Journal of Machine Learning Research 27 (2026) 1-59 Submitted 6/24; Revised 1/26; Published 6/26

Extrapolation-Aware Nonparametric Statistical Inference

Niklas Pfister NIKLAS.PFISTER@GMAIL.COM
Department of Mathematical Sciences
University of Copenhagen
Copenhagen, Denmark

Peter Bühlmann PETER.BUHLMANN@STAT.MATH.ETHZ.CH
Seminar für Statistik
ETH Zürich
Zürich, Switzerland

Editor: Christophe Giraud

Abstract

We define extrapolation as statistical inference on a conditional function (e.g., a conditional expectation or conditional quantile) evaluated outside the support of the conditioning variable. This type of extrapolation occurs in many data analysis applications and can invalidate the conclusions if not taken into account. While extrapolation is straightforward in parametric models, it becomes challenging in nonparametric models. In this work, we extend the nonparametric statistical model to explicitly allow for extrapolation and introduce a class of extrapolation assumptions that can be combined with existing inference techniques to draw extrapolation-aware conclusions. The proposed extrapolation assumptions stipulate that the conditional function attains its minimal and maximal directional derivative, in each direction, within the observed support. We illustrate how the framework applies to several statistical applications including prediction and uncertainty quantification. We furthermore propose a consistent estimation procedure to adjust existing nonparametric estimates for extrapolation by providing lower and upper extrapolation bounds. The procedure is empirically evaluated on simulated and real-world data.

Keywords: extrapolation, nonparametric statistical inference, conditional function, prediction and confidence intervals, causality

1. Introduction

In the natural sciences, the term extrapolation broadly refers to any process that extends conclusions about observed settings to unseen settings. For example, we extrapolate if we learn the gravitational constant on earth in a controlled experiment and later use it as part of a model (in this case based on the laws of physics) to predict the energy required to launch a rocket into space. While extrapolation with a known mechanistic model generally works well, it becomes more challenging in more noisy, complex, or chaotic settings where full mechanistic knowledge is unavailable. For instance, given data from a randomized control trial for a specific drug based on an adult cohort, we might want to extrapolate the study results to infants. In this case the underlying mechanism is not fully understood and some additional knowledge is required to draw reliable conclusions. In this work, we focus on extrapolation when only a statistical model without mechanistic knowledge is

available. To make this precise, let P_0 be a distribution over $(X, Y) \in \mathcal{X} \times \mathbb{R}$ where X are covariates and Y is a response and let $\text{supp}(X)$ denote the support of X. Furthermore, assume we are interested in a conditional (on X) function $\Phi_0 : \mathcal{X} \to \mathbb{R}$, e.g., a conditional expectation. Extrapolation throughout this work refers to statistical inference on $\Phi_0(x)$ for $x \in \mathcal{X} \setminus \text{supp}(X)$. Since the conditional function is generally defined through P_0 alone, it is not well-defined outside of $\text{supp}(X)$. Hence, extrapolation can only be meaningful under appropriate assumptions.

Existing works have used various assumptions to render extrapolation a well-defined inference task. These can be roughly categorized into three groups. Firstly, *global parametric assumptions*, which assume the conditional function Φ_0 is parametric on all of $\mathcal{X}$ (e.g., linear or polynomial) in a way that ensures the parameters can be (partially) identified from the observed distribution P_0 and hence used to either identify or bound the behavior of Φ_0 on all of $\mathcal{X}$. All parametric statistical models fall into this category, as well as many semiparametric models. Secondly, *functional constraint assumptions*, which assume that Φ_0 has specific properties that transfer from $\text{supp}(X)$ to all of $\mathcal{X}$ (e.g., monotonicity, periodicity, or smoothness). For nonparametric regression in particular, there are many methods that make such assumptions implicitly, for example by assuming Φ_0 extrapolates constantly (e.g., tree ensembles) or linearly outside of the support (Li and Heckman, 2003; Christiansen et al., 2021). Other works make more explicit assumptions on patterns or periodicity (Wilson and Adams, 2013; Wang et al., 2024) or assume additive nonlinear functions (Dong and Ma, 2022). Thirdly, *mechanistic assumptions*, which assume an underlying mechanistic model (often causal) that ensures that Φ_0 is (partially) identified on $\mathcal{X}$. For example, physical laws can sometimes be used to constrain the model class, which can then improve how well a model extrapolates (e.g., Pfister et al., 2019). Recent works have also assumed causal structure such as independent additive noise to evaluate nonlinear functions outside of the support (Shen and Meinshausen, 2025; Saengkyongam et al., 2024).

While extrapolating under global parametric assumptions is straightforward, it becomes more challenging for functional constraint and mechanistic assumptions. Most nonparametric works that explicitly consider extrapolation have focused on prediction of Y at a point x. We suspect one reason for this is that in nonparametric statistics, the target of inference is conventionally a quantity identified from the data generating distribution P_0. However, when extrapolating, the target of inference – $\Phi_0(x)$ for $x \in \mathcal{X} \setminus \text{supp}(X)$ – is a priori not a function of P_0. Prediction is therefore a natural task to consider as it can be viewed model-free (as commonly done in the machine learning community), ensuring a well-defined extrapolation task without explicitly defining Φ_0. In this work, we propose a framework, based on Markov kernels, that extends the nonparametric approach in a way that explicitly allows us to consider extrapolation of any conditional function Φ_0 defined via the Markov kernel. We further ensure partial identifiability by assuming that Φ_0 has directional derivatives on $\mathcal{X}$ that are bounded by its directional derivatives on $\text{supp}(X)$. This functional constraint assumption allows us to apply Taylor's theorem to construct extrapolation bounds on Φ_0 on all of $\mathcal{X}$ that are identified by P_0. We show that these bounds can be useful in a range of statistical inference tasks and can be estimated consistently from data. Our framework can be applied with any existing nonparametric estimate of Φ_0 on $\text{supp}(X)$ as long as the extrapolation assumption for Φ_0 holds. Importantly, the resulting inference is extrapolation-aware in the sense that (in the large sample limit) it remains

unchanged if no extrapolation occurs, since the lower and upper extrapolation bounds coincide on $\text{supp}(X)$. So even if the extrapolation assumptions are incorrect, the inference remains valid on $\text{supp}(X)$. A further benefit of the proposed approach is that it does not require knowledge of $\text{supp}(X)$ which is often unknown and non-trivial to estimate from data, particularly if X is multi-dimensional. The extrapolation bounds automatically adapt to become wider in regions of the X-space with few (or no) observations reasonably close by and where there is a large extrapolation uncertainty in the conditional function Φ_0. We use this property to derive a score that quantifies extrapolation, which can be used to quantify extrapolation uncertainty.

There is a large body of literature on related types of extrapolation. The fields of domain adaptation and generalization, which aim to find predictive models that perform well on a test distribution different from the training distribution, have obvious connections that we discuss in Section 3.1. In most of the literature, however, one assumes that the test and training distributions have overlapping supports, which excludes extrapolation as defined here. One exception is distributionally robust optimization methods that use the Wasserstein distance, which allows for the supports to be disjoint (e.g., Sinha et al., 2018). A further related area is causal inference, where the term extrapolation is sometimes used to refer to generalizing from the observational to a previously unseen interventional distribution. Again, most works in this area explicitly exclude extrapolation as defined here by assuming overlapping supports. Nevertheless, as we discuss in Section 3.4, since our framework extends the conventional nonparametric model, it can also be applied to causal inference tasks with the same causal assumptions required as in the non-extrapolation setting.

Our main contributions are as follows. We introduce a nonparametric statistical framework that explicitly accounts for extrapolation using Markov kernels. We propose a class of extrapolation assumptions based on bounded directional derivatives and use Taylor's theorem to derive extrapolation bounds. We show how these bounds apply to prediction, uncertainty quantification, and quantifying extrapolation. Finally, we propose a consistent estimation procedure and evaluate it empirically.

The remainder of the paper is structured as follows. Section 2 uses Taylor's theorem to derive extrapolation bounds for differentiable functions. Section 3 introduces the nonparametric statistical framework and discusses three applications: out-of-distribution prediction, extrapolation-aware uncertainty quantification, and quantifying extrapolation. Section 3.4 discusses a causal perspective on extrapolation. Section 4 proposes a consistent estimation approach for the extrapolation bounds. Section 5 empirically evaluates the estimation procedure on simulated data and illustrates extrapolation-aware prediction intervals on two real-world data sets.

Notation Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a fixed domain, denote by $C^q(\mathcal{X})$ the set of q-times continuously differentiable functions and define $\mathcal{B} \coloneqq \{x \in \mathbb{R}^d \mid \|x\|_2 = 1\}$. For all $f \in C^q(\mathcal{X})$ and all $v \in \mathcal{B} \cup \{0\}$ define the directional derivative $D_v f : \mathcal{X} \to \mathbb{R}$ for all $x \in \mathcal{X}$ by

$$D_v f(x) \coloneqq \lim_{h \to 0} \frac{f(x + h \cdot v) - f(x)}{h}.$$

Moreover, for all $\ell \in \{1, \dots, q\}$, define the ℓ-th directional derivative recursively by $D_v^\ell f \coloneqq D_v(D_v^{\ell-1} f)$, where $D_v^0 f = f$. For all multi-indices $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $|\boldsymbol{\alpha}| \coloneqq \sum_{j=1}^d \boldsymbol{\alpha}^j \leq q$

and all functions $f \in C^q(\mathcal{X})$ define $\partial^{\boldsymbol{\alpha}} f \coloneqq D_{e_1}^{\alpha^1} \cdots D_{e_d}^{\alpha^d} f$ and for all $j \in \{1, \ldots, d\}$ define $\partial_j f = D_{e_j}^1 f$, where e_j denotes the j-th canonical unit vector. Additionally, define the function $\overline{v} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^d$ which maps all points $x_0, x_1 \in \mathcal{X}$ to the unit vector pointing in the same direction as the vector from x_0 to x_1, that is, for all $x_0, x_1 \in \mathcal{X}$ define $\overline{v}(x_0, x_1) \coloneqq \frac{x_1 - x_0}{\|x_1 - x_0\|_2} \mathbb{1}_{\{x_0 \neq x_1\}}$. Lastly, we denote by $\mathfrak{B}(\mathbb{R})$ the Borel-sigma algebra on $\mathbb{R}$.

2. Extrapolation via Taylor's theorem

We start by considering extrapolation from a fully deterministic perspective. Assume we are given a function $f \in C^q(\mathcal{X})$ but are only able to evaluate it on a closed domain $\mathcal{D} \subseteq \mathcal{X}$. Since, f is q-times continuously differentiable, knowing the function on $\mathcal{D}$ can help constrain how the function can behave on $\mathcal{X} \setminus \mathcal{D}$. Formally, using Taylor's theorem (Taylor, 1715), we get for all $x_0 \in \mathcal{D}$ and all $x_1 \in \mathcal{X}$ that there exists a $c \in [0, 1]$ such that for $\xi \coloneqq cx_1 + (1-c)x_0$ it holds that

$$f(x_1) = \sum_{\ell=0}^{q-1} D_{\overline{v}(x_0,x_1)}^{\ell} f(x_0) \frac{\|x_1 - x_0\|_2^{\ell}}{\ell!} + D_{\overline{v}(x_0,x_1)}^{q} f(\xi) \frac{\|x_1 - x_0\|_2^{q}}{q!}. \tag{1}$$

The only quantity in this equation which can – depending on c – rely on evaluating the function outside of $\mathcal{D}$ is the value of $D_{\overline{v}(x_0,x_1)}^{q} f(\xi)$. However, if we are willing to assume that the q-th directional derivative of f is bounded on all of $\mathcal{X}$, we can use (1) to construct upper and lower bounds on $f(x_1)$. Our approach is based on assuming that f – the function we want to extrapolate – behaves at most as extreme on $\mathcal{X}$ as on $\mathcal{D}$. To be at most as extreme in our setting, means that the directional derivatives of f on $\mathcal{X}$ are bounded by its directional derivatives on $\mathcal{D}$, for all possible directions.

Definition 1 (q-th derivative dominated) *Let $f \in C^q(\mathcal{X})$ be a function and $\mathcal{D} \subseteq \mathcal{X}$ be a non-empty closed set. A function $g \in C^q(\mathcal{X})$ is called* q-th derivative dominated by f over $\mathcal{D}$*, denoted by $g \triangleleft_{\mathcal{D}}^q f$, if it holds for all $v \in \mathcal{B}$ that*

$$\inf_{x \in \mathcal{X}} D_v^q g(x) \geq \inf_{x \in \mathcal{D}} D_v^q f(x) \quad \text{and} \quad \sup_{x \in \mathcal{X}} D_v^q g(x) \leq \sup_{x \in \mathcal{D}} D_v^q f(x).$$

In words a function g is q-th derivative dominated by f on $\mathcal{D}$ if all its directional derivatives of order q are bounded on all of $\mathcal{X}$ by the corresponding directional derivatives of f on $\mathcal{D}$. Based on this definition we now consider functions $f \in C^q(\mathcal{X})$ that satisfy

$$f \triangleleft_{\mathcal{D}}^q f. \tag{2}$$

This formalizes the previously mentioned intuitive notion of behaving at most as extreme on $\mathcal{X}$ as on $\mathcal{D}$. Whenever a function satisfies (2), we can use (1) to provide bounds on its behavior on $\mathcal{X}$ that only depend on the values it attains on $\mathcal{D}$.

Definition 2 (Extrapolation bounds) *For all $f \in C^q(\mathcal{X})$ and all non-empty closed $\mathcal{D} \subseteq \mathcal{X}$ define the* extrapolation bounds $B_{q,f,\mathcal{D}}^{\text{lo}}, B_{q,f,\mathcal{D}}^{\text{up}} : \mathcal{X} \to [-\infty, \infty]$ *given for all $x \in \mathcal{X}$ by*

$$B_{q,f,\mathcal{D}}^{\text{lo}}(x) \coloneqq \sup_{x_0 \in \mathcal{D}} \left(\sum_{\ell=0}^{q-1} D_{\overline{v}(x_0,x)}^{\ell} f(x_0) \frac{\|x - x_0\|_2^{\ell}}{\ell!} + \inf_{z \in \mathcal{D}} D_{\overline{v}(x_0,x)}^{q} f(z) \frac{\|x - x_0\|_2^{q}}{q!} \right)$$

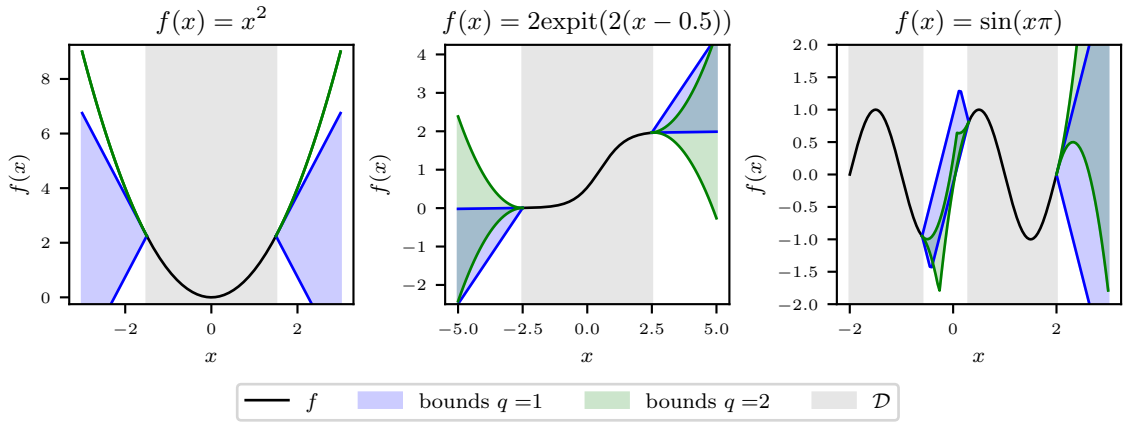


Figure 1: Visualization of the extrapolation bounds given in Definition 2 for three different functions f and domains $\mathcal{D}$. The shaded gray area corresponds to $\mathcal{D}$ on which the function we would like to extrapolate is given in black. Blue corresponds to the first order upper and lower extrapolation bounds $B^{\text{lo}}_{1,f,\mathcal{D}}$ and $B^{\text{up}}_{1,f,\mathcal{D}}$ and green to the second order extrapolation bounds $B^{\text{lo}}_{2,f,\mathcal{D}}$ and $B^{\text{up}}_{2,f,\mathcal{D}}$. For this visualization, we approximate the bounds by sampling points uniformly in $\mathcal{D}$, which is consistent by Theorem 11 below.

and

$$B^{\text{up}}_{q,f,\mathcal{D}}(x) \coloneqq \inf_{x_0 \in \mathcal{D}} \left(\sum_{\ell=0}^{q-1} D^{\ell}_{\overline{v}(x_0,x)} f(x_0) \frac{\|x - x_0\|_2^{\ell}}{\ell!} + \sup_{z \in \mathcal{D}} D^{q}_{\overline{v}(x_0,x)} f(z) \frac{\|x - x_0\|_2^{q}}{q!} \right).$$

The extrapolation bounds are constructed using (1) and then replacing the highest order derivative with the worst possible directional derivative f attains in $\mathcal{D}$. Since the resulting bound is valid for any anchor point $x_0 \in \mathcal{D}$, we select the one that results in the tightest bound. From this construction it can be shown that for all $x \in \mathcal{D}$ the bounds satisfy

$$B^{\text{lo}}_{q,f,\mathcal{D}}(x) = f(x) = B^{\text{up}}_{q,f,\mathcal{D}}(x),$$

as long as the q-th directional derivatives of f in all directions are bounded on $\mathcal{D}$. The bounds are visualized for three one dimensional functions in Figure 1. Different orders q capture different aspects of the function, for example, for $q = 1$ monotone behavior is captured in Figure 1 (middle). Moreover, as seen in Figure 1 (left), the extrapolation bounds only bound the true function if it indeed satisfies $f \triangleleft^{q}_{\mathcal{D}} f$. The following theorem provides a formal connection between the extrapolation bounds for a function $f \in C^q(\mathcal{X})$ and functions that are q-th derivative dominated by f.

Theorem 3 (Properties of extrapolation bounds) *Let $f \in C^q(\mathcal{X})$ be a function and $\mathcal{D} \subseteq \mathcal{X}$ be a non-empty closed set. Then, the following two statements hold:*

(i) For all $g \in C^q(\mathcal{X})$ satisfying for all $x \in \mathcal{D}$ that $g(x) = f(x)$ and $g \lhd^q_{\mathcal{D}} f$ it holds that

$$\forall x \in \mathcal{X}: \quad B^{\text{lo}}_{q,f,\mathcal{D}}(x) \leq g(x) \leq B^{\text{up}}_{q,f,\mathcal{D}}(x).$$

(ii) If $\mathcal{X}$ is compact, there exists, for all $\star \in \{\text{lo}, \text{up}\}$, a sequence $(g^\star_n)_{n\in\mathbb{N}} \subseteq C^q(\mathcal{X})$ satisfying

$$\lim_{n\to\infty} \sup_{x\in\mathcal{X}} \left|B^\star_{q,f,\mathcal{D}}(x) - g^\star_n(x)\right| = 0 \quad \textit{and} \quad \forall n \in \mathbb{N}: \, g^\star_n \lhd^q_{\mathcal{D}} f.$$

A proof is given in Supplementary Material D.1. Part (i) implies that for all $f \in C^q(\mathcal{X})$ satisfying $f \lhd^q_{\mathcal{D}} f$ the extrapolation bounds bound the values of f on all of $\mathcal{X}$, i.e.,

$$\forall x \in \mathcal{X}: \quad B^{\text{lo}}_{q,f,\mathcal{D}}(x) \leq f(x) \leq B^{\text{up}}_{q,f,\mathcal{D}}(x).$$

Part (ii) provides a partial converse of this statement. Specifically, it states that the extrapolation bounds can be uniformly approximated by a sequence of functions that are q-th derivative dominated by f. As a side result, part (ii) further implies that the extrapolation bounds are uniformly continuous.

3. Extrapolation in statistical applications

We now apply the Taylor-based extrapolation bounds from the previous section to statistical settings, where we link data-generating distributions to conditional functions. We assume q is fixed and drop it from the notation.[1] We begin by formalizing extrapolation within a nonparametric statistical model.

Let P_0 be a data-generating distribution over $(X, Y) \in \mathcal{X} \times \mathbb{R}$ and define $\mathcal{D}_{\text{in}} := \text{supp}(X)$ and $\mathcal{D}_{\text{out}} := \mathcal{X} \setminus \mathcal{D}_{\text{in}}$. Let $Q_0 : \mathcal{X} \times \mathfrak{B}(\mathbb{R}) \to [0, 1]$ be a Markov kernel that generates a distribution over Y for all distributions over X and satisfies for all $B \in \mathfrak{B}(\mathbb{R})$ that

$$P^Y_0(B) = \int_{\mathcal{X}} Q_0(x, B) P^X_0(dx), \tag{3}$$

where P^X_0 and P^Y_0 denote the marginal distributions of X and Y under P_0, respectively. The conditional distribution of $Y|X = x$ is only defined on $\mathcal{D}_{\text{in}}$, so extrapolation requires extending this notion to all of $\mathcal{X}$. The Markov kernel Q_0 provides such an extension by defining a distribution over Y for all $x \in \mathcal{X}$. As we discuss in Section 3.4, one can alternatively define extrapolation using a causal model, since interventional conditionals, unlike observational conditionals, immediately extend to the entire domain $\mathcal{X}$. We choose to avoid the overhead of a causal model and instead use the Markov kernel Q_0, which is well-defined albeit not necessarily fully identified. Our goal is to perform inference on a conditional function $\Phi_0 : \mathcal{X} \to \mathbb{R}$ describing some aspect of Q_0. We focus on the conditional expectation and quantile functions, which cover a range of interesting applications, but the theory extends to any conditional function.

1. In practice, we focus on $q = 1$. While the theory extends to arbitrary q, higher-order derivatives are difficult to estimate reliably, and our estimation procedure in Section 4 requires $q = 1$ when $d > 1$.

Definition 4 (Conditional expectations and quantiles) *We call the function* $\Psi_0 : \mathcal{X} \to \mathbb{R}$ *defined for all* $x \in \mathcal{X}$ *by*

$$\Psi_0(x) \coloneqq \int_{\mathbb{R}} y Q_0(x, dy) \tag{4}$$

the conditional expectation function*. Similarly, for all* $\alpha \in (0,1)$ *the function* $\mathcal{T}_0^\alpha : \mathcal{X} \to \mathbb{R}$ *defined for all* $x \in \mathcal{X}$ *by*

$$\mathcal{T}_0^\alpha(x) \coloneqq \inf\{t \in \mathbb{R} \mid \textstyle\int_{\mathbb{R}} \mathbb{1}(y \le t) Q_0(x, dy) \ge \alpha\}. \tag{5}$$

is called conditional α-quantile function*.*

Both the conditional expectation and the conditional quantile are defined on all of $\mathcal{X}$. However, a priori, they might not be fully identified by P_0. To see this, observe that, using (3), it holds P_0-a.s. that $\Psi_0(X) = \mathbb{E}[Y|X]$ and $\mathcal{T}_0^\alpha(X) = \inf\{t \in \mathbb{R} \mid \mathbb{P}(Y \le t|X) \ge \alpha\}$. Hence – without additional assumptions – they are not identified at values $x \in \mathcal{D}_{\text{out}}$. For a conditional function Φ_0 of interest, we therefore distinguish two types of statistical inference: (i) *interpolation*, which is inference on $\Phi_0(x)$ for $x \in \mathcal{D}_{\text{in}}$ and (ii) *extrapolation*, which is inference on $\Phi_0(x)$ for $x \in \mathcal{D}_{\text{out}}$. The distinguishing feature between interpolation and extrapolation is that interpolation is feasible with conventional nonparametric assumptions (e.g., smoothness or shape constraints), while extrapolation is impossible without additional extrapolation assumptions. To render extrapolation feasible, we first specify extrapolation assumptions that are reasonable in practice. The following example illustrates two common extrapolation assumptions – parametric and periodic.

Example 1 (Parametric and periodic extrapolation) *The most common approach to making extrapolation meaningful is to assume that the conditional function of interest is parametric. For example, one could assume a linear structural equation model (SEM) given by*

$$Y = \theta^\top X + \varepsilon \quad \textit{and} \quad \varepsilon \perp\!\!\!\perp X,$$

where $\theta \in \mathbb{R}^d$ *and* $\varepsilon \sim \mu$ *for some distribution* μ*. This model implies a Markov kernel* Q_0 *that satisfies for all* $x \in \mathcal{X}$ *and all* $B \in \mathcal{B}(\mathbb{R})$ *that* $Q_0(x, B) = \mu(B - \theta^\top x)$*. The conditional expectation is then simply* $\Psi_0 : x \mapsto \theta^\top x$ *and hence fully identified by the parameter* θ *which is identified as long as* P_0 *is a non-degenerate distribution.*

A further approach is to assume a periodic extrapolation model. For example, again using a SEM model, one could assume a model of the form

$$Y = f(X) + \varepsilon \quad \textit{and} \quad \varepsilon \perp\!\!\!\perp X,$$

where f *is a measurable function satisfying for all* $x \in [0,1)$ *and all* $k \in \mathbb{Z}$ *that* $f(x) = f(x+k)$ *and* $\varepsilon \sim \mu$ *for some distribution* μ*. The Markov kernel* Q_0 *implied by this model is given for all* $x \in \mathcal{X}$ *and* $B \in \mathcal{B}(\mathbb{R})$ *by* $Q_0(x, B) = \mu(B - f(x))$*. The conditional* α*-quantile* $\mathcal{T}_0^\alpha$ *is then given for all* $x \in \mathcal{X}$ *by* $\mathcal{T}_0^\alpha(x) = f(x) + \inf\{t \in \mathbb{R} \mid \mathbb{P}(\varepsilon \le t) \ge \alpha\}$ *which is identifiable as long as* f *is identifiable from* P_0*. Since* f *is periodic, identifiablility of* f *again does not require* X *to have full support on* $\mathcal{X}$*.*

In the parametric example, the extrapolation assumption constrains the data-generating distribution within the support, making it incompatible with nonparametric approaches. The periodic example can be combined with nonparametric approaches, but only applies to settings where periodic extrapolation is appropriate. We instead propose a smoothness-based extrapolation assumption directly on the conditional function of interest, which can be combined with existing nonparametric approaches without further restricting their interpolation behavior. Specifically, we assume that Φ_0 is q-th derivative dominated by itself.

Definition 5 (q-th derivative extrapolating) *A conditional function $\Phi_0 : \mathcal{X} \to \mathbb{R}$ is called* q-th derivative extrapolating *if $\Phi_0 \in C^q(\mathcal{X})$ and*

$$\Phi_0 \lhd^q_{\mathcal{D}_{\text{in}}} \Phi_0.$$

This condition depends on both the conditional distribution of Y given X (via Φ_0) and the marginal distribution of X (via $\mathcal{D}_{\text{in}}$). Importantly, when $\mathcal{D}_{\text{in}} = \mathcal{X}$, it reduces to requiring that Φ_0 is q-times continuously differentiable – a standard nonparametric assumption. Hence, for interpolation tasks, the usual smoothness assumption suffices, and our method does not require knowing whether a task involves interpolation or extrapolation. For extrapolation, the condition constrains Φ_0 on $\mathcal{D}_{\text{out}}$ sufficiently to provide meaningful inference, albeit with added uncertainty, see Remark 6.

Remark 6 (Scope of q-th derivative extrapolating assumption) *The q-th derivative extrapolating assumption only constrains the behavior on $\mathcal{D}_{\text{out}}$; within $\mathcal{D}_{\text{in}}$, it reduces to standard smoothness. Outside the observed support, the assumption implies that Φ_0 is bounded by order-q polynomials anchored at points in $\mathcal{D}_{\text{in}}$. The assumption is violated when the absolute values of the directional derivatives are larger outside the observed domain than within, for instance, for an exponential function observed only on a bounded interval. However, under the strictly weaker assumption that the ordering of directional changes in the q-th derivative is preserved from $\mathcal{D}_{\text{in}}$ to $\mathcal{D}_{\text{out}}$ (with small and large changes on $\mathcal{D}_{\text{in}}$ corresponding to small and large changes on $\mathcal{D}_{\text{out}}$, respectively), diverging extrapolation bounds indicate directions with extrapolation uncertainty. Thus, even when the q-th derivative extrapolating assumption is violated, the extrapolation bounds and the resulting extrapolation score (Section 3.3) may remain useful for identifying regions where predictions rely on extrapolation, though the bounds may be optimistic about the true function's behavior.*

Under the assumption that Φ_0 is q-th derivative extrapolating, it is possible to bound Φ_0 on $\mathcal{D}_{\text{out}}$ by only knowing its values on $\mathcal{D}_{\text{in}}$. More specifically, Theorem 3 ensures that Φ_0 lies in the set of *feasible conditional functions* defined by

$$\mathcal{F}_{\Phi_0} \coloneqq \left\{ \phi \in C^0(\mathcal{X}) \,\middle|\, \forall x \in \mathcal{X} : \, B^{\text{lo}}_{\Phi_0, \mathcal{D}_{\text{in}}}(x) \leq \phi(x) \leq B^{\text{up}}_{\Phi_0, \mathcal{D}_{\text{in}}}(x) \right\}, \tag{6}$$

which is identifiable from P_0.

There are high-level connections between this formulation and several other approaches. First, from an empirical Bayes perspective (Robbins, 1992), the q-th derivative extrapolating assumption corresponds to assuming that the target function behaves similarly on $\mathcal{D}_{\text{out}}$ as on $\mathcal{D}_{\text{in}}$. To see this, consider a prior supported on functions from $\mathcal{X}$ to $\mathbb{R}$ with q-th derivatives bounded by unknown C_{low} and C_{upp}. Estimating these bounds from data on

$\mathcal{D}_{\text{in}}$ and performing a posterior update yields, by construction, a posterior supported on q-th derivative extrapolating functions.

Second, distributionally robust optimization (DRO) considers ambiguity sets over candidate distributions (Rahimian and Mehrotra, 2019), whereas $\mathcal{F}_{\Phi_0}$ can be viewed as an ambiguity set over candidate conditional functions; for DRO to address extrapolation as defined here, the ambiguity set must include distributions with disjoint supports for $\mathcal{D}_{\text{in}}$ and $\mathcal{D}_{\text{out}}$, which rules out f-divergence-based sets but is permitted by Wasserstein-based sets (Sinha et al., 2018). A key advantage of our approach is that extrapolation uncertainty depends on directional derivative variation, enabling extrapolation far from $\mathcal{D}_{\text{in}}$ when derivatives vary little. In contrast, DRO becomes increasingly conservative with distance, e.g., when using large Wasserstein balls. In addition, formulating assumptions directly on the conditional function has much better interpretability: the q-th derivative extrapolating assumption provides a direct constraint on the target of inference, whereas equivalent distributional assumptions are more obscure. Causal approaches to distribution generalization (Christiansen et al., 2021; Ahuja et al., 2020; Peters et al., 2016) also formulate assumptions on functions rather than distributions, though they do not consider extrapolation outside the support.

Finally, there is a tenuous relation to shape constraints such as monotonicity or convexity (e.g., Groeneboom and Jongbloed, 2014), which restrict the function class globally through shape rather than smoothness. Like smoothness assumptions in nonparametric methods, shape constraints provide sufficient regularity to efficiently estimate the target function on $\mathcal{D}_{\text{in}}$. However, they similarly do not necessarily yield identifiability on $\mathcal{D}_{\text{out}}$, even though they may more substantially restrict the possible values on $\mathcal{D}_{\text{out}}$. In contrast, if the target function is both smooth and monotone, our framework infers monotonicity from estimated derivatives and captures this in the extrapolation bounds; more directly leveraging shape constraints for extrapolation remains an interesting direction for future work.

In the following sections, we discuss three applications that leverage the fact that $\mathcal{F}_{\Phi_0}$ is identifiable from P_0, allowing us to perform inference on the extrapolation bounds rather than the conditional function directly: (i) out-of-support prediction, (ii) extrapolation-aware uncertainty quantification, and (iii) quantifying extrapolation. We present results for arbitrary q, though in practice we focus on the case $q = 1$.

3.1 Out-of-support prediction

Consider a setting in which n i.i.d. observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ from P_0 are observed and we want to learn a prediction function $\hat{f}$ for Y given $X = x$ for some $x \in \mathcal{X}$. A standard nonparametric approach is to estimate the conditional expectation Ψ_0 and use it as $\hat{f}$. Under mild regularity conditions Φ_0 minimizes the mean squared prediction error under P_0 and hence is optimal in the mean squared sense. However, since Φ_0 is only identified on $\text{supp}(X)$ without extrapolation assumptions, this guarantee is no longer valid when considering predictions at points $x \in \mathcal{D}_{\text{out}}$. To avoid such problems, we need explicit extrapolation assumptions.

Using the terminology of the previous section, assume that Ψ_0 is q-th derivative extrapolating. Then, by Theorem 3, it holds for all $x \in \mathcal{X}$ that

$$B^{\text{lo}}_{\Psi_0, \mathcal{D}_{\text{in}}}(x) \leq \Psi_0(x) \leq B^{\text{up}}_{\Psi_0, \mathcal{D}_{\text{in}}}(x). \tag{7}$$

Since the extrapolation bounds are identified from P_0, this allows us to bound the conditional expectation on all of $\mathcal{X}$. We can use this for prediction by constructing a point estimate based on the extrapolation bounds. As the conditional expectation is not necessarily identified everywhere on $\mathcal{X}$, the performance of the point estimate may depend on the underlying Markov kernel Q_0. We therefore attempt to find a prediction function that is worst-case optimal in the sense that it minimizes the worst-case mean-squared prediction error among all Markov kernels Q that are equal to Q_0 on $\mathcal{D}_{\text{in}}$ and for which $x \mapsto \int_{\mathbb{R}} yQ(x, dy)$ is q-th derivative extrapolating. Formally, we define the set of feasible Markov kernels by

$$\mathcal{Q}_0 \coloneqq \big\{ Q : \mathcal{X} \times \mathfrak{B}(\mathbb{R}) \to [0,1] \text{ Markov kernel} \,\big|\, \forall x \in \mathcal{D}_{\text{in}} : \, Q(x,\cdot) = Q_0(x,\cdot) \text{ and } x \mapsto \textstyle\int_{\mathbb{R}} yQ(x,dy) \text{ is } q\text{-th deriv. extr.} \big\}. \tag{8}$$

For all $Q \in \mathcal{Q}_0$ and all $x \in \mathcal{X}$, we denote by Y_x a random variable with distribution $Q(x, \cdot)$. Using this notation, our goal is to find a prediction function $\hat{f}$, such that for all $x \in \mathcal{X}$, it minimizes the worst-case mean squared prediction error

$$\sup_{Q \in \mathcal{Q}_0} \mathbb{E}_{Q(x,\cdot)}[(Y_x - \hat{f}(x))^2].$$

This guarantees that $\hat{f}$ also performs well in cases where the true underlying Markov kernel Q_0 is chosen adversarial among all observationally equivalent q-th derivative extrapolating Markov kernels.

Proposition 7 (Worst-case optimal prediction under extrapolation) *Let $\mathcal{X}$ be compact, assume the Markov kernel Q_0 is such that the conditional expectation Ψ_0 is q-th derivative extrapolating and define $f^* : \mathcal{X} \to \mathbb{R}$ for all $x \in \mathcal{X}$ by*

$$f^*(x) \coloneqq \frac{1}{2}\left(B^{\text{lo}}_{\Psi_0, \mathcal{D}_{\text{in}}}(x) + B^{\text{up}}_{\Psi_0, \mathcal{D}_{\text{in}}}(x) \right). \tag{9}$$

Then, it holds for all $x \in \mathcal{X}$ that

$$\inf_{f \in C^0(\mathcal{X})} \sup_{Q \in \mathcal{Q}_0} \mathbb{E}_{Q(x,\cdot)}[(Y_x - f(x))^2] = \sup_{Q \in \mathcal{Q}_0} \mathbb{E}_{Q(x,\cdot)}[(Y_x - f^*(x))^2]. \tag{10}$$

A proof is given in Supplementary Material D.2. The type of guarantee in Proposition 7 is common in the field of distribution generalization, where one assumes the training data was generated under P_0 but wants to predict Y under a potentially different test distribution P_{test}. Without further assumptions on P_{test} this is impossible. A well-established assumption is the covariate-shift assumption (Sugiyama et al., 2007), which assumes that the conditional expectation under P_0 and P_{test} remains fixed. For this assumption to suffice without additional extrapolation assumptions, one requires

$$\text{supp}(P^X_{\text{test}}) \subseteq \text{supp}(P^X_0) = \mathcal{D}_{\text{in}},$$

otherwise predictions can be arbitrarily wrong outside $\mathcal{D}_{\text{in}}$. Proposition 7 extends this by allowing the test distribution to have arbitrary support, provided the conditional $Y|X$ under P_{test} is generated by a Markov kernel $Q \in \mathcal{Q}_0$.

3.2 Extrapolation-aware uncertainty quantification

Quantifying uncertainty is important in real-world applications. In this section, we consider two approaches for uncertainty quantification when predicting a real-valued response Y from predictors X: confidence intervals, which quantify the uncertainty in the estimation of the regression function and prediction intervals, which quantify uncertainty in the prediction itself. Existing methods for nonparametric regression only apply within $\mathcal{D}_{\text{in}}$ and therefore cannot provide coverage guarantees when extrapolating. This is particularly troubling in applications where it is difficult or impossible to determine whether and to what degree extrapolation occurs. The extrapolation assumptions discussed above provide a solution. We show that the extrapolation bounds derived above can be combined with existing nonparametric approaches to construct extrapolation-aware confidence and prediction intervals.

We begin with confidence intervals for the conditional expectation function Ψ_0 – the same can be done for other conditional functions Φ_0. Consider a setting in which we want to estimate the conditional expectation Ψ_0 from i.i.d. observations $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P_0$. Assuming that Ψ_0 is q-th derivative extrapolating, it follows from Theorem 7 that

$$\Psi_0(x) \in \left[B^{\text{lo}}_{\Psi_0, \mathcal{D}_{\text{in}}}(x),\, B^{\text{up}}_{\Psi_0, \mathcal{D}_{\text{in}}}(x) \right]. \tag{11}$$

Since the extrapolation bounds are identifiable from P_0, we can use any procedure that produces asymptotically valid confidence intervals for both the lower and upper extrapolation bound and combine them to get extrapolation bounds that are asymptotically valid on all of $\mathcal{X}$.

Proposition 8 (Extrapolation-aware confidence interval coverage) *Fix $\alpha \in (0,1)$ and assume the conditional expectation function Ψ_0 is q-th derivative extrapolating. For both $\star \in \{\text{lo}, \text{up}\}$, all $x \in \mathcal{X}$ and all $\gamma \in (0,1)$ let $\widehat{G}^{\star}_n(\gamma, x)$ be an estimation procedure based on n i.i.d. observations from P_0 satisfying $\lim_{n\to\infty} \mathbb{P}\left(B^{\star}_{\Psi_0, \mathcal{D}_{\text{in}}}(x) \leq \widehat{G}^{\star}_n(\gamma, x) \right) = \gamma$. Define for all $x \in \mathcal{X}$ the confidence intervals*

$$\widehat{\mathrm{C}}^{\text{conf}}_{n;\alpha}(x) \coloneqq \left[\widehat{G}^{\text{lo}}_n(\tfrac{\alpha}{2}, x),\, \widehat{G}^{\text{up}}_n(1 - \tfrac{\alpha}{2}, x) \right].$$

Then, it holds for all $x \in \mathcal{X}$ that

$$\liminf_{n\to\infty} \mathbb{P}(\Psi_0(x) \in \widehat{\mathrm{C}}^{\text{conf}}_{n;\alpha}(x)) \geq 1 - \alpha.$$

The proof follows from a direct application of (11) and is provided in Supplementary Material D.3 for completeness. The estimates $\widehat{G}^{\star}_n(\gamma, x)$ in Proposition 8 can be constructed by a combination of an estimate of $B^{\star}_{\Psi_0, \mathcal{D}_{\text{in}}}(x)$ (discussed in Section 4) and a method to estimate the quantile of the estimator distribution. This could for example be a nonparametric bootstrap procedure, e.g., the percentile bootstrap (Efron, 1981). Since the lower and upper extrapolation bounds are equal to Ψ_0 on $\mathcal{D}_{\text{in}}$, the confidence intervals for conditional expectation resulting from such a procedure are extrapolation-aware in the sense that they become larger on $\mathcal{D}_{\text{out}}$ while remaining tight on $\mathcal{D}_{\text{in}}$ as shown in Example 2.

As the bounds $B^{\star}_{\Psi_0, \mathcal{D}_{\text{in}}}(x)$ are obtained by applying a non-smooth inf-sup map to an estimate of Ψ_0 and its derivatives, the standard boostrap can be inconsistent (Fang and

Santos, 2019). On $\mathcal{D}_{\text{in}}$ this is not a concern, as the bounds reduce to Φ_0 and the procedure to a standard nonparametric confidence interval. On $\mathcal{D}_{\text{out}}$, the map is differentiable whenever the optimizing anchor point is unique; non-uniqueness arises mainly under symmetry and is the case in which exact coverage is not guaranteed. A rigorous treatment of the non-unique optimizer case is beyond our scope and would likely require resampling schemes adapted to non-smooth functional (Fang and Santos, 2019, or subsampling). The conservative coverage observed in our experiments (e.g., Figure 9) is consistent with the bounds being mildly over-covered rather than under-covered on $\mathcal{D}_{\text{out}}$.

Example 2 (Extrapolation-aware bootstrap confidence intervals) *In Figure 2 (left), the data come from a one-dimensional linear model. In this case, assuming a linear model is sufficient for extrapolation and a linear regression can be used to predict values outside of the support. However, in Figure 2 (right) the data come from a model with a nonlinear conditional expectation. Even though a linear regression leads to a good fit, it does not extrapolate well. If one instead assumes that the model is first order Ψ-extrapolating, which is satisfied in both examples, extrapolation-aware confidence intervals based on the extrapolation bounds and a nonparametric percentile bootstrap (blue, solid) using a random forest estimate of Ψ_{P_0} (red, solid) are able to preserve coverage also while extrapolating. In particular, the confidence intervals are extrapolation-aware: they detect the linearity in the linear setting resulting in tight bounds and capture the indeterminacy in the behavior of Ψ_{P_0} outside of the support in the nonlinear setting. Details on the data generating distribution and a coverage experiment are provided in Appendix B.1.*

We now consider prediction intervals. Instead of estimating the conditional expectation function, we want to estimate an interval that contains $Y|X = x$ with a pre-specified probability $1 - \alpha$. One nonparametric approach is to estimate quantiles of the conditional distribution of Y and use them to construct prediction intervals, that is, estimate $[\mathcal{T}_0^{\alpha/2}(x), \mathcal{T}_0^{1-\alpha/2}(x)]$. We can modify this approach to be extrapolation-aware as follows.

Proposition 9 (Extrapolation-aware prediction interval coverage) *Fix $\alpha \in (0, 1)$ and assume the Markov kernel Q_0 is such that the conditional quantiles $\mathcal{T}_0^{\alpha/2}$ and $\mathcal{T}_0^{1-\alpha/2}$ are both q-th derivative extrapolating. Define for all $x \in \mathcal{X}$ the prediction intervals*

$$\mathrm{C}_\alpha^{\text{pred}}(x) := \left[B^{\text{lo}}_{\mathcal{T}_0^{\alpha/2}, \mathcal{D}_{\text{in}}}(x),\ B^{\text{up}}_{\mathcal{T}_0^{1-\alpha/2}, \mathcal{D}_{\text{in}}}(x) \right].$$

Then, it holds for all $x \in \mathcal{X}$ that

$$\mathbb{P}_{Q_0(x,\cdot)}(Y_x \in \mathrm{C}_\alpha^{\text{pred}}(x)) \geq 1 - \alpha.$$

A proof is given in Supplementary Material D.4. In practice, the extrapolation bounds must be estimated, making the prediction intervals random. Consistency of the extrapolation bound estimates is discussed in Section 4, leading to (pointwise) asymptotically valid prediction intervals (see Corollary 12). The extrapolation-aware confidence and prediction intervals can be conservative for points in $\mathcal{D}_{\text{out}}$. We see this as a positive feature: it protects against potentially misleading conclusions due to accidental extrapolation. Using extrapolation-aware bounds is therefore – at least asymptotically – always preferable, as they coincide with standard behavior on $\mathcal{D}_{\text{in}}$ regardless of whether the extrapolation assumption holds, while providing at least some guarantee outside the support.

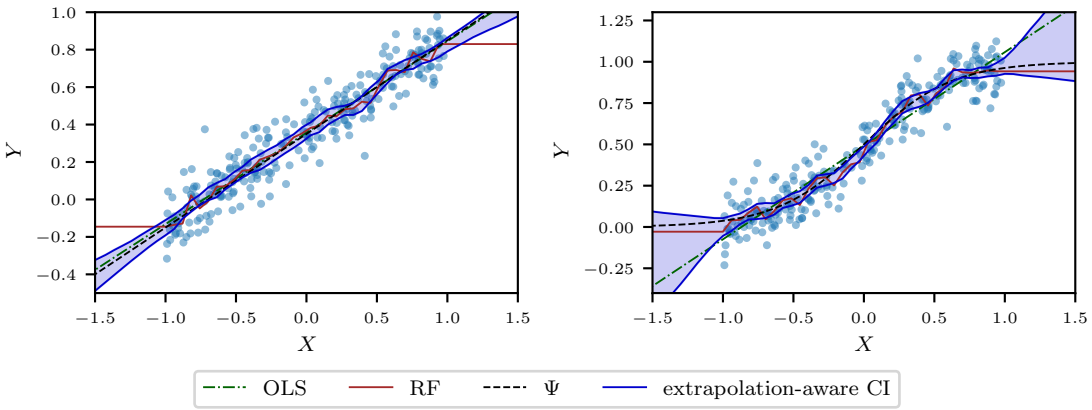


Figure 2: Linear (left) and nonlinear (right) conditional expectation models. In both cases, OLS regression (green, dotdashed) fits the data well (RMSE = 0.31 and RMSE = 0.30 respectively). However, the OLS only extrapolates well for the linear setting, while it is misleading in the nonlinear setting. Extrapolation-aware confidence intervals constructed using a percentile bootstrap and estimates of the extrapolation bounds based on random forest (RF) are able to capture the extrapolating behavior in both cases. In the linear setting it closely matches the OLS prediction with tight bounds, while in the nonlinear setting it captures the uncertainty outside of the data support.

3.3 Quantifying extrapolation

Beyond out-of-support prediction and uncertainty quantification, it is often useful to quantify the level of extrapolation itself. Simply measuring the minimal distance to observed samples does not always provide a good indicator of whether extrapolation is problematic for a given point x, particularly if X is multi-dimensional. Extrapolation bounds provide a way to quantify extrapolation that accounts for both the target of inference and the extrapolation assumptions. By Theorem 3, if the conditional function of interest Φ_0 is q-th derivative extrapolating, it holds for all $x \in \mathcal{X}$ that $\Phi_0(x)$ is identifiable from P_0 if

$$B^{\text{up}}_{\Phi_0,\mathcal{D}_{\text{in}}}(x) - B^{\text{lo}}_{\Phi_0,\mathcal{D}_{\text{in}}}(x) = 0.$$

We can therefore use the difference between the upper and lower extrapolation bounds to quantify extrapolation. This explicitly accounts for the target of inference Φ_0. For example, if Φ_0 only varies in the first coordinate and is constant in the remaining coordinates, the extrapolation bounds do not changes as long as the first coordinate is kept fixed. Hence, the bounds can overlap even if x is far away from any observed samples in Euclidean distance.

Depending on the conditional function under consideration, we propose to use different variations of this score. For example, when estimating a conditional expectation Ψ_0, we suggest using the extrapolation score $S : \mathcal{X} \to [0, \infty)$ defined for all $x \in \mathcal{X}$ by

$$S(x) := \frac{B^{\text{up}}_{\Psi_0,\mathcal{D}_{\text{in}}}(x) - B^{\text{lo}}_{\Psi_0,\mathcal{D}_{\text{in}}}(x)}{\sqrt{\mathbb{E}[(Y - \Psi_0(X))^2]}}.$$

At a point $x \in \mathcal{X}$ with $S(x) = 0$ there is no extrapolation, while for $S(x) > 0$ there is extrapolation. The normalization by the residual standard deviation allows us to interpret the score values more directly: if $S(x) = 1$ the extrapolation uncertainty in the conditional expectation function Ψ_0 is equal to the standard deviation of the residual noise. In particular for scores greater than one, we should be cautious when interpreting any point estimates as the error due to unidentifiablity of $\Phi_0(x)$ may be larger than the noise level.

3.4 A causal perspective on extrapolation

When extrapolating a conditional function Φ_0, we implicitly assume a two-step generative model that first selects X and then samples Y according to the distribution $Q_0(X, \cdot)$ specified by the Markov kernel. This generative model does not need to correspond to an underlying causal mechanism and should not, in general, be interpreted as such. If, however, one is interested in causal quantities, where X (or a part of it) is actively set to specific values rather than merely conditioned on, causal models provide a rigorous mathematical framework. Within this framework, the nonparametric approach to causal inference formalizes assumptions under which causal quantities can be expressed as functions of P_0 and hence estimated using observational methods. Since our extrapolation framework extends the conventional nonparametric model, it can also be applied to causal inference, enabling reasoning about causal quantities outside the observed support.

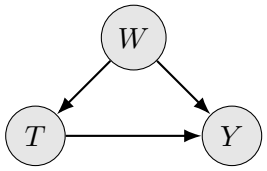
Remark 10 (Causal extrapolation) *In causal inference, the term extrapolation is sometimes used to refer to the task of generalizing from the observational to a previously unseen*

interventional distribution. To avoid confusion, we call such inference tasks causal extrapolation*. Causal extrapolation is in general different from the notion of extrapolation considered here because it does not necessarily (and in fact mostly does not) correspond to evaluating a conditional function outside of* $\operatorname{supp}(X)$*.*

We now illustrate how our extrapolation framework combines with causal models using a treatment-response example. We use structural causal models (SCMs) (Pearl, 2009) as they allow us to specify the function classes naturally, but the ideas transfer readily to other causal frameworks, e.g., the potential outcome model (Rubin, 2005).

Example 3 (Treatment-response extrapolation) *Let* M_0 *be an SCM over* $(X, Y) \in \mathcal{X} \times \mathbb{R}$*, where the covariates are divided into* $X = (T, W)$ *with treatment variables* $T \in \mathcal{T}$ *and pre-treatment covariates* $W \in \mathcal{W}$*:*

$$M_0 : \begin{cases} W \leftarrow \varepsilon_W \\ T \leftarrow h(W, \varepsilon_T) \\ Y \leftarrow g(W, T) + \varepsilon_Y. \end{cases}$$



The SCM M_0 *induces an observed distribution* P_0 *over* (T, W, Y) *and for all* $t \in \mathcal{T}$ *the interventional distributions* $P_{\operatorname{do}(T=t)}$ *over* (T, W, Y) *corresponding to the intervention* $\operatorname{do}(T = t)$ *that assigns treatment* t*. Given the SCM* M_0 *we can define causal quantities of interest. For example, if* T *is continuous, we could consider the dose-response curve* $\Phi_0^{\mathrm{DRC}} : \mathcal{T} \to \mathbb{R}$*, also called the average treatment function, defined for all* $t \in \mathcal{T}$ *by*

$$\Phi_0^{\mathrm{DRC}}(t) \coloneqq \mathbb{E}_{\operatorname{do}(T=t)}[Y],$$

where the subscript $\operatorname{do}(T = t)$ *denotes that the expectation is taken with respect to the distribution* $P_{\operatorname{do}(T=t)}$*. Similarly, if* T *is binary, we could consider the conditional average treatment effect* $\Phi_0^{\mathrm{CATE}} : \mathcal{W} \to \mathbb{R}$ *defined for all* $w \in \mathcal{W}$ *by*

$$\Phi_0^{\mathrm{CATE}}(w) \coloneqq \mathbb{E}_{\operatorname{do}(T=1)}[Y|W = w] - \mathbb{E}_{\operatorname{do}(T=0)}[Y|W = w].$$

Using the g-computation formula (Robins, 1986), it follows from the causal assumptions encoded in M_0 *that both quantities can be expressed as functions of* P_0*:*

$$\Phi_0^{\mathrm{DRC}}(t) = \mathbb{E}[\mathbb{E}[Y|T = t, W]] \quad \textit{and} \quad \Phi_0^{\mathrm{CATE}}(w) = \mathbb{E}[Y|T = 1, W = w] - \mathbb{E}[Y|T = 0, W = w].$$

This reduction from interventional to observational quantities is sometimes called causal extrapolation (see Remark 10). However, Φ_0^{DRC} *and* Φ_0^{CATE} *are only identified on* $\operatorname{supp}(T)$ *and* $\operatorname{supp}(W)$*, respectively. Our extrapolation framework enables inference beyond these supports if we assume the causal quantity of interest is* q*-th derivative extrapolating.*

The approach illustrated in Example 3 extends naturally to other causal inference settings. The key insight is that once a causal quantity is identified as a functional of the observed distribution P_0, our framework can be applied to extrapolate it outside the observed support, provided the identified functional satisfies the q-th derivative extrapolating assumption.

This applies beyond conditional means to other functionals. For instance, quantile treatment effects can be handled similarly (Chernozhukov and Hansen, 2005), replacing the conditional expectation with conditional quantiles. The framework also extends to settings with different identification strategies, such as nonparametric instrumental variable identification (Newey and Powell, 2003) or frontdoor identification (Pearl, 2009). In each case, the identification formula expresses the causal quantity as a function of observables, and our extrapolation bounds can then be applied to this identified functional. The extrapolation assumption must, of course, be assessed for the specific identified quantity, which may differ in structure from a simple conditional expectation.

4. Estimating extrapolation bounds

Assume n i.i.d. observations $(X_1, Y_1), \dots, (X_n, Y_n)$ from the distribution P_0 are observed and we want to estimate the extrapolation bounds $B^{\text{lo}}_{\Phi_0, \mathcal{D}_{\text{in}}}$ and $B^{\text{up}}_{\Phi_0, \mathcal{D}_{\text{in}}}$. By definition the extrapolation bounds are completely identified by P_0 since they can be exactly computed from the conditional function Φ_0 on the set $\mathcal{D}_{\text{in}}$. A natural plugin estimator is therefore given by first estimating Φ_0 with a q-times differentiable function $\widehat{\Phi}$ and then directly evaluating the bounds. In practice this can be difficult for two reasons: (i) Directly computing the extrapolation bounds involves two optimizations over the potentially unknown set $\mathcal{D}_{\text{in}}$ and (ii) existing nonparametric estimation procedures for estimating Φ_0 might not result in q-times differentiable functions and even if they do the derivatives may be ill-behaved.

We consider these two problems separately. First, in Section 4.1, we start by assuming access to a q-times differentiable estimate $\widehat{\Phi}_n$ that approximates Φ_0 and its derivatives sufficiently well. In that case, we show that the extrapolation bounds can be estimated consistently, by performing the two optimizations over the sample points $X_1, \dots, X_n$ only. Second, in Section 4.2, we propose a procedure based on random forests and local polynomials that uses only $(X_1, \widehat{\Phi}_n(X_1)), \dots, (X_n, \widehat{\Phi}_n(X_n))$ to estimate directional derivatives $D^k_v \widehat{\Phi}_n(X_1), \dots D^k_v \widehat{\Phi}_n(X_1)$ for arbitrary directions v and orders k. Finally, in Section 4.3, we combine the procedures from Sections 4.1 and 4.2 in a computationally efficient way.

4.1 Extrapolation bounds from differentiable estimates

Let $\widehat{\Phi}_n$ be a q-times differentiable estimate of Φ_0 based on the data $(X_1, Y_1), \dots, (X_n, Y_n)$. Then, at an arbitrary point $x \in \mathcal{X}$, we propose to estimate the extrapolation bounds by plugging the differentiable estimate $\widehat{\Phi}_n$ into the definition and optimizing only over the observed samples $X_1, \dots, X_n$. More formally, define

$$\widehat{B}^{\text{lo}}_n(x) \coloneqq B^{\text{lo}}_{\widehat{\Phi}_n, \{X_1, \dots, X_n\}}(x) \quad \text{and} \quad \widehat{B}^{\text{up}}_n(x) \coloneqq B^{\text{up}}_{\widehat{\Phi}_n, \{X_1, \dots, X_n\}}(x). \tag{12}$$

Using multi-index notation, the lower extrapolation bound estimate can be expressed as

$$\widehat{B}^{\text{lo}}_n(x) \coloneqq \max_{i \in \{1, \dots, n\}} \left(\sum_{\ell=0}^{q-1} \sum_{|\boldsymbol{\alpha}|=\ell} \partial^{\boldsymbol{\alpha}} \widehat{\Phi}_n(X_i) \tfrac{(x - X_i)^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!} + \min_{k \in \{1, \dots, n\}} \sum_{|\boldsymbol{\alpha}|=q} \partial^{\boldsymbol{\alpha}} \widehat{\Phi}_n(X_k) \tfrac{(x - X_i)^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!} \right). \tag{13}$$

Similarly, the upper extrapolation bound estimate can be expressed as

$$\widehat{B}_n^{\mathrm{up}}(x) \coloneqq \min_{i\in\{1,\dots,n\}} \left(\sum_{\ell=0}^{q-1} \sum_{|\boldsymbol{\alpha}|=\ell} \partial^{\boldsymbol{\alpha}} \widehat{\Phi}_n(X_i) \tfrac{(x-X_i)^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!} + \max_{k\in\{1,\dots,n\}} \sum_{|\boldsymbol{\alpha}|=q} \partial^{\boldsymbol{\alpha}} \widehat{\Phi}_n(X_k) \tfrac{(x-X_i)^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!} \right). \tag{14}$$

From these expressions, it can be seen that, instead of evaluating all possible directional derivatives, the estimates can be computed by only evaluating $\partial^{\boldsymbol{\alpha}}\widehat{\Phi}_n(X_i)$ once at every observation X_i and every partial derivative $\partial^{\boldsymbol{\alpha}}$. If $q = 1$, for example, this means it is sufficient to evaluate all $\partial^j\widehat{\Phi}_n(X_i)$ corresponding to nd evaluations instead of n^2 evaluations for $D_{\overline{v}(x,X_j)}\widehat{\Phi}(X_i)$. This becomes particularly beneficial if the bounds are evaluated at many target points $x \in \mathcal{X}$. As shown in the following theorem, the estimates are consistent if $\widehat{\Phi}_n$ and all partial derivatives $\partial^{\boldsymbol{\alpha}}\widehat{\Phi}_n$ up to order q are uniformly consistent on $\mathcal{D}_{\mathrm{in}}$, a condition satisfied, for example, by local polynomial estimators (Masry, 1996, Theorem 6).

Theorem 11 (Consistency of extrapolation bound estimates) *Assume $\mathcal{X}$ is compact and let $\Phi_0 : \mathcal{X} \to \mathbb{R}$ be a conditional function satisfying $\Phi_0 \in C^{q+1}(\mathcal{X})$. For all $n \in \mathbb{N}$, let $\widehat{\Phi}_n$ be a q-times differentiable estimate of Φ_0 based on n i.i.d. observations $(X_1, Y_1), \dots, (X_n, Y_n) \sim P_0$ satisfying for all $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $|\boldsymbol{\alpha}| \leq q$ that*

$$\sup_{x\in\mathcal{D}_{\mathrm{in}}} \left|\widehat{\Phi}_n(x) - \Phi_0(x)\right| \xrightarrow{P_0} 0 \quad \textit{and} \quad \sup_{x\in\mathcal{D}_{\mathrm{in}}} \left|\partial^{\boldsymbol{\alpha}}\widehat{\Phi}_n(x) - \partial^{\boldsymbol{\alpha}}\Phi_0(x)\right| \xrightarrow{P_0} 0 \quad \textit{as } n \to \infty.$$

Additionally, for all $n \in \mathbb{N}$ and all $x \in \mathcal{X}$, let $\widehat{B}_n^{\mathrm{lo}}(x)$ and $\widehat{B}_n^{\mathrm{up}}(x)$ be defined as in (12). Then, for all $x \in \mathcal{X}$, it holds that

$$\left|B^{\mathrm{lo}}_{\Phi_0,\mathcal{D}_{\mathrm{in}}}(x) - \widehat{B}_n^{\mathrm{lo}}(x)\right| \xrightarrow{P_0} 0 \quad \textit{and} \quad \left|B^{\mathrm{up}}_{\Phi_0,\mathcal{D}_{\mathrm{in}}}(x) - \widehat{B}_n^{\mathrm{up}}(x)\right| \xrightarrow{P_0} 0 \quad \textit{as } n \to \infty.$$

A proof of Theorem 11 is given in Supplementary Material D.5. Optimal rates for (non-adaptive) derivative estimation are of order $n^{-(p-q)/(2p+d)}$ for estimating the q-th derivative of a p-times differentiable function (Stone, 1982), illustrating that larger q requires more samples to achieve reliable estimates, particularly in higher dimensions. An immediate implication of the consistency in Theorem 11 is asymptotic validity of the prediction intervals discussed in Section 3.2.

Corollary 12 *Assume $\mathcal{X}$ is compact, fix $\alpha \in (0, 1)$ and assume the Markov kernel Q_0 is such that the conditional quantiles $\mathcal{T}_0^{\alpha/2}$ and $\mathcal{T}_0^{1-\alpha/2}$ are both q-th derivative extrapolating. Denote by $\widehat{B}^{\mathrm{lo}}_{n;\alpha}$ and $\widehat{B}^{\mathrm{up}}_{n;\alpha}$ estimates of the lower extrapolation bound of $\mathcal{T}_0^{\alpha/2}$ and the upper extrapolation bound of $\mathcal{T}_0^{1-\alpha/2}$, respectively. Furthermore, assume that the estimates satisfies the consistency in Theorem 11 and define for all $x \in \mathcal{X}$ the intervals*

$$\widehat{C}^{\mathrm{pred}}_{n;\alpha}(x) \coloneqq \left[\widehat{B}^{\mathrm{lo}}_{n;\alpha}(x), \widehat{B}^{\mathrm{up}}_{n;\alpha}(x)\right].$$

Then, it holds for all $x \in \mathcal{X}$ that

$$\liminf_{n\to\infty} \mathbb{P}_{Q_0(x,\cdot)}\left(Y_x \in \widehat{C}^{\mathrm{pred}}_{n;\alpha}(x)\right) \geq 1 - \alpha.$$

The assumptions on uniform consistency over $\mathcal{D}_{\text{in}}$ can be relaxed by making a stronger extrapolation assumption: for example, one may require only uniform consistency over the interior of $\mathcal{D}_{\text{in}}$ but then require that the derivatives outside the interior are dominated by the ones within. We skip the straightforward modifications.

Whenever a q-times differentiable estimator $\widehat{\Phi}_n$ is available the above estimates for the lower and upper bounds can be used. For conditional expectations, multiple methods have been proposed that either provide differentiable estimates (e.g., Härdle and Stoker, 1989; Mack and Müller, 1989; Wahba, 1990) or that estimate the corresponding derivatives separately (e.g., Wang and Lin, 2015; Dai et al., 2016). These methods are, however, generally constructed only for conditional expectations and do not apply to other conditional functions Φ_0. Furthermore, they are targeted towards specific estimation procedures and often only work for univariate X, making them inapplicable to modern applications, where state-of-the-art performance is achieved with nonparametric machine learning procedures. Unfortunately, those machine learning estimates, in general, cannot be used directly as they are either not smooth (e.g., random forests or boosted trees) or the derivatives of the resulting estimates are ill-behaved without additional regularization (e.g., neural networks or support vector machines) (e.g., De Brabanter et al., 2013). One possible solution is to use procedures that start from a potentially non-differentiable pilot estimate $\widehat{\Phi}_n$ and smooth the estimate such that the smoothed estimate has well-behaved derivatives. Such a procedure based on kernel-smoothing has been proposed by Klyne and Shah (2023). We propose a related approach but instead of directly smoothing $\widehat{\Phi}_n$ we only use the predictions $\widehat{\Phi}_n(X_1), \dots, \widehat{\Phi}_n(X_n)$ to estimate the required derivatives. A related procedure was also used by Lundborg and Pfister (2023).

4.2 Estimating directional derivatives

We now introduce a procedure for estimating directional derivatives based only on

$$(X_1, \widehat{\Phi}_n(X_1)), \dots, (X_n, \widehat{\Phi}_n(X_n)), \tag{15}$$

where $\widehat{\Phi}_n$ is a potentially non-differentiable estimate of the conditional function Φ_0. We omit a detailed statistical analysis of the procedure as this goes beyond the scope of this article and instead argue heuristically and empirically (see Section 5) that the proposed procedure has several properties making it amenable to our application.

Throughout this section we fix a direction[2] $v \in \mathcal{B}$ and an order $k \in \{1, \dots, q\}$ and aim to estimate the directional derivatives $D_v^k \Phi_0(X_1), \dots, D_v^k \Phi_0(X_n)$ from (15). Derivative estimation is well-known to be statistically challenging, particularly in high-dimensions. Our proposal aims to overcome these challenges by combining random forests with local polynomials. Local polynomials are among the most prominent methods used for derivative estimation. The idea is to estimate the derivative $D_v^k \Phi_0(X_i)$ using a polynomial p of order $q+1$ defined for all $x \in \mathcal{X}$ by

$$p(x) = \sum_{j=0}^{q+1} \beta_j ((x - X_i)^\top v)^j.$$

2. For our proposed implementation of the first-order derivative estimation, we only need the directions $v \in \{e_1, \dots, e_d\}$, where e_j denotes the j-th unit vector.

Then, if for all $\ell \in \{1, \dots, q+1\}$ it holds that $\beta_\ell = \frac{D_v^\ell \Phi_0(X_i)}{\ell!}$, Taylor's theorem implies that p is a good approximation of Φ_0 around X_i in the direction v or, more formally, for all $h \in \mathbb{R}$ close to zero, $\Phi_0(X_i + hv) = p(X_i + hv) + \mathcal{O}(h^{q+2})$. Based on this observation, we estimate coefficients $\beta_0, \dots, \beta_{q+1}$ such that the polynomial p is a good local approximation of Φ_0 around X_i and then use them to estimate the directional derivative. More concretely, for a given weight matrix $W \in \mathbb{R}^{n \times n}$, we minimize the weighted mean squared loss

$$\hat{\beta}(i) \coloneqq \operatorname*{arg\,min}_{\beta \in \mathbb{R}^{q+1}} \sum_{\ell=1}^{n} \left(\widehat{\Phi}_n(X_\ell) - \sum_{j=0}^{q+1} \beta_j ((X_\ell - X_i)^\top v)^j \right)^2 W_{i,\ell}. \tag{16}$$

Then, using the estimated coefficients $\hat{\beta}(i)$, we can estimate $D_v^k \Phi_0(X_i)$ as

$$\widehat{D_v^k \Phi}_n(X_i) \coloneqq k! \, \hat{\beta}_k(i). \tag{17}$$

Using local polynomials to estimate derivatives of conditional expectations has been analyzed extensively in the literature (e.g., Masry and Fan, 1997; De Brabanter et al., 2013), however using Y_ℓ's instead of $\widehat{\Phi}_n(X_\ell)$'s in (16). Most existing approaches use kernel weights, i.e., $W_{i,\ell} = k((X_i - X_\ell)/\sigma)$ for a kernel function k and a bandwidth $\sigma > 0$. For our purposes, kernel weights are not ideal for two reasons. Firstly, kernel weights can perform poorly in higher dimensions and secondly, require careful tuning of the bandwidth parameter.

To avoid these issues, we instead suggest to use weights constructed by a random forest with a modified splitting rule (Lin and Jeon, 2006; Meinshausen, 2006; Athey et al., 2019). Intuitively, the weights at an observation i, that is $W_{i,1}, \dots, W_{i,n}$, should upweight a large set of observations $\ell_1, \dots, \ell_m$ for which $(v^\top X_{\ell_1}, \widehat{\Phi}_n(X_{\ell_1})), \dots, (v^\top X_{\ell_m}, \widehat{\Phi}_n(X_{\ell_m}))$ and $(v^\top X_i, \widehat{\Phi}_n(X_i))$ all can be (approximately) described by the same polynomial of order $q+1$. Such weights can be constructed in a greedy fashion by using a random forest with the following modified splitting rule: For each proposed split, fit a polynomial of order $q+1$ with $v^\top X$ as argument on each child node and use the residual sum of squares across both child nodes as impurity measure. We denote this type of random forest with polynomial splitting in direction v by `rfpoly-`v. The fitted random forest regression function $\widehat{\mu}$ can be expressed as

$$\widehat{\mu}(x) = \sum_{i=1}^{n} \widehat{w}_i(x) \widehat{\Phi}_n(X_i),$$

where $\widehat{w}_i : \mathcal{X} \to [0, 1]$ are weight functions that are given by

$$\widehat{w}_i(x) = \frac{1}{M} \sum_{k=1}^{M} \frac{\mathbb{1}(i \in \widehat{\mathcal{L}}_k(x))}{|\widehat{\mathcal{L}}_k(x)|},$$

with M the number of trees in the random forest and $\widehat{\mathcal{L}}_k(x)$ the sample indices specified by the k-th tree's terminal node in which x lies. Based on these weights, we then define for all $i, \ell \in \{1, \dots, n\}$ the weights $W_{i,\ell} \coloneqq \widehat{w}_i(X_\ell)$ and use them in the local polynomial derivative estimation. The full procedure is detailed in Algorithm 1 which includes an additional regularization step discussed in the following section.

Algorithm 1: `RFLocPol`

Input : Data $(X_1, \widehat{\Phi}_n(X_1)), \dots, (X_n, \widehat{\Phi}_n(X_n))$, order k, direction v
Tuning : Penalty λ, `rf` parameters Γ
Output: Directional derivative estimates $\widehat{D_v^k \Phi}_n(X_1), \dots, \widehat{D_v^k \Phi}_n(X_n)$

$\widehat{\mu} \leftarrow$ `rfpoly`-v on $(X_1, \widehat{\Phi}_n(X_1)), \dots, (X_n, \widehat{\Phi}_n(X_n))$ with parameters Γ
Extract weight matrix $W = (w_i(X_\ell))_{i,\ell}$ from $\widehat{\mu}$
for $i \in \{1, \dots, n\}$ **do**
  $\hat{\beta} \leftarrow$ coefficients of order $q+1$ local polynomial fit in (18) with penalty λ
  $\widehat{D_v^k \Phi}_n(X_i) \leftarrow k!\, \hat{\beta}_{i,k}$
end

4.2.1 Additional regularization and tuning of hyperparameters

Since the function Φ_0 is assumed to be continuously differentiable up to order q, it can be beneficial to regularize the local polynomial estimate in (16) to ensure the derivatives become smoother. We propose to do this by estimating all coefficients $\hat{\beta} = (\hat{\beta}(1), \dots, \hat{\beta}(n))$ simultaneously by minimizing the penalized weighted mean squared loss

$$\hat{\beta} \coloneqq \underset{\beta \in \mathbb{R}^{n \times (q+1)}}{\arg\min} \sum_{i,\ell=1}^{n} \left(\widehat{\Phi}_n(X_\ell) - \sum_{j=0}^{q+1} \beta_{i,j} ((X_\ell - X_i)^\top v)^j \right)^2 W_{i,\ell} + \lambda P_W(\beta), \tag{18}$$

where the penalty term P_W is defined for all $\beta \in \mathbb{R}^{n \times (q+1)}$ by

$$P_W(\beta) \coloneqq \sum_{i=1}^{n} \sum_{j=1}^{q+1} \left(\sum_{\ell=1}^{n} (j!\, \beta_{i,j} - j!\, \beta_{\ell,j}) W_{i,\ell} \right)^2.$$

By (17) the term $j!\, \beta_{i,j}$ parametrizes the derivative $D_v^j \Phi_0(X_i)$, which implies that the penalty term P_W penalizes large differences between the derivatives at each point i and the locally averaged derivatives close to i. This penalty therefore enforces smoothness of the derivatives.

Including this penalization the full `RFLocPol` procedure depends on two types of hyperparameters; the penalty parameter λ from the penalized local polynomial and the random forest parameters Γ used to fit the random forests `rfpoly`-v. Both parameters substantially affect the performance of the overall procedure and need to be selected carefully. We suggest a heuristic tuning procedure that selects optimal parameters (Γ^*, λ^*) from a K-tuple $(\Gamma_1, \dots, \Gamma_K)$ of random forest parameters and a L-tuple $(\lambda_1, \dots, \lambda_L)$ of penalty parameters. For this we assume that the tuples are both ordered with decreasing regularization stength, i.e., Γ_i regularizes more than Γ_j and $\lambda_i > \lambda_j$ for all $i < j$. For the random forest parameters, we could for example use an increasing sequence of maximal depths or decreasing minimal node sizes. We then apply `RFLocPol` for all different parameter settings and select the parameters for which the local polynomial estimates $\hat{\beta}_0(1), \dots, \hat{\beta}_0(n)$ are not significantly worse than $\widehat{\Phi}_n(X_1), \dots, \widehat{\Phi}_n(X_n)$, measured by a given loss function. Full details on this tuning procedure are provided in Algorithm 3 in Supplementary Material A.

4.3 Xtrapolation

We now adapt the plug-in estimates for the extrapolation bounds from Section 4.1 to use the forest-weighted local polynomial derivative estimates from Section 4.2 in a computationally efficient way. This leads to a procedure, which we call `Xtrapolation`, that can estimate the extrapolation bounds from arbitrary and potentially non-differentiable pilot estimates $\widehat{\Phi}_n$.

Since the `RFLocPol` procedure estimates directional derivatives it does not directly apply to the plug-in estimates in (13) and (14) which are expressed in terms of partial derivatives. A workaround is to consider plug-in estimates based on directional derivatives instead, however this relies on computing directional derivatives in n different directions which would involve n random forest fits. As this is computationally infeasible in practice, we only focus on two special cases: The order-one case (i.e., $q = 1$ and arbitrary d) and the one-dimensional case (i.e., $d = 1$ and arbitrary q).[3] In both cases the partial derivatives correspond to directional derivatives and hence the plug-in estimates can be combined with `RFLocPol`. More specifically, for the order-one case the plug-in estimates only involve first order partial derivatives $\partial_1\widehat{\Phi}_n, \dots, \partial_d\widehat{\Phi}_n$ which are equal to the directional derivatives in the directions $v \in \{e_1, \dots, e_d\}$. Similarly, for the one-dimensional case all involved partial derivatives correspond to the directional derivatives in the direction $v = 1$.

When using `RFLocPol` to estimate derivatives it can happen that the derivative estimates are not equal to the derivatives of the original estimate (assuming they even exist). As a consequence, it is no longer guaranteed that the lower extrapolation bound estimate is smaller than the upper bound estimate. To enforce this constraint, we propose to check at a specific target point whether the lower estimate is indeed smaller than the upper estimate and if not to set both estimates to the average of the lower and upper extrapolation bound estimate. The full `Xtrapolation` procedure for the order-one case is detailed in Algorithm 2. The version for the one-dimensional case is very similar and provided in Algorithm 4 in Supplementary Material A.

4.3.1 Computational speed up

We now consider a modification of the default `Xtrapolation` procedure that speeds up the computation for large sample sizes n and many target points m. The default procedure has $\mathcal{O}(nm)$ complexity, which can be reduced by considering only a subset of all possible anchor points (i.e., the loop in line 6 of Algorithm 2). A naive approach would be to subsample random anchor points, but as is clear in the one-dimensional case bounds are generally tighter for anchor points close to the target point. It can therefore be beneficial to select anchor points based on closeness to the target points. While in one-dimensional settings using the Euclidean distance is an obvious choice, it becomes more subtle in multi-dimensional settings. This is because points that are far away in Euclidean distance may have tight bounds if they are only far away in directions in which the variance of the observed directional derivatives is small. Therefore, to capture a more meaningful notion of closeness (for $q = 1$), we propose a first-order derivative scaled Euclidean distance. Denote by $\widehat{\nabla\Phi}_n(X) \in \mathbb{R}^{n\times d}$ the matrix with (i, j)-th entry $\widehat{\partial_j\Phi}_n(X_i)$ and let $V\Sigma V^\top$ be the

3. The remaining cases, in which both $q > 1$ and $d > 1$, are computationally and statistically difficult and we do not expect direct application of our framework to work well in these cases.

Algorithm 2: `Xtrapolation` (order-one version)

Input : Estimates $\widehat{\Phi}_n(X_1), \dots, \widehat{\Phi}_n(X_n)$, data $X_1, \dots, X_n$, target points $\bar{x}_1, \dots, \bar{x}_m$
Tuning : Penalty λ, `rf` parameters Γ
Output: Extrapolation bound estimates $\widehat{B}^{\mathrm{lo}}(\bar{x}_1), \dots, \widehat{B}^{\mathrm{lo}}(\bar{x}_m)$, $\widehat{B}^{\mathrm{up}}(\bar{x}_1), \dots, \widehat{B}^{\mathrm{up}}(\bar{x}_m)$

$\mathcal{D} \leftarrow (X_1, \widehat{\Phi}_n(X_1)), \dots, (X_1, \widehat{\Phi}_n(X_n))$
for $j \in \{1, \dots, d\}$ **do**
  $\widehat{\partial_j \Phi}_n(X_1), \dots, \widehat{\partial_j \Phi}_n(X_n) \leftarrow$ `RFLocPol`$(\mathcal{D}, k = 1, v = e_j, \lambda = \lambda, \Gamma = \Gamma)$
end
for $j \in \{1, \dots, d\}$ **do**
  $\widehat{\partial_j \Phi}_n(X_1), \dots, \widehat{\partial_j \Phi}_n(X_n) \leftarrow$ `RFLocPol`$(\mathcal{D}, k = 1, v = e_j, \lambda = \lambda, \Gamma = \Gamma)$
end
for $\ell \in \{1, \dots, m\}$ **do**
  for $i \in \{1, \dots, n\}$ **do**
    $S \leftarrow \left\{ \sum_{j=1}^d \widehat{\partial_j \Phi}_n(X_1)^\top (\bar{x}_\ell^j - X_i^j), \dots, \sum_{j=1}^d \widehat{\partial_j \Phi}_n(X_n)^\top (\bar{x}_\ell^j - X_i^j) \right\}$
    $B_i^{\mathrm{lo}} \leftarrow \widehat{\Phi}_n(X_i) + \min(S) \quad \text{and} \quad B_i^{\mathrm{up}} \leftarrow \widehat{\Phi}_n(X_i) + \max(S)$
  end
  $\widehat{B}^{\mathrm{lo}}(\bar{x}_\ell) \leftarrow \max_i B_i^{\mathrm{lo}} \quad \text{and} \quad \widehat{B}^{\mathrm{up}}(\bar{x}_\ell) \leftarrow \min_i B_i^{\mathrm{up}}$
  if $\widehat{B}^{\mathrm{lo}}(\bar{x}_\ell) > \widehat{B}^{\mathrm{up}}(\bar{x}_\ell)$ **then**
    $\widehat{B}^{\mathrm{lo}}(\bar{x}_\ell) \leftarrow \big(\widehat{B}^{\mathrm{lo}}(\bar{x}_\ell) + \widehat{B}^{\mathrm{up}}(\bar{x}_\ell)\big)/2$
    $\widehat{B}^{\mathrm{up}}(\bar{x}_\ell) \leftarrow \widehat{B}^{\mathrm{lo}}(\bar{x}_\ell)$
  end
end

eigenvalue decomposition of the estimated covariance

$$\widehat{\nabla\Phi}_n(X)^\top \widehat{\nabla\Phi}_n(X) - \left(\frac{1}{n}\sum_{i=1}^n \widehat{\nabla\Phi}_n(X_i)\right)^\top \left(\frac{1}{n}\sum_{i=1}^n \widehat{\nabla\Phi}_n(X_i)\right).$$

To measure closeness of a sample point X_i to a target point $\bar{x}_k$, we use the distance $\|V\Sigma^{1/2}(X_i - \bar{x}_k)\|_2$. This distance is larger in directions where the derivatives vary substantially and smaller in directions where they remain fixed. Alternatively, one can use a distance measure induced by random forest weights.

4.3.2 Allowing for categorical covariates

In some applications, not all covariates X are continuous, which means that the framework does not apply directly. However, it can be adapted in settings where $X = (Z, W)$ with $Z \in \mathcal{Z} \subseteq \mathbb{R}^{d_Z}$ continuous and $W \in \mathcal{W} \subseteq \mathbb{R}^{d_W}$ categorical, as long as no extrapolation occurs in the categorical predictors. The idea is to apply the framework conditional on W. More specifically, we can assume that for all $w \in \mathcal{W}$ the conditional function Φ_0 satisfies that $z \mapsto \Phi_0((z, w))$ is q-derivative extrapolating and derive the same extrapolation bounds but conditional on $W = w$. A straightforward modification of Algorithm 2 is to select the anchor points used in the loop of line 6 based on random forest weights. If the random forest is grown sufficiently deep, one can expect that samples with different W values for which Φ_0 is sufficiently different will have small weights and hence not be used as anchor points. This approach is used in the real data example in Section 5.2 below.

5. Numerical experiments

We now present numerical experiments in which we investigate the performance of the proposed estimation procedure (Section 5.1) and demonstrate possible applications of the extrapolation bounds on real data (Section 5.2). All experiments can be reproduced using the publicly available code at `https://github.com/NiklasPfister/ExtrapolationAware-Inference`, which contains an implementation of `Xtrapolation` based on the Python package `adaXT` (Brodersen and Pfister, 2024). For the regressions and cross-validation we used the Python packages `scikit-learn` (Pedregosa et al., 2011) and `quantile-forest` (Johnson, 2024).

5.1 Simulation experiments

We begin by empirically analyzing the proposed `Xtrapolation` procedure on simulated data. To this end, we consider random data generating models for different sample sizes n and dimensions d. For each simulation, we randomly choose $\mathcal{D}_{\text{in}} \subseteq [-2, 2]^d$ and $f : [-2, 2]^d \to \mathbb{R}$ and then generate n i.i.d. copies $(X_1, Y_1), \ldots (X_n, Y_n)$ of $(X, Y) \in [-2, 2]^d \times \mathbb{R}$ defined via

$$X \sim \text{Unif}(\mathcal{D}_{\text{in}}) \quad \text{and} \quad Y = f(X) + \tfrac{1}{10}\varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, 1)$ independent of X. The way we select the set $\mathcal{D}_{\text{in}}$ and the function f is such that the extrapolation assumptions are satisfied and it is easy to interpret the results. More specifically, we select $\mathcal{D}_{\text{in}}$ and f sequentially as follows.

(1) *Selection of $\mathcal{D}_{\text{in}}$:* Define the intervals $I_1 \coloneqq [-2,-1)$, $I_2 \coloneqq [-1,0)$, $I_3 \coloneqq [0,1)$ and $I_4 \coloneqq [1,2]$. Sample d sets $C_1,\dots,C_d$ uniformly from $\{[-2,2]\setminus I_1,\dots,[-2,2]\setminus I_4\}$ (with replacement) and define $\mathcal{D}_{\text{in}} \coloneqq \bigtimes_{j=1}^{d} C_j$ and $\mathcal{D}_{\text{out}} \coloneqq [-2,2]^d \setminus \mathcal{D}_{\text{in}}$.

(2) *Selection of f:* Let f be a piecewise linear function in the first coordinate such that for all $x \in [-2,2]^d$ it holds

$$f(x) = \sum_{j=1}^{4} (s_j x^1 + c_j) \mathbb{1}_{I_j}(x^1),$$

where $s_1,\dots,s_4$ are drawn randomly (see Supplementary Material C.1) and $c_1,\dots,c_4$ are selected such that f is continuous and $f(-2) = 0$. Importantly, the slopes $s_1,\dots,s_4$ are drawn in such a way that $f \lhd^1_{\mathcal{D}_{\text{in}}} f$ almost surely and $\text{var}(f(U)) = 1$ for U uniform on $[-2,2]^d$.

This sampling procedure leads to models for which the conditional expectation Ψ_0 of Y given X is first derivative extrapolating almost everywhere. Furthermore, there are two types of extrapolation scenarios that can happen depending on $\mathcal{D}_{\text{in}}$ and f. Either the lower and upper extrapolation bounds are equal on $\mathcal{D}_{\text{out}}$ implying that the extrapolation assumption identifies Ψ_0 on $\mathcal{D}_{\text{out}}$ or they do not coincide on $\mathcal{D}_{\text{out}}$ in which case Ψ_0 is not identifiable on $\mathcal{D}_{\text{out}}$. The first case occurs if one of the inner intervals (i.e., I_2 or I_3) is missing in the first coordinate and f has the smallest or largest slope on that interval. Examples are shown in Figure 13 in Supplementary Material C.1.

For the following experiments, we generate 50 datasets for all combinations of $n \in \{100, 200, 400, 600, 800, 1600\}$ and $d \in \{2, 8\}$, where each dataset is sampled with randomly selected $\mathcal{D}_{\text{in}}$ and f as described above. We then consider four regression procedures: Random forest regression (`rf`), support vector regression (`svr`), neural network regression (`mlp`) and ordinary least square regression (`ols`). For all procedures – except `ols` – we tune hyperparameters using a 5-fold cross-validation and additionally screen for variables using random forest based Gini impurity (see Supplementary Material C.2). On top of each regression fit, we then apply `Xtrapolation` with order $q = 1$ and the parameter tuning discussed in Section 4.2.1 (see Supplementary Material C.3) to predict the lower and upper extrapolation bounds $\widehat{B}^{\text{lo}}_n(x)$ and $\widehat{B}^{\text{up}}_n(x)$ for all $x \in \widehat{\mathcal{D}}_{\text{in}} \cup \widehat{\mathcal{D}}_{\text{out}}$, where $\widehat{\mathcal{D}}_{\text{in}}$ consists of 200 uniformly sampled points on $\mathcal{D}_{\text{in}}$ and $\widehat{\mathcal{D}}_{\text{out}}$ consists of 200 uniformly sampled points on $\mathcal{D}_{\text{out}}$.

Estimation accuracy of extrapolation bounds We first assess how accurately the extrapolation bounds for different regression procedures are estimated and empirically validate Theorem 11. We evaluate the accuracy of the estimated bounds by comparing them with oracle extrapolation bounds (i.e., using the true function f but only optimizing over the anchor points $X_1,\dots,X_n$) evaluated at the same new observations. Formally, we consider the root mean squared error (RMSE) given by

$$\sqrt{\frac{1}{200}\sum_{x\in\mathcal{D}} \left(\widehat{B}^{\text{lo}}_n(x) - B^{\text{lo}}_{f,\{X_1,\dots,X_n\}}(x)\right)^2} + \sqrt{\frac{1}{200}\sum_{x\in\mathcal{D}} \left(\widehat{B}^{\text{up}}_n(x) - B^{\text{up}}_{f,\{X_1,\dots,X_n\}}(x)\right)^2}, \quad (19)$$

where $\mathcal{D} = \widehat{\mathcal{D}}_{\text{in}}$ or $\mathcal{D} = \widehat{\mathcal{D}}_{\text{out}}$. The results are shown in Figure 3. As expected the extrapolation bounds based on `ols` do not converge. In contrast, for all three nonparametric

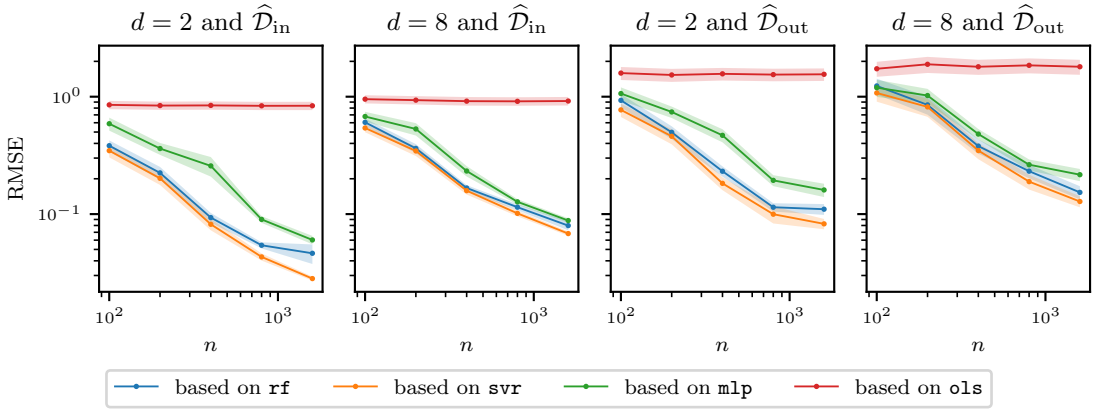


Figure 3: Accuracy of estimated extrapolation bounds measured using RMSE given in (19). For all three nonparametric regression procedures the RMSE decays with increasing n. As expected the extrapolation bounds based on `ols` do not decay as a linear regression cannot approximate the piecewise linear conditional expectations on $\mathcal{D}_{\text{in}}$.

estimators `rf`, `svr` and `mlp` the extrapolation bounds converge both on $\mathcal{D}_{\text{in}}$ and $\mathcal{D}_{\text{out}}$. Moreover, while the increased dimension leads to slightly worse accuracy the RMSE decays at a similar rate. This is particularly interesting as the `Xtrapolation` procedure does not explicitly take sparsity into account, but appears to automatically adapt to the sparsity in the regression estimates (due to the variable screening). We see this as promising empirical evidence that the random forest weights ensure that the derivatives are estimated well even in multiple dimensions.

Out-of-support prediction Next we show how the extrapolation bounds can be used to construct regression-agnostic predictions on $\mathcal{D}_{\text{out}}$ that are worst-case optimal as discussed in Section 3.1. We use the same simulations with $n = 1600$ and $d = 2$ but now additionally estimate (9) in Section 3.1 by

$$\widehat{f}_{\text{xtra}}(x) = \frac{\widehat{B}_n^{\text{lo}}(x) + \widehat{B}_n^{\text{up}}(x)}{2}. \tag{20}$$

We then compare this estimate with regression estimates $\widehat{f}_{\text{reg}}$ resulting from the plain regressions. We evaluate the performance using the worst-case RMSE given by

$$\frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \sup_{Q \in \mathcal{Q}_0} \sqrt{\mathbb{E}_{Q(x,\cdot)}[(Y_x - \widehat{f}(x))^2]}, \tag{21}$$

where $\widehat{f} \in \{\widehat{f}_{\text{xtra}}, \widehat{f}_{\text{reg}}\}$ and $\mathcal{D} \in \{\widehat{D}_{\text{in}}, \widehat{D}_{\text{out}}\}$. As its not directly possible to evaluate this loss, we use that the worst-case $Q \in \mathcal{Q}_0$ is attained at the true extrapolation bounds (see proof of Proposition 7) and then approximate the loss using the oracle extrapolation bounds. The results are shown in Figure 4, We additionally distinguish between identifiable

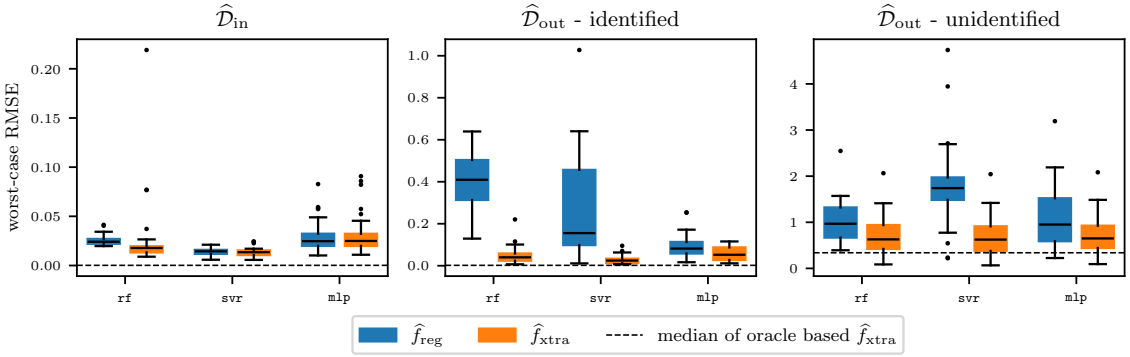


Figure 4: Comparison of worst-case RMSE (see (21)) for $\widehat{f}_{\text{xtra}}$ and $\widehat{f}_{\text{reg}}$ on $\mathcal{D}_{\text{in}}$ (left) and $\mathcal{D}_{\text{out}}$ (middle and right). The out-of-support predictions are separated into simulations for which the oracle lower and upper bound agree, i.e., the conditional expectation is identified (middle) and those that are not (right). While the out-of-support predictions of plain regression $\widehat{f}_{\text{reg}}$ depend on the used regression procedure this is not the case for the estimates $\widehat{f}_{\text{xtra}}$ based on `Xtrapolation`.

and non-identifiable extrapolation settings (i.e., where Ψ_0 is identified on $\mathcal{D}_{\text{out}}$ and where not) by splitting the 50 simulations into either identifiable (22 simulations) and unidentifiable settings (28 simulations) depending on whether the oracle bounds are approximately equal on the evaluated points or not. We observe that on $\mathcal{D}_{\text{in}}$ both plain regression and `Xtrapolation` perform similarly. The largest difference occurs for `rf`, which makes sense as `Xtrapolation` smooths the estimates which has almost no effect on the already smooth `svr` and `mlp` estimates but slightly improves the non-smooth `rf` estimates. Furthermore, while the regression estimates extrapolate differently on $\mathcal{D}_{\text{out}}$, the differences disappear after applying `Xtrapolation` which is expected as the extrapolation estimates target the same quantities in all cases.

Quantifying extrapolation Finally, we consider the proposed extrapolation score from Section 3.3. To this end, we consider the simulations with $n = 1600$ and $d \in \{2, 8\}$. We estimate the extrapolation scores for all $x \in \widehat{\mathcal{D}}_{\text{in}} \cup \widehat{\mathcal{D}}_{\text{out}}$ by

$$\widehat{S}(x) := \frac{\widehat{B}_n^{\text{up}}(x) - \widehat{B}_n^{\text{lo}}(x)}{\widehat{\sigma}_{\text{CV}}},$$

where $\widehat{\sigma}_{\text{CV}}$ is the square root of the cross-validation generalization error computed for the regression method used to estimate the extrapolation bounds. As a benchmark, we additionally compute a minimal Euclidean distance defined for all $x \in \widehat{\mathcal{D}}_{\text{in}} \cup \widehat{\mathcal{D}}_{\text{out}}$ by

$$\widehat{E}(x) := \min_{i \in \{1,\dots,n\}} \|x - X_i\|_2.$$

To compare how well $\widehat{S}$ and $\widehat{E}$ capture extrapolation, we compute for all thresholds $\lambda \in [0, \infty)$ (i) the fraction of observations in $\widehat{\mathcal{D}}_{\text{in}} \cup \widehat{\mathcal{D}}_{\text{out}}$ which have an extrapolation score below

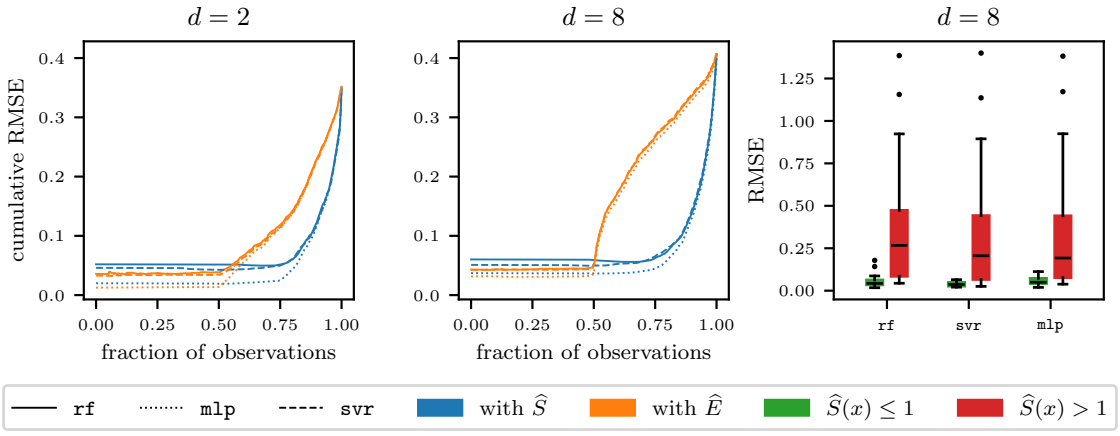


Figure 5: (left and middle) Comparison of extrapolation score $\widehat{S}$ and a Euclidean based benchmark score $\widehat{E}$. While $\widehat{E}$ is only able to separate 50% of the samples (which corresponds to the points in $\mathcal{D}_{\text{in}}$) before the cumulative RMSE increases, the extrapolation score $\widehat{S}$ is able to separate approximately 75% (hence also points in $\mathcal{D}_{\text{out}}$) before the cumulative RMSE increases. (right) Comparison of the RMSE for points with $\widehat{S} \leq 1$ versus points with $\widehat{S} > 1$ (the value 1 corresponds to points where the extrapolation error is equal to the residual noise level).

λ and (ii) the cumulative RMSE of the predictions $\widehat{f}_{\text{xtra}}$ defined in (20) at all points with a score below λ, i.e.,

$$\sqrt{|\{x \in \widehat{\mathcal{D}}_{\text{in}} \cup \widehat{\mathcal{D}}_{\text{out}} \mid \widehat{\text{score}}(x) \leq \lambda\}|^{-1} \sum_{x \in \widehat{\mathcal{D}}_{\text{in}} \cup \widehat{\mathcal{D}}_{\text{out}}} \left(\widehat{f}_{\text{xtra}}(x) - f(x)\right)^2 \mathbb{1}(\widehat{\text{score}}(x) \leq \lambda)},$$

where $\widehat{\text{score}} = \widehat{S}$ or $\widehat{\text{score}} = \widehat{E}$. We use $\widehat{f}_{\text{xtra}}$, here as the plain regression estimates may behave arbitrary outside of $\mathcal{D}_{\text{out}}$. The results are shown in Figure 5 (left and middle). For all regression methods the cumulative RMSE increases sharply after 0.5 when sorted according to $\widehat{E}$. This makes sense as $\widehat{E}$ only separates $\mathcal{D}_{\text{in}}$ from $\mathcal{D}_{\text{out}}$ but does not take into account whether the function might also be accurate on $\mathcal{D}_{\text{out}}$. In contrast, the extrapolation score $\widehat{S}$ also separates points on $\mathcal{D}_{\text{out}}$ for which the predictions are expected to be good. We further separate points with $\widehat{S} \leq 1$ for which the extrapolation error is of smaller order than residual noise level and points with $\widehat{S} > 1$ for which extrapolation error is of larger order. The aggregated RMSEs for these splits are shown in Figure 5 (right). As expected the RMSE is small (and on the order of the residual noise level) when $\widehat{S} \leq 1$ and becomes large if $\widehat{S} > 1$.

5.2 Extrapolation-aware prediction intervals on real data

In this section we illustrate how taking into account extrapolation explicitly can improve uncertainty quantification and help detect when nonparametric prediction procedures are

extrapolating. We consider two datasets: (i) The `biomass` dataset due to Hiernaux et al. (2023), where the task is predicting the foilage dry mass of a tree from its crown area and (ii) the well-known `abalone` dataset from the UCI ML repository (Nash et al., 1995), where the task is to predict the age of abalone shells from several phenotype measurements (sex, length, diameter, height, whole weight, shucked weight, viscera weight and shell weight).

Throughout this section we fix $\alpha = 0.2$. We consider four different standard nonparametric methods to construct predictions intervals: (i) Quantile regression forests (Meinshausen, 2006), denoted by `qrf`, (ii) quantile neural networks (Taylor, 2000), denoted by `qnn`, (iii) conformalized quantile regression forests, denoted by `cpqrf`, and (iv) conformalized quantile neural networks, denoted by `cpqnn`. The two conformalized methods are based on Romano et al. (2019) and calibrate the prediction intervals from the corresponding quantile regression to have a finite sample exact unconditional coverage guarantee similar to conventional conformal prediction (Balasubramanian et al., 2014). We then compare each of these methods with its extrapolation-aware counterpart, denoted by `xtra-qrf`, `xtra-qnn`, `xtra-cpqrf` and `xtra-cpqnn` respectively, which is constructed by applying `Xtrapolation` to the conditional quantiles. More specifically, for all $\star \in \{\texttt{qrf}, \texttt{qnn}, \texttt{cpqrf}, \texttt{cpqnn}\}$ we compare

$$\widehat{C}^{\text{pred}}_{\star}(x) \coloneqq \left[\widehat{\mathcal{T}}^{\alpha/2}_{\star}(x), \widehat{\mathcal{T}}^{1-\alpha/2}_{\star}(x)\right] \quad \text{with} \quad \widehat{C}^{\text{pred}}_{\texttt{xtra-}\star}(x) \coloneqq \left[\widehat{B}^{\text{lo}}_{\widehat{\mathcal{T}}^{\alpha/2}_{\star}}(x), \widehat{B}^{\text{up}}_{\widehat{\mathcal{T}}^{1-\alpha/2}_{\star}}(x)\right],$$

where $\widehat{\mathcal{T}}^{\alpha/2}_{\star}$ and $\widehat{\mathcal{T}}^{1-\alpha/2}_{\star}$ are estimates based on quantile regression forest, $\widehat{B}^{\text{lo}}_{\widehat{\mathcal{T}}^{\alpha/2}_{\star}}(x)$ is the estimate of the lower extrapolation bound for $\widehat{\mathcal{T}}^{\alpha/2}_{\star}$ and $\widehat{B}^{\text{up}}_{\widehat{\mathcal{T}}^{1-\alpha/2}_{\star}}(x)$ is the estimate of the upper extrapolation bound for $\widehat{\mathcal{T}}^{1-\alpha/2}_{\star}$. As the data contains point masses (the age variable in the `abalone` dataset is discrete), we use averaged randomized prediction intervals to calibrate the coverage to the precise level α (see Supplementary Material C.4 for details).

We generate two types of train and test splits for both datasets: (i) *Random splits* that randomly split the data into 8 approximately equally sized sets and (ii) *extrapolation splits* that split the data into 8 approximately equally sized sets according to a predictor variable (crown area for `biomass` and length for `abalone`). The resulting coverage on each split is given in Figure 6 (top for `biomass` bottom for `abalone`). While the standard prediction interval estimates (blue markers) have good coverage for random splits, they under cover for some of the extrapolation splits. The reason is that they are not intended to work outside of the support and will behave differently depending on the underlying regression procedure (e.g., tree-based models extrapolate constant and neural networks linearly). For `qrf` applied to `biomass` this can be seen in Figure 7, where we plot the estimated quantiles for both `qrf` and `xtra-qrf` (similar plots for the other methods are provided in Supplementary Material B). The extrapolation-aware prediction intervals (green markers) perform similar to standard prediction intervals on random splits but avoid under coverage on the extrapolation splits. The slightly conservative behavior of the extrapolation-aware prediction intervals is expected since the extrapolation bounds account for the uncertainty due to the extrapolation. The fact that the coverage is preserved provides empirical evidence that the proposed extrapolation assumption is indeed satisfied on both datasets.

We now show how the difference between lower and upper extrapolation bounds can be used as an extrapolation score. We compute, in an 8-fold cross-validation style (using the

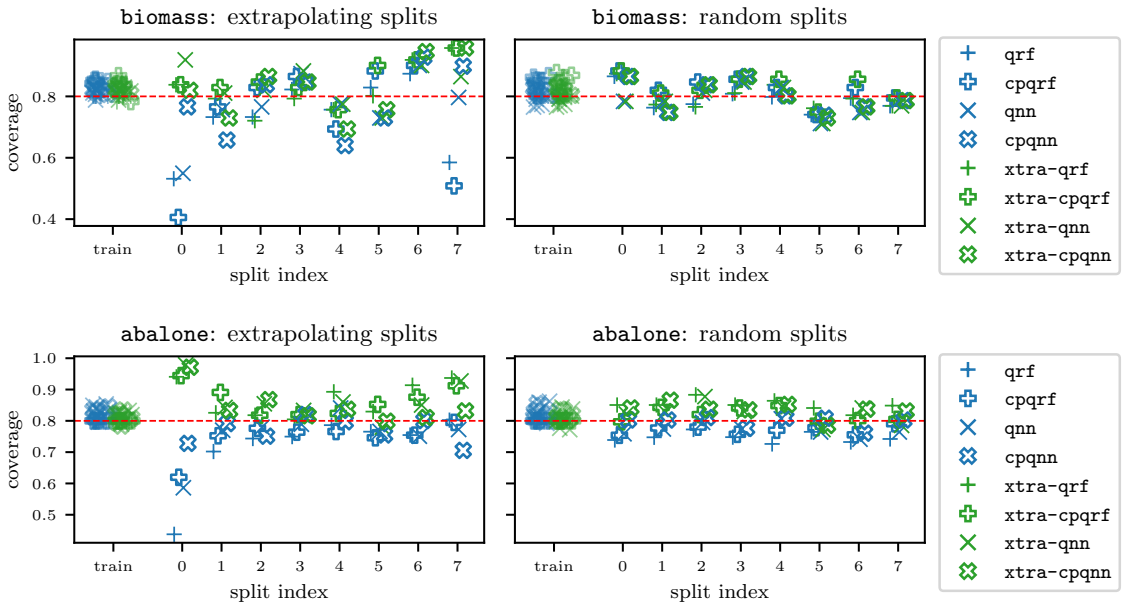


Figure 6: Coverage of prediction intervals on `biomass` and `abalone` datasets. Extrapolating splits are train test splits that leave out certain ranges of tree crown areas for `biomass` and length for `abalone`, while random splits are randomly drawn splits. While the standard prediction intervals (blue markers) under cover for some extrapolating splits, the extrapolation-aware counterparts (blue markers) guard against such under coverage.

extrapolation splits), for all $\star \in \{\texttt{qrf}, \texttt{qnn}, \texttt{cpqrf}, \texttt{cpqnn}\}$ and for each sample point X_i the extrapolation scores

$$\left(\widehat{B}^{\mathrm{up}}_{\widehat{\mathcal{T}}_\star^{\alpha/2}}(X_i) - \widehat{B}^{\mathrm{lo}}_{\widehat{\mathcal{T}}_\star^{\alpha/2}}(X_i)\right) + \left(\widehat{B}^{\mathrm{up}}_{\widehat{\mathcal{T}}_\star^{1-\alpha/2}}(X_i) - \widehat{B}^{\mathrm{lo}}_{\widehat{\mathcal{T}}_\star^{1-\alpha/2}}(X_i)\right), \tag{22}$$

where the estimates in this expression are computed on all splits not containing X_i. We then sort the samples according to this score from small to large and estimate the coverage using a rolling window (size 100 for `biomass` and size 400 for `abalone`). The result is shown in Figure 8. For both datasets and all methods the coverage remains similar for small extrapolation scores but starts diverging for larger ones. Importantly, while the standard prediction intervals start to under cover, their extrapolation-aware counterparts only become conservative. Figure 6 (extrapolating split 7) and Figure 8 (for high extrapolation scores) further show that the extrapolation behavior on `abalone` is substantially different between the `qrf` and `qnn` based intervals (likely due to the difference in the underlying function classes). In contrast, the extrapolation-aware counterparts appear to extrapolate similarly regardless of the underlying method.

Overall, we find that none of the standard methods or methodologies for constructing prediction intervals (e.g., quantile regression or conformalized versions) are reliable for ex-

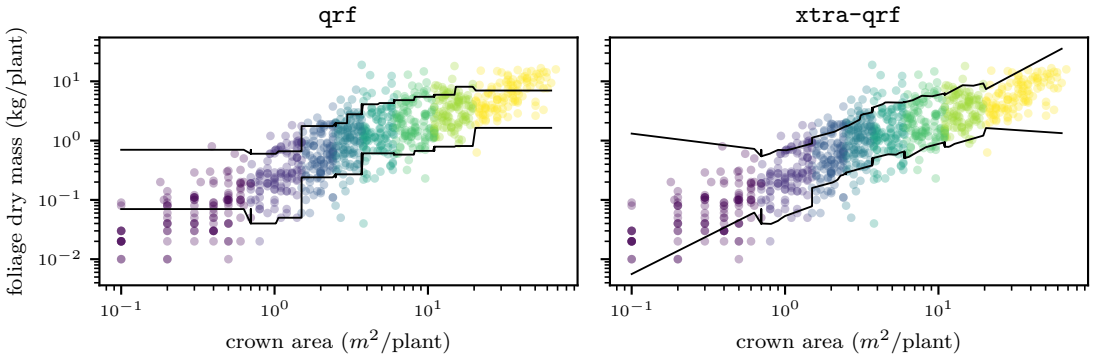


Figure 7: Estimated prediction intervals for `qrf` (left) and `xtra-qrf` (right) for the extrapolation split. Colors correspond to different extrapolation splits. While `qrf` always extrapolates constantly, the extrapolation-aware versions adapt to the changes observed on the training data.

trapolation and may fail severely when extrapolating. Our extrapolation-aware framework, in contrast, works well but is conservative. Any mathematical guarantee for extrapolation must rely on assumptions that cannot be verified directly. However, the fact that conditional coverage is preserved in these experiments provides empirical support for the q-th derivative extrapolating assumption on real-world data. Furthermore, applying `Xtrapolation` does not significantly affect conclusions when interpolating, which suggests that adding it does not harm inference while protecting against erroneous conclusions.

6. Discussion

We defined extrapolation as the process of performing inference on a conditional function outside of the support of the conditioning variable X. This type of extrapolation is not well-defined in a conventional nonparametric sense as the data generating distribution P_0 does not specify the conditional distribution outside of the support of X. We therefore assumed the existence of a Markov kernel that fully specifies the conditional – also outside of the support of X – but which might not be identified by P_0 alone. Then, by assuming that the conditional function behaves at most as extreme on the entire domain as it does on the support of X, we were able to construct extrapolation bounds on the conditional function that are identified by P_0. We proposed to perform inference on these extrapolation bounds instead of on the conditional function directly. We emphasize that an extrapolation assumption is needed: ours, assuming at most as extreme derivatives as in the observed domain, seems natural and we gave some additional interpretation after its Definition 5.

A key feature of our framework is that performing inference on the extrapolation bounds instead of the conditional function (at least on a population level) only affects the nonparametric analysis if extrapolation occurs, since the lower and upper bound are both equal to the conditional function inside support of X. This ensures that the analysis is extrapolation-aware, which guards against potential errors and can provide additional insights when one

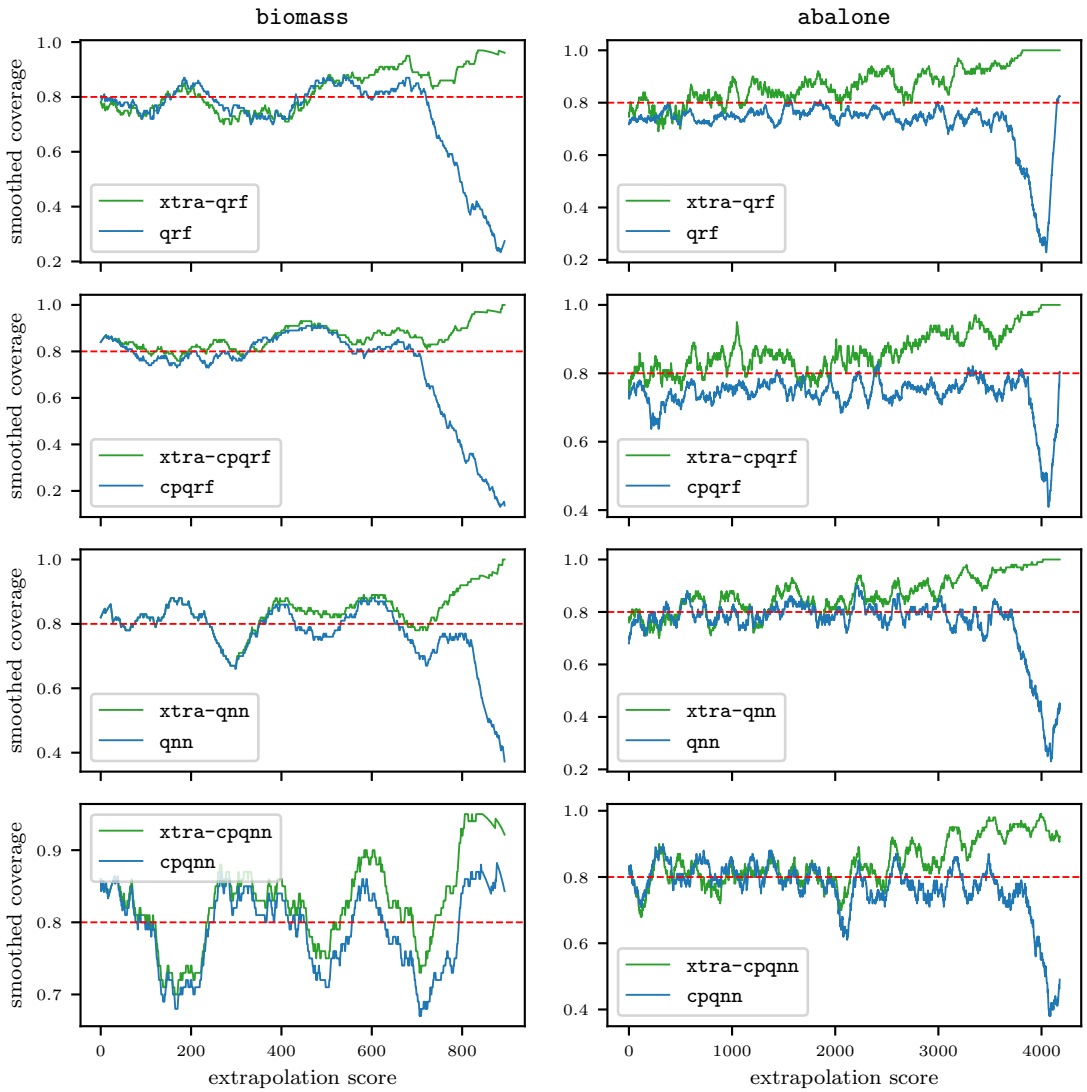


Figure 8: Coverage computed based on a rolling window across observations sorted by the extrapolation score in (22). For observations with a small extrapolation score the coverage of standard prediction intervals is close to their extrapolation-aware counterparts. However, for large extrapolation scores, the coverages diverge. The extrapolation-aware prediction intervals are expected to remain valid or become conservative (i.e., above 0.8) when the extrapolation score becomes large, while we have no guarantees on the extrapolation behavior of standard prediction intervals.

is indeed extrapolating. Here, we considered three specific applications, out-of-support prediction, uncertainty quantification and quantifying extrapolation, but there are likely many other inference tasks that could benefit from an extrapolation-aware analysis. Even

though the extrapolation bounds are identified, they may be difficult to estimate in practice. We propose a method that is able to estimate the bounds in an estimator-agnostic way, which performs well in empirical experiments. Importantly, this estimation procedure only takes the estimated conditional function at the sample points as input (i.e., $(X_1, \widehat{\Phi}(X_1)), \ldots, (X_n, \widehat{\Phi}(X_n))$), allowing practitioners to use their preferred nonparametric estimates $\widehat{\Phi}$ to estimate the extrapolation bounds.

We hope the proposed framework inspires new developments related to extrapolation, which – given its importance – has received too little attention in the wider statistics and machine learning community so far. An important direction of future work is to consider other types of extrapolation assumptions. For example, one could consider shape constraints on the conditional function, which can likely be incorporated similarly to the derivative assumptions considered here. Additionally, it would be interesting to see the proposed framework applied across several applied domains in the style of the real data analysis reported in Figure 8. This would provide valuable insights into how realistic the extrapolation assumptions are and whether there are variations that are more amenable in certain applications.

Acknowledgements

We thank Anton Rask Lundborg for helpful discussions as well as Christian Igel and Pierre Hiernaux for providing the biomass data. NP was supported by a research grant (0069071) from Novo Nordisk Fonden. The research was partially conducted during NP's research stay at the Institute for Mathematical Research at ETH Zürich (FIM). PB has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 786461).

Supplementary Material

- Section A: Additional algorithms
- Section B: Additional results from numerical experiments
- Section C: Details on numerical experiments
- Section D: Proofs
- Section E: Auxiliary results

Appendix A. Additional algorithms

Algorithm 3: `ParameterTuning`

Input : Data $(X_1, \widehat{\Phi}_n(X_1)), \dots, (X_n, \widehat{\Phi}_n(X_n))$, parameter lists $(\lambda_1, \dots, \lambda_L)$ and $(\Gamma_1, \dots, \Gamma_K)$

Tuning : Tolerance `tol`, number of folds K

Output: Optimal parameters λ^* and Γ^*

Split indices $\{1, \dots, n\}$ into M disjoint sets $I_1, \dots, I_M$ of (roughly) equal size
for $m \in \{1, \dots, M\}$ **do**
  for $k \in \{1, \dots, K\}$ **do**
    $\widehat{\mu} \leftarrow$ `rfpoly-`v on $(X_i, \widehat{\Phi}_i(X_i))_{i \in I_{-m}}$ with parameters Γ_k
    Extract weight matrix $W(k) = (w_i(X_j))_{i,j}$ from $\widehat{\mu}$
    for $\ell \in \{1, \dots, L\}$ **do**
      $\hat{f}_m \leftarrow$ local polynomial fit with (18) and penalty $\lambda = \lambda_\ell$ using $(X_i, \widehat{\Phi}_i(X_i))_{i \in I_{-m}}$ and $W(k)$
      $\hat{Y}_i^{k,\ell} \leftarrow \hat{f}_m(X_i)$ for all $i \in I_m$
    end
  end
end
$\overline{E}_{k,\ell} \leftarrow \frac{1}{n} \sum_{i=1}^n \mathrm{Loss}(\hat{Y}_i^{k,\ell}, \widehat{\Phi}_n(X_i))$ for all $(k, \ell) \in \{1, \dots, K\} \times \{1, \dots, L\}$
$(\bar{k}, \bar{\ell}) \leftarrow \arg\min\{\overline{E}_{k,\ell} \mid (k, \ell) \in \{1, \dots, K\} \times \{1, \dots, L\}\}$
for $(k, \ell) \in \{1, \dots, K\} \times \{1, \dots, L\}$ **do**
  $S_{k,\ell} \leftarrow \left(\frac{1}{n} \sum_{i=1}^n (E_{\bar{k},\bar{\ell},i} - E_{k,\ell,i})^2\right)^{\frac{1}{2}} / \sqrt{n}$
end
$k^* \leftarrow \min\{k \in \{1, \dots, K\} \mid \exists \ell \in \{1, \dots, L\} : \overline{E}_{k,\ell} \le \overline{E}_{\bar{k},\bar{\ell}} + \mathtt{tol} \cdot S_{k,\ell}\}$
$\ell^* \leftarrow \min\{\ell \in \{1, \dots, L\} \mid \overline{E}_{k^*,\ell} \le \overline{E}_{\bar{k},\bar{\ell}} + \mathtt{tol} \cdot S_{k^*,\ell}\}$
$\Gamma^* \leftarrow \Gamma_{k^*}$ and $\lambda^* \leftarrow \lambda_{\ell^*}$

Algorithm 4: `Xtrapolation` (one-dimensional version)

Input : Estimates $\widehat{\Phi}_n(X_1), \dots, \widehat{\Phi}_n(X_n)$, data $X_1, \dots, X_n$, target points $\bar{x}_1, \dots, \bar{x}_m$, order k

Tuning : Penalty λ, `rf` parameters Γ

Output: Extrapolation bound estimates $\widehat{B}^{\text{lo}}(\bar{x}_1), \dots, \widehat{B}^{\text{lo}}(\bar{x}_m)$, $\widehat{B}^{\text{up}}(\bar{x}_1), \dots, \widehat{B}^{\text{up}}(\bar{x}_m)$

$\mathcal{D} \leftarrow (X_1, \widehat{\Phi}_n(X_1)), \dots, (X_1, \widehat{\Phi}_n(X_n))$
for $k \in \{1, \dots, q\}$ **do**
  $\widehat{\partial^k \Phi}_n(X_1), \dots, \widehat{\partial^k \Phi}_n(X_n) \leftarrow \texttt{RFLocPol}(\mathcal{D}, k = k, v = 1, \lambda = \lambda, \Gamma = \Gamma)$
end
for $\ell \in \{1, \dots, m\}$ **do**
  for $i \in \{1, \dots, n\}$ **do**
    $S \leftarrow \left\{ \widehat{\partial^q \Phi}_n(X_1) \frac{(\bar{x}_\ell - X_i)^q}{q!}, \dots, \widehat{\partial^q \Phi}_n(X_n) \frac{(\bar{x}_\ell - X_i)^q}{q!} \right\}$
    $B_i^{\text{lo}} \leftarrow \sum_{k=1}^{q-1} \widehat{\partial^k \Phi}_n(X_i) \frac{(\bar{x}_\ell - X_i)^k}{k!} + \min(S)$
    $B_i^{\text{up}} \leftarrow \sum_{k=1}^{q-1} \widehat{\partial^k \Phi}_n(X_i) \frac{(\bar{x}_\ell - X_i)^k}{k!} + \max(S)$
  end
  $\widehat{B}^{\text{lo}}(\bar{x}_\ell) \leftarrow \max_i B_i^{\text{lo}}$ and $\widehat{B}^{\text{up}}(\bar{x}_\ell) \leftarrow \min_i B_i^{\text{up}}$
  if $\widehat{B}^{\text{lo}}(\bar{x}_\ell) > \widehat{B}^{\text{up}}(\bar{x}_\ell)$ **then**
    $\widehat{B}^{\text{lo}}(\bar{x}_\ell) \leftarrow \big(\widehat{B}^{\text{lo}}(\bar{x}_\ell) + \widehat{B}^{\text{up}}(\bar{x}_\ell)\big)/2$ and $\widehat{B}^{\text{up}}(\bar{x}_\ell) \leftarrow \big(\widehat{B}^{\text{lo}}(\bar{x}_\ell) + \widehat{B}^{\text{up}}(\bar{x}_\ell)\big)/2$
  end
end

Appendix B. Additional results from numerical experiments

B.1 Coverage of extrapolation-aware bootstrap confidence intervals

To illustrate the coverage guarantee in Proposition 8, we perform a coverage experiment using Example 2. We consider two fixed distributions over (X, Y). A linear model given by

$$Y = 0.5X + 0.35 + \epsilon \quad \text{and} \quad X \perp\!\!\!\perp \epsilon,$$

and a nonlinear model given by

$$Y = \frac{1}{1 + e^{-3.25X}} + \epsilon \quad \text{and} \quad X \perp\!\!\!\perp \epsilon,$$

where in both cases $X \sim \text{Unif}(-1, 1)$ and $\epsilon \sim \mathcal{N}(0, 0.1^2)$. The parameters are chosen to make the two distributions similar. We then draw a dataset $(X_1, Y_1), \ldots, (X_n, Y_n)$ with $n = 250$ for each distribution. For each setting, we construct nonparametric bootstrap confidence intervals using a percentile bootstrap as follows: (1) draw a bootstrap sample of size n with replacement, (2) fit a random forest and apply `Xtrapolation` with parameter tuning to construct extrapolation bounds, and (3) repeat steps (1)–(2) $B = 50$ times and select the 5% quantile of the lower and 95% quantile of the upper extrapolation bounds as bootstrap bounds. These bounds are shown in Figure 2. To empirically verify the coverage guarantee in Proposition 8, we repeat the same experiment 200 times and compute the coverage at 100 evenly spaced points in $[-3, 3]$. The results are shown in Figure 9.

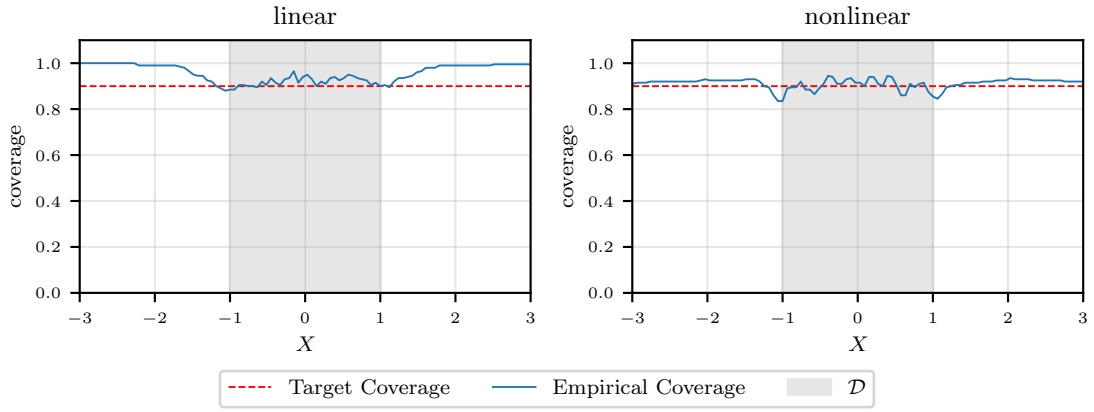


Figure 9: Pointwise coverage of extrapolation-aware bootstrap confidence intervals for the linear example (left) and nonlinear example (right). In the linear example, the true conditional mean lies in the center of the extrapolation bounds (see Figure 2 left), leading to overcoverage. In the nonlinear example, the conditional mean lies on the boundary of the extrapolation bounds (see Figure 2 right), resulting in less overcoverage.

B.2 Extrapolation plots for `biomass`

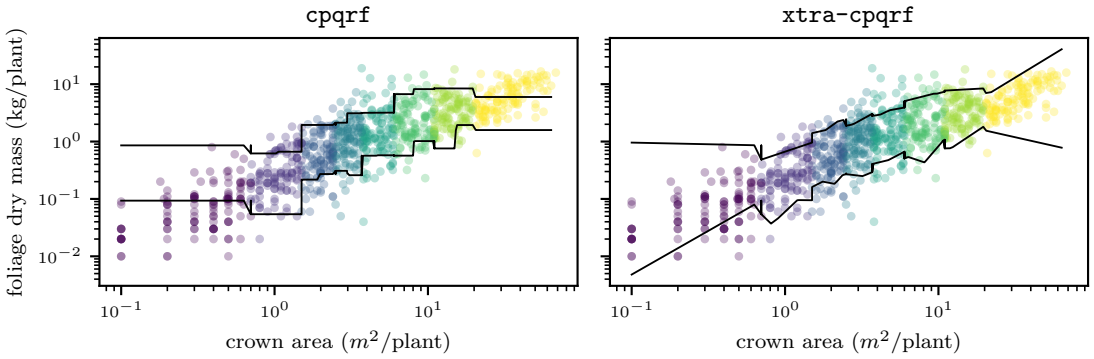


Figure 10: Estimated prediction intervals for `cpqrf` (left) and `xtra-cpqrf` (right) for the extrapolation split. Colors correspond to different extrapolation splits.

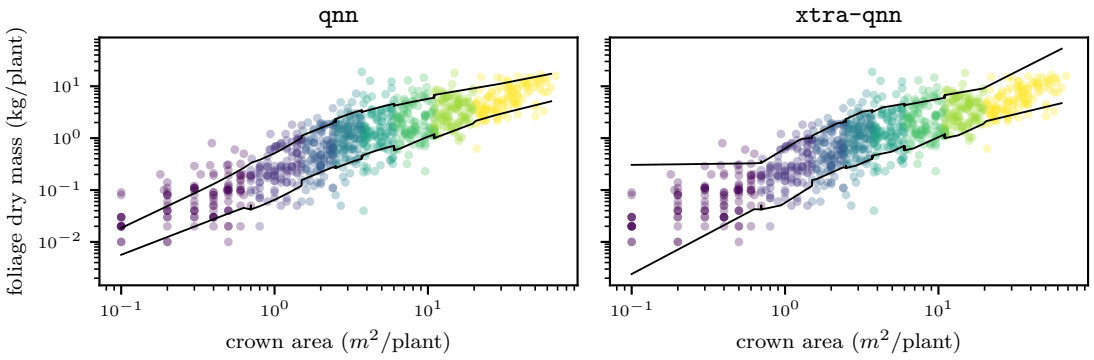


Figure 11: Estimated prediction intervals for `qnn` (left) and `xtra-qnn` (right) for the extrapolation split. Colors correspond to different extrapolation splits.

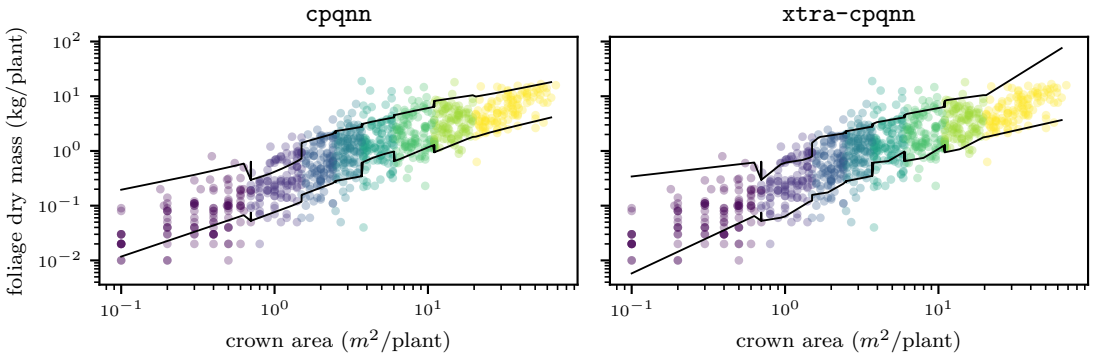


Figure 12: Estimated prediction intervals for `cpqnn` (left) and `xtra-cpqnn` (right) for the extrapolation split. Colors correspond to different extrapolation splits.

Appendix C. Details on numerical experiments

C.1 Sampling of slopes in simulation

Let $j \in \{1, \ldots, 4\}$ be the index of the interval I_j that was left out from C_1. The slopes s_1, s_2, s_3, s_4 are then sampled as follows. First, for all $\ell \in \{1, \ldots, 4\} \setminus \{j\}$ independently of each other sample $s_\ell \sim \mathrm{Unif}([-10, 10])$. Then, randomly draw $j^* \sim \mathrm{Unif}(\{1, \ldots, 4\} \setminus \{j\})$ and set $s_j := s_{j^*}$. This ensures that the slope corresponding to the left out region of the first coordinate I_j are all observed in $\mathcal{D}_{\mathrm{in}}$, which further guarantees that $f \triangleleft^1_{\mathcal{D}_{\mathrm{in}}} f$ (almost every).

C.2 Hyperparameter selection for regression procedures

Firstly, for the simulation experiments in Section 5.1, we combine each regression procedure with a variable screening step and tune the hyperparameters with cross-validation, to

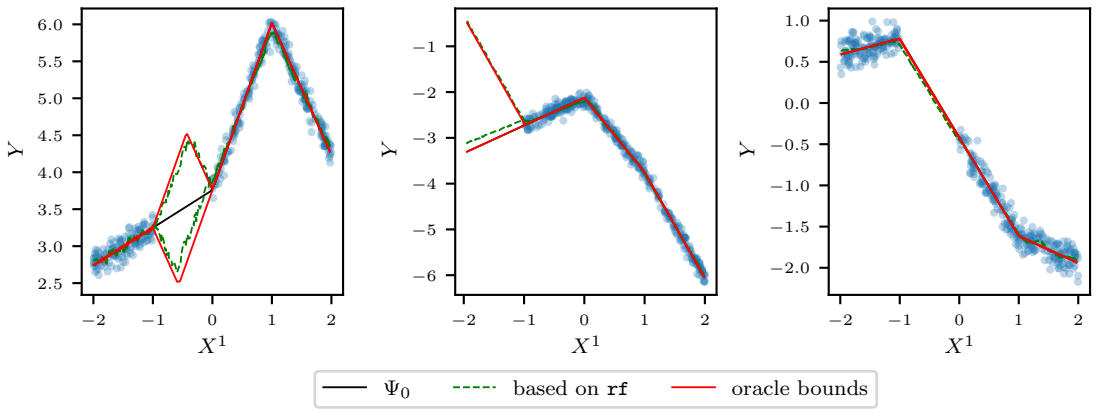


Figure 13: Three simulations generated according to the model introduced in Section 5.1. Since Ψ_0 is first derivative extrapolating, it is identified also on $\mathcal{D}_{\text{out}}$ in the example on the right.

ensure that the regressions procedures perform well across a large range of settings. More specifically, for the variable selection we fit a random forest and only keep features with Gini importance that is larger than the mean of the Gini importance across all coordinates. The variable screening captures the sparsity in the simulation setup and improves the predictive performance if X is multi-dimensional. This variable screening step is always performed on the same training data on which the subsequent regression procedure is fitted. For each regressions procedure, we then construct grids of potential hyperparameters and select the optimal one based on a 5-fold cross-validation using the mean squared error as a score:

- `rf`: We used 500 trees and choose an optimal `min_sample_leafs` from $\{128, 64, 32, 16, 8, 4, 2, 1\}$.

- `svr`: We used a radial basis function kernel and choose an optimal bandwidth `gamma` in $\{\frac{10^{-3}}{d}, \frac{10^{-2}}{d}, \frac{10^{-1}}{d}, \frac{1}{d}, \frac{10^{1}}{d}, \frac{10^{2}}{d}, \frac{10^{3}}{d}\}$ and an optimal penalty `C` in $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$. Furthermore, we scaled the data before applying the support vector regression.

- `mlp`: We fit a neural network with 'relu' activations using the 'adam' solver and with early stopping. We chose the optimal architecture `hidden_layer_sizes` in $\{(100,), (20, 20, 20,)\}$ and the optimal penalty `alpha` in $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$.

Secondly, in the real data applications in Section 5.2, we tuned the hyperparameters of the regression models using a 5-fold cross-validation with the average pinball loss for the two target quantiles as a score. We used the following hyperparameter grids for the different methods:

- `qrf` and `cpqrf`: We used the grid $\{1, 5, 10, 20, 40, 80, 160\}$ for `min_samples_leaf`.

- `qnn` and `cpqnn`: We used the grids $\{10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}\}$ and $\{64, 128, 256\}$ for `weight_decay` (specifying the penalty) and `hidden_size` (specifying the number of hidden variables), respectively.

In all cases we selected the most regularized parameter that was at most one standard deviation worse in score than the best model (similar in spirit to Algorithm 3 with `tol` $= 1$). We used the quantile neural network and the conformalized regression procedures included in the code of Romano et al. (2019). The remaining parameters were left at their default values. In all cases we included a scaling step that scaled the training data to have mean zero and variance one.

C.3 Hyperparameter selection for `Xtrapolation`

When applying extrapolation (for $q = 1$) as described in Algorithm 2, we need to select the penalty parameter λ and the `rf` parameters Γ. To select the optimal parameter we used the parameter tuning described in Algorithm 3 with `tol` $= 1$ and $K = 5$ and parameter grids that depend on the experiment:

- Simulation experiment (Section 5.1): We selected λ among $\{10, 1, 0.1, 0.01, 0.001, 0\}$ and Γ among default trees but with `impurity_tol` among $\{100, 10, 1, 0.1, 0.01\}$.
- Section 5.2 `biomass` data: We set $\lambda = 0$ and selected Γ among default trees but with `min_samples_leaf` among $\{40, 30, 20, 10\}$.
- Section 5.2 `abalone` data: We selected λ among $\{0.1, 0.01, 0.001\}$ and Γ among default trees but with `min_samples_leaf` among $\{10, 5, 1\}$.

To speed up the computation, we further only went over a subselection of all possible anchor points (see Section 4.3.1). More specifically, for the experiments in Section 5.1 we used the $n/2$ closest points in Euclidean distance. For the `biomass` application we used all possible anchor points and for the `abalone` application we used the 100 closest (but non-zero weighted) points based on a random forest closeness measure (required here as the covariate 'sex' is binary and hence not included in the extrapolation).

C.4 Randomized prediction intervals

Since in both the `biomass` and `abalone` data there are samples with the exact same Y values (relatively common in `abalone` since years are counts), we randomized the prediction intervals in the experiments. This ensures that its always possible to reach the exact coverage on average. To formalize randomized prediction intervals, we introduce a random variable $U \sim \mathrm{Ber}(p)$. Then for any interval $C = [C_{\mathrm{lo}}, C_{\mathrm{up}}]$ we can define a corresponding randomized interval C^{rand} by

$$C^{\mathrm{rand}}(U) \coloneqq \begin{cases} [C_{\mathrm{lo}}, C_{\mathrm{up}}] & \text{if } U = 1 \\ (C_{\mathrm{lo}}, C_{\mathrm{up}}) & \text{if } U = 0. \end{cases}$$

For a given prediction interval $\widehat{C}$ which does not have the correct level due to atoms on the boundaries of the interval, we can then use the corresponding randomized version $\widehat{C}^{\mathrm{rand}}(U)$,

where we choose the probability p to calibrate the prediction interval on the training data to have the correct coverage. For all results, we then report the expected coverage of such randomized prediction intervals.

Appendix D. Proofs

D.1 Proof of Theorem 3

Proof We prove the two parts separately.

Part (i): First, fix arbitrary $x \in \mathcal{X}$ and let $g \in C^q(\mathcal{X})$ satisfy for all $x \in \mathcal{D}$ that $g(x) = f(x)$ and for all $v \in \mathcal{B}$ that

$$\inf_{x \in \mathcal{X}} D_v^q g(x) \geq \inf_{x \in \mathcal{D}} D_v^q f(x) \quad \text{and} \quad \sup_{x \in \mathcal{X}} D_v^q g(x) \leq \sup_{x \in \mathcal{D}} D_v^q f(x). \tag{23}$$

Then, using (1), it holds for all $x_0 \in \mathcal{D}$ that there exists $\xi_{x_0} \in \mathcal{X}$ such that

$$\begin{aligned} g(x) &= \sum_{\ell=0}^{q-1} D^\ell_{\overline{v}(x_0,x)} g(x_0) \frac{\|x - x_0\|_2^\ell}{\ell!} + D^q_{\overline{v}(x_0,x)} g(\xi_{x_0}) \frac{\|x - x_0\|_2^q}{q!} \\ &= \sum_{\ell=0}^{q-1} D^\ell_{\overline{v}(x_0,x)} f(x_0) \frac{\|x - x_0\|_2^\ell}{\ell!} + D^q_{\overline{v}(x_0,x)} g(\xi_{x_0}) \frac{\|x - x_0\|_2^q}{q!}, \end{aligned} \tag{24}$$

where we used that $g = f$ on $\mathcal{D}$. Moreover, using (23), we get for all $x_0 \in \mathcal{D}$ that

$$D^q_{\overline{v}(x_0,x)} g(\xi_{x_0}) \leq \sup_{z \in \mathcal{X}} D^q_{\overline{v}(x_0,x)} g(z) \leq \sup_{z \in \mathcal{D}} D^q_{\overline{v}(x_0,x)} f(z). \tag{25}$$

Hence, together with (24) we have for all $x_0 \in \mathcal{D}$ that

$$g(x) \leq \sum_{\ell=0}^{q-1} D^\ell_{\overline{v}(x_0,x)} f(x_0) \frac{\|x - x_0\|_2^\ell}{\ell!} + \sup_{z \in \mathcal{D}} D^q_{\overline{v}(x_0,x)} f(z) \frac{\|x - x_0\|_2^q}{q!}. \tag{26}$$

Finally, taking the infimum over $x_0 \in \mathcal{D}$ on both sides results in

$$g(x) \leq B^{\text{up}}_{q,f,\mathcal{D}}(x).$$

The same argument also applies to the lower extrapolation bound, which completes the proof of part (i).

Part (ii): We only prove the result for the upper bound, the same arguments apply to the lower bound. Since we assume that $\mathcal{X}$ is compact it holds that $\mathcal{D}$ is also compact. We can therefore define the function $\xi : \mathcal{X} \to \mathcal{D}$ such that for all $x \in \mathcal{X}$ it holds that $\xi(x) \in \mathcal{D}$ satisfies

$$B^{\text{up}}_{q,f,\mathcal{D}}(x) = \sum_{\ell=0}^{q-1} D^\ell_{\overline{v}(\xi(x),x)} f(\xi(x)) \frac{\|x - \xi(x)\|_2^\ell}{\ell!} + \sup_{z \in \mathcal{D}} D^q_{\overline{v}(\xi(x),x)} f(z) \frac{\|x - \xi(x)\|_2^q}{q!}.$$

Moreover, define function $\nu : \mathcal{X} \to \mathcal{D}$ such that for all $x \in \mathcal{X}$ it holds that $\nu(x) \in \mathcal{D}$ satisfies

$$\sup_{z \in \mathcal{D}} D^q_{\overline{v}(\xi(x),x)} f(z) = D^q_{\overline{v}(\xi(x),x)} f(\nu(x)).$$

In the following, we construct a q-times differentiable sequence $(g_n^{\mathrm{up}})_{n\in\mathbb{N}}$ which approximates the extrapolation bound $B^{\mathrm{up}}_{q,f,\mathcal{D}}$, is q-times continuously differentiable and satisfies $g_n^{\mathrm{up}} \lhd^q_{\mathcal{D}} f$. Constructing such a sequence directly is difficult due to the two optimizations ξ and ν. We therefore construct the approximation in two steps.

First approximation step: Fix arbitrary $\varepsilon > 0$. For all $z \in \mathcal{X}$, using multi-index notation, define the functions $B_z : \mathbb{R}^d \to \mathbb{R}$ for all $x \in \mathbb{R}^d$ by

$$B_z(x) \coloneqq \sum_{|\boldsymbol{\alpha}|<q} \partial^{\boldsymbol{\alpha}} f(\xi(z)) \cdot \frac{(x-\xi(z))^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!} + \sum_{|\boldsymbol{\alpha}|=q} \partial^{\boldsymbol{\alpha}} f(\nu(z)) \cdot \frac{(x-\xi(z))^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!}. \tag{27}$$

Since for fixed $z \in \mathcal{X}$ the points $\xi(z)$ and $\nu(z)$ are fixed it is easier to analyze B_z, which is multivariate polynomial of degree at most q. Moreover, for all $v \in \mathcal{B}$, $\ell \in \{0,\dots,d\}$ and $x \in \mathcal{X}$ the directional derivative satisfies $D_v^\ell f(x) = \sum_{|\boldsymbol{\alpha}|=\ell} \frac{\ell!}{\boldsymbol{\alpha}!} \partial^{\boldsymbol{\alpha}} f(x) v^{\boldsymbol{\alpha}}$. This implies for all $z \in \mathcal{B}$ that

$$B_z(z) = B^{\mathrm{up}}_{q,f,\mathcal{D}}(z). \tag{28}$$

More properties of B_z are listed in Lemma 14.

Consider now the collection of open sets $\{\mathcal{U}_\varepsilon(x) \mid x \in \mathcal{X}\}$, where $\mathcal{U}_\varepsilon(x) \coloneqq \{y \in \mathbb{R}^d \mid \|x-y\|_2 < \varepsilon\}$. Since, $\mathcal{X}$ is compact, there exists a finite set of points $x_1^\varepsilon, \dots, x_{M_\varepsilon}^\varepsilon$ such that $\{\mathcal{U}_\varepsilon(x_\ell^\varepsilon) \mid \ell \in \{1,\dots,M_\varepsilon\}\}$ covers $\mathcal{X}$. We can therefore define the function $g_\varepsilon : \mathbb{R}^d \to \mathbb{R}$ for all $x \in \mathbb{R}^d$ by

$$g_\varepsilon(x) \coloneqq \frac{1}{w_\varepsilon(x)} \sum_{\ell=1}^{M_\varepsilon} B_{x_\ell^\varepsilon}(x) \mathbb{1}_{\mathcal{U}_\varepsilon(x_\ell^\varepsilon)}(x),$$

where $w_\varepsilon(x) \coloneqq |\{\ell \in \{1,\dots,M_\varepsilon\} \mid x \in \mathcal{U}_\varepsilon(x_\ell^\varepsilon)\}|$. We now prove that there exists $K_1 > 0$ such that

$$\sup_{x\in\mathcal{X}} \left| B^{\mathrm{up}}_{q,f,\mathcal{D}}(x) - g_\varepsilon(x) \right| \le K_1 \varepsilon. \tag{29}$$

To see this, fix $z \in \mathcal{X}$, then using the triangle inequality and (28) it holds that

$$\begin{aligned}
\sup_{x\in\mathcal{U}_\varepsilon(z)} \left| B^{\mathrm{up}}_{q,f,\mathcal{D}}(x) - B_z(x) \right| &\le \sup_{x\in\mathcal{U}_\varepsilon(z)} \left| B^{\mathrm{up}}_{q,f,\mathcal{D}}(x) - B^{\mathrm{up}}_{q,f,\mathcal{D}}(z) \right| + \sup_{x\in\mathcal{U}_\varepsilon(z)} \left| B_z(z) - B_z(x) \right| \\
&\le (c_1 + c_2)\varepsilon,
\end{aligned} \tag{30}$$

where $c_1 > 0$ and $c_2 > 0$ exists by Lemma 13 and Lemma 14 (ii). Using (30) we further get

$$\begin{aligned}
\sup_{x\in\mathcal{X}} \left| B^{\mathrm{up}}_{q,f,\mathcal{D}}(x) - g_\varepsilon(x) \right| &= \sup_{x\in\mathcal{X}} \left| B^{\mathrm{up}}_{q,f,\mathcal{D}}(x) - \frac{1}{w_\varepsilon(x)} \sum_{\ell=1}^{M_\varepsilon} B_{x_\ell^\varepsilon}(x) \mathbb{1}_{\mathcal{U}_\varepsilon(x_\ell^\varepsilon)}(x) \right| \\
&\le \sup_{x\in\mathcal{X}} \left(\frac{1}{w_\varepsilon(x)} \sum_{\ell=1}^{M_\varepsilon} \left| B^{\mathrm{up}}_{q,f,\mathcal{D}}(x) - B_{x_\ell^\varepsilon}(x) \right| \mathbb{1}_{\mathcal{U}_\varepsilon(x_\ell^\varepsilon)}(x) \right) \\
&\le \sup_{x\in\mathcal{X}} \left(\frac{1}{w_\varepsilon(x)} \sum_{\ell=1}^{M_\varepsilon} \mathbb{1}_{\mathcal{U}_\varepsilon(x_\ell^\varepsilon)}(x) \right) (c_1 + c_2)\varepsilon \\
&= \underbrace{(c_1 + c_2)}_{=:K_1} \varepsilon,
\end{aligned}$$

which proves (29). Furthermore, notice that the set $\mathcal{W} \coloneqq \mathbb{R}^d \setminus (\partial\mathcal{U}_\varepsilon(x_1^\varepsilon) \cup \cdots \cup \partial\mathcal{U}_\varepsilon(x_{M_\varepsilon}^\varepsilon))$ is open and hence for all $z \in \mathcal{W}$ there exists $\kappa > 0$ such that $\mathcal{U}_\kappa(z) \subseteq \mathcal{W}$. Moreover, for all $x \in \mathcal{U}_\kappa(z)$ it holds that

$$g_\varepsilon(x) = \frac{1}{w_\varepsilon(x)} \sum_{\ell=1}^{M_\varepsilon} B_{x_\ell^\varepsilon}(x) \mathbb{1}_{\mathcal{U}_\varepsilon(x_\ell^\varepsilon)}(x) = \frac{1}{w_\varepsilon(z)} \sum_{\ell=1}^{M_\varepsilon} B_{x_\ell^\varepsilon}(x) \mathbb{1}_{\mathcal{U}_\varepsilon(x_\ell^\varepsilon)}(z). \tag{31}$$

Since $B_{x_\ell^\varepsilon}$ is q-times continuously differentiable and since $\mathbb{R}^d \setminus \mathcal{W}$ is a Lebesgue null-set this implies that g_ε is Lebesgue almost everywhere q-times continuously differentiable. In particular, using (31) and Lemma 14 (i), it holds for every multi-index $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $|\boldsymbol{\alpha}| = q$ and every $x \in \mathcal{W}$ that

$$\partial^{\boldsymbol{\alpha}} g_\varepsilon(x) = \frac{1}{w_\varepsilon(x)} \sum_{\ell=1}^{M_\varepsilon} \partial^{\boldsymbol{\alpha}} f(\nu(x_\ell^\varepsilon)) \mathbb{1}_{\mathcal{U}_\varepsilon(x_\ell^\varepsilon)}(x). \tag{32}$$

Second approximation step: Again let $\varepsilon > 0$ be arbitrary. Denote by $\eta_\varepsilon : \mathbb{R}^d \to \mathbb{R}$ the standard mollifier defined for all $x \in \mathbb{R}^d$ by

$$\eta_\varepsilon(x) \coloneqq \begin{cases} C \exp\left(1/\left(\|x\|_2^2 - \varepsilon^2\right)\right) & \text{if } x \in \mathcal{U}_\varepsilon(0) \\ 0 & \text{otherwise,} \end{cases}$$

where $C > 0$ is chosen such that $\int_{\mathbb{R}^d} \eta_\epsilon(x) \mathrm{d}x = 1$. For any integrable function $h : \mathbb{R}^d \to \mathbb{R}$, we define the mollified function $h * \eta_\varepsilon$ for all $x \in \mathbb{R}^d$ by the convolution

$$(h * \eta_\varepsilon)(x) \coloneqq \int_{\mathcal{U}_\varepsilon(0)} h(x - y) \eta_\varepsilon(y) \mathrm{d}y.$$

We now define the approximation function $\bar{g}_\varepsilon : \mathbb{R}^d \to \mathbb{R}$ for all $x \in \mathbb{R}^d$ by

$$\bar{g}_\varepsilon(x) \coloneqq (g_\varepsilon * \eta_\varepsilon)(x).$$

Using that the convolution with a mollifier is smooth (Hörmander, 2015, Theorem 1.3.1), we get that $\bar{g}_\varepsilon$ is q-times continuously differentiable. Furthermore, using Jensen's inequality and the triangle inequality, we can bound the approximation error of the mollification as

follows,

$$
\begin{aligned}
&\sup_{x\in\mathcal{X}} |\bar{g}_\varepsilon(x) - g_\varepsilon(x)| \\
&\quad = \sup_{x\in\mathcal{X}} |(g_\varepsilon * \eta_\varepsilon)(x) - g_\varepsilon(x)| \\
&\quad \leq \sup_{x\in\mathcal{X}} \int_{\mathcal{U}_\varepsilon(0)} |g_\varepsilon(x-y) - g_\varepsilon(x)|\, \eta_\varepsilon(y)\mathrm{d}y \\
&\quad \leq \sup_{\substack{x,y\in\mathcal{X}:\\ \|x-y\|_2\leq\varepsilon}} |g_\varepsilon(y) - g_\varepsilon(x)| \\
&\quad = \sup_{\substack{x,y\in\mathcal{X}:\\ \|x-y\|_2\leq\varepsilon}} \left| g_\varepsilon(y) - B^{\mathrm{up}}_{q,f,\mathcal{D}}(y) - g_\varepsilon(x) + B^{\mathrm{up}}_{q,f,\mathcal{D}}(x) + B^{\mathrm{up}}_{q,f,\mathcal{D}}(y) - B^{\mathrm{up}}_{q,f,\mathcal{D}}(x) \right| \\
&\quad \leq 2 \sup_{x\in\mathcal{X}} \left| g_\varepsilon(x) - B^{\mathrm{up}}_{q,f,\mathcal{D}}(x) \right| + \sup_{\substack{x,y\in\mathcal{X}:\\ \|x-y\|_2\leq\varepsilon}} \left| B^{\mathrm{up}}_{q,f,\mathcal{D}}(x) - B^{\mathrm{up}}_{q,f,\mathcal{D}}(y) \right| \\
&\quad \leq \underbrace{(2K_1 + c_1)}_{=:K_2}\, \varepsilon, \qquad (33)
\end{aligned}
$$

where we used (29) and Lemma 13 in the last step.

Let $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $|\boldsymbol{\alpha}| = q$, then using that all partial derivatives of $g_\varepsilon * \eta_\varepsilon$ can be bounded on $\overline{\mathcal{U}_\varepsilon(x)}$, we can pull derivatives under the integral (e.g., Folland, 1999, Theorem 2.27). Together with (32), we hence get

$$
\begin{aligned}
\partial^{\boldsymbol{\alpha}} \bar{g}_\varepsilon(x) &= \partial^{\boldsymbol{\alpha}} (g_\varepsilon * \eta_\varepsilon)(x) \\
&= \partial^{\boldsymbol{\alpha}} \int_{\mathcal{U}_\varepsilon(0)} g_\varepsilon(x-y)\eta_\varepsilon(y)\mathrm{d}y \\
&= \int_{\mathcal{U}_\varepsilon(0)} (\partial^{\boldsymbol{\alpha}} g_\varepsilon)(x-y)\eta_\varepsilon(y)\mathrm{d}y \\
&= \int_{\mathcal{U}_\varepsilon(0)} \frac{1}{w_\varepsilon(x-y)} \sum_{\ell=1}^{M_\varepsilon} \partial^{\boldsymbol{\alpha}} f(\nu(x_\ell^\varepsilon)) \mathbb{1}_{\mathcal{U}_\varepsilon(x_\ell^\varepsilon)}(x-y)\eta_\varepsilon(y)\mathrm{d}y \\
&= \sum_{\ell=1}^{M_\varepsilon} \partial^{\boldsymbol{\alpha}} f(\nu(x_\ell^\varepsilon)) \underbrace{\int_{\mathcal{U}_\varepsilon(0)} \frac{1}{w_\varepsilon(x-y)} \mathbb{1}_{\mathcal{U}_\varepsilon(x_\ell^\varepsilon)}(x-y)\eta_\varepsilon(y)\mathrm{d}y}_{=:\bar{w}_\ell(x)}.
\end{aligned}
$$

Hence expressing the directional derivative in terms of partial derivatives we get for all $v \in \mathcal{B}$ and all $x \in \mathcal{X}$ that

$$\begin{aligned} D_v^q \bar{g}_\varepsilon(x) &= \sum_{|\boldsymbol{\alpha}|=q} \frac{q!}{\boldsymbol{\alpha}!} \partial^{\boldsymbol{\alpha}} \bar{g}_\varepsilon(x) v^{\boldsymbol{\alpha}} \\ &= \sum_{\ell=1}^{M_\varepsilon} \sum_{|\boldsymbol{\alpha}|=q} \frac{q!}{\boldsymbol{\alpha}!} \partial^{\boldsymbol{\alpha}} f(\nu(x_\ell^\varepsilon)) v^{\boldsymbol{\alpha}} \bar{w}_\ell(x) \\ &= \sum_{\ell=1}^{M_\varepsilon} D_v^q f(\nu(x_\ell^\varepsilon)) \bar{w}_\ell(x) \end{aligned}$$

This further implies for all $v \in \mathcal{B}$ and all $x \in \mathcal{X}$ that

$$\inf_{z \in \mathcal{D}} D_v^q f(z) = \inf_{z \in \mathcal{D}} D_v^q f(z) \sum_{\ell=1}^{M_\varepsilon} \bar{w}_\ell(x) \leq D_v^q \bar{g}_\varepsilon(x) \leq \sup_{z \in \mathcal{D}} D_v^q f(z) \sum_{\ell=1}^{M_\varepsilon} \bar{w}_\ell(x) = \sup_{z \in \mathcal{D}} D_v^q f(z).$$

Hence, we have shown that

$$\bar{g}_\varepsilon \lhd_{\mathcal{D}}^q f. \tag{34}$$

Conclude by constructing g_n^{up}: Define for all $n \in \mathbb{N}$ the functions

$$g_n^{\text{up}} \coloneqq \bar{g}_{\varepsilon_n} \quad \text{with} \quad \varepsilon_n \coloneqq \frac{1}{n(K_1 + K_2)}.$$

Then, with (29) and (33) it holds that

$$\begin{aligned} \sup_{x \in \mathcal{X}} |g_n^{\text{up}}(x) - B_{q,f,\mathcal{D}}^{\text{up}}(x)| &= \sup_{x \in \mathcal{X}} |\bar{g}_{\varepsilon_n}(x) - B_{q,f,\mathcal{D}}^{\text{up}}(x)| \\ &\leq \sup_{x \in \mathcal{X}} |\bar{g}_{\varepsilon_n}(x) - g_{\varepsilon_n}(x)| + \sup_{x \in \mathcal{X}} |g_{\varepsilon_n}(x) - B_{q,f,\mathcal{D}}^{\text{up}}(x)| \\ &\leq \varepsilon_n K_2 + \varepsilon_n K_1 = \frac{1}{n}. \end{aligned} \tag{35}$$

By (34) it further holds for all $n \in \mathbb{N}$ that $g_n^{\text{up}} \lhd_{\mathcal{D}}^q f$, which completes the proof of part (ii) and hence also of Theorem 3. ■

D.2 Proof of Proposition 7

Proof Fix $x \in \mathcal{X}$ throughout the proof. For all $Q \in \mathcal{Q}_0$, denote by $\Psi_Q : \mathcal{X} \to \mathbb{R}$ the conditional expectation function corresponding to the Markov kernel Q. Then for all $Q \in \mathcal{Q}_0$ there exists a mean-zero random variable U_x such that $Y_x = \Psi_Q(x) + U_x$.

We now first construct an upper bound on the right-hand side of (10). To this end, fix $Q \in \mathcal{Q}_0$, then it holds that $\Psi_Q(x) \in \left[B_{\Psi_0,\mathcal{D}_{\text{in}}}^{\text{lo}}(x),\, B_{\Psi_0,\mathcal{D}_{\text{in}}}^{\text{up}}(x)\right]$. Hence, we get that

$$\left| \Psi_Q(x) - \frac{1}{2}\left(B_{\Psi_0,\mathcal{D}_{\text{in}}}^{\text{lo}}(x) + B_{\Psi_0,\mathcal{D}_{\text{in}}}^{\text{up}}(x) \right) \right| \leq \frac{1}{2}\left(B_{\Psi_0,\mathcal{D}_{\text{in}}}^{\text{up}}(x) - B_{\Psi_0,\mathcal{D}_{\text{in}}}^{\text{lo}}(x) \right).$$

Using this bound and $\mathbb{E}_Q[U_x] = 0$, we further get

$$\begin{aligned}\mathbb{E}_Q[(Y_x - f^*(x))^2] &= (\Psi_Q(x) - f^*(x))^2 + \mathbb{E}_Q[U_x^2] \\ &= \left(\Psi_Q(x) - \frac{1}{2}\left(B^{\mathrm{lo}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x) + B^{\mathrm{up}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(X)\right)\right)^2 + \mathbb{E}_Q[U_x^2] \\ &\leq \frac{1}{4}\left(B^{\mathrm{up}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x) - B^{\mathrm{lo}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x)\right)^2 + \mathbb{E}_Q[U_x^2].\end{aligned}$$

Since $Q \in \mathcal{Q}_0$ was arbitrary this implies

$$\sup_{Q\in\mathcal{Q}_0} \mathbb{E}_Q[(Y_x - f^*(x))^2] \leq \frac{1}{4}\left(B^{\mathrm{up}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x) - B^{\mathrm{lo}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x)\right)^2 + \sup_{Q\in\mathcal{Q}_0} \mathbb{E}_Q[U_x^2]. \tag{36}$$

Next, we derive a lower bound. To this end, fix an arbitrary $f \in C^0(\mathcal{X})$ and $Q \in \mathcal{Q}_0$. The triangle inequality implies

$$\left|B^{\mathrm{up}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x) - f(x)\right| + \left|B^{\mathrm{lo}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x) - f(x)\right| \geq \left|(B^{\mathrm{up}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x) - B^{\mathrm{lo}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x)\right|. \tag{37}$$

By Theorem 3 it holds that $B^{\mathrm{lo}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}, B^{\mathrm{up}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}} \in C^0(\mathcal{X})$ and hence there exist $Q_{\mathrm{lo}}, Q_{\mathrm{up}} \in \mathcal{Q}_0$ such that $\Psi_{Q_{\mathrm{lo}}} \equiv B^{\mathrm{lo}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}$ and $\Psi_{Q_{\mathrm{up}}} \equiv B^{\mathrm{up}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}$. This, implies that

$$\begin{aligned}\sup_{Q\in\mathcal{Q}_0} (\Psi_Q(x) - f(x)) &\geq \max\left((B^{\mathrm{up}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x) - f(x))^2, (B^{\mathrm{lo}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x) - f(x))^2\right) \\ &\geq \frac{1}{2}\left((B^{\mathrm{up}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x) - f(x))^2 + (B^{\mathrm{lo}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x) - f(x))^2\right) \\ &\geq \frac{1}{4}\left(|B^{\mathrm{up}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x) - f(x)| + |B^{\mathrm{lo}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x) - f(x)|\right)^2 \\ &\geq \frac{1}{4}\left(B^{\mathrm{up}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x) - B^{\mathrm{lo}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x)\right)^2,\end{aligned} \tag{38}$$

where for the third inequality we used $(a+b)^2 \leq 2a^2 + 2b^2$ and (37) for the last inequality. Therefore together with (38) it holds that

$$\sup_{Q\in\mathcal{Q}_0} \mathbb{E}[(Y_x - f(x))^2] \geq \frac{1}{4}\left(B^{\mathrm{up}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x) - B^{\mathrm{lo}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x)\right)^2 + \sup_{Q\in\mathcal{Q}_0} \mathbb{E}_Q[U_x]. \tag{39}$$

Combining (36) and (39) completes the proof of Proposition 7. ■

D.3 Proof of Proposition 8

Proof Since Ψ_0 is assumed to be q-th derivative extrapolating it holds by Theorem 3 that $\Psi_0 \in [B^{\mathrm{lo}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x), B^{\mathrm{up}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x)]$. We can use this together with standard probability

inequalities to get

$$\begin{aligned}
&\mathbb{P}(\Psi_0(x) \in \widehat{\mathrm{C}}^{\mathrm{conf}}_{n;\alpha}(x)) \\
&\quad = \mathbb{P}\left(\Psi_0(x) \in \left[\widehat{G}^{\mathrm{lo}}_n(\tfrac{\alpha}{2}, x),\, \widehat{G}^{\mathrm{up}}_n(1-\tfrac{\alpha}{2}, x)\right]\right) \\
&\quad = \mathbb{P}\left(\Psi_0(x) \geq \widehat{G}^{\mathrm{lo}}_n(\tfrac{\alpha}{2}, x),\, \Psi_0(x) \leq \widehat{G}^{\mathrm{up}}_n(1-\tfrac{\alpha}{2}, x)\right) \\
&\quad \geq \mathbb{P}\left(B^{\mathrm{lo}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x) > \widehat{G}^{\mathrm{lo}}_n(\tfrac{\alpha}{2}, x),\, B^{\mathrm{up}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x) \leq \widehat{G}^{\mathrm{up}}_n(1-\tfrac{\alpha}{2}, x)\right) \\
&\quad \geq 1 - \mathbb{P}\left(B^{\mathrm{lo}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x) \leq \widehat{G}^{\mathrm{lo}}_n(\tfrac{\alpha}{2}, x)\right) - \mathbb{P}\left(B^{\mathrm{up}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x) > \widehat{G}^{\mathrm{up}}_n(1-\tfrac{\alpha}{2}, x)\right) \\
&\quad = -\mathbb{P}\left(B^{\mathrm{lo}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x) \leq \widehat{G}^{\mathrm{lo}}_n(\tfrac{\alpha}{2}, x)\right) + \mathbb{P}\left(B^{\mathrm{up}}_{\Psi_0,\mathcal{D}_{\mathrm{in}}}(x) \leq \widehat{G}^{\mathrm{up}}_n(1-\tfrac{\alpha}{2}, x)\right).
\end{aligned}$$

Next, taking the $\liminf$ as n goes to infinity on both sides together with the assumptions on $\widehat{G}^{\star}_n$, we get that

$$\liminf_{n\to\infty} \mathbb{P}(\Psi_0(x) \in \widehat{\mathrm{C}}^{\mathrm{conf}}_{n;\alpha}(x)) \geq 1 - \alpha.$$

This completes the proof of Proposition 8. ■

D.4 Proof of Proposition 9

Proof By Theorem 3, since $\mathcal{T}_0^{\alpha/2}$ and $\mathcal{T}_0^{1-\alpha/2}$ are both q-th derivative extrapolating, it holds that

$$\left[\mathcal{T}_0^{\alpha/2}(x),\, \mathcal{T}_0^{1-\alpha/2}(x)\right] \subseteq \mathrm{C}^{\mathrm{pred}}_{\alpha}(x).$$

Using the definition of the conditional quantile function, we therefore get that

$$\mathbb{P}_{Q_0(x,\cdot)}(Y_x \in \mathrm{C}^{\mathrm{pred}}_{\alpha}(x)) \geq \mathbb{P}_{Q_0(x,\cdot)}(Y_x \in [\mathcal{T}_0^{\alpha/2}(x),\, \mathcal{T}_0^{1-\alpha/2}(x)]) = 1 - \alpha.$$

This completes the proof of Proposition 9. ■

D.5 Proof of Theorem 11

Proof We first show that since $\mathcal{X}$ is compact it holds that

$$\Lambda_n \coloneqq \sup_{z\in\mathcal{D}_{\mathrm{in}}} \min_{k\in\{1,\ldots,n\}} \|X_k - z\|_2 \xrightarrow{P_0} 0 \quad \text{as} \quad n \to \infty. \tag{40}$$

To see this, first observe that $\mathcal{D}_{\mathrm{in}} = \mathrm{supp}(X) \coloneqq \{x \in \mathcal{X} \mid P_0^X(\{z \in \mathcal{X} \mid \|x - z\|_2 < r\}) > 0 \text{ for all } r > 0\}$ is closed and hence compact (since $\mathcal{X}$ is compact). Next, fix $\epsilon > 0$, then by compactness, finitely many balls with radius $\epsilon/2$ and centers $c_1 \ldots, c_M \in \mathcal{D}_{\mathrm{in}}$ cover $\mathcal{D}_{\mathrm{in}}$. Define for all $j \in \{1, \ldots, M\}$ the probability $p_j \coloneqq P_0^X(\{x \in \mathcal{X} \mid \|x - c_j\|_2 < \epsilon/2\})$, then by the definition of the support, $p_j > 0$. Let A_n denote the event where each of these balls contains at least one sample from $X_1, \ldots, X_n$. Then, on A_n it holds for all $z \in \mathcal{D}_{\mathrm{in}}$ that there exists $j \in \{1, \ldots, M\}$ such that $\|z - c_j\| < \epsilon/2$ and hence by the triangle inequality it

holds that $\|X_k - z\|_2 \leq \|X_k - c_j\|_2 + \|z - c_j\|_2 < \epsilon$. This implies that $\mathbb{P}(A_n) \leq \mathbb{P}(\Lambda_n \leq \epsilon)$. Therefore,

$$\mathbb{P}(\Lambda_n > \epsilon) \leq \mathbb{P}(A_n^c) = \sum_{j=1}^{M} (1 - p_j)^n \leq M \left(1 - \min_{j \in \{1,\dots,M\}} p_j\right)^n,$$

which implies (40).

Now for the main proof, recall that the directional derivative can be expressed in multi-index notation as $D_v^\ell f(z) = \sum_{\boldsymbol{\alpha}=\ell} \partial^{\boldsymbol{\alpha}} f(z) v^{\boldsymbol{\alpha}} \frac{\ell!}{\boldsymbol{\alpha}!}$. Furthermore, by compactness of $\mathcal{X}$ it holds that $K \coloneqq \sup_{x,y \in \mathcal{X}} \|x - y\|_2 < \infty$. Throughout, the proof we fix $x \in \mathcal{X}$. We begin by defining for all $w, z \in \mathcal{X}$ the functions

$$F_1(w) \coloneqq \sum_{\ell=0}^{q-1} \sum_{|\boldsymbol{\alpha}|=\ell} \partial^{\boldsymbol{\alpha}} \Phi_0(w) \frac{(x-w)^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!} \quad \text{and} \quad F_2(w,z) \coloneqq \sum_{|\boldsymbol{\alpha}|=q} \partial^{\boldsymbol{\alpha}} \Phi_0(z) \frac{(x-w)^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!}.$$

Then, it holds, using the multi-index expression of the directional derivative, that

$$B^{\mathrm{up}}_{\Phi_0, \mathcal{D}_{\mathrm{in}}}(x) = \inf_{w \in \mathcal{D}_{\mathrm{in}}} \left(F_1(w) + \sup_{z \in \mathcal{D}_{\mathrm{in}}} F_2(w,z) \right).$$

Similarly define for all $w, z \in \mathcal{X}$ the functions

$$\widehat{F}_1(w) \coloneqq \sum_{\ell=0}^{q-1} \sum_{|\boldsymbol{\alpha}|=\ell} \partial^{\boldsymbol{\alpha}} \widehat{\Phi}_n(w) \frac{(x-w)^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!} \quad \text{and} \quad \widehat{F}_2(w,z) \coloneqq \sum_{|\boldsymbol{\alpha}|=q} \partial^{\boldsymbol{\alpha}} \widehat{\Phi}_n(z) \frac{(x-w)^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!}.$$

Then, it holds that

$$\widehat{B}^{\mathrm{up}}_n(x) = \min_{i \in \{1,\dots,n\}} \left(\widehat{F}_1(X_i) + \max_{k \in \{1,\dots,n\}} \widehat{F}_2(X_i, X_k) \right).$$

Lastly, define for all $w \in \mathcal{D}_{\mathrm{in}}$ the functions

$$G(w) \coloneqq F_1(w) + \sup_{z \in \mathcal{D}_{\mathrm{in}}} F_2(w,z) \quad \text{and} \quad \widehat{G}(w) \coloneqq \widehat{F}_1(w) + \max_k \widehat{F}_2(w, X_k).$$

For the remaining proof we make use of the following properties:

(i) F_1 is Lipschitz with constant $L_1 > 0$.

(ii) There exists a constant $L_2 > 0$ such that for all $z \in \mathcal{X}$ the function $w \mapsto F_2(w,z)$ is Lipschitz with constant L_2.

(iii) There exists a constant $L_3 > 0$ such that for all $w \in \mathcal{X}$ the function $z \mapsto F_2(w,z)$ is Lipschitz with constant L_3.

(iv) For all functions $f : \mathcal{D} \to \mathbb{R}$ and $g : \mathcal{D} \to \mathbb{R}$ it holds that

$$|\sup_{z \in \mathcal{D}} f(z) - \sup_{z \in \mathcal{D}} g(z)| \leq \sup_{z \in \mathcal{D}} |f(z) - g(z)| \quad \text{and} \quad |\inf_{z \in \mathcal{D}} f(z) - \inf_{z \in \mathcal{D}} g(z)| \leq \sup_{z \in \mathcal{D}} |f(z) - g(z)|.$$

Property (i) holds since any continuously differentiable function on a compact set is Lipschitz. To prove property (ii) we use that by compactness of $\mathcal{X}$ there exists $\bar{K}^1 \in \mathbb{R}$ such that for all $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $|\boldsymbol{\alpha}| = q$, it holds that $\sup_{z \in \mathcal{X}} |\partial^{\boldsymbol{\alpha}} \Phi_0(z)| < \bar{K}^1$. Moreover, for all $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $|\boldsymbol{\alpha}| = q$, the function $w \mapsto (x - w)^{\boldsymbol{\alpha}}$ is Lipschitz with constant $\bar{L}^1_{\boldsymbol{\alpha}}$ (since it is differentiable and $\mathcal{X}$ is compact). Then, for all $z, w, w' \in \mathcal{X}$ we can apply the triangle inequality and bound each term to get

$$|F_2(w, z) - F_2(w', z)| \leq \sum_{|\boldsymbol{\alpha}|=q} \frac{\bar{K}^1 \bar{L}^1_{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!} \|w - w'\|_2,$$

which proves (ii) since the resulting constant does not depend on z. For (iii), we additionally use that since Φ_0 is $(q+1)$-differentiable it holds for all $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $|\boldsymbol{\alpha}| = q$ that $\partial^{\boldsymbol{\alpha}} \Phi_0$ is Lipschitz with constant $\bar{L}^2_{\boldsymbol{\alpha}}$ (since $\mathcal{X}$ is compact). Then, for $w, z, z' \in \mathcal{X}$ it holds using the triangle inequality and bounding the terms that

$$|F_2(w, z) - F_2(w, z')| \leq \sum_{|\boldsymbol{\alpha}|=q} \frac{K^q \bar{L}^2_{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!} \|z - z'\|_1,$$

which proves (iii) since the resulting constant does not depend on w. Finally, for (iv) observe that for all $w \in \mathcal{D}$ it holds

$$f(w) \leq \sup_{z \in \mathcal{D}} |f(z) - g(z)| + g(w).$$

Now taking either the sup or inf on both sides, rearranging and then swapping the roles of f and g proves the result.

We are now ready to prove the main result in three steps:

- Step 1: Show that

$$\sup_{w \in \mathcal{D}_{\text{in}}} \left| F_1(w) - \widehat{F}_1(w) \right| \xrightarrow{P_0} 0 \quad \text{as } n \to \infty. \tag{41}$$

- Step 2: Show that

$$\sup_{w \in \mathcal{D}_{\text{in}}} \left| \sup_{z \in \mathcal{D}_{\text{in}}} F_2(w, z) - \max_k \widehat{F}_2(w, X_k) \right| \xrightarrow{P_0} 0 \quad \text{as } n \to \infty. \tag{42}$$

- Step 3: Conclude the proof using G and $\widehat{G}$.

Step 1: To see this, we use the triangle inequality and bound the resulting terms as follows

$$\begin{aligned}
&\sup_{w \in \mathcal{D}_{\text{in}}} \left| F_1(w) - \widehat{F}_1(w) \right| \\
&= \sup_{w \in \mathcal{D}_{\text{in}}} \left| \sum_{\ell=0}^{q-1} \sum_{|\boldsymbol{\alpha}|=\ell} \left(\partial^{\boldsymbol{\alpha}} \Phi_0(w) - \partial^{\boldsymbol{\alpha}} \widehat{\Phi}_n(w) \right) \frac{(x - w)^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!} \right| \\
&\leq \sum_{\ell=0}^{q-1} \sum_{|\boldsymbol{\alpha}|=\ell} \sup_{w \in \mathcal{D}_{\text{in}}} \left| \partial^{\boldsymbol{\alpha}} \Phi_0(w) - \partial^{\boldsymbol{\alpha}} \widehat{\Phi}_n(w) \right| \frac{K^q}{\boldsymbol{\alpha}!}.
\end{aligned}$$

Since each of the summands converges in probability to zero by assumption, this proves (41).

Step 2: We begin by fixing $w \in \mathcal{D}_{\text{in}}$. Then it holds that

$$\begin{aligned}
&\left| \sup_{z \in \mathcal{D}_{\text{in}}} F_2(w,z) - \max_k \widehat{F}_2(w, X_k) \right| \\
&\quad \leq \left| \max_k \widehat{F}_2(w, X_k) - \max_k F_2(w, X_k) \right| + \left| \sup_{z \in \mathcal{D}_{\text{in}}} F_2(w,z) - \max_k F_2(w, X_k) \right| \\
&\quad \leq \max_k \left| \widehat{F}_2(w, X_k) - F_2(w, X_k) \right| + \left| \sup_{z \in \mathcal{D}_{\text{in}}} F_2(w,z) - \max_k F_2(w, X_k) \right| \\
&\quad \leq \sup_{z \in \mathcal{D}_{\text{in}}} \left| \widehat{F}_2(w,z) - F_2(w,z) \right| + \left| \sup_{z \in \mathcal{D}_{\text{in}}} F_2(w,z) - \max_k F_2(w, X_k) \right|, \qquad (43)
\end{aligned}$$

where for the second inequality, we used property (iv). Next, we bound each summand separately. First, using the triangle inequality and bounding the terms we get

$$\begin{aligned}
\sup_{z \in \mathcal{D}_{\text{in}}} \left| \widehat{F}_2(w,z) - F_2(w,z) \right| &= \sup_{z \in \mathcal{D}_{\text{in}}} \left| \sum_{|\boldsymbol{\alpha}|=q} \left(\partial^{\boldsymbol{\alpha}} \widehat{\Phi}_n(z) - \partial^{\boldsymbol{\alpha}} \Phi_0(z) \right) \frac{(x-w)^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!} \right| \\
&\leq \sum_{|\boldsymbol{\alpha}|=q} \sup_{z \in \mathcal{D}_{\text{in}}} \left| \partial^{\boldsymbol{\alpha}} \widehat{\Phi}_n(z) - \partial^{\boldsymbol{\alpha}} \Phi_0(z) \right| \frac{K^q}{\boldsymbol{\alpha}!}. \qquad (44)
\end{aligned}$$

Second, fix $w \in \mathcal{D}_{\text{in}}$, then since $\mathcal{X}$ is compact there exists $z_w \in \mathcal{D}_{\text{in}}$ (since $\mathcal{D}_{\text{in}}$ is closed) such that $\sup_{z \in \mathcal{D}_{\text{in}}} F_2(w,z) = F_2(w, z_w)$. Next, recall $\Lambda_n = \sup_{z \in \mathcal{D}_{\text{in}}} \min_k \|X_k - z\|_2$ and use that $X_1, \ldots, X_n \in \mathcal{D}_{\text{in}}$ to get

$$\begin{aligned}
\left| \sup_{z \in \mathcal{D}_{\text{in}}} F_2(w,z) - \max_k F_2(w, X_k) \right| &= F_2(w, z_w) - \max_k F_2(w, X_k) \\
&\leq F_2(w, z_w) - \inf_{\substack{z \in \mathcal{D}_{\text{in}}: \\ \|z - z_w\|_2 \leq \Lambda_n}} F_2(w,z) \\
&\leq F_2(w, z_w) - (F_2(w, z_w) - \Lambda_n L_3) \\
&= \Lambda_n L_3, \qquad (45)
\end{aligned}$$

where fore the last inequality we used that F_2 is Lipschitz with constant L_3 in the second argument by property (iii). Since the bound does not depend on w this implies

$$\sup_{w \in \mathcal{D}_{\text{in}}} \left| \sup_{z \in \mathcal{D}_{\text{in}}} F_2(w,z) - \max_k F_2(w, X_k) \right| \leq \Lambda_n L_3. \qquad (46)$$

Using (44) and (46) to further bound (43) we get

$$\begin{aligned}
&\sup_{w \in \mathcal{D}_{\text{in}}} \left| \sup_{z \in \mathcal{D}_{\text{in}}} F_2(w,z) - \max_k \widehat{F}_2(w, X_k) \right| \\
&\quad \leq \sum_{|\boldsymbol{\alpha}|=q} \sup_{z \in \mathcal{D}_{\text{in}}} \left| \partial^{\boldsymbol{\alpha}} \widehat{\Phi}_n(z) - \partial^{\boldsymbol{\alpha}} \Phi_0(z) \right| \frac{K^q}{\boldsymbol{\alpha}!} + \Lambda_n L_3.
\end{aligned}$$

Since each summand converges in probability to zero by assumption, this implies (42).
Step 3: Since by the triangle inequality it holds that

$$\sup_{w\in\mathcal{D}_{\text{in}}}\left|G(w)-\widehat{G}(w)\right| \le \sup_{w\in\mathcal{D}_{\text{in}}}\left|F_1(w)-\widehat{F}_1(w)\right| + \sup_{w\in\mathcal{D}_{\text{in}}}\left|\sup_{z\in\mathcal{D}_{\text{in}}} F_2(w,z) - \max_k \widehat{F}_2(w,X_k)\right|,$$

we get by (41) and (42) that

$$\sup_{w\in\mathcal{D}_{\text{in}}}\left|G(w)-\widehat{G}(w)\right| \xrightarrow{P_0} 0 \quad \text{as } n\to\infty. \tag{47}$$

Similar to the argument in (43), we get

$$\begin{aligned}\left|\inf_{w\in\mathcal{D}_{\text{in}}} G(w) - \min_k \widehat{G}(X_k)\right| &\le \left|\min_k \widehat{G}(X_k) - \min_k G(X_k)\right| + \left|\inf_{w\in\mathcal{D}_{\text{in}}} G(w) - \min_k G(X_k)\right| \\ &\le \max_k \left|\widehat{G}(X_k) - G(X_k)\right| + \left|\inf_{w\in\mathcal{D}_{\text{in}}} G(w) - \min_k G(X_k)\right| \\ &\le \sup_{w\in\mathcal{D}_{\text{in}}}\left|\widehat{G}(w) - G(w)\right| + \left|\inf_{w\in\mathcal{D}_{\text{in}}} G(w) - \min_k G(X_k)\right|, \end{aligned} \tag{48}$$

where for the second inequality we used property (iv). Moreover, for all $x, y \in \mathcal{D}_{\text{in}}$ we can use properties (i), (ii) and (iv) to get

$$\begin{aligned}|G(x)-G(y)| &\le |F_1(x)-F_1(y)| + \left|\sup_{z\in\mathcal{D}_{\text{in}}} F_2(x,z) - \sup_{z\in\mathcal{D}_{\text{in}}} F_2(y,z)\right| \\ &\le L_1\|x-y\|_2 + \sup_{z\in\mathcal{D}_{\text{in}}}|F_2(x,z)-F_2(y,z)| \\ &\le L_1\|x-y\|_2 + L_2\|x-y\|_2 .\end{aligned}$$

Hence, letting $w^* \in \mathcal{D}_{\text{in}}$ be such that $\inf_{w\in\mathcal{D}_{\text{in}}} G(w) = G(w^*)$, we get by a similar argument as in (45) that

$$\begin{aligned}\left|\inf_{w\in\mathcal{D}_{\text{in}}} G(w) - \min_k G(X_k)\right| &= \min_k G(X_k) - G(w^*) \\ &\le \sup_{\substack{w\in\mathcal{D}_{\text{in}}: \\ \|w-w^*\|_2\le\Lambda_n}} G(w) - G(w^*) \\ &\le G(w^*) + \Lambda_n(L_1+L_2) - G(w^*) \\ &= \Lambda_n(L_1+L_2).\end{aligned}$$

Using this in (48), leads to

$$\left|\inf_{w\in\mathcal{D}_{\text{in}}} G(w) - \min_k \widehat{G}(X_k)\right| \le \sup_{w\in\mathcal{D}_{\text{in}}}\left|\widehat{G}(w) - G(w)\right| + \Lambda_n(L_1+L_2).$$

Since both summands converge in probability to zero by (47) and (40), this implies

$$\left|B^{\text{up}}_{\Phi_0,\mathcal{D}_{\text{in}}}(x) - \widehat{B}^{\text{up}}_n(x)\right| = \left|\inf_{w\in\mathcal{D}_{\text{in}}} G(w) - \min_k \widehat{G}(X_k)\right| \xrightarrow{P_0} 0 \quad \text{as } n\to\infty,$$

which completes the proof of Theorem 11. ■

Appendix E. Auxiliary results

Lemma 13 *Assume $\mathcal{X}$ is compact, let $f \in C^q(\mathcal{X})$ and $\mathcal{D} \subseteq \mathcal{X}$ non-empty closed. Then, there exists a constant $C > 0$ such that for all $\varepsilon > 0$ it holds that*

$$\sup_{\substack{x,y\in\mathcal{X}:\\ \|x-y\|_2<\varepsilon}} \left|B^{\mathrm{up}}_{q,f,\mathcal{D}}(x) - B^{\mathrm{up}}_{q,f,\mathcal{D}}(y)\right| < C\,\varepsilon.$$

Proof First, introduce for all $x_0 \in \mathcal{D}$ the functions $B^{x_0} : \mathcal{X} \to \mathbb{R}$ defined for all $x \in \mathcal{X}$ by

$$B^{x_0}(x) \coloneqq \sum_{\ell=0}^{q-1} D^{\ell}_{\overline{v}(x_0,x)} f(x_0) \cdot \frac{\|x - x_0\|_2^{\ell}}{\ell!} + \sup_{z\in\mathcal{D}} D^{q}_{\overline{v}(x_0,x)} f(z) \cdot \frac{\|x - x_0\|_2^{q}}{q!}.$$

Fix $\varepsilon > 0$ and $x^*, y^* \in \mathcal{X}$ with $\|x^* - y^*\| < \varepsilon$. Since $\mathcal{D}$ is compact, we can define a function $\xi : \mathcal{X} \to \mathcal{D}$ such that for all $x \in \mathcal{X}$ it holds that $\xi(x) \in \mathcal{D}$ satisfies

$$B^{\mathrm{up}}_{q,f,\mathcal{D}}(x) = B^{\xi(x)}(x).$$

By construction this implies that

$$\left|B^{\mathrm{up}}_{q,f,\mathcal{D}}(x^*) - B^{\mathrm{up}}_{q,f,\mathcal{D}}(y^*)\right| = \left|B^{\xi(x^*)}(x^*) - B^{\xi(y^*)}(y^*)\right|.$$

Furthermore, since the points $\xi(x^*)$ and $\xi(y^*)$ are chosen as minimizers, it holds that

$$\left|B^{\xi(x^*)}(x^*) - B^{\xi(y^*)}(y^*)\right| \leq \begin{cases} \left|B^{\xi(x^*)}(x^*) - B^{\xi(x^*)}(y^*)\right| & \text{if } B^{\xi(x^*)}(x^*) \leq B^{\xi(y^*)}(y^*) \\ \left|B^{\xi(y^*)}(x^*) - B^{\xi(y^*)}(y^*)\right| & \text{otherwise.} \end{cases}$$

Without loss of generality we now assume $B^{\xi(x^*)}(x^*) \leq B^{\xi(y^*)}(y^*)$ and bound the expression in this case. The second case follows analogous with the role of x^* and y^* switched. First, by the triangle inequality we get

$$\begin{aligned}
&\left|B^{\xi(x^*)}(x^*) - B^{\xi(x^*)}(y^*)\right| \\
&\quad= \Bigg| \sum_{\ell=0}^{q-1} \frac{1}{\ell!} \left(D^{\ell}_{\overline{v}(\xi(x^*),x^*)} f(\xi(x^*)) \|x^* - \xi(x^*)\|_2^{\ell} - D^{\ell}_{\overline{v}(\xi(x^*),y^*)} f(\xi(x^*)) \|y^* - \xi(x^*)\|_2^{\ell} \right) \\
&\qquad\quad + \frac{1}{q!} \left(\sup_{z\in\mathcal{D}} D^{q}_{\overline{v}(\xi(x^*),x^*)} f(z) \|x^* - \xi(x^*)\|_2^{q} - \sup_{z\in\mathcal{D}} D^{q}_{\overline{v}(\xi(x^*),y^*)} f(z) \|y^* - \xi(x^*)\|_2^{q} \right) \Bigg| \\
&\quad\leq \sum_{\ell=0}^{q-1} \frac{1}{\ell!} \left| D^{\ell}_{\overline{v}(\xi(x^*),x^*)} f(\xi(x^*)) \|x^* - \xi(x^*)\|_2^{\ell} - D^{\ell}_{\overline{v}(\xi(x^*),y^*)} f(\xi(x^*)) \|y^* - \xi(x^*)\|_2^{\ell} \right| \\
&\qquad\quad + \frac{1}{q!} \left| \sup_{z\in\mathcal{D}} D^{q}_{\overline{v}(\xi(x^*),x^*)} f(z) \|x^* - \xi(x^*)\|_2^{q} - \sup_{z\in\mathcal{D}} D^{q}_{\overline{v}(\xi(x^*),y^*)} f(z) \|y^* - \xi(x^*)\|_2^{q} \right|.
\end{aligned} \tag{49}$$

We now consider the summands individually. Firstly, for all $\ell \in \{1, \dots, q-1\}$, it holds that

$$\begin{aligned}
&\left| D^\ell_{\overline{v}(\xi(x^*),x^*)} f(\xi(x^*)) \|x^* - \xi(x^*)\|_2^\ell - D^\ell_{\overline{v}(\xi(x^*),y^*)} f(\xi(x^*)) \|y^* - \xi(x^*)\|_2^\ell \right| \\
&\quad \leq \left| D^\ell_{\overline{v}(\xi(x^*),x^*)} f(\xi(x^*)) \right| \left| \|x^* - \xi(x^*)\|_2^\ell - \|y^* - \xi(x^*)\|_2^\ell \right| \\
&\qquad + \left| D^\ell_{\overline{v}(\xi(x^*),x^*)} f(\xi(x^*)) - D^\ell_{\overline{v}(\xi(x^*),y^*)} f(\xi(x^*)) \right| \|y^* - \xi(x^*)\|_2^\ell. \qquad (50)
\end{aligned}$$

We separate two cases: (i) $\|x^* - \xi(x^*)\|_2 < 2\varepsilon$ and (ii) $\|x^* - \xi(x^*)\|_2 \geq 2\varepsilon$. For case (i), it holds by the triangle inequality that $\|y^* - \xi(x^*)\|_2 \leq \|y^* - x^*\|_2 + \|x^* - \xi(x^*)\|_2 \leq 4\varepsilon$. Combined with (50) and using that all directional derivatives of f up to order q are bounded by a constant $K > 0$ this implies in case (i) that

$$\left| D^\ell_{\overline{v}(\xi(x^*),x^*)} f(\xi(x^*)) \cdot \|x^* - \xi(x^*)\|_2^\ell - D^\ell_{\overline{v}(\xi(x^*),y^*)} f(\xi(x^*)) \cdot \|y^* - \xi(x^*)\|_2^\ell \right| \leq 14K\varepsilon. \quad (51)$$

For case (ii), consider $\mathcal{X}_\varepsilon \coloneqq \{x \in \mathcal{X} \mid \inf_{x_0 \in \mathcal{D}} \|x - x_0\|_2 \geq \varepsilon\}$ which is compact since $\mathcal{X}$ is compact. Next, observe that for all $x_0 \in \mathcal{D}$ the function $x \mapsto \|x - x_0\|_2^\ell$ is Lipschitz continuous on $\mathcal{X}_\varepsilon$ with a constant $0 < L^\ell_{x_0} < \infty$, which implies that

$$\left| \|x^* - \xi(x^*)\|_2^\ell - \|y^* - \xi(x^*)\|_2^\ell \right| \leq \sup_{x_0 \in \mathcal{D}} L^\ell_{x_0} \|x^* - y^*\|_2. \quad (52)$$

Since $\mathcal{D}$ is compact the supremum is attained and $\bar{L}^\ell \coloneqq \sup_{x_0 \in \mathcal{D}} L^\ell_{x_0} < \infty$. Furthermore, expressing the directional derivative in terms of partial derivatives (using multi-index notation) and letting $M > 0$ be an upper bound on all partial derivatives of f up to order q, it holds that

$$\begin{aligned}
&\left| D^\ell_{\overline{v}(\xi(x^*),x^*)} f(\xi(x^*)) - D^\ell_{\overline{v}(\xi(x^*),y^*)} f(\xi(x^*)) \right| \\
&\quad = \left| \sum_{\boldsymbol{\alpha} \in \mathbb{N}^d : |\boldsymbol{\alpha}| = \ell} \partial^{\boldsymbol{\alpha}} f(\xi(x^*)) \tfrac{\ell!}{\boldsymbol{\alpha}!} \left(\tfrac{x^* - \xi(x^*)}{\|x^* - \xi(x^*)\|_2} \right)^{\boldsymbol{\alpha}} - \sum_{\boldsymbol{\alpha} \in \mathbb{N}^d : |\boldsymbol{\alpha}| = \ell} \partial^{\boldsymbol{\alpha}} f(\xi(x^*)) \tfrac{\ell!}{\boldsymbol{\alpha}!} \left(\tfrac{y^* - \xi(x^*)}{\|y^* - \xi(x^*)\|_2} \right)^{\boldsymbol{\alpha}} \right| \\
&\quad \leq \sum_{\boldsymbol{\alpha} \in \mathbb{N}^d : |\boldsymbol{\alpha}| = \ell} M \tfrac{\ell!}{\boldsymbol{\alpha}!} \left| \left(\frac{x^* - \xi(x^*)}{\|x^* - \xi(x^*)\|_2} \right)^{\boldsymbol{\alpha}} - \left(\frac{y^* - \xi(x^*)}{\|y^* - \xi(x^*)\|_2} \right)^{\boldsymbol{\alpha}} \right|. \qquad (53)
\end{aligned}$$

Next, observe that for all $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $|\boldsymbol{\alpha}| = \ell$ and all $x_0 \in \mathcal{D}$ it holds that the function $x \mapsto \left(\frac{x - x_0}{\|x - x_0\|} \right)^{\boldsymbol{\alpha}}$ is Lipschitz continuous on $\mathcal{X}_\varepsilon$ since it is differentiable everywhere on $\mathcal{X}_\varepsilon$ and has bounded derivatives. Denote by $N^\ell_{x_0}$ the corresponding Lipschitz constant, then

$$\left| D^\ell_{\overline{v}(\xi(x^*),x^*)} f(\xi(x^*)) - D^\ell_{\overline{v}(\xi(x^*),y^*)} f(\xi(x^*)) \right| \leq \left(\textstyle\sum_{\boldsymbol{\alpha} \in \mathbb{N}^d : |\boldsymbol{\alpha}| = \ell} M \frac{\ell!}{\boldsymbol{\alpha}!} \right) \sup_{x_0 \in \mathcal{D}} N^\ell_{x_0} \|x^* - y^*\|_2. \quad (54)$$

Again, since $\mathcal{D}$ is compact the supremum is attained and

$$\bar{N}^\ell \coloneqq \left(\sum_{\boldsymbol{\alpha} \in \mathbb{N}^d : |\boldsymbol{\alpha}| = \ell} M \tfrac{\ell!}{\boldsymbol{\alpha}!} \right) \sup_{x_0 \in \mathcal{D}} N^\ell_{x_0} < \infty.$$

Hence, combining (50), (52) and (54) we get in case (ii) that

$$\left| D^{\ell}_{\overline{v}(\xi(x^*),x^*)} f(\xi(x^*)) \cdot \|x^* - \xi(x^*)\|_2^{\ell} - D^{\ell}_{\overline{v}(\xi(x^*),y^*)} f(\xi(x^*)) \cdot \|y^* - \xi(x^*)\|_2^{\ell} \right| \leq K \bar{L}^{\ell} \varepsilon + \bar{N}^{\ell} C \varepsilon, \tag{55}$$

where $C \coloneqq \sup_{x,y \in \mathcal{X}} \|x - y\| < \infty$ (since $\mathcal{X}$ is compact). Combining the bound (51) from case (i) and the bound (55) from case (ii) implies

$$\begin{aligned} &\left| D^{\ell}_{\overline{v}(\xi(x^*),x^*)} f(\xi(x^*)) \cdot \|x^* - \xi(x^*)\|_2^{\ell} - D^{\ell}_{\overline{v}(\xi(x^*),y^*)} f(\xi(x^*)) \cdot \|y^* - \xi(x^*)\|_2^{\ell} \right| \\ &\quad \leq \max\left(14K, K\bar{L}^{\ell} + \bar{N}^{\ell} C\right) \varepsilon. \end{aligned} \tag{56}$$

The only summand in (49) that remains to be bounded is the one involving supremum terms. For this term the exact same bounds as above apply except for the bound in (53). To bound this term, use compactness of $\mathcal{D}$ to define the function $\nu : \mathcal{X} \to \mathcal{D}$ which satisfies for all $x \in \mathcal{X}$ that

$$\sup_{z \in \mathcal{D}} D^{q}_{\overline{v}(\xi(x^*),x)} f(z) = D^{q}_{\overline{v}(\xi(x^*),x)} f(\nu(x)).$$

Then, using the triangle inequality we get

$$\begin{aligned} &\big| \sup_{z \in \mathcal{D}} D^{q}_{\overline{v}(\xi(x^*),x^*)} f(z) - \sup_{z \in \mathcal{D}} D^{q}_{\overline{v}(\xi(x^*),y^*)} f(z) \big| \\ &\quad = \big| D^{q}_{\overline{v}(\xi(x^*),x^*)} f(\nu(x^*)) - D^{q}_{\overline{v}(\xi(x^*),y^*)} f(\nu(y^*)) \big| \\ &\quad \leq \big| D^{q}_{\overline{v}(\xi(x^*),x^*)} f(\nu(x^*)) - D^{q}_{\overline{v}(\xi(x^*),y^*)} f(\nu(x^*)) \big| + \big| D^{q}_{\overline{v}(\xi(x^*),y^*)} f(\nu(y^*)) - D^{q}_{\overline{v}(\xi(x^*),x^*)} f(\nu(y^*)) \big| \\ &\qquad\quad + \big| D^{q}_{\overline{v}(\xi(x^*),y^*)} f(\nu(x^*)) - D^{q}_{\overline{v}(\xi(x^*),x^*)} f(\nu(y^*)) \big|. \end{aligned}$$

Now for the last term we can observe that if $D^{q}_{\overline{v}(\xi(x^*),y^*)} f(\nu(x^*)) \leq D^{q}_{\overline{v}(\xi(x^*),x^*)} f(\nu(y^*))$, we can use that by definition of ν it holds that $D^{q}_{\overline{v}(\xi(x^*),x^*)} f(\nu(y^*)) \leq D^{q}_{\overline{v}(\xi(x^*),x^*)} f(\nu(x^*))$ and hence

$$\big| D^{q}_{\overline{v}(\xi(x^*),y^*)} f(\nu(x^*)) - D^{q}_{\overline{v}(\xi(x^*),x^*)} f(\nu(x^*)) \big|.$$

Similarly if $D^{q}_{\overline{v}(\xi(x^*),y^*)} f(\nu(x^*)) \geq D^{q}_{\overline{v}(\xi(x^*),x^*)} f(\nu(y^*))$ we get

$$\big| D^{q}_{\overline{v}(\xi(x^*),y^*)} f(\nu(y^*)) - D^{q}_{\overline{v}(\xi(x^*),x^*)} f(\nu(y^*)) \big|.$$

Hence, combined we get that

$$\begin{aligned} &\big| \sup_{z \in \mathcal{D}} D^{q}_{\overline{v}(\xi(x^*),x^*)} f(z) - \sup_{z \in \mathcal{D}} D^{q}_{\overline{v}(\xi(x^*),y^*)} f(z) \big| \\ &\quad \leq 2 \big| D^{q}_{\overline{v}(\xi(x^*),x^*)} f(\nu(x^*)) - D^{q}_{\overline{v}(\xi(x^*),y^*)} f(\nu(x^*)) \big| + 2 \big| D^{q}_{\overline{v}(\xi(x^*),y^*)} f(\nu(y^*)) - D^{q}_{\overline{v}(\xi(x^*),x^*)} f(\nu(y^*)) \big|. \end{aligned}$$

Using the same argument as for (54) this results in

$$\big| \sup_{z \in \mathcal{D}} D^{q}_{\overline{v}(\xi(x^*),x^*)} f(z) - \sup_{z \in \mathcal{D}} D^{q}_{\overline{v}(\xi(x^*),y^*)} f(z) \big| \leq 4 \bar{N}^{q} \|x^* - y^*\|.$$

In total, we get the following bound for the supremum term

$$\begin{aligned}&\left|\sup_{z\in\mathcal{D}} D^q_{\overline{v}(\xi(x^*),x^*)} f(z)\|x^*-\xi(x^*)\|_2^q - \sup_{z\in\mathcal{D}} D^q_{\overline{v}(\xi(x^*),y^*)} f(z)\|y^*-\xi(x^*)\|_2^q\right| \\ &\quad\le \max\left(14K, K\bar{L}^q + 4\bar{N}^q C\right)\varepsilon. \end{aligned} \tag{57}$$

Finally, using the bounds (56) and (57) in (49) leads to

$$\begin{aligned}&\left|B^{\xi(x^*)}(x^*) - B^{\xi(x^*)}(y^*)\right| \\ &\quad\le \varepsilon \underbrace{\left(\sum_{\ell=0}^{q-1}\frac{1}{\ell!}\max\left(14K, K\bar{L}^\ell + \bar{N}^\ell C\right) + \max\left(14K, K\bar{L}^q + 4\bar{N}^q C\right)\right)}_{=:c}. \end{aligned} \tag{58}$$

Since, $c > 0$ only depends on K, $\bar{L}^\ell$, $\bar{N}^\ell$, C and q, we have shown for all $\varepsilon > 0$ and all $x^*, y^* \in \mathcal{X}$ with $\|x^* - y^*\|_2 < \varepsilon$ that

$$\left|B^{\text{up}}_{q,f,\mathcal{D}}(x^*) - B^{\text{up}}_{q,f,\mathcal{D}}(y^*)\right| \le c\,\varepsilon,$$

which completes the proof of Lemma 13. ■

Lemma 14 *Assume $\mathcal{X}$ is compact, let $f \in C^q(\mathcal{X})$ and $\mathcal{D} \subseteq \mathcal{X}$ non-empty closed. For all $z \in \mathcal{X}$ let $B_z : \mathbb{R}^d \to \mathbb{R}$ be the functions defined in (27) in the proof of Theorem 3. Then, the following statements hold:*

(i) For all $z \in \mathcal{X}$ the function B_z is infinitely often continuously differentiable and for all $x \in \mathbb{R}$ and all $\boldsymbol{\beta} \in \mathbb{N}^d$ with $|\boldsymbol{\beta}| \le q$ it holds that

$$\partial^{\boldsymbol{\beta}} B_z(x) = \sum_{|\boldsymbol{\alpha}|<q} \partial^{\boldsymbol{\alpha}} f(\xi(z)) \tfrac{(x-\xi(z))^{\boldsymbol{\alpha}-\boldsymbol{\beta}}}{(\boldsymbol{\alpha}-\boldsymbol{\beta})!}\mathbb{1}(\boldsymbol{\beta}\le\boldsymbol{\alpha}) + \sum_{|\boldsymbol{\alpha}|=q} \partial^{\boldsymbol{\alpha}} f(\nu(z)) \tfrac{(x-\xi(z))^{\boldsymbol{\alpha}-\boldsymbol{\beta}}}{(\boldsymbol{\alpha}-\boldsymbol{\beta})!}\mathbb{1}(\boldsymbol{\beta}\le\boldsymbol{\alpha})$$

In particular, for $|\boldsymbol{\beta}| = q$ this implies

$$\partial^{\boldsymbol{\beta}} B_z(x) = \partial^{\boldsymbol{\beta}} f(\nu(z)).$$

(ii) There exists a constant $C > 0$ such that for all $\varepsilon > 0$ and all $z \in \mathcal{X}$ it holds that

$$\sup_{\substack{x,y\in\mathcal{X}:\\ \|x-y\|_2<\varepsilon}} \left|B_z(x) - B_z(y)\right| < C\,\varepsilon.$$

Proof Since B_z is a polynomial of order q it is infinitely often continuously differentiable. Moreover, a direct computation shows for all $x \in \mathbb{R}^d$ and all $\boldsymbol{\beta} \in \mathbb{N}^d$ with $|\boldsymbol{\beta}| \le q$ that

$$\partial^{\boldsymbol{\beta}} B_z(x) = \sum_{|\boldsymbol{\alpha}|<q} \partial^{\boldsymbol{\alpha}} f(\xi(z)) \tfrac{(x-\xi(z))^{\boldsymbol{\alpha}-\boldsymbol{\beta}}}{(\boldsymbol{\alpha}-\boldsymbol{\beta})!}\mathbb{1}(\boldsymbol{\beta}\le\boldsymbol{\alpha}) + \sum_{|\boldsymbol{\alpha}|=q} \partial^{\boldsymbol{\alpha}} f(\nu(z)) \tfrac{(x-\xi(z))^{\boldsymbol{\alpha}-\boldsymbol{\beta}}}{(\boldsymbol{\alpha}-\boldsymbol{\beta})!}\mathbb{1}(\boldsymbol{\beta}\le\boldsymbol{\alpha}).$$

For $|\boldsymbol{\beta}| = q$ all summands are zero except the one with $\boldsymbol{\alpha} = \boldsymbol{\beta}$, hence

$$\partial^{\boldsymbol{\beta}} B_z(x) = \partial^{\boldsymbol{\beta}} f(\nu(z)).$$

Next, we prove (ii). Fix $x, y, z \in \mathcal{X}$. Then, it holds that

$$\begin{aligned}\big|B_z(x) - B_z(y)\big| \leq \sum_{|\boldsymbol{\alpha}|<q} \tfrac{1}{\boldsymbol{\alpha}!}\big|\partial^{\boldsymbol{\alpha}} f(\xi(z))\big| \cdot \big|(x-\xi(z))^{\boldsymbol{\alpha}} - (y-\xi(z))^{\boldsymbol{\alpha}}\big| \\ + \sum_{|\boldsymbol{\alpha}|=q} \tfrac{1}{\boldsymbol{\alpha}!}\big|\partial^{\boldsymbol{\alpha}} f(\nu(z))\big| \cdot \big|(x-\xi(z))^{\boldsymbol{\alpha}} - (y-\xi(z))^{\boldsymbol{\alpha}}\big|. \qquad (59)\end{aligned}$$

Next, we define the polynomial $p_z^{\boldsymbol{\alpha}}$ for all $w \in \mathcal{X}$ by $p_z^{\boldsymbol{\alpha}}(w) \coloneqq (w-\xi(z))^{\boldsymbol{\alpha}}$. Its gradient is given by

$$\nabla p_z^{\boldsymbol{\alpha}}(w) = \begin{pmatrix} (w-\xi(z))^{\boldsymbol{\alpha}-e_1} \\ \vdots \\ (w-\xi(z))^{\boldsymbol{\alpha}-e_d} \end{pmatrix}.$$

Hence, it holds that

$$\begin{aligned} L_z^{\boldsymbol{\alpha}} &\coloneqq \sup_{w\in\mathcal{X}} \|\nabla p_z(w)\|_2 \\ &= \sup_{w\in\mathcal{X}} \left(\sum_{j=1}^d \big((w-\xi(z))^{\boldsymbol{\alpha}-e_j}\big)^2\right)^{\frac{1}{2}} \\ &\leq \left(\sum_{j=1}^d \left(\big(\sup_{w\in\mathcal{X}} |w^1 - \xi(z)^1|, \ldots, \sup_{w\in\mathcal{X}} |w^d - \xi(z)^d|\big)^\top\right)^{2\boldsymbol{\alpha}-2e_j}\right)^{\frac{1}{2}} \\ &\leq \left(\sum_{j=1}^d \left(\big(\sup_{w,v\in\mathcal{X}} |w^1 - v^1|, \ldots, \sup_{w,v\in\mathcal{X}} |w^d - v^d|\big)^\top\right)^{2\boldsymbol{\alpha}-2e_j}\right)^{\frac{1}{2}} =: \bar{L}^{\boldsymbol{\alpha}}. \end{aligned}$$

Moreover, by compactness of $\mathcal{X}$ it holds that $\sup_{v,w\in\mathcal{X}} |w^j - v^j| < \infty$ and therefore also $\bar{L}^{\boldsymbol{\alpha}} < \infty$. This implies for all $x, z \in \mathcal{X}$ that

$$\big|(x-\xi(z))^{\boldsymbol{\alpha}} - (y-\xi(z))^{\boldsymbol{\alpha}}\big| = |p_z(x) - p_z(y)| \leq L_z^{\boldsymbol{\alpha}} \|x-y\|_2 \leq \bar{L}^{\boldsymbol{\alpha}} \|x-y\|_2.$$

Together with (59) and since all partial derivatives of f up to order q are bounded by a constant $K > 0$ this yields

$$\big|B_z(x) - B_z(y)\big| \leq \underbrace{\left(\sum_{|\boldsymbol{\alpha}|<q} \tfrac{1}{\boldsymbol{\alpha}!} K \cdot \bar{L}^{\boldsymbol{\alpha}} + \sum_{|\boldsymbol{\alpha}|=q} \tfrac{1}{\boldsymbol{\alpha}!} K \cdot \bar{L}^{\boldsymbol{\alpha}}\right)}_{=:C<\infty} \|x-y\|_2.$$

As $x, y, z \in \mathcal{X}$ were arbitrary and C does not depend on them, this implies for all $z \in \mathcal{X}$ and all $\varepsilon > 0$ that

$$\sup_{\substack{x,y\in\mathcal{X}:\\ \|x-y\|_2<\varepsilon}} \big|B_z(x) - B_z(y)\big| < C\,\varepsilon. \qquad (60)$$

This completes the proof of Lemma 14. ■

References

Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, pages 145–155. PMLR, 2020.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148 – 1178, 2019.

Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

Simon Brodersen and Niklas Pfister. adaXT: Fast adaptable and extendable trees for research. `https://github.com/NiklasPfister/adaXT`, 2024. Python package.

Victor Chernozhukov and Christian Hansen. An iv model of quantile treatment effects. *Econometrica*, 73(1):245–261, 2005.

Rune Christiansen, Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco, and Jonas Peters. A causal framework for distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6614–6630, 2021.

Wenlin Dai, Tiejun Tong, and Marc G Genton. Optimal estimation of derivatives in nonparametric regression. *Journal of Machine Learning Research*, 17(1):5700–5724, 2016.

Kris De Brabanter, Joseph De Brabanter, Irene Gijbels, and Bart De Moor. Derivative estimation with local polynomial fitting. *Journal of Machine Learning Research*, 14(1): 281–301, 2013.

Kefan Dong and Tengyu Ma. First steps toward understanding the extrapolation of nonlinear models to unseen domains. In *The Eleventh International Conference on Learning Representations*, 2022.

Bradley Efron. Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics*, 9(2):139–158, 1981.

Zheng Fang and Andres Santos. Inference on directionally differentiable functions. *The Review of Economic Studies*, 86(1):377–412, 2019.

Gerald B Folland. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

Piet Groeneboom and Geurt Jongbloed. *Nonparametric estimation under shape constraints*. Number 38. Cambridge University Press, 2014.

Wolfgang Härdle and Thomas M. Stoker. Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84(408): 986–995, 1989.

Pierre Hiernaux, Hassane Bil-Assanou Issoufou, Christian Igel, Ankit Kariryaa, Moussa Kourouma, Jérôme Chave, Eric Mougin, and Patrice Savadogo. Allometric equations to estimate the dry mass of sahel woody plants mapped with very-high resolution satellite imagery. *Forest Ecology and Management*, 529:120653, 2023.

Lars Hörmander. *The analysis of linear partial differential operators I: Distribution theory and Fourier analysis*. Springer, 2015.

Reid A. Johnson. quantile-forest: A python package for quantile regression forests. *Journal of Open Source Software*, 9(93):5976, 2024. doi: 10.21105/joss.05976. URL `https://doi.org/10.21105/joss.05976`.

Harvey Klyne and Rajen D. Shah. Average partial effect estimation using double machine learning. *arXiv preprint arXiv:2308.09207*, 2023.

Xiaochun Li and Nancy E Heckman. Local linear extrapolation. *Journal of Nonparametric Statistics*, 15(4-5):565–578, 2003.

Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006.

Anton Rask Lundborg and Niklas Pfister. Perturbation-based analysis of compositional data. *arXiv preprint arXiv:2311.18501*, 2023.

YP Mack and Hans-Georg Müller. Derivative estimation in nonparametric regression with random predictor variable. *Sankhyā: The Indian Journal of Statistics, Series A*, 51(1): 59–72, 1989.

Elias Masry. Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of Time Series Analysis*, 17(6):571–599, 1996.

Elias Masry and Jianqing Fan. Local polynomial estimation of regression functions for mixing processes. *Scandinavian Journal of Statistics*, 24(2):165–179, 1997.

Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(35):983–999, 2006.

Warwick Nash, Tracy Sellers, Simon Talbot, Andrew Cawthorn, and Wes Ford. Abalone. UCI Machine Learning Repository, 1995. DOI: https://doi.org/10.24432/C55C7W.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

Niklas Pfister, Stefan Bauer, and Jonas Peters. Learning stable and predictive structures in kinetic systems. *Proceedings of the National Academy of Sciences*, 116(51):25405–25411, 2019.

Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

Herbert E Robbins. An empirical bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and basic theory*, pages 388–394. Springer, 1992.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Sorawit Saengkyongam, Elan Rosenfeld, Pradeep Ravikumar, Niklas Pfister, and Jonas Peters. Identifying representations for intervention extrapolation. In *Proceedings of the 12th International Conference on Learning Representations*, 2024.

Xinwei Shen and Nicolai Meinshausen. Engression: extrapolation through the lens of distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(3):653–677, 2025.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053, 1982.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8 (35):985–1005, 2007.

Brook Taylor. Methodus incrementorum directa et inversa [direct and reverse methods of incrementation]. *Pearsonianis prostant apud Gul. Innys*, 1715.

James W Taylor. A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of forecasting*, 19(4):299–311, 2000.

Grace Wahba. *Spline models for observational data*. SIAM, 1990.

Meijia Wang, Jingyu He, and P Richard Hahn. Local Gaussian process extrapolation for BART models with applications to causal inference. *Journal of Computational and Graphical Statistics*, 33(2):724–735, 2024.

Wen Wu Wang and Lu Lin. Derivative estimation based on difference sequence via locally weighted least squares regression. *Journal of Machine Learning Research*, 16(1):2617–2641, 2015.

Andrew Wilson and Ryan Adams. Gaussian process kernels for pattern discovery and extrapolation. In *International conference on machine learning*, pages 1067–1075. PMLR, 2013.