is not immediately available. However, many estimation strategies have been proposed in the literature [24]. We do not recommend a particular choice, but rather note that any estimator satisfying the mild convergence rate requirement in Condition (iii) of Theorem 2 can be used. Alternatively, a weighted bootstrap strategy could be used to construct an approximation process $\widehat{Z}(\mathbf{x}, q)$, which avoids estimation of the covariance function. See the discussion at the end of Section 6.2 for more details.

## 7.2 Example 2: Generalized Conditional Distribution Regression

The loss function is $\rho(y, \eta; q) = (\mathbb{1}(y \le q) - \eta)^2$ with a possibly non-identity inverse link function $\eta(\cdot)$. The derivative function is $\psi(y, \eta; q) = -2(\mathbb{1}(y \le q) - \eta)$. The following proposition, proven in the supplemental appendix (Proposition H.4), verifies our high-level assumptions under mild regularity conditions on the conditional distribution function of $Y$ given $\boldsymbol{X}$.

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

Implementation of uniform inference in this example follows the blueprint described in Section 6.2. To construct the prerequisite estimators, note that in this case $S_{q,\tilde{q}}(\mathbf{x}_i) = 4F_{Y|X}(q \wedge \tilde{q}|\mathbf{x}_i)\big(1 - F_{Y|X}(q \vee \tilde{q}|\mathbf{x}_i)\big)$. Therefore, a simple plug-in estimator of $\bar{\mathbf{\Sigma}}_{q,\tilde{q}}$ is

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

In addition, a plug-in estimator of the matrix $\bar{\boldsymbol{Q}}_q$ is $\widehat{\boldsymbol{Q}}_q = 2\mathbb{E}_n[\eta^{(1)}(\widehat{\mu}(\mathbf{x}_i, q))^2\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^{\mathsf{T}}]$. A bootstrap-based method, as briefly discussed in Section 6.2, could be used as an alternative implementation approach.

### 7.3 Example 3: Generalized $L_p$ Regression

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

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

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

For implementation, we can follow the blueprint in Section 6.2. Since $\mathcal{Q}$ is a singleton, dependence on $q$ can be dropped. Direct plug-in choices for estimating the prerequisite matrices take the form

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

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

### 7.4 Example 4: Logistic Regression

For this final example, the loss function is $\rho(y,\eta) = -y\log\eta - (1-y)\log(1-\eta)$, the inverse link function is $\eta(\theta) = 1/(1+e^{-\theta})$, and the derivative function is $\psi(y,\eta) = -y/\eta + (1-y)/(1-\eta)$, and the loss function does not depend on $q \in \mathcal{Q}$. The following proposition, proven in the supplemental appendix (Proposition H.7), gives simple primitive conditions verifying the high-level assumptions for our general theoretical results.

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

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

$$\mathbb{E}[\widehat{Z}(\mathbf{x})\widehat{Z}(\tilde{\mathbf{x}})|\boldsymbol{D}_n] = \widehat{\ell}(\mathbf{x})^\intercal\widehat{\boldsymbol{\Sigma}}\widehat{\ell}(\tilde{\mathbf{x}}), \qquad \widehat{\ell}(\mathbf{x})^\intercal = \frac{\mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\intercal\widehat{\boldsymbol{Q}}^{-1}}{\sqrt{\widehat{\Omega}_{\mathbf{v}}(\mathbf{x})}}$$

where $\widehat{\Omega}_{\mathbf{v}}(\mathbf{x}) = \mathbf{p}^{(\mathbf{v})}(\mathbf{x})^\intercal\widehat{\boldsymbol{Q}}^{-1}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{Q}}^{-1}\mathbf{p}^{(\mathbf{v})}(\mathbf{x})$, with $\widehat{\boldsymbol{Q}}$ and $\widehat{\boldsymbol{\Sigma}}$ estimators of $\bar{\boldsymbol{Q}}$ and $\bar{\boldsymbol{\Sigma}}$, respectively. More precisely, standard choices are

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

where $\widehat{\eta}_i = \eta(\widehat{\mu}(\mathbf{x}_i))$ and $\widehat{\epsilon}_i = y_i - \widehat{\eta}_i$. A bootstrap strategy could also be used. See Section 6.2 for more discussion.

## 8 Other Parameters of Interest

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

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

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

- *Conditional Treatment Effect Estimator*: $\eta(\widehat{\mu}_1(\mathbf{x},q)) - \eta(\widehat{\mu}_2(\mathbf{x},q))$.

Uniform consistency of the three estimators immediately follows from uniform consistency of $\widehat{\mu}(\mathbf{x},q)$ (Corollary 1) because the transformation function $\eta$ is twice continuously differentiable. Furthermore, a Bahadur representation for each of the transformation estimators is readily available by Theorem 1 and a Taylor expansion. For example, for the level estimator,

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

with

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

and for the marginal effect of the $k$th regressor,

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

with

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

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

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

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

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

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

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

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

with strong approximation rate $r_{\mathtt{SAME}}$ as in Theorem 2. These results are formalized in Theorem J.1 in the supplemental appendix. Of course, an analogous result holds for the conditional treatment effect estimator.

Finally, for implementation we can construct feasible processes to approximate $Z_{\mathtt{LE}}(\mathbf{x},q)$ and $Z_{\mathtt{ME}}(\mathbf{x},q)$ via plug-in or bootstrap methods as discussed in Section 6.2, and illustrated in Section 7, which then can be employed to characterize distributions of the entire level process $(\eta(\widehat{\mu}(\mathbf{x},q)) : (\mathbf{x},q) \in \mathcal{X}\times\mathcal{Q})$, marginal effect process $(\frac{\partial}{\partial x_k}\eta(\widehat{\mu}(\mathbf{x},q)) : (\mathbf{x},q) \in \mathcal{X}\times\mathcal{Q})$, and conditional treatment effect process $(\eta(\widehat{\mu}_2(\mathbf{x},q)) - \eta(\widehat{\mu}_1(\mathbf{x},q)) : (\mathbf{x},q) \in \mathcal{X}\times\mathcal{Q})$.

# 9 Conclusion

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

# Acknowledgements

We thank Richard Crump, Max Farrell, Will Underwood, and Rae Yu for helpful comments and discussions. Cattaneo gratefully acknowledges financial support from the National Science Foundation through grants DMS-2210561 and SES-2241575. Feng gratefully acknowledges financial support from the National Natural Science Foundation of China (NSFC) through grants 72203122, 72133002, and 72250064.

# References

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

[2] 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.

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

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

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

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

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

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

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

[10] 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.

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

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

[13] 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.

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

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

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

[17] Dudley, R. M. (2014). *Uniform central limit theorems*, *142*: Cambridge university press.

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

[19] Györfi, L., Kohler, M., Krzyzak, A., Walk, H. *et al.* (2002). *A distribution-free theory of nonparametric regression*, *1*: Springer.

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

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

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

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

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

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

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

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

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

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

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

[31] 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.

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

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

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

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

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

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

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