Essays in Nonparametric Econometrics

Inauguraldissertation

zur Erlangung des Grades eines Doktors
der Wirtschaftswissenschaften
durch die
Rechts- und Staatswissenschaftliche Fakultät
der Rheinischen Friedrich-Wilhelms-Universität
Bonn

vorgelegt von
Jan Scherer
aus Lörrach

2026

Dekan:	Prof. Dr. Martin Böse
Erstreferent:	Prof. Dr. Joachim Freyberger
Zweitreferent:	Prof. Dr. Michael Vogt
Tag der mündlichen Prüfung:	13. August 2025

Abstract

In econometrics, the choice of model specification plays a central role in shaping empirical analysis and inference. Traditional parametric models impose specific functional forms on the relationships among economic variables, such as linear regressions or logit models, which can provide interpretability and high statistical power if the model is correctly specified but may lead to wrong conclusions if the assumed functional forms are incorrect. In response, nonparametric and semiparametric models have emerged as flexible alternatives which relax functional form assumptions and allow the data to determine the shape of the relationship between variables.

This dissertation consists of three chapters on inference in various non- and semiparametric problems. The chapters that are self-contained and can be read separately. Each chapter ends with an appendix that collects the proofs and technical details.

In the first chapter, we study inference on parameters of the form $\phi(\theta_0)$, where ϕ is a known directionally differentiable transformation and θ_0 is an unknown parameter. We focus on settings, where θ_0 is an unknown function estimated using some nonparametric estimator $\hat{\theta}_n$. As many nonparametric estimators do not converge in distribution, existing extensions to the Delta method are not applicable in our setting. We propose to use strong approximations to the distribution of $\hat{\theta}_n$ as an alternative concept to convergence in distribution. Further, we present a notion of directional differentiability which is sufficiently flexible to handle the irregularity of nonparametric estimators. These concepts enable us to derive a new Delta method which approximates the distribution of the plug-in estimator $\phi(\hat{\theta}_n)$. Since these distributional approximations are rarely pivotal, we suggest a simulation-based estimator and provide conditions for its consistency. Confidence intervals based on this estimator are shown to provide local size control under conditions on the directional derivative of ϕ. We illustrate the applicability of our results in two examples and study its finite sample performance in a simulation study.

Anti-concentration bounds play an important role in the modern theory on confidence intervals and testing in settings such as high-dimensional and nonparametric statistics. In the second chapter, we establish such a bound for sublinear and continuous functionals of tight Gaussian random vectors in real-valued Banach spaces. The bound is dimension-free and therefore equally applies to finite- as well as infinite-dimensional settings. It imposes only weak restrictions on the covariance structure

of the Gaussian vectors. As an application of our anti-concentration bound, we derive Berry-Esseen type bounds for sublinear and continuous functionals of high-dimensional mean vectors and kernel-type estimators.

The last chapter, which is joint work with Michael Vogt, studies estimation and inference in the high-dimensional partially linear model $Y = \delta + m(T) + X^\intercal\beta + \varepsilon$, where m is a smooth unknown function and β a sparse unknown regression parameter. The dimension of the covariates X is allowed to increase with the sample size and in particular is allowed to be larger than the sample size. We propose an estimator of β which attains the same rates as the infeasible Lasso estimator which knows the unknown function m. Further, we show that ad-hoc estimators of m might be biased due to the estimation of the high-dimensional parameter β and propose an orthogonalized Nadaraya-Watson estimator of m which effectively decreases this high-dimensional bias. This estimator is shown to converge at the same rates as an infeasible Nadaraya-Watson estimator which knows the true value of β. Based on this estimator, we propose a test for the hypothesis that $m = 0$ which generalizes the idea of significance testing in linear models to allow for general nonlinear effects of T on Y. Moreover, we propose a consistent multiplier bootstrap in order to set the critical values and show uniform consistency of the resulting set against local Hölder balls. We study the finite sample performance of our proposed test in a simulation study and demonstrate its good debiasing and power properties.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisors, Joachim Freyberger and Michael Vogt. Their unwavering support, guidance, and constructive feedback have been instrumental throughout my studies. I am truly fortunate to have had two such dedicated mentors, whose deep understanding of econometrics and statistics enriched my research and helped shape this dissertation. Their encouragement, patience, and generosity with their time made a lasting impact on both my academic and personal development.

I would also like to thank Alois Kneip. The numerous discussions with him were very insightful, and I have profited a lot from his knowledge about nonparametric statistics and statistics in general.

I am very indebted to my office mate Björn Höppner. Our uncountable discussions reignited my passion for the field time and time again and served as a constant source of motivation. I am truly grateful for all the support when stuck at a problem and for all the laughter along the way.

I have always appreciated the familiar atmosphere at the Econometrics and Statistics Group and have received valuable feedback from many faculty members and visitors including Christoph Breunig, Lena Janys, Philip Ketz, Dominik Liebl, Konrad Menzel, Vlad Morozov, Claudia Noack, Dennis Schroers, Jörg Stoye and Chris Walsh. Moreover, I would like to thank my fellow graduate students at the BGSE. They have contributed a lot towards making Bonn a lively and pleasant environment for me. Financial support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - EXC 2047/1 is gratefully acknowledged.

Finally, I am especially grateful to my family and friends. They have always stuck by me whatever has happened. Thank you for that!

Contents

List of Figures

List of Tables

Chapter 1

Inference on Directionally Differentiable Functions of Nonparametric Estimators

1.1 Introduction

The Delta method is a fundamental tool of asymptotic statistics and enables researchers to derive the asymptotic distribution of nonlinear functionals of estimators. The classical Delta method studies parameters of the form $\phi(\theta_0)$, where ϕ is a known differentiable transformation and θ_0 is an unknown but estimable parameter. It has been extended independently by Shapiro (1991) and Dümbgen (1993) to transformations which are only directionally differentiable. Such settings arise frequently across statistics and econometrics, including examples such as inference on eigenvalues of covariance matrices (Dümbgen, 1993), empirical Wasserstein distances (Sommerfeld and Munk, 2018), linear and quadratic programming (Hsieh et al., 2022), conditional moment inequality models (Chernozhukov et al. (2013b), Andrews and Shi (2014)), sensitivity analysis (Masten et al., 2020) and inference on value functions (Firpo et al., 2019).

In this paper, we focus on settings where θ_0 is an unknown function estimated using some nonparametric estimator $\hat{\theta}_n$ such as kernel, series or machine learning based procedures. Such nonparametric estimators have the advantage that they can flexibly estimate the unknown function and therefore guard against potential model misspecification. However, this flexibility comes at a price in that many

nonparametric estimators do not converge in distribution when interpreted as an estimator of a function and therefore existing extensions to the Delta method do not apply. This phenomenon has e.g. been observed by Firpo et al. (2019) and Masten et al. (2020) and lead these authors to restrict attention to regular estimators which possess a limiting distribution. Others, as e.g. Freyberger and Larsen (2021), rely on subsampling to perform inference. However, subsampling cannot be easily applied in this setting as it is hard to show that the plug-in estimator converges in distribution without relying on convergence in distribution of the preliminary estimator.

Following Chernozhukov et al. (2013b), we propose to use strong approximations to the distribution of $\hat{\theta}_n$ as an alternative concept to convergence in distribution. Such strong approximation results approximate the distribution of the scaled estimator $r_n(\hat{\theta}_n - \theta_0)$ by a sequence of penultimate processes Z_n satisfying

$$\|r_n(\hat{\theta}_n - \theta_0) - Z_n\| = o_p(1).$$

These penultimate processes usually have an intuitive appeal as approximate distributions and there is a growing literature deriving such strong approximation results (e.g. Chernozhukov et al. (2013b), Belloni et al. (2015), Chen and Christensen (2018), Belloni et al. (2019b), Cattaneo et al. (2022) and Singh and Vijaykumar (2023)).

We derive a new Delta method which provides a strong approximation to the distribution of the plug-in estimator $\phi(\hat{\theta}_n)$. This Delta method relies on a strengthened notion of Fréchet directional differentiability which is sufficiently flexible to handle the irregularity of nonparametric estimators. This flexibility is needed since the plug-in estimator might converge at a different rate than the preliminary estimator $\hat{\theta}_n$ as is well known from the semiparametric statistics literature. In the important special case when ϕ is a real-valued functional and Z_n is Gaussian, we derive conditions implying that the distribution of the plug-in estimator converges to our proposed approximate distribution in Kolmogorov-Smirnov distance. This result is of particular interest in our setting, as it does not rely on the knowledge whether the approximate distribution has a non-degenerate limit or not. And indeed, when ϕ is only directionally differentiable, it is often hard to determine whether the plug-in estimator converges in distribution. Besides the conditions of our Delta method, this result only depends on convexity of the derivative and a lower bound on the variance of the approximate distribution.

While the Delta method allows to study the asymptotic properties of plug-in estimators, it is of limited use for the construction of inference procedures as the derived approximate distribution might depend on unknown parameters both through the penultimate process Z_n and the derivative of ϕ. Following Fang and Santos (2018), we demonstrate how a consistent bootstrap estimator of the approximate distribution of the plug-in estimator can be constructed. This construction relies on two ingredients. First we require a bootstrap estimator for the distribution of the penultimate process Z_n, and secondly we require a consistent estimator of the derivative of ϕ. Further we analyze subsampling, the rescaled bootstrap (Dümbgen, 1993) and the numerical Delta method (Hong and Li, 2020) in our context and derive sufficient conditions for their validity.

The asymptotic approximation implied by our Delta method may depend discontinuously on the parameter θ_0, when θ_0 is a point where ϕ fails to be continuously differentiable in θ. This suggests that resulting inference procedures might fail to be robust against perturbations of the data generating process. We therefore complement our results and study the plug-in estimator in a local asymptotic framework. We find a different approximate distribution in this local framework. In the case where ϕ is a real-valued functional, we show that convexity of the derivative is sufficient to ensure local size control of one-sided confidence intervals. Our results can be seen as extensions of the results in Dümbgen (1993) and Fang and Santos (2018) to our setting.

We illustrate the applicability of our Delta method with two examples. Our first example studies inference on the maximum of a conditional mean function and is inspired by Chernozhukov et al. (2013b). The second example is due to Freyberger and Larsen (2021). They study partial identification bounds in an alternating-offer bargaining model. This example is non-trivial in that the plug-in estimator exhibits multiple qualitatively different limiting distributions with various different rates of convergence. We further study the finite sample performance of our proposed methods using a Monte Carlo simulation.

Related literature: Our proposed Delta method contributes to the large literature on extensions of the Delta method. Many of the proposed extensions to the Delta method require convergence in distribution of the preliminary estimator $\hat{\theta}_n$. Examples include Beutner and Zähle (2010), Phillips (2012), Belloni et al. (2017) and Kasy (2018). As discussed above, the assumption of convergence in distribution precludes the applicability of these Delta methods to many nonparametric estima-

tors. Our Delta method is closest to the results in Shapiro (1991) and Dümbgen (1993). They also allow for directionally differentiable transformations ϕ while assuming convergence of the preliminary estimator in distribution. On the other hand, they impose Hadamard directional differentiability of ϕ which is different to our assumed notion of Fréchet directional differentiability. In particular, neither concept of directional differentiability nests the other. Therefore, our Delta method can be seen as a complement to the results in Shapiro (1991) and Dümbgen (1993) which allows for nonparametric estimators.

We further contribute to the literature on inference on smooth functionals of nonparametric estimators by allowing ϕ to be only directionally differentiable. Examples include the Delta methods by Newey (1997) and Chen and Christensen (2015) for series estimators, Shen (1997) and Chen and Shen (1998) for sieve and penalized MLE, Chen and Pouzo (2015) in the context of semi/nonparametric conditional moment models, Chen and Christensen (2018) for 2SLS in nonparametric instrumental variable estimation and Koltchinskii (2022) in the Gaussian sequence model. All of these papers consider transformations with a linear derivative excluding transformations which are only directionally differentiable. Further, they also do not require that the preliminary estimator converges in distribution but instead derive conditions which imply that the linearized estimator converges in distribution.

Our Delta method is also related to Chernozhukov et al. (2013b). They study inference on suprema and infima of conditional moment equations which are estimated using kernel- or series-based estimators. This problem can be interpreted as a directionally differentiable functional of an unknown curve and therefore fits into our setup. Our motivation to use strong approximations stems from their analysis. Besides that, their analysis differs to our approach. While we approximate the problem through the directional derivative of the transformation, Chernozhukov et al. (2013b) leverage the specific properties of suprema and infima to construct one-sided projection confidence intervals.

Finally, we contribute to the literature on simulation based inference in directionally differentiable problems. Our proposed construction of the bootstrap follows closely the ideas in Fang and Santos (2018) and complements their analysis by allowing for nonparametric preliminary estimators $\hat{\theta}_n$. Similarly, our analysis of the rescaled bootstrap / numerical Delta method complement the results in Dümbgen (1993) and Hong and Li (2020) by allowing for nonparametric estimators $\hat{\theta}_n$. Further, we contribute to the literature studying subsampling of directionally differ-

entiable parameters as e.g. in Andrews and Guggenberger (2009b) and Andrews and Guggenberger (2010). Andrews' and Guggenberger's results differ to ours as they study semiparametric problems where the nonparametric component has no influence on the limiting distribution, while we focus on settings where the approximate distribution is driven by the nonparametric estimator. On the other hand, we restrict attention to local robustness properties, while Andrews and Guggenberger (2009b) and Andrews and Guggenberger (2010) study global uniformity.

Outline: The remainder of the article is organized as follows. In Section 1.2, we provide motivating examples and discuss the Delta method by Shapiro (1991) and Dümbgen (1993), highlighting challenges in the context of nonparametric estimation. Further, we present here our Delta method results. In Section 1.3, we construct consistent simulation based estimators of the approximating distribution and present sufficient conditions for the validity of subsampling and the rescaled bootstrap / numerical Delta method. In Section 1.4, we study local properties of the plug-in estimator and present our results of a small Monte Carlo study in Section 1.5.

1.2 Setup and the Delta method

In this section, we briefly introduce three examples illustrating potential applications, discuss a Delta method for directionally differentiable functionals due to Shapiro (1991) and Dümbgen (1993) and propose a new Delta method which allows θ_0 to be estimated nonparametrically.

1.2.1 Motivating examples

The first example is mainly expository and is used throughout the text to clarify ideas. It may be seen as a simplified version of the problem that arises in inference on intersection bounds as in Chernozhukov et al. (2013b).

Example 1 (Maximum of a conditional mean function). Let X and Y be scalar random variables such that $X \in [0,1]$ and suppose we wish to estimate the parameter

$$\phi(\theta_0) = \max_{x \in [0,1]} \mathrm{E}[Y \mid X = x].$$

Here $\theta_0(x) = \mathrm{E}[Y \mid X = x]$ and the transformation ϕ is given by $\phi(\theta) = \max_{x \in [0,1]} \theta(x)$.

Our second example is taken from Freyberger and Larsen (2021).[1] This example arises in a partial identification analysis in an alternating-offer bargaining model. Similar bounds arise under monotone treatment response, monotone treatment selection and monotone instrumental variables assumptions as in Manski (1997) and Manski and Pepper (2000).

Example 2 (Bargaining bounds). Given some scalar random variables $X \in [0,1]$ and $Y \in [0,1]$, let $(y,x) \mapsto F_{Y|X}(y,x) = \mathrm{P}(Y \leq y \,|\, X = x)$ and ϕ be given by

$$\phi(F_{Y|X}, f_X) = \int \max_{x' \geq x} F_{Y|X}(y,x') f_X(x) dx,$$

where f_X denotes the marginal pdf of X. In this example $\theta_0 = (F_{Y|X}, f_X)$.

Our final example is taken from Kwon and Mbakop (2021) in the context of estimation of the number of components in nonparametric mixture models. Similar parameters of interest arise in rank tests or inference on the singular values of an integral operator.

Example 3 (Singular values of an integral operator). Let X and Y be continuously distributed random variables with $X, Y \in [0,1]$ and denote by f their joint pdf. Let T denote the integral operator given by

$$u \mapsto (Tu)(\cdot) = \int_0^1 f(x,\cdot) u(x) dx$$

and suppose we are interested in the jth singular value, σ_j, of T. In this case, θ_0 corresponds to the joint density f and ϕ to the mapping which maps f to the jth singular value of T.

Common to all of the above examples is that θ_0 is a function-valued parameter and that the maps ϕ depend on the whole function. Further, the studied ϕ are all directionally differentiable in the sense that there exists a function ϕ'_θ such that for any direction h

$$\lim_{t \downarrow 0} \frac{\phi(\theta + th) - \phi(\theta)}{t} = \phi'_\theta(h).$$

[1] Cf. equation (14) there.

In the above examples, the directional derivative $h \mapsto \phi'_\theta(h)$ is in general a nonlinear function. In fact, the nonlinearity of the directional derivative is even a characterizing feature of directional differentiability in the sense that ϕ is (fully) differentiable if the directional derivative is linear.

1.2.2 Discussion of the Delta methods by Shapiro (1991) and Dümbgen (1993)

Shapiro (1991) and Dümbgen (1993) derived independently a Delta method which extends the functional Delta method of Reeds (1976) to allow for directionally differentiable transformations ϕ. In this section, we present their Delta method and highlight challenges which arise when the preliminary estimator is a nonparametric estimator. In order to make the presentation as clear as possible, we slightly simplify the original assumptions.

The Delta method by Shapiro (1991) and Dümbgen (1993) allows for $\theta_0 \in \mathbb{D}$ for some Banach space $\mathbb{D}$ and for transformations ϕ mapping $\mathbb{D}$ to another Banach space $\mathbb{E}$. It relies on two basic assumptions. First, they assume that there is some preliminary or first-step estimator $\hat{\theta}_n$ of θ_0 satisfying

$$r_n(\hat{\theta}_n - \theta_0) \xrightarrow{d} Z$$

for some $\mathbb{D}$-valued tight random variable Z and sequence of real numbers $r_n \to \infty$. The convergence in distribution is meant in the Hoffmann-Jørgenson sense.[2] Secondly, they assume that ϕ is Hadamard directionally differentiable at θ_0 with derivative ϕ'_{θ_0}. Under these conditions, Shapiro (1991) and Dümbgen (1993) show that a Delta method applies, i.e.,

$$r_n(\phi(\hat{\theta}_n) - \phi(\theta_0)) \xrightarrow{d} \phi'_{\theta_0}(Z).$$

There are two problems with the above Delta method, when the preliminary estimator $\hat{\theta}_n$ is a nonparametric estimator. First, many nonparametric estimators do not converge in distribution when $\mathbb{D}$ is a function space such as the space of continuous functions $\mathcal{C}$ or the space of square integrable functions L_2. In order to illustrate this point, suppose that $x \mapsto \theta_0(x)$ denotes a conditional mean function

[2] See e.g. chapter 1.3 in van der Vaart and Wellner (1996) for an introduction to this notion of convergence in distribution.

which we want to estimate nonparametrically using an estimator $\hat{\theta}_n$. For any fixed x, it can be shown for many estimators that $r_n(\hat{\theta}_n(x) - \theta_0(x))$ is asymptotically normally distributed. However, convergence in distribution of the whole process $x \mapsto r_n(\hat{\theta}_n(x) - \theta_0(x))$ usually fails as $r_n(\hat{\theta}_n - \theta_0)$ fails to be asymptotically tight.[3] Secondly, the above Delta method suggests that the plug-in estimator converges at the same rate as the preliminary estimator. However, well-known examples such as estimation of average treatment effects or average derivatives show that the plug-in estimator may converge at a much faster rate than the preliminary estimator.

There have been considered different solutions to the problem with convergence in distribution in the literature. Newey (1997), Chen and Christensen (2015), Chen and Christensen (2018) and Chen and Pouzo (2015) do not rely on convergence in distribution of the preliminary estimator and instead show directly that $r_n\phi'_{\theta_0}(\hat{\theta}_n - \theta_0)$ converges in distribution. This works particularly well when the derivative is linear and $\hat{\theta}_n$ is asymptotically linear as e.g. kernel- or sieve-based estimators. In our setting, where the derivative might fail to be linear, it is much harder to show that $\phi'_{\theta_0}(r_n(\hat{\theta}_n - \theta_0))$ converges in distribution without relying on convergence in distribution of the preliminary estimator $\hat{\theta}_n$.

Another potential solution to the convergence in distribution problem has been used in Chernozhukov et al. (2013b) in the context of inference on intersection bounds. Instead of relying on a limiting distribution, the authors propose to approximate the distribution of the scaled estimator $r_n(\hat{\theta}_n - \theta_0)$ by a sequence of penultimate random vectors Z_n. For series- or kernel-based estimators, they construct a sequence of Gaussian processes Z_n such that

$$\|r_n(\hat{\theta}_n - \theta_0) - Z_n\|_{\mathbb{D}} = o_p(1). \tag{1.1}$$

Such constructions are called strong approximation theorems or coupling constructions in the probability theory literature and there is a growing literature in nonparametric statistics and econometrics deriving such penultimate processes for a variety of different nonparametric estimators. Examples include Rio (1994) for kernel density estimation, Chernozhukov et al. (2013b) and Cattaneo et al. (2022) for kernel-based estimators, Chernozhukov et al. (2013b) and Belloni et al. (2015)

[3] One could argue that the estimator converges in distribution if we choose a different norm or topology on $\mathbb{D}$. However, choosing a different norm also affects the smoothness properties of ϕ. Hence, there is a tradeoff between smoothness of ϕ and distributional convergence of $\hat{\theta}_n$, which we discuss in more detail in Lemma 42 in the supplementary material.

for series estimators, Chen and Christensen (2018) for 2SLS series estimators in nonparametric instrumental variables models, Belloni et al. (2019b) for conditional quantile estimation using series estimators and Singh and Vijaykumar (2023) for kernel-ridge estimators. Furthermore, there are general strong approximation results for empirical processes (e.g. Rio (1994) and Koltchinskii (1994)) and high-dimensional sums of i.i.d. random vectors (e.g. Belloni et al. (2019b) and Cattaneo et al. (2022)) which allow deriving Gaussian approximations to other nonparametric and machine-learning estimators as well. The concept of strong approximations therefore seems to be suitably flexible for a Delta method for general nonparametric estimator.

In order to illustrate this concept, consider again the setting of Example 1.

Example 4 (Maximum of conditional mean continued). For concreteness, we use the local polynomial estimator in order to estimate the unknown conditional mean function θ_0. Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel function, $h_n > 0$ a sequence of bandwidths and $\ell \geq 0$ an integer. Then, the local polynomial estimator $\hat{\beta}_n$ is given by

$$\hat{\beta}_n(x) = \underset{\beta \in \mathbb{R}^{\ell+1}}{\operatorname{argmin}} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n}\right)\left(Y_i - \sum_{j=0}^{\ell} \beta_j \frac{(X_i - x)^j}{j! h^j}\right)^2$$

and the local polynomial estimator of the conditional mean function is $\hat{\theta}_n(x) := \hat{\beta}_{0,n}(x)$. Under fairly standard regularity conditions, Chernozhukov et al. (2013b) show that there exists a sequence $\{Z_n\}$ of zero mean Gaussian processes on $[0, 1]$ satisfying

$$\sup_{x \in [0,1]} |\sqrt{nh_n}\{\hat{\theta}_n(x) - \theta_0(x)\} - Z_n(x)| = O_p\left(\left(\frac{\log^{24} n}{nh_n^2}\right)^{1/4} + \sqrt{nh_n^{2\ell+3}}\right).$$

The covariance function of Z_n is given by

$$\Sigma_n(x, x') = \frac{1}{h_n f_X(x) f_X(x')} \mathrm{E}[\mathbf{K}_h(X_i - x)\mathbf{K}_h(X_i - x')\varepsilon_i^2],$$

where f_X denotes the marginal density of X, $\mathbf{K}$ denotes the asymptotically equivalent kernel, i.e., $\mathbf{K}(x) = U(0)^\intercal S^{-1} K(x) U(x)$, $\mathbf{K}_h(x) = \mathbf{K}(x/h)$, $U(x) = (1, x, x^2/2!, \ldots, x^\ell/\ell!)^\intercal$ and $S \in \mathbb{R}^{(\ell+1)\times(\ell+1)}$ with elements $S_{j,k} = \int K(u) u^{j+k-2} du$, $j, k = 1, \ldots, \ell + 1$.

This covariance function is a first order approximation of the covariance function

of $\sqrt{nh_n}\{\hat{\theta}_n - \theta_0\}$ and can better capture the local dependence structure than an asymptotic approximation based on $\lim_{n\to\infty} \Sigma_n$. Indeed, under standard regularity conditions, the limit of $\Sigma_n(x, x')$ is zero for all $x \neq x'$ and nonzero if $x = x'$ while Σ_n is continuous. In this sense, Z_n has an intuitive appeal as an approximate distribution and can be thought of as a Gaussian random variable which approximately matches the first two moments of the scaled estimator.

The inflexibility of the above Delta method to capture differing rates of the plug-in and the first-stage estimator can be attributed to the choice of the differentiability concept. While the assumed Hadamard differentiability is a rather mild restriction, its implied rate for the approximation of ϕ through its derivative is rather slow and therefore limits the ability of the above Delta method to adapt to a faster rate of the plug-in estimator. In order to illustrate this point, consider the following characterization of Hadamard directional differentiability.[4]

Definition 1. *The map $\phi : \mathbb{D}_\phi \subset \mathbb{D} \to \mathbb{E}$ is Hadamard directionally differentiable at $\theta \in \mathbb{D}_\phi$, if there is a continuous map $\phi'_\theta : \mathbb{D} \to \mathbb{E}$ such that*

$$\lim_{n\to\infty} \sup_{h\in K} \left\| \frac{\phi(\theta + t_n h) - \phi(\theta)}{t_n} - \phi'_\theta(h) \right\|_{\mathbb{E}} = 0$$

for any compact set $K \subset \mathbb{D}$ and all sequences $t_n \downarrow 0$ such that $\theta + t_n h \in \mathbb{D}_\phi$ for all n and any $h \in K$.

For the sake of argument, suppose further that $r_n(\hat{\theta}_n - \theta_0)$ is asymptotically tight. Then

$$r_n(\phi(\hat{\theta}_n) - \phi(\theta_0)) = \phi'_{\theta_0}(r_n(\hat{\theta}_n - \theta_0)) + \underbrace{r_n\left(\phi\left(\theta_0 + \frac{1}{r_n} r_n(\hat{\theta}_n - \theta_0)\right) - \phi(\theta_0)\right)}_{=:R_n}.$$

Here R_n measures the approximation error due to an approximation of ϕ by its directional derivative. Since $r_n(\hat{\theta}_n - \theta_0)$ is asymptotically tight, it concentrates on a compact set $K \subset \mathbb{D}$ with arbitrarily high probability and therefore Hadamard directional differentiability implies $R_n = o_p(1)$. Thus, Hadamard directional differentiability ties the rate of the approximation error R_n to the tightness properties of the preliminary estimator and therefore is not sufficient to analyze settings where

[4] See Shapiro (1990) for other equivalent characterizations of Hadamard directional differentiability.

the plug-in estimator converges faster than the tightness rate of the preliminary estimator.

As a more flexible notion of directional differentiability, we propose to assume that ϕ is γ-Fréchet directionally differentiable as given in the following definition.

Definition 2. *Let $\gamma > 1$. $\phi : \mathbb{D}_\phi \subset \mathbb{D} \to \mathbb{E}$ is called γ-Fréchet directionally differentiable at θ in the interior of $\mathbb{D}_\phi$, if there exists a positively homogeneous function $\phi'_\theta : \mathbb{D} \to \mathbb{E}$ such that,*

$$\|\phi(\theta + h) - \phi(\theta) - \phi'_\theta(h)\|_E = O(\|h\|^\gamma), \tag{1.2}$$

for all h s.t. $\theta + h \in \mathbb{D}_\phi$ and as $\|h\| \to 0$.

In comparison to Hadamard directional differentiability, γ-Fréchet directional differentiability implies

$$\phi(\hat{\theta}_n) - \phi(\theta_0) = \phi'_{\theta_0}(\hat{\theta}_n - \theta_0) + O_p(\|\hat{\theta}_n - \theta_0\|^\gamma)$$

and therefore the approximation error converges faster to zero than the rate of boundedness of the preliminary estimator, allowing for different rates of the plug-in and preliminary estimator. The value of γ measures the speed at which the approximation error converges to zero. The higher γ, the stronger the implied notion of differentiability. We show in Lemma 14 in the Appendix that if ϕ is twice Fréchet directionally differentiable with continuous derivatives, then ϕ satisfies (1.2) for $\gamma \geq 2$. In this sense, γ-Fréchet directional differentiability can be thought of as requiring ϕ to be γ-times Fréchet directionally differentiable.

A further difference of the proposed differentiability concept is that it controls the approximation error via the boundedness properties of the preliminary estimator instead of tightness properties. This further increases the rate of the approximation error as the rate of boundedness typically is faster than the rate of tightness for nonparametric estimators. For example, if $\mathbb{D}$ corresponds to the space of continuous functions on $[0, 1]$ endowed with the supremum norm, the rate of $\|\hat{\theta}_n - \theta_0\|$ differs from the pointwise rate of convergence only by a logarithmic term. On the other hand, $r_n(\hat{\theta}_n - \theta_0)$ is asymptotically tight if and only if it is asymptotically stochastically equicontinuous,[5] meaning that $r_n(\hat{\theta}_n - \theta_0)$ is stochastically bounded and for any

[5] Cf. Theorem 7.3 in Billingsley (1999).

$\varepsilon, \eta > 0$, there is some $\delta > 0$ such that

$$\limsup_{n\to\infty} \mathrm{P}\left(\sup_{|x-x'|\leq\delta} r_n |\{\hat{\theta}_n(x) - \theta_0(x)\} - \{\hat{\theta}_n(x') - \theta_0(x')\}| > \varepsilon \right) < \eta.$$

Heuristically, asymptotic stochastic equicontinuity demands that r_n is chosen in such a way that the derivative of $r_n(\hat{\theta}_n - \theta_0)$ is stochastically bounded, requiring a considerably slower rate r_n as is needed for stochastic boundedness alone.

Alternatively to γ-Frèchet directional differentiability, one might impose a notion of directional differentiability which is more closely related to Hadamard directional differentiability. For example, suppose that

$$\sup_{h\in K} \left\| \frac{\phi(\theta + t_n h) - \phi(\theta)}{t_n} - \phi'_\theta(h) \right\|_E = O(t_n^{\gamma-1}), \qquad \text{as } t_n \downarrow 0 \tag{1.3}$$

for any compact $K \subset \mathbb{D}$ and some $\gamma > 1$. We call any ϕ that satisfies (1.3), γ-Hadamard directionally differentiable. While this concept imposes weaker conditions on the smoothness of ϕ, it bounds the approximation error using the rate of tightness of the preliminary estimator instead of the rate of boundedness as discussed above. Thus, there is a trade-off between the smoothness properties of ϕ and the convergence rates of $\hat{\theta}_n$. Since all the presented examples in Section 1.2.1 are γ-Fréchet directionally differentiable for a suitably chosen norm, we prefer to use γ-Fréchet directional differentiability for our Delta method.

Similar notions of differentiability for fully differentiable transformations with linear derivatives have been proposed in the literature. In the context of the functional Delta method, Dudley (1992), Dudley (1994) and Dudley and Norvaiša (1999) propose to use γ-Fréchet differentiability with respect to the p-variation norm. They show that many commonly used operators such as function composition, multiplication and the quantile function are all γ-Fréchet differentiable for some $\gamma > 1$. In the literature on smooth functionals of nonparametric estimators and extremum or GMM estimation, Newey (1994b), Newey (1997), Shen (1997), Chen and Shen (1998), Chen et al. (2003) and Chen and Christensen (2015) assume γ-Fréchet differentiability with respect to various different norms. In most cases, γ is assumed to be 2. Other papers assume a bound on the approximation error in 2 which can be written as a product of two norms $\|\cdot\|_1^\alpha \cdot \|\cdot\|_2^\beta$ for some $\alpha, \beta > 0$. For example, Newey (1994a) and Chen and Christensen (2018) use such a differentiability concept with the restriction $\alpha = \beta = 1$, while Chen and Liao (2015) and Chen and Pouzo (2015)

do not restrict α and β.

We illustrate the above discussion on the differentiability concept within the context of Example 1:

Example 5 (Maximum of conditional mean function continued). For $\mathbb{D}$ equal to the space of continuous functions endowed with the supremum norm $\|\cdot\|_\infty$, it has been shown in Fang and Santos (2018) that the maximum functional is Hadamard directionally differentiable at any θ_0 with derivative

$$\phi'_{\theta_0}(h) = \max_{x \in \Psi(\theta_0)} h(x),$$

where $\Psi(\theta_0) = \text{argmax}_{x\in[0,1]}\, \theta_0(x)$, which is a nonlinear transformation of h whenever the maximum of θ_0 is non-unique. When $\mathbb{D}$ is taken as the space of Lipschitz continuous functions from $[0,1]$ to $\mathbb{R}$ endowed with the bounded Lipschitz norm $\|\cdot\|_{BL}$ given by

$$\|f\|_{BL} = \max\left\{\|f\|_\infty, \sup_{x\neq y} \frac{|f(x)-f(y)|}{|x-y|}\right\},$$

we can show γ-Fréchet directional differentiability of ϕ under further restrictions on θ_0. For instance, consider functions θ_0 which satisfy the following well-separatedness condition:[6] there exist constants $c, \rho, \delta > 0$ such that for all $x \in [0,1]$,

$$\max_{x' \in [0,1]} \{\theta_0(x')\} - \theta_0(x) \geq (cd(x, \Psi(\theta_0))^\rho) \wedge \delta. \tag{1.4}$$

Many functions satisfy such a condition. For example, if the argmax set is finite and contained in the interior and θ is a smooth function with second derivatives bounded away from zero by a constant c, then this condition is satisfied for some δ and $\rho = 2$. But also examples with a continuum of maximizers or maximizers at the boundary satisfy this condition.

Given that θ_0 satisfies the well-separatedness condition (1.4), we can show, for $\rho > 1$, that ϕ is $(\rho/(\rho-1))$-Fréchet directionally differentiable, i.e.,[7]

$$\phi(\theta_0 + h) = \phi(\theta_0) + \phi'_{\theta_0}(h) + O\big(\|h\|_{BL}^{\frac{\rho}{\rho-1}}\big),$$

[6] This condition can also be interpreted as a degree of identifiability of the argmax set $\Psi(\theta_0)$. See Condition V and its discussion in Chernozhukov et al. (2013b).

[7] See Proposition 1 in the Appendix.

and that this γ is tight. This result crucially depends on the chosen norm on $\mathbb{D}$. When this norm is chosen to be the sup-norm, ϕ fails to be Fréchet and γ-Hadamard directionally differentiable.[8] This phenomenon occurs more generally: if $|\cdot|$ is another norm on $\mathbb{D}$ which is even stronger than $\|\cdot\|$, then γ-Fréchet directional differentiability with respect to $\|\cdot\|$ implies γ-Fréchet directional differentiability with respect to the stronger norm $|\cdot|$. Heuristically, Fréchet directional differentiability with respect to a stronger norm requires (1.2) to only hold for a smaller set of directions and therefore is a weaker condition. The same applies for γ-Hadamard directional differentiability.

1.2.3 The Delta method

Motivated by the above discussion, our Delta method relies on the following two assumptions.

Assumption 1 (On ϕ)**.** *(i)* $\mathbb{D}$ *and* $\mathbb{E}$ *are real Banach spaces with norms* $\|\cdot\|_{\mathbb{D}}$ *and* $\|\cdot\|_{\mathbb{E}}$.

(ii) $\phi : \mathbb{D}_\phi \subset \mathbb{D} \to \mathbb{E}$ *is Borel measurable and* γ*-Fréchet directionally differentiable at* θ_0 *with Borel measurable derivative* $\phi'_{\theta_0} : \mathbb{D} \to \mathbb{E}$. $\mathbb{D}_\phi \subset \mathbb{D}$ *has non-empty interior and* $\theta_0 \in \text{int}(\mathbb{D}_\phi)$. ϕ'_{θ_0} *is not the zero-mapping.*

Assumption 2 (On $\hat{\theta}_n$)**.** *(i) The estimator* $\hat{\theta}_n$ *is a measurable function, mapping a sequence of random variables* $\{X_i\}_{i=1}^n$ *into* $\mathbb{D}_\phi$.

(ii) There exists a sequence of real numbers $a_n \to 0$ *such that* $\|\hat{\theta}_n - \theta_0\| = O_p(a_n)$.

(iii) There exists a sequence of real numbers $b_n \to 0$ *and a sequence of* $\mathbb{D}$*-valued random vectors* $\{Z_n\}_{i=1}^n$ *such that*

$$\|\phi'_{\theta_0}(\hat{\theta}_n - \theta_0) - \phi'_{\theta_0}(Z_n)\|_{\mathbb{E}} = O_p(b_n). \tag{1.5}$$

Assumption 1 summarizes our smoothness requirements on the map ϕ. It requires that the map ϕ is γ-Fréchet directional differentiability at θ_0 only. We assume for simplicity that θ_0 lies in the interior of the domain of ϕ. Instead, we could allow θ_0 to lie on the boundary of $\mathbb{D}_\phi$, by restricting the domain of the derivative to the span of all $h \in \mathbb{D}$ such that $\theta_0 + h \in \mathbb{D}_\phi$ and assuming that Z_n is supported

[8] See Lemma 16 in the Appendix.

on this set. See also Remark 2.1 in Fang and Santos (2018) for further technical details on this point. Further, our notion of directional differentiability differs to the proposals in Shapiro (1991) and Dümbgen (1993) in that we do not formulate a notion of Fréchet differentiability tangential to a set. Such a tangential formulation allows restricting the set of directions to lie in a subset of $\mathbb{D}$ and therefore weakens the required smoothness on ϕ. Our definition can be modified to allow for such a restriction of the directions. In order to facilitate the exposition and since this extension is not needed in our examples, we decided not to include this formulation. Moreover, we assume that the derivative is not degenerate, i.e., not the zero mapping. If this is the case, the approximate distribution $\phi'_{\theta_0}(Z_n)$ would be degenerate and therefore of little use for inference. Similarly to the setting of the classical Delta method, we might then perform a higher order Delta method instead. This has been investigated by Chen and Fang (2019) in the setting of the Delta method by Shapiro (1991) and Dümbgen (1993). While the derivation of a higher order Delta method would be interesting in our setting, it is out of the scope of the current paper.

The measurability assumption on the estimator in Assumption 2 (i) might seem restrictive in comparison to the Delta method in Dümbgen (1993) and Shapiro (1991) who only require asymptotic measurability instead. $\hat{\theta}_n$ can fail to be measurable in as simple examples as the empirical distribution function.[9] On the other hand, when $\hat{\theta}_n$ is some kernel or series-based curve estimator, $\hat{\theta}_n$ is typically measurable under mild assumptions on the used kernel or system of basis functions. Nevertheless, our Delta method extends to a setting where the preliminary estimator fails to be measurable while the approximate distribution sequence Z_n is measurable as we demonstrate in Appendix A.5.

In Assumption 2 (iii), we require the existence of an approximate distribution sequence for the derivative ϕ'_{θ_0} applied to the estimator. Alternatively, we could have assumed that the derivative is Lipschitz continuous and that we have a strong approximation for the whole process $(\hat{\theta}_n - \theta_0)$. This difference can have a large impact on the requirements on the estimator $\hat{\theta}_n$ as well as the rate b_n. For example, Chernozhukov et al. (2014a) derived a coupling for the supremum of a series empirical process under weak requirements on the number of basis functions, while strong approximations for the whole series empirical process often require much stronger restrictions on the number of basis functions (cf. Chernozhukov et al. (2013b), Belloni et al. (2015), Chen and Christensen (2018) and Belloni et al. (2019b)). Moreover,

[9] See Chapter 1.1 in van der Vaart and Wellner (1996).

note that the assumed approximate distribution Z_n is not scaled. An appropriate scaling of Z_n will be later determined by the rate of the plug-in estimator.

We proof the following Delta method in the Appendix.

Theorem 1. *Let Assumptions 1 and 2 hold. Then,*

$$\phi(\hat{\theta}_n) - \phi(\theta_0) = \phi'_{\theta_0}(Z_n) + O_p(a_n^\gamma + b_n).$$

Theorem 1 formalizes in what sense $\phi'_{\theta_0}(Z_n)$ can be thought of as an approximate distribution of the plug-in estimator. It quantifies the size of the approximation error and separates it into the two components a_n^γ and b_n. a_n^γ measures the analytical approximation error caused by the approximation of ϕ through its derivative and b_n measures the stochastic approximation error induced by the distributional approximation of the preliminary estimator. The result might seem unconventional for a Delta method due to the chosen mode of convergence. However, if we further know that the approximand converges in distribution, say $r_n\phi'_{\theta_0}(Z_n) \xrightarrow{d} G$ for some rate r_n and limiting distribution G, Theorem 1 implies that $r_n(\phi(\hat{\theta}_n) - \phi(\theta_0)) \xrightarrow{d} G$ whenever $r_n(a_n^\gamma + b_n) \to 0$. Moreover, when $\mathbb{E}$ is finite dimensional, say $\mathbb{E} = \mathbb{R}$ for simplicity, Theorem 1 implies that the studentized plug-in estimator converges in distribution along subsequences under further moment conditions. Indeed, we show in Lemma 5 in the Appendix that there exists a subsequence (n_k) such that

$$\frac{\phi(\hat{\theta}_{n_k}) - \phi(\theta_0) - \mathrm{E}[\phi'_{\theta_0}(Z_{n_k})]}{\sqrt{\mathrm{Var}(\phi'_{\theta_0}(Z_{n_k}))}} \xrightarrow{d} G$$

for some non-degenerate limiting distribution G as long as the approximation error $(a_n^\gamma + b_n)$ is of smaller order than the standard deviation of $\phi'_{\theta_0}(Z_n)$.

The imposed assumptions of Theorem 1 are rather mild. For example, in the parametric setup, when both $\mathbb{D}$ and $\mathbb{E}$ are finite dimensional and $r_n(\hat{\theta}_n - \theta_0)$ converges in distribution to some limit distribution G, Theorem 1 implies $r_n(\phi(\hat{\theta}_n) - \phi(\theta_0)) \xrightarrow{d} \phi'_{\theta_0}(G)$ as long as ϕ is γ-Fréchet directionally differentiable for some $\gamma > 1$ with continuous derivative $h \mapsto \phi'_{\theta_0}(h)$. Indeed, by tightness of $r_n(\hat{\theta}_n - \theta_0)$, a_n can be chosen as r_n^{-1}, Strassen's Theorem[10] implies that there is a sequence $Z_n \sim G/r_n$ with $b_n = o(r_n^{-1})$ since $r_n(\hat{\theta}_n - \theta_0)$ converges in distribution and continuity of the derivative implies $r_n\phi'_{\theta_0}(Z_n) \xrightarrow{d} G$.

[10] Cf. Chapter 10, Theorem 8 in Pollard (2001).

Regarding the construction of confidence intervals, Theorem 1 is strong enough to bound the quantiles of functionals of the plug-in estimator by the quantiles of the approximate distribution. Indeed, for any Lipschitz continuous functional $\nu : \mathbb{E} \to \mathbb{R}$ such as point evaluation, an average or the norm, Theorem 1 implies for any $\varepsilon > 0$, that there is some K such that for sufficiently large n,

$$\begin{aligned} Q_{p-\varepsilon}(\nu(\phi'_{\theta_0}(Z_n))) - K(a_n^\gamma + b_n) \leq & Q_p(\nu(\phi(\hat{\theta}_n) - \phi(\theta_0))) \\ \leq & Q_{p+\varepsilon}(\nu(\phi'_{\theta_0}(Z_n))) + K(a_n^\gamma + b_n), \end{aligned}$$

where $Q_p(Z) = \inf\{t \in \mathbb{R} : \mathrm{P}(Z \leq t) \geq p\}$.[11] In particular, when the approximation error $(a_n^\gamma + b_n)$ converges faster than the standard deviation of the approximate distribution $\sigma_n = \sqrt{\mathrm{Var}(\nu(\phi'_{\theta_0}(Z_n)))}$, Theorem 1 implies

$$Q_{p-\varepsilon}(\nu(\phi'_{\theta_0}(Z_n))) - \varepsilon\sigma_n \leq Q_p(\nu(\phi(\hat{\theta}_n) - \phi(\theta_0))) \leq Q_{p+\varepsilon}(\nu(\phi'_{\theta_0}(Z_n))) + \varepsilon\sigma_n$$

for any $\varepsilon > 0$ and sufficiently large n and therefore that the quantiles of the approximate distribution are close to the quantiles of the plug-in estimator in comparison to the standard deviation of the approximate distribution.

The asymptotic behavior of the quantiles can be strengthened under equicontinuity and monotony conditions on the sequence of distribution functions of $\phi'_{\theta_0}(Z_n)$. While in general it is hard to check such conditions, under further analytic assumptions on the derivative and if the penultimate process Z_n is Gaussian, we can strengthen our Delta method in Theorem 1 to a bound in the Kolmogorov distance.

Theorem 2. *Suppose that Assumptions 1 and 2 hold with $\mathbb{E} = \mathbb{R}$. Further, suppose that ϕ'_{θ_0} is sublinear and continuous and that Z_n is a sequence of centered and tight Gaussian random vectors. Let $\sigma_n^2 = \mathrm{Var}(\phi'_{\theta_0}(Z_n)) > 0$. If*

$$\frac{a_n^\gamma + b_n}{\sigma_n} \to 0,$$

then

$$\lim_{n\to\infty} \sup_{t>0} \left| \mathrm{P}(\phi(\hat{\theta}_n) - \phi(\theta_0) \leq t) - \mathrm{P}\big(\phi'_{\theta_0}(Z_n) \leq t\big) \right| = 0.$$

[11] See Lemma 6 in the Appendix.

If further $\mathcal{F} = \{f \in \mathbb{D}^* \,|\, \forall h \in \mathbb{D} : f(h) \leq \phi'_{\theta_0}(h),\, \mathrm{Var}(f(Z_n)) = 0\} = \emptyset$,[12] *then*

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}} \big|\mathrm{P}(\phi(\hat{\theta}_n) - \phi(\theta_0) \leq t) - \mathrm{P}\big(\phi'_{\theta_0}(Z_n) \leq t\big)\big| = 0.$$

Theorem 2 implies that the approximate distribution $\phi'_{\theta_0}(Z_n)$ can approximate the finite sample distribution of the plug-in estimator. One appealing aspect of an approximation in Kolmogorov distance is that we do not need to know whether the plug-in estimator converges in distribution or not. We only need to know that the Delta method approximation error $(a_n^\gamma + b_n)$ is of smaller order than the standard deviation of $\phi'_{\theta_0}(Z_n)$ as measured by σ_n. In particular, we do not need to know at which rate the plug-in estimator converges; a lower bound on the standard deviation is sufficient. Moreover, Theorem 2 implies that the quantiles of the approximate distribution converge to the quantiles of the plug-in estimator in the following sense: for any $p \in (0,1)$ such that $Q_p(\phi'_{\theta_0}(Z_n)) > 0$,[13] we have

$$\lim_{n\to\infty} \mathrm{P}(\phi(\hat{\theta}_n) - \phi(\theta_0) \leq Q_p(\phi'_{\theta_0}(Z_n))) = p,$$

which can be used in order to show pointwise validity of one-sided confidence sets. A similar result holds when $\mathbb{E}$ is a general Banach space, and we are only interested in a functional $\nu : \mathbb{E} \to \mathbb{R}$ of $\phi(\hat{\theta}_n) - \phi(\theta_0)$ such as point evaluation or the norm. Then Theorem 2 can be shown to hold as long as $\nu \circ \phi'_{\theta_0}$ is continuous, positively homogeneous and sublinear and if one substitutes σ_n with the standard deviation of $\nu(\phi'_{\theta_0}(Z_n))$. This is for example satisfied when the derivative is continuous and linear and $\nu(\cdot) = \|\cdot\|_{\mathbb{E}}$, which is useful for the construction of uniform confidence bands on function spaces.

The proof of Theorem 2 relies on an anti-concentration bound for continuous and sublinear functionals of tight Gaussian random variables in our companion paper Scherer (2024). We require centered Gaussian random vectors in order to obtain a cleaner result although the aforementioned anti-concentration bound also allows for non-centered Gaussian random variables. We only require tightness of Z_n for each n instead of asymptotic tightness of the sequence (Z_n) which is likely to be violated for nonparametric estimators. The tightness of Z_n is e.g. guaranteed when $\mathbb{D}$ is separable. In the case $\mathcal{F} \neq \emptyset$, we can only bound the distance between the distribution functions on $(0, \infty)$ since the cdf of $\phi'_{\theta_0}(Z_n)$ might have a jump discontinuity at

[12] Here, $\mathbb{D}^*$ denotes the topological dual space of $\mathbb{D}$.

[13] If $\mathcal{F} = \emptyset$, the sign restriction on the quantile is not needed.

zero. In Scherer (2024), we show that $\phi'_{\theta_0}(Z_n)$ can be represented as a supremum of a Gaussian process indexed by continuous linear functionals. If $\mathcal{F} \neq \emptyset$, this Gaussian process attains the value zero for some functional and therefore bounds the supremum from below by zero. On the other hand, when $\mathcal{F} = \emptyset$, the cdf of $\phi'_{\theta_0}(Z_n)$ is continuous, and we can obtain a bound on the Kolmogorov distance over $\mathbb{R}$.

Examples revisited

We discuss here only the example of the maximum of the conditional mean and refer to the Appendix for the example concerning the bargaining bound.

Example 6 (Maximum of conditional mean continued). Under fairly standard regularity conditions (Assumption (11) in the appendix), the local polynomial estimator $\hat{\theta}_n$ satisfies Assumption 2. In the appendix, we show that

$$\|\hat{\theta}_n - \theta_0\|_{BL} = O_p\Bigg(\underbrace{\frac{\sqrt{\log n}}{h_n}\left\{\sqrt{\frac{\log n}{nh_n}} + h^{\ell+1}\right\}}_{=:a_n}\Bigg).$$

As we presented in Example 4, there exists a sequence $\{Z_n\}$ of zero mean Gaussian processes on $[0,1]$ satisfying[14]

$$\sup_{x\in[0,1]} |\{\hat{\theta}_n(x) - \theta_0(x)\} - Z_n(x)| = O_p\Bigg(\underbrace{\left(\frac{\log^8 n}{nh_n^{4/3}}\right)^{3/4} + h_n^{\ell+1}}_{=b_n}\Bigg).$$

If further the well-separatedness condition (1.4) holds with $\rho \in \{1,2\}$, ϕ is 2-Frèchet directionally differentiable with a continuous and sublinear derivative and therefore satisfies Assumption 1. Theorem 2 implies

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}} \left| \mathrm{P}\left(\max_{x\in[0,1]} \hat{\theta}_n(x) - \max_{x\in[0,1]} \theta_0(x) \leq t \right) - \mathrm{P}\left(\max_{x\in\Psi_0} Z_n(x) \leq t \right) \right| = 0$$

under the rate requirements

$$\frac{\log^{24} n}{nh_n^2} \to 0, \qquad \frac{\log^2 n}{nh_n^5} \to 0 \qquad \text{and} \qquad nh_n^{2\ell+1} \to 0.$$

This result can recover similar rate requirements as in Müller (1985) and Ziegler

[14] See Lemma 19 in the Appendix.

(2002) who restricted attention to conditional mean functions having a unique maximizer. Our result on the other hand only restricts the set of maximizers through the well-separatedness condition and allows for example for finitely many well-separated maximizers as well as a continuum of maximizers. The presented results are similar to the results in section 3.6 in Chernozhukov et al. (2013b). Their condition S on the equicontinuity radius is closely related to our requirement on the bounded Lipschitz norm of the estimator as both measure the equicontinuity of the process. Moreover, their condition V is the same as our well-separatedness condition (1.4). However, our approach differs to their proposal. Firstly, they do not propose a plug-in estimator but instead their estimator can be interpreted as a corrected plug-in estimator $\phi(\hat{\theta}_n + \hat{c}_n)$, where $\hat{c}_n$ is a correction term. Using our Delta method, one can achieve a similar correction using an estimate of the derivative as we will show in our local analysis in Section 1.4. Secondly, their asymptotic approximation uses instead of a derivative the intermediate approximation

$$\phi(\theta_0 + h_n) - \phi(\theta_0) = \max_{x \in \Psi_0^{\delta_n}} h_n(x)$$

for some $\delta_n \to 0$. This approximation allows them to analyze their estimator even for functions that do not satisfy the well-separatedness condition, which we cannot do relying on our Delta method.

1.3 Bootstrap

While the Delta method allows to study the asymptotic properties of plug-in estimators, it is of limited use for the construction of inference procedures as the derived approximate distribution might depend on unknown parameters both through the penultimate process Z_n and the derivative. In this section, we demonstrate how a consistent bootstrap estimator of the approximate distribution of the plug-in estimator can be constructed following similar ideas as in Fang and Santos (2018). Further, in section 1.3.2, we discuss the validity of some general purpose bootstrap techniques which have been proposed in the literature.

The construction of our bootstrap estimator relies on two components. Firstly, we require a bootstrap estimator $\hat{Z}_n$ for the distribution of the penultimate process Z_n in Assumption 3, and secondly we require a consistent estimator $\hat{\phi}'_n$ of the derivative ϕ'_{θ_0} in Assumption 4. Given these estimators, we can then estimate the

approximate distribution of $\phi'_{\theta_0}(Z_n)$ using $\hat{\phi}'_n(\hat{Z}_n)$.

Assumption 3. *(On the Bootstrap $\hat{Z}_n$)*

(i) $\hat{Z}_n : \{X_i, W_i\}_{i=1}^n \to \mathbb{D}$ *measurable with* $\{W_i\}_{i=1}^n$ *independent of* $\mathcal{D}_n = \{X_i\}_{i=1}^n$.

(ii) There is some sequence $d_n \to 0$ *such that* $\|\hat{Z}_n\| = O_p(d_n)$.

(iii) There is some sequence $s_n \to 0$ *and some* $Z_n^* | \mathcal{D}_n \sim Z_n$ *such that for any* $\varepsilon > 0$, *there exist* M *such that*

$$\lim_{n\to\infty} \mathrm{P}\big(\mathrm{P}(\|\phi'_{\theta_0}(\hat{Z}_n) - \phi'_{\theta_0}(Z_n^*)\|_{\mathbb{E}} > M s_n \,|\, \mathcal{D}_n) > \varepsilon\big) = 0.$$

Assumption 3 mirrors the assumptions imposed on the estimator in Assumption 2. We require the bootstrap to be measurable and that we know some rate for the norm of the bootstrap distribution. Note that this rate is assumed to hold unconditionally. A sufficient condition is that a similar bound holds conditionally on the data:[15]

(ii') for any $\varepsilon > 0$, there exist M such that

$$\lim_{n\to\infty} \mathrm{P}\big(\mathrm{P}(\|\hat{Z}_n\| > M d_n \,|\, \mathcal{D}_n) > \varepsilon\big) = 0. \tag{1.6}$$

As Z_n might fail to have a limiting distribution, we cannot use the usual notion of a consistent bootstrap procedure. Instead, we require in Assumption 3(iii) a similar coupling construction as in Assumption 2 (iii). Such results are readily available for various nonparametric estimators. In particular, Chernozhukov et al. (2013b) and Belloni et al. (2015) give a bootstrap for series estimators, Chernozhukov et al. (2013b) and Cattaneo et al. (2017) for kernel regression and local polynomial estimators and Belloni et al. (2019b) for conditional quantile estimation based on series regression.[16] Moreover, we implicitly assume in part (iii) that the bootstrap method can estimate the mean of the asymptotic distribution correctly. Since the mean of Z_n is closely related to the bias of $\hat{\theta}_n$, this means that we usually need to apply some debiasing to $\hat{\theta}_n$ or some undersmoothing of the tuning parameter. Finally, we want to stress some technicality with couplings. Usually the coupled

[15] For a proof, see Lemma 8 in the Appendix.

[16] Most of these results derive a bound on the bounded Lipschitz distance and imply a coupling result as in 3(iii) by Strassen's coupling theorem (cf. Pollard (2001) Theorem 8 in Chapter 10.3).

distribution Z_n^* in Assumption 3(iii) depends on the specific M and ε. Instead, we assume that Z_n^* does neither depend on M nor ε. One can usually obtain such a uniform coupling by slightly increasing s_n. See Remark 2.1 in Cattaneo et al. (2022) for further details on this point.

Besides a bootstrap estimator $\hat{Z}_n$, we further need an estimator $\hat{\phi}'_n$ for the unknown derivative ϕ'_{θ_0}. We assume:

Assumption 4. *(On the estimator $\hat{\phi}'_n$)*
The map $\hat{\phi}'_n : \mathbb{D} \to \mathbb{E}$ is a measurable function of $\{X_i\}_{i=1}^n$, which is positively homogeneous and satisfies for a sequence $\eta_n \to 0$

$$\sup_{\|h\|\leq 1} \|\hat{\phi}'_n(h) - \phi'_{\theta_0}(h)\|_{\mathbb{E}} = O_p(\eta_n).$$

In Assumption 4, we assume that the estimator of the derivative is a uniformly consistent estimator of the derivative at θ_0. Uniform convergence is needed as we not only need to approximate the derivative along a fixed deterministic direction h but along a sequence of random variables which may take on many different values. Further, a simple plug-in estimator of the form $\hat{\phi}_n = \phi'_{\hat{\theta}_n}$ easily fails to satisfy Assumption 4 unless the derivative $\theta \mapsto \phi'_\theta(h)$ is sufficiently smooth. However, when ϕ is only directionally differentiable, the derivative often even fails to be continuous in θ.

Given these assumptions, we can show a notion of consistency of the bootstrap for the approximate distribution $\phi'_{\theta_0}(Z_n)$.

Theorem 3. *Let the Assumptions of Theorem 1, Assumptions 3 and 4 hold. Then, for any $\varepsilon > 0$, there exist M such that*

$$\lim_{n\to\infty} \mathrm{P}\big(\mathrm{P}(\|\hat{\phi}'_n(\hat{Z}_n) - \phi'_{\theta_0}(Z_n^*)\|_{\mathbb{E}} > M(d_n\eta_n \vee s_n) \,|\, \mathcal{D}_n) > \varepsilon\big) = 0.$$

Theorem 3 implies that the bootstrap error $\|\hat{\phi}'_n(\hat{Z}_n) - \phi'_{\theta_0}(Z_n^*)\|_{\mathbb{E}}$ converges with rate $d_n\eta_n \vee s_n$ conditionally on the data $\mathcal{D}_n$. The rate separates into $d_n\eta_n$, which measures the approximation error of the derivative estimator $\hat{\phi}'_n$, and the coupling error s_n. Theorem 3 has similar implications on the quantiles as Theorem 1. Let $\nu : \mathbb{E} \to \mathbb{R}$ be a Lipschitz continuous functional and let $\sigma^2_{n,\nu} = \mathrm{Var}(\nu(\phi'_{\theta_0}(Z_n)))$. If the bootstrap error is of smaller order than the $\sigma_{n,\nu}$, i.e., $(d_n\eta_n \vee s_n)/\sigma_{n,\nu} \to 0$, then

Theorem 3 implies for every $\varepsilon > 0$ and sufficiently large n

$$Q_{p-\varepsilon}(\nu(\hat{\phi}'_n(\hat{Z}_n))|\mathcal{D}_n) - \varepsilon\sigma_n \leq Q_p(\nu(\phi(\hat{\theta}_n) - \phi(\theta_0))) \leq Q_{p+\varepsilon}(\nu(\hat{\phi}'_n(\hat{Z}_n))|\mathcal{D}_n) + \varepsilon\sigma_n$$

with probability converging to one and where $Q_p(Z|\mathcal{D}_n) = \inf\{t \in \mathbb{R} : \mathrm{P}(Z \leq t|\mathcal{D}_n) \geq p\}$.[17] This bound can be used to justify confidence intervals when a consistent estimator of the standard deviation is available. For instance, suppose we have some $\hat{\sigma}_n$ satisfying

$$\frac{\hat{\sigma}_n - \sigma_n}{\sigma_n} \xrightarrow{p} 0.$$

Then, the above bounds imply for any $\varepsilon > 0$ and $\alpha \in (0, 1/2)$

$$\liminf_{n\to\infty} \mathrm{P}(\phi(\hat{\theta}_n) - \phi(\theta_0) \leq Q_{1-\alpha+\varepsilon}(\nu(\hat{\phi}'_n(\hat{Z}_n))|\mathcal{D}_n) + \varepsilon\hat{\sigma}_n) \geq 1 - \alpha.$$

As in the discussion of Theorem 1, we can improve Theorem 3 to a bound in Kolmogorov distance under further analytical assumptions on the derivative and Gaussianity of the penultimate process Z_n as we show in the following Theorem.

Theorem 4. *Suppose that the Assumptions of Theorem 2, Assumptions 3 and 4 hold. Then, if* $\{(d_n\eta_n) \vee s_n\} \operatorname{Var}(\phi'_{\theta_0}(Z_n))^{-1/2} \to 0$,

$$\sup_{t>0} \left|\mathrm{P}\left(\hat{\phi}'_n(\hat{Z}_n) \leq t \,\middle|\, \mathcal{D}_n\right) - \mathrm{P}\left(\phi(\hat{\theta}_n) - \phi(\theta_0) \leq t\right)\right| = o_p(1).$$

If further $\mathcal{F} = \{f \in \mathbb{D}^* \,|\, \forall h \in \mathbb{D} : f(h) \leq \phi'_{\theta_0}(h),\ \operatorname{Var}(f(Z_n)) = 0\} = \emptyset$, *then*

$$\sup_{t\in\mathbb{R}} \left|\mathrm{P}\left(\hat{\phi}'_n(\hat{Z}_n) \leq t \,\middle|\, \mathcal{D}_n\right) - \mathrm{P}\left(\phi(\hat{\theta}_n) - \phi(\theta_0) \leq t\right)\right| = o_p(1).$$

Theorem 4 can be used to prove pointwise validity of confidence intervals. We illustrate this here for confidence intervals of the following kind: For some $\alpha \in (0, 1)$, let $\hat{q}_n(1-\alpha)$ denote the $(1-\alpha)$-quantile of the bootstrapped distribution $\hat{\phi}'_n(\hat{Z}_n)$ conditionally on the data, i.e.,

$$\hat{q}_n(1-\alpha) = \inf\{q : \mathrm{P}(\hat{\phi}'_n(\hat{Z}_n) \leq q \,|\, \mathcal{D}_n) \geq 1 - \alpha\} \tag{1.7}$$

[17] See Lemma 11 in Chernozhukov et al. (2013b).

and define the confidence interval as

$$\widehat{CI}_n = \{\varphi \in \mathbb{E} : \phi(\hat{\theta}_n) - \varphi \leq \hat{q}_n(1-\alpha)\}. \tag{1.8}$$

Then, Theorem 4 readily implies pointwise validity as we show in the following Corollary.

Corollary 1. *Suppose that the Assumptions in Theorem 4 hold. Further, suppose that for $\alpha \in (0,1)$, $\hat{q}_n(1-\alpha) > \tau$ with probability approaching one, where $\tau = 0$ if $\mathcal{F} \neq \emptyset$ and $\tau = -\infty$ otherwise. Then,*

$$\liminf_{n\to\infty} \mathrm{P}\big(\phi(\theta_0) \in \widehat{CI}_n\big) \geq 1-\alpha.$$

The assumption that the bootstrapped quantiles $\hat{q}_n(1-\alpha)$ are larger than τ might be hard to show in applications. Therefore, we present simple sufficient conditions in Lemma 9 in the Appendix involving analytical properties of the derivative as well as properties of the approximating Gaussian process Z_n.

1.3.1 Examples revisited

Example 7 (Maximum of conditional mean continued:). We use the following multiplier bootstrap in order to estimate the quantiles of the local polynomial estimator. Let η_i, $i = 1, \ldots, n$, be i.i.d. standard normally distributed random variables which are independent of the data $\mathcal{D}_n = \{(Y_i, X_i) : i = 1, \ldots, n\}$ and set

$$\hat{Z}_n(x) = \frac{1}{nh_n\hat{f}_n(x)} \sum_{i=1}^{n} \eta_i U(0)^\intercal S^{-1} K_h(X_i - x) U\left(\frac{X_i - x}{h_n}\right)\hat{\varepsilon}_i,$$

where $\hat{f}_n(x)$ denotes a kernel density estimator of $f(x)$ and $\hat{\varepsilon}_i = Y_i - \hat{\theta}_n(x)$. Conditionally on the data, $\hat{Z}_n = \{\hat{Z}_n(x) : x \in [0,1]\}$ is a centered Gaussian process whose covariance function (conditionally on the data) is a consistent estimator of the covariance function of Z_n. Therefore, this multiplier bootstrap can also be thought of as a Monte Carlo simulation of the approximate limiting distribution based on a consistent estimator of the covariance function. In the appendix, we show that this bootstrap satisfies Assumption 3 with $d_n = \sqrt{\log n/(nh_n^3)}$ and $s_n = o((nh_n \log n)^{-1/2})$.

As an estimator for the derivative, we use

$$\hat{\phi}'_n(h) = \max_{x \in \hat{\Psi}_n} h(x),$$

where

$$\hat{\Psi}_n = \{x : \hat{\theta}_n(x) + \hat{q}_{\gamma_n}\hat{\sigma}_n(x) \geq \max_{\tilde{x}} \hat{\theta}_n(\tilde{x}) - \hat{q}_{\gamma_n}\hat{\sigma}_n(\tilde{x})\}.$$

Here, $\hat{\sigma}_n$ denotes an estimator of the standard error of $\hat{\theta}_n$ and $\hat{q}_{\gamma_n}$ is an estimator of the γ_n-quantile of $\|(\hat{\theta}_n - \theta_0)/\sigma_n\|_\infty$. $\gamma_n \to 1$ slowly enough so that $\hat{q}_{\gamma_n}\|\hat{\sigma}_n\|_\infty \to 0$.

$\hat{\Psi}_n$ is an estimator of the argmax set of θ_0. Graphically, it is constructed as follows: $\hat{q}_{\gamma_n}$ is chosen such that $\hat{\theta}_n \pm \hat{q}_{\gamma_n}\hat{\sigma}_n(x)$ is a γ_n uniform confidence band for θ_0. $\hat{\Psi}_n$ now consists of all x for which the upper band lies above the maximum of the lower band. Thus, it consists of all potential maximizers which are justified by the uniform confidence band and in particular covers the true set of maximizers whenever the confidence band covers the true conditional mean function.

We show in the appendix, that the rate of this estimator can be bounded by

$$\sup_{\|h\|_{BL} \leq 1} |\hat{\phi}'_n(h) - \phi'_{\theta_0}(h)| = O_p\left(\frac{(\hat{q}_{\gamma_n}\|\hat{\sigma}_n\|_\infty)^{1/\rho}}{c}\right).$$

This rate depends on the well-separatedness condition (1.4) on the maximum of the conditional mean function. Heuristically, when the conditional mean function is rather flat around its maximum as measured by the exponent ρ, it is harder to estimate the set of maximizers therefore reducing the rate of convergence of the derivative estimator.

Hence, this bootstrap of the plug-in estimator satisfies our general bootstrap assumptions and Theorem 4 applies under the rate requirement $\log^3 n/(nh_n^5) \to 0$ for $\rho = 2$. This is the same requirement as needed for the Delta method in Theorem 2 to hold.

1.3.2 Validity of general-purpose bootstrap procedures

The literature on inference on directionally differentiable functionals has proposed several generally applicable bootstrap procedures including the standard bootstrap, subsampling and the numerical Delta method. These procedures have been

analyzed in the setting of the Delta method by Shapiro (1991) and Dümbgen (1993), and it is not obvious that they yield valid inference in our setup. We therefore discuss in this section under which conditions these approaches yield consistent estimators in the context of our proposed Delta method.

Standard / Naive Bootstrap: First, we consider what has been called the standard bootstrap by Fang and Santos (2018) and the naive bootstrap by Dümbgen (1993) and Hong and Li (2018).[18] It goes back to proposals by Krinsky and Robb (1986), Krinsky and Robb (1990) and Runkle (1987). The idea is to draw bootstrapped samples $\hat{\theta}_n^*$ of the preliminary estimator $\hat{\theta}_n$ and use the distribution of $\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n)$ conditionally on the data as an estimator of the distribution of the plug-in estimator. This approach can be slightly more generally motivated as follows. The plug-in estimator can be written as

$$\phi(\hat{\theta}_n) - \phi(\theta_0) = \phi(\theta_0 + \{\hat{\theta}_n - \theta_0\}) - \phi(\theta_0).$$

The standard bootstrap now replaces the unknown θ_0 by its estimator $\hat{\theta}_n$ and $(\hat{\theta}_n - \theta_0)$ by a bootstrap $\hat{Z}_n$ for its approximate distribution Z_n. Then, the quantiles of the plug-in estimator are estimated by the quantiles of $\phi(\hat{\theta}_n + \hat{Z}_n) - \phi(\hat{\theta}_n)$ conditionally on the sample $\{X_i\}_{i=1}^n$.

The following Lemma gives an asymptotic approximation of the distribution of the standard bootstrap.

Lemma 1. *Suppose that Assumption 3 and the Assumptions of Theorem 1 hold with Assumption 3(iii) replaced with*

$$\lim_{n\to\infty} \mathrm{P}\big(\mathrm{P}(\|\phi'_{\theta_0}((\hat{\theta}_n - \theta_0) + \hat{Z}_n) - \phi'_{\theta_0}(Z_n + Z_n^*)\|_{\mathbb{E}} > M(s_n \vee b_n) \,|\, \mathcal{D}_n) > \varepsilon\big) = 0.$$

Then, for any $\varepsilon > 0$, there exist M such that

$$\lim_{n\to\infty} \mathrm{P}\big(\mathrm{P}\big(\|(\phi(\hat{\theta}_n + \hat{Z}_n) - \phi(\hat{\theta}_n)) - (\phi'_{\theta_0}(Z_n + Z_n^*) - \phi'_{\theta_0}(Z_n))\|_{\mathbb{E}} > M(b_n \vee s_n \vee a_n^\gamma \vee d_n^\gamma) \,\big|\, \mathcal{D}_n\big) > \varepsilon\big) = 0.$$

Heuristically, this result shows that the standard bootstrap does not approximate

[18] It is also called the parametric or asymptotic distribution bootstrap. See Woutersen and Ham (2014) for further details on this procedure.

$\phi'_{\theta_0}(Z_n)$ but instead

$$\phi(\hat{\theta}_n + \hat{Z}_n) - \phi(\hat{\theta}_n) \approx \phi'_{\theta_0}(Z_n + Z_n^*) - \phi'_{\theta_0}(Z_n).$$

Thus, if the derivative ϕ'_{θ_0} is linear in h, the standard bootstrap yields a consistent estimator. On the other hand, when the derivative is non-linear, the standard bootstrap might fail to be consistent. This observation is in line with the results in Fang and Santos (2018). They show in the context of the Delta method by Shapiro (1991) and Dümbgen (1993) that, when the limiting distribution is Gaussian, the standard bootstrap is consistent if and only if the derivative is linear.

Subsampling: Two common approaches when dealing with non-standard bootstrap settings are the m-out of n bootstrap and subsampling (in the context of inference on directionally differentiable functionals, see e.g. Romano and Shaikh (2008), Romano and Shaikh (2010), Andrews and Guggenberger (2009b) and Andrews and Guggenberger (2010)). Even in our setting, these procedures are consistent if $\phi'_{\theta_0}(Z_n)$ converges in distribution. Since the analysis of subsampling and the m out of n bootstrap is similar, we focus on subsampling in the following.

The subsampling approach is based on simulation of

$$\frac{1}{a_m}\big(\phi(\hat{\theta}_n + a_m \hat{Z}_m^*) - \phi(\hat{\theta}_n)\big)$$

where $\hat{Z}_m$ is based on a subsample of size $m \ll n$ and a_n was given in Assumption 2. This can be interpreted as using implicitly the derivative estimator

$$\hat{\phi}'_n(h) = \frac{1}{a_m}\big(\phi(\hat{\theta}_n + a_m h) - \phi(\hat{\theta}_n)\big). \tag{1.9}$$

Under the further assumption that $h \mapsto \phi'_{\theta_0}(h)$ is Lipschitz continuous, we show in the Appendix that this estimator satisfies

$$\sup_{\|h\| \leq 1} \|\hat{\phi}'_n(h) - \phi'_{\theta_0}(h)\|_{\mathbb{E}} = O_p\left(\frac{a_n}{a_m} + a_m^{\gamma - 1}\right).$$

Therefore, subsampling fits into our general framework given above and implies:

Lemma 2. *Suppose that Assumption 3 and the assumptions of Theorem 2 hold. If*

m is chosen so that

$$\frac{\{d_m(a_n/a_m + a_m^{\gamma-1})\} \vee s_m}{\sigma_m} \to 0,$$

where $\sigma_n^2 = \mathrm{Var}(\phi'_{\theta_0}(Z_n))$, then

$$\sup_{t>\tau} \left| \mathrm{P}\big(\hat{\phi}'_n(\hat{Z}_m) \leq t \,\big|\, \mathcal{D}_n\big) - \mathrm{P}\big(\phi(\hat{\theta}_m) - \phi(\theta_0) \leq t\big) \right| = o_p(1).$$

Here $\tau = 0$ if $\mathcal{F} \neq \emptyset$ and $\tau = -\infty$ else. Moreover, if $r_n \phi'_{\theta_0}(Z_n) \xrightarrow{d} G$ for some sequence r_n and continuous limiting distribution G, then

$$\sup_{t\in\mathbb{R}} \left| \mathrm{P}\big(r_m \hat{\phi}'_n(\hat{Z}_m) \leq t \,\big|\, \mathcal{D}_n\big) - \mathrm{P}\big(r_n\{\phi(\hat{\theta}_n) - \phi(\theta_0)\} \leq t\big) \right| = o_p(1).$$

This result suggests that the subsampling estimator is a consistent estimator of the distribution of $\phi'_{\theta_0}(Z_m^*)$ at sample size $m \ll n$. This might be problematic as the covariance functions of Z_n and Z_m might differ considerably. In particular, when Z_n is the approximating distribution sequence of a local polynomial estimator as discussed in Example 4, the covariance function changes considerably with the bandwidth choice. Since $h_n \ll h_m$, Z_m might fail to approximate Z_n well and therefore also $\phi'_{\theta_0}(Z_m)$ might be a poor approximation of $\phi'_{\theta_0}(Z_n)$. On the other hand, when there is a rate r_n such that $r_n \phi'_{\theta_0}(Z_n)$ has a limiting distribution, a simple rescaling ensures that subsampling yields valid inference.

Rescaled bootstrap / numerical Delta method: Another idea would be to use the same numerical derivative estimator as in subsampling but to use some general bootstrap $\hat{Z}_n$ of Z_n instead of the subsampling distribution based on $m \ll n$ observations. This approach has been proposed by Dümbgen (1993) as the rescaled bootstrap and by Hong and Li (2020) as the numerical Delta method. Formally, the rescaled bootstrap / numerical Delta method uses the distribution of

$$\frac{1}{a_m}\big(\phi(\hat{\theta}_n + a_m \hat{Z}_n) - \phi(\hat{\theta}_n)\big)$$

conditionally on the data as an estimator of the distribution of the plug-in estimator.

As this method uses the same estimator of the derivative as subsampling, the rescaled bootstrap / numerical Delta method also can be analyzed using Theorem 3 and 4 above. In particular, when the derivative is a continuous and sublinear

functional, we show in the Appendix:

Lemma 3. *Suppose that the assumptions of Theorem 2 hold. If m is chosen so that*

$$\frac{\{d_n(a_n/a_m + a_m^{\gamma-1})\} \vee s_n}{\sigma_n} \to 0,$$

where $\sigma_n^2 = \mathrm{Var}(\phi'_{\theta_0}(Z_n))$, then, for $\hat{\phi}'_n$ given in (1.9),

$$\sup_{t>\tau} \left| \mathrm{P}\big(\hat{\phi}'_n(\hat{Z}_n) \leq t \,\big|\, \mathcal{D}_n\big) - \mathrm{P}\big(\phi(\hat{\theta}_n) - \phi(\theta_0) \leq t\big) \right| = o_p(1).$$

Here $\tau = 0$ if $\mathcal{F} \neq \emptyset$ and $\tau = -\infty$ else.

1.4 Local analysis

The non-smoothness of only directionally differentiable transformations leads to challenges for estimation and inference as has been shown in Dümbgen (1993), Hirano and Porter (2012) and Fang and Santos (2018). In particular, Hirano and Porter (2012) have shown in a local asymptotic framework that when ϕ is only directionally differentiable, there do not exist asymptotically quantile-unbiased estimators, severely limiting the ability to perform asymptotically valid inference. We therefore study in this section the robustness of our results to local perturbations of the data generating process. We model this by considering parameter sequences θ_n which converge to θ_0. More generally, we allow the data generating process to change with the sample size and denote this by using P_n instead of a fixed probability measure P. We write o_{p_n} and O_{p_n} in order to denote convergence in probability and stochastic boundedness with respect to the sequence P_n.

Assumption 5. *(i) The estimator $\hat{\theta}_n$ is a measurable function mapping a sequence of random variables $\{X_i\}_{i=1}^n$ into $\mathbb{D}_\phi$.*

(ii) There exists a sequence $a_n \to 0$ such that $\theta_n = \theta_0 + h_n \in \mathbb{D}_\theta \subseteq \mathbb{D}_\phi$ with $\|h_n\| = O(a_n)$.

(iii) $\|\hat{\theta}_n - \theta_n\| = O_{p_n}(a_n)$.

(iv) There exists a sequence of real numbers $b_n \to 0$ and a sequence of $\mathbb{D}$-valued

random vectors $\{Z_{n,h_n}\}_{i=1}^n$ *such that* $Z_{n,h_n} \sim Z_n$, $n \in \mathbb{N}$, *and*

$$\|\phi'_{\theta_0}(\{\hat{\theta}_n - \theta_n\} + h_n) - \phi'_{\theta_0}(Z_{n,h_n} + h_n)\|_{\mathbb{E}} = O_{p_n}(b_n).$$

Assumption 5 formalizes our local asymptotic framework and requires the preliminary estimator $\hat{\theta}_n$ to be insensitive to the local perturbations of the data generating process. We allow the local parameters θ_n to lie in a subset of $\mathbb{D}_\phi$ in order to allow for further restrictions on the parameter such as smoothness restrictions. This might be needed for nonparametric preliminary estimators in order to control their bias. Assumptions (iii) and (iv) can be interpreted as locally uniform versions of the corresponding assumptions in Assumption 2. For instance, (iii) requires that the rate a_n is unaffected by local deviations. In (iv), we further require that the approximate distribution Z_{n,h_n} does not change with h_n. This can be interpreted as a notion of asymptotic equivariance in law which is a common starting point for a local asymptotic analysis. In particular, it is satisfied for regular estimators $\hat{\theta}_n$. The required locally robust coupling is for example satisfied when the derivative is Lipschitz continuous and there is a coupling of $\hat{\theta}_n - \theta_0$ to Z_{n,h_n}. Such local asymptotic couplings can usually be derived along the same lines as the pointwise results since the used strong approximation results are finite sample bounds.

Our two main results of this section can be seen as extensions of local results in Fang and Santos (2018) and Dümbgen (1993) to our setting:

Theorem 5. *Suppose Assumption 1 and 5 hold. Then*

$$\phi(\hat{\theta}_n) - \phi(\theta_n) = \phi'_{\theta_0}(Z_{n,h_n} + h_n) - \phi'_{\theta_0}(h_n) + O_{p_n}\big(a_n^\gamma + b_n\big).$$

Theorem 6. *Suppose that Assumptions 1 and 5 hold with* $\mathbb{E} = \mathbb{R}$. *Suppose that* ϕ'_{θ_0} *is sublinear and continuous and that* Z_n *is a sequence of centered and tight Gaussian random vectors. If*

$$\frac{a_n^\gamma + b_n}{\sigma_{n,h}} \to 0,$$

where $\sigma_{n,h}^2 = \mathrm{Var}(\sup_{f \in \mathcal{F}_+} f(Z_{n,h_n} + h_n))$ *with* $\mathcal{F}_+ = \{f \in \mathbb{D}^* \,|\, \forall h \in \mathbb{D} : f(h) \leq \phi'_{\theta_0}(h)\} \setminus \mathcal{F}$, *then*

$$\lim_{n\to\infty} \sup_{t > \tau_n} \big| \mathrm{P}_n(\phi(\hat{\theta}_n) - \phi(\theta_n) \leq t) - \mathrm{P}_n\big(\phi'_{\theta_0}(Z_{n,h_n} + h_n) - \phi'_{\theta_0}(h_n) \leq t\big) \big| = 0,$$

where $\tau_n = \sup_{f \in \mathcal{F}} f(h_n)$.

If instead $\mathcal{F} = \{f \in \mathbb{D}^* \,|\, \forall h \in \mathbb{D} : f(h) \leq \phi'_{\theta_0}(h),\ \mathrm{Var}(f(Z_n)) = 0\} = \emptyset$ *and*

$$\frac{a_n^\gamma + b_n}{\sigma_{n,h}} \to 0,$$

where $\sigma_{n,h}^2 = \mathrm{Var}(\phi'_{\theta_0}(Z_{n,h_n} + h_n))$, *then*

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}} \big|\mathrm{P}_n(\phi(\hat{\theta}_n) - \phi(\theta_n) \leq t) - \mathrm{P}_n\big(\phi'_{\theta_0}(Z_n + h_n) - \phi'_{\theta_0}(h_n) \leq t\big)\big| = 0.$$

The result in the above Theorems indicate how the finite sample distribution of the plug-in estimator changes when the true parameter is close to a parameter θ_0 where ϕ fails to be fully differentiable. In this case, the approximate distribution depends on the perturbation h_n and the pointwise asymptotic approximation in Theorems 1 and 2 might be misleading. Importantly, Theorems 5 and 6 allow us to study when our proposed procedures deliver reliable size control.

For example, in the setting of Theorem 6 and one-sided confidence intervals as constructed in (1.8), Theorem 6 implies

$$\mathrm{P}(\phi(\hat{\theta}_n) - \phi(\theta_n) \leq \hat{q}_n(1-\alpha)) \geq \mathrm{P}\big(\phi'_{\theta_0}(Z_n + h_n) - \phi'_{\theta_0}(h_n) \leq \hat{q}_n(1-\alpha)\big) - o(1).$$

In order to show that this confidence interval provides local size control, it is sufficient to show that the probability on the right-hand side is larger than $1-\alpha$. While this might be hard to show in general, sublinearity of $h \mapsto \phi'_{\theta_0}(h)$ implies

$$\phi'_{\theta_0}(Z_n + h_n) - \phi'_{\theta_0}(h_n) \leq \phi'_{\theta_0}(Z_n).$$

Thus, the quantiles of $\phi'_{\theta_0}(Z_n)$ provide an upper bound on the quantiles of the local asymptotic approximate distribution and therefore the confidence set achieves local size control. We summarize this finding in the following Corollary.

Corollary 2. *Suppose the conditions of Theorem 6 and Assumption 6 below hold. Then, if* $\mathcal{F} = \emptyset$, *it holds for any* $\alpha \in (0,1)$ *that*

$$\liminf_{n\to\infty} \mathrm{P}_n(\phi(\hat{\theta}_n) - \phi(\theta_n) \leq \hat{q}_n(1-\alpha)) \geq 1-\alpha,$$

where $\hat{q}_n(1-\alpha) = \inf\{q : \mathrm{P}(\hat{\phi}'_n(\hat{Z}_n) \leq q \,|\, \mathcal{D}_n) \geq 1-\alpha\}$.

The proof of Corollary 2 requires validity of the bootstrapped quantiles also in this local asymptotic framework. We therefore impose the following adaption of Assumptions 3 and 4.

Assumption 6. *(i)* $\hat{Z}_n : \{X_i, W_i\}_{i=1}^n \to \mathbb{D}$ *measurable with* $\{W_i\}_{i=1}^n$ *independent of* $\mathcal{D}_n = \{X_i\}_{i=1}^n$.

(ii) There is some sequence $d_n \to 0$ *such that* $\|\hat{Z}_n\| = O_{p_n}(d_n)$.

(iii) There is some sequence $s_n \to 0$ *and some* $Z^*_{n,h_n}|\mathcal{D}_n \sim Z_n$ *such that for any* $\varepsilon > 0$*, there exist* M *such that*

$$\lim_{n\to\infty} \mathrm{P}_n\big(\mathrm{P}_n(\|\phi'_{\theta_0}(\hat{Z}_n) - \phi'_{\theta_0}(Z^*_{n,h_n})\|_{\mathbb{E}} > Ms_n \,|\, \mathcal{D}_n) > \varepsilon\big) = 0.$$

(iv) The map $\hat{\phi}'_n : \mathbb{D} \to \mathbb{R}$ *is a measurable function of* $\{X_i\}_{i=1}^n$*, which is positively homogeneous and satisfies for a sequence* $\eta_n \to 0$ *that for any* $\varepsilon > 0$*, there is a* K *such that*

$$\limsup_{n\to\infty} \mathrm{P}_n\left(\sup_{\|h\|\leq 1} \phi'_{\theta_0}(h) - \hat{\phi}'_n(h) > K\eta_n\right) < \varepsilon.$$

Assumptions (i)-(iii) are analogs to the corresponding assumptions in Assumption 3 and require the bootstrap $\hat{Z}_n$ to be insensitive to local perturbations. Assumption (iv), on the other hand, allows the estimator of the derivative to be biased by local perturbations but restricts the direction of this bias. Heuristically, it imposes that $\hat{\phi}'_n$ consistently estimates an upper bound of the derivative at θ_0 and is weaker than assuming that $\hat{\phi}'_n$ consistently estimates ϕ'_{θ_0}. We allow for such a bias since $\theta \mapsto \phi'_\theta$ easily fails to be continuous when ϕ is only directionally differentiable and therefore locally robust estimators of the derivative might be hard to construct. On the other hand, (iv) is for example satisfied by the implicit derivative estimator of the subsampling estimator.

The result of Corollary 2 can be interpreted as a local asymptotic half-quantile-unbiased-ness property of the bootstrapped quantiles. While this property implies local size control, it does not rule out that the resulting confidence sets are conservative. This potential for conservativeness can be attributed to two factors. On the one hand, we bounded the quantiles of the local approximate distribution using the sublinearity of ϕ'_θ and secondly Assumption 6 (iv) allows the bootstrapped quantiles to be themselves an upward biased estimator of the quantiles of $\phi'_{\theta_0}(Z_n)$.

The existence of locally asymptotically half-quantile-unbiased estimators is also consistent with the impossibility results in Hirano and Porter (2012). While the authors show that there cannot exist locally asymptotically quantile-unbiased estimators, this does not rule out the existence of locally asymptotically half-quantile-unbiased estimators. This has also already been observed in the literature on moment inequalities as e.g. in Chernozhukov et al. (2013b) and Andrews and Shi (2013). There, the half-quantile-unbiasedness property is also used to construct estimators which are asymptotically half-median-unbiased which translates in our setting to the property that

$$\liminf_{n\to\infty} \mathrm{P}_n(\phi(\hat{\theta}_n) - \hat{q}_n(1/2) \leq \phi(\theta_n)) \geq \frac{1}{2}$$

and in this sense $\phi(\hat{\theta}_n) - \hat{q}_n(1/2)$ is a locally asymptotically half-median-unbiased estimator of $\phi(\theta_0)$. Further, the observation that sublinearity is sufficient for local size control when the derivative is only directionally differentiable has already been shown by Fang and Santos (2018). In this sense, Corollary 2 can be interpreted as an application of their idea to our setting.

While sublinearity of the derivative holds in all of our examples, it can fail to hold. In this scenario, one-sided confidence intervals as constructed in (1.8) may fail to provide local size control and need further investigation. On the other hand, Theorems 5 and 6 imply that an upper bound on the quantiles of $\phi'_{\theta_0}(Z_n + h_n) - \phi'_{\theta_0}(h_n)$ is sufficient for local size control. Sometimes such an upper bound can be motivated by analyzing the analytical expression of the derivative as e.g. in Example 2.2 in Fang and Santos (2018). Alternatively, motivated by the size-corrected fixed-critical-values in Andrews and Guggenberger (2009a), one may compute the worst case bound

$$\phi^*_{\theta_0}(Z_n) := \sup_{h\in\mathbb{D}_\theta} \phi'_{\theta_0}(Z_n + h) - \phi'_{\theta_0}(h).$$

The function $\phi^*_{\theta_0}$ is a sublinear majorant of the derivative ϕ'_{θ_0} and reduces to ϕ'_{θ_0} when the derivative is sublinear. By definition of ϕ^*_θ, the quantiles of $\phi^*_{\theta_0}(Z_n)$ imply an upper bound on the quantiles of the local asymptotic approximate distribution. Thus, an alternative strategy to construct one-sided confidence intervals may be based on an estimator $\hat{\phi}^*_n$ of this upper bound $\phi^*_{\theta_0}$ instead of the derivative ϕ'_{θ_0}. In particular, if ϕ is a continuous functional and Z_n a sequence of centered and tight

Gaussians, we can obtain a similar result for the bootstrapped quantiles of $\phi^*_{\theta_0}(Z_n)$ as in Corollary 2.

Corollary 3. *Suppose the conditions of Theorem 5 and Assumption 6 hold when (iii) and (iv) in Assumption 6 are replaced by*

(iii') *There is some sequence $s_n \to 0$ and some $Z^*_{n,h_n}|\mathcal{D}_n \sim Z_n$ such that for any $\varepsilon > 0$, there exist M such that*

$$\lim_{n\to\infty} \mathrm{P}_n\big(\mathrm{P}_n(\|\phi^*_{\theta_0}(\hat{Z}_n) - \phi^*_{\theta_0}(Z^*_{n,h_n})\|_{\mathbb{E}} > Ms_n \,|\, \mathcal{D}_n) > \varepsilon\big) = 0.$$

(iv') *The map $\hat{\phi}^*_n : \mathbb{D} \to \mathbb{R}$ is a measurable function of $\{X_i\}_{i=1}^n$, which is positively homogeneous and satisfies for a sequence $\eta_n \to 0$ that for any $\varepsilon > 0$, there is a K such that*

$$\limsup_{n\to\infty} \mathrm{P}_n\left(\sup_{\|h\|\leq 1} \phi^*_{\theta_0}(h) - \hat{\phi}^*_n(h) > K\eta_n\right) < \varepsilon.$$

Then, if

$$\frac{a_n^\gamma + b_n}{\sigma_{n,*}} \to 0,$$

where $\sigma^2_{n,*} = \mathrm{Var}(\phi^*_{\theta_0}(Z_n))$, and $\mathcal{F} = \{f \in \mathbb{D}^* \,|\, \forall h \in \mathbb{D} : f(h) \leq \phi^*_{\theta_0}(h),\ \mathrm{Var}(f(Z_n)) = 0\} = \emptyset$, it holds for any $\alpha \in (0,1)$ that

$$\liminf_{n\to\infty} \mathrm{P}_n(\phi(\hat{\theta}_n) - \phi(\theta_n) \leq q^*_n(1-\alpha)) \geq 1 - \alpha,$$

where $q^*_n(1-\alpha) = \inf\{q : \mathrm{P}(\hat{\phi}^*_n(\hat{Z}_n) \leq q \,|\, \mathcal{D}_n) \geq 1-\alpha\}$.

1.5 Numerical Results

In this section, we present results of a Monte Carlo study to illustrate the finite-sample relevance of our theoretical results. We focus here on the setting in Example 1, i.e. the maximum of a conditional mean function, and present further simulations in the setting of Example 2 in the Appendix. We employ the same Monte Carlo designs as in Chernozhukov et al. (2013b). This allows us to compare our simulation results to the results in the aforementioned paper. We compare 4 different methods

which can be briefly described as follows: the first method follows the construction of a bootstrap method proposed in Section 1.3 and was already outlined in Example 7. The second method utilizes the standard bootstrap as discussed in Section 1.3.2. The third method is a one-sided projection interval and the final method is the method proposed in Chernozhukov et al. (2013b). All methods are compared in terms of coverage probability of the true underlying parameter and a close false parameter.

1.5.1 Data Generating Processes

We consider random samples from the model

$$X_i \sim \text{Unif}[-2,2], \quad \varepsilon_i = \min\{\max\{-3, \sigma\tilde{\varepsilon}_i\}, 3\}, \quad \text{and}$$
$$Y_i = L\varphi(X_i) + \varepsilon_i,$$

where $\varphi(\cdot)$ is the standard normal density function, $\tilde{\varepsilon}_i \sim \mathcal{N}(0,1)$, and L and σ are constants. We consider four data-generating processes (DGP) taken from Chernozhukov et al. (2013b):

$$\text{DGP1}: \quad L = 0 \text{ and } \sigma = 0.1; \quad \text{DGP2}: \quad L = 1 \text{ and } \sigma = 0.1;$$
$$\text{DGP3}: \quad L = 5 \text{ and } \sigma = 0.1; \quad \text{DGP4}: \quad L = 5 \text{ and } \sigma = 0.01.$$

The first design DGP1 is completely flat, and therefore the functional is only directionally differentiable. The second design DGP2 is nonflat and smooth around the maximum with a unique maximizer at zero. In this DGP the functional is fully differentiable. DGP3 and DGP4 are also nonflat and similar to DGP2 with a peaked conditional mean function with a unique maximizer. DGP3 and DGP4 only differ in the conditional variance of Y_i which is a hundred times smaller for DGP4. In our simulations, we considered sample sizes $n = 500$ and $n = 1000$ for all of the above DGPs.

1.5.2 Inference Procedures

Our implemented inference procedures use the local linear estimator as described in Example 4 as an estimator for the conditional mean function. We follow the tuning parameter choice in Chernozhukov et al. (2013b) and use the quartic kernel

$K(x) = \frac{15}{16}(1-x^2)^2\mathbb{1}(|x| \leq 1)$ as a kernel function and the rule-of-thumb bandwidth

$$h = \hat{h}_{\text{ROT}} \times \hat{s}_x \times n^{1/5} \times n^{-2/7},$$

where $\hat{s}_x$ is the sample standard deviation of the X_i and $\hat{h}_{\text{ROT}}$ is rule of thumb bandwidth as described in Section 4.2 of Fan and Gijbels (1996).[19] The factor $n^{1/5} \times n^{-2/7}$ is multiplied to ensure that the smoothing bias is asymptotically negligible. For the evaluation of the estimates, we consider only x within an interval between the 0.05 and 0.95 sample quantiles of X_i's to avoid influence of outlier at the boundary of the support of X_i.

We implement three procedures for inference. The first method follows the construction of a bootstrap method proposed in Section 1.3 and will be called the derivative estimation based procedure. We implemented this method following the construction in Example 7 and chose $\gamma_n = 1 - 0.1/\log n$ for the tuning parameter of the argmax set estimator $\hat{\Psi}_n$. The second method, which we call the standard bootstrap, uses the same bootstrap for the local linear estimator as described in Example 7 and then applies the standard bootstrap as presented in Section 1.3.2. Finally, the third method is a one-sided projection interval. It employs the bootstrap from Example 7 to simulate the 95% quantile $\hat{q}_n$ of

$$\sup_x \frac{\hat{Z}_n(x)}{\hat{\sigma}_n(x)},$$

where $\hat{\sigma_n}(\cdot)$ denotes the same estimator of the standard error of $\hat{\theta}_n$ as used in the construction of the argmax set estimator. The resulting set

$$\hat{CB}_n = \{\theta(x) \,|\, \forall x : \theta(x) \geq \hat{\theta}_n(x) - \hat{q}_n\hat{\sigma}_n(x)\}$$

is a uniform one-sided confidence band for θ_0. The third method consists of using the following projection interval

$$\hat{C}_n = \{\phi(\theta) \,|\, \theta \in \hat{CB}_n\} = \big\{\phi(\theta) \,\big|\, \phi(\theta) \geq \max_{\tilde{x}}\{\hat{\theta}_n(\tilde{x}) - \hat{q}_n\hat{\sigma}_n(\tilde{x})\}\big\}.$$

Note that this interval coincides with the conservative nonstochastic choice $\hat{V} = \mathcal{V}$ described in section 7.1. in Chernozhukov et al. (2013b). Further, we implement

[19] For further details, see section 7.3 in Chernozhukov et al. (2013b).

the bootstrap in all of the above methods using 1000 draws.

1.5.3 Simulation Results

We evaluate our simulation results through the coverage probability (CP) at the true maximum $\phi(\theta_0)$ with nominal level 95% and the false coverage probability (FCP) evaluated at $\phi(\theta_0) - 0.02$. We perform 1000 independent replications for each experiment. The results are summarized in Table 1.1.

We first consider the performance of the derivative estimator based procedure. It performs reasonably well for DGP1 and undercovers only slightly. The argmax set estimator covers approximately the full support of the X_is which is the true argmax set for this DGP. For the other DGPs, the derivative based method is quite conservative with CP ranging from 99% to 99.7%. This can be in part attributed to the argmax set estimator; it overestimates the size of the true argmax set and therefore results in a derivative estimator which is larger than the true derivative. This in turn shifts the quantiles upwards and leads to increased coverage probabilities. In DGP2 and DGP3, the derivative based method has non-trivial power as indicated by the FCP and which decreases as the sample size increases. In comparison to DGP2, DGP3 has a more peaked conditional mean function which results in a shorter estimated argmax set showing that the argmax set estimator adapts to the DGP. DGP3 and DGP4 only differ in the conditional variance of Y_i which is a hundred times smaller in DGP4. This results in more precise estimates of the conditional mean function, reduces the width of the estimated argmax set and overall improves power so that the FCP is zero under DGP4. Further, the derivative based method shows a similar performance as the proposed method in Chernozhukov et al. (2013b). Its CP are larger than the corresponding values for CLR across all DGPs resulting in a loss of power. On the other hand, for DGP1 the derivative based method undercovers less than CLR and is closer to the nominal level of 95%. Overall, neither approach dominates the other.

The performance of the standard bootstrap based procedure is summarized in Table 1.1. For DGP1, the standard bootstrap severely undercovers with a CP of 76.1% and 71.2%. For DGP2 and DGP3, the CP is close to the nominal level resulting in a higher power as compared to the other methods in Table 1.1. Finally, for DGP4, the standard bootstrap also overcovers but with a smaller CP as compared to the other methods. These observations are consistent with our theoretical results

DGP	Sample Size	Method	Cov. Prob.	False Cov. Prob.	Ave. Argmax Set Min.	Ave. Argmax Set Max.
1	500	Deriv	0.941	0.122	-1.771	1.773
1	500	CLR	0.923	0.064	-1.799	1.792
1	500	Project	0.923	0.064	-1.799	1.792
1	500	Standard	0.761	0.026	-	-
1	1000	Deriv	0.942	0.008	-1.777	1.777
1	1000	CLR	0.936	0.003	-1.801	1.796
1	1000	Project	0.936	0.003	-1.801	1.796
1	1000	Standard	0.712	0.000	-	-
2	500	Deriv	0.993	0.822	-0.702	0.701
2	500	CLR	0.989	0.808	-0.890	0.892
2	500	Project	0.995	0.871	-1.799	1.792
2	500	Standard	0.944	0.555	-	-
2	1000	Deriv	0.991	0.695	-0.611	0.610
2	1000	CLR	0.990	0.675	-0.776	0.776
2	1000	Project	0.996	0.779	-1.801	1.796
2	1000	Standard	0.944	0.372	-	-
3	500	Deriv	0.990	0.916	-0.352	0.352
3	500	CLR	0.986	0.876	-0.426	0.424
3	500	Project	0.995	0.943	-1.799	1.792
3	500	Standard	0.953	0.745	-	-
3	1000	Deriv	0.990	0.841	-0.306	0.305
3	1000	CLR	0.986	0.816	-0.380	0.377
3	1000	Project	0.992	0.907	-1.801	1.796
3	1000	Standard	0.943	0.584	-	-
4	500	Deriv	0.997	0.000	-0.136	0.136
4	500	CLR	0.981	0.000	-0.142	0.142
4	500	Project	0.991	0.000	-1.799	1.792
4	500	Standard	0.988	0.000	-	-
4	1000	Deriv	0.997	0.000	-0.108	0.108
4	1000	CLR	0.991	0.000	-0.127	0.127
4	1000	Project	0.997	0.000	-1.801	1.796
4	1000	Standard	0.986	0.000	-	-

Table 1.1: Results for the Monte Carlo experiment. The column "method" indicates the chosen method. "Deriv" denotes the derivative based method, "CLR" the proposed method in Chernozhukov et al. (2013b), "Standard" is based on the standard bootstrap and "Project" denotes the one-sided projection interval. The results for "CLR" and "Project" are taken from Table III in Chernozhukov et al. (2013b).

in Section 1.3.2 as the derivative is nonlinear under DGP1 and therefore the standard bootstrap is inconsistent, while the derivative is linear under the remaining DGPs.

The results for the projection interval are presented in Table 1.1. Under DGP1, it has a lower CP as compared to the derivative based method and yields the same coverage properties as method CLR. For DGPs 2 and 3, the projection interval has the highest CP and the lowest power among all the presented methods. DGP4 is an exception. Here, the projection interval is less conservative than the derivative based method for $n = 500$ and achieves the same coverage properties as the derivative based method for $n = 1000$. This suggests that the derivative based method is preferable to the projection interval, but no method strictly dominates the other.

1.6 Conclusion

In this paper, we study inference on parameters of the form $\phi(\theta_0)$, where ϕ is a known directionally differentiable transformation and θ_0 is an unknown parameter. We focus on settings, where θ_0 is an unknown function estimated using some nonparametric or machine-learning estimator $\hat{\theta}_n$. As many nonparametric or machine-learning estimators do not converge in distribution, existing extensions to the Delta method are not applicable in our setting. Similarly, subsampling cannot be easily applied in this setting as it is hard to show that the plug-in estimator converges in distribution without relying on convergence in distribution of the preliminary estimator. We propose to use strong approximations to the distribution of $\hat{\theta}_n$ as an alternative concept to convergence in distribution. Such strong approximations provide a sequence of approximate distributions instead of a single limiting distribution and are readily available for a wide-range of estimators. Further, we present a higher-order notion of Fréchet directional differentiability which is sufficiently flexible to handle the irregularity of nonparametric estimators and allows analyzing settings where the plug-in estimator converges at a faster rate than the preliminary estimator. These concepts enable us to derive a new Delta method which implies a strong approximation to the distribution of the plug-in estimator $\phi(\hat{\theta}_n)$. Further, we derive Berry-Esseen type Delta method under further analytical conditions on the transformation and the approximate distributions of the preliminary estimator.

Since the distributional approximations implied by our Delta method are rarely pivotal, we suggest a simulation-based estimator of the approximate distribution of the plug-in estimator and provide conditions for its consistency. Besides that,

we also study the validity of the standard bootstrap, subsampling and the rescaled bootstrap / numerical Delta method. We study local approximations to the distribution of the plug-in estimator and provide sufficient conditions on the directional derivative of ϕ implying local size control for one-sided confidence intervals. We illustrate the applicability of our results in the context of inference on the maximum of a conditional mean function. In the appendix, we further apply our results to an example taken from Freyberger and Larsen (2021) which illustrates that our results can be applied in more complicated examples. Finally, we illustrate the finite sample performance in these examples using Monte-Carlo simulations. In the context of the maximum of a conditional mean function, we find that our proposed method performs comparably to the proposal in Chernozhukov et al. (2013b) and that it improves upon projection bands. For the example from Freyberger and Larsen (2021), we find that the resulting bands are conservative but improve upon projection band.

Appendix

1.A Proofs of the main theorems

1.A.1 The Delta method

Proof of Theorem 1: By γ-Fréchet directional differentiability of ϕ at θ_0, there exist $\delta > 0$ and $C < \infty$ such that

$$\|\phi(\theta_0 + h) - \phi(\theta_0) - \phi'_{\theta_0}(h)\|_{\mathbb{E}} \leq C\|h\|^{\gamma} \tag{1.10}$$

for all h satisfying $\theta_0 + h \in \mathbb{D}_\phi$ and $\|h\| \leq \delta$. By Assumption 2(ii), for any $\varepsilon > 0$, there exists a K_ε and an N_1 such that the event $A_n = \{\|\hat{\theta}_n - \theta_0\| \leq a_n K_\varepsilon\}$ has probability $\mathrm{P}(A_n) > 1 - \varepsilon$ for all $n \geq N$. By Assumption 2(i), $\hat{\theta}_n \in \mathbb{D}_\phi$ and therefore $\phi(\hat{\theta}_n)$ is well-defined in this case. Further, this implies that there exists an $N_3 \geq N_1 \vee N_2$ such that on A_n, $h_n := \hat{\theta}_n - \theta_0$ satisfies $\theta_0 + h_n \in \mathbb{D}_\phi$ and $\|h_n\| \leq \delta$ for all $n \geq N_3$. Thus, by (1.10),

$$1 - \varepsilon < \mathrm{P}(A_n) \leq \mathrm{P}\big(\|\phi(\hat{\theta}_n) - \phi(\theta_0) - \phi'_{\theta_0}(h_n)\|_{\mathbb{E}} \leq C\|h_n\|_c^{\gamma}\big)$$

for all $n \geq N_3$ and therefore

$$\big\|\phi(\hat{\theta}_n) - \phi(\theta_0) - \phi'_{\theta_0}(\hat{\theta}_n - \theta_0)\big\|_{\mathbb{E}} = O_p\big(\|\hat{\theta}_n - \theta_0\|_c^{\gamma}\big) = O_p(a_n^{\gamma}).$$

The claim now follows by the assumed coupling in Assumption 2(iii). □

The proof of Theorem 2 relies on a slight modification of Le Cam's Lemma (cf. Le Cam (2012), p.402) which is proven in our companion paper Scherer (2024).

Lemma 4. *For X, Z arbitrary real-valued random variables, $\tau \in \mathbb{R}$ and $\lambda > 0$,*

$$\sup_{t>\tau} |\mathrm{P}(X \leq t) - \mathrm{P}(Z \leq t)| \leq \mathrm{P}(|X - Z| > \lambda) + \zeta_\lambda(X) \vee \zeta_\lambda(Z),$$

where $\zeta_\lambda(V) = \sup_{t>\tau} \mathrm{P}(t \leq V \leq t + \lambda)$ for any real-valued random variable $V \in \mathbb{R}$.

Proof of Theorem 2: Since ϕ'_{θ_0} is Lipschitz continuous and sublinear[20] and Z_n is a centered and tight $\mathbb{D}$-valued Gaussian random vector, we can apply the anti-concentration bound in Scherer (2024) which implies

$$\sup_{t>\tau} \mathrm{P}\big(t \leq \phi'_{\theta_0}(Z_n) \leq t + \varepsilon\big) \leq \frac{\varepsilon\sqrt{12}}{\sqrt{\sigma_n^2 + \varepsilon^2/12}}, \qquad \forall \varepsilon > 0$$

for $\tau = 0$ if $\mathcal{F} \neq \emptyset$ and $\tau = -\infty$ otherwise. By Le Cam's Lemma 4,

$$\begin{aligned} &\sup_{t>\tau} \big|\mathrm{P}(\phi(\hat{\theta}_n - \theta_0) \leq t) - \mathrm{P}\big(\phi'_{\theta_0}(Z_n) \leq t\big)\big| \\ \leq &\inf_{\eta>0} \mathrm{P}\big(\big|(\phi(\hat{\theta}_n) - \phi(\theta_0)) - \phi'_{\theta_0}(Z_n)\big| > \eta\big) + \frac{\eta\sqrt{12}}{\sigma_n}, \end{aligned} \tag{1.11}$$

where we used that $\varepsilon/\sqrt{\sigma_n^2 + \varepsilon^2/12} \leq \varepsilon/\sigma_n$. By Theorem 1, we have

$$\big|(\phi(\hat{\theta}_n) - \phi(\theta_0)) - \phi'_{\theta_0}(Z_n)\big| = O_p\big(a_n^\gamma + b_n\big).$$

Further, by assumption, $(a_n^\gamma + b_n)/\sigma_n \to 0$ as $n \to \infty$ and therefore, for any ε, there is some N such that for $\eta_n = \sigma_n\varepsilon/(8\sqrt{3})$

$$\mathrm{P}\big(\big|(\phi(\hat{\theta}_n) - \phi(\theta_0)) - \phi'_{\theta_0}(Z_n)\big| > \eta_n\big) \leq \frac{\varepsilon}{2}$$

for all $n \geq N$. This implies for (1.11)

$$\begin{aligned} &\sup_{t>\tau} \big|\mathrm{P}((\phi(\hat{\theta}_n) - \phi(\theta_0)) \leq t) - \mathrm{P}\big(\phi'_{\theta_0}(Z_n) \leq t\big)\big| \\ \leq &\mathrm{P}\big(\big|(\phi(\hat{\theta}_n) - \phi(\theta_0)) - \phi'_{\theta_0}(Z_n)\big| > \eta_n\big) + \frac{\eta_n\sqrt{12}}{\sigma_n} \leq \varepsilon, \end{aligned}$$

for all $n \geq N$. The claim follows. □

[20] Lipschitz continuity follows by continuity and sublinearity as mentioned in the discussion in Section 1.2.

Auxiliary results

Lemma 5. *Let the assumptions of Theorem 1 hold. Further, suppose that there exists some* $\delta > 0$ *such that* $\mathrm{E}[|\phi'_{\theta_0}(Z_n)|^{2+\delta}] < \infty$ *exists for all* n, $\sigma_n^2 = \mathrm{Var}(\phi'_{\theta_0}(Z_n)) > 0$ *and that*

$$\frac{E[|\phi'_{\theta_0}(Z_n) - \mu_n|^{2+\delta}]}{\sigma_n^{2+\delta}} \leq M, \qquad \forall n \in \mathbb{N}$$

where $\mu_n = \mathrm{E}[\phi'_{\theta_0}(Z_n)]$. *If* $a_n^\gamma + b_n = o(\sigma_n)$, *then for every subsequence there is a further subsequence* (n_k) *satisfying*

$$\frac{\phi(\hat{\theta}_{n_k}) - \phi(\theta_0) - \mu_{n_k}}{\sigma_{n_k}} \xrightarrow{d} G$$

for some non-degenerate limiting distribution G *which might depend on the subsubsequence.*

Proof. Consider the studentized approximate distribution implied by Theorem 1

$$G_n := \frac{\phi'_{\theta_0}(Z_n) - \mu_n}{\sigma_n}.$$

G_n is asymptotically tight by Chebyshev's inequality and hence by Prokhorov's Theorem[21] (G_n) is relatively compact in the topology of weak convergence of probability measures. By relative compactness, any subsequence has a further subsequence (n_k) such that $G_{n_k} \xrightarrow{d} G$ for some limiting distribution G which might depend on the chosen subsubsequence.

By Theorem 1 and since $a_n^\gamma + b_n = o(\sigma_n)$,

$$\frac{\phi(\hat{\theta}_n) - \phi(\theta_0) - \mu_n}{\sigma_n} = G_n + o_p(1)$$

and therefore also for every subsequence (n_k)

$$\frac{\phi(\hat{\theta}_{n_k}) - \phi(\theta_0) - \mu_{n_k}}{\sigma_{n_k}} = G_{n_k} + o_p(1).$$

[21] See e.g. Theorem 5.1 in Billingsley (1999).

Thus, for any subsequence satisfying $G_{n_k} \xrightarrow{d} G$, by Slutsky's theorem

$$\frac{\phi(\hat{\theta}_{n_k}) - \phi(\theta_0) - \mu_{n_k}}{\sigma_{n_k}} \xrightarrow{d} G.$$

It remains to show that the potential limiting distributions are non-degenerate. To this end, note that by Strassen's Theorem[22] there exist $Z_{n_k} \sim G$ such that $G_{n_k} = Z_{n_k} + o_p(1)$. Since (G_n) is uniformly square integrable, this implies $\mathrm{Var}(Z_{n_k}) \to \mathrm{Var}(G_{n_k}) = 1$ and therefore G is non-degenerate. As G was chosen arbitrarily among the potential limiting distributions, the claim follows. □

The following result is a slight alteration of Lemma 11 in Chernozhukov et al. (2013b).

Lemma 6. *Let X, Y be real-valued random variables, satisfying $\mathrm{P}(|X-Y| > \eta) \leq \varepsilon$ for some $\eta, \varepsilon > 0$. Denote by $q_Z(p)$ the quantile function of a real-valued random variable Z, i.e., $q_Z(p) = \inf\{t \in \mathbb{R} : \mathrm{P}(Z \leq t) \geq p\}$. Then, for $p \in (0,1)$*

$$q_Y(p-\varepsilon) - \eta \leq q_X(p) \leq q_Y(p+\varepsilon) + \eta.$$

Proof. Let $B = \{|X - Y| > \eta\}$. For any $t > 0$, it holds

$$\begin{aligned}
\mathrm{P}(X \leq t) &\leq \mathrm{P}(Y \leq t + |X - Y|) \\
&\leq \mathrm{P}(\{Y \leq t + |X - Y|\} \cap B^c) + \mathrm{P}(B) \\
&\leq \mathrm{P}(Y \leq t + \eta) + \varepsilon.
\end{aligned}$$

This implies by definition of the quantile function

$$q_X(p) \geq q_{Y-\eta}(p - \varepsilon) = q_Y(p - \varepsilon) - \eta.$$

For the other direction, by the same arguments as above

$$\begin{aligned}
\mathrm{P}(Y \leq t - \eta) &\leq \mathrm{P}(X \leq t + |X - Y| - \eta) \\
&\leq \mathrm{P}(\{X \leq t + |X - Y| - \eta\} \cap B^c) + \mathrm{P}(B) \\
&\leq \mathrm{P}(X \leq t) + \varepsilon
\end{aligned}$$

implying $q_X(p) \leq q_Y(p + \varepsilon) + \eta$. □

[22] Cf. Chapter 10, Theorem 8 in Pollard (2001).

Sufficient conditions for Assumption 2(iii) when the derivative is linear.

Lemma 7. *Suppose that Assumptions 1 holds with* $\mathbb{E} = \mathbb{R}$ *and* $h \mapsto \phi'_{\theta_0}(h)$ *linear. Further, suppose that* $X_1, \dots, X_n$ *is a random sample and that there exist* ψ_n *with* $\mathrm{E}[\phi'_{\theta_0}(\psi_n(X_i))] = 0$ *and* $E[|\phi'_{\theta_0}(\psi_n(X_i))|^3] \leq b_{2,n} < \infty$ *such that*

$$\hat{\theta}_n = \theta_0 + \sum_{i=1}^{n} \psi_n(X_i) + O_p(b_{1,n}). \tag{1.12}$$

Then, there exists a sequence $Y_n \sim \mathcal{N}(0, n \operatorname{Var}(\phi'_{\theta_0}(\psi_n(X_i))))$ *satisfying*

$$\phi'_{\theta_0}(\hat{\theta}_n - \theta_0) = Y_n + O_p\big((nb_{2,n})^{1/3} \log n + b_{1,n}\big).$$

If moreover $\mathbb{D}$ *is separable, there exists a sequence of centered Gaussian random vectors* Z_n *with values in* $\mathbb{D}$ *such that*

$$\phi'_{\theta_0}(\hat{\theta}_n - \theta_0) = \phi'_{\theta_0}(Z_n) + O_p\big((nb_{2,n})^{1/3} \log n + b_{1,n}\big)$$

and $\phi'_{\theta_0}(Z_n) \sim \mathcal{N}(0, n \operatorname{Var}(\phi'_{\theta_0}(\psi_n(X_i))))$.

Proof. By Yurinskii's Coupling (see e.g. Chapter 10 Theorem 10 in Pollard (2001)), there exists, for any $\delta > 0$ and n, a $Y_n(\delta) \sim \mathcal{N}(0, n \operatorname{Var}(\phi'_{\theta_0}(\psi_n(X_i))))$ satisfying

$$\mathrm{P}\left(\left|\sum_{i=1}^{n} \phi'_{\theta_0}(\psi_n(X_i)) - Y_n(\delta)\right| > 3\delta\right) \leq C_0 \frac{nb_{2,n}}{\delta^3}(1 + \log(|\delta^3/(nb_{2,n})|)),$$

for some universal constant C_0. Choosing $\delta_n = (nb_{2,n})^{1/3} \log(n)$ implies

$$\mathrm{P}\left(\left|\sum_{i=1}^{n} \phi'_{\theta_0}(\psi_n(X_i)) - Y_n(\delta_n)\right| > 3\delta_n\right) = O\left(\frac{\log(\log n)}{\log^3(n)}\right)$$

and in particular for $Y_n := Y_n(\delta_n)$,

$$\sum_{i=1}^{n} \phi'_{\theta_0}(\psi_n(X_i)) = Y_n + O_p(\delta_n).$$

Together with (1.12) and Lipschitz continuity of the derivative, this implies

$$\phi'_{\theta_0}(\hat{\theta}_n - \theta_0) = Y_n + O_p\big((nb_{2,n})^{1/3} \log n + b_{1,n}\big).$$

Now suppose that $\mathbb{D}$ is separable. In this case, it remains to show the existence of Z_n. Consider a centered Gaussian random vector Z_n on a probability space $(\Omega, \mathcal{A}, \mathrm{P})$ with values in $\mathbb{D}$ and whose covariance function Σ_n is given by

$$\Sigma_n(f,g) = n \operatorname{Cov}(f(\psi_n(X_i)), g(\psi_n(X_i))), \qquad f,g \in \mathbb{D}^*.$$

Since Z_n takes its values in a separable Banach space, it is Radon. By Proposition 3.6 in Ledoux and Talagrand (1991), there is an orthonormal basis (g_i) of the $L_2(\Omega, \mathcal{A}, \mathrm{P})$ closure of the variables of the form $f(Z_n)$ with $f \in \mathbb{D}^*$, such that $Z_n = \sum_i g_i x_i$ $\mathrm{P}-a.s.$ with $x_i = \mathrm{E}[g_i Z_n]$. The basis can be chosen to contain ϕ'_{θ_0}, or at least a properly normalized version of it. Without loss of generality, we can assume that g_1 corresponds to this normalized version. Then, $\phi'_{\theta_0}(g_1 x_1) \sim Y_n$ and $g_1 x_1 \perp\!\!\!\perp \sum_{i=2}^{\infty} g_i x_i$. Thus, the image measure of $(\phi'_{\theta_0}(g_1 x_1), \phi'_{\theta_0}(\sum_{i=2}^{\infty} g_i x_i))$ is a product measure on $\mathbb{R} \times \mathbb{R}$. These thoughts show that we can realize Z_n as the image of $\mathbb{R}^{\mathbb{N}}$ under the map $\sum_i g_i x_i$. Moreover, since $\phi'_{\theta_0}(Z_n) = g_1 \phi'_{\theta_0}(x_1)$ is a one-to-one mapping from g_1 to $\mathbb{R}$, we can set $g_1 = Y_n / \phi'_{\theta_0}(x_1)$, where we have used that $\phi'_{\theta_0}(x_1) \neq 0$ as $\operatorname{Var}(\phi'_{\theta_0}(Z_n)) > 0$. Thus, if we extend the original probability space, we can construct Z_n such that $Y_n = \phi'_{\theta_0}(Z_n)$. The claim follows. □

1.A.2 The bootstrap

Proof of Theorem 3: By the triangle inequality,

$$\|\hat{\phi}'_n(\hat{Z}_n) - \phi'_{\theta_0}(Z_n^*)\|_{\mathbb{E}} \leq \|\hat{\phi}'_n(\hat{Z}_n) - \phi'_{\theta_0}(\hat{Z}_n)\|_{\mathbb{E}} + \|\phi'_{\theta_0}(\hat{Z}_n) - \phi'_{\theta_0}(Z_n^*)\|_{\mathbb{E}}. \tag{1.13}$$

We will first derive a bound on the first term on the right-hand side. The second term can be bounded by the assumed coupling on the bootstrap process.

Fix some arbitrary $\varepsilon, \delta > 0$. Then, there exist $M_1, M_2, M_3, N_1, N_2, N_3$ such that the events

$$\begin{aligned} B_n &= \Big\{ \sup_{\|h\| \leq 1} \|\hat{\phi}'_n(h) - \phi'_{\theta_0}(h)\|_{\mathbb{E}} > M_1 \eta_n \Big\} \\ C_n &= \big\{ \|\hat{Z}_n\| > M_2 d_n \big\} \\ D_n &= \big\{ \mathrm{P}(\|\phi'_{\theta_0}(\hat{Z}_n) - \phi'_{\theta_0}(Z_n^*)\|_{\mathbb{E}} > M_3 s_n \,|\, \mathcal{D}_n) > \varepsilon/2 \big\} \end{aligned}$$

satisfy $\mathrm{P}(B_n) < \delta/4$ for all $n \geq N_1$, $\mathrm{P}(C_n) < \delta/4$ for all $n \geq N_2$ and $\mathrm{P}(D_n) < \delta/2$

for all $n \geq N_3$. On $B_n^c \cap C_n^c$ and by positive homogeneity of $\hat{\phi}'_n$ and ϕ'_{θ_0},

$$\begin{aligned}\|\hat{\phi}'_n(\hat{Z}_n) - \phi'_{\theta_0}(\hat{Z}_n)\|_{\mathbb{E}} =& \|\hat{Z}_n\| \left\| \hat{\phi}'_n\left(\frac{\hat{Z}_n}{\|\hat{Z}_n\|}\right) - \phi'_{\theta_0}\left(\frac{\hat{Z}_n}{\|\hat{Z}_n\|}\right)\right\|_{\mathbb{E}} \\ \leq& M_2 d_n \sup_{\|h\|\leq 1} \|\hat{\phi}'_n(h) - \phi'_{\theta_0}(h)\|_{\mathbb{E}} \leq M_1 M_2 d_n \eta_n.\end{aligned}$$

Hence,

$$\begin{aligned}&\mathrm{P}\big(\mathrm{P}(\|\hat{\phi}'_n(\hat{Z}_n) - \phi'_{\theta_0}(\hat{Z}_n)\|_{\mathbb{E}} > (M_1 M_2 d_n \eta_n) \vee (M_3 s_n) \,|\, \mathcal{D}_n) > \varepsilon/2\big) \\ \leq& \mathrm{P}\big(\big\{\mathrm{P}(\|\hat{\phi}'_n(\hat{Z}_n) - \phi'_{\theta_0}(\hat{Z}_n)\|_{\mathbb{E}} > (M_1 M_2 d_n \eta_n) \vee (M_3 s_n) \,|\, \mathcal{D}_n) > \varepsilon/2\big\} \cap (B_n^c \cap C_n^c)\big) \\ &\quad + \mathrm{P}(B_n \cup C_n) \\ \leq& \mathrm{P}(B_n) + \mathrm{P}(C_n) < \frac{\delta}{2}\end{aligned}$$

for all $n \geq N_1 \vee N_2$. Thus, we have shown that for any $\varepsilon > 0$, there exist $M = M_3 \vee M_1 M_2$ such that

$$\lim_{n\to\infty} \mathrm{P}\big(\mathrm{P}(\|\hat{\phi}'_n(\hat{Z}_n) - \phi'_{\theta_0}(\hat{Z}_n)\|_{\mathbb{E}} > M(d_n \eta_n \vee s_n) \,|\, \mathcal{D}_n) > \varepsilon\big) = 0.$$

Finally, by (1.13),

$$\begin{aligned}&\mathrm{P}(\|\hat{\phi}'_n(\hat{Z}_n) - \phi'_{\theta_0}(Z_n^*)\|_{\mathbb{E}} > 2(M_1 M_2 d_n \eta_n) \vee (M_3 s_n) \,|\, \mathcal{D}_n) \\ \leq& \mathrm{P}(\|\hat{\phi}'_n(\hat{Z}_n) - \phi'_{\theta_0}(\hat{Z}_n)\|_{\mathbb{E}} > (M_1 M_2 d_n \eta_n) \vee (M_3 s_n) \,|\, \mathcal{D}_n) \\ &\quad + \mathrm{P}(\|\phi'_{\theta_0}(\hat{Z}_n) - \phi'_{\theta_0}(Z_n^*)\|_{\mathbb{E}} > (M_1 M_2 d_n \eta_n) \vee (M_3 s_n) \,|\, \mathcal{D}_n).\end{aligned}$$

This implies

$$\begin{aligned}&\mathrm{P}\big(\mathrm{P}(\|\hat{\phi}'_n(\hat{Z}_n) - \phi'_{\theta_0}(Z_n^*)\|_{\mathbb{E}} > 2(M_1 M_2 d_n \eta_n) \vee (M_3 s_n) \,|\, \mathcal{D}_n) > \varepsilon\big) \\ \leq& \mathrm{P}\big(\mathrm{P}(\|\hat{\phi}'_n(\hat{Z}_n) - \phi'_{\theta_0}(\hat{Z}_n^*)\|_{\mathbb{E}} > (M_1 M_2 d_n \eta_n) \vee (M_3 s_n) \,|\, \mathcal{D}_n) > \varepsilon/2\big) \\ &\quad + \mathrm{P}\big(\mathrm{P}(\|\phi'_{\theta_0}(\hat{Z}_n) - \phi'_{\theta_0}(Z_n^*)\|_{\mathbb{E}} > M_3 s_n \,|\, \mathcal{D}_n) > \varepsilon/2\big) \\ <& \frac{\delta}{2} + \frac{\delta}{2} = \delta,\end{aligned}$$

for all $n \geq N_1 \vee N_2 \vee N_3$, where we have used that the first term on the right-hand side is less than $\delta/2$ by the above arguments and the second term is smaller than $\delta/2$ since $\mathrm{P}(D_n) < \delta/2$. Thus, setting $M = M_1 M_2 \vee M_3$, the claim follows. □

Proof of Theorem 4: Fix arbitrary $\varepsilon, \delta > 0$ and let

$$\Delta_n = \sup_{t>0} \left| \mathrm{P}\big(\hat{\phi}'_n(\hat{Z}_n) \leq t \,\big|\, \mathcal{D}_n\big) - \mathrm{P}\big(\phi'_{\theta_0}(Z_n^*) \leq t \,\big|\, \mathcal{D}_n\big) \right|.$$

Since $Z_n^* | \mathcal{D}_n \sim Z_n$, it holds $\mathrm{P}\big(\phi'_{\theta_0}(Z_n^*) \leq t \,\big|\, \mathcal{D}_n\big) = \mathrm{P}\big(\phi'_{\theta_0}(Z_n) \leq t\big)$ and by Theorem 2 there exist N_1 such that

$$\sup_{t>0} \left| \mathrm{P}\big(\phi'_{\theta_0}(Z_n) \leq t\big) - \mathrm{P}\big(\phi(\hat{\theta}_n) - \phi(\theta_0) \leq t\big) \right| < \varepsilon/2$$

for all $n \geq N_1$. Thus, conditionally on $\mathcal{D}_n$,

$$\begin{aligned} &\sup_{t>0} \left| \mathrm{P}\big(\hat{\phi}'_n(\hat{Z}_n) \leq t \,\big|\, \mathcal{D}_n\big) - \mathrm{P}\big(\phi(\hat{\theta}_n) - \phi(\theta_0) \leq t\big) \right| \\ \leq &\Delta_n + \sup_{t>0} \left| \mathrm{P}\big(\phi'_{\theta_0}(Z_n) \leq t\big) - \mathrm{P}\big(\phi(\hat{\theta}_n) - \phi(\theta_0) \leq t\big) \right| < \Delta_n + \frac{\varepsilon}{2} \end{aligned}$$

almost surely for all $n \geq N_1$. By Le Cam's Lemma 4, for any $\lambda > 0$

$$\Delta_n \leq \mathrm{P}\big(|\hat{\phi}'_n(\hat{Z}_n) - \phi'_{\theta_0}(Z_n^*)| > \lambda \,\big|\, \mathcal{D}_n\big) + \zeta_\lambda(\phi'_{\theta_0}(Z_n)),$$

where we have used that $Z_n^* | \mathcal{D}_n \sim Z_n$. By Theorem 1 in our companion paper Scherer (2024), $\zeta_\lambda(\phi'_{\theta_0}(Z_n)) \leq \sqrt{12}\lambda/\sigma_n$, where $\sigma_n = \sqrt{\mathrm{Var}(\phi'_{\theta_0}(Z_n))}$.

By Theorem 3, there exist M and N_2 such that the event

$$A_n = \left\{ \mathrm{P}\big(\|\hat{\phi}'_n(\hat{Z}_n) - \phi'_{\theta_0}(Z_n^*)\|_{\mathbb{E}} > M(d_n\eta_n \vee s_n) \,\big|\, \mathcal{D}_n\big) > \frac{\varepsilon}{2} \right\}$$

satisfies $\mathrm{P}(A_n) < \delta$ for all $n \geq N_2$. Moreover, by the rate requirement $\{(d_n\eta_n) \vee s_n\}$ $\mathrm{Var}(\phi'_{\theta_0}(Z_n))^{-1/2} \to 0$, there exist N_3 such that, on A_n^c,

$$\mathrm{P}\big(\|\hat{\phi}'_n(\hat{Z}_n) - \phi'_{\theta_0}(Z_n^*)\|_{\mathbb{E}} > \lambda_n \,\big|\, \mathcal{D}_n\big) \leq \frac{\varepsilon}{4}$$

for all $n \geq N_2 \vee N_3$ and for $\lambda_n < \varepsilon\sigma_n/(8\sqrt{3})$. Taking this λ_n in the Le Cam's bound above implies, on A_n^c,

$$\begin{aligned} \Delta_n \leq &\mathrm{P}\big(|\hat{\phi}'_n(\hat{Z}_n) - \phi'_{\theta_0}(Z_n^*)| > \lambda_n \,\big|\, \mathcal{D}_n\big) + \zeta_{\lambda_n}(\phi'_{\theta_0}(Z_n)) \\ \leq &\mathrm{P}\big(\|\hat{\phi}'_n(\hat{Z}_n) - \phi'_{\theta_0}(Z_n^*)\|_{\mathbb{E}} > \lambda_n \,\big|\, \mathcal{D}_n\big) + \frac{\sqrt{12}\lambda_n}{\sigma_n} < \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{\varepsilon}{2}, \end{aligned}$$

for all $n \geq N_2 \vee N_3$. Thus, on A_n^c,

$$B_n := \sup_{t>0} \left| \mathrm{P}\big(\hat{\phi}_n'(\hat{Z}_n) \leq t \,\big|\, \mathcal{D}_n\big) - \mathrm{P}\big(\phi(\hat{\theta}_n) - \phi(\theta_0) \leq t\big) \right| < \Delta_n + \frac{\varepsilon}{2} < \varepsilon,$$

for all $n \geq N_1 \vee N_2 \vee N_3$. This implies

$$\mathrm{P}(B_n > \varepsilon) \leq \underbrace{\mathrm{P}(\{B_n > \varepsilon\} \cap A_n^c)}_{=0} + \mathrm{P}(A_n) < \delta,$$

for all $n \geq N_1 \vee N_2 \vee N_3$ and therefore the claim follows by taking $N = N_1 \vee N_2 \vee N_3$. □

In the following, we proof a slightly more flexible result than Corollary 1. This result shows validity for sequences of confidence levels α_n. This flexibility is needed in our examples in the construction of the estimators of the derivative.

Corollary 4. *Suppose that the Assumptions in Theorem 4 hold. Further, suppose that for $\bar{\alpha} \in (0,1)$, $\hat{q}_n(1-\bar{\alpha}) > \tau$ with probability approaching one. Here $\tau = 0$ if $\mathcal{F} \neq \emptyset$ and $\tau = -\infty$ otherwise. Then, for any sequence (α_n) with $\alpha_n \to \alpha$ and $\alpha_n \leq \bar{\alpha}$, $n \in \mathbb{N}$,*

$$\liminf_{n\to\infty} \mathrm{P}\big(\phi(\hat{\theta}_n) - \phi(\theta_0) \leq \hat{q}_n(1-\alpha_n)\big) \geq 1 - \alpha.$$

Moreover, let $\mathrm{Med}(\phi_{\theta_0}'(Z_n))$ denote the median of $\phi_{\theta_0}'(Z_n)$ and $\sigma_{n,weak}^2$ denote the weak variance of $\phi_{\theta_0}'(Z_n)$, that is,

$$\sigma_{n,weak}^2 = \sup_{f \in \mathcal{F}_{\leq} : \|f\|^* = 1} \mathrm{Var}(f(Z_n)).$$

Here $\mathcal{F}_{\leq} = \{f \in \mathbb{D}^* : \forall h \in \mathbb{D} : f(h) \leq \phi_{\theta_0}'(h)\}$. Then,

$$\hat{q}_n(1-\alpha_n) = O_p(\mathrm{Med}(\phi_{\theta_0}'(Z_n)) + \sigma_{n,weak}\sqrt{\log(1/(2\alpha_n))}).$$

Proof of Corollary 4: Let Δ_n denote the Kolmogorov distance between the bootstrap distribution and the sample distribution, i.e.,

$$\Delta_n = \sup_{t>\tau} \left| \mathrm{P}\big(\hat{\phi}_n'(\hat{Z}_n) \leq t \,\big|\, \mathcal{D}_n\big) - \mathrm{P}\big(\phi(\hat{\theta}_n) - \phi(\theta_0) \leq t\big) \right|.$$

As by assumption, $\Delta_n = o_p(1)$, there exists a sequence $\varepsilon_n \downarrow 0$ such that $\mathrm{P}(\Delta_n \leq$

$\varepsilon_n) \to 1$ as $n \to \infty$. Moreover, by monotonicity of $t \mapsto \hat{q}_n(t)$, also $\hat{q}_n(1-\alpha_n) > \tau$ with probability converging to one. Thus, there exists an Ω_n such that $\mathrm{P}(\Omega_n) \to 1$ and, for all $\mathcal{D}_n \in \Omega_n$, $\Delta_n \leq \varepsilon_n$, $\alpha_n \leq \alpha + \varepsilon_n$ as well as $\hat{q}_n(1-\alpha_n) > \tau$ and therefore

$$\begin{aligned}\mathrm{P}\big(\phi(\hat{\theta}_n) - \phi(\theta_0) \leq \hat{q}_n(1-\alpha_n)\big) \geq& \mathrm{P}\big(\hat{\phi}'_n(\hat{Z}_n) \leq \hat{q}_n(1-\alpha_n) \,\big|\, \mathcal{D}_n\big) - \Delta_n \\ \geq& 1 - \alpha_n - \varepsilon_n \geq 1 - \alpha - 2\varepsilon_n.\end{aligned}$$

Hence, on Ω_n, $\hat{q}_n(1-\alpha_n) \geq q_n(1-\alpha-2\varepsilon_n)$, where $q_n(1-\alpha-2\varepsilon_n)$ denotes the $(1-\alpha-2\varepsilon_n)$-quantile of $\phi(\hat{\theta}_n) - \phi(\theta_n)$. This implies that unconditionally

$$\begin{aligned}\mathrm{P}\big(\phi(\hat{\theta}_n) - \phi(\theta_0) > \hat{q}_n(1-\alpha_n)\big) \leq& \mathrm{P}\big(\{\phi(\hat{\theta}_n) - \phi(\theta_0) > \hat{q}_n(1-\alpha_n)\} \cap \Omega_n\big) + \mathrm{P}(\Omega_n^c) \\ \leq& \mathrm{P}\big(\phi(\hat{\theta}_n) - \phi(\theta_0) > q_n(1-\alpha-2\varepsilon_n)\big) + \mathrm{P}(\Omega_n^c) \\ \leq& \alpha + 2\varepsilon_n + \mathrm{P}(\Omega_n^c) \to \alpha,\end{aligned}$$

since $\varepsilon_n \downarrow 0$ and $\mathrm{P}(\Omega_n) \to 1$, proving the claim.

Moreover, by a similar argument as above, on Ω_n, $\hat{q}_n(1-\alpha_n) \leq \tilde{q}_n(1-\alpha+\varepsilon_n)$, where $\tilde{q}_n$ denotes the quantile function of $\phi'_{\theta_0}(Z_n)$. As the derivative is Lipschitz continuous and sublinear and Z_n is a tight Gaussian random vector on a Banach space, the same arguments as in the Lipschitz concentration inequality for Gaussian random vectors in Lemma 3.1 in Ledoux and Talagrand (1991), imply that

$$\mathrm{P}(\phi'_{\theta_0}(Z_n) > \mathrm{Med}(\phi'_{\theta_0}(Z_n)) + \sigma_{n,weak} t) \leq \frac{1}{2}\exp(-t^2/2), \qquad t > 0.$$

Taking $t = \sqrt{2\log(1/(2\alpha_n))}$, this inequality implies

$$\mathrm{P}(\phi'_{\theta_0}(Z_n) \leq \mathrm{Med}(\phi'_{\theta_0}(Z_n)) + \sigma_{n,weak} t) \geq 1 - \alpha_n$$

and therefore $q_n(1-\alpha_n) = O(\mathrm{Med}(\phi'_{\theta_0}(Z_n)) + \sigma_{n,weak}\sqrt{\log(1/(2\alpha_n))})$. The claim follows. □

Auxiliary results

Lemma 8. *Suppose that (1.6) holds, then* $\|\hat{Z}_n\| = O_p(d_n)$.

Proof. Fix $\varepsilon > 0$ and take $\delta = \varepsilon/2$. Then, there exists M and N so that

$$\mathrm{P}\big(\underbrace{\mathrm{P}(\|\hat{Z}_n\| > Md_n \,|\, \mathcal{D}_n) < \varepsilon/2}_{=:B_n}\big) > 1 - \frac{\varepsilon}{2}, \qquad \forall n \geq N.$$

By the law of iterated expectations,

$$\begin{aligned}
\mathrm{P}(\|\hat{Z}_n\| > Md_n) &= \mathrm{E}\big[\mathrm{P}(\|\hat{Z}_n\| > Md_n \,|\, \mathcal{D}_n)\big] \\
&= \mathrm{E}[\mathbb{1}(B_n)\mathrm{P}(\|\hat{Z}_n\| > Md_n \,|\, \mathcal{D}_n) + \mathbb{1}(B_n^c)\mathrm{P}(\|\hat{Z}_n\| > Md_n \,|\, \mathcal{D}_n)] \\
&< \frac{\varepsilon}{2} + \mathrm{P}(B_n^c) < \varepsilon
\end{aligned}$$

for all $n \geq N$. The claim follows. □

The following Lemma gives sufficient conditions on the derivative ϕ'_{θ_0} and the approximate distribution sequence Z_n so that $\hat{q}_n(1-\alpha) > \tau$.

Lemma 9. *Suppose that the Assumptions in Theorem 4 hold. Let* $\mathcal{F} = \{f \in \mathbb{D}^* \,|\, \forall x \in \mathbb{D} : f(x) \leq \phi'_{\theta_0}(x)\}$*, where* $\mathbb{D}^*$ *denotes the topological dual space of* $\mathbb{D}$*. Further, let* Z_n *be centered and set* $\mathcal{F}_n = \{f \in \mathcal{F} : \mathrm{Var}(f(Z_n)) = 0\}$*. Then,* $\tau = 0$ *and* $\hat{q}_n(1-\alpha) > 0$ *with probability approaching one for all* $\alpha < 1/2$*. If further* $\mathcal{F}_n = \emptyset$*, for all* $n \in \mathbb{N}$*, then* $\tau = -\infty$ *and* $\hat{q}_n(1-\alpha) > \tau$ *with probability approaching one for all* $\alpha \in (0,1)$*.*

Proof. The values of τ follow from the same arguments as in the proof of Theorem 2. Further, by the same arguments as in Lemma 11 in Appendix B in Chernozhukov et al. (2013b), the coupling of the bootstrap process imply that $\hat{q}_n(1-\alpha) \geq q_n(1-\alpha-\delta_n) - \varepsilon_n$ with probability approaching one for some $\delta_n, \varepsilon_n \downarrow 0$, where $q_n(t)$ denotes the t-quantile of $\phi'_{\theta_0}(Z_n)$. Since $h \mapsto \phi'_{\theta_0}(h)$ is not the zero functional by assumption, $\mathcal{F} \setminus \mathcal{F}_n \neq \emptyset$. Thus, for any $\alpha < 1/2$,

$$1 - \alpha \leq \mathrm{P}\big(\phi'_{\theta_0}(Z_n) \leq q_n(1-\alpha)\big) \leq \mathrm{P}(f(Z_n) \leq q_n(1-\alpha)),$$

for any $f \in \mathcal{F} \setminus \mathcal{F}_n$. Since $f(Z_n) \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma^2 \in (0, \infty)$, this set of inequalities implies $q_n(1-\alpha) > 0$. Since $\delta_n, \varepsilon_n \downarrow 0$, this implies $\hat{q}_n(1-\alpha) > 0$ with probability converging to one, proving the claim. □

General Purpose Bootstrap Procedures

Proof of Lemma 1: We can decompose

$$\begin{aligned}\phi(\hat{\theta}_n + \hat{Z}_n) - \phi(\hat{\theta}_n) =& \phi'_{\theta_0}((\hat{\theta}_n - \theta_0) + \hat{Z}_n) - \phi'_{\theta_0}(\hat{\theta}_n - \theta_0) \\ &+ \big(\phi\big(\theta_0 + \big\{(\hat{\theta}_n - \theta_0) + \hat{Z}_n\big\}\big) - \phi(\theta_0) - \phi'_{\theta_0}((\hat{\theta}_n - \theta_0) + \hat{Z}_n)\big) \\ &- \big(\phi\big(\theta_0 + \big\{\hat{\theta}_n - \theta_0\big\}\big) - \phi(\theta_0) - \phi'_{\theta_0}(\hat{\theta}_n - \theta_0)\big).\end{aligned}$$

Here, the first line corresponds to the derivative and the second and third lines to the analytical approximation errors. By the same arguments as in the proof of Theorem 1,

$$\phi(\hat{\theta}_n + \hat{Z}_n) - \phi(\hat{\theta}_n) = \phi'_{\theta_0}((\hat{\theta}_n - \theta_0) + \hat{Z}_n) - \phi'_{\theta_0}(Z_n) + O_p\big(a_n^\gamma + d_n^\gamma + b_n\big).$$

Let A_n given by

$$A_n = \{\|\phi(\hat{\theta}_n + \hat{Z}_n) - \phi(\hat{\theta}_n) - [\phi'_{\theta_0}((\hat{\theta}_n - \theta_0) + \hat{Z}_n) - \phi'_{\theta_0}(Z_n)]\|_{\mathbb{E}} \leq C(a_n^\gamma + d_n^\gamma + b_n)\}.$$

By the above argument, A_n has probability arbitrarily close to one for all sufficiently large n, by choosing C appropriately. By the triangle inequality, we have

$$\begin{aligned}&\|\phi(\hat{\theta}_n + \hat{Z}_n) - \phi(\hat{\theta}_n) - [\phi'_{\theta_0}(Z_n + Z_n^*) - \phi'_{\theta_0}(Z_n)]\|_{\mathbb{E}} \\ \leq& \|\phi(\hat{\theta}_n + \hat{Z}_n) - \phi(\hat{\theta}_n) - [\phi'_{\theta_0}((\hat{\theta}_n - \theta_0) + \hat{Z}_n) - \phi'_{\theta_0}(Z_n)]\|_{\mathbb{E}} \\ &+ \|\phi'_{\theta_0}((\hat{\theta}_n - \theta_0) + \hat{Z}_n) - \phi'_{\theta_0}(Z_n + Z_n^*)\|_{\mathbb{E}}\end{aligned}$$

and therefore for any K and all $\mathcal{D}_n \in A_n$

$$\begin{aligned}&\mathrm{P}\big(\|\phi(\hat{\theta}_n + \hat{Z}_n) - \phi(\hat{\theta}_n) - \phi'_{\theta_0}(Z_n + Z_n^*) - \phi'_{\theta_0}(Z_n)\|_{\mathbb{E}} > K(b_n \vee s_n \vee a_n^\gamma \vee d_n^\gamma) \,\big|\, \mathcal{D}_n\big) \\ \leq& \mathrm{P}\big(\|\phi'_{\theta_0}((\hat{\theta}_n - \theta_0) + \hat{Z}_n) - \phi'_{\theta_0}(Z_n + Z_n^*)\|_{\mathbb{E}} > K(b_n \vee s_n \vee a_n^\gamma \vee d_n^\gamma)/2 \,\big|\, \mathcal{D}_n\big) \\ &+ \mathrm{P}\big(\|\phi(\hat{\theta}_n + \hat{Z}_n) - \phi(\hat{\theta}_n) - [\phi'_{\theta_0}((\hat{\theta}_n - \theta_0) + \hat{Z}_n) - \phi'_{\theta_0}(Z_n)]\|_{\mathbb{E}} \\ &\qquad > K(b_n \vee s_n \vee a_n^\gamma \vee d_n^\gamma)/2 \,\big|\, \mathcal{D}_n\big).\end{aligned}$$

If K is chosen sufficiently large, the second term on the right-hand size is exactly zero on A_n. Therefore, for any $\varepsilon > 0$,

$$\mathrm{P}\big(\mathrm{P}\big(\|\phi(\hat{\theta}_n + \hat{Z}_n) - \phi(\hat{\theta}_n) - \phi'_{\theta_0}(Z_n + Z_n^*) - \phi'_{\theta_0}(Z_n)\|_{\mathbb{E}}$$

$$
\begin{aligned}
&> K(b_n \vee s_n \vee a_n^\gamma \vee d_n^\gamma) \,\big|\, \mathcal{D}_n\big) > \varepsilon\big) \\
\leq &\mathrm{P}\big(\mathrm{P}\big(\|\phi(\hat\theta_n + \hat Z_n) - \phi(\hat\theta_n) - \phi'_{\theta_0}(Z_n + Z_n^*) - \phi'_{\theta_0}(Z_n)\|_{\mathbb{E}} \\
&> K(b_n \vee s_n \vee a_n^\gamma \vee d_n^\gamma) \,\big|\, \mathcal{D}_n\big) > \varepsilon, A_n\big) + \mathrm{P}(A_n^c) \\
\leq &\mathrm{P}\big(\mathrm{P}\big(\|\phi'_{\theta_0}((\hat\theta_n - \theta_0) + \hat Z_n) - \phi'_{\theta_0}(Z_n + Z_n^*)\|_{\mathbb{E}} \\
&> K(b_n \vee s_n \vee a_n^\gamma \vee d_n^\gamma)/2 \,\big|\, \mathcal{D}_n\big)\big) + \mathrm{P}(A_n^c).
\end{aligned}
$$

The second term on the right-hand side can be made arbitrarily small by choosing C and K sufficiently large. Next for this given K, the first term on the right-hand side converges to zero by assumption. The claim follows. □

Proof of Lemma 2: Subsampling can be interpreted as using implicitly the derivative estimator

$$
\hat\phi'_n(h) = \frac{1}{a_m}\big(\phi(\hat\theta_n + a_m h) - \phi(\hat\theta_n)\big).
$$

By γ-Fréchet directional differentiability, for any h with $\|h\| \leq 1$,

$$
\begin{aligned}
\hat\phi'_n(h) =& \frac{1}{a_m}\left(\phi\left(\theta_0 + a_m\left\{h + \frac{a_n}{a_m}\frac{1}{a_n}(\hat\theta_n - \theta_0)\right\}\right) - \phi(\theta_0)\right) \\
&- \frac{1}{a_m}\left(\phi\left(\theta_0 + \frac{a_n}{a_m}\frac{1}{a_n}(\hat\theta_n - \theta_0)\right) - \phi(\theta_0)\right) \\
=& \phi'_{\theta_0}\left(h + \frac{a_n}{a_m}\frac{1}{a_n}(\hat\theta_n - \theta_0)\right) - \phi'_{\theta_0}\left(\frac{a_n}{a_m}\frac{1}{a_n}(\hat\theta_n - \theta_0)\right) + O_p(a_m^{\gamma-1}).
\end{aligned}
$$

By Lipschitz continuity of the derivative and $\|\hat\theta_n - \theta_0\| = O_p(a_n)$,

$$
\begin{aligned}
\phi'_{\theta_0}\left(h + \frac{a_n}{a_m}\frac{1}{a_n}(\hat\theta_n - \theta_0)\right) =& \phi'_{\theta_0}(h) + O_p(a_n/a_m) \\
\phi'_{\theta_0}\left(\frac{a_n}{a_m}\frac{1}{a_n}(\hat\theta_n - \theta_0)\right) =& O_p(a_n/a_m).
\end{aligned}
$$

Note that the above holds uniformly over bounded sets B and therefore

$$
\sup_{h\in B}\|\hat\phi'_n(h) - \phi'_{\theta_0}(h)\|_{\mathbb{E}} = O_p\left(\frac{a_n}{a_m} + a_m^{\gamma-1}\right).
$$

The claim now follows by applying the general bootstrap theorem for $\hat Z_m^*$. □

Proof. By the proof of Lemma 2, for any bounded set B,

$$\sup_{h\in B}\|\hat{\phi}'_n(h) - \phi'_{\theta_0}(h)\|_{\mathbb{E}} = O_p\left(\frac{a_n}{a_m} + a_m^{\gamma-1}\right).$$

The remainder of the claim follows along the same lines as the proof of Lemma 2 by using the different rates for $\hat{Z}_n$. □

1.A.3 Local analysis

Proof of Theorem 5. By γ-Fréchet directional differentiability at θ_0,

$$\begin{aligned}\phi(\hat{\theta}_n) - \phi(\theta_n) =& \phi(\theta_0 + \{\hat{\theta}_n - \theta_n\} + h_n) - \phi(\theta_0)\\ &+ \phi(\theta_0 + h_n) - \phi(\theta_0)\\ =& \phi'_{\theta_0}(\{\hat{\theta}_n - \theta_n\} + h_n) - \phi'_{\theta_0}(h_n) + O_p(\|\hat{\theta}_n - \theta_n\|^\gamma + \|h_n\|^\gamma),\end{aligned}$$

where we used that $f(x) = x^\gamma$ for $\gamma > 1$, $x \in \mathbb{R}$ with $x > 0$ is convex and therefore $f(x+y) \leq 2^{\gamma-1}(f(x)+f(y))$ for all $x, y \geq 0$. The claim now follows by the assumed coupling. □

Proof of Theorem 6. The claim follows along the same lines as the proof of Theorem 2. We only have to discuss the anti-concentration bound of the approximate distribution.

Since $\phi'_{\theta_0}(h_n)$ is a deterministic real number

$$\begin{aligned}&\sup_{t>\tau} \mathrm{P}(t \leq \phi'_{\theta_0}(Z_{n,h_n} + h_n) - \phi'_{\theta_0}(h_n) \leq t + \eta)\\ =&\sup_{t>\tau} \mathrm{P}(t \leq \phi'_{\theta_0}(Z_{n,h_n} + h_n) \leq t + \eta).\end{aligned}$$

By sublinearity of ϕ'_{θ_0} and since $Z_{n,h_n} + h_n$ is Gaussian, we can apply the anti-concentration bound in our companion paper Scherer (2024). In the case $\mathcal{F} \neq \emptyset$, we have $\tau = \sup_{f\in\mathcal{F}} f(h_n)$ and $\sigma^2_{n,h}$ as given in the Theorem. If instead $\mathcal{F} = \emptyset$, we have $\tau = -\infty$ and $\sigma^2_{n,h} = \mathrm{Var}(\phi'_{\theta_0}(Z_n + h_n))$. In either cases, the implied anti-concentration bound has exactly the same form as in the proof of Theorem 2. Thus, by the rate requirements and the same arguments as in the latter proof, we obtain

$$\lim_{n\to\infty} \sup_{t>\tau} \left|\mathrm{P}(\phi(\hat{\theta}_n) - \phi(\theta_n) \leq t) - \mathrm{P}\left(\phi'_{\theta_0}(Z_{n,h_n} + h_n) - \phi'_{\theta_0}(h_n) \leq t\right)\right| = 0,$$

for $\tau \in \{\tau_n, -\infty\}$. □

Validity of the confidence interval

The proof of Corollary 2 relies on a one-sided modification of Le Cam's Lemma 4:

Lemma 10. *For $X, Z \in \mathbb{R}$ arbitrary random variables and $\lambda > 0$,*

$$\sup_{t\in\mathbb{R}} \mathrm{P}(X \leq t) - \mathrm{P}(Z \leq t) \leq \mathrm{P}(Z - X > \lambda) + \zeta_\lambda(X) \vee \zeta_\lambda(Z),$$

where $\zeta_\lambda(V) = \sup_{t\in\mathbb{R}} \mathrm{P}(t \leq V \leq t + \lambda)$ for any real-valued random variable $V \in \mathbb{R}$.

Proof. For any $t \in \mathbb{R}$ and $\lambda > 0$,

$$\begin{aligned}\mathrm{P}(X \leq t) &\leq \mathrm{P}(X \leq t, Z - X \leq \lambda) + \mathrm{P}(Z - X > \lambda) \\ &\leq \mathrm{P}(Z \leq t + \lambda) + \mathrm{P}(Z - X > \lambda) \\ &\leq \mathrm{P}(Z \leq t) + \mathrm{P}(t \leq Z \leq t + \lambda) + \mathrm{P}(Z - X > \lambda).\end{aligned}$$

The proof follows by rearranging this inequality and taking the supremum over t. □

Proof of Corollary 2: The proof follows along the same lines as indicated in section 1.4. It only remains to show the validity of the bootstrap in this triangular array setup. But this follows by the same arguments as in the proof of Theorems 3, 4 and Corollary 1. In particular, we have by similar arguments as in Theorem 3 for any $\varepsilon > 0$ exists M such that

$$\lim_{n\to\infty} \mathrm{P}\big(\mathrm{P}(\phi'_{\theta_0}(Z_n^*) - \hat{\phi}'_n(\hat{Z}_n) > M(d_n\eta_n \vee s_n) \,|\, \mathcal{D}_n) > \varepsilon\big) = 0.$$

This implies by Le Cam's one-sided Lemma 10 and the arguments in the proof of Theorem 4

$$\sup_{t\in\mathbb{R}} \mathrm{P}(\hat{\phi}'_n(\hat{Z}_n) \leq t \,|\, \mathcal{D}_n) - \mathrm{P}_n(\phi'_{\theta_0}(Z_{n,h_n}) \leq t) = o_p(1),$$

which implies for any $t \in \mathbb{R}$ and $\varepsilon > 0$ with probability converging to one

$$\mathrm{P}_n(\phi'_{\theta_0}(Z_{n,h_n}) \leq t) \geq \mathrm{P}_n(\hat{\phi}'_n(\hat{Z}_n) \leq t \,|\, \mathcal{D}_n) - \varepsilon$$

and in particular for $t = \hat{q}_{n,1-\alpha}$,

$$\mathrm{P}_n(\phi'_{\theta_0}(Z_{n,h_n}) \leq \hat{q}_{n,1-\alpha}) \geq 1 - \alpha - \varepsilon.$$

This implies by the same arguments as in the proof of Corollary 1

$$\liminf_{n\to\infty} \mathrm{P}_n(\phi'_{\theta_0}(Z_{n,h_n}) \leq \hat{q}_{n,1-\alpha}) \geq 1 - \alpha$$

and the claim follows. □

1.A.4 Results on the differentiability concept

Before proving properties of γ-Fréchet directional differentiability, we prove some equivalent characterization of it.

Lemma 11. *ϕ is γ-Fréchet directionally differentiable at θ if and only if for any bounded set $K \subset \mathbb{D}$, and $t_n \downarrow 0$ such that $\theta + t_n h \in \mathbb{D}_\phi$, for all $h \in K$, $n \in \mathbb{N}$,*

$$\sup_{h\in K} \left\| \frac{\phi(\theta + t_n h) - \phi(\theta)}{t_n} - \phi'_\theta(h) \right\|_{\mathbb{E}} = O(t_n^{\gamma-1}). \tag{1.14}$$

Proof. First suppose that (1.14) holds. Let h_n be any sequence such that $\|h_n\| \to 0$. Set $t_n = \|h_n\|$ and $g_n = h_n/\|h_n\|$. Then $t_n \downarrow 0$ and $\|g_n\| = 1$. In particular, $K = \{g_n : n \in \mathbb{N}\}$ is bounded. It holds

$$\|\phi(\theta + h_n) - \phi(\theta) - \phi'_\theta(h_n)\|_{\mathbb{E}} \leq t_n \sup_{g\in K} \left\| \frac{\phi(\theta + t_n g) - \phi(\theta)}{t_n} - \phi'_\theta(g) \right\|_{\mathbb{E}} = O(t_n^\gamma).$$

Hence, by definition of t_n, ϕ is γ-Fréchet directionally differentiable at θ.

Now, suppose ϕ does not satisfy the (1.14). Then there exists a bounded set K and $t_n \downarrow 0$ so that for any $C > 0$ and all $n \in \mathbb{N}$

$$\sup_{h\in K} \|\phi(\theta + t_n h) - \phi(\theta) - \phi'_\theta(t_n h)\|_{\mathbb{E}}/t_n^\gamma > C.$$

Thus, there also exists a sequence $(h_n) \subset K$ so that

$$\|\phi(\theta + t_n h_n) - \phi(\theta) - \phi'_\theta(t_n h_n)\|_{\mathbb{E}}/(t_n\|h_n\|)^\gamma > C/\sup_{h\in K}\|h\|.$$

Set $g_n = t_n h_n$. Then, $\|g_n\| \to 0$ but

$$\frac{\|\phi(\theta + g_n) - \phi(\theta) - \phi'_\theta(g_n)\|_{\mathbb{E}}}{\|g_n\|^\gamma} > C / \sup_{h \in K} \|h\|.$$

As this holds for any C and all n, ϕ cannot be γ-Fréchet directionally differentiable at θ. The claim follows. □

For the arguments below, we need that the domain $\mathbb{D}_\phi$ is sufficiently large in order to ensure that there exist sufficiently many sequences $t_n \downarrow$ so that $\theta + t_n h_n \in \mathbb{D}_\phi$. A sufficient condition is that $\mathbb{D}_\phi$ has non-empty interior and that θ is contained in the interior of $\mathbb{D}_\phi$.

Lemma 12. *Suppose that $\phi : \mathbb{D}_\phi \to \mathbb{E}$ is γ-Fréchet directionally differentiable at $\theta \in \mathbb{D}_\phi$ with derivative $\phi'_\theta : \mathbb{D} \to \mathbb{E}$. Further, suppose that θ lies in the interior of $\mathbb{D}_\phi$. Then,*

(i) ϕ'_θ is positively homogeneous, i.e., for all $\lambda \geq 0$ and $h \in \mathbb{D}_0$ so that $\lambda h \in \mathbb{D}_0$, $\phi'_\theta(\lambda h) = \lambda \phi'_\theta(h)$.

(ii) If ϕ is Lipschitz continuous with respect to $\|\cdot\|$, then ϕ'_θ is also Lipschitz continuous with respect to $\|\cdot\|$.

Proof. **(i):** Fix $\lambda \geq 0$ and $h \in \mathbb{D}$ so that $\lambda h \in \mathbb{D}_0$. Then, for any sequence $t_n \downarrow 0$ so that $\theta + t_n h \in \mathbb{D}_\phi$, we have that also $s_n := t_n / \lambda \downarrow 0$ as well as $\theta + s_n \lambda h \in \mathbb{D}_\phi$ and therefore by the equivalence Lemma 11

$$\phi'_\theta(\lambda h) = \lim_{n \to \infty} \frac{\phi(\theta + s_n \lambda h) - \phi(\theta)}{s_n} = \lambda \lim_{n \to \infty} \frac{\phi(\theta + t_n h) - \phi(\theta)}{t_n} = \lambda \phi'_\theta(h).$$

(ii): The proof follows along the same lines as in the discussion before Proposition 3.5 in Shapiro (1990).

Again using the equivalence Lemma 11, let $h, h' \in \mathbb{D}_0$ and $t_n \downarrow 0$ so that $\theta + t_n h \in \mathbb{D}_\phi$ and $\theta + t_n h' \in \mathbb{D}_\phi$ for all $n \in \mathbb{N}$. Such sequences exist since θ lies in the interior of $\mathbb{D}_\phi$ and therefore by choosing t_n sufficiently small, $\theta + t_n h$ and $\theta + t_n h'$ lie in a sufficiently small ball around θ. Given this construction, we have

$$\begin{aligned} \phi'_\theta(h) - \phi'_\theta(h') = \lim_{n \to \infty} & \frac{\phi(\theta + t_n h) - \phi(\theta + t_n h')}{t_n} \\ & + \phi'_\theta(h) - \frac{\phi(\theta + t_n h) - \phi(\theta)}{t_n} + \frac{\phi(\theta + t_n h') - \phi(\theta)}{t_n} - \phi'_\theta(h'). \end{aligned}$$

By Fréchet directional differentiability, the terms in the second line converge to zero. Further, by Lipschitz continuity of ϕ, there is some L such that

$$\left\| \frac{\phi(\theta + t_n h) - \phi(\theta + t_n h')}{t_n} \right\|_{\mathbb{E}} \leq L \|h - h'\|_{\mathbb{D}},$$

implying that also ϕ'_θ is Lipschitz continuous. □

We close this section with a uniqueness result for the derivative function.

Lemma 13. *Suppose that $\phi : \mathbb{D} \to \mathbb{E}$ is Gâteaux directionally differentiable at θ. Then, the derivative ϕ'_θ is uniquely determined.*

Proof. Suppose that ϕ'_θ and g both satisfy, for all $h \in \mathbb{D}$,

$$\lim_{t \downarrow 0} \left\| \frac{\phi(\theta + th) - \phi(h)}{t} - \phi'_\theta(h) \right\|_{\mathbb{E}} = 0$$
$$\lim_{t \downarrow 0} \left\| \frac{\phi(\theta + th) - \phi(h)}{t} - g(h) \right\|_{\mathbb{E}} = 0.$$

Thus, for any $t > 0$,

$$\|\phi'_\theta(h) - g(h)\|_{\mathbb{E}} \leq \left\| \frac{\phi(\theta + th) - \phi(h)}{t} - \phi'_\theta(h) \right\|_{\mathbb{E}} + \left\| \frac{\phi(\theta + th) - \phi(h)}{t} - g(h) \right\|_{\mathbb{E}}.$$

The right-hand side converges to zero as $t \downarrow 0$. Since the left-hand side does not depend on t, this implies $\|\phi'_\theta(h) - g(h)\|_{\mathbb{E}} = 0$ for all h, i.e., $\phi'_\theta = g$. □

Second order derivative

We propose here a notion of second order Fréchet directional differentiability similar to the proposal in Definition 2.2 in Hong and Li (2020).

Definition 3. *ϕ is twice Fréchet directionally differentiable at θ_0, if there are continuous maps $\phi'_{\theta_0}, \phi''_{\theta_0} : \mathbb{D} \to \mathbb{R}$ with ϕ'_{θ_0} positively homogeneous and ϕ''_{θ_0} positively homogeneous of order 2 such that*

$$\phi(\theta_0 + h) = \phi(\theta_0) + \phi'_{\theta_0}(h) + \frac{1}{2}\phi''_{\theta_0}(h) + o(\|h\|_c^2)$$

as $\|h\| \to 0$.

This notion of second order Fréchet directional derivative can also be justified by considering the Fréchet directional derivative of the first derivative as done in Yamamuro (2006) in the context of full Fréchet differentiability. The continuity of the derivative function is not necessary for directional differentiability per se, we however rely on it in the following result.

Lemma 14. *Suppose that ϕ is twice Fréchet directionally differentiable at θ_0 and that its second derivative ϕ''_{θ_0} is continuous at zero. Then, it is also 2-Fréchet directionally differentiable at θ_0.*

Proof. By continuity of the second derivative at zero, for any $\varepsilon > 0$ there is some $\delta > 0$ such that when $\|h\| \leq \delta$ it holds $\|\phi''_{\theta_0}(h)\|_{\mathbb{E}} < \varepsilon$. In particular, this implies by positive homogeneity of order 2

$$\sup_{\|h\|\leq 1} \|\phi''_{\theta_0}(h)\|_{\mathbb{E}} = \frac{1}{\delta^2} \sup_{\|h\|\leq 1} \|\phi''_{\theta_0}(\delta h)\|_{\mathbb{E}} = \frac{1}{\delta^2} \sup_{\|h\|\leq \delta} \|\phi''_{\theta_0}(h)\|_{\mathbb{E}} \leq \frac{\varepsilon}{\delta^2} < \infty.$$

But this implies as ϕ''_{θ_0} is positively homogeneous of order 2

$$\|\phi''_{\theta_0}(h)\|_{\mathbb{E}} \leq \|h\|^2 \sup_{\|h\|=1} \|\phi''_{\theta_0}(h)\|_{\mathbb{E}} = O(\|h\|^2),$$

proving the claim. □

1.A.5 Extension of our main results to the non-measurable case

In this section, we extend our main results to allow the preliminary estimator $\hat{\theta}_n$ to be non-measurable with respect to the Borel σ-algebra on $\mathbb{D}$. Measurability problems occur for as fundamental estimators as the empirical distribution function. Indeed, it can be shown that the empirical distribution function interpreted as an estimator in the space of càdlàg functions endowed with the supremum norm fails to be Borel measurable (cf. Chapter 18 in Billingsley (1999)). By the same arguments, the Nadaraya-Watson estimator of the conditional distribution function fails to be measurable with respect to the Borel σ-algebra on $\mathbb{D}$ as defined in Example 2.

We follow conceptually the ideas in the Hoffmann-Jørgenson theory of weak convergence as discussed in detail in van der Vaart and Wellner (1996). This theory defines a notion of weak convergence of probability measures which allows the sequence to not be measurable while assuming that its limit is measurable. Similarly,

we do not assume that the preliminary estimator is measurable but only that the coupled random vectors Z_n are.

For the following, further notation is needed. Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space and $T : \Omega \to \bar{\mathbb{R}}$ an arbitrary map, where $\bar{\mathbb{R}}$ denotes the extended reals. Let $\mathrm{E}^*[T]$ and $\mathrm{E}_*[T]$ denote the outer (inner) integral of T given by

$$\mathrm{E}^*[T] = \inf\{\mathrm{E}[U] : U \geq T, U : \Omega \to \bar{\mathbb{R}} \text{ measurable }, \mathrm{E}[U] \text{ exists } \}$$
$$\mathrm{E}_*[T] = \sup\{\mathrm{E}[U] : U \leq T, U : \Omega \to \bar{\mathbb{R}} \text{ measurable }, \mathrm{E}[U] \text{ exists } \},$$

where $\mathrm{E}[U]$ exists if at least one of $\mathrm{E}[\max\{U, 0\}]$ or $\mathrm{E}[\max\{-U, 0\}]$ is finite. Similarly, denote by $\mathrm{P}^*(B)$ and $\mathrm{P}_*(B)$ the outer (inner) measure of $B \subset \Omega$, i.e.,

$$\mathrm{P}^*(B) = \inf\{\mathrm{P}(A) : A \supset B, A \in \mathcal{A}\}$$
$$\mathrm{P}_*(B) = \sup\{\mathrm{P}(A) : A \subset B, A \in \mathcal{A}\}.$$

Synonymously, we will call the outer (inner) measure an outer (inner) probability. Finally, for a sequence of maps $X_n : \Omega \to \bar{\mathbb{R}}$, we set $X_n = o_{p^*}(1)$ if X_n converges in outer probability to zero and $X_n = O_{p^*}(1)$ if X_n is stochastically bounded with respect to the outer measure. The corresponding notation for the inner measure is defined analogously.

The Delta method

Given this notation, we can formulate our weakened assumptions on the preliminary estimator.

Assumption 7 (On $\hat{\theta}_n$)**.** *(i) The estimator $\hat{\theta}_n$ is a function mapping a sequence of random variables $\{X_i\}_{i=1}^n$ into $\mathbb{D}_\phi$.*

(ii) There exists a sequence $a_n \to 0$ such that $\|\hat{\theta}_n - \theta_0\| = O_{p^*}(a_n)$.

(iii) There exists a sequence of real numbers $b_n \to 0$ and a sequence of $\mathbb{D}$-valued random vectors $\{Z_n\}_{i=1}^n$ such that

$$\|\phi'_{\theta_0}(\hat{\theta}_n - \theta_0) - \phi'_{\theta_0}(Z_n)\|_{\mathbb{E}} = O_p(b_n).$$

In comparison to Assumption 2, we do not assume $\hat{\theta}_n$ to be measurable anymore and assume that the bound on the norm holds in outer probability. Note that when

(ii) holds, we also have $\|\hat{\theta}_n - \theta_0\| = O_{p_*}(a_n)$ as $\mathrm{P}_*(B) \leq \mathrm{P}^*(B)$ for all $B \subset \Omega$. We keep the statement on the strong approximation bound in (iii) unchanged as such coupling bounds typically do not rely on measurability of the estimator.[23] Assumption 7 captures a similar notion of asymptotic measurability as in the Hoffmann-Jørgenson theory of weak convergence. As there, we do not assume that the estimator is measurable and only that the approximating random vector Z_n is.

While outer integrals and probability are not that commonly encountered in other areas of econometrics and statistics, we want to stress that (ii) is not necessarily harder to show than for measurable estimators. This is because usually such bounds are based on upper bounds which are measurable anyway and to which the usual probabilistic arguments apply.

Given this weakened assumption, we can proof the following Delta method:

Theorem 7. *Let Assumptions 1 and 7 hold. Then,*

$$\phi(\hat{\theta}_n) - \phi(\theta_0) = \phi'_{\theta_0}(Z_n) + O_{p^*}(a_n^\gamma + b_n).$$

Its proof follows along the same lines as the proof of Theorem 1 and is therefore omitted. One only has to swap the corresponding statements in probability with statements in outer probability.

Moreover, Theorem 2 can also be shown under Assumption 7:

Theorem 8. *Suppose that Assumptions 1 and 7 hold with $\mathbb{E} = \mathbb{R}$. Suppose that ϕ'_{θ_0} is sublinear and that Z_n is a sequence of centered and tight Gaussian random vectors. If*

$$\frac{a_n^\gamma + b_n}{\sigma_n} \to 0,$$

where $\sigma_n^2 = \mathrm{Var}(\phi'_{\theta_0}(Z_n))$, then

$$\lim_{n\to\infty} \sup_{t>0} \left| \mathrm{P}^*(\phi(\hat{\theta}_n) - \phi(\theta_0) \leq t) - \mathrm{P}\big(\phi'_{\theta_0}(Z_n) \leq t\big) \right| = 0.$$

If further $\mathcal{F} = \{f \in \mathbb{D}^* \,|\, \forall h \in \mathbb{D} : f(h) \leq \phi'_{\theta_0}(h),\ \mathrm{Var}(f(Z_n)) = 0\} = \emptyset$, then

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}} \left| \mathrm{P}^*(\phi(\hat{\theta}_n) - \phi(\theta_0) \leq t) - \mathrm{P}\big(\phi'_{\theta_0}(Z_n) \leq t\big) \right| = 0.$$

[23] See the discussion in Dudley and Philipp (1983).

The proof of Theorem 8 follows along the same lines as Theorem 2 and is omitted. One only has to swap Le Cam's Lemma with the following version for non-measurable mappings:

Lemma 15. *Let* $(\Omega, \mathcal{A}, \mathrm{P})$ *be a probability space,* $X : \Omega \to \bar{\mathbb{R}}$ *and* Z *a real random variable. For any* $\tau \in \mathbb{R}$ *and* $\lambda > 0$*,*

$$\sup_{t>\tau} |\mathrm{P}^*(X \leq t) - \mathrm{P}(Z \leq t)| \leq \mathrm{P}^*(|X - Z| > \lambda) + \zeta_\lambda(Z),$$
$$\sup_{t>\tau} |\mathrm{P}^*(X > t) - \mathrm{P}(Z > t)| \leq \mathrm{P}^*(|X - Z| > \lambda) + \zeta_\lambda(Z),$$

where $\zeta_\lambda(V) = \sup_{t>\tau} \mathrm{P}(t \leq V \leq t + \lambda)$ *for real-valued* V*.*

Proof. The claim follows along the same lines as the proof for measurable maps as it only uses monotonicity and subadditivity of probability measures which also hold for outer measures. Therefore, this result does not apply to inner measures as they are superadditive instead of subadditive.

For any $t \in \mathbb{R}$ and any $\lambda > 0$,

$$\begin{aligned}
\mathrm{P}^*(X \leq t) \leq& \mathrm{P}^*(X \leq t, |X - Z| \leq \lambda) + \mathrm{P}^*(|X - Z| > \lambda) \\
\leq& \mathrm{P}(Z \leq t + \lambda) + \mathrm{P}^*(|X - Z| > \lambda) \\
\leq& \mathrm{P}(Z \leq t) + \mathrm{P}(t \leq Z \leq t + \lambda) + \mathrm{P}^*(|X - Z| > \lambda).
\end{aligned}$$

Moreover, it holds

$$\begin{aligned}
\mathrm{P}(Z \leq t) \leq& \mathrm{P}(Z \leq t - \lambda) + \mathrm{P}(t - \lambda \leq Z \leq t) \\
\leq& \mathrm{P}^*(Z \leq t, |X - Z| \leq \lambda) + \mathrm{P}^*(|X - Z| > \lambda) + \mathrm{P}(t - \lambda \leq Z \leq t) \\
\leq& \mathrm{P}^*(X \leq t) + \mathrm{P}^*(|X - Z| > \lambda) + \mathrm{P}(t - \lambda \leq Z \leq t).
\end{aligned}$$

Thus, we have shown

$$\sup_{t \geq \tau} |\mathrm{P}^*(X \leq t) - \mathrm{P}(Z \leq t)| \leq \mathrm{P}^*(|X - Z| > \lambda) + \sup_{t \geq \tau} \mathrm{P}(t \leq Z \leq t + \lambda).$$

The second inequality follows using similar arguments. □

The bootstrap

Assumption 8. *(On the Bootstrap* $\hat{Z}_n$*)*

(i) $\hat{Z}_n : \{X_i, W_i\}_{i=1}^n \to \mathbb{D}$ *with* $\{W_i\}_{i=1}^n$ *independent of* $\mathcal{D}_n = \{X_i\}_{i=1}^n$.

(ii) There is some sequence $d_n \to 0$ *such that* $\|\hat{Z}_n\| = O_{p^*}(d_n)$

(iii) There is some sequence $s_n \to 0$ *and some* $Z_n^* | \mathcal{D}_n \sim Z_n$ *such that for any* $\varepsilon > 0$, *there exist* M *such that*

$$\lim_{n\to\infty} \mathrm{P}\big(\mathrm{P}(\|\phi'_{\theta_0}(\hat{Z}_n) - \phi'_{\theta_0}(Z_n^*)\|_{\mathbb{E}} > M s_n \,|\, \mathcal{D}_n) > \varepsilon\big) = 0.$$

As in Assumption 7, we do not assume that the bootstrap is measurable and only assume a bound on the norm in outer probability, while keeping the coupling in probability. Assumption 4 is left unchanged.

Theorem 9. *Let the Assumptions of Theorem 1, Assumptions 4 and 8 hold. Then, for any* $\varepsilon > 0$, *there exist* M *such that*

$$\lim_{n\to\infty} \mathrm{P}^*\big(\mathrm{P}^*(\|\hat{\phi}'_n(\hat{Z}_n) - \phi'_{\theta_0}(Z_n^*)\|_{\mathbb{E}} > M(d_n\eta_n \vee s_n) \,|\, \mathcal{D}_n) > \varepsilon\big) = 0.$$

Theorem 9 follows along the same lines as Theorem 3 while replacing probability statements with the corresponding statements in outer probability. Moreover, we have to clarify what is meant by a conditional outer measure. First note that the proof only works with conditional distributions of real-valued functions given the data which are measurable. Since the real numbers are a Polish space, these conditional distributions can be defined as regular conditional probability distributions and in particular are probability measures for each fixed realization of the data $\mathcal{D}_n$. In this setting, we can define the outer conditional distribution of a regular conditional probability distribution of Y given X as

$$\mathrm{P}^*_{Y|X}(B \,|\, x) = \inf\{\mathrm{P}(Y \in A \,|\, X = x) : A \supset B, A \in \mathcal{A}\}, \qquad \forall B \subset \mathbb{R}, \forall x.$$

As a regular conditional probability distribution is in general only almost surely unique, also this version of a conditional outer measure is only almost surely unique. Therefore, one also has to adapt the proof slightly by carrying around a set of measure zero.

Similarly as for Theorem 8, we can extend Theorem 4 to the non-measurable case by replacing statements in probability with statements in outer probability and Lemma 15.

Theorem 10. *Suppose that the Assumptions of Theorem 2, Assumptions 8 and 4 hold. Then, if* $\{(d_n\eta_n) \vee s_n\} \operatorname{Var}(\phi'_{\theta_0}(Z_n))^{-1/2} \to 0$,

$$\sup_{t>0} \left| \mathrm{P}^*\big(\hat{\phi}'_n(\hat{Z}_n) \leq t \,\big|\, \mathcal{D}_n\big) - \mathrm{P}\big(\phi(\hat{\theta}_n) - \phi(\theta_0) \leq t\big) \right| = o_{p^*}(1).$$

If further $\mathcal{F} = \{f \in \mathbb{D}^* \,|\, \forall h \in \mathbb{D} : f(h) \leq \phi'_{\theta_0}(h),\ \operatorname{Var}(f(Z_n)) = 0\} = \emptyset$, *then*

$$\sup_{t\in\mathbb{R}} \left| \mathrm{P}^*\big(\hat{\phi}'_n(\hat{Z}_n) \leq t \,\big|\, \mathcal{D}_n\big) - \mathrm{P}\big(\phi(\hat{\theta}_n) - \phi(\theta_0) \leq t\big) \right| = o_{p^*}(1).$$

An extension of Corollary 1 to the non-measurable setting is a bit trickier. The estimated quantile as defined in (1.7) is not necessarily defined when the bootstrap fails to be measurable. However, we can still define a similar quantile estimator as the limit of a Monte Carlos simulation. That is, let $\{W_b = (W_{1b}, \ldots, W_{nb})^\intercal : b = 1, \ldots, B\}$ denote a random sample of the randomization variables W_b. Then, we can estimate the conditional distribution $\mathrm{P}(\hat{\phi}'_n(\hat{Z}_n(\mathcal{D}_n, W_b)) \leq t \,|\, \mathcal{D}_n)$ via

$$\hat{P}_B(\hat{\phi}'_n(\hat{Z}_n(\mathcal{D}_n, W_b)) \leq t \,|\, \mathcal{D}_n) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\big\{\hat{\phi}'_n(\hat{Z}_n(\mathcal{D}_n, W_b)) \leq t\big\}$$

and define the quantile estimator $\hat{q}_{n,B}(1-\alpha)$ as

$$\hat{q}_{n,B}(1-\alpha) = \inf\{q : \hat{P}_B(\hat{\phi}'_n(\hat{Z}_n(\mathcal{D}_n, W_b)) \leq t \,|\, \mathcal{D}_n) \geq 1-\alpha\}.$$

Then, using essentially the same arguments as in the proof of Corollary 1, one can show:

Corollary 5. *Suppose that the Assumptions in Theorem 10 hold. Further, suppose that for* $\alpha \in (0,1)$, $\hat{q}_{n,B}(1-\alpha) > \tau$ *with probability approaching one. Here* $\tau = 0$ *if* $\mathcal{F} \neq \emptyset$ *and* $\tau = -\infty$ *otherwise. Then,*

$$\liminf_{n\to\infty} \mathrm{P}^*\big(\phi(\hat{\theta}_n) - \phi(\theta_0) > \hat{q}_{n,B}(1-\alpha_n)\big) \leq \alpha.$$

Moreover, let $\operatorname{Med}(\phi'_{\theta_0}(Z_n))$ *denote the median of* $\phi'_{\theta_0}(Z_n)$ *and* $\sigma^2_{n,weak}$ *denote the weak variance of* $\phi'_{\theta_0}(Z_n)$, *that is,*

$$\sigma^2_{n,weak} = \sup_{f\in\mathcal{F}_{\leq}:\|f\|^*=1} \operatorname{Var}(f(Z_n)).$$

Here $\mathcal{F}_{\leq} = \{f \in \mathbb{D}^* : \forall h \in \mathbb{D} : f(h) \leq \phi'_{\theta_0}(h)\}$. *Then,*

$$\hat{q}_{n,B}(1-\alpha) = O_{p^*}(\text{Med}(\phi'_{\theta_0}(Z_n)) + \sigma_{n,weak}\sqrt{\log(1/(2\alpha))}),$$

Local analysis

For our local analysis, we need for any n a potentially different outer measure. As we do not work with minimal measurable majorants or essential infima, this does not introduce any problems.[24]

As for the Delta method and bootstrap, we need to change the assumption on the preliminary estimator:

Assumption 9. *(i) The estimator* $\hat{\theta}_n$ *is a function mapping a sequence of random variables* $\{X_i\}_{i=1}^n$ *into* $\mathbb{D}_\phi$.

(ii) $\theta_n = \theta_0 + h_n$ *with* $\|h_n\| = O(a_n)$.

(iii) There exists a sequence $a_n \to 0$ *such that* $\|\hat{\theta}_n - \theta_n\| = O_{p_n^*}(a_n)$.

(iv) There exists a sequence of real numbers $b_n \to 0$ *and a sequence of* $\mathbb{D}$*-valued random vectors* Z_{n,h_n} *such that* $Z_{n,h_n} \sim Z_n$, $n \in \mathbb{N}$, *and*

$$\|\phi'_{\theta_0}(\{\hat{\theta}_n - \theta_n\} + h_n) - \phi'_{\theta_0}(Z_{n,h_n} + h_n)\|_{\mathbb{E}} = O_{p_n}(b_n).$$

Given this assumption, the same arguments as in the proofs of Theorem 5 and Theorem 6 apply when swapping probability statements with statements in outer probability.

Theorem 11. *Suppose Assumption 1 and 5 hold. Then*

$$\phi(\hat{\theta}_n) - \phi(\theta_n) = \phi'_{\theta_0}(Z_{n,h_n} + h_n) - \phi'_{\theta_0}(h_n) + O_{p_n^*}\big(a_n^\gamma + b_n\big).$$

Theorem 12. *Suppose that Assumptions 1 and 5 hold with* $\mathbb{E} = \mathbb{R}$. *Suppose that* ϕ'_{θ_0} *is sublinear and that* Z_n *is a sequence of centered and tight Gaussian random vectors. If*

$$\frac{a_n^\gamma + b_n}{\sigma_{n,h}} \to 0,$$

[24] See Chapter 1.2 in van der Vaart and Wellner (1996) for a discussion.

where $\sigma_{n,h}^2 = \mathrm{Var}(\sup_{f\in\mathcal{F}_+} f(Z_{n,h}+h_n))$ *with* $\mathcal{F}_+ = \{f \in \mathbb{D}^* \,|\, \forall h \in \mathbb{D} : f(h) \leq \phi'_{\theta_0}(h)\} \setminus \mathcal{F}$*, then*

$$\lim_{n\to\infty} \sup_{t>\tau_n} \left|\mathrm{P}_n^*(\phi(\hat{\theta}_n) - \phi(\theta_n) \leq t) - \mathrm{P}_n\big(\phi'_{\theta_0}(Z_{n,h}+h_n) - \phi'_{\theta_0}(h_n) \leq t\big)\right| = 0,$$

where $\tau_n = \sup_{f\in\mathcal{F}} f(h_n)$*.*

If instead $\mathcal{F} = \{f \in \mathbb{D}^* \,|\, \forall h \in \mathbb{D} : f(h) \leq \phi'_{\theta_0}(h), \mathrm{Var}(f(Z_n)) = 0\} = \emptyset$ *and*

$$\frac{a_n^\gamma + b_n}{\sigma_{n,h}} \to 0,$$

where $\sigma_{n,h}^2 = \mathrm{Var}(\phi'_{\theta_0}(Z_{n,h}+h_n))$*, then*

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}} \left|\mathrm{P}_n^*(\phi(\hat{\theta}_n) - \phi(\theta_n) \leq t) - \mathrm{P}_n\big(\phi'_{\theta_0}(Z_n+h_n) - \phi'_{\theta_0}(h_n) \leq t\big)\right| = 0.$$

For the extension of Corollary 2, we need a slightly altered form of Assumption 6:

Assumption 10. *(i)* $\hat{Z}_n : \{X_i, W_i\}_{i=1}^n \to \mathbb{D}$ *with* $\{W_i\}_{i=1}^n$ *independent of* $\mathcal{D}_n = \{X_i\}_{i=1}^n$*.*

(ii) There is some sequence $d_n \to 0$ *such that* $\|\hat{Z}_n\| = O_{p_n^*}(d_n)$

(iii) There is some sequence $s_n \to 0$ *and some* $Z_n^*|\mathcal{D}_n \sim Z_n$ *such that for any* $\varepsilon > 0$*, there exist* M *such that*

$$\lim_{n\to\infty} \mathrm{P}_n\big(\mathrm{P}_n(\|\phi'_{\theta_0}(\hat{Z}_n) - \phi'_{\theta_0}(Z_n^*)\|_{\mathbb{E}} > M s_n \,|\, \mathcal{D}_n) > \varepsilon\big) = 0.$$

(iv) The map $\hat{\phi}'_n : \mathbb{D} \to \mathbb{E}$ *is a measurable function of* $\{X_i\}_{i=1}^n$*, which is positively homogeneous and satisfies for a sequence* $\eta_n \to 0$ *and any bounded* $K \subset \mathbb{D}$*,*

$$\sup_{h\in K} \|\hat{\phi}'_n(h) - \phi'_{\theta_0}(h)\|_{\mathbb{E}} = O_{p_n}(\eta_n).$$

Corollary 6. *Suppose the conditions of Theorem 12 and Assumption 10 hold. Then, if* $\mathcal{F} = \emptyset$*, it holds for any* $\alpha \in (0,1)$ *that*

$$\liminf_{n\to\infty} \mathrm{P}_n^*(\phi(\hat{\theta}_n) - \phi(\theta_n) > \hat{q}_{n,B}(1-\alpha)) \leq \alpha.$$

As for the other results in this section, the proof follows along the same lines as the original one and therefore is omitted.

1.B Proofs concerning the examples

1.B.1 Maximum of conditional mean function

Proof of Differentiability:

Proposition 1. *Let $\phi : \mathcal{C}([0,1]) \to \mathbb{R}$ be given by $\phi(\theta) = \max_{x\in[0,1]} \theta(x)$ and denote by $\Psi(\theta) = \operatorname{argmax}_x \theta(x)$. Further, suppose that θ satisfies the well-separatedness condition (1.4). Then, for $\rho > 1$ and if $\|h\|_{BL} < \delta$,*

$$|\phi(\theta + h) - \phi(\theta) - \phi'_\theta(h)| \leq \frac{1}{c}\|h\|_{BL}^{\frac{\rho}{\rho-1}}$$

and if $\rho = 1$ as well as $\|h\|_{BL} < \delta$,

$$|\phi(\theta + h) - \phi(\theta) - \phi'_\theta(h)| \leq \frac{2}{c}\|h\|_{BL}^2.$$

Proof. First suppose that $\rho > 1$. The case $\rho = 1$ will be treated later.

Let $\Psi(\theta) := \operatorname{argmax}_{x\in[0,1]} \theta(x)$ and denote by $d(x, \Psi(\theta)) = \inf_{x^*\in\Psi(\theta)}|x - x^*|$. Further, for any $x \in [0,1]$, let $x^*(x) \in \operatorname{argmin}_{x'\in[0,1]}|x - x'|$. It holds, for any $x \in [0,1]$,

$$\begin{aligned} \theta(x) + h(x) &\leq \theta(x) + h(x^*(x)) + \|h\|_{BL} d(x, \Psi(\theta)) \\ \max_x\{\theta(x) + h(x)\} &\geq \phi(\theta) + \phi'_\theta(h) \geq \phi(\theta) + h(x^*(x)). \end{aligned}$$

Hence, for any maximizer $x_h \in \Psi(\theta + h) := \operatorname{argmax}_{x\in[0,1]}\{\theta(x) + h(x)\}$,

$$\theta(x_h) + h(x^*(x_h)) + \|h\|_{BL} d(x_h, \Psi(\theta)) \geq \theta(x_h) + h(x_h) \geq \phi(\theta) + h(x^*(x_h))$$

and therefore

$$\phi(\theta) - \theta(x_h) \leq \|h\|_{BL} d(x_h, \Psi(\theta)).$$

By the well-separatedness condition,

$$\phi(\theta) - \theta(x_h) \geq (c\, d(x_h, \Psi(\theta))^\rho) \wedge \delta.$$

If $\|h\|_{BL} < \delta$, $\phi(\theta) - \theta(x_h) \leq \|h\|_{BL} d(x_h, \Psi(\theta)) < \delta$ and therefore the well-separatedness condition implies

$$\|h\|_{BL} d(x_h, \Psi(\theta)) \geq c\, d(x_h, \Psi(\theta))^\rho$$

or equivalently,

$$d(x_h, \Psi(\theta)) \leq \frac{\|h\|_{BL}^{\frac{1}{\rho-1}}}{c} =: \varepsilon(h)$$

implying $\Psi(\theta + h) \subset \Psi(\theta)^{\varepsilon(h)}$. Thus, if $\|h\|_{BL} < \delta$,

$$\begin{aligned}\phi(\theta + h) - \phi(\theta) &= \max_{x \in \Psi(\theta)^{\varepsilon(h)}} \{\theta(x) + h(x)\} - \max_{x \in \Psi(\theta)} \theta(x) \\ &= \max_{x \in \Psi(\theta)^{\varepsilon(h)}} \{\theta(x) + h(x)\} - \max_{x \in \Psi(\theta)} \{\theta(x) + h(x)\} + \phi'_\theta(h)\end{aligned}$$

where we have used that θ is constant on $\Psi(\theta)$. The first term on the right-hand side satisfies

$$\begin{aligned}0 < &\max_{x \in \Psi(\theta)^{\varepsilon(h)}} \{\theta(x) + h(x)\} - \max_{x \in \Psi(\theta)} \{\theta(x) + h(x)\} \\ \leq &\phi(\theta) + \max_{x \in \Psi(\theta)^{\varepsilon(h)}} h(x) - \max_{x \in \Psi(\theta)} \{\theta(x) + h(x)\} \\ \leq &\max_{x \in \Psi(\theta)^{\varepsilon(h)}} h(x) - \max_{x \in \Psi(\theta)} h(x).\end{aligned}$$

Let $\tilde{x} \in \operatorname{argmax}_{x \in \Psi(\theta)^{\varepsilon(h)}} h(x)$. Then, since $x^*(\tilde{x}) \in \Psi(\theta)$,

$$\begin{aligned}&\max_{x \in \Psi(\theta)^{\varepsilon(h)}} h(x) - \max_{x \in \Psi(\theta)} h(x) \leq h(\tilde{x}) - h(x^*(\tilde{x})) \\ \leq &\sup_{|x - x'| \leq \varepsilon(h)} |h(x) - h(x')| \leq \|h\|_{BL} \varepsilon(h) = \frac{1}{c} \|h\|_{BL}^{\frac{\rho}{\rho-1}}.\end{aligned}$$

Thus, the claim for $\rho > 1$ follows. For $\rho = 1$, it holds for any $x_h \in \Psi(\theta + h)$,

$$\theta(x_h) + h(x_h) \leq \theta(x_h) + \|h\|_\infty$$

$$\theta(x_h) + h(x_h) \geq \phi(\theta) + \phi'_\theta(h) \geq \phi(\theta) - \|h\|_\infty$$

and therefore $\phi(\theta) - \theta(x_h) \leq 2\|h\|_{BL}$ as $\|h\|_\infty \leq \|h\|_{BL}$. By similar arguments as for $\rho > 1$, if $\|h\|_{BL} < \delta$,

$$d(x_h, \Psi(\theta)) \leq \frac{2\|h\|_{BL}}{c} =: \varepsilon(h)$$

implying $\Psi(\theta + h) \subset \Psi(\theta)^{\varepsilon(h)}$. By the same arguments as for $\rho > 1$, we obtain

$$|\phi(\theta + h) - \phi(\theta) - \phi'_\theta(h)| \leq \frac{2}{c}\|h\|^2_{BL},$$

proving the claim for $\rho = 1$. □

The following lemma discusses tightness of Proposition 1 above.

Lemma 16. *Suppose that the conditions of Proposition 1 hold.*

(i) There exists a function $\theta \in \mathcal{C}([0,1])$ satisfying the well-separatedness condition with $\rho > 1$ and a sequence of functions h_n with $\|h_n\|_{BL} \to 0$ such that the bound given in Proposition 1 holds for all $\rho > 1$ and sufficiently large n.

(ii) ϕ is not Fréchet directionally differentiable with respect to the supremum norm.

(iii) There does not exist a $\gamma > 1$ such that ϕ satisfying the well-separatedness condition with $\rho > 1$ is γ-Hadamard directionally differentiable in the sense of (1.3).

Proof. **(i):** Suppose that $\theta(x) = 1 - x^\ell$ for $x \in [0,1]$, $\ell > 1$, and $h_n(x) = b_n x$ for $b_n > 0$. Then, $\Psi(\theta) = \{0\}$ and $|\theta^{(s)}(0)| = 0$ for all $s < \ell$ and $|\theta^{(\ell)}(0)| \neq 0$. Therefore, $\phi'_\theta(h_n) = 0$. Moreover, $\theta(x) + h_n(x) = 1 - x^\ell + b_n x$ has a unique maximum at $x^* = (b_n/\ell)^{1/\ell - 1}$ with value given by

$$\begin{aligned}\phi(\theta + h_n) =& 1 - \left(\frac{b_n}{\ell}\right)^{\ell/(\ell-1)} + b_n\left(\frac{b_n}{\ell}\right)^{1/(\ell-1)} = 1 - (\ell^{-\ell/(\ell-1)} - \ell^{1/(\ell-1)})b_n^{\ell/(\ell-1)} \\ =& 1 - \ell^{-\ell/(\ell-1)}(\ell - 1)b_n^{\ell/(\ell-1)}\end{aligned}$$

Moreover, Thus,

$$\phi(\theta + h_n) = \phi(\theta) + \phi'_\theta(h_n) - \ell^{-\ell/(\ell-1)}(\ell - 1)b_n^{\ell/(\ell-1)}$$

Note that $b_n = \|h_n\|_L$, and therefore we obtain exactly the rate of the approximation error as claimed by the proposition.

(ii): No, as can be seen by the following example. Let again $\theta(x) = 1 - x^\rho$ for $x \in [0,1]$, $\rho > 1$, and $h_c(x) = \min\{cx, 1\}$, $c \geq 1$. Then $B := \{h_c : c \geq 1\}$ is bounded with respect to the supremum norm and

$$\phi(\theta + th_c) = 1 - (ct/\rho)^\rho + t$$

with unique maximizer $x^* = c/\rho t$. Now take $t = \rho/c^2$ and let $c \to \infty$. Then $x^* = 1/c$ and $\phi(\theta + th_c) = 1 - c^{-\rho} + 1$. Further, $\phi(\theta) = 1$ and $\phi'_\theta(h) = 0$, for all c and therefore

$$\phi(\theta + th_c) - \phi(\theta) - t\phi'_\theta(h_c) = 1 - c^{-\rho}$$

or

$$\frac{\phi(\theta + th_c) - \phi(\theta)}{t} - \phi'_\theta(h_c) = \frac{c^2 - c}{\rho} \to \infty$$

Hence, there is a bounded set and a sequence $t_n \downarrow 0$ such that

$$\lim_{n\to\infty} \sup_{h\in B} \left| \frac{\phi(\theta + th) - \phi(\theta)}{t} - \phi'_\theta(h) \right| = \infty.$$

(iii): Consider the same θ as in (i) and (ii) but now take $h_\beta(x) = x^\beta$, for $\beta \in (0,1)$. Then, $\theta + th_\beta$ is maximized at $x^* = (t/\rho)^{1/(\rho-\beta)}$ with maximum

$$\phi(\theta + th_\beta) = 1 - \left(\frac{t}{\rho}\right)^{\frac{\rho}{\rho-\beta}} + \rho^{-\frac{\beta}{\rho-\beta}} t^{\frac{\rho}{\rho-\beta}}$$

and therefore

$$\frac{\phi(\theta + th_\beta) - \phi(\theta)}{t} - \phi'_\theta(h_\beta) = \left(\rho^{-\frac{\beta}{\rho-\beta}} - \rho^{-\frac{\rho}{\rho-\beta}}\right) t^{\frac{\beta}{\rho-\beta}}$$

and the exponent can be made arbitrarily close to zero. Hence, ϕ is not γ-Hadamard directionally differentiable. □

Assumptions

The assumptions are based on the conditions in Example 7 in Appendix F in Chernozhukov et al. (2013b). This allows us to use the coupling results on both the

estimator and the multiplier bootstrap.

Assumption 11. *(i)* (Y_i, X_i), $i = 1, \dots, n$ *are i.i.d.*

(ii) X_i *has a continuously differentiable density* $f : \mathbb{R} \to \mathbb{R}$ *which is bounded away from zero and infinity on* $[0, 1]$

$$0 < c_f \leq f(x) \leq C_f < \infty, \qquad \forall x \in [0, 1]$$

(iii) ε_i *is bounded and the density of* ε_i *conditionally on* X_i *exists and is uniformly bounded from above and below. The conditional variance* $\sigma_\varepsilon^2(x) := \mathrm{E}[\varepsilon_i^2 \mid X_i = x]$ *is bounded away from zero and twice continuously differentiable.*

(iv) θ *is* $(\ell+1)$ *times continuously differentiable whose* $(\ell+1)$*th derivative is Lipschitz continuous. Further,* θ *satisfies the well-separatedness condition (1.4) for some* $c, \delta > 0$ *and* $\rho \in \{1, 2\}$.

(v) K *is a twice continuously differentiable second order kernel with support* $[-1, 1]$.

(vi) $h_n \to 0$, $nh_n^3 \to \infty$, $nh_n^5 / \log^{24} n \to \infty$, $nh_n^{2\ell+3} \to 0$, $\sqrt{n^{-1}h_n^{-2}} \to 0$ *at polynomial rates in* n.

We differ here slightly from the presentation in the main text in that we assume that X is supported on a potentially larger set than $[0, 1]$, and we only focus for inference on $[0, 1]$. This assumption is imposed to shorten the proofs as we can rely on the results in Kong et al. (2010) and Chernozhukov et al. (2013b). In particular, this assumption simplifies the equivalent kernel of the local polynomial estimator so that it does not change at the boundary.

The rate requirements are imposed in order to apply the uniform Bahadur expansion in Kong et al. (2010). The Lipschitz continuity of the $(\ell + 1)$th derivative of θ is used in the derivation of the rate of the bounded Lipschitz distance. The assumption that ε_i is bounded is used in the Rio-Massart coupling and can be weakened by assuming smoothness assumptions on the conditional quantile function of ε_i given X_i. The rate requirement $nh_n^{2\ell+3} \to 0$ is an undersmoothing condition and implies that the approximating Gaussian process is centered. Moreover, the rate conditions require higher order smoothness $\ell > 1$, ruling out the local linear estimator. The well-separatedness condition is required for bounding the analytic approximation error made by the Delta method.

Bounded Lipschitz distance

The local polynomial estimator can be written as a weighted average of the Y_i as

$$\hat{\theta}_n(x) = \sum_{i=1}^{n} w_i(x) Y_i$$

with weights

$$w_i(x) = U(0)^\intercal B_n(x)^{-1} a_{n,i}(x)$$

where

$$\begin{aligned} a_{n,i}(x) &= \frac{1}{nh} K\left(\frac{X_i - x}{h}\right) U\left(\frac{X_i - x}{h}\right) \\ B_n(x) &= \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) U\left(\frac{X_i - x}{h}\right) U^\intercal\left(\frac{X_i - x}{h}\right) \\ U(x) &= (1, x, x^2/2!, \ldots, x^\ell/\ell!)^\intercal. \end{aligned}$$

Lemma 17. *Under Assumption 11,*

$$\|\hat{\theta}_n - \theta\|_{BL} = O_p\left(\sqrt{\frac{\log n}{nh^3}} + h^{\ell+1}\right).$$

Proof. By standard arguments, one can show that

$$\|\hat{\theta}_n - \theta\|_{\infty} = O_p\left(\sqrt{\frac{\log n}{nh}} + h^{\ell+1}\right).$$

In the following we will derive an upper bound on the rate of convergence of the Lipschitz constant of $\hat{\theta}_n - \theta$. As it turns out, this rate is slower than the rate in sup norm.

In order to bound the Lipschitz constant, we split

$$\begin{aligned} &\hat{\theta}_n(x) - \theta(x) - \{\hat{\theta}_n(x') - \theta(x')\} \\ =&\sum_{i=1}^{n} w_i(x)(\theta(X_i) - \theta(x)) - \sum_{i=1}^{n} w_i(x')(\theta(X_i) - \theta(x')) \end{aligned}$$

$$+\sum_{i=1}^{n}\{w_i(x)-w_i(x')\}\varepsilon_i.$$

For the bias terms, by a Taylor expansion, there exist intermediate values ξ_i and ζ_i such that

$$\begin{aligned}\sum_{i=1}^{n} w_i(x)(\theta(X_i)-\theta(x)) =&\frac{\theta^{(\ell+1)}(x)}{(\ell+1)!}\sum_{i=1}^{n} w_i(x)(X_i-x)^{\ell+1}\\ &+\sum_{i=1}^{n} w_i(x)(X_i-x)^{\ell+1}\frac{\{\theta^{(\ell+1)}(\xi_i)-\theta^{(\ell+1)}(x)\}}{(\ell+1)!}\\ \sum_{i=1}^{n} w_i(x')(\theta(X_i)-\theta(x')) =&\frac{\theta^{(\ell+1)}(x')}{(\ell+1)!}\sum_{i=1}^{n} w_i(x')(X_i-x')^{\ell+1}\\ &+\sum_{i=1}^{n} w_i(x')(X_i-x')^{\ell+1}\frac{\{\theta^{(\ell+1)}(\zeta_i)-\theta^{(\ell+1)}(x')\}}{(\ell+1)!}.\end{aligned}$$

Regarding the third and fourth term, we have by Lemma 18 (iii)

$$\sum_{i=1}^{n} w_i(x)(X_i-x)^{\ell+1}\frac{\{\theta^{(\ell+1)}(\xi_i)-\theta^{(\ell+1)}(x)\}}{\ell!}=O(h^{\ell+1+\beta})$$

Regarding the first terms on the right-hand side, we further split into

$$\begin{aligned}&\frac{\theta^{(\ell+1)}(x)}{(\ell+1)!}\sum_{i=1}^{n} w_i(x)(X_i-x)^{\ell+1}-\frac{\theta^{(\ell+1)}(x')}{(\ell+1)!}\sum_{i=1}^{n} w_i(x')(X_i-x')^{\ell+1}\\ =&\frac{\theta^{(\ell+1)}(x)-\theta^{(\ell+1)}(x')}{(\ell+1)!}\sum_{i=1}^{n} w_i(x)(X_i-x)^{\ell+1} &(1.15)\\ &-\frac{\theta^{(\ell+1)}(x')}{(\ell+1)!}\sum_{i=1}^{n} w_i(x')\big\{(X_i-x')^{\ell+1}-(X_i-x)^{\ell+1}\big\} &(1.16)\\ &-\frac{\theta^{(\ell+1)}(x')}{(\ell+1)!}\sum_{i=1}^{n}\{w_i(x)-w_i(x')\}(X_i-x)^{\ell+1}. &(1.17)\end{aligned}$$

For (1.15), we have by Hölder continuity of the derivative and Lemma 18 (i) and (iii)

$$\frac{\theta^{(\ell+1)}(x)-\theta^{(\ell+1)}(x')}{(\ell+1)!}\sum_{i=1}^{n} w_i(x)(X_i-x)^{\ell+1}=O_p(|x-x'|^{\beta}h^{\ell+1}).$$

For (1.16), we have by the same arguments as in the proof of Lemma 18

$$\frac{\theta^{(\ell+1)}(x')}{(\ell+1)!}\sum_{i=1}^{n} w_i(x')\big\{(X_i-x')^{\ell+1}-(X_i-x)^{\ell+1}\big\} = O_p\big(|x-x'|^{\ell+1}\wedge h^{\ell+1}\big).$$

For (1.17), we have by Lemma 18 (i) and (iv)

$$\frac{\theta^{(\ell+1)}(x')}{(\ell+1)!}\sum_{i=1}^{n}\{w_i(x)-w_i(x')\}(X_i-x)^{\ell+1} = O_p(|x-x'|h^{\ell+1})$$

Hence, uniformly over x, x'

$$\hat\theta_n(x)-\theta(x)-\{\hat\theta_n(x')-\theta(x')\} = \sum_{i=1}^{n}\{w_i(x)-w_i(x')\}\varepsilon_i + O(|x-x'|^{\beta}h^{\ell+1}).$$

By discretization and conditional sub-Gaussianity of ε_i,

$$\sup_{x\neq x'}\left|\sum_{i=1}^{n}\frac{w_i(x)-w_i(x')}{|x-x'|}\varepsilon_i\right| = O_p\left(\sqrt{\log n\sup_{x\neq x'}\sum_{i=1}^{n}\mathbb{1}_i\left(\frac{w_i(x)-w_i(x')}{\|x-x'\|_{\mathcal{X}}}\right)^2}\right),$$

where $\mathbb{1}_i = \mathbb{1}(|X_i-x|\le h)+\mathbb{1}(|X_i-x'|\le h)$. This implies together with Lemma 18 (iv)

$$\sup_{x\neq x'}\left|\sum_{i=1}^{n}\frac{w_i(x)-w_i(x')}{|x-x'|}\varepsilon_i\right| = O_p\left(\sqrt{\frac{\log n}{nh^3}}\right).$$

The claim follows. □

Lemma 18. *Suppose that Assumption 11 hold. Then,*

(i) for all x so that $|X_i-x|>h$, $w_{n,i}(x)=0$ and

$$\sup_x\sum_{i=1}^{n}|w_{n,i}(x)| = O_p(1)$$

(ii) For all $x\in[0,1]$, $\sum_i w_{n,i}(x)=1$ and for $j=1,\dots,\ell$,

$$\sum_{i=1}^{n} w_{n,i}(x)(X_i-x)^j = 0.$$

(iii) For all $x \in [0,1]$,

$$\sum_{i=1}^{n} w_{n,i}(x)(X_i - x)^{\ell+1} = O_p(h^{\ell+1}).$$

(iv) The weights $w_{n,i}(x)$ *satisfy the following Lipschitz property*

$$\sup_{x \neq x'} \frac{|w_{n,i}(x) - w_{n,i}(x')|}{|x - x'|} = O_p\left(\frac{1}{nh^2}\right).$$

Proof. We work in the following on the event $\Lambda_n = \{\forall x \in [0,1], \forall v \in \mathbb{R}^{\ell+1} : \|B_n(x)v\|_2 \geq \underline{\lambda}\|v\|_2\}$. We later show that there is some $\underline{\lambda} > 0$ so that $\mathrm{P}(\Lambda_n) \to 1$.

(i): The first part directly follows since K has support $[-1,1]$. For the second part, note that on Λ_n, for all $x \in [0,1]$,

$$\begin{aligned}
|w_{n,i}(x)| \leq & \|U(0)\|_2 \|B_n(x)\|_2 \|a_{n,i}(x)\|_2 \\
\leq & \frac{C\|K\|_\infty}{nh\underline{\lambda}} \|U((X_i - x)/h)\|_2 \mathbb{1}(|X_i - x| \leq h) \\
\leq & \frac{C}{nh} \mathbb{1}(|X_i - x| \leq h),
\end{aligned}$$

where we have used that $\|U(\cdot)\|_2$ is uniformly bounded on $[-1,1]$. Now, by standard consistency results for the kernel density estimator with rectangular kernel

$$\sup_{x \in [0,1]} \sum_{i=1}^{n} |w_{n,i}(x)| \leq C \sup_{x \in [0,1]} \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}(|X_i - x| \leq h) = O_p(1).$$

(ii): Let δ_{st} denote the Kronecker-delta, i.e., $\delta_{st} = 1$ if $s = t$ and 0 else. Then, for any $j = 1, \ldots, \ell + 1$,

$$\begin{aligned}
& \frac{1}{j!} \sum_{i=1}^{n} w_{n,i}(x) \left(\frac{X_i - x}{h}\right)^j \\
= & U(0)^\intercal B_n(x)^{-1} \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) U\left(\frac{X_i - x}{h}\right) U_j\left(\frac{X_i - x}{h}\right) = \delta_{1j}.
\end{aligned}$$

(iii): (i) implies

$$\left|\sum_{i=1}^{n} w_{n,i}(x)(X_i - x)^{\ell+1}\right| = h^{\ell+1} \left|\sum_{i=1}^{n} w_{n,i}(x)(X_i - x)^{\ell+1}\right|$$

$$\leq h^{\ell+1} \sum_{i=1}^{n} |w_{n,i}(x)| \sup_{|X_i - x\| \leq h} \left| \frac{X_i - x}{h} \right|^{\ell+1} = O_p(h^{\ell+1}).$$

(iv): For any $x, x' \in [0,1]$ satisfying $|x - x'| \leq 2h$, decompose

$$w_{n,i}(x) - w_{n,i}(x') = U(0)^\intercal \{B_n(x)^{-1} - B_n(x')^{-1}\} a_{n,i}(x) \tag{1.18}$$

$$+ U(0)^\intercal B_n(x') \{a_{n,i}(x) - a_{n,i}(x')\}. \tag{1.19}$$

For any square matrix A, denote by $\|A\|_2 = \sup_{v \neq 0} \|Av\|_2 / \|v\|_2$. Using that $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ for any invertible matrices A and B, we can bound (1.18) by

$$\begin{aligned}
&|U(0)^\intercal \{B_n(x)^{-1} - B_n(x')^{-1}\} a_{n,i}(x)| \\
\leq &\|U(0)\|_2 \|B_n(x)^{-1}\|_2 \|B_n(x') - B_n(x)\|_2 \|B_n(x')^{-1}\|_2 \|a_{n,i}(x)\|_2 \\
\leq &\underline{\lambda}^{-2} \|B_n(x') - B_n(x)\|_2 \|a_{n,i}(x)\|_2.
\end{aligned}$$

It holds for some constant C

$$\|a_{n,i}(x)\|_2 \leq \frac{\|K\|_\infty}{nh} \|U\|_2 \leq \frac{C}{nh},$$

where we used that $\|U((X_i - x)/h)\|_2$ is bounded uniformly over X_i, x and h since $\|X_i - x\| \leq h$. Secondly, decompose

$$\begin{aligned}
&B_n(x') - B_n(x) \\
= &\frac{1}{nh} \sum_{i=1}^{n} \left\{ K\left(\frac{X_i - x'}{h}\right) - K\left(\frac{X_i - x}{h}\right) \right\} U\left(\frac{X_i - x'}{h}\right) U^\intercal \left(\frac{X_i - x'}{h}\right) && (1.20) \\
&+ \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \left\{ U\left(\frac{X_i - x'}{h}\right) - U\left(\frac{X_i - x}{h}\right) \right\} U^\intercal \left(\frac{X_i - x'}{h}\right) && (1.21) \\
&+ \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) U\left(\frac{X_i - x}{h}\right) \left\{ U\left(\frac{X_i - x'}{h}\right) - U\left(\frac{X_i - x}{h}\right) \right\}^\intercal && (1.22)
\end{aligned}$$

For (1.20), we bound

$$\|(1.20)\|_2 \leq \frac{C}{nh} \sum_{i=1}^{n} \left| K\left(\frac{X_i - x'}{h}\right) - K\left(\frac{X_i - x}{h}\right) \right|,$$

where we have again used that the Euclidean norm of the U vector is uniformly bounded. Decompose

$$\begin{aligned}
&\frac{1}{nh}\sum_{i=1}^{n}\left|K\left(\frac{X_i-x'}{h}\right)-K\left(\frac{X_i-x}{h}\right)\right| \\
=&\frac{1}{nh}\sum_{i=1}^{n}\left|K\left(\frac{X_i-x'}{h}\right)-K\left(\frac{X_i-x}{h}\right)\right|-\mathrm{E}\left[\left|K\left(\frac{X_i-x'}{h}\right)-K\left(\frac{X_i-x}{h}\right)\right|\right] \\
&+\frac{1}{h}\,\mathrm{E}\left[\left|K\left(\frac{X_i-x'}{h}\right)-K\left(\frac{X_i-x}{h}\right)\right|\right].
\end{aligned}$$

By standard arguments and Lipschitz continuity of K,

$$\begin{aligned}
&\sup_{x\neq x'}\left|\frac{1}{nh|x-x'|}\sum_{i=1}^{n}\left|K\left(\frac{X_i-x'}{h}\right)-K\left(\frac{X_i-x}{h}\right)\right|\right. \\
&\qquad\left.-\mathrm{E}\left[\left|K\left(\frac{X_i-x'}{h}\right)-K\left(\frac{X_i-x}{h}\right)\right|\right]\right|=O_p\left(\sqrt{\frac{\log n}{nh^2}}\right) \\
&\sup_{x\neq x'}\frac{1}{h|x-x'|}\,\mathrm{E}\left[\left|K\left(\frac{X_i-x'}{h}\right)-K\left(\frac{X_i-x}{h}\right)\right|\right]=O(h^{-1}).
\end{aligned}$$

Since $\log n/(nh)\to 0$, $\sup_{x\neq x'}\|(1.20)\|_2/|x-x'|=O_p(h^{-1})$. Regarding (1.21) and (1.22), by the binomial theorem, for any $k=0,1,\ldots,\ell$,

$$\left(\frac{X_i-x}{h}\right)^k=\sum_{s=0}^{k}\binom{k}{s}\left(\frac{X_i-x'}{h}\right)^{k-s}\left(\frac{x'-x}{h}\right)^s$$

and therefore

$$\left(\frac{X_i-x}{h}\right)^k-\left(\frac{X_i-x'}{h}\right)^k=\sum_{s=1}^{k}\binom{k}{s}\left(\frac{X_i-x'}{h}\right)^{k-s}\left(\frac{x'-x}{h}\right)^s.$$

This implies

$$\sup_{x\neq x'}\frac{1}{|x-x'|}\left\|U\left(\frac{X_i-x'}{h}\right)-U\left(\frac{X_i-x}{h}\right)\right\|_2=O(h^{-1})$$

and since the kernel density estimator is asymptotically stochastically bounded,

$$\sup_{x\neq x'}\frac{\|(1.21)\|_2}{|x-x'|}=O_p(h^{-1})\quad\text{and}\quad\sup_{x\neq x'}\frac{\|(1.22)\|_2}{|x-x'|}=O_p(h^{-1}).$$

This implies the following bound on (1.18)

$$\sup_{x\neq x'} \frac{|U(0)^\intercal\{B_n(x)^{-1} - B_n(x')^{-1}\}a_{n,i}(x)|}{|x-x'|} = O_p\left(\frac{1}{nh^2}\right).$$

For (1.19), we have on Λ_n,

$$|U(0)^\intercal B_n(x')\{a_{n,i}(x) - a_{n,i}(x')\}| \leq C\underline{\lambda}^{-1}\|a_{n,i}(x) - a_{n,u}(x')\|_2.$$

By the same arguments as above

$$\sup_{x\neq x'} \frac{\|a_{n,i}(x) - a_{n,i}(x')\|_2}{|x-x'|} = O\left(\frac{1}{nh^2}\right)$$

implying for (1.19)

$$\sup_{x\neq x'} \frac{|U(0)^\intercal B_n(x')\{a_{n,i}(x) - a_{n,i}(x')\}|}{|x-x'|} = O\left(\frac{1}{nh^2}\right).$$

Thus, we have shown that

$$\sup_{x\neq x'} \frac{|w_{n,i}(x) - w_{n,i}(x')|}{|x-x'|} = O_p\left(\frac{1}{nh^2}\right).$$

□

The penultimate process

The following result provides an approximating distribution sequence as required by Assumption 2(iii).

Lemma 19. *Suppose that Assumptions 11 hold. Then, there exist a sequence of centered Gaussian processes Z_n with covariance function Σ_n satisfying*

$$\begin{aligned}&nh_n\Sigma_n(x,x')\\ =&U(0)^\intercal S^{-1}\frac{1}{h_n}\,\mathrm{E}\left[\frac{K_h(X_i-x)}{f(x)}\frac{K_h(X_i-x')}{f(x')}\sigma^2(X_i)U\left(\frac{X_i-x}{h_n}\right)U^\intercal\left(\frac{X_i-x}{h_n}\right)\right]S^{-1}U(0),\end{aligned}$$

so that

$$\sup_{x\in[0,1]}|\{\hat{\theta}_n(x) - \theta(x)\} - Z_n(x)| = O_p\left(\left(\frac{\log^8 n}{nh_n^{4/3}}\right)^{3/4} + h_n^{\ell+1}\right).$$

Note that since the derivative is Lipschitz continuous, Lemma 19 implies Assumption 2(iii) with the same rate.

Proof. Assumptions 11 are sufficient to imply (13) in Kong et al. (2010) (see Appendix F in Chernozhukov et al. (2013b) for a proof) and imply

$$\sup_{x\in[0,1]} |\{\hat{\theta}_n(x) - \theta(x)\} - \theta_n^*(x)| = O_p\bigg(\frac{\log n}{nh_n} + h_n^{\ell+1}\bigg),$$

where

$$\theta_n^*(x) = U(0)^\intercal S_n^{-1}(x)\frac{1}{nh_n}\sum_{i=1}^n K_h(X_i - x)U((X_i - x)/h)\varepsilon_i$$

and $S_n(x) \in \mathbb{R}^{(\ell+1)\times(\ell+1)}$ with elements $S_{n,j,k}(x) = \int K(u)u^{j+k-2}f(x+hu)du$, $j,k = 1,\ldots,\ell+1$. By Lemma 8 in Kong et al. (2010),

$$\sup_{x\in[0,1]} |S_n(x) - f(x)S| = O_p\bigg(\sqrt{\frac{\log n}{nh_n}} + h_n\bigg),$$

where $S \in \mathbb{R}^{(\ell+1)\times(\ell+1)}$ with elements $S_{j,k} = \int K(u)u^{j+k-2}du$, $j,k = 1,\ldots,\ell+1$. Thus, by standard arguments, for $\mathbf{K}(x) = U(0)^\intercal S^{-1}K(x)U(x)$ and $\mathbf{K}_h(x) = \mathbf{K}(x/h)$,

$$\sup_{x\in[0,1]}\bigg|\{\hat{\theta}_n(x) - \theta(x)\} - \frac{1}{nh}\sum_{i=1}^n \frac{\mathbf{K}_h(X_i - x)}{f(x)}\varepsilon_i\bigg| = O_p\bigg(\frac{\log n}{nh_n} + h_n^{\ell+1}\bigg).$$

Further, let

$$g_x(\varepsilon, X) = \frac{1}{\sqrt{h_n}}\frac{\mathbf{K}_h(X - x)}{f(x)}\varepsilon, \qquad \forall x \in [0,1].$$

Then, by Theorem 8 in Appendix G in Chernozhukov et al. (2013b), there exists a Brownian Bridge B_n on $\{g_x : x \in [0,1]\}$ such that for all $t \geq C\log n$ for some constant C

$$\mathrm{P}\bigg(\sup_{x\in[0,1]}\bigg|\frac{1}{\sqrt{nh_n}}\sum_{i=1}^n \frac{\mathbf{K}_h(X_i - x)}{f(x)}\varepsilon_i - B_n(g_x)\bigg| \geq C\bigg[\sqrt{\frac{t}{\sqrt{nh_n^2}}} + t\sqrt{\frac{\log n}{nh_n}}\bigg]\bigg) \leq e^{-t},$$

or equivalently, for $Z_n(x) = B_n(g_x)/\sqrt{nh_n}$

$$\mathrm{P}\left(\sup_{x\in[0,1]} \left| \frac{1}{nh_n} \sum_{i=1}^{n} \frac{\mathbf{K}_h(X_i - x)}{f(x)} \varepsilon_i - Z_n(x) \right| \geq C \left[\sqrt{\frac{t}{(nh_n^{4/3})^{3/2}}} + t\frac{\sqrt{\log n}}{nh_n} \right] \right) \leq e^{-t}.$$

Taking $t_n = C \log^2 n$ for some sufficiently large C, implies

$$\sup_{x\in[0,1]} \left| \frac{1}{nh_n} \sum_{i=1}^{n} \frac{\mathbf{K}_h(X_i - x)}{f(x)} \varepsilon_i - Z_n(x) \right| = O_p\left(\left(\frac{\log^8 n}{nh_n^{4/3}} \right)^{3/4} \right).$$

Together with the bound on the linearization error of the local polynomial estimator above, we have

$$\sup_{x\in[0,1]} |\{\hat{\theta}_n(x) - \theta(x)\} - Z_n(x)| = O_p\left(\left(\frac{\log^8 n}{nh_n^{4/3}} \right)^{3/4} + h_n^{\ell+1} \right).$$

The claim follows. □

Lemma 20. *Z_n can be chosen to have almost surely Lipschitz continuous sample paths and there is a version of Z_n such that $Z_n \in \mathbb{D}$.*

Proof. Consider the integral operator $\mathcal{K}$ related to the covariance function Σ_n of Z_n given by

$$\begin{aligned} f \mapsto (\mathcal{K}f)(\cdot) &= \int_0^1 \Sigma_n(x, \cdot) f(x) dx \\ &= \mathrm{E}\left[\frac{\mathbf{K}_h(X_i - \cdot)}{f_X(\cdot)} \sigma^2(X_i) \int \frac{1}{h_n} \frac{\mathbf{K}_h(X_i - x)}{f_X(x)} f(x) dx \right]. \end{aligned}$$

By properties of the convolution of two functions and smoothness properties of $\mathbf{K}$ and f_X, $\mathcal{K}f$ is Lipschitz continuous with Lipschitz constant proportional to $\|f\|_{L_2}/h_n$. Therefore, all of its eigenfunctions are Lipschitz continuous with at most the above Lipschitz constant.

Further, since

$$\int \Sigma_n(x, x) dx < \infty,$$

$\mathcal{K}$ is trace class and hence the square root of its eigenvalues are absolutely summable. Now, denote by $\lambda_1 \geq \lambda_2 \geq \ldots$ the eigenvalues of $\mathcal{K}$ and by $e_1, e_2, \ldots$ the corresponding orthonormal system of eigenfunctions. By the Karhunen-Loève Theorem, we can

represent Z_n as

$$Z_n(x) = \sum_{j=1}^{\infty} \eta_j \sqrt{\lambda_j} e_j(x),$$

for some sequence (η_j) of i.i.d. standard normal random variables. Here, the convergence is in the mean square sense. Further, Z_n is a.s. absolutely summable. Indeed, since

$$|Z_n(x)| \leq \sum_{j=1}^{\infty} \sqrt{\lambda_j}|\eta_j||e_j(x)| \leq \sqrt{\sum_{j=1}^{\infty} \sqrt{\lambda_j}\eta_j^2} \sqrt{\sum_{j=1}^{\infty} \sqrt{\lambda_j}|e_j(x)|^2}$$

and since the eigenfunctions are uniformly bounded and $(\sqrt{\lambda_j})$ absolutely summable, the term on the right-hand side is a.s. finite and hence $Z_n(x)$ is a.s. absolutely summable. We have, for any x, x',

$$\begin{aligned} |Z_n(x) - Z_n(x')| \leq & \sqrt{\sum_{j=1}^{\infty} \sqrt{\lambda_j}\eta_j^2} \sqrt{\sum_{j=1}^{\infty} \sqrt{\lambda_j}(e_j(x) - e_j(x'))^2} \\ \leq & \frac{L}{h_n} \sqrt{\sum_{j=1}^{\infty} \sqrt{\lambda_j}\eta_j^2} \sqrt{\sum_{j=1}^{\infty} \sqrt{\lambda_j}}|x - x'|. \end{aligned}$$

Again, since $(\sqrt{\lambda_j})$ is absolutely summable,

$$\mathrm{E}\left[\sqrt{\sum_{j=1}^{\infty} \sqrt{\lambda_j}\eta_j^2}\right] \leq \sum_{j=1}^{\infty} \sqrt{\lambda_j} < \infty$$

and hence, the representations Z_n are almost surely Lipschitz continuous.

The above construction shows that any Gaussian process with the covariance function Σ_n has a version with a.s. Lipschitz continuous sample paths. In order to show that the Z_n constructed using the Rio-Massart coupling has such a version, we slightly modify the Karhunen-Loève construction from above. Let

$$\eta_j = \int_0^1 Z_n(x) e_j(x) dx.$$

Then, the resulting process (which we now call $Y(x)$ instead) satisfies $\mathrm{Var}(Z_n(x) -$

$Y(x)) = 0$ (by mean square convergence, i.e. the same projection argument as in the proof of the Karhunen-Loève theorem). Finally, note that

$$\mathrm{P}(\exists x \in \mathbb{Q} \cap [0,1] : Z_n(x) \neq Y(x)) \leq \sum_{x \in \mathbb{Q} \cap [0,1]} \mathrm{P}(Z_n(x) \neq Y(x)) = 0,$$

and therefore we can identify Z_n and Y almost surely. □

Lemma 21. *Suppose that Assumptions 11 hold. Then for the Gaussian process Z_n given in Lemma 19, it holds for any continuous θ_0*

$$\frac{1}{\sqrt{\mathrm{Var}(\phi'_{\theta_0}(Z_n))}} = O\big(\sqrt{nh_n}\big).$$

Proof. Since ϕ'_{θ_0} is a sublinear and Lipschitz continuous functional and Z_n a centered Gaussian process with almost surely continuous sample paths, Lemma 3 in Scherer (2024) applies. We distinguish two cases. If the argmax set of θ_0 is a singleton, say $\Psi(\theta_0) = x^*$, then

$$\mathrm{Var}(\phi'_{\theta_0}(Z_n)) = \Sigma(x^*, x^*) = \frac{1}{nh_n} \frac{\sigma^2(x^*)}{f(x^*)} \int K(u)^2 du (1 + o(1))$$

and the claim follows for this case. Next, suppose that the argmax set of θ_0 contains at least two elements. By the same arguments as above, we can lower bound $\bar{\sigma}^2$ to be of order $(nh_n)^{-1}$. Moreover, by standard arguments, $\mathrm{E}[\phi'_{\theta_0}(Z_n)] \leq \sqrt{\log n}$. Hence by Lemma 3, in both cases the variance is lower bounded of order $(nh_n)^{-1}$ and the claim follows. □

Application of the Delta method

Lemma 22. *Suppose that Assumptions 11 hold. Then*

$$\lim_{n \to \infty} \sup_{t \in \mathbb{R}} |\mathrm{P}(\phi(\hat{\theta}_n) - \phi(\theta_0) \leq t) - \mathrm{P}(\phi'_{\theta_0}(Z_n) \leq t)| = 0.$$

Proof. By Lemma 17

$$\|\hat{\theta}_n - \theta_0\|_{BL} = O_p\left(\sqrt{\frac{\log n}{nh_n^3}} + h_n^{\ell+1}\right)$$

and therefore by our Taylor result

$$\phi(\hat{\theta}_n) - \phi(\theta_0) = \phi'_{\theta_0}(\hat{\theta}_n - \theta_0) + O_p\bigg(\frac{\log n}{nh_n^3} + h_n^{2(\ell+1)}\bigg).$$

Further, by Lemma 19,

$$\sup_{x\in[0,1]} |\{\hat{\theta}_n(x) - \theta(x)\} - Z_n(x)| = O_p\bigg(\bigg(\frac{\log^8 n}{nh_n^{4/3}}\bigg)^{3/4} + h_n^{\ell+1}\bigg)$$

and by Lemma 21

$$\frac{1}{\sqrt{\mathrm{Var}(\phi'_{\theta_0}(Z_n))}} = O\big(\sqrt{nh_n}\big).$$

Further, since Z_n has mean zero and $\mathrm{Var}(Z_n(x)) > 0$ for all $x \in [0,1]$, the claim follows by Theorem 2 since

$$\begin{aligned}
\frac{a_n^{-\gamma}}{\sqrt{\mathrm{Var}(\phi'_{\theta_0}(Z_n))}} =& O\bigg(\sqrt{nh_n}\bigg\{\frac{\log n}{nh_n^3} + h_n^{2(\ell+1)}\bigg\}\bigg) \\
=& O\bigg(\frac{\log n}{\sqrt{nh_n^5}} + \sqrt{nh_n^{4\ell+3}}\bigg) \\
\frac{r_n^{-1}}{\sqrt{\mathrm{Var}(\phi'_{\theta_0}(Z_n))}} =& O\bigg(\sqrt{nh_n}\bigg\{\bigg(\frac{\log^8 n}{nh_n^{4/3}}\bigg)^{3/4} + h_n^{\ell+1}\bigg\}\bigg) \\
=& O\bigg(\frac{\log^6 n}{(nh_n^2)^{1/4}} + \sqrt{nh_n^{2\ell+3}}\bigg)
\end{aligned}$$

and by the assumed rate requirements in Assumption 11 (vi). □

Consistency of the bootstrap for the preliminary estimator

Lemma 23. *Suppose that Assumptions 11 hold. Then*

$$\|\hat{Z}_n\|_{BL} = O_p\bigg(\sqrt{\frac{\log n}{nh_n^3}}\bigg).$$

Proof. Let

$$\tilde{w}_i(x) = \frac{1}{nh_n\hat{f}_n(x)}\mathbf{K}_h(X_i - x)\hat{\varepsilon}_i.$$

Note that the equivalent kernel $\mathbf{K}$ is an $(\ell+1)$th order kernel with support $[-1,1]$. Moreover, it is Lipschitz continuous. Further, by standard arguments, $\hat{f}_n$ is Lipschitz continuous and the event $A_n = \{\min_x \hat{f}_n(x) \geq c\}$ converges in probability to one for sufficiently small constant $c > 0$. Moreover, the event $B_n = \{\max_i|\hat{\varepsilon}_i| \leq b\}$ converges in probability to one for a sufficiently large constant b, since Y_i is bounded and $\|\hat{\theta}_n - \theta_0\|_\infty \to 0$. Therefore, on $A_n \cap B_n$, and for $|x - x'| \leq h_n$

$$\begin{aligned}
|\tilde{w}_i(x) - \tilde{w}_i(x')| \leq& \frac{b}{nh_n\hat{f}_n(x)}|\mathbf{K}_h(X_i - x) - \mathbf{K}_h(X_i - x')| \\
&+ \frac{b}{nh_n}\left|\frac{1}{\hat{f}_n(x)} - \frac{1}{\hat{f}_n(x')}\right||\mathbf{K}_h(X_i - x')| \\
\leq& \frac{C}{nh_n^2}|x - x'|\mathbb{1}(|X_i - x| \leq h_n)\mathbb{1}(|X_i - x'| \leq h_n),
\end{aligned}$$

where C denotes some constant that potentially changes in every occasion. Further, we have used in the second inequality that the distance between $\hat{f}_n(x)^{-1}$ and $\hat{f}_n(x')^{-1}$ can be upper bounded by the distance between $\hat{f}_n(x)$ and $\hat{f}_n(x')$ on $A_n \cap B_n$. This implies

$$\sup_{|x-x'|\leq h_n} \frac{|\tilde{w}_i(x) - \tilde{w_i(x')}|}{|x - x'|} \leq \frac{C}{nh_n}\mathbb{1}(\|X_i - x\| \leq h_n)\mathbb{1}(\|X_i - x'| \leq h_n),$$

and therefore, since $A_n \cap B_n$ with probability converging to one, we have

$$\sup_{|x-x'|\leq h_n} |\hat{Z}_n(x) - \hat{Z}_n(x')| = \sup_{|x-x'|\leq h_n}\left|\sum_{i=1}^n \frac{\tilde{w}_i(x) - \tilde{w}_i(x')}{|x - x'|}\eta_i\right| = O_p\left(\sqrt{\frac{\log n}{nh_n^3}}\right).$$

Next, for x, x' such that $|x - x'| > h_n$, we have

$$\sup_{|x-x'|>h_n} |\hat{Z}_n(x) - \hat{Z}_n(x')| \leq 2\frac{\|\hat{Z}_n\|_\infty}{h_n} = O_p\left(\sqrt{\frac{\log n}{nh_n^3}}\right),$$

where we have used that $\|\hat{Z}_n\|_\infty = O_p(\sqrt{\log n/(nh_n)})$. The claim follows. □

The following Theorem relies on the coupling constructed in Theorem 9 in Chernozhukov et al. (2013b). As the authors of this paper, we assume a sufficiently rich probability space which contains a uniformly distributed random variable which is independent of the data. For details, consider the discussion in before Theorem 9 and in the Appendix A in Chernozhukov et al. (2013b).

Lemma 24. *Suppose that Assumptions 11 hold. Then,*

$$\sup_{t\in\mathbb{R}}\left|\mathrm{P}\left(\sup_{x\in[0,1]}\left|\frac{\hat{Z}_n(x)}{\mathrm{std}(\hat{Z}_n(x)\,|\,\mathcal{D}_n)}\right|\leq t\,\middle|\,\mathcal{D}_n\right)-\mathrm{P}\left(\sup_{x\in[0,1]}\left|\frac{\hat{\theta}_n(x)-\theta(x)}{\mathrm{std}(\hat{\theta}_n(x))}\right|\leq t\right)\right|=o_p(1),$$

where $\mathrm{std}(\hat{Z}_n(x)\,|\,\mathcal{D}_n)$ *denotes the standard deviation of* $\hat{Z}_n(x)$ *conditionally on the data, and*

$$\sup_{t\in\mathbb{R}}|\mathrm{P}(\phi'_{\theta_0}(\hat{Z}_n)\leq t\,|\,\mathcal{D}_n)-\mathrm{P}(\phi(\hat{\theta}_n)-\phi(\theta_0)\leq t)|=o_p(1).$$

In particular, this Lemma implies by Corollary 1 that the bootstrapped quantile

$$\hat{q}_{n,1-\alpha}=\inf\left\{q\in\mathbb{R}\,:\,\mathrm{P}\left(\sup_{x\in[0,1]}\left|\frac{\hat{Z}_n(x)}{\mathrm{std}(\hat{Z}_n(x)\,|\,\mathcal{D}_n)}\right|\leq t\,\middle|\,\mathcal{D}_n\right)\geq 1-\alpha\right\}$$

satisfies

$$\mathrm{P}\left(\sup_{x\in[0,1]}\left|\frac{\hat{\theta}_n(x)-\theta(x)}{\mathrm{std}(\hat{\theta}_n(x))}\right|\leq\hat{q}_{n,1-\alpha}\right)\geq 1-\alpha+\varepsilon_n$$

for some $\varepsilon_n\downarrow 0$ uniformly over $\alpha\in(0,\bar{\alpha})$ for some $\bar{\alpha}<1/2$.

Proof. By Theorem 9 in Appendix H.4 in Chernozhukov et al. (2013b) and standard arguments[25], there exists an identical copy Z_n^* of Z_n, independent of the data $\mathcal{D}_n$ such that

$$\mathrm{P}\left(\sup_{x\in[0,1]}|\hat{Z}_n(x)-Z_n^*(x)|>o\left(\frac{1}{\sqrt{nh_n\log n}}\right)\middle|\mathcal{D}_n\right)=o_p\left(\frac{1}{\log n}\right).$$

[25]Theorem 9 implies a coupling for an unfeasible multiplier bootstrap which relies on the unknown marginal density f and residuals ε_i. It remains to show that the multiplier bootstrap which uses estimates of these also satisfies this coupling bound. This follows by standard arguments on the estimation properties of the marginal density and residuals and therefore is omitted.

By Lipschitz continuity of the derivative ϕ'_{θ_0}, this implies

$$\mathrm{P}\left(|\phi'_{\theta_0}(\hat{Z}_n) - \phi'_{\theta_0}(Z_n^*)| > o\left(\frac{1}{\sqrt{nh_n \log n}}\right) \middle| \mathcal{D}_n\right) = o_p\left(\frac{1}{\log n}\right).$$

By Le Cam's Lemma,

$$\begin{aligned}\Delta_n =& \sup_{t\in\mathbb{R}} |\mathrm{P}(\phi'_{\theta_0}(\hat{Z}_n) \leq t \,|\, \mathcal{D}_n) - \mathrm{P}(\phi(\hat{\theta}_n) - \phi(\theta_0) \leq t)| \\ \leq& \inf_{\eta>0} \left\{ \mathrm{P}\left(|\phi'_{\theta_0}(\hat{Z}_n) - \phi'_{\theta_0}(Z_n^*)| > \eta | \mathcal{D}_n\right) + \sup_{t\in\mathbb{R}} \mathrm{P}(t \leq \phi'_{\theta_0}(Z_n) \leq t + \eta) \right\}\end{aligned}$$

By choosing $\eta = \varepsilon \sqrt{\mathrm{Var}(\phi'_{\theta_0}(Z_n))}$, for $\varepsilon > 0$, the first term on the right-hand side converges to zero with probability approaching one and the second term is smaller than a constant multiple of ε with probability one. Thus, for any $\varepsilon > 0$, $\mathrm{P}(\Delta_n > \varepsilon) \to 0$.

It remains to show the first bound. Let $s_n(x) = \mathrm{std}(\hat{\theta}_n(x))$, $\sigma_n(x) = \mathrm{std}(Z_n(x))$ and $\hat{\sigma}_n(x) = \mathrm{std}(\hat{Z}_n(x))$. By standard arguments,

$$\sup_{x\in[0,1]} \left| \frac{\sigma_n(x)}{s_n(x)} - 1 \right| = o(1) \qquad \text{and} \qquad \sup_{x\in[0,1]} \left| \frac{\sigma_n(x)}{\hat{\sigma}_n(x)} - 1 \right| = o_p(1).$$

Further, $\sqrt{nh_n}\sigma_n(x)$ is bounded away from zero uniformly over $x \in [0,1]$. This implies

$$\sup_{x\in[0,1]} \left| \frac{\hat{\theta}_n(x) - \theta_0(x)}{s_n(x)} - \frac{Z_n(x)}{\sigma_n(x)} \right| \leq \frac{1}{\inf_x \sigma_n(x)} \|\{\hat{\theta}_n - \theta_0\} - Z_n\|_\infty (1 + o_p(1))$$

and therefore Lemma 19 and Lemma 21 imply

$$\begin{aligned}\sup_{x\in[0,1]} \left| \frac{\hat{\theta}_n(x) - \theta_0(x)}{s_n(x)} \right| - \sup_{x\in[0,1]} \left| \frac{Z_n(x)}{\sigma_n(x)} \right| &= O_p\left(\frac{\log n}{(nh_n^2)^{1/4}}\right) \\ \frac{1}{\sqrt{\mathrm{Var}(\sup_x |Z_n(x)/\sigma_n(x)|)}} &\leq \log n.\end{aligned}$$

By the same arguments as above, this implies

$$\sup_{t\in\mathbb{R}} \left| \mathrm{P}\left(\sup_{x\in[0,1]} \left| \frac{\hat{Z}_n(x)}{\mathrm{std}(\hat{Z}_n(x) \,|\, \mathcal{D}_n)} \right| \leq t \,\middle|\, \mathcal{D}_n \right) - \mathrm{P}\left(\sup_{x\in[0,1]} \left| \frac{\hat{\theta}_n(x) - \theta(x)}{\mathrm{std}(\hat{\theta}_n(x))} \right| \leq t \right) \right| = o_p(1).$$

□

Consistency of the derivative estimator

Lemma 25. *Suppose that Assumptions 11 and the well-separatedness condition (1.4) hold. Further suppose that* $1 - \gamma_n = O(1/\log n)$. *Then,*

$$\sup_{\|h\|_{BL}\leq 1} |\hat{\phi}'_n(h) - \phi'_{\theta_0}(h)| = O_p\left(\left(\frac{\log n}{nh_n}\right)^{\frac{1}{2\rho}}\right).$$

Proof. We will show that $\hat{\Psi}_n$ is a consistent estimator of the argmax set of θ_0, Ψ_0, in the sense that both $\Psi_0 \subset \hat{\Psi}_n \subset \Psi_n$ with probability converging to one, for some set Ψ_n which converges to Ψ_0. In order to see this, note that, on the event $A_n = \{\|(\hat{\theta}_n - \theta_0)/\sigma_n\|_\infty \leq \hat{q}_{\gamma_n}\}$,

$$\begin{aligned}
&\hat{\theta}_n(x) + \hat{q}_{\gamma_n}\hat{\sigma}_n(x) - \phi(\hat{\theta}_n - \hat{q}_{\gamma_n}\hat{\sigma}_n) \\
=&\theta_0(x) + \big(\hat{\theta}_n(x) - \theta_0(x) + \hat{q}_{\gamma_n}\hat{\sigma}_n(x)\big) - \max_{x'}\big\{\theta_0(x') + [\hat{\theta}_n(x') - \theta_0(x') - \hat{q}_{\gamma_n}\hat{\sigma}_n(x')]\big\} \\
\geq&\theta_0(x) + \big(\hat{\theta}_n(x) - \theta_0(x) + \hat{q}_{\gamma_n}\hat{\sigma}_n(x)\big) - \phi(\theta_0) - \max_{x'}\big\{\hat{\theta}_n(x') - \theta_0(x') - \hat{q}_{\gamma_n}\hat{\sigma}_n(x')\big\} \\
\geq&\theta_0(x) - \phi(\theta_0)
\end{aligned}$$

and therefore $\Psi_0 = \{x : \theta_0(x) \geq \phi(\theta_0)\} \subset \hat{\Psi}_n$. For the other direction, note that by similar arguments, on A_n,

$$\hat{\theta}_n(x) + \hat{q}_{\gamma_n}\hat{\sigma}_n(x) - \phi(\hat{\theta}_n - \hat{q}_{\gamma_n}\hat{\sigma}_n) \leq \theta_0(x) + 2\hat{q}_{\gamma_n}\hat{\sigma}_n(x) - \phi(\theta_0)$$

and therefore $\hat{\Psi}_n \subset \{x : \theta_0(x) \geq \phi(\theta_0) - 2\hat{q}_{\gamma_n}\hat{\sigma}_n(x)\} =: \Psi_n$.

Consider the event $B_n = \{\hat{q}_{\gamma_n}\|\hat{\sigma}_n\|_\infty < \delta\}$ for the δ given in the well-separatedness condition on θ. Since $\hat{q}_{\gamma_n}\|\hat{\sigma}_n\|_\infty = o_p(1)$, $\mathrm{P}(B_n) \to 1$. On B_n, we have by the well-separatedness condition, for all $x \in \Psi_n$,

$$\begin{aligned}
cd(x, \Psi_0)^\rho &\leq \hat{q}_{\gamma_n}\|\hat{\sigma}_n\|_\infty \\
d(x, \Psi_0) &\leq \frac{(\hat{q}_{\gamma_n}\|\hat{\sigma}_n\|_\infty)^{1/\rho}}{c} =: \varepsilon_n,
\end{aligned}$$

implying $\hat{\Psi}_n \subset \Psi_0^{\varepsilon_n}$. Therefore, on $A_n \cap B_n$, for any h with $\|h\|_{BL} \leq 1$

$$|\hat{\phi}'_n(h) - \phi'_{\theta_0}(h)| \leq \sup_{|x-x'|\leq\varepsilon_n} |h(x) - h(x')| \leq \|h\|_{BL}\varepsilon_n.$$

Since both $\mathrm{P}(A_n) \to 1$ and $\mathrm{P}(B_n) \to 1$, it holds

$$\sup_{\|h\|_{BL}\leq 1} |\hat{\phi}'_n(h) - \phi'_{\theta_0}(h)| = \left(\frac{\log n}{nh_n}\right)^{\frac{1}{2\rho}},$$

where the last statement follows since $\hat{q}_{\gamma_n} = O_p(\sqrt{\log n} + \sqrt{-\log(1-\gamma_n)})$, $\|\hat{\sigma}_n\|_\infty = O_p((nh_n)^{-1/2})$ by standard arguments and $1 - \gamma_n = O(1/\log n)$. □

Consistency of the bootstrap for the plug-in estimator

Lemma 26. *Suppose that Assumption 11 and (1.4) hold for $\rho = 2$. Then,*

$$\sup_{t\in\mathbb{R}} \left|\mathrm{P}\left(\hat{\phi}'_n(\hat{Z}_n) \leq t \,\middle|\, \mathcal{D}_n\right) - \mathrm{P}\left(\phi(\hat{\theta}_n) - \phi(\theta_0) \leq t\right)\right| = o_p(1).$$

Proof. We apply Theorem 4. Assumptions 3 follows by Lemma 24 and Assumption 4 by Lemma 25. It only remains to check the further rate requirements. We have

$$\frac{d_n\eta_n}{\sigma_n} = \sqrt{nh_n}\sqrt{\frac{\log n}{nh_n^3}}\left(\frac{\log n}{nh_n}\right)^{1/4} = \left(\frac{\log^3 n}{nh_n^5}\right)^{1/4} \to 0$$
$$\frac{s_n}{\sigma_n} = \sqrt{nh_n}\frac{1}{\sqrt{nh_n \log n}} = \frac{1}{\sqrt{\log n}} \to 0$$

and the claim follows by Theorem 4. □

1.B.2 Bargaining bounds

Differentiability:

The differentiability properties in this example are qualitatively similar to Example 1. The derivative is given by

$$\phi'_{F_{Y|X},f_X}(h,g) = \phi'_{F_{Y|X}}(h) + \phi'_{f_X}(g)$$

where

$$\phi'_{F_{Y|X}}(h)(y) = \int \max_{x' \in \Psi(y,x)} h(y,x') f_X(x) dx,$$
$$\phi'_{f_X}(\theta, g)(y) = \int \max_{x' \geq x} F_{Y|X}(y,x') g(x) dx$$

and $\Psi(y,x) = \text{argmax}_{x' \geq x} F_{Y|X}(y,x')$. Under a similar well-separatedness condition given in Section 1.B.2 below, we can show that ϕ is 2-Fréchet directionally differentiable with respect to the norm $\|h\|_{BL} \vee \|g\|_\infty$, where

$$\|h\|_{BL} = \sup_y \left\{ \sup_x |h(y,x)| \vee \sup_{x \neq x'} \frac{|h(y,x) - h(y,x')|}{|x - x'|} \right\}.$$

The well-separatedness condition allows for many qualitatively different conditional distribution functions. As we show in Lemma 27, the well-separatedness condition is for example satisfied when $x \mapsto F_{Y|X}(y,x)$ is strictly increasing or decreasing with a lower bound on the derivative. In the increasing case, the derivative reduces to a point-evaluation functional

$$\phi'_{F_{Y|X}}(h)(y) = h(y,1),$$

while in the decreasing case

$$\phi'_{F_{Y|X}}(h)(y) = \int h(y,x) f_X(x) dx$$

the derivative averages over $h(y,\cdot)$. Another example of interest is when Y and X are independent so that the conditional distribution function $x \mapsto F_{Y|X}(y,x)$ is flat for all y. In this case, the derivative reduces to ϕ

$$\phi'_{F_{Y|X}}(h)(y) = \int \max_{x' \geq x} h(y,x') f_X(x) dx.$$

Finally, the well-separatedness condition can be shown to hold in location scale models $Y = m(X) + s(X)\varepsilon$, with $X \perp\!\!\!\perp \varepsilon$, under suitable bounds on the derivatives of the conditional mean m and conditional standard deviation function s.

Application of the Delta Method:

Consider the Nadaraya-Watson estimator of the cumulative distribution function of

Y_i conditional on X_i

$$\hat{F}_n(y \,|\, x) = \frac{1}{\hat{f}_X(x)} \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n}\right) \mathbb{1}(Y_i \leq y)$$

where $\hat{f}_X$ denotes the KDE of f_X, the marginal density of X_i. We consider two separate strong approximations of this estimator. First in Section 1.B.2, when the derivative is nonlinear, we use the Rio-Massart coupling to approximate the whole process $(x, y) \mapsto (\hat{F}_n(y, x) - F_{Y|X}(y, x), \hat{f}_n(x) - f_X(x))$. This yields a centered Gaussian process $Z_n = (Z_{F,n}, Z_{f,n})$ with continuous sample paths such that

$$\begin{aligned} &\|(\hat{F}_n - F_{Y|X}) - Z_{F,n}\|_\infty \vee \|(\hat{f}_n - f_X) - Z_{f,n}\|_\infty \\ =&O_p\left(\left(\frac{\log n}{n^3 h_n^4}\right)^{1/4} + \frac{\sqrt{\log^3 n}}{nh_n} + \sqrt{\frac{h_n \log n}{n}} + h_n^\ell\right). \end{aligned}$$

Second in Section 1.B.2, when the derivative is linear, we directly couple the derivative evaluated at the preliminary estimator to obtain a Gaussian process Z_n with the same distribution as above satisfying

$$\sup_y |\phi'_{\theta_0}(\hat{F}_n - F_{Y|X}, \hat{f}_n - f_X) - \phi'_{\theta_0}(Z_{F,n}, Z_{f,n})| = O_p(b_n),$$

for some b_n which depends on whether the derivative is mean-square continuous or not. Moreover, we can show that

$$\|\hat{F}_n - F_{Y|X}\|_{BL} = O_p\left(\sqrt{\frac{\log n}{nh_n^3}} + h_n^\ell\right),$$

where ℓ quantifies the smoothness of $F_{Y|X}$. Using these results, we apply in Section 1.B.2 both Delta methods Theorem 1 and 2. As Theorem 2 is directed towards univariate functionals, we apply it here to the point-evaluation functionals for $y \in \mathbb{R}$.[26] Theorem 2 implies for both of the above couplings

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}} \left|\mathrm{P}(\phi(\hat{\theta}_n)(y) - \phi(\theta_0)(y) \leq t) - \mathrm{P}\big(\phi'_{\theta_0}(Z_n)(y) \leq t\big)\right| = 0, \qquad \forall y \in [0, 1]$$

under the undersmoothing condition $\sqrt{nh_n^{2\ell}} \to 0$ and $nh_n^6 \to \infty$ when the deriva-

[26] We show in Lemma 29 in the Appendix that these point-evaluations are sublinear and Lipschitz continuous as well as the supremum over the point-evaluations.

tive is linear and mean-square continuous and under the undersmoothing condition $\sqrt{nh_n^{2\ell+1}} \to 0$ and $nh_n^5 \to \infty$ in the other cases. To the best of our knowledge, these conditions are the only ones currently available. Moreover, other methods of analysis as for example in Chernozhukov et al. (2013b), Andrews and Shi (2014), Firpo et al. (2019) and Semenova (2023) cannot be directly applied here.

We want to stress that this distributional approximation holds regardless of the underlying conditional distribution function as long as the rate requirements and the well-separatedness condition hold. This is a demanding property as the approximate distribution changes considerably under different conditional distribution functions. For example, when $x \mapsto F_{Y|X}(y,x)$ is increasing, we have $\phi_F'(Z_n)(y) = Z_{1,n}(y,1)$, while in the decreasing case, $\phi_F'(Z_n)(y) = \int Z_{1,n}(y,x) f_X(x) dx$. In the first case, the approximate distribution converges at the usual nonparametric rate and at the parametric rate in the second case. Moreover, if for example $Y \perp\!\!\!\perp X$, $\phi_F'(Z_n) = \phi(Z_{1,n}, f_X)(y)$ in which case it is hard to derive a precise rate of convergence.

The rate requirements on the bandwidth in the case when the derivative is linear and mean square continuous is restrictive. Since ϕ is even 2-Fréchet directionally differentiable with respect to the sup-norm instead of the bounded Lipschitz norm, we can also apply Theorem 2 with respect to the sup-norm. In this case,

$$a_n = \sqrt{\frac{\log n}{nh_n}}, \quad b_n = \left(\frac{\log^4 n}{n^3}\right)^{1/4} + \sqrt{\frac{\log^5 n}{n^2}} + \sqrt{\frac{h_n \log n}{n}} + h_n^\ell$$

and

$$\sigma_n^{-1} = O(\sqrt{n}).$$

Theorem 2 now requires

$$\frac{\log^2 n}{nh_n^2} \to 0 \quad \text{and} \quad nh_n^{2\ell} \to 0$$

improving over the aforementioned rate requirements. This illustrates the dependence of our Delta method results with respect to the chosen norm on $\mathbb{D}$.

Consistency Bootstrap:
We use the following Gaussian multiplier bootstrap in order to estimate the distribution of the Nadaraya-Watson conditional distribution estimator and the kernel density estimator. Let η_i, $i = 1, \ldots, n$, be i.i.d. standard normally distributed ran-

dom variables which are independent of the data $\mathcal{D}_n = \{(Y_i, X_i) : i = 1, \dots, n\}$ and set

$$\hat{Z}_{n,F}(x,y) = \frac{1}{nh_n\hat{f}_n(x)} \sum_{i=1}^{n} \eta_i K_h(X_i - x)\big(\mathbb{1}\{Y_i \le y\} - \hat{F}_{Y|X,n}(y,x)\big)$$

$$\hat{Z}_{n,f}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} \eta_i \big(K_h(X_i - x) - \hat{f}_n(x)\big).$$

We show in Lemma 41 that Assumption 3 holds with $d_n = a_n$, $s_n = o((nh_n \log n)^{-1/2})$ in the nonlinear derivative case and $s_n = o((n \log n)^{-1/2})$ when the derivative is linear and mean square continuous. This Gaussian multiplier

As an estimator of the derivative we choose

$$\hat{\phi}'_{n,F}(h)(y) = \int \max_{x' \in \hat{\Psi}_n(y,x)} h(y,x')\hat{f}_n(x)dx$$

$$\hat{\phi}'_{n,f}(g)(y) = \int \max_{x' \ge x} \hat{F}_n(y,x')g(x)dx$$

where

$$\hat{\Psi}_n(y,x) = \{x' \ge x \,|\, \hat{F}_n(y,x') + \hat{z}_n\hat{\sigma}_n(y,x') \ge \max_{\tilde{x} \ge x} \hat{F}_n(y,\tilde{x}) - \hat{z}_n\hat{\sigma}_n(y,\tilde{x})\}.$$

Here, $\hat{\sigma}_n(y,x)$ denotes an estimator of the standard deviation of $\hat{F}_n(y,x) - F_{Y|X}(y,x)$ and $\hat{z}_n$ is an estimator of the β_n quantile of $\sup_{y,x}|\hat{F_{Y|X}}_n(y,x) - F_{Y|X}(y,x)|/\hat{\sigma}_n(y,x)$ with $\beta_n \to 1$ sufficiently slowly. In our simulations, we took $\beta_n = 1 - 0.01/\log n$. We show in Lemma 40 in the Appendix, that this estimator satisfies Assumption 4 with $\eta_n = (\log n/(nh_n))^{1/4}$. Hence, this bootstrap of the plug-in estimator satisfies our general bootstrap assumptions and Theorem 4 applies under the rate requirement $\log^5 n/(nh_n^5) \to 0$ when the derivative is nonlinear or linear but mean square discontinuous and $\log^3 n/(nh_n^7) \to 0$ when the derivative is linear and mean square continuous. In the nonlinear case, these are similar requirement as needed for the Delta method in Theorem 2 to hold but in the linear and mean square continuous case, we require here a larger bandwidth. The latter is again due to the chosen norm on $\mathbb{D}$. If we endow $\mathbb{D}$ with the sup-norm instead, we can improve the rate requirements for the bootstrap as indicated in our discussion of the Delta method above.

Simulation Results

In this section, we present results of a small Monte Carlo study to illustrate the finite-sample performance of our proposed procedures in the setting of the example on bargaining bounds from Freyberger and Larsen (2021).

We consider four data generating processes corresponding to the examples discussed in Section 1.B.2. Across all designs X is sampled from a uniform distribution on $[0,1]$ and Y is sampled from a truncated normal on $[0,1]$ with standard deviation $1/3$ and design-specific mean. In our first design, we take the mean $m_{dec}(x) = x$, resulting in a decreasing conditional mean function. In the second design, $m_{inc}(x) = 1 - x$ and the conditional distribution function is increasing in x. For our third design, we choose the conditional mean to be constant, $m_{ind}(x) = 1/2$, corresponding to independence of Y and X. In our last design, we take

$$m_4(x) = -\frac{1}{3}\sin(4\pi x) + \frac{x^2}{3} + 0.65.$$

Our last design corresponds to the case where the partial derivatives $x \mapsto \partial_x F_{Y|X}(y,x)$ have finitely many zeros as discussed in Section 1.B.2. As discussed there, these four DGPs should correspond to different behavior of the plug-in estimator. As we do not know any of these shape information in practice, we use for our estimator the same bandwidths in any of the above DGPs. Potentially, one can improve considerably by using a data-dependent bandwidth selection rule, but this is left for future research.

For all DGPs, we generated 200 independent samples of sample size $n = 1000$, and we implemented both the Nadaraya-Watson estimator and the kernel density estimator using the rectangular kernel and the same bandwidth. The bandwidth h_n is chosen according to the rule

$$h_n = \frac{\int K(v)^2 dv}{(\int v^2 K(v) dv)^2} n^{-3/7}.$$

The rate factor $n^{-3/7}$ is chosen to guarantee undersmoothing and the first factor is a ratio measuring the kernel specific parts of the first order leading terms of the MSE of the Nadaraya-Watson. This choice of bandwidth is rather arbitrary and has not the aim to be optimal.

We simulate pointwise one-sided 95%-confidence intervals using the standard

bootstrap, our Delta method based approach[27] as proposed in Section 1.B.2 and projection confidence intervals. The projection intervals are based on uniform bands for the conditional distribution and density. Further details are given in Section 1.B.2. The performance of these confidence intervals is measured by their pointwise coverage probability and the average distance of the lower confidence bound to the plug-in estimator, which we call average length in the following.

Our results are depicted in Figures 1.B.1 and 1.B.2. The standard bootstrap and Delta method based confidence intervals seem to be preferable compared to the projection intervals. The projection intervals are very conservative, having 100% coverage across all DGPs and lead to much wider confidence bands than the other two approaches. While both the standard bootstrap and Delta method based confidence interval are severely conservative in the second and fourth DGP, the standard bootstrap intervals undercover both in the decreasing and independent DGP. In these cases, the Delta method based intervals perform much better and only slightly undercover in the independent DGP, while having a similar average length as the standard intervals across all DGPs.

Our results suggest that the standard bootstrap and Delta method based confidence intervals can adapt to the DGPs. The confidence intervals are roughly half as wide in DGPs 1 and 4 as compared to DGPs 2 and 3. From the theoretical analysis one might expect even smaller confidence intervals in DGP 1 as this DGP allows for $\sqrt{n}$-convergence while the other DGPs only converge at nonparametric rates. Therefore, the adaption to the DGP seems to be limited. It might be interesting to see whether data-dependent choices of the bandwidth can lead to a better adaptiveness of the confidence intervals.

Proof of differentiability

The Gâteaux directional derivative of ϕ is given by

$$\phi'_{F_{Y|X},f}(h,g) = \phi'_{F_{Y|X}}(f,h) + \phi'_f(F_{Y|X},g)$$

where

$$\phi'_{F_{Y|X}}(f,h) = \int \max_{x' \in \Phi(y,x)} h(y,x') f(x) dx$$

[27] We depict some realizations of the resulting confidence bands in Section 1.C.2.

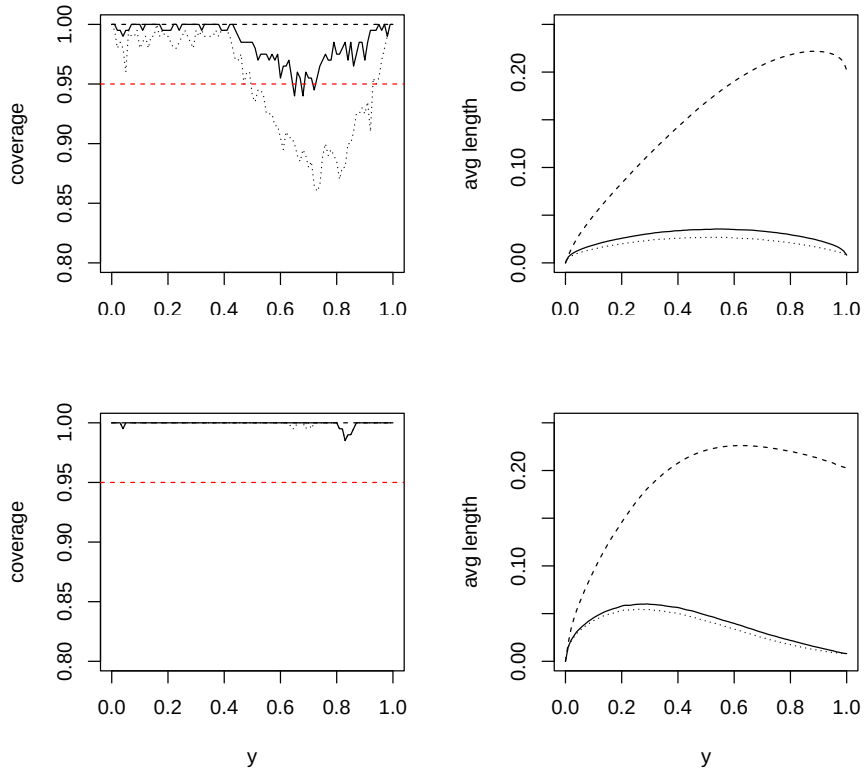

Figure 1.B.1: The left column presents estimates of the pointwise coverage probabilities and the right column the average length of the confidence intervals. The solid line corresponds to the Delta method confidence intervals, the dashed line to projection intervals and the dotted line to intervals based on the standard bootstrap. The first row presents results in the decreasing DGP and the second row for the increasing DGP.

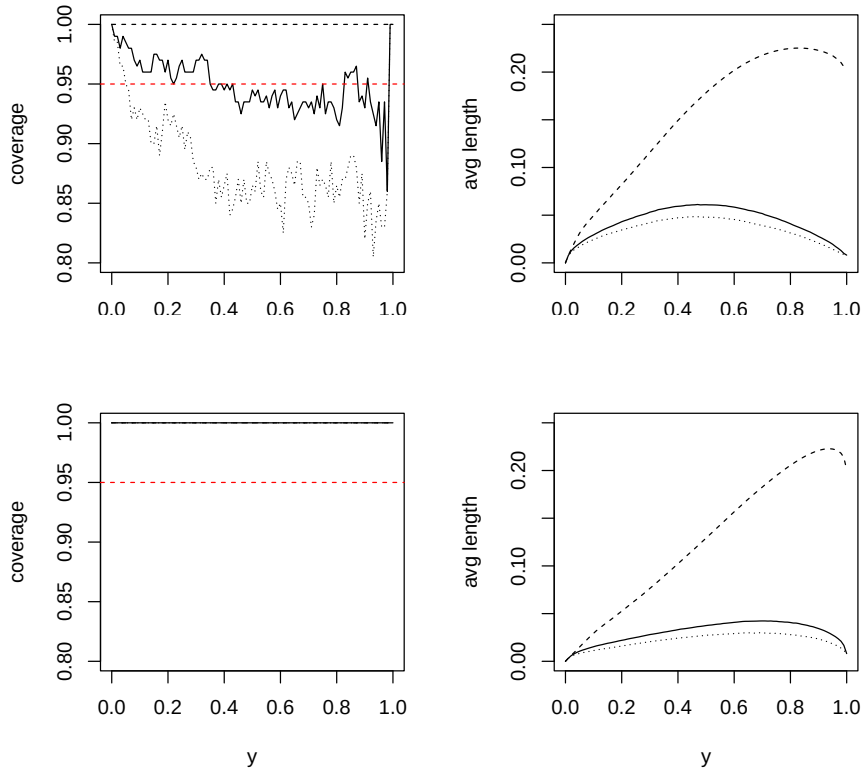

Figure 1.B.2: The left column presents estimates of the pointwise coverage probabilities and the right column the average length of the confidence intervals. The solid line corresponds to the Delta method confidence intervals, the dashed line to projection intervals and the dotted line to intervals based on the standard bootstrap. The first row presents results in the independent DGP and the second row for the fourth DGP.

and

$$\phi'_f(F_{Y|X}, g) = \int \max_{x' \geq x} F_{Y|X}(y, x')g(x)dx.$$

Note that in general, the Gâteaux directional derivative does not need to be additively separable as in the above form. If the Gâteaux directional derivative is continuous in $(F_{Y|X}, f)$, then this additive separability can be reassured. Moreover, in this continuous case, the derivative will be linear and therefore the mapping fully differentiable instead of only directionally differentiable.

For our differentiability result, we need some notion of well-separability of the local maxima of $x \mapsto F_{Y|X}(y, x)$. This condition can also be interpreted as a measure of identifiability of the restricted argmax sets $\Psi(y, x)$. It is similar to the well-separatedness condition used in the maximum of a conditional mean function example, but allows for regions of weak identifiability.

Assumption 12 (Well-separatedness Condition)**.** *Let* $\Psi(y, x) := \text{argmax}_{x' \geq x} F_{Y|X}(y, x')$ *and* Ξ *be some finite subset of* $[0, 1]$*. For some constants* $c_1, c_2, \delta > 0$ *and* $\rho, \alpha \geq 1$ *independent of* (y, x)*, it holds for all* y*,* $x \in [0, 1]$ *and* $\tilde{x} \geq x$

$$\max_{x' \geq x} \{F_{Y|X}(y, x')\} - F_{Y|X}(y, \tilde{x}) \geq (c_1 d(\tilde{x}, \Psi(y, x))^\rho) \wedge (c_2 d(x, \Xi)^\alpha) \wedge \delta. \tag{1.23}$$

While we assume this condition to hold everywhere, it is sufficient to only assume this condition to hold Lebesgue almost everywhere since ϕ averages over the restricted maximum.

In the following Lemma, we give examples of conditional distribution functions satisfying this requirement.

Lemma 27. *(i) Suppose that* $F_{Y|X}(y, x)$ *is strictly increasing in* x *for all* y *with* $\partial_x F_{Y|X}(y, x) \geq c > 0$ *for all* x *and* y*. Then* $F_{Y|X}$ *satisfies (1.23) with* $\rho = 1$*,* $c_1 = c$ *and* c_2, α *arbitrary. Further,* ϕ *is* 2*-Frèchet directionally differentiable and its derivative reduces to* $\phi'_{F_{Y|X}}(h, f)(y) = h(y, 1)$*.*

(ii) Suppose that $F_{Y|X}(y, x)$ *is strictly decreasing in* x *for all* y *with* $\partial_x F_{Y|X}(y, x) \leq c < 0$ *for all* x *and* y*. Then* $F_{Y|X}$ *satisfies (1.23) with* $\rho = 1$*,* $c_1 = c$ *and* c_2, α *arbitrary. Further,* ϕ *is* 2*-Fréchet directionally differentiable and its derivative*

reduces to

$$\phi'_{F_{Y|X}}(h,f)(y) = \int h(y,x)f(x)d(x).$$

(iii) Suppose that Y and X are independent. Then $F_{Y|X}$ satisfies (1.23) with c_1, c_2, ρ and α arbitrary. Further, ϕ is 2-Fréchet directionally differentiable.

(iv) Suppose that $Y = m(X) + \varepsilon$ with $\varepsilon \,|\, X \sim G$. G has a density g with respect to the Lebesgue measure on $[0,1]$ and g is twice continuously differentiable and bounded away from zero and infinity. m is two times continuously differentiable with a Lipschitz continuous second derivative. m has a finite number of local minimizers $0 < x_1 < x_2 < \cdots < x_m < 1$ with corresponding values $m(x_1) < m(x_2) < \cdots < m(x_m)$. These local minimizers are well separated in the sense that $m(x_j) - m(x_{j-1}) \leq -\delta$ for some $\delta > 0$ and all $j = 2, \ldots, m$. Further, the second derivative of m has at most $2(m-1)$ zeros and satisfies $\min_j m''(x_j) \geq c$.

Then $F_{Y|X}$ satisfies (1.23) with $\rho = 2$, c_1 and c_2 some constants bounded away from zero, $\alpha = 1$ and Ξ is finite. Further, ϕ is 2-Fréchet directionally differentiable.

Proof. **(i):** In this case $\Psi(y,x) = \{1\}$, for all (y,x), and by the lower bound on the derivative, for any $\tilde{x} \geq x$,

$$F_{Y|X}(y,1) - F_{Y|X}(y,\tilde{x}) = \int_{\tilde{x}}^{1} \frac{\partial F_{Y|X}(y,t)}{\partial x} dt \geq c(1-\tilde{x}) = c\, d(\tilde{x}, \Psi(y,x)).$$

(ii): Here, $\Psi(y,x) = \{x\}$, for all (y,x), and again by the lower bound on the derivative, for any $\tilde{x} \geq x$,

$$F_{Y|X}(y,x) - F_{Y|X}(y,\tilde{x}) = \int_{x}^{\tilde{x}} -\frac{\partial F_{Y|X}(y,t)}{\partial x} dt \geq c\, d(\tilde{x}, \Psi(y,x)).$$

(iii): Under independence, the conditional distribution function does not change in x, i.e., there is some function $F_{Y|X}(y)$ such that $F_{Y|X}(y,x) = F_{Y|X}(y)$ for all x. Therefore, the argmax set satisfies $\Psi(y,x) = [x,1]$ in this case and $\max_{x' \geq x} F_{Y|X}(y,x') - F_{Y|X}(y,\tilde{x}) = 0$ for all $\tilde{x} \geq x$.

(iv) It holds $F_{Y|X}(y,x) = G(y - m(x))$. Then,

$$\frac{\partial F_{Y|X}(y,x)}{\partial x} = -g(y - m(x))m'(x)$$

$$\frac{\partial^2 F_{Y|X}(y,x)}{\partial x^2} = g'(y - m(x))(m'(x))^2 - g(y - m(x))m''(x)$$

Note that any local minimizer of m corresponds to a local maximizer of $F_{Y|X}$ uniformly over y. Moreover, by the assumed structure on the local minimizers of m, the restricted argmax set $\Psi(y,x)$ satisfies

$$\Psi(y,x) = \begin{cases} \{x_1\} & \text{, if } x \in [0, x_1] \\ \{x\} & \text{, if } x \in (x_1, \xi_1) \\ \{x, x_2\} & \text{, if } x = \xi_1 \\ \{x_2\} & \text{, if } x \in (\xi_1, x_2] \\ \{x\} & \text{, if } x \in (x_2, \xi_2) \\ \{x, x_3\} & \text{, if } x = \xi_2 \\ \dots & \\ \{x_{m-1}\} & \text{, if } x \in (\xi_{m-2}, x_{m-1}] \\ \{x\} & \text{, if } x \in (x_{m-1}, \xi_{m-1}) \\ \{x, x_m\} & \text{, if } x = \xi_{m-1} \\ \{x_m\} & \text{, if } x \in (\xi_{m-1}, x_m] \\ \{x\} & \text{, if } x \in (x_m, 1] \end{cases}$$

for some values ξ_j, $j = 1, \dots, m-1$ which are the solution to the equation

$$F_{Y|X}(y, \xi_{j-1}) = F_{Y|X}(y, x_j).$$

These ξ_j do not depend on y, since G is invertible and therefore the above equation is equivalent to $m(\xi_{j-1}) = m(x_j)$. There are only $m-1$ many of these values by the assumed upper bound on the zeros of the second derivative. In order to see this, note that for any x sufficiently close to x_{j-1}, we have $m'(x) > 0$ and for any x sufficiently close to x_j, it holds $m'(x) < 0$. As the derivative is continuous, the derivative has to attain zero in the interior of $[x_{j-1}, x_j]$. This also implies that the second derivative has to have attained at least twice the value zero in the interior of $[x_{j-1}, x_j]$. By the assumed upper bound on the zeros of the second derivative, it attains zero exactly twice. This then implies that the first derivative attains zero exactly once in the interior. Denote this zero by ζ. Since $m'(x) > 0$ for all $x < \zeta$

and $m'(x) < 0$ for all $x > \zeta$, ξ_{j-1} has to lie to the left of ζ and $m(\zeta) > m(x_j)$. This implies by the intermediate value theorem that there exists exactly one ξ_{j-1}. Finally, since there are $m-1$ intervals of the form $[x_{j-1}, x_j]$, the claim follows.

The above arguments show that the first derivative of m is bounded from below on the interval $[x_{j-1}+\varepsilon, \xi_{j-1}]$, where $\varepsilon > 0$. That is, $m'(x) \geq c_1(\varepsilon)$ for all $x \in [x_{j-1}+\varepsilon, \xi_{j-1}]$. Taking ε sufficiently small, we can make this lower bound uniform over j. This implies for any $x \in [x_{j-1}+\varepsilon, \xi_{j-1}]$

$$m(x) \leq m(\xi_{j-1}) + c_1(\varepsilon)(x - \xi_{j-1}),$$

i.e., $m(x_j) - m(x) \geq c_1(\varepsilon) d(x, \xi_j)$. This implies the following well-separatedness conditions for $x \in [x_{j-1}+\varepsilon, \xi_{j-1}]$ and $x' \in [x, \xi_{j-1}]$

$$F_{Y|X}(y,x) - F_{Y|X}(y,x') = \int_x^{x'} g(y - m(t)) m'(t) dt \geq \underline{g} c_1(\varepsilon) d(x', \Psi(y,x)),$$

where $\underline{g}$ denotes a lower bound on g. Moreover, for any $x' > \xi_{j-1}$, $F_{Y|X}(y,x') \leq F_{Y|X}(y,x_j)$ since x_j corresponds to the next smallest local minimum of m. Hence, for any $x \in [x_{j-1}+\varepsilon, \xi_{j-1}]$ and $x' \geq x$, we have

$$F_{Y|X}(y,x) - F_{Y|X}(y,x') \geq F_{Y|X}(y,x) - F_{Y|X}(y,\xi_{j-1}) \geq \underline{g} c_1(\varepsilon) d(x, \xi_{j-1})$$

We can make this bound hold uniformly over j, by defining $\Xi = \{\xi_j : j = 1, \ldots, m-1\}$. Then, for any j and any $x \in [x_{j-1}+\varepsilon, \xi_{j-1}]$ and $x' \geq x$, we have

$$F_{Y|X}(y,x) - F_{Y|X}(y,x') \geq \underline{g} c_1(\varepsilon) d(x, \Xi).$$

Thus, if $x \in [x_{j-1}+\varepsilon, \xi_{j-1}]$ and $\tilde{x} \geq x$,

$$\max_{x' \geq x} F_{Y|X}(y,x') - F_{Y|X}(y,\tilde{x}) \geq \underline{g} c_1(\varepsilon) \big(d(\tilde{x}, \Psi(y,x)) \wedge d(x, \Xi) \big).$$

The case $x \in (x_m, 1]$ follows analogously with a qualitatively similar bound and is omitted in order to preserve space.

If $x \in [0, x_1]$ or $(\xi_{j-1}, x_j]$, after potentially decreasing the ε from above, there is a constant C which only depends on c and the bounds of the density g such that

$$\sup_{d(x',X) \leq \varepsilon} \frac{\partial^2 F_{Y|X}(y,x')}{\partial x^2} \leq -\frac{C}{2}. \tag{1.24}$$

Here X denotes the collection of local minimizers, i.e., $X = \{x_j : j = 1, \dots, m\}$. Thus, for any j and x such that $|x' - x_j| \leq \varepsilon$, by a second order Taylor expansion

$$F_{Y|X}(y, x') - F_{Y|X}(y, x_j) \leq -\frac{c}{4} d(x', \Psi(y, x))^2$$

or equivalently, $\max_{\tilde{x} \geq x} F_{Y|X}(y, \tilde{x}) - F_{Y|X}(y, x') \geq C/4 d(x', \Psi(y, x))^2$. Next, for any x' such that $|x' - x_j| > \varepsilon$, we have either $F_{Y|X}(y, x') \leq F_{Y|X}(y, x_{j+1}) \leq F_{Y|X}(y, x_j) - \delta$ or $F_{Y|X}(y, x_{j+1}) \leq F_{Y|X}(y, x')$. In the latter case, we have by the same arguments as above that $F_{Y|X}(y, x')$ decreases at least linearly for all $|x' - x_j| > \varepsilon$ with $F_{Y|X}(y, x') > F_{Y|X}(y, x_{j+1})$. Hence, by potentially lowering δ, we have

$$\max_{\tilde{x} \geq x} F_{Y|X}(y, \tilde{x}) - F_{Y|X}(y, x') \geq \delta,$$

for all x' satisfying $|x' - x_j| > \varepsilon$. Thus, if $x \in [0, x_1]$ or $(\xi_{j-1}, x_j]$, for all $x' \geq x$,

$$\max_{\tilde{x} \geq x} F_{Y|X}(y, \tilde{x}) - F_{Y|X}(y, x') \geq \left(C/4 d(x', \Psi(y, x))^2\right) \wedge \delta.$$

Next, let $x \in (x_j, x_j + \varepsilon]$, $j = 1, \dots m$. For any $\eta \in (0, \varepsilon]$, by (1.24) and $m'(x_j) = 0$

$$\frac{\partial F_{Y|X}(y, x_j + \eta)}{\partial x} = \frac{\partial F_{Y|X}(y, x_j)}{\partial x} + \int_{x_j}^{x_j + \eta} \frac{\partial^2 F_{Y|X}(y, t)}{\partial x^2} dt \leq -\frac{C}{2}\eta.$$

Hence, for any $\tilde{x} \in [x, x_j + \varepsilon]$,

$$F_{Y|X}(y, x) - F_{Y|X}(y, \tilde{x}) - = \int_{\tilde{x}}^{x} \frac{\partial F_{Y|X}(y, t)}{\partial x} dt \geq \frac{C}{4}[\tilde{x}^2 - x^2] \geq \frac{C x_1}{2} d(\tilde{x}, \Psi(y, x)).$$

Moreover, by L-Lipschitz continuity of $F_{Y|X}$,

$$F_{Y|X}(y, x_j) - F_{Y|X}(y, x_j + \varepsilon) \leq L\varepsilon$$

and as $F_{Y|X}(y, x) \geq F_{Y|X}(y, x_j + \varepsilon)$, we have for all $\tilde{x} \geq x$

$$\begin{aligned} F_{Y|X}(y, x) - F_{Y|X}(y, \tilde{x}) \geq & F_{Y|X}(y, x_j) - F_{Y|X}(y, \tilde{x}) + F_{Y|X}(y, x_j + \varepsilon) - F_{Y|X}(y, x_j) \\ \geq & \delta - L\varepsilon, \end{aligned}$$

where we used that $F_{Y|X}(y, x_j) - F_{Y|X}(y, \tilde{x}) \geq \delta$ by the above argument. Thus, for

any $x \in (x_j, x_j + \varepsilon]$, $j = 1, \dots m$, and any $\tilde{x} \geq x$

$$F_{Y|X}(y, x) - F_{Y|X}(y, \tilde{x}) \geq \left(\frac{Cx_1}{2} d(\tilde{x}, \Psi(y, x)) \right) \wedge \big(\delta - L\varepsilon \big).$$

To conclude, we have shown that there exists a finite collection Ξ and some constants c_1, c_2, c_3 and δ, so that for any y, $x \in [0, 1]$ and any $\tilde{x} \geq x$,

$$\begin{aligned} &\max_{x' \geq x} F_{Y|X}(y, x') - F_{Y|X}(y, \tilde{x}) \\ &\geq \begin{cases} \{c_1 d(\tilde{x}, \Psi(y, x))^2\} \wedge \delta & \text{, if } x \in [0, x_1] \text{ or } x \in (\xi_{j-1}, x_j] \\ \{c_2 d(\tilde{x}, \Psi(y, x))\} \wedge \delta & \text{, if } x \in (x_j, x_j + \varepsilon] \\ \{c_3 d(\tilde{x}, \Psi(y, x))\} \wedge c_3 d(x, \Xi) & \text{, if } x \in (x_j + \varepsilon, \xi_j). \end{cases} \end{aligned}$$

Since $d(\tilde{x}, \Psi(y, x)) \geq d(\tilde{x}, \Psi(y, x))^2$ for all $\tilde{x}, x, y$, we have that for all y, any $x \in [0, 1]$ and $\tilde{x} \geq 1$,

$$\max_{x' \geq x} F_{Y|X}(y, x') - F_{Y|X}(y, \tilde{x}) \geq \big(c_4 d(\tilde{x}, \Psi(y, x))^2 \big) \wedge \big(c_3 d(x, \Xi) \big) \wedge \delta,$$

where c_4 denotes the minimum of c_1, c_2 and c_3. The claim follows. □

We want to give two further examples to illustrate the behavior of the numbers of zero of the partial derivative with respect to x. Consider for example (Y, X) with $\varepsilon = Y - m(X)$, $\mathrm{E}[Y \mid X] = m(X)$ and $\varepsilon | X \sim F$. Then

$$\theta(y, x) = F(y - m(x))$$

and

$$\partial_x \theta(y, x) = -f(y - m(x)) m'(x)$$

which is zero exactly when $m'(x) = 0$. Thus, the number and locations of the zeros are fully determined by $m(x)$. Moreover,

$$\partial_{xx} \theta(y, x) = f'(y - m(x))(m'(x))^2 - f(y - m(x)) m''(x).$$

In particular, for any x such that $m'(x) = 0$, the sign of the curvature is fully determined by $m''(x)$. Furthermore, we can bound the size of the curvature at any

such zero by assuming a bound on the density f and m'' at these zeros.

Consider on the other hand a location scale model, $Y = m(X) + \sigma(X)\varepsilon$ with $\varepsilon|X \sim F$. Here, $\mathrm{E}[Y \mid X] = m(X)$, $\mathrm{Var}(Y|X) = \sigma^2(X)$ and

$$\theta(y,x) = F\left(\frac{Y - m(X)}{\sigma(X)}\right)$$

with

$$\partial_x \theta(y,x) = -f\left(\frac{y-m(x)}{\sigma(x)}\right)\left\{\frac{y-m(x)}{\sigma(x)}\frac{\sigma'(x)}{\sigma(x)} + \frac{m'(x)}{\sigma(x)}\right\}$$

Here, not only the location of the zeros may depend on y but also the number.

The following Lemma shows $(\gamma, \|\cdot\|_*)$-Fréchet directional differentiability of ϕ. Here, we use the norm $\|(h,g)\|_* = \|h\|_{BL} \vee \|g\|_\infty$, where

$$\|h\|_{BL} = \sup_y \left\{ \sup_x |h(y,x)| \vee \sup_{x \neq x'} \frac{|h(y,x) - h(y,x')|}{|x - x'|} \right\}.$$

Lemma 28. *For any (y,x), let $\Psi(y,x) = \mathrm{argmax}_{x' \geq x} F_{Y|X}(y,x')$ denote the argmax set at (y,x). Further, suppose that the well-separatedness condition (1.23) holds. Then, ϕ is $(\frac{\rho}{\rho-1} \wedge \frac{\alpha+1}{\alpha} \wedge 2)$-Frèchet directionally differentiable at $(F_{Y|X}, f)$ with derivative $\phi'_{F_{Y|X},f}(h,g) = \phi'_{F_{Y|X}}(f,h) + \phi'_f(F_{Y|X}, g)$, where*

$$\phi'_{F_{Y|X}}(h,f) = \int \max_{x' \in \Phi(y,x)} h(y,x') f(x) dx$$

and

$$\phi'_f(F_{Y|X}, g) = \int \max_{x' \geq x} F_{Y|X}(y,x') g(x) dx.$$

Proof of Lemma 28: We only discuss the case $\rho > 1$. The case $\rho = 1$ can be handled using the same arguments as in the proof of Proposition 1.

For any bounded and Lipschitz continuous function h and bounded continuous function g,

$$\begin{aligned}\phi(F_{Y|X} + h, f + g) = &\int \max_{x' \geq x} \{F_{Y|X}(y,x') + h(y,x')\} f(x) dx \\ &+ \int \max_{x' \geq x} \{F_{Y|X}(y,x') + h(y,x')\} g(x) dx.\end{aligned}$$

We can further decompose

$$\int \max_{x' \geq x} \{F_{Y|X}(y, x') + h(y, x')\} f(x) dx = \phi(F_{Y|X}, f) + \phi'_{F_{Y|X}}(h, f) + R_1$$

and

$$\begin{aligned} &\int \max_{x' \geq x} \{F_{Y|X}(y, x') + h(y, x')\} g(x) dx \\ =&\phi'_f(F_{Y|X}, g) + \int \max_{x' \in \Phi(y,x)} h(y, x') g(x) dx + R_2, \end{aligned}$$

where

$$\begin{aligned} R_1 &= \int \left\{ \max_{x' \geq x} \{F_{Y|X}(y, x') + h(y, x')\} - \max_{x' \in \Psi(y,x)} \{F_{Y|X}(y, x') + h(y, x')\} \right\} f(x) dx \\ R_2 &= \int \left\{ \max_{x' \geq x} \{F_{Y|X}(y, x') + h(y, x')\} - \max_{x' \in \Psi(y,x)} \{F_{Y|X}(y, x') + h(y, x')\} \right\} g(x) dx. \end{aligned}$$

In (i), we will show that for all $x \in [0, 1]$

$$\sup_y \left| \max_{x' \geq x} \{F_{Y|X}(y, x') + h(y, x')\} - \max_{x' \geq x} F_{Y|X}(y, x') - \max_{x' \in \Psi(y,x)} h(y, x') \right| \leq \frac{1}{c_1} \|h\|_{BL}^{\frac{\rho}{\rho - 1}}, \tag{1.25}$$

whenever $\|h\|_{BL} < c_2 x^\alpha$. In (ii), we show that this implies

$$\sup_y |R_1| \leq \left(\frac{1}{c_1} \|h\|_{BL}^{\frac{\rho}{\rho-1}} + \frac{1}{c_2} \|h\|_{BL}^{\frac{\alpha+1}{\alpha}} \right) \|f\|_\infty \tag{1.26}$$

$$\sup_y |R_2| \leq \left(\frac{1}{c_1} \|h\|_{BL}^{\frac{\rho}{\rho-1}} + \frac{1}{c_2} \|h\|_{BL}^{\frac{\alpha+1}{\alpha}} \right) \|g\|_\infty. \tag{1.27}$$

Further, by the Cauchy-Schwarz inequality and properties of the sup-norm,

$$\left| \int \max_{x' \in \Phi(y,x)} h(y, x') g(x) dx \right| \leq \|h\|_{BL} \|g\|_\infty.$$

Combining these results, yields

$$\begin{aligned} &\sup_y |\phi(F_{Y|X} + th, f + tg) - \phi(F_{Y|X}, f) + \phi'_{F_{Y|X}}(h, f) + \phi'_f(F_{Y|X}, g)| \\ \leq &\left(\frac{1}{c_1} \|h\|_{BL}^{\frac{\rho}{\rho-1}} + \frac{1}{c_2} \|h\|_{BL}^{\frac{\alpha+1}{\alpha}} \right) \left\{ \|f\|_\infty + \|g\|_\infty \right\} + \|h\|_{BL} \|g\|_\infty. \end{aligned}$$

As $\|h\|_{BL}, \|g\|_\infty \to 0$ and by the elementary inequality $ab \le (a \vee b)^2$ for $a, b \ge 0$

$$\begin{aligned}&\left(\frac{1}{c_1}\|h\|_{BL}^{\frac{\rho}{\rho-1}} + \frac{1}{c_2}\|h\|_{BL}^{\frac{\alpha+1}{\alpha}}\right)\{\|f\|_\infty + \|g\|_\infty\} + \|h\|_{BL}\|g\|_\infty \\ =&O\big((\|h\|_{BL} \vee \|g\|_\infty)^{\frac{\rho}{\rho-1}\wedge\frac{\alpha+1}{\alpha}\wedge 2}\big).\end{aligned}$$

It remains to show (1.25)-(1.27).

(i) This follows essentially by the same arguments as in the proof of Proposition 1. Fix some $y \in \mathbb{R}$ and $x \in [0,1]$. Let $\Psi(y,x) = \operatorname{argmax}_{x' \ge x} F_{Y|X}(y,x')$ and $\Psi_h(y,x) = \operatorname{argmax}_{x' \ge x}\{F_{Y|X}(y,x') + h(y,x')\}$. Further, let $x^*(x) \in \operatorname{argmin}_{x' \in \Psi(y,x)}|x - x'|$. Now, for any $x_h \in \Psi_h(y,x)$,

$$\begin{aligned}F_{Y|X}(y,x_h) + h(y,x_h) \le& F_{Y|X}(y,x_h) + h(y,x^*(x_h)) + \|h\|_{BL} d(x_h, \Psi(y,x)) \\ F_{Y|X}(y,x_h) + h(y,x_h) \ge& \bar{F}_{Y|X}(y) + \max_{x' \in \Psi(y,x)} h(y,x') \ge \bar{F}_{Y|X}(y) + h(y,x^*(x_h)),\end{aligned}$$

where $\bar{F}_{Y|X}(y) = \max_{x' \in \Psi(y,x)} F_{Y|X}(y,x')$. This implies

$$\bar{F}_{Y|X}(y) - F_{Y|X}(y,x_h) \le \|h\|_{BL} d(x_h, \Psi(y,x)).$$

Whenever $\|h\|_{BL} < c_2 d(x,\Xi)^\alpha \wedge \delta =: \delta(x)$, by the well-separatedness condition

$$d(x_h, \Psi(y,x)) \le \frac{1}{c_1}\|h\|_{BL}^{\frac{1}{\rho-1}} =: \varepsilon(h).$$

Now, (1.25) follows by the same computations as in the proof of Proposition 1.

(ii) We only show (1.26). (1.27) follows by the same arguments replacing f by g. Now, let $A = \{x : \|h\| < \delta(x)\}$ and denote by $A^c = [0,1] \setminus A$. We can decompose R_1 into

$$\begin{aligned}&\int\left\{\max_{x' \ge x}\{F_{Y|X}(y,x') + h(y,x')\} - \max_{x' \in \Psi(y,x)}\{F_{Y|X}(y,x') + h(y,x')\}\right\} f(x)dx \\ =&\int_A\left\{\max_{x' \ge x}\{F_{Y|X}(y,x') + h(y,x')\} - \max_{x' \in \Psi(y,x)}\{F_{Y|X}(y,x') + h(y,x')\}\right\} f(x)dx \\ &+ \int_{A^c}\left\{\max_{x' \ge x}\{F_{Y|X}(y,x') + h(y,x')\} - \max_{x' \in \Psi(y,x)}\{F_{Y|X}(y,x') + h(y,x')\}\right\} f(x)dx.\end{aligned}$$

On A, we can bound the integrand by (1.25)

$$\begin{aligned}&\left|\int_A \left\{\max_{x'\geq x}\{F_{Y|X}(y,x')+h(y,x')\}-\max_{x'\in\Psi(y,x)}\{F_{Y|X}(y,x')+h(y,x')\}\right\}f(x)dx\right|\\ \leq&\frac{1}{c_1}\|h\|_{BL}^{\frac{\rho}{\rho-1}}\|f\|_\infty.\end{aligned}$$

On A^c, by properties of the maximum and since $x' \mapsto F_{Y|X}(y,x')$ is constant on $\Psi(y,x)$,

$$\begin{aligned}0\leq&\max_{x'\geq x}\{F_{Y|X}(y,x')+h(y,x')\}-\max_{x'\in\Psi(y,x)}\{F_{Y|X}(y,x')+h(y,x')\}\\ \leq&F^-_{Y|X}(y)+\max_{x'\geq x}h(y,x')-F^-_{Y|X}(y)-\max_{x'\in\Psi(y,x)}h(y,x')\leq\|h\|_{BL}.\end{aligned}$$

Hence,

$$\begin{aligned}&\left|\int_{A^c}\left\{\max_{x'\geq x}\{F_{Y|X}(y,x')+h(y,x')\}-\max_{x'\in\Psi(y,x)}\{F_{Y|X}(y,x')+h(y,x')\}\right\}f(x)dx\right|\\ \leq&\|h\|_{BL}\|f\|_\infty\mathrm{Leb}(A^c),\end{aligned}$$

where $\mathrm{Leb}(A)$ denotes the Lebesgue measure of set A. Further, since $A^c = \{x : d(x,\Xi) \leq \|h\|_{BL}^{1/\alpha}/c_2\}$, $\mathrm{Leb}(A^c) \leq \|h\|_{BL}^{1/\alpha}\#\Xi/c_2$ implying

$$\sup_y|R_1| \leq \left(\frac{1}{c_1}\|h\|_{BL}^{\frac{\rho}{\rho-1}}+\frac{\#\Xi}{c_2}\|h\|_{BL}^{\frac{\alpha+1}{\alpha}}\right)\|f\|_\infty.$$

The claimed result follows. □

Lemma 29. *The derivative $\phi'_{F_{Y|X},f}(h,g)$ in Proposition 28 is 1-Lipschitz continuous and convex for each y. Further, $\sup_y \phi'_{F_{Y|X},f}(h,g)(y)$ is sublinear and 1-Lipschitz continuous.*

Proof. Recall that $\phi'_{F_{Y|X},f}(h,g) = \phi'_{F_{Y|X}}(f,h) + \phi'_f(F_{Y|X},g)$, where

$$\phi'_{F_{Y|X}}(h,f) = \int \max_{x'\in\Phi(y,x)} h(y,x')f(x)dx$$

and

$$\phi'_f(F_{Y|X},g) = \int \max_{x'\geq x} F_{Y|X}(y,x')g(x)dx.$$

Since for any h, h' and g, g'

$$\begin{aligned}&|\phi'_{F_{Y|X},f}(h,g) - \phi'_{F_{Y|X},f}(h',g')|\\ \leq&|\phi'_{F_{Y|X}}(h,f) - \phi'_{F_{Y|X}}(h',f)| + |\phi'_f(F_{Y|X},g) - \phi'_f(F_{Y|X},g')|\end{aligned}$$

and

$$\begin{aligned}|\phi'_{F_{Y|X}}(h,f) - \phi'_{F_{Y|X}}(h',f)| \leq& \int \left| \max_{x' \in \Phi(y,x)} h(y,x') - \max_{x' \in \Phi(y,x)} h(y,x') \right| f(x)dx\\ \leq& \|h - h'\|_\infty \int f(x)dx = \|h - h'\|_\infty\end{aligned}$$

as well as

$$|\phi'_f(F_{Y|X},g) - \phi'_f(F_{Y|X},g')| \leq \|g - g'\|_\infty,$$

where we used that $F_{Y|X}(y,x) \leq 1$ as it is a conditional distribution function. Thus, $|\phi'_{F_{Y|X},f}(h,g) - \phi'_{F_{Y|X},f}(h',g')| \leq \|h - h'\|_\infty + \|g - g'\|_\infty$, i.e., $\phi'_{F_{Y|X},f}$ is 1-Lipschitz continuous.

Secondly, for any y, $\phi'_{F_{Y|X}}$ is convex since for all h, h' and $\lambda \in [0,1]$

$$\begin{aligned}&\phi'_{F_{Y|X}}(\lambda h + (1-\lambda)h', f)\\ =& \int \max_{x' \in \Phi(y,x)} \{\lambda h(y,x') + (1-\lambda)h'(y,x')\} f(x)dx\\ \leq& \lambda \int \max_{x' \in \Phi(y,x)} h(y,x') f(x)dx + (1-\lambda) \int \max_{x' \in \Phi(y,x)} h'(y,x') f(x)dx\\ =& \lambda \phi'_{F_{Y|X}}(h,f) + (1-\lambda)\phi'_{F_{Y|X}}(h',f).\end{aligned}$$

Further, ϕ'_f is convex as it is linear in g. Thus, $\phi'_{F_{Y|X},f}$ is convex as it is the sum of two convex functions. □

The next lemma gives some further analytic properties of the derivative when it is linear. These properties are needed for the strong approximations in Lemma 36.

Lemma 30. *Suppose that the derivative is linear. Then, for any y, there exists a unique monotonely increasing càdlàg function $x^* : [0,1] \to [0,1]$ with $x^*(x) \in [x,1]$ so that*

$$\phi'_{F_{Y|X}}(h)(y) = \int h(y, x^*(x)) f_X(x)dx$$

for all h. Finally, if x^* is strictly increasing, $x^*(x) = x$.

Proof. The derivative is linear if and only if $\Psi(y, x)$ is almost everywhere unique when holding y fixed. That is, for any y, there exist selections $\tilde{x}(x) \in \Psi(y, x)$, for all x, which are a.e. unique and satisfy, for all h,

$$\phi'_{F_{Y|X}}(h)(y) = \int h(y, \tilde{x}(x)) f_X(x) dx.$$

Let $x^*(x)$ denote the selection which always chooses $\max \Psi(y, x)$. Note that it is well-defined by continuity of $F_{Y|X}$. By definition of the argmax set $\Psi(y, x)$, we have for any $x < x'$, $x^*(x) \leq x^*(x')$ and therefore this selection is monotonely increasing. It remains to show that x^* is càdlàg. For the continuity from the right, let $x_n \downarrow x$ for some arbitrary x. If $\max \Psi(y, x) \in (x, 1)$, then there exist some N so that for all $n \geq N$, we have $\Psi(y, x_n) = \Psi(y, x)$ and therefore $x^*(x_n) = x^*(x)$. If otherwise $\max \Psi(y, x) = x$, then by the well-separatedness condition, there exists a $\delta(x)$ such that for any $\tilde{x}$ satisfying $|\tilde{x} - x| < \delta(x)$, we have $\|\Psi(y, x) - \max \Psi(y, \tilde{x})| < \delta(x)$ and therefore x^* is continuous from the right.

For the existence of a limit from left, let $x_n \uparrow x$ for some arbitrary $x \in [0, 1]$. If $\Psi(y, x) = x$, then for all $\tilde{x}$ satisfying $\tilde{x} < x$, it holds $\Psi(y, \tilde{x}) \in \{\tilde{x}, x\}$. When $\Psi(y, \tilde{x}) = x$, it holds for all $x' \in (\tilde{x}, x)$, $\Psi(y, x') = x$ and therefore $x^*(x')$ is constant on $[\tilde{x}, x]$ and clearly the left limit exists. Now suppose that $\Psi(y, \tilde{x}) = \tilde{x}$ and fix some $\varepsilon < x - \tilde{x}$. Since $x_n \uparrow x$, there is some N such that for all $n \geq N$, $x - x_n < \varepsilon$. By monotony of x^*, it holds $x^*(x_n) \in [x_n, x]$ for all n and therefore $\lim_{n\to\infty} x^*(x_n) = x$. Hence, x^* is even continuous in this case. If otherwise $\Psi(y, x) > x$, then either there exist some $\tilde{x}$ so that x^* is constant on $(\tilde{x}, x]$ and therefore the limit exists or by the same argument as above, we can show that $x^*(x_n)$ converges to x.

In order to see the final claim, suppose by contradiction that x^* is discontinuous at x. Since x^* is càdlàg, it has to be discontinuous from the left. That is, there is a sequence $x_n \uparrow x$ such that $x^*(x_n)$ does not converge to $x^*(x)$. As we have seen above, x^* can only be discontinuous from the left when $x^*(x) > x$. But when $x^*(x) > x$ for some x, then x^* is constant on $[x, x^*(x)]$ and therefore x^* cannot be strictly increasing. Finally, note that since $x^*(x) \geq x$ and $x^*(x)$ cannot be larger than x, it holds $x^*(x) = x$ in this case. □

For the application of Theorem 2, we require a lower bound on the variance of the derivative. Our derivation of such a lower bound depends on a representation

result for the derivative which we prove in the following Lemma.

Lemma 31. *The derivative ϕ'_{θ_0} can be represented for any fixed $y \in (0,1)$ as*

$$\phi'_{\theta_0}(h,g)(y) = \sup_{\psi \leq \phi'_{\theta_0}} \psi(h,g),$$

where the set $\{\psi \leq \phi'_{\theta_0}\}$ is a shorthand for

$$\{\psi \leq \phi'_{\theta_0}\} = \{\psi \in \mathbb{D}^* | \forall h, g : \psi(h,g) \leq \phi'_{\theta_0}(h,g)(y)\}.$$

The sets $\{\psi \leq \phi_{F_{Y|X}}\}$ and $\{\psi \leq \phi'_f\}$ are similarly defined. Then, it holds $\{\psi \leq \phi'_f\} = \{\phi'_f\}$ and

$$\{\psi \leq \phi_{F_{Y|X}}\} \supset \left\{h \mapsto \int h(y, x^*(x)) f_X(x) dx : x^*(x) \in \Psi(y,x) \forall x \in [0,1]\right\}$$
$$\{\psi \leq \phi_{F_{Y|X}}\} \subset \left\{h \mapsto \int h(y,x) d\mu(x) : \mu \text{ positive and finite Borel measure}\right\}.$$

Proof. Fix some arbitrary $y \in (0,1)$. The first part of the claim follows by Lemma 2 in our companion paper Scherer (2024). For the remainder, note that ϕ'_f evaluated at $y \in (0,1)$ is a linear functional and therefore $\{\psi \leq \phi'_f\} = \{\phi'_f\}$. Therefore, it only remains to characterize the linear functionals which are dominated by the derivative with respect to $F_{Y|X}$. For one direction, let $x^*(x) \in \Psi(y,x)$ for all x and fixed $y \in (0,1)$. Then, $x^*(x)$ satisfies $h(y, x^*(x)) \leq \max_{x' \in \Psi(y,x)} h(y,x')$ and therefore the functional ψ_x given by

$$\psi_x(h) = \int h(y, x^*(x)) f_X(x) dx \leq \int \max_{x' \in \Psi(y,x)} h(y,x') f_X(x) dx = \phi'_{F_{Y|X}}(h).$$

Since h was arbitrary, this implies $\{\psi_{x^*} : x^*(x) \in \Psi(y,x) \forall x\} \subset \{\psi \leq \phi'_{F_{Y|X}}\}$ and therefore $\{\psi_{x^*} + \phi'_{f_X} : x^*(x) \in \Psi(y,x) \forall x\} \subset \{\psi \leq \phi'_{\theta_0}\}$.

For the second claim, note that for fixed y, $h(y,\cdot) \in \mathcal{C}([0,1])$. We first argue that any $\psi \in \{\psi \leq \phi'_{F_{Y|X}}\}$ necessarily is a positive linear functional. Suppose by contradiction that ψ is not positive, that is, there exists some h satisfying $h(y,x) \geq 0$ for all x and $\psi(h) < 0$. By linearity, this also implies $\psi(-h) > 0$. However, since $\phi'_{F_{Y|X}}(-h)(y) \leq 0$, this implies $\phi_{F_{Y|X}}(-h)(y) < \psi(-h)$, a contradiction.

This implies by the Riesz-Markov-Kakutani representation theorem, that for any $\psi \in \{\psi \leq \phi'_{F_{Y|X}}\}$, there is a unique positive and finite Borel measure μ on the Borel

σ-field $\mathcal{B}([0,1])$ such that

$$\psi(h) = \int_0^1 h(y,x)d\mu(x).$$

Finally, evaluating ψ at $h(y,x) = 1$ for all $x \in [0,1]$, implies $\psi(h) = \mu([0,1]) \leq \phi_{F_{Y|X}}(h)(y) = 1$. The claim follows. □

Assumptions

Assumption 13. *(i) Let U_1, U_2 be i.i.d. uniformly distributed on $[0,1]$ and set $X = F_X^{-1}(U_1)$ and $Y = F_{Y|X}^{-1}(U_2 \,|\, X_1)$. Further, let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be a random sample following the same construction for a random sample of $(U_{i,1}, U_{i,2})$, $i = 1, \ldots, n$.*

(ii) X_i has an $(\ell+1)$-times continuously differentiable density $f : \mathbb{R} \to \mathbb{R}$ which is bounded away from zero and infinity on $[0,1]$

$$0 < c_f \leq f(x) \leq C_f < \infty, \qquad \forall x \in [0,1]$$

(iii) $F_{Y|X}$ is $(\ell+1)$-times continuously differentiable whose $(\ell+1)$th derivative is Lipschitz continuous, satisfies the well-separatedness condition (1.23) for $\rho = 2$, $\alpha = 1$ and

$$\iint_{\mathbb{R}^2} \left| \frac{\partial f_{Y|X}(Y \,|\, X)}{\partial X} \right| dY\, dX < \infty.$$

(iv) K is a twice continuously differentiable and symmetric ℓth order kernel with support $[-1,1]$.

(v) $h_n \to 0$, $nh_n^3 \to \infty$, $nh_n^5 / \log^{24} n \to \infty$, $nh_n^{2\ell+3} \to 0$, $\sqrt{n^{-1}h_n^{-2}} \to 0$ at polynomial rates in n.

These assumptions are rather standard assumptions in nonparametric regression. Assumption (i) can be thought of as requiring an i.i.d. sample only. It is written in this way in order to fit into the framework of the Rio-Massart coupling. The well-separatedness condition in (1.23) is both needed for the differentiability of ϕ and also in the derivation of the rate of the derivative estimator.

Strong Approximation when the derivative is nonlinear

In this section, we derive a joint strong approximation of the NW estimator of the conditional distribution and the KDE. While marginal strong approximation results would be sufficient for the application of Theorem 1, we need a joint coupling in order to verify the rate requirements in Theorem 2.

We construct the strong approximation using the Rio-Massart coupling. This coupling requires that the probability space $(\Omega, \mathcal{A}, \mathrm{P})$ is sufficiently rich in that it supports a random variable which is uniformly distributed on $[0,1]$ and which is independent of the data. This is not a strong restriction as we can always enrich an original space by taking the product with $[0,1]$ equipped with the uniform measure over Borel sets of $[0,1]$. Therefore, we will assume in the following that $(\Omega, \mathcal{A}, \mathrm{P})$ is sufficiently rich in the above sense.

We construct a strong approximation of the estimators instead of the derivative evaluated at the estimators. While we can represent the derivative in this example as a supremum over linear functionals of the estimators, the entropy of these linear functionals is in general too large to be a VC class as is required by the Rio-Massart coupling. While the couplings in Koltchinskii (1994) and Chernozhukov et al. (2014a) are also applicable function classes with such a large entropy, they result in slower rates of convergence.

Lemma 32. *Suppose Assumption 13 holds. Then, there exists a centered Gaussian process $Z_n = (Z_{F,n}, Z_{f,n})$ with covariance functions*

$$\begin{aligned}
\Sigma_F(y,x,y',x') &= \mathrm{Cov}(Z_{F,n}(y,x), Z_{F,n}(y',x')) \\
&= \frac{1}{nh_n^2 f_X(x) f_X(x')} \mathrm{E}[K_{h_n}(X_i - x) K_{h_n}(X_i - x') \varepsilon_i(y) \varepsilon_i(y')] \qquad (1.28) \\
\Sigma_f(x,x') &= \mathrm{Cov}(Z_{f,n}(x), Z_{f,n}(x')) \\
&= \frac{1}{nh_n^2} \{ \mathrm{E}[K_{h_n}(X_i - x) K_{h_n}(X_i - x')] \\
&\qquad - \mathrm{E}[K_{h_n}(X_i - x)] \, \mathrm{E}[K_{h_n}(X_i - x')] \} \qquad (1.29) \\
\Sigma_{F,f}(y,x,x') &= \mathrm{Cov}(Z_{F,n}(y,x), Z_{f,n}(x')) = 0, \qquad (1.30)
\end{aligned}$$

and almost surely continuous sample paths so that

$$\|(\hat{F}_n - F_{Y|X}) - Z_{F,n}\|_\infty \vee \|(\hat{f}_n - f_X) - Z_{f,n}\|_\infty$$

$$=O_p\bigg(\bigg(\frac{\log^4 n}{n^3 h_n^4}\bigg)^{1/4}+\frac{\sqrt{\log^5 n}}{nh_n}+\sqrt{\frac{h_n\log n}{n}}+h_n^\ell\bigg).$$

Proof. Using the construction in Assumption 13, we can reduce the coupling problem to a coupling of the uniform empirical process $h_n^{-1}\mathbb{G}_n(h_{x,y,d})$, where

$$\begin{aligned}h_{x,y,F}(u_1,u_2)=&\frac{1}{f_X(x)}K_{h_n}(F_X^{-1}(u_1)-x)\big(\mathbb{1}\{F_{Y|X}^{-1}(u_2\,|\,F_X^{-1}(u_1))\le y\}-u_2\big)\\ h_{x,y,f}(u_1)=&K_{h_n}(F_X^{-1}(u_1)-x)-\mathrm{E}[K_{h_n}(F_X^{-1}(u_1)-x)].\end{aligned}$$

We collect all these functions in $\mathcal{H}_n=\mathcal{H}_{n,F}\cup\mathcal{H}_{n,f}$, where $\mathcal{H}_{n,d}=\{h_{x,y,d}:(x,y)\in[0,1]^2\}$, $d\in\{f,F\}$.

Indeed, by Lemma 33 and standard bounds on the bias of the KDE,

$$\begin{aligned}\sup_{x,y}\big|(\hat{F}_n(y,x)-F_{Y|X}(y,x))-(nh_n^2)^{-1/2}\mathbb{G}_n(h_{x,y,F})\big|&=O_p\bigg(\sqrt{\frac{h_n\log n}{n}}+h_n^\ell\bigg)\\ \sup_{x}\big|(\hat{f}_n(x)-f_X(x))-(nh_n^2)^{-1/2}\mathbb{G}_n(h_{x,y,f})\big|&=O_p\big(h_n^\ell\big).\end{aligned}$$

By Lemma 34, $\mathcal{H}_n$ is a VC class with functions which are uniformly bounded over n. By Lemma 35, the total variation norm of $\mathcal{H}_n$ is of order h_n^{-1}. Further, $\mathcal{H}_n$ clearly satisfies the measurability conditions of Theorem 1.1 in Rio (1994). Thus, by Theorem 1.1 in Rio (1994), there exists a Brownian bridge B_n indexed by $\mathcal{H}_n$ with almost surely continuous trajectories on $(\mathcal{H}_n,\|\cdot\|_{L_1([0,1]^2)})$ such that, for any positive $t\ge C\log n$,

$$\mathrm{P}\bigg(\sqrt{n}\sup_{h\in\mathcal{H}_n}|\mathbb{G}_n(h)-B_n(h)|\ge C\sqrt{n^{1/2}\sup_{h\in\mathcal{H}_n}\|h\|_{TV}t}+C\sqrt{\log n t}\bigg)\le\exp(-t),$$

where C is a positive constant depending only on $d(\mathcal{H}_n)$ and $C(\mathcal{H}_n)$. Take $\sqrt{nh_n}Z_{F,n}(y,x)=h_n^{-1/2}B_n(h_{x,y,F})$ and $\sqrt{nh_n}Z_{f,n}(x)=h_n^{-1/2}B_n(h_{x,y,f})$, for all y,x and define

$$\begin{aligned}\Delta_n=\sup_{y,x}|&h_n^{-1/2}\mathbb{G}_n(h_{x,y,F})-\sqrt{nh_n}Z_{F,n}(y,x)|\\ &\vee\sup_x|h_n^{-1/2}\mathbb{G}_n(h_{x,y,f})-\sqrt{nh_n}Z_{f,n}(x)|.\end{aligned}$$

Then, the above result implies for all $\eta > 0$ sufficiently large

$$\mathrm{P}\left(\Delta_n \geq C\left\{\sqrt{\frac{\eta \log n}{\sqrt{n h_n^2}}} + \eta\sqrt{\frac{\log^3 n}{n h_n}}\right\}\right) \leq n^{-\eta}.$$

In particular, for η a sufficiently large multiple of $\log n$,

$$\sup_{y,x}|h_n^{-1/2}\mathbb{G}_n(h_{x,y}) - \sqrt{n h_n} Z_n(y,x)| = O_p\left(\left(\frac{\log^4 n}{n h_n^2}\right)^{1/4} + \sqrt{\frac{\log^5 n}{n h_n}}\right)$$

The rate of the coupling now follows together with the linearization in Lemma 33. The form of the covariance functions in (1.28) and (1.29) follow directly from the empirical process representation. The covariance function (1.30) satisfies

$$\Sigma_{F,f}(y,x,x') = \frac{1}{n h_n^2 f_X(x)} \mathrm{E}[K_{h_n}(X_i - x)\varepsilon_i(y,x)K_{h_n}(X_i - x')] = 0$$

by the law of iterated expectations.

Finally, we show that Z_n has almost surely continuous sample paths. We use the substitution

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = T(X,Y) = \begin{pmatrix} F_X(X) \\ F_{Y|X}^{-1}(Y \mid X) \end{pmatrix}$$

with Jacobian

$$DT(X,Y) = \begin{pmatrix} \frac{\partial F_X(X)}{\partial X} & 0 \\ \frac{\partial F_{Y|X}(Y \mid X)}{\partial X} & \frac{\partial F_{Y|X}(Y \mid X)}{\partial Y} \end{pmatrix} = \begin{pmatrix} f_X(X) & 0 \\ \frac{\partial F_{Y|X}(Y \mid X)}{\partial X} & f_{Y|X}(Y \mid X) \end{pmatrix}.$$

Thus, by the transformation formula, for any $y, y', x, x' \in [0,1]$

$$\begin{aligned} \|h_{x,y,F} - h_{x',y',F}\|_{L_1([0,1]^2)} &= \mathrm{E}\left[\left|\frac{1}{f_X(x)}K_{h_n}(X_i - x)\varepsilon_i(y,x)\right.\right. \\ &\qquad \left.\left. - \frac{1}{f_X(x')}K_{h_n}(X_i - x')\varepsilon_i(y',x')\right|\right] \\ \|h_{x,f} - h_{x',f}\|_{L_1([0,1]^2)} &= \mathrm{E}[|K_{h_n}(X_i - x) - K_{h_n}(X_i - x')|]. \end{aligned}$$

The second term is Lipschitz continuous and therefore $Z_{f,n}$ has almost surely con-

tinuous sample paths. For the first term, we bound

$$\begin{aligned}
&\mathrm{E}\left[\left|\frac{1}{f_X(x)}K_{h_n}(X_i-x)\varepsilon_i(y,x)-\frac{1}{f_X(x')}K_{h_n}(X_i-x')\varepsilon_i(y',x')\right|\right]\\
\leq&\,\mathrm{E}\left[\left|\frac{1}{f_X(x)}K_{h_n}(X_i-x)\varepsilon_i(y,x)-\frac{1}{f_X(x)}K_{h_n}(X_i-x')\varepsilon_i(y',x)\right|\right]\\
&+\mathrm{E}\left[\left|\frac{1}{f_X(x)}K_{h_n}(X_i-x)\varepsilon_i(y',x)-\frac{1}{f_X(x')}K_{h_n}(X_i-x')\varepsilon_i(y',x')\right|\right]
\end{aligned}$$

For the first term on the right-hand side, we have

$$\begin{aligned}
&\mathrm{E}\left[\left|\frac{1}{f_X(x)}K_{h_n}(X_i-x)\varepsilon_i(y,x)-\frac{1}{f_X(x)}K_{h_n}(X_i-x')\varepsilon_i(y',x)\right|\right]\\
=&\,\mathrm{E}\left[\left|\frac{1}{f_X(x)}K_{h_n}(X_i-x)\{\varepsilon_i(y,x)-\varepsilon_i(y',x)\}\right|\right]\leq\frac{\|K\|_\infty\|f_{Y,X}\|_\infty}{h_n\min_x f_X(x)}|y-y'|
\end{aligned}$$

and the second term on the right-hand side is similarly Lipschitz continuous uniformly over y' by smoothness properties of $f_X, F_{Y|X}$ and K. Thus, $Z_{F,n}$ also has almost surely continuous sample paths. The claim follows. □

Auxiliary results for the Rio-Massart Coupling

The following Lemma shows that the Nadaraya-Watson estimator is asymptotically linear.

Lemma 33. *Suppose that Assumption 13 holds. Then,*

$$\sup_{x,y}\left|\sqrt{nh_n}(\hat{F}_n(y\,|\,x)-F_{Y|X}(y\,|\,x))-h_n^{-1/2}\mathbb{G}_n(h_{x,y})\right|=O_p\big(h_n\sqrt{\log n}+\sqrt{nh_n^{2\ell+1}}\big).$$

Proof. Let $\varepsilon_i(y)=\mathbb{1}(Y_i\leq y)-F_{Y|X}(y\,|\,X_i)$, for $i=1,\dots,n$ and decompose the estimator,

$$\begin{aligned}
\hat{F}_n(y|x)=&\frac{1}{nh_n\hat{f}_n(x)}\sum_{i=1}^n K_h(X_i-x)F_{Y|X}(y\,|\,X_i)+\frac{1}{nh_n\hat{f}_n(x)}\sum_{i=1}^n K_h(X_i-x)\varepsilon_i(y)\\
=&\frac{1}{nh_nf_X(x)}\sum_{i=1}^n\big\{K_h(X_i-x)\big\{F_{Y|X}(y\,|\,X_i)-F_{Y|X}(y\,|\,x)\big\}\\
&\qquad\qquad-\mathrm{E}[K_h(X_i-x)\{F_{Y|X}(y\,|\,X_i)-F_{Y|X}(y\,|\,x)\}]\big\}\\
&+\frac{1}{nh_nf_X(x)}\sum_{i=1}^n K_h(X_i-x)\varepsilon_i(y)
\end{aligned}$$

$$
\begin{aligned}
&+ F_{Y|X}(y \,|\, x) \\
&+ \frac{1}{h_n f_X(x)} \mathrm{E}[K_h(X_i - x)\{F_{Y|X}(y \,|\, X_i) - F_{Y|X}(y \,|\, x)\}] \\
&+ \left(\frac{f_X(x)}{\hat{f}_n(x)} - 1\right) \frac{1}{n h_n f_X(x)} \sum_{i=1}^{n} K_h(X_i - x)\big\{F_{Y|X}(y \,|\, X_i) - F_{Y|X}(y \,|\, x)\big\} \\
&+ \left(\frac{f_X(x)}{\hat{f}_n(x)} - 1\right) \frac{1}{n h_n f_X(x)} \sum_{i=1}^{n} K_h(X_i - x)\varepsilon_i(y).
\end{aligned}
$$

By boundedness and standard arguments, uniformly over y and x,

$$
\begin{aligned}
&\sup_{y,x}\left|\frac{1}{n h_n f_X(x)} \sum_{i=1}^{n} K_h(X_i - x)\big\{F_{Y|X}(y \,|\, X_i) - F_{Y|X}(y \,|\, x)\big\}\right. \\
&\qquad \left. - \frac{1}{h_n f_X(x)} \mathrm{E}[K_h(X_i - x)\{F_{Y|X}(y \,|\, X_i) - F_{Y|X}(y \,|\, x)\}]\right| = O_p\left(\sqrt{\frac{h_n \log n}{n}}\right) \\
&\sup_{y,x}\left|\frac{1}{h_n f_X(x)} \mathrm{E}[K_h(X_i - x)\{F_{Y|X}(y \,|\, X_i) - F_{Y|X}(y \,|\, x)\}]\right| = O_p(h_n^{\ell}) \\
&\sup_{y,x}\left|\frac{1}{n h_n f_X(x)} \sum_{i=1}^{n} K_h(X_i - x)\varepsilon_i(y)\right| = O_p\left(\sqrt{\frac{\log n}{n h_n}}\right) \\
&\sup_{x}\left|\frac{f_X(x)}{\hat{f}_n(x)} - 1\right| = O_p\left(\sqrt{\frac{\log n}{n h_n}} + h_n^{\ell}\right)
\end{aligned}
$$

implying

$$
\begin{aligned}
&\sup_{y,x}\left|\sqrt{n h_n}\big(\hat{F}_n(y|x) - F_{Y|X}(y \,|\, x)\big) - \frac{1}{\sqrt{n h_n} f_X(x)} \sum_{i=1}^{n} K_h(X_i - x)\varepsilon_i(y)\right| \\
=&O_p\big(h_n \sqrt{\log n} + \sqrt{n h_n^{2s+1}}\big).
\end{aligned}
$$

□

The following Lemma gives a bound on the VC dimension used in the coupling of the Nadaraya-Watson estimator.

Lemma 34. *There exist constants $0 < C, d < \infty$ so that, for all n,*

$$\sup_{Q} N(\varepsilon, \mathcal{H}_n, L_1(Q)) \leq C\varepsilon^{-d}.$$

Proof. This proof follows closely the arguments in the proof of Theorem 8 in Chernozhukov et al. (2013b). Note that $\mathcal{H}_{n,F}$ is the product of $\{\mathbb{1}(F_{Y|X}^{-1}(u_2 \,|\, F_X^{-1}(u_1)) \leq$

$1) - u_2 : y \in \mathbb{R}\}$ with VC-dimension 1 and $\{K_h(F_X^{-1}(u_1) - x)/f_X(x) : x \in [0,1]\}$ is a VC class uniformly over n by Lemma 4.1 in Rio (1994). Thus, the $\mathcal{H}_{n,F}$ is a VC class by Lemma A.1 in Ghosal et al. (2000).

Further, $\mathcal{H}_{n,f}$ is a VC class by Lemma 4.1 in Rio (1994). Finally, since the covering numbers of a union of two classes can be bounded by the product of the bounds on the separate covering numbers, it follows that $\mathcal{H}_n$ is a VC class. □

The last Lemma in this section bounds the total variation norm. For a function $f : \mathbb{R}^d \to \mathbb{R}$, let $\|f\|_{TV}$ denote the total variation norm which is given by

$$\|f\|_{TV} := \sup_{g \in \mathcal{D}_c([0,1]^d)} \int_{\mathbb{R}^d} f(x) \frac{\text{div} g(x)}{\|g\|_\infty} dx,$$

where $\mathcal{D}_c([0,1]^d)$ is the space of $\mathcal{C}^\infty$ functions taking values in $\mathbb{R}^d$ with compact support included in $[0,1]^d$, and where $\text{div} g(x)$ is the divergence of $g(x)$.

Lemma 35. *Suppose that f_X is bounded away from zero, K is symmetric and continuously differentiable with compact support and*

$$\iint_{\mathbb{R}^2} \left| \frac{\partial f_{Y|X}(Y \mid X)}{\partial X} \right| dY dX < \infty.$$

Then, there exists a constant $C < \infty$ which only depends on the lower bound on f_X, the kernel K and the smoothness condition on the conditional density of Y given X such that

$$\sup_{h \in \mathcal{H}_n} \|h\|_{TV} \leq \frac{C}{h_n}.$$

Proof. For any $g = (g_1, g_2) \in \mathcal{D}_c([0,1]^2)$,

$$\begin{aligned} &\int_0^1 \int_0^1 h_{x,y}(u_1, u_2) \frac{\text{div} g(u_1, u_2)}{\|g\|_\infty} du_2 du_1 \\ =&\frac{1}{\|g\|_\infty} \int_0^1 \int_0^1 h_{x,y}(u_1, u_2) \left\{ \frac{\partial g_1(u_1, u_2)}{\partial u_1} + \frac{\partial g_2(u_1, u_2)}{\partial u_2} \right\} du_2 du_1 \end{aligned}$$

First consider

$$\int_0^1 \int_0^1 h_{x,y}(u_1, u_2) \frac{\partial g_1(u_1, u_2)}{\partial u_1} du_2 du_1$$

$$= \frac{1}{f_X(x)} \int_0^1 \int_0^1 K_{h_n}(F_X^{-1}(u_1) - x) \mathbb{1}\{F_{Y|X}^{-1}(u_2 \mid F_X^{-1}(u_1)) \leq y\} \frac{\partial g_1(u_1, u_2)}{\partial u_1} du_2 du_1.$$

We use the substitution

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = T(X, Y) = \begin{pmatrix} F_X(X) \\ F_{Y|X}^{-1}(Y \mid X) \end{pmatrix}$$

with Jacobian

$$DT(X, Y) = \begin{pmatrix} \frac{\partial F_X(X)}{\partial X} & 0 \\ \frac{\partial F_{Y|X}(Y \mid X)}{\partial X} & \frac{\partial F_{Y|X}(Y \mid X)}{\partial Y} \end{pmatrix} = \begin{pmatrix} f_X(X) & 0 \\ \frac{\partial F_{Y|X}(Y \mid X)}{\partial X} & f_{Y|X}(Y \mid X) \end{pmatrix}.$$

Thus, by the transformation formula

$$\begin{aligned}
&\int_0^1 \int_0^1 K_{h_n}(F_X^{-1}(u_1) - x) \mathbb{1}\{F_{Y|X}^{-1}(u_2 \mid F_X^{-1}(u_1)) \leq y\} \frac{\partial g_1(u_1, u_2)}{\partial u_1} du_2 du_1 \\
&= \iint_{\mathbb{R}^2} K_{h_n}(X - x) \mathbb{1}\{Y \leq y\} \frac{\partial g_1(F_X(X), F_{Y|X}(Y \mid X))}{\partial u_1} f_{Y|X}(Y \mid X) f_X(X) dY dX \\
&= \mathrm{E}\left[K_{h_n}(X - x) \mathbb{1}\{Y \leq y\} \frac{\partial g_1(F_X(X), F_{Y|X}(Y \mid X))}{\partial u_1} \right].
\end{aligned}$$

Further, it holds

$$\begin{aligned}
\frac{\partial g_1(F_X(X), F_{Y|X}(Y \mid X))}{\partial X} =& \frac{\partial g_1(F_X(X), F_{Y|X}(Y \mid X))}{\partial u_1} f_X(X) \\
&+ \frac{\partial g_1(F_X(X), F_{Y|X}(Y \mid X))}{\partial u_2} \frac{\partial F_{Y|X}(Y \mid X)}{\partial X} \\
\frac{\partial g_1(F_X(X), F_{Y|X}(Y \mid X))}{\partial Y} =& \frac{\partial g_1(F_X(X), F_{Y|X}(Y \mid X))}{\partial u_2} f_{Y|X}(Y \mid X)
\end{aligned}$$

and therefore

$$\begin{aligned}
&\mathrm{E}\left[K_{h_n}(X - x) \mathbb{1}\{Y \leq y\} \frac{\partial g_1(F_X(X), F_{Y|X}(Y \mid X))}{\partial u_1} \right] \\
&= \mathrm{E}\left[K_{h_n}(X - x) \mathbb{1}\{Y \leq y\} \left\{ \frac{\partial g_1(F_X(X), F_{Y|X}(Y \mid X))}{\partial u_1} \right.\right. \\
&\qquad \left.\left. + \frac{\partial g_1(F_X(X), F_{Y|X}(Y \mid X))}{\partial u_2} \frac{\partial F_{Y|X}(Y \mid X)}{\partial X} \frac{1}{f_X(X)} \right\} \right] \qquad (1.31)
\end{aligned}$$

$$- \mathrm{E}\bigg[K_{h_n}(X-x)\mathbb{1}\{Y \le y\}\frac{\partial g_1(F_X(X), F_{Y|X}(Y \,|\, X))}{\partial u_2}\frac{\partial F_{Y|X}(Y \,|\, X)}{\partial X}\frac{1}{f_X(X)}\bigg]. \tag{1.32}$$

In order to bound (1.31), we decompose

$$\begin{aligned}
&\mathrm{E}\bigg[K_{h_n}(X-x)\mathbb{1}\{Y \le y\}\bigg\{\frac{\partial g_1(F_X(X), F_{Y|X}(Y \,|\, X))}{\partial u_1} \\
&\qquad + \frac{\partial g_1(F_X(X), F_{Y|X}(Y \,|\, X))}{\partial u_2}\frac{\partial F_{Y|X}(Y \,|\, X)}{\partial X}\frac{1}{f_X(X)}\bigg\}\bigg] \\
&= \iint K_{h_n}(X-x)\mathbb{1}\{Y \le y\}\frac{\partial g_1(F_X(X), F_{Y|X}(Y \,|\, X))}{\partial X} f_{Y|X}(Y \,|\, X)dY\,dX \\
&= \iint K_{h_n}(X-x)\mathbb{1}\{Y \le y\}\bigg\{\frac{\partial g_1(F_X(X), F_{Y|X}(Y \,|\, X))}{\partial X} f_{Y|X}(Y \,|\, X) \\
&\qquad + g_1(F_X(X), F_{Y|X}(Y \,|\, X))\frac{\partial f_{Y|X}(Y \,|\, X)}{\partial X}\bigg\}dY\,dX \qquad (1.33) \\
&\quad - \iint K_{h_n}(X-x)\mathbb{1}\{Y \le y\} g_1(F_X(X), F_{Y|X}(Y \,|\, X))\frac{\partial f_{Y|X}(Y \,|\, X)}{\partial X}dY\,dX. \qquad (1.34)
\end{aligned}$$

Note that

$$\begin{aligned}
&\frac{\partial g_1(F_X(X), F_{Y|X}(Y \,|\, X)) f_{Y|X}(Y \,|\, X)}{\partial X} \\
=&\frac{\partial g_1(F_X(X), F_{Y|X}(Y \,|\, X))}{\partial X} f_{Y|X}(Y \,|\, X) + g_1(F_X(X), F_{Y|X}(Y \,|\, X))\frac{\partial f_{Y|X}(Y \,|\, X)}{\partial X}
\end{aligned}$$

Thus, we can simplify (1.33) by partial integration

$$\begin{aligned}
&\iint K_{h_n}(X-x)\mathbb{1}\{Y \le y\}\bigg\{\frac{\partial g_1(F_X(X), F_{Y|X}(Y \,|\, X))}{\partial X} f_{Y|X}(Y \,|\, X) \\
&\qquad + g_1(F_X(X), F_{Y|X}(Y \,|\, X))\frac{\partial f_{Y|X}(Y \,|\, X)}{\partial X}\bigg\}dY\,dX \\
=&K_{h_n}(X-x)\int_{-\infty}^{y} g_1(F_X(X), F_{Y|X}(Y \,|\, X)) f_{Y|X}(Y \,|\, X)dY\bigg|_{\mathbb{R}} \\
&\qquad - \iint \frac{\partial K_{h_n}(X-x)}{\partial X}\mathbb{1}\{Y \le y\} g_1(F_X(X), F_{Y|X}(Y \,|\, X)) f_{Y|X}(Y \,|\, X)dY\,dX \\
=&- \iint \frac{\partial K_{h_n}(X-x)}{\partial X}\mathbb{1}\{Y \le y\} g_1(F_X(X), F_{Y|X}(Y \,|\, X)) f_{Y|X}(Y \,|\, X)dY\,dX,
\end{aligned}$$

where we used that the first term in the integration by parts formula vanishes since the kernel has compact support. We can upper bound the absolute value of the

remaining term with

$$\begin{aligned}&\|g_1\|_\infty \iint \left|\frac{\partial K_{h_n}(X-x)}{\partial X}\right| \mathbb{1}\{Y \leq y\} f_{Y|X}(Y \mid X) dY dX \\ =&\|g_1\|_\infty \int \left|\frac{\partial K_{h_n}(X-x)}{\partial X}\right| F_{Y|X}(y \mid X) dX \leq \frac{C}{h_n} \|g_1\|_\infty \|\partial K/\partial x\|_\infty.\end{aligned}$$

Similarly, we can upper bound the absolute value of (1.34) by

$$\begin{aligned}&\|g_1\|_\infty \iint |K_{h_n}(X-x)| \mathbb{1}\{Y \leq y\} \left|\frac{\partial f_{Y|X}(Y \mid X)}{\partial X}\right| dY dX \\ \leq&\|g_1\|_\infty \|K\|_\infty \iint \left|\frac{\partial f_{Y|X}(Y \mid X)}{\partial X}\right| dY dX.\end{aligned}$$

It remains to bound (1.32). It holds,

$$\begin{aligned}&\mathrm{E}\left[K_{h_n}(X-x)\mathbb{1}\{Y \leq y\} \frac{\partial g_1(F_X(X), F_{Y|X}(Y \mid X))}{\partial u_2} \frac{\partial F_{Y|X}(Y \mid X)}{\partial X} \frac{1}{f_X(X)}\right] \\ =&\iint K_{h_n}(X-x)\mathbb{1}\{Y \leq y\} \frac{\partial g_1(F_X(X), F_{Y|X}(Y \mid X))}{\partial u_2} \frac{\partial F_{Y|X}(Y \mid X)}{\partial X} f_{Y|X}(Y \mid X) dY dX \\ =&\iint K_{h_n}(X-x)\mathbb{1}\{Y \leq y\} \frac{\partial g_1(F_X(X), F_{Y|X}(Y \mid X))}{\partial Y} \frac{\partial F_{Y|X}(Y \mid X)}{\partial X} dY dX.\end{aligned}$$

By integration by parts,

$$\begin{aligned}&\int_{-\infty}^{y} \mathbb{1}\{Y \leq y\} \frac{\partial g_1(F_X(X), F_{Y|X}(Y \mid X))}{\partial Y} \frac{\partial F_{Y|X}(Y \mid X)}{\partial X} dY \\ =&g_1(F_X(X), F_{Y|X}(Y \mid X)) \frac{\partial F_{Y|X}(Y \mid X)}{\partial X} \\ &- \int_{-\infty}^{y} g_1(F_X(X), F_{Y|X}(Y \mid X)) \frac{\partial f_{Y|X}(Y \mid X)}{\partial X} dY\end{aligned}$$

implying by similar arguments as above

$$\begin{aligned}&\left|\iint K_{h_n}(X-x)\mathbb{1}\{Y \leq y\} \frac{\partial g_1(F_X(X), F_{Y|X}(Y \mid X))}{\partial Y} \frac{\partial F_{Y|X}(Y \mid X)}{\partial X} dY dX\right| \\ \leq&2\|g_1\|_\infty \|K\|_\infty \iint \left|\frac{\partial f_{Y|X}(Y \mid X)}{\partial X}\right| dY dX.\end{aligned}$$

Thus, we have shown that

$$\int_0^1 \int_0^1 h_{x,y}(u_1, u_2) \frac{\partial g_1(u_1, u_2)}{\partial u_1} du_2 du_1 \leq C \|g_1\|_\infty$$

for a constant C which depends only on the kernel, the lower bound on f_X and the smoothness of the conditional density of Y given X.

It remains to bound

$$\begin{aligned}
&\int_0^1 \int_0^1 h_{x,y}(u_1, u_2) \frac{\partial g_2(u_1, u_2)}{\partial u_2} du_2 du_1 \\
=&\frac{1}{f_X(x)} \int_0^1 \int_0^1 K_{h_n}(F_X^{-1}(u_1) - x) \mathbb{1}\{F_{Y|X}^{-1}(u_2 \,|\, F_X^{-1}(u_1)) \leq y\} \frac{\partial g_2(u_1, u_2)}{\partial u_2} du_2 du_1 \\
=&\frac{1}{f_X(x)} \iint_{\mathbb{R}^2} K_{h_n}(X - x) \mathbb{1}\{Y \leq y\} \frac{\partial g_2(F_X(X), F_{Y|X}(Y \,|\, X))}{\partial u_2} f_{Y|X}(Y \,|\, X) f_X(X) dY dX \\
=&\frac{1}{f_X(x)} \iint_{\mathbb{R}^2} K_{h_n}(X - x) \mathbb{1}\{Y \leq y\} \frac{\partial g_2(F_X(X), F_{Y|X}(Y \,|\, X))}{\partial Y} f_X(X) dY dX \\
=&\frac{1}{f_X(x)} \int_{\mathbb{R}^2} K_{h_n}(X - x) g_2(F_X(X), F_{Y|X}(y \,|\, X)) f_X(X) dX,
\end{aligned}$$

where we have used the same transformations as above. Moreover, by similar bounds as above, we have

$$\int_0^1 \int_0^1 h_{x,y}(u_1, u_2) \frac{\partial g_2(u_1, u_2)}{\partial u_2} du_2 du_1 \leq C \|g_2\|_\infty$$

where the constant only depends on the lower bound of f_X and the kernel K. Thus,

$$\int_0^1 \int_0^1 h_{x,y}(u_1, u_2) \frac{\operatorname{div} g(u_1, u_2)}{\|g\|_\infty} du_2 du_1 \leq C \frac{\|g_1\|_\infty + \|g_2\|_\infty}{\|g\|_\infty} \leq C$$

proving the claim. □

Strong approximation when the derivative is linear

In the case when the derivative is linear, i.e., ϕ is fully differentiable at θ_0, we construct again the coupling using the Rio-Massart coupling, but this time we directly couple the derivative. This is feasible since in this case the complexity of the implied function class is sufficiently small. In some settings, such as for a decreasing conditional distribution function as discussed in Section 1.B.2, the

complexity is so small that even a Donsker theorem applies, and we could derive a limiting distribution. However, in other settings such as for an increasing conditional distribution function, the complexity is still too large to justify a Donsker theorem.

Lemma 36. *Suppose Assumption 13 holds and that the derivative is linear. Then, there exists a centered Gaussian $Z_n = (Z_{1,n}, Z_{2,n})$ with the same covariance functions as in Lemma 32 and almost surely continuous sample paths so that*

$$\begin{aligned}&\sup_y |\phi'_{\theta_0}(\hat{F}_n - F_{Y|X}, \hat{f}_n - f_X) - \phi'_{\theta_0}(Z_{1,n}, Z_{2,n})| \\ =&O_p\bigg(\bigg(\frac{\log^4 n}{n^3}\bigg)^{1/4} + \sqrt{\frac{\log^5 n}{n^2}} + \sqrt{\frac{h_n \log n}{n}} + h_n^\ell\bigg).\end{aligned}$$

If the derivative is L_2 continuous and

$$\begin{aligned}&\sup_y |\phi'_{\theta_0}(\hat{F}_n - F_{Y|X}, \hat{f}_n - f_X) - \phi'_{\theta_0}(Z_{1,n}, Z_{2,n})| \\ =&O_p\bigg(\bigg(\frac{\log^4 n}{n^3 h_n^4}\bigg)^{1/4} + \sqrt{\frac{\log^5 n}{n^2 h_n^2}} + \sqrt{\frac{h_n \log n}{n}} + h_n^\ell\bigg)\end{aligned}$$

if it is not.

Proof. If the derivative is linear, then there exists a function $x^*(x) \in [x, 1]$ for all $x \in [0, 1]$ so that

$$\phi'_{F_{Y|X}}(h)(y) = \int h(y, x^*(x)) f_X(x) dx.$$

By Lemma 30, this selection x^* is càdlàg and weakly monotone. Moreover, x^* is strictly increasing iff the derivative is L_2-continuous. Further, by Lemma 33, we can linearize the conditional distribution estimator

$$\sup_{x,y} \big|\sqrt{nh_n}(\hat{F}_n(y \,|\, x) - F_{Y|X}(y \,|\, x)) - h_n^{-1/2}\mathbb{G}_n(h_{x,y})\big| = O_p\big(h_n\sqrt{\log n} + \sqrt{nh_n^{2\ell+1}}\big).$$

This implies

$$\begin{aligned}&\int (\hat{F}_n(y, x^*(x)) - F_{Y|X}(y, x^*(x))) f_X(x) dx \\ =&\frac{1}{n}\sum_{i=1}^n \int \frac{1}{h_n} K_{h_n}(X_i - x^*(x)) \{\mathbb{1}(Y_i \le y) - F_{Y|X}(y, x^*(x))\} \frac{f_X(x)}{f_X(x^*(x))} dx\end{aligned}$$

$$+ O_p\left(\sqrt{\frac{h_n \log n}{n}} + h_n^\ell\right)$$

and

$$\begin{aligned}&\int F_{Y|X}(y,x)\{\hat{f}_n(x) - f_X(x)\}dx \\ =&\frac{1}{n}\sum_{i=1}^n \int \frac{1}{h_n}\{K_{h_n}(X_i - x) - \mathrm{E}[K_{h_n}(X_i - x)]\}F_{Y|X}(y,x)dx + O_p(h_n^\ell),\end{aligned}$$

where we used the usual bound on the bias of the KDE in the second equation. Note that this linearized process has mean zero. Further, the function classes

$$\begin{aligned}\mathcal{F}_n = \Bigg\{(X_i,Y_i) \mapsto &\int \frac{1}{h_n}K_{h_n}(X_i - x^*(x))\{\mathbb{1}(Y_i \leq y) - F_{Y|X}(y,x^*(x))\}\frac{f_X(x)}{f_X(x^*(x))}dx \\ &+ \int \frac{1}{h_n}\{K_{h_n}(X_i - x) - \mathrm{E}[K_{h_n}(X_i - x)]\}F_{Y|X}(y,x)dx : y \in [0,1]\Bigg\}\end{aligned} \tag{1.35}$$

are of VC-type uniformly over n as shown in Lemma 37 below. However, for the bound on the total variation norm, we need a case distinction. If x^* is strictly increasing, the total variation norm is uniformly bounded over n and if x^* is not strictly increasing, the total variation norm is of order h_n^{-1}. Both claims follow by similar arguments as in Lemma 35. Hence, we can apply the Rio-Massart coupling as in the proof of Lemma 32. Consider first the case that x^* is strictly increasing. In this case, by Theorem 1.1 in Rio (1994), there exists a Brownian bridge B_n indexed by $\mathcal{F}_n$ such that, for any positive $t \geq C \log n$,

$$\mathrm{P}\left(\sqrt{n}\sup_{f\in\mathcal{F}_n}|\mathbb{G}_n(f) - B_n(f)| \geq C\sqrt{n^{1/2}t} + C\sqrt{\log n t}\right) \leq \exp(-t).$$

Take $\sqrt{n}Y_n(y) = B_n(f_y)$, for all $y \in [0,1]$. Then, the above result implies for all t sufficiently large

$$\sup_{y\in[0,1]}|\mathbb{G}_n(f_y) - Y_n(y)| = O_p\left(\left(\frac{\log^4 n}{n}\right)^{1/4} + \sqrt{\frac{\log^5 n}{n}}\right).$$

Together with the above linearization, this implies

$$\begin{aligned}&\sup_y|\phi'_{\theta_0}(\hat{F}_n - F_{Y|X}, \hat{f}_n - f_X)(y) - Y_n(y)|\\ =&O_p\bigg(\bigg(\frac{\log^4 n}{n^3}\bigg)^{1/4} + \sqrt{\frac{\log^5 n}{n}} + \sqrt{\frac{h_n \log n}{n}} + h_n^\ell\bigg).\end{aligned}$$

In the case where x^* is not strictly increasing, the same arguments imply the existence a Gaussian process Y_n such that

$$\begin{aligned}&\sup_y|\phi'_{\theta_0}(\hat{F}_n - F_{Y|X}, \hat{f}_n - f_X)(y) - Y_n(y)|\\ =&O_p\bigg(\bigg(\frac{\log^4 n}{n^3 h_n^4}\bigg)^{1/4} + \sqrt{\frac{\log^5 n}{n h_n}} + \sqrt{\frac{h_n \log n}{n}} + h_n^\ell\bigg).\end{aligned}$$

Finally, the existence of the claimed process Z_n can be shown using the same arguments as in the ATE example. The claim follows. □

Lemma 37. *The function classes $\mathcal{F}_n$ defined in (1.35) is a VC class.*

Proof. First, consider the functions

$$\begin{aligned}&\int \frac{1}{h_n} K_{h_n}(X_i - x)\mathbb{1}(Y_i \leq y)\frac{f_X(x)}{f_X(x^*(x))}dx\\ =&\int \frac{1}{h_n} K_{h_n}(X_i - x)\frac{f_X(x)}{f_X(x^*(x))}dx\mathbb{1}(Y_i \leq y).\end{aligned}$$

This is just a product of a single function and an indicator function. Such indicator functions are well discussed in the literature. For instance, Theorem 2.6.4 and Example 2.6.1 in van der Vaart and Wellner (1996) imply

$$\sup_Q N(\varepsilon, \{\mathbb{1}(\cdot \leq y) : y \in [0,1]\}, L_1(Q)) \leq \frac{C}{\varepsilon},$$

for some universal constant C. Thus, these functions form a VC class by Lemma A.1 in Ghosal et al. (2000).

Next, consider the functions

$$g_y(X_i, Y_i) := -\int \frac{1}{h_n} K_{h_n}(X_i - x)F_{Y|X}(y, x)\frac{f_X(x)}{f_X(x^*(x))}dx$$

$$+ \int \frac{1}{h_n}\{K_{h_n}(X_i - x) - \mathrm{E}[K_{h_n}(X_i - x)]\}F_{Y|X}(y, x)dx$$

Since these are L-Lipschitz continuous in y, for some constant L independent of n, we have for any probability measure Q with finite support in $[0, 1]^2$,

$$\|g_y - g_{y'}\|_{L_1(Q)} \leq L|y - y'|.$$

Now, fix $\varepsilon > 0$ and let $0 = y_0, y_1, \ldots, y_N = 1$ be a uniform grid of $[0, 1]$ with mesh width $1/N$. Then, if $L/N \leq \varepsilon$, or equivalently, $N \geq L\varepsilon^{-1}$, there is for any $y \in [0, 1]$, some j so that

$$\|g_y - g_{y_j}\|_{L_1(Q)} \leq \varepsilon.$$

As ε and Q were arbitrary, we have

$$\sup_Q N(\varepsilon, \{g_y : y \in [0, 1]\}, L_1(Q)) \leq \frac{L}{\varepsilon}.$$

Finally, take minimal cover of the class of indicators of size $\lceil C/\varepsilon \rceil$ and a minimal cover of the g_y functions of size $\lceil L/\varepsilon \rceil$. Then we can approximate any function $f \in \mathcal{F}_n$ by

$$\begin{aligned}&\min_{i,j} \|f_y - \mathbb{1}(\cdot \leq y_i) - g_{y_j}\|_{L_1(Q)} \\ \leq &\min_i \|\mathbb{1}(\cdot \leq y) - \mathbb{1}(\cdot \leq y_i)\|_{L_1(Q)} + \min_j \|g_y - g_{y_j}\|_{L_1(Q)} \leq 2\varepsilon.\end{aligned}$$

Hence,

$$\sup_Q N(\varepsilon, \mathcal{F}_n, L_1(Q)) \leq \frac{CL}{(2\varepsilon)^2}$$

and the claim follows. □

Bounded Lipschitz norm of the NWE

Lemma 38. *Suppose Assumption 13 holds. Then,*

$$\|\hat{F}_n - F_{Y|X}\|_{BL} = O_p\left(\sqrt{\frac{\log n}{nh_n^3}}\right).$$

Proof. Fix some arbitrary $y \in [0,1]$ and decompose for any $x \in [0,1]$

$$\{\hat{F}_n(y,x) - F_{Y|X}(y,x)\} = \frac{1}{nh_n\hat{f}_n(x)} \sum_{i=1}^{n} K_{h_n}(X_i - x)\varepsilon_i(y) + \frac{1}{nh_n\hat{f}_n(x)} \sum_{i=1}^{n} K_{h_n}(X_i - x)\{F_{Y|X}(y,X_i) - F_{Y|X}(y,x)\},$$

where $\varepsilon_i(y) = \mathbb{1}(Y_i \le y) - F_{Y|X}(y,X_i)$.

Then, for any $x, x' \in [0,1]$,

$$\begin{aligned}
&|\{\hat{F}_n(y,x) - F_{Y|X}(y,x)\} - \{\hat{F}_n(y,x') - F_{Y|X}(y,x')\}| \\
\le &\left|\frac{1}{nh_n}\sum_{i=1}^{n}\left\{\frac{K_{h_n}(X_i - x)}{\hat{f}_n(x)} - \frac{K_{h_n}(X_i - x')}{\hat{f}_n(x')}\right\}\varepsilon_i(y)\right| && (1.36)\\
&+ \left|\frac{1}{nh_n}\sum_{i=1}^{n}\left\{\frac{K_{h_n}(X_i - x)}{\hat{f}_n(x)}\{F_{Y|X}(y,X_i) - F_{Y|X}(y,x)\}\right.\right. && (1.37)\\
&\qquad \left.\left. - \frac{K_{h_n}(X_i - x')}{\hat{f}_n(x')}\{F_{Y|X}(y,X_i) - F_{Y|X}(y,x')\}\right\}\right|. && (1.38)
\end{aligned}$$

For the first term on the right-hand side, (1.36), we have

$$\begin{aligned}
&\left|\frac{1}{nh_n}\sum_{i=1}^{n}\left\{\frac{K_{h_n}(X_i - x)}{\hat{f}_n(x)} - \frac{K_{h_n}(X_i - x')}{\hat{f}_n(x')}\right\}\varepsilon_i(y)\right| \\
\le &\left|\frac{1}{nh_n\hat{f}_n(x)}\sum_{i=1}^{n}\big\{K_{h_n}(X_i - x) - K_{h_n}(X_i - x')\big\}\varepsilon_i(y)\right| && (1.39)\\
&+ \left|\frac{1}{\hat{f}_n(x)} - \frac{1}{\hat{f}_n(x')}\right|\left|\frac{1}{nh_n}\sum_{i=1}^{n} K_{h_n}(X_i - x)\varepsilon_i(y)\right| && (1.40)
\end{aligned}$$

Regarding (1.39), by similar arguments as in the proof of Lemma 17,

$$\begin{aligned}
&\sup_{x \ne x'}\left|\frac{1}{nh_n\hat{f}_n(x)}\sum_{i=1}^{n}\big\{K_{h_n}(X_i - x) - K_{h_n}(X_i - x')\big\}\varepsilon_i(y)\right|/|x - x'| \\
&= O_p\left(\sqrt{\frac{\log n}{nh_n^3}} + h_n^{\ell}\right).
\end{aligned}$$

Regarding (1.40), by similar arguments as in the proof of Lemma 17,

$$\sup_{x\neq x'}\left|\frac{1}{|x-x'|}\frac{1}{\hat{f}_n(x)}-\frac{1}{\hat{f}_n(x')}\right| = O_p\left(\sqrt{\frac{\log n}{nh_n^3}}+h_n^\ell\right)$$
$$\sup_{y,x}\left|\frac{1}{nh_n}\sum_{i=1}^n K_{h_n}(X_i-x)\varepsilon_i(y)\right| = O_p\left(\sqrt{\frac{\log n}{nh_n}}+h_n^\ell\right).$$

We can deal with (1.38) similarly. We split

$$(1.38) \leq \left|\frac{1}{nh_n}\sum_{i=1}^n \frac{K_{h_n}(X_i-x)-K_{h_n}(X_i-x')}{\hat{f}_n(x)}\{F_{Y|X}(y,X_i)-F_{Y|X}(y,x)\}\right| \quad (1.41)$$
$$+\left|\frac{1}{nh_n}\sum_{i=1}^n \frac{K_{h_n}(X_i-x')}{\hat{f}_n(x)}\{F_{Y|X}(y,x')-F_{Y|X}(y,x)\}\right| \quad (1.42)$$
$$+\left|\frac{1}{\hat{f}_n(x)}-\frac{1}{\hat{f}_n(x')}\right|\left|\frac{1}{nh_n}\sum_{i=1}^n K_{h_n}(X_i-x')\{F_{Y|X}(y,X_i)-F_{Y|X}(y,x')\}\right|. \quad (1.43)$$

By the same arguments as above, it can be shown that (1.41) - (1.43) are of the same order as claimed in the Lemma. □

Application of the Delta Method

Theorem 13. *Suppose that Assumptions 13 hold. Fix $y \in [0,1]$. Then*

$$\lim_{n\to\infty}\sup_{t\in\mathbb{R}}|\mathrm{P}(\phi(\hat{F}_n,\hat{f}_n)(y)-\phi(F_{Y|X},f_X)(y)\leq t)-\mathrm{P}(\phi'_{\theta_0}(Z_n)(y)\leq t)|=0,$$

holds, where (Z_n) is a sequence of mean-zero Gaussian random variables with values in $\mathcal{C}([0,1]^2)\times\mathcal{C}([0,1])$ and covariance function given in (1.28)-(1.30), if

(i) ϕ'_{θ_0} is nonlinear and

$$\frac{\log^3 n}{nh_n^5}\to 0 \qquad \text{and} \qquad nh_n^{2\ell+1}\log n\to 0;$$

(ii) ϕ'_{θ_0} is linear and L_2-continuous and

$$\frac{\log^2 n}{nh_n^6}\to 0 \qquad \text{and} \qquad nh_n^{2\ell}\to 0;$$

(iii) ϕ'_{θ_0} *is linear but not* L_2*-continuous and*

$$\frac{\log^2 n}{nh_n^5} \to 0 \qquad \textit{and} \qquad nh_n^{2\ell+1} \to 0.$$

Proof. We only need to check the conditions of Theorem 2. ϕ is $(2, \|\cdot\|_*)$-Fréchet directionally differentiable at $(F_{Y|X}, f_X)$ by the well-separatedness condition (1.23) and Lemma 28 and therefore Assumption 1 holds. Assumption 2 (ii) holds by Lemma 38 with

$$a_n = \sqrt{\frac{\log n}{nh_n^3}}.$$

Moreover, Z_n is a centered Gaussian random vector in $\mathcal{C}([0,1]^2) \times \mathcal{C}([0,1])$ and tight since the latter is separable. Further $\mathcal{F} = \emptyset$ by Lemma 39. Thus, it only remains to verify Assumption 2 (iii) and the rate requirement $(a_n^2 + b_n)/\sigma_n \to 0$.

(i): If ϕ'_{θ_0} is nonlinear, Assumption 2 (iii) holds by Lemma 32 and Lipschitz continuity of the derivative with

$$b_n = \left(\frac{\log^4 n}{n^3 h_n^4}\right)^{1/4} + \frac{\sqrt{\log^5 n}}{nh_n} + \sqrt{\frac{h_n \log n}{n}} + h_n^\ell.$$

By Lemma 39 (iii), it holds $\sigma_n^{-1} = O(\sqrt{nh_n \log n})$. Since

$$\frac{a_n^2}{\sigma_n} = O\left(\sqrt{\frac{\log^3 n}{nh_n^5}}\right) = o(1)$$

and

$$\frac{b_n}{\sigma_n} = O\left(\left(\frac{\log^6 n}{nh_n^2}\right)^{1/4} + \sqrt{\frac{\log^6 n}{nh_n}} + \sqrt{h_n^2 \log n} + \sqrt{nh_n^{2\ell+1} \log n}\right) = o(1),$$

the claim follows by Theorem 2.

(ii): If ϕ'_{θ_0} is linear and L_2-continuous, Assumption 2 (iii) holds by Lemma 36 with

$$b_n = \left(\frac{\log^4 n}{n^3}\right)^{1/4} + \sqrt{\frac{\log^5 n}{n^2}} + \sqrt{\frac{h_n \log n}{n}} + h_n^\ell.$$

By Lemma 39 (i), it holds $\sigma_n^{-1} = O(\sqrt{n})$. Since

$$\frac{a_n^2}{\sigma_n} = O\left(\sqrt{\frac{\log^2 n}{nh_n^6}}\right) = o(1)$$

and

$$\frac{b_n}{\sigma_n} = O\left(\left(\frac{\log^4 n}{n}\right)^{1/4} + \sqrt{\frac{\log^5 n}{n}} + \sqrt{h_n \log n} + \sqrt{nh_n^{2\ell}}\right) = o(1),$$

the claim follows by Theorem 2.

(iii): If ϕ'_{θ_0} is linear but not L_2-continuous, Assumption 2 (iii) holds by Lemma 36 with

$$b_n = \left(\frac{\log^4 n}{n^3 h_n^4}\right)^{1/4} + \sqrt{\frac{\log^5 n}{n^2 h_n^2}} + \sqrt{\frac{h_n \log n}{n}} + h_n^{\ell}.$$

By Lemma 39 (ii), it holds $\sigma_n^{-1} = O(\sqrt{nh_n})$. Since

$$\frac{a_n^2}{\sigma_n} = O\left(\sqrt{\frac{\log^2 n}{nh_n^5}}\right) = o(1)$$

and

$$\frac{b_n}{\sigma_n} = O\left(\left(\frac{\log^4 n}{nh_n^2}\right)^{1/4} + \sqrt{\frac{\log^5 n}{nh_n}} + \sqrt{h_n^2 \log n} + \sqrt{nh_n^{2\ell+1}}\right) = o(1),$$

the claim follows by Theorem 2. □

Auxiliary results for the Delta method

The following Lemma gives lower bounds on the variance of the derivative and is needed in the application of Theorem 2.

Lemma 39. *Suppose that Assumptions 13 hold. Let* $\sigma_n^2 = \mathrm{Var}(\phi'_{\theta_0}(Z_n)(y))$, *for* $y \in [0,1]$. *If the derivative* ϕ'_{θ_0} *is*

(i) linear and x^* *is strictly increasing, we have* $\sigma_n^{-1} = O(\sqrt{n})$;

(ii) linear and x^* *is not strictly increasing, it holds* $\sigma_n^{-1} = O(\sqrt{nh_n})$;

(iii) nonlinear, we have $\sigma_n^{-1} = O(\sqrt{nh_n \log n})$.

In all cases $\mathcal{F} = \emptyset$*, for the set* $\mathcal{F}$ *given in Theorem 2.*

Proof. Fix $y \in (0,1)$.

(i): The variance is given by

$$\begin{aligned}\sigma_n^2 &= \operatorname{Var}(\phi'_{F_{Y|X}}(Z_{F,n})(y) + \phi'_f(Z_{f,n})(y)) \\ &= \operatorname{Var}(\phi'_{F_{Y|X}}(Z_{F,n})(y)) + \operatorname{Var}(\phi'_f(Z_{f,n})(y)) + 2\operatorname{Cov}(\phi'_{F_{Y|X}}(Z_{F,n})(y), \phi'_f(Z_{f,n})(y)).\end{aligned}$$

For the variance of the derivative with respect to the conditional distribution function, by Fubini's theorem

$$\begin{aligned}&\operatorname{Var}(\phi'_{F_{Y|X}}(Z_{F,n})(y)) \\ =&\iint \Sigma_F(y, x^*(x), y, x^*(x'))f_X(x)f_X(x')dxdx' \\ =&\frac{1}{nh_n^2}\iint \mathrm{E}[K_{h_n}(X_i - x^*(x))K_{h_n}(X_i - x^*(x'))\varepsilon_i^2(y)]\frac{f_X(x)}{f_X(x^*(x))}\frac{f_X(x')}{f_X(x^*(x'))}dxdx' \\ =&\frac{1}{n}\mathrm{E}\left[\left(\frac{1}{h_n}\int K_{h_n}(X_i - x^*(x))\frac{f_X(x)}{f_X(x^*(x))}dx\right)^2 F_{Y|X}(y, X_i)\{1 - F_{Y|X}(y, X_i)\}\right].\end{aligned}$$

For fixed $y \in (0,1)$, the conditional variance of ε_i is bounded away from zero as the joint density of X and Y is bounded away from zero. Further, since the selection is strictly increasing $x^*(x) = x$ for all $x \in [0,1]$ by Lemma 30 and the integral within the expectation satisfies

$$\frac{1}{h_n}\int K_{h_n}(X_i - x^*(x))\frac{f_X(x)}{f_X(x^*(x))}dx = \frac{1}{h_n}\int K_{h_n}(X_i - x)dx = 1$$

and therefore, for some $c > 0$ which depends on y,

$$\operatorname{Var}(\phi'_{F_{Y|X}}(Z_{F,n})(y)) \geq \frac{c}{n}.$$

For our purposes, a crude lower bound on the variance is enough. Therefore, we lower bound the variance of the derivative with respect to the marginal density by zero. The covariance of the two derivatives is

$$\operatorname{Cov}(\phi'_{F_{Y|X}}(Z_{F,n}), \phi'_f(Z_{f,n})) = \iint \underbrace{\Sigma_{F,f}(y, x, x')}_{=0} f_X(x)F_{Y|X}(y, x^*(x'))dxdx' = 0,$$

where the first equality follows by Fubini's theorem. This show's the first claim.

(ii): By the same arguments as in (i) we can focus on the variance of the derivative with respect to the conditional distribution function. The difference is that now there are intervals[28] $I_m \subset [0,1]$ such that $x^*(x) = \max I_m =: x_m$ for all $x \in I_m$ and $x^*(x) = x$ for all x which are not included in any I_m. Therefore,

$$\begin{aligned}\phi'_{F_{Y|X}}(Z_{F,n})(y) &= \int Z_{F,n}(y, x^*(x)) f_X(x) dx \\ &= \sum_{m=1}^{M} Z_{F,n}(y, x_m) \mathrm{P}(X_i \in I_m) + \int_{[0,1]\setminus \cup_m I_m} Z_{F,n}(y,x) f_X(x) dx\end{aligned}$$

where M is either finite or countably infinite. Consider the covariances between the different summands. For $m \neq k$,

$$\mathrm{Cov}(Z_{F,n}(y, x_m), Z_{F,n}(y, x_k)) = 0$$

for sufficiently small h, implying

$$\sum_{m=1}^{M} \sum_{k \neq m} \mathrm{Cov}(Z_{F,n}(y, x_m), Z_{F,n}(y, x_k)) \mathrm{P}(X_i \in I_m) \mathrm{P}(X_i \in I_k) = 0$$

for sufficiently small h, or equivalently, for sufficiently large n. Further, for any m, by standard arguments

$$nh_n \sum_{m=1}^{M} \mathrm{Cov}\left(Z_{F,n}(y, x_m), \int_{[0,1]\setminus \cup_m I_m} Z_{F,n}(y,x) f_X(x) dx \right) = O(h_n).$$

This implies for the variance,

$$\begin{aligned}\mathrm{Var}(\phi'_{F_{Y|X}}(Z_{F,n})(y)) &\geq \sum_{m=1}^{M} \mathrm{Var}(Z_{F,n}(y, x_m)) \mathrm{P}(X_i \in I_m)^2 \\ &\quad + \mathrm{Var}\left(\int_{[0,1]\setminus \cup_m I_m} Z_{F,n}(y,x) f_X(x) dx \right) - O_p\left(\frac{1}{n}\right) \\ &\geq \sum_{m=1}^{M} \mathrm{Var}(Z_{F,n}(y, x_m)) \mathrm{P}(X_i \in I_m)^2 - O_p(n^{-1}).\end{aligned}$$

[28] Since x^* is monotonely increasing, I_m has to be connected. Moreover, since $x^*(x)$ is càdlàg, there can be at most countably many such intervals.

Finally,

$$\begin{aligned}\mathrm{Var}(Z_{F,n}(y,x_m)) =& \frac{1}{nh_n^2 f_X(x_m)^2}\,\mathrm{E}[K_{h_n}(X_i-x_m)^2\varepsilon_i(y)^2]\\ =& \frac{1}{nh_n f_X(x_m)^2}\int K(v)^2\,\mathrm{E}[\varepsilon_i^2(y)\mid X_i = x+h_n v]f_X(x+h_n v)dv\\ \geq& \frac{\mathrm{E}[\varepsilon_i^2(y)\mid X_i=x]}{nh_n f_X(x_m)}\int K(v)^2dv - o_p\bigg(\frac{1}{nh_n}\bigg).\end{aligned}$$

For fixed y, the conditional variance of ε_i is bounded away from zero since the joint density of X and Y is bounded away from zero and therefore there is a constant $c>0$ such that

$$\mathrm{Var}(\phi'_{F_{Y|X}}(Z_{F,n})(y)) \geq \frac{c}{nh_n}.$$

(iii): The claim essentially follows by Lemma 3 in our companion paper Scherer (2024) and the representation result in Lemma 31. In order to apply Lemma 3 in our companion paper, we need a lower and upper bound on the weak variance as well as an upper bound on $\mathrm{E}[\phi'_{\theta_0}(Z_n)]$.

For the upper bound on the weak variance, it holds for any $\psi \leq \phi'_{F_{Y|X}}$

$$\mathrm{Var}(\psi(Z_{F,n})+\phi'_f(Z_{f,n})) = \mathrm{Var}(\psi(Z_{F,n})) + \mathrm{Var}(\phi'_f(Z_{f,n})),$$

where we have used that $Z_{F,n}$ and $Z_{f,n}$ are uncorrelated. The variance of the functional ψ satisfies

$$\begin{aligned}\mathrm{Var}(\psi(Z_{F,n})) =& \iint \Sigma_F(y,x,y,x')d\mu(x)d\mu(x')\\ =& \frac{1}{nh_n^2}\,\mathrm{E}\bigg[\bigg(\int K_{h_n}(X_i-x)d\mu(x)\bigg)^2\varepsilon_i(y)^2\bigg]\\ \leq& \frac{1}{4nh_n^2}\,\mathrm{E}\bigg[\bigg(\int K_{h_n}(X_i-x)d\mu(x)\bigg)^2\bigg].\end{aligned}$$

Using Fubini another time, implies

$$\begin{aligned}&\frac{1}{4nh_n^2}\,\mathrm{E}\bigg[\bigg(\int K_{h_n}(X_i-x)d\mu(x)\bigg)^2\bigg]\\ =&\frac{1}{4nh_n^2}\iiint K_{h_n}(X_i-x)K_{h_n}(X_i-x')f_X(X_i)dX_i d\mu(x)d\mu(x')\end{aligned}$$

$$=\frac{1}{4nh_n}\iiint K(v)K\left(v+\frac{x'-x}{h_n}\right)f_X(x+h_nv)dvd\mu(x)d\mu(x') \leq \frac{\|K\|_\infty^2\|f_X\|_\infty}{4nh_n},$$

where we have used that μ is a positive Borel measure μ with $\mu([0,1]) \leq 1$. Further, the variance of the derivative with respect to the marginal density satisfies by standard arguments

$$\begin{aligned}
&\mathrm{Var}(\phi_f'(Z_{f,n})) \\
=&\iint \max_{x_1\in\Psi(y,x)} F_{Y|X}(y,x_1) \max_{x_2\in\Psi(y,x')} F_{Y|X}(y,x_2)\Sigma_f(x,x')dxdx' \\
=&\frac{1}{n}\,\mathrm{E}\left[\left(\frac{1}{h_n}\int \max_{x'\in\Psi(y,x)} F_{Y|X}(y,x')\{K_{h_n}(X_i-x)-\mathrm{E}[K_{h_n}(X_i-x)]\}dx\right)^2\right] \leq \frac{C}{n}
\end{aligned}$$

and therefore $\mathrm{Var}(\psi(Z_{F,n})+\phi_f'(Z_{f,n})) \leq C/(nh_n)$, for some constant which does not depend on ψ, implying

$$\bar{\sigma}^2 := \sup_{\psi\leq\phi_{\theta_0}'} \mathrm{Var}(\psi(Z_n)) \leq \frac{C}{nh_n}.$$

For the lower bound on the weak variance, we use that by Lemma 31, $\psi(h) = \int h(y,x^*(x))f_X(x)dx$ is dominated by $\phi_{F_{Y|X}}'$ for some selection $x^*(x) \in \Psi(y,x)$, $x \in [0,1]$. As the derivative is nonlinear, there exists a continuum of x so that $\Psi(y,x)$ is not a singleton. In particular, there exists some interval $I \subset [0,1]$ such that $\max I \in \Psi(y,x)$ for all $x \in I$. In order to see this, let x be such that $\Psi(y,x)$ is not a singleton. Then, there exists an $\tilde{x} \in \Psi(y,x)$ with $\tilde{x} > x$. By definition of $\Psi(y,x)$ as an argmax set, it then holds that $\tilde{x} \in \Psi(y,x')$ for all $x' \in [x,\tilde{x}]$. Therefore, we can find a selection which is monotone and càdlàg a.e. which we denote by x^* in the following. Now, by the same arguments as in part (ii), there is a constant $c > 0$ such that

$$\mathrm{Var}\left(\int Z_{F,n}(y,x^*(x))f_X(x)dx + \phi_f'(Z_{f,n})\right) \geq \frac{c}{nh_n},$$

implying $\bar{\sigma}^2 \geq c/(nh_n)$.

We further need an upper bound on $\mathrm{E}[\phi_{\theta_0}'(Z_n)]$. By the proof of Lemma 2 in our

companion paper Scherer (2024), $\{\psi \leq \phi'_{\theta_0}\} \subset \{\psi \in \mathbb{D}^* : \|\psi\|_{\mathbb{D}^*} \leq 1\}$. Thus,

$$\begin{aligned}\mathrm{E}[\phi'_{\theta_0}(Z_n)] &= \mathrm{E}\bigg[\sup_{\psi \leq \phi'_{\theta_0}} \psi(Z_n)\bigg] = \mathrm{E}\bigg[\sup_{\psi \leq \phi'_{F_{Y|X}}} \psi(Z_{F,n})\bigg] + \underbrace{\mathrm{E}[\phi'_f(Z_{f,n})]}_{=0} \\ &\leq \mathrm{E}[\sup_{\|\psi\| \leq 1} \psi(Z_{F,n})] = \mathrm{E}\Big[\sup_{y,x} Z_{F,n}(y,x)\Big] = O\bigg(\sqrt{\frac{\log n}{nh_n}}\bigg),\end{aligned}$$

where the last bound follows by standard arguments.

Now, we are ready to apply Lemma 3 in our companion paper. Therefore, either

$$\mathrm{Var}(\sqrt{nh_n}\phi'_{F_{Y|X}}(Z_{F,n})) \geq \frac{nh_n\bar{\sigma}^2}{32} \geq c > 0$$

or

$$\mathrm{Var}(\sqrt{nh_n}\phi'_{F_{Y|X}}(Z_{F,n})) \geq \frac{1}{5}\frac{nh_n\bar{\sigma}^2}{(2nh_n\bar{\sigma}^2 + (\sqrt{nh_n}\,\mathrm{E}[\phi'_{F_{Y|X}}(Z_{F,n})] + 1)^2)} \geq \frac{c}{\log n} > 0.$$

By taking the smaller of the lower bounds, we have

$$\mathrm{Var}(\phi'_{F_{Y|X}}(Z_{F,n})) \geq \frac{c}{nh_n \log n}.$$

The claim now follows since

$$\begin{aligned}\mathrm{Var}(\phi'_{\theta_0}(Z_n)) &= \mathrm{Var}(\phi'_{F_{Y|X}}(Z_{F,n})) + \mathrm{Var}(\phi'_f(Z_{f,n})) + 2\,\mathrm{Cov}(\phi'_{F_{Y|X}}(Z_{F,n}), \phi'_f(Z_{f,n})) \\ &\geq \frac{c}{nh_n \log n},\end{aligned}$$

where we have used that the derivatives are uncorrelated.

Finally, we show that $\mathcal{F} = \emptyset$. By Lemma 31, it is sufficient to show that, for every positive and finite Borel measure μ,

$$\mathrm{Var}\bigg(\int_0^1 Z_{F,n}(y,x)d\mu(x) + \phi'_f(Z_{f,n})(y)\bigg) > 0.$$

Since these summands are uncorrelated by the same arguments as in part (i) of this proof,

$$\mathrm{Var}\bigg(\int_0^1 Z_{F,n}(y,x)d\mu(x) + \phi'_f(Z_{f,n})(y)\bigg) \geq \mathrm{Var}(\phi'_f(Z_{f,n})(y)).$$

Further, by Fubini's theorem and (1.29)

$$\begin{aligned}
&\operatorname{Var}(\phi'_f(Z_{f,n})(y)) \\
=&\iint \max_{x_1\in\Psi(y,x)} F_{Y|X}(y,x_1) \max_{x_2\in\Psi(y,x')} F_{Y|X}(y,x_2)\Sigma_f(x,x')dxdx' \\
=&\frac{1}{n}\operatorname{E}\left[\left(\frac{1}{h_n}\int \max_{x'\in\Psi(y,x)} F_{Y|X}(y,x')\{K_{h_n}(X_i-x)-\operatorname{E}[K_{h_n}(X_i-x)]\}dx\right)^2\right].
\end{aligned}$$

The latter is strictly positive since $F_{Y|X}$ is bounded away from zero for fixed y and K integrates to one. □

Consistency of the derivative estimator

Lemma 40. *Suppose Assumptions 13 hold. Then,*

$$\sup_{\|h\|_{BL}\leq 1,\|g\|_\infty\leq 1} |\hat{\phi}_n(h,g)-\phi'_{\theta_0}(h,g)| = O_p\left(\left(\frac{\log n}{nh_n}\right)^{1/4}\right).$$

Proof. Regarding consistency of the derivative estimator. Consider first the derivative with respect to F.

$$\begin{aligned}
\hat{\phi}'_{n,F}(h)(y)-\phi'_F(h)(y) &= \int \max_{x'\in\hat{\Psi}_n(y,x)} h(y,x')\hat{f}_n(x)dx - \int \max_{x'\in\Psi(y,x)} h(y,x')f(x)dx \\
&= \int\left\{\max_{x'\in\hat{\Psi}_n(y,x)} h(y,x') - \max_{x'\in\Psi(y,x)} h(y,x')\right\}\hat{f}_n(x)dx \\
&\quad + \int \max_{x'\in\Psi(y,x)} h(y,x')\{\hat{f}_n(x)-f(x)\}dx
\end{aligned}$$

We show below that $\Psi(y,x)\subset\hat{\Psi}_n(y,x)\subseteq\Psi(y,x)^{\delta_n}$ uniformly in (y,x) for some $\delta_n\to 0$. Then, we can bound the first term by

$$\begin{aligned}
&\left|\int\left\{\max_{x'\in\hat{\Psi}_n(y,x)} h(y,x') - \max_{x'\in\Psi(y,x)} h(y,x')\right\}\hat{f}_n(x)dx\right| \\
\leq&\sup_{x,y}\left|\max_{x'\in\hat{\Psi}_n(y,x)} h(y,x') - \max_{x'\in\Psi(y,x)} h(y,x')\right| \leq \|h\|_{BL}\delta_n.
\end{aligned}$$

For the second part, we have

$$\left|\int \max_{x'\in\Psi(y,x)} h(y,x')\{\hat{f}_n(x)-f_X(x)\}dx\right| \leq \|h\|_{BL}\|\hat{f}_n-f_X\|_\infty.$$

Hence, on the event $\Psi(y,x) \subset \hat{\Psi}_n(y,x) \subseteq \Psi(y,x)^{\delta_n}$, we have

$$\|\hat{\phi}'_{n,F}(h) - \phi'_F(h)\|_\infty \leq \|h\|_{BL}(\delta_n + \|\hat{f}_n - f\|_\infty).$$

For the derivative with respect to the marginal density of X, we have

$$\begin{aligned}|\hat{\phi}'_{n,f}(g)(y) - \phi'_f(g)(y)| =& \left| \int \left\{ \max_{x' \geq x} \hat{F}_n(y,x') - \max_{x' \geq x} F_{Y|X}(y,x') \right\} g(x) dx \right| \\ \leq& \|\hat{F}_n - F_{Y|X}\|_\infty \|g\|_\infty.\end{aligned}$$

Combining the above results implies

$$\begin{aligned}&\sup_{\|h\|_{BL} \leq 1, \|g\|_\infty \leq 1} |\hat{\phi}_n(h,g) - \phi'_{\theta_0}(h,g)| \\ \leq& C(\|h\|_{BL} \vee \|g\|_\infty)\{\delta_n + \|\hat{F}_n - F_{Y|X}\|_\infty + \|\hat{f}_n - f_X\|_\infty\}\end{aligned}$$

on B_n. Since $\mathrm{P}(B_n) \to 1$, the claim follows.

It remains to show that $\Psi(y,x) \subset \hat{\Psi}_n(y,x) \subseteq \Psi(y,x)^{\delta_n}$ with high probability. Recall that

$$\hat{\Psi}_n(y,x) = \{x' \geq x \,|\, \hat{F}_n(y,x') + \hat{z}_n \hat{\sigma}_n(y,x') \geq \max_{\tilde{x} \geq x} \hat{F}_n(y,\tilde{x}) - \hat{z}_n \hat{\sigma}_n(y,\tilde{x})\}$$

and bound

$$\begin{aligned}&\hat{F}_n(y,x') + 2\hat{z}_n \hat{\sigma}_n(y,x') - \max_{\tilde{x} \geq x} \hat{F}_n(y,\tilde{x}) \\ =& F_{Y|X}(y,x') - \max_{\tilde{x} \geq x} F_{Y|X}(y,\tilde{x}) \\ &+ \{\hat{F}_n(y,x') - F_{Y|X}(y,x')\} \\ &- \left\{ \max_{\tilde{x} \geq x} \hat{F}_n(y,\tilde{x}) - \max_{\tilde{x} \geq x} F_{Y|X}(y,\tilde{x}) \right\} \\ &+ 2\hat{z}_n \hat{\sigma}_n(y,x')\end{aligned}$$

On the event $B_n = \{\|(\hat{F}_n - F_{Y|X})/\hat{\sigma}_n\|_\infty \leq \hat{z}_n\}$, we have

$$\hat{F}_n(y,x') + 2\hat{z}_n \hat{\sigma}_n(y,x') - \max_{\tilde{x} \geq x} \hat{F}_n(y,\tilde{x})$$

$$\begin{cases} \geq F_{Y|X}(y,x') - \max_{\tilde{x}\geq x} F_{Y|X}(y,\tilde{x}) \\ \leq F_{Y|X}(y,x') - \max_{\tilde{x}\geq x} F_{Y|X}(y,\tilde{x}) + 4\hat{z}_n \|\hat{\sigma}_n\|_\infty. \end{cases}$$

This implies

$$\begin{aligned} &\{x' \geq x : F_{Y|X}(y,x') - \max_{\tilde{x}\geq x} F_{Y|X}(y,\tilde{x}) \geq 0\} \\ \supset &\{x' \geq x : F_{Y|X}(y,x') - \max_{\tilde{x}\geq x} F_{Y|X}(y,\tilde{x}) \geq 0\} \\ = &\{x' \geq x : F_{Y|X}(y,x') = \max_{\tilde{x}\geq x} F_{Y|X}(y,\tilde{x})\} = \Psi(y,x) \end{aligned}$$

and

$$\hat{\Psi}_n(y,x) \subset \{x' \geq x : F_{Y|X}(y,x') - \max_{\tilde{x}\geq x} F_{Y|X}(y,\tilde{x}) + 4\hat{z}_n\|\hat{\sigma}_n\|_\infty \geq 0\} \subset \Psi(y,x)^{\delta_n}$$

with $\delta_n = 1/c_1(4\hat{z}_n\|\hat{\sigma}_n\|_\infty)^{\frac{\rho}{\rho-1}\vee\frac{\alpha+1}{\alpha}}$, which follows by the same arguments as in the proof of Lemma 28. □

Consistency of the bootstrap

Lemma 41. *Suppose that Assumption 13 holds. Then,*

$$\sup_{t\in\mathbb{R}} \left| \mathrm{P}\left(\hat{\phi}'_n(\hat{Z}_n) \leq t \,\middle|\, \mathcal{D}_n\right) - \mathrm{P}\left(\phi(\hat{\theta}_n) - \phi(\theta_0) \leq t\right) \right| = o_p(1).$$

Proof. We apply Theorem 4 separately to the cases of a linear and nonlinear derivative. First consider the nonlinear case. The rate of the bootstrap can be shown to be

$$d_n = \sqrt{\frac{\log n}{nh_n^3}}$$

by the same arguments as in Lemmas 38 and 23. Moreover, the coupling for the bootstrap follows using the same arguments as in the setting of the maximum of a conditional mean function in Lemma 24 building on the proof of Theorem 9 in the supplementary material of Chernozhukov et al. (2013b). It results in an approxima-

tion

$$\mathrm{P}\left(\sup_{x\in[0,1]}|\hat{Z}_n(x)-Z_n^*(x)|>o\left(\frac{1}{\sqrt{nh_n\log^2 n}}\right)\middle|\mathcal{D}_n\right)=o_p\left(\frac{1}{\log n}\right),$$

and therefore we have $s_n = o((nh_n \log^2 n)^{-1/2})$. Additionally, Lemma 40 implies

$$\eta_n = \left(\frac{\log n}{nh_n}\right)^{1/4}.$$

It only remains to check the further rate requirements. We have

$$\begin{aligned}\frac{d_n\eta_n}{\sigma_n} &= \sqrt{nh_n\log n}\sqrt{\frac{\log n}{nh_n^3}}\left(\frac{\log n}{nh_n}\right)^{1/4} = \left(\frac{\log^5 n}{nh_n^5}\right)^{1/4} \to 0\\ \frac{s_n}{\sigma_n} &= \sqrt{nh_n\log n}\frac{1}{\sqrt{nh_n\log^2 n}} = \frac{1}{\sqrt{\log n}} \to 0\end{aligned}$$

and the claim follows by Theorem 4.

For the case of a linear mean square discontinuous derivative, the result follows along the same lines as above (up to log terms). For the final case of a linear and mean square continuous derivative, it still holds

$$d_n = \sqrt{\frac{\log n}{nh_n^3}} \qquad \text{and} \qquad \eta_n = \left(\frac{\log n}{nh_n}\right)^{1/4}.$$

The coupling for the bootstrap follows again using the arguments as in the proof of Theorem 9 in the supplementary material of Chernozhukov et al. (2013b) applied to the coupling construction in Lemma 36 and yield the rate $s_n = o((n \log n)^{-1/2})$. It only remains to check the further rate requirements. We have

$$\begin{aligned}\frac{d_n\eta_n}{\sigma_n} &= \sqrt{n}\sqrt{\frac{\log n}{nh_n^3}}\left(\frac{\log n}{nh_n}\right)^{1/4} = \left(\frac{\log^3 n}{nh_n^7}\right)^{1/4} \to 0\\ \frac{s_n}{\sigma_n} &= \sqrt{n}\frac{1}{\sqrt{n\log n}} = \frac{1}{\sqrt{\log n}} \to 0\end{aligned}$$

and the claim follows by Theorem 4. □

Projection Confidence bands

Let $[F_\ell(y,x), F_u(y,x)]$ be a confidence band for $F_{Y|X}$ and $[f_\ell(x), f_u(x)]$ a band for f_X. Then the projection bands are $[\phi(F_\ell, f_\ell)(y), \phi(F_u, f_u)(y)]$ which follows by monotony of ϕ with respect to the pointwise ordering. In order to see this monotony, let $F \leq G$ and $f \leq g$, then

$$\begin{aligned}\phi(F,f) &= \int \max_{x' \geq x} F(y,x') f(x) dx \\ &\leq \int \max_{x' \geq x} G(y,x') g(x) dx = \phi(G,g)\end{aligned}$$

since $F(y,x') \geq 0$.

1.C Additional results

1.C.1 Topological thoughts on the Delta method

Suppose $\hat{\theta}_n, \theta \in \mathbb{D}$ and $\phi(\mathbb{D}) \subset \mathbb{E}$. Now consider two topologies $\mathcal{O}$ and $\mathcal{V}$ on $\mathbb{D}$ satisfying $\mathcal{O} \subset \mathcal{V}$, i.e., $\mathcal{V}$ is finer than $\mathcal{O}$ (or $\mathcal{O}$ is coarser than $\mathcal{V}$). Further, denote by $\mathcal{K}(\mathcal{O})$ the collection of compact sets in the topology $\mathcal{O}$.

In the following Lemma, we collect some topological results:

Lemma 42. *(i) If $\|x\|_\mathcal{O} \leq \|x\|_\mathcal{V}$ for all $x \in \mathbb{D}$, then $\mathcal{O} \subset \mathcal{V}$.*

(ii) If $\mathcal{O} \subseteq \mathcal{V}$, then $\mathcal{K}(\mathcal{V}) \subseteq \mathcal{K}(\mathcal{O})$.

(iii) If $\|x\|_\mathcal{O} \leq \|x\|_\mathcal{V}$ for all $x \in \mathbb{D}$, then $\mathcal{V}$-bounded set is also $\mathcal{O}$-bounded.

(iv) If $\mathcal{O} \subseteq \mathcal{V}$ and $f : \mathbb{D} \to Y$ is $\mathcal{O}$-continuous, then f is also $\mathcal{V}$-continuous.

(v) If $\mathcal{O} \subseteq \mathcal{V}$ and $X_n \xrightarrow{d} X$ in $\mathcal{V}$, then $X_n \xrightarrow{d} X$ in $\mathcal{O}$.

(vi) If $\mathcal{O} \subseteq \mathcal{V}$ and ϕ is $\mathcal{O}$-Hadamard-differentiable, then ϕ is also $\mathcal{V}$-Hadamard-differentiable.

(vii) If $\mathcal{O} \subseteq \mathcal{V}$ and ϕ is $\mathcal{O}$-Fréchet-differentiable, then ϕ is also $\mathcal{V}$-Fréchet-differentiable.

In words: (i) allows the other results to be applied to a setting where we have norms of differing strengths and therefore gives a common setting for the other

results. (ii) states that the finer the topology, the smaller the number of compact sets. This also implies that a random variable that is tight in the coarse topology does not have to be tight in the finer topology. (iii) states that the finer the topology, the smaller the number of bounded sets. (iv) implies that the finer the topology, the larger the number of continuous functions. This almost directly implies (v), that a sequence which converges in distribution in a coarser topology does not necessarily need to converge in a finer topology. (v)-(vii) combined reveal a trade-off when choosing the topology for the Delta method: The finer the topology, the harder it gets for the random variable to converge in distribution but the easier it is for the transformation to be differentiable.

Proof. **(i):** Suppose that $\mathcal{O}$, $\mathcal{V}$ are generated by the norms $\|\cdot\|_{\mathcal{O}}$ and $\|\cdot\|_{\mathcal{V}}$ respectively. Further, suppose that

$$\|x\|_{\mathcal{O}} \leq \|x\|_{\mathcal{V}}$$

for any $x \in \mathbb{D}$ (implying that the identity mapping $id : (\mathbb{D}, \mathcal{O}) \to (\mathbb{D}, \mathcal{V})$ is continuous). This implies

$$\{y \in \mathbb{D} : \|x - y\|_{\mathcal{V}} < \varepsilon\} \subseteq \{y \in \mathbb{D} : \|x - y\|_{\mathcal{O}} < \varepsilon\}$$

for any $x \in \mathbb{D}$ and $\varepsilon > 0$. Now, let $O \in \mathcal{O}$. Then, for any $x \in O$, there exists $\varepsilon > 0$ such that

$$\{y \in \mathbb{D} : \|x - y\|_{\mathcal{V}} < \varepsilon\} \subseteq O$$

and hence $O \in \mathcal{V}$, proving $\mathcal{O} \subseteq \mathcal{V}$.

(ii): Every $\mathcal{V}$-compact set K is also $\mathcal{O}$-compact. Let $\mathcal{K}(\mathcal{O})$ denote the compact sets wrt the topology $\mathcal{O}$, then $\mathcal{K}(\mathcal{V}) \subseteq \mathcal{K}(\mathcal{O})$. Indeed, any $\mathcal{O}$-open cover of $\mathbb{D}$ is also a $\mathcal{V}$-open cover of $\mathbb{D}$ and the remaining argument is simple. Thus, the finer a topology the (weakly) smaller the number of compact sets.

(iii): Follows directly by definition.

(iv): A function $f : X \to Y$ is continuous, if for any $O \in \mathcal{O}_Y$ it holds $f^{-1}(O) = \{x \in X : f(x) \in O\}$. Slightly abusing notation: $f^{-1}(\mathcal{O}_Y) \subseteq \mathcal{O}_X$. Now, if $\mathcal{O} \subseteq \mathcal{V}$ and $f : \mathbb{D} \to Y$ (for an arbitrary topology on Y) is $\mathcal{O}$-continuous, then f is also $\mathcal{V}$-continuous.

(v): Recall convergence in distribution for a $\mathbb{D}$-valued random variables X_n to X if

$$\mathrm{E}^* f(X_n) \to \mathrm{E}\, f(X)$$

for all continuous and bounded $f : \mathbb{D} \to \mathbb{R}$. Hence, by the observation on the number of continuous functions in (iv), (v) follows.

(vi): This follows directly by the relation between Hadamard and compact differentiability as well as (ii).

(vii): This follows directly by the relation between Fréchet and bounded differentiability as well as (iii). □

1.C.2 Further Plots for the Simulation

The following figures depict the first four realizations of our Monte Carlo simulation across the four DGPs. The black curve depicts the plug-in estimator, the blue dashed line the lower confidence interval and the red line shows the true curve.

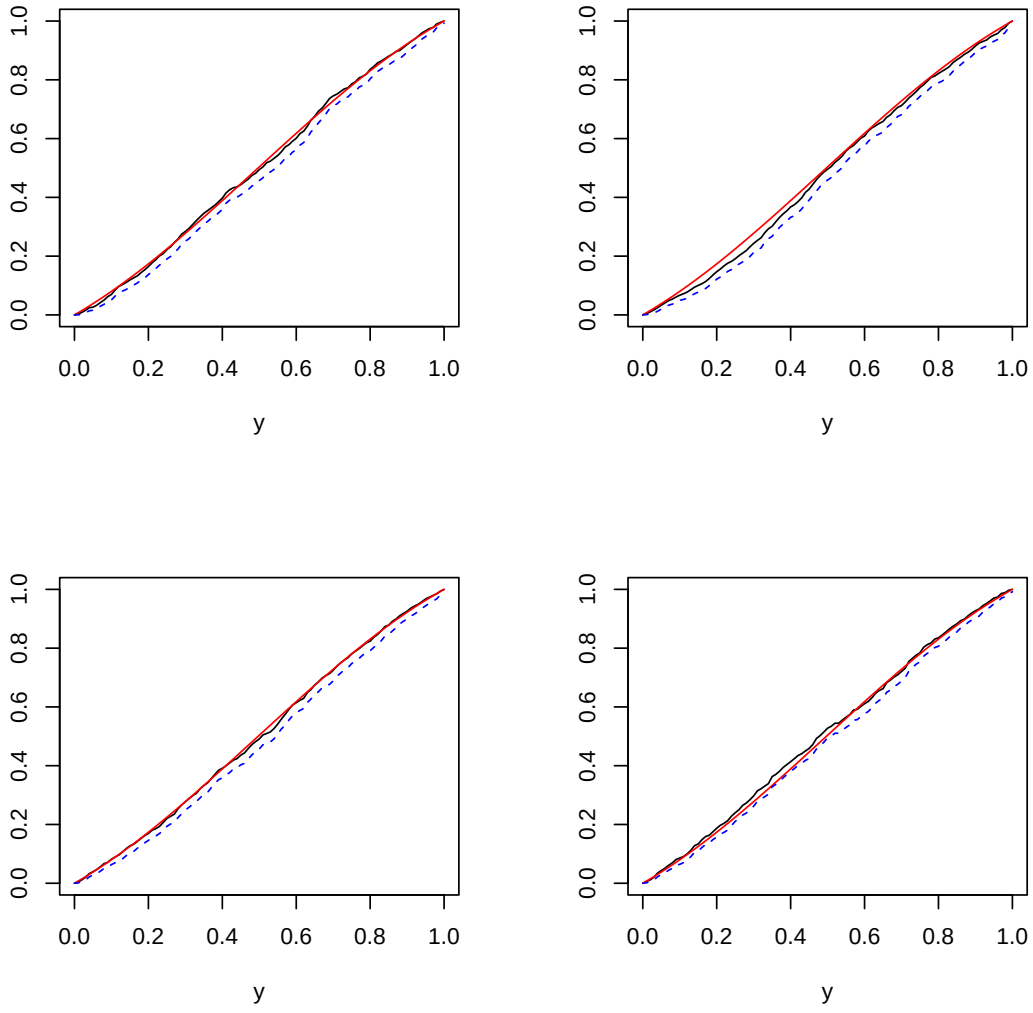

Figure 1.C.1: DGP decreasing

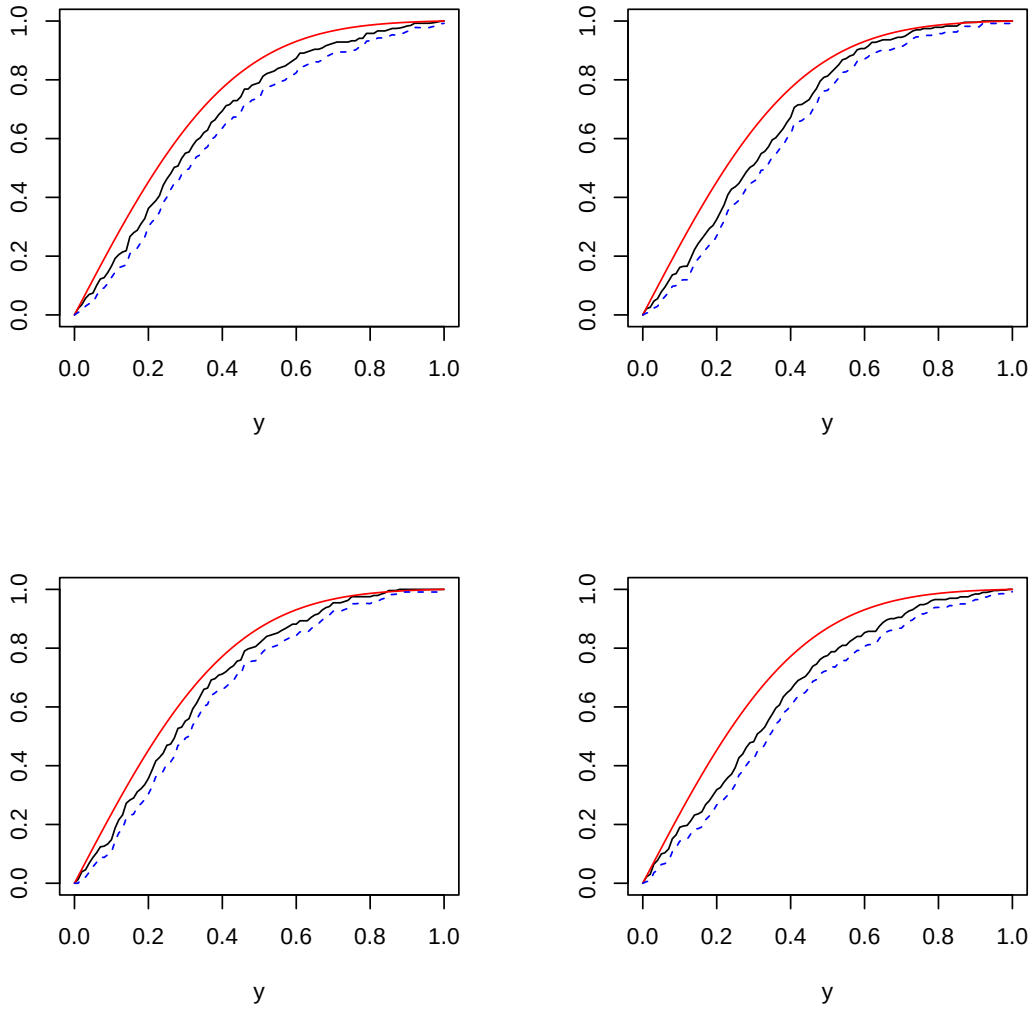

Figure 1.C.2: DGP increasing

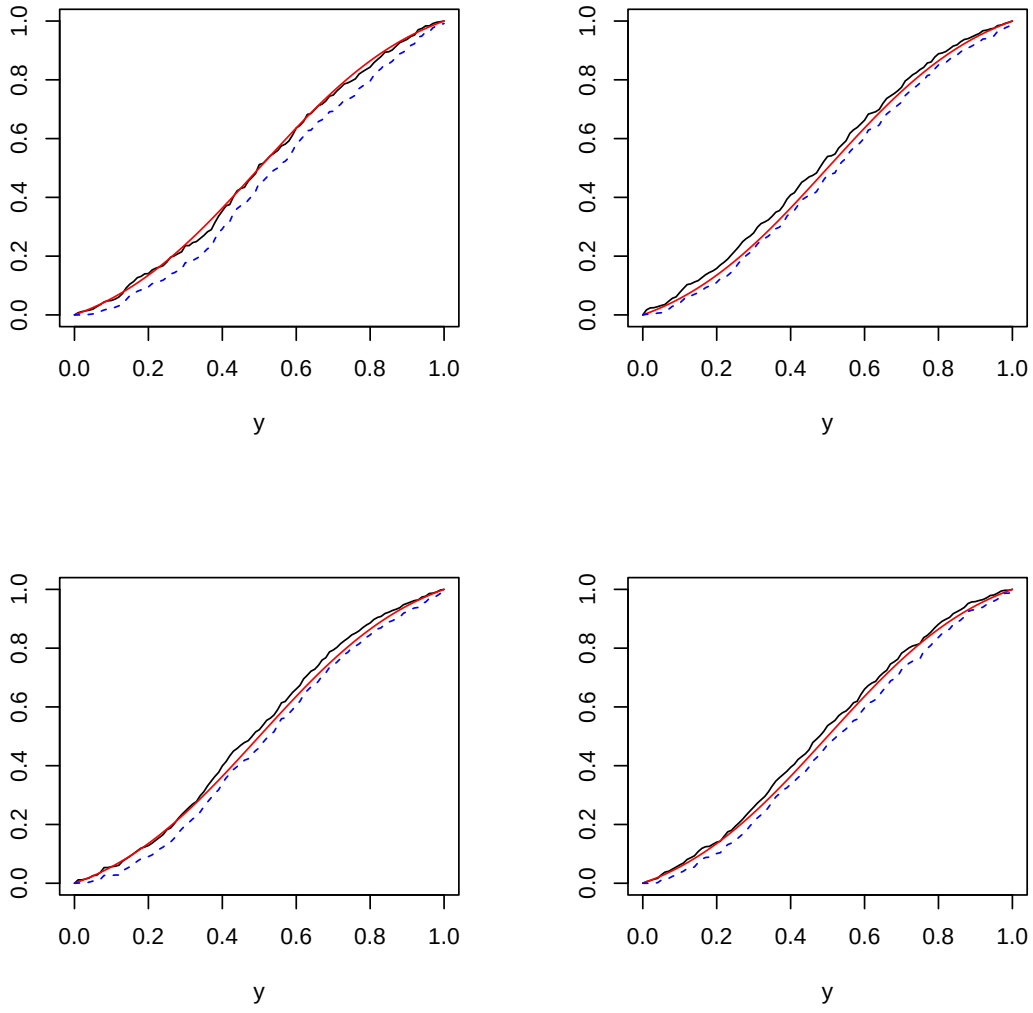

Figure 1.C.3: DGP independent

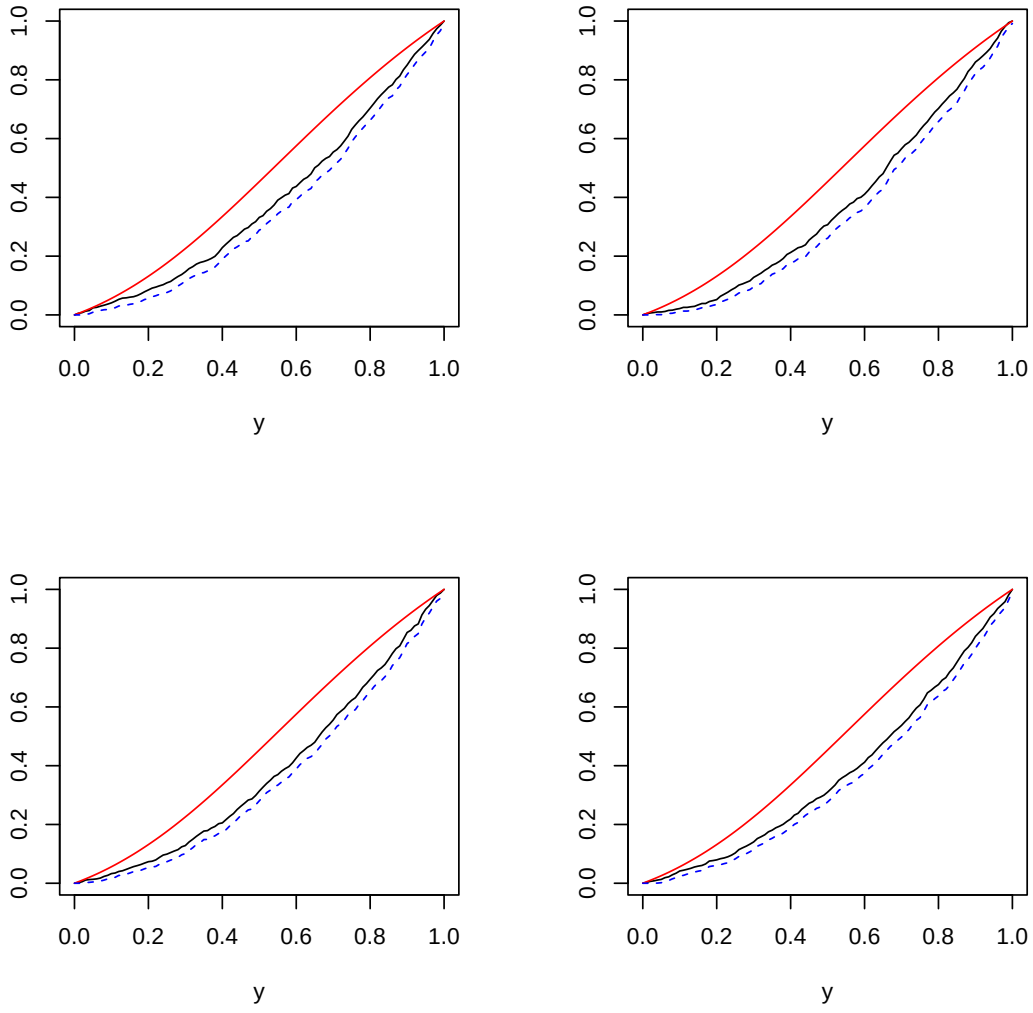

Figure 1.C.4: DGP 4

Chapter 2

An Anti-Concentration Bound for Sublinear and Continuous Functionals of Gaussian Random Vectors

2.1 Introduction

Anti-concentration bounds play an important role in the modern theory on confidence intervals and testing in settings such as high-dimensional and nonparametric statistics. In these settings, it is often hard to derive limiting distributions of estimators or even no limiting distribution might exist. Here, anti-concentration bounds help as they provide non-asymptotic bounds leading to distributional approximations in Kolmogorov distance. More broadly, anti-concentration bounds have been used in random matrix theory to study the behavior of smallest singular values (e.g. Rudelson and Vershynin (2008) and Rudelson and Vershynin (2009)), in machine learning to characterize the computational complexity of learning sets under the Gaussian distribution (Klivans et al. (2008)) and in nonparametric Bayesian statistics to show validity of credible sets from a frequentist perspective (Götze et al. (2019)).

Anti-concentration bounds refer to inequalities bounding the probability that a real-valued random variable, say X, lies in a small interval. Formally, they give upper bounds on $\mathrm{P}(|X - x| \leq \varepsilon)$ for arbitrary $x \in \mathbb{R}$ and small $\varepsilon > 0$. More

commonly known are concentration bounds, such as Markov's inequality, which give a lower bound on this quantity and where x typically is the mean or median of X.

In this paper, we derive an anti-concentration bound for sublinear and continuous functionals of tight Gaussian random vectors. The Gaussian random vectors are assumed to take values in a real-valued Banach space, therefore allowing for finite dimensional as well as infinite dimensional applications. Sublinear functionals arise naturally in many different settings. Two important examples are (semi)norms or linear functionals. Besides that, they also come up in multiple one-sided testing problems such as moment inequality models (see, e.g., Canay and Shaikh (2016) and Molinari (2020) for an overview of this literature) or estimation under shape restrictions (see Chetverikov et al. (2018) for a survey). Moreover, sublinear functionals appear frequently in inferential problems on directionally differentiable functionals (see, Fang and Santos (2018) for examples).

The proof of our anti-concentration bound relies on two observations. Firstly, any continuous sublinear functional can be represented as a supremum over a subset of the topological dual of the Banach space under consideration and secondly any Gaussian random vector in a Banach space $\mathbb{D}$ leads to a Gaussian process over the topological dual of $\mathbb{D}$. These observations allow us to recast our setup to the study of the supremum of Gaussian processes. For the latter, we adapt the proof of an anti-concentration inequality due to Giessing (2023a). Giessing derives a dimension-free anti-concentration bound for suprema of separable Gaussian processes with non-degenerate marginal distributions. We adapt Giessing's bound to allow for non-separable Gaussian processes with potentially degenerate marginal distributions so that we do not need to restrict the Banach space nor the continuous sublinear functional. The extension to non-separable Gaussian processes allows our bound to be applied to spaces such as the space of continuous or absolutely integrable functions, and the extension to degenerate marginal distributions allows for applications where the Gaussian random vector has a singular covariance matrix or where the functional is non-negative.

The proposed anti-concentration bound contributes to the growing literature on anti-concentration bounds for suprema of separable Gaussian processes. Closest to our bound is the proposal by Giessing (2023a). As in Giessing (2023a), our bounds are sharp up to a multiplicative constant and can recover the anti-concentration bounds in Chernozhukov et al. (2014b), Chernozhukov et al. (2015a), Chernozhukov et al. (2017) and Deng and Zhang (2020) when restricted to our setting. Notably,

Deng and Zhang (2020) also propose an anti-concentration bound which does not depend on the smallest marginal variance, but their results are not dimension-free and cannot be applied as easily to infinite-dimensional settings as ours. Our bounds are also related to the bounds in Ball (1993), Nazarov (2003) and Klivans et al. (2008) who derive reverse isoperimetric inequalities for standard Gaussian measures of balls, half-spaces, convex polytopes, cones as well as general convex bodies. Importantly, their bounds are dimension-dependent and therefore do not readily apply to infinite-dimensional applications. Since norms are particular examples of sublinear continuous functionals, our bound is also related to the results in Götze et al. (2019) who study Gaussian comparison and anti-concentration bounds for squared norms on Hilbert spaces. Their bounds are qualitatively different to ours as they derive an anti-concentration bound for the squared norm which is not a sublinear functional. Finally, we want to mention the anti-concentration bounds in Belloni et al. (2019a) and Peccati and Turchi (2023) on min-max statistics of finite-dimensional vectors and Kozbur (2021) and Giessing (2023a) on order statistics. These are partially more general in that they consider more general transformations than a supremum but other than that are not related to the setup in this paper.

We apply our anti-concentration bound to derive non-asymptotic bounds on the Kolmogorov distance for sublinear and continuous functionals of sums of independent high-dimensional random vectors. By this, we contribute to the literature on high-dimensional CLTs by allowing for more general transformations of high-dimensional sums. Our bound is related to the literature on Berry-Esseen type bounds for convex and symmetric sets since any convex set that includes a closed ball can be represented by a sublinear and Lipschitz continuous Minkowski functional. Thus, our results are related to Bentkus (2003), Bentkus (2005), Raič (2019) and Fang and Koike (2024) who derived Berry-Esseen type bounds for general convex and symmetric sets as well as Euclidean balls and to Chernozhukov et al. (2013a), Chernozhukov et al. (2017), Chernozhuokov et al. (2022), Fang and Koike (2021), Lopes (2022) and Bong et al. (2023) who derived bounds for hyperrectangles. Similarly, Kuchibhotla and Rinaldo (2020) bound the distance between cdfs of sums of high-dimensional vectors, which can be represented as a particular collection of hyperrectangles. Further, since any norm is sublinear and Lipschitz continuous, our bounds also relate to the literature on Kolmogorov bounds for norms of high-dimensional sums as in Lopes et al. (2020) who studied the maximum norm and Giessing and Fan (2020), Giessing (2023b) and Giessing and Fan (2023) who stud-

ied general ℓ_p-norms. The general applicability of our bounds comes at the price of requiring often stronger restrictions on the dimension of the random vectors as in the above-mentioned papers. This suggests that our bounds are best used for functionals which are not covered by the above results.

Moreover, we apply our anti-concentration bound to derive distributional approximations for kernel-type estimators. We derive two separate results that differ in the assumed complexity of the functional. In the case of a rather small complexity, we rely on a coupling for empirical processes developed in Chernozhukov et al. (2014a) and in the case of large complexity, we use the Rio-Massart coupling derived in Chernozhukov et al. (2013b). As for the Kolmogorov bounds on high-dimensional vectors, we contribute to the literature by allowing for more general transformations of the estimators. Related to our bounds are Chernozhukov et al. (2013b) and Chernozhukov et al. (2014a) who study the supremum of kernel-based empirical processes, Cattaneo et al. (2024) who derive bounds for the sup-norm of boundary adaptive local polynomial density estimators and Cheng and Chen (2019) who derive bounds for the sup-norm of debiased kernel density and local linear estimators. I am not aware of any other work considering more general transformations than the sup-norm.

The rest of the paper is organized as follows. In Section 2.2, we discuss motivating examples and discuss their relation to the results in this paper. In Section 2.3, we derive our anti-concentration bound and lower bound on the variance of suprema of separable Gaussian processes. Finally, in Section 2.4, we apply our anti-concentration bound to derive Berry-Esseen type bounds for sums of independent high-dimensional vectors and kernel-type estimators. We collect all of our proofs in the Appendix.

Notation. For any measure Q on a measurable space $(S, \mathcal{S})$ and any measurable function $f : S \to \mathbb{R}$, we use the notation $\|f\|_{Q,p} = (\int |f|^p dQ)^{1/p}$, $p \in [1, \infty]$. When Q denotes the counting measure on $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$, where $\mathcal{P}(\mathbb{N})$ denotes the power set of $\mathbb{N}$, we write just $\|f\|_p = (\sum_{n\in\mathbb{N}} |f_n|^p)^{1/p}$. We also denote by $\|\cdot\|_p$ the ℓ_p norms on $\mathbb{R}^d$ for any $d \in \mathbb{N}$. For an arbitrary set T, let $\mathcal{C}(T)$ denote the space of continuous functions $T \to \mathbb{R}$ endowed with the uniform norm $\|f\|_T := \sup_{t\in T} |f(t)|$. Unless otherwise stated, $c, C > 0$ denote universal constants of which the values may change from place to place. For $a, b \in \mathbb{R}$, we use the notation $a \vee b = \max\{a, b\}$ and $a_+ = a \vee 0$.

2.2 Motivation

In this section, we briefly introduce two motivating examples illustrating potential applications and discuss their relation to our anti-concentration bounds.

Example 8 (Testing many inequalities). Let $X_1, \dots, X_n \in \mathbb{R}^k$ be i.i.d. random vectors with unknown mean $\mu_0 \in \mathbb{R}^k$ for some k which is allowed to increase with n. Consider testing the set of linear inequalities

$$H_0 : \mu_0 \leq 0 \qquad \text{vs.} \qquad H_1 : \mu_{0,j} > 0, \text{ for some } j = 1, \dots, k.$$

This testing problem has been analyzed for example by Chernozhukov et al. (2019) and Bai et al. (2019). Testing problems of this kind arise in different examples in economics such as in market entry games as in Ciliberto and Tamer (2009), in discrete choice models with endogeneity in Chesher et al. (2013) and in dynamic models with imperfect competition as in Bajari et al. (2007).

In order to develop a test statistic, let $\hat{\mu}_j$ and $\hat{\sigma}_j^2$ denote the sample mean and variance of $X_{1j}, \dots, X_{nj}$ respectively, for $j = 1, \dots, k$, that is

$$\hat{\mu}_{n,j} = \frac{1}{n} \sum_{i=1}^{n} X_{ij} \quad \text{and} \quad \hat{\sigma}_j^2 = \frac{1}{n} \sum_{i=1}^{n} (X_{ij} - \hat{\mu}_j)^2.$$

There are many different test statistics in order to test H_0 against H_1. It is natural to consider test statistics that take large values when some $\hat{\mu}_j$ is large. For example consider

$$T_{n,\infty} = \max_{1 \leq j \leq k} \sqrt{n} \frac{\hat{\mu}_j}{\hat{\sigma}_j} \quad \text{and} \quad T_{n,2} = \sqrt{\sum_{j=1}^{k} \left(\sqrt{n} \frac{\hat{\mu}_j}{\hat{\sigma}_j} \vee 0 \right)^2}.$$

Large values of $T_{n,\infty}$ or $T_{n,2}$ indicate that H_0 is violated, and therefore it is natural to consider a test of the form

$$T_{n,\ell} > c_\ell \quad \Longrightarrow \quad \text{reject } H_0, \qquad \ell = 2, \infty$$

where c_ℓ denotes a critical value that is suitably chosen so that the test is approximately of size $\alpha \in (0, 1)$. These two tests lead to different power properties. Heuristically, $T_{n,2}$ is preferred to $T_{n,\infty}$ when many inequalities are violated simultaneously, while $T_{n,\infty}$ is preferred when only relatively few inequalities are violated.

In order to construct critical values c_ℓ, note that under H_0,

$$T_{n,2} \leq \sqrt{\sum_{j=1}^{k} \left(\sqrt{n} \frac{\hat{\mu}_j - \mu_j}{\hat{\sigma}_j} \vee 0 \right)^2} =: S_{n,2}, \tag{2.1}$$

where equality holds when all μ_j are zero. A similar inequality holds for $T_{n,\infty}$. Hence, in order to construct a critical value for $T_{n,\ell}$, it is sufficient to construct critical values as the $(1-\alpha)$-quantile of the distribution of $S_{n,\ell}$. While this is straightforward in the finite dimensional case when k is fixed, it is challenging to derive a distributional approximation in the high-dimensional setting when k is allowed to increase with n. Even though the sup-based statistic has been studied in the high-dimensional setting, to the best of my knowledge, there is no distributional approximation result for the ℓ_2-based statistic in the high-dimensional setting in the literature. As we demonstrate in Section 2.4, our results allow for such an application.

Alternatively, one may construct critical values based on an inequality selection mechanism as e.g. in Chernozhukov et al. (2019) instead of the bound (2.1). Heuristically, such inequality selection procedures first discard uninformative inequalities which are far from binding and base the critical values on the reduced set of inequalities which improves the power of the test. As for the simpler test outlined above, one may derive a distributional approximation result for this refined testing procedure using our results.

Our next example is about one-sided testing in nonparametric curve estimation. Such one-sided tests arise in different settings such as inference on shape restrictions or in partial identification analysis. See e.g. Firpo et al. (2019) for specific examples.

Example 9 (One-sided testing in curve estimation). Suppose our object of interest is $\theta_0 \in \mathcal{C}(0,1)$, and we are interested in testing the hypothesis: $H_0 : \forall x \in [0,1] : \theta_0(x) \leq 0$

$$H_0 : \forall x \in [0,1] : \theta_0(x) \leq 0 \quad \text{vs.} \quad H_1 : \exists x \in [0,1] : \theta_0(x) > 0.$$

Further, suppose there is some nonparametric estimator $\hat{\theta}_n$ of θ_0 which can be approximated by a suitable Gaussian distribution as $n \to \infty$. A natural choice for a test statistic rejects the null hypothesis when $\hat{\theta}_n(x)$ is positive and large. For

example one may use a sup- or L_2-based test statistic

$$\hat{T}_{n,\infty} = \sup_{x\in[0,1]} \frac{\hat{\theta}_n(x)}{\hat{\sigma}_n(x)} \qquad \text{or} \qquad \hat{T}_{n,2} = \sqrt{\int_0^1 \left(\frac{\hat{\theta}_n(x)}{\hat{\sigma}_n(x)}\right)_+^2 dx},$$

where $\sigma_n(x)$ denotes an estimator of the standard error of $\hat{\theta}_n(x)$. Test statistics of this form have been investigated by several authors, see for example Chernozhukov et al. (2013b) and references therein for the sup-based test statistic and Lee et al. (2013) for the L_2-based test statistic. Alternatively, one may use a type of one-sided Lorentz norm

$$\hat{T}_{n,w,q} = \left(\int_0^\infty w(t)\left(\frac{\hat{\theta}_n(x)}{\hat{\sigma}_n(x)}\right)_+^*(t)^q dt\right)^{\frac{1}{q}},$$

where w is a decreasing non-negative locally integrable weight function on $\mathbb{R}_+$, $q \in [1,\infty)$ and the $(\cdot)_+^*$ operator denotes the one-sided decreasing rearrangement operator, i.e., for each f on $[0,1]$, f_+^* is given by

$$f_+^*(t) = \inf\{y : \text{Leb}(x \in [0,1] : f(x) > y) \leq t\}.$$

For $q = 2$ and $w = 1$, this test statistic would coincide with $\hat{T}_{n,2}$. But for different choices of the weight function w, this test statistic allows putting more weight on large deviations from the null and therefore may be seen as a test statistic which is between $\hat{T}_{n,2}$ and $\hat{T}_{n,\infty}$.

As in Example 8, $\hat{T}_{n,2}$ may be preferred to $\hat{T}_{n,\infty}$ when θ_0 is positive on a relatively large interval, while $\hat{T}_{n,\infty}$ may be preferred when θ_0 only deviates from the null on a small interval. Depending on the weight function w and q, the test statistics $\hat{T}_{n,w,q}$ may interpolate between power properties of $\hat{T}_{n,2}$ and $\hat{T}_{n,\infty}$ and therefore constitute an interesting alternative to these test statistics. Further, all of the above test statistics can be written as a functional ψ of the differences $\hat{\theta}_n/\hat{\sigma}_n$. For instance, $\hat{T}_{n,w,q} = \psi(\hat{\theta}_n - \mu)$, where ψ is given by

$$\psi(f) = \left(\int_0^\infty w(t)(f(x))_+^q dx\right)^{1/q}.$$

ψ is sublinear and continuous with respect to both the sup- and the L_q-norm.[1]

[1]This follows by the same arguments as in Chapter 2 §2 in DeVore and Lorentz (1993) applied

However, ψ is not absolutely homogeneous and therefore is no seminorm. The same holds true for the sup- and L_2-based test statistic.

Common to both examples is that we are interested in inference of some sublinear and continuous functional ψ applied to a random vector X_n. Also, common to both examples is that inference is non-standard in the sense that the estimators under consideration rarely converge in distribution to a useful limiting object. Heuristically, this can be seen as follows. If the non-studentized sample means $\sqrt{n}(\hat{\mu} - \mu)$ converge in distribution, then they have to be asymptotically tight. In particular, the norm $\sqrt{n}\|\hat{\mu} - \mu\|$ has to be asymptotically stochastically bounded. On the other hand, this norm usually diverges as the dimension of the vector increases (unless the vectors are highly correlated across their components). A similar argument applies to many nonparametric estimators such as kernel or series estimators. Thus, one cannot justify inference in the above examples by appealing to a limiting distribution approximation and the continuous mapping theorem.

While the above reasoning does not preclude the possibility that the test statistics converge in distribution, showing that $\psi(X_n)$ actually converges in distribution is a hard task and may need to rely on arguments specific to both the functional ψ and the random vectors X_n. On the other hand, there is a generic argument for approximation of these test statistics based on the following result due to Le Cam (Le Cam (2012), p. 402) which we adapted to our setting here.

Lemma 43. *For X, Z arbitrary random vectors and $\varepsilon > 0$,*

$$\sup_{t \in \mathbb{R}} |\mathrm{P}(\psi(X) \leq t) - \mathrm{P}(\psi(Z) \leq t)| \leq \mathrm{P}(|\psi(X) - \psi(Z)| > \varepsilon) + \zeta_\varepsilon(\psi(Z)),$$

where $\zeta_\varepsilon(V) = \sup_{x \in \mathbb{R}} \mathrm{P}(|V - x| \leq \varepsilon)$ for a real-valued random variable V.

In particular, this Lemma can be used to justify a Gaussian approximation of the distribution of $\psi(X_n)$ in the sense that we can approximate the distribution of $\psi(X_n)$ by the distribution of $\psi(Z_n)$ for some Gaussian random vector Z_n. For this purpose, this Lemma requires two ingredients. First, we need a bound on $\mathrm{P}(|\psi(X_n) - \psi(Z_n)| > \varepsilon)$. Such bounds are readily available for sums of high-dimensional random vectors (see e.g. Zaitsev (2013) for a survey) and for empirical processes (e.g. Massart (1989), Rio (1994), Koltchinskii (1994), Chernozhukov et al.

to $f\mathbb{1}\{f \geq 0\}$.

(2014a), Chernozhukov et al. (2016), Cattaneo et al. (2022) and Giessing (2023b)). Secondly, we need a bound on the Lévy concentration function $\zeta_\varepsilon(\psi(Z_n))$, i.e. an anti-concentration bound for sublinear and continuous functionals ψ of Gaussian random vectors Z_n. While there are related anti-concentration bounds for suprema of separable Gaussian processes (see e.g. Chernozhukov et al. (2014b) and Giessing (2023a)), these bounds do not allow for applications where Z_n takes its values in the space of continuous functions or the space of absolutely integrable functions. To the best of my knowledge, there is no anti-concentration bound in this setting which applies to general Banach spaces which motivates the results of the following sections.

2.3 The Anti-Concentration Bound

Let $(\mathbb{D}, \|\cdot\|)$ be a real Banach space with continuous dual space $(\mathbb{D}, \|\cdot\|^*)$ and $\psi : \mathbb{D} \to \mathbb{R}$ be continuous at zero and sublinear, i.e.,

$$\begin{aligned} \psi(\lambda x) &= \lambda \psi(x), && \lambda \geq 0, x \in \mathbb{D} \\ \psi(x+y) &\leq \psi(x) + \psi(y), && x, y \in \mathbb{D}. \end{aligned}$$

Moreover, suppose that Z is a tight Gaussian random vector in $\mathbb{D}$, i.e., $f(Z)$ is measurable for every $f \in \mathbb{D}^*$, and that every finite linear combination $\sum_i \alpha_i f_i(Z), \alpha_i \in \mathbb{R}, f_i \in \mathbb{D}^*$, is Gaussian.

In order to derive an anti-concentration bound for $\psi(Z)$, we will reduce the problem to an anti-concentration bound for a supremum over a not necessarily separable Gaussian process. For the latter, we adapt the anti-concentration for suprema of separable Gaussian processes derived by Giessing (2023a). For this purpose, we use the following representation result for Lipschitz continuous and sublinear functionals. Its proof essentially follows by the Hahn-Banach dominated extension theorem.

Lemma 44. *Let $\psi : \mathbb{D} \to \mathbb{R}$ be sublinear and continuous at zero. Then it can be represented as*

$$\psi(x) = \sup\{f(x) \,|\, f \in \mathbb{D}^* : \forall y \in \mathbb{D} : f(y) \leq \psi(y)\}.$$

Conversely, any supremum of a uniformly bounded family of continuous linear func-

tionals is a Lipschitz continuous and sublinear functional.

Lemma 44 lets us recast the problem as follows. Denote by $\mathcal{F} = \{f \in \mathbb{D}^* \,|\, \forall x \in \mathbb{D} : f(x) \leq \psi(x)\}$. Then $\psi(Z) = \sup_{f\in\mathcal{F}} f(Z)$ is the supremum over the Gaussian process $G := \{f(Z) : f \in \mathcal{F}\}$. This process has many nice properties such as almost surely bounded and Lipschitz continuous sample paths. However, it might fail to have a separable version since the dual space $\mathbb{D}^*$ might not be separable as for example when $\mathbb{D}$ is the space of continuous functions from $[0,1] \to \mathbb{R}$ endowed with the supremum norm. Instead, by tightness of Z and special properties of $\mathcal{F}$, we can approximate $\psi(Z)$ by a supremum over a countable set $\tilde{\mathcal{F}}$ and therefore have sufficient regularity to apply the arguments in Giessing (2023a).

Theorem 14. *Let $\psi : \mathbb{D} \to \mathbb{R}$ be a sublinear and continuous functional and let Z be a tight $\mathbb{D}$-valued Gaussian random vector. Let $\mathcal{F}_0 = \{f \in \mathcal{F} : \mathrm{Var}(f(Z)) = 0\} \neq \emptyset$ and define $\bar{\mu} = \sup_{f\in\mathcal{F}_0} \mathrm{E}[f(Z)]$. Further, let $Y := \sup_{f\in\mathcal{F}\setminus\mathcal{F}_0} f(Z)$, then, for any $t \in \mathbb{R}$ and $\varepsilon > 0$,*

$$\mathrm{P}(t \leq \psi(Z) \leq t+\varepsilon) \leq \frac{\varepsilon\sqrt{12}}{\sqrt{\mathrm{Var}(Y) + \varepsilon^2/12}} + \mathrm{P}(\psi(Z) = \bar{\mu})\mathbb{1}(\bar{\mu} \in [t, t+\varepsilon]).$$

If $\mathcal{F}_0 = \emptyset$, this bound reduces to

$$\mathrm{P}(t \leq \psi(Z) \leq t+\varepsilon) \leq \frac{\varepsilon\sqrt{12}}{\sqrt{\mathrm{Var}(\psi(Z)) + \varepsilon^2/12}}.$$

We defer the proof to section 2.A.2 in the Appendix.

Remark 1. *Suppose that Z is centered in the sense that $\mathrm{E}[f(Z)] = 0$ for all $f \in \mathbb{D}^*$. Then we can replace $\bar{\mu}$ in the upper bound by zero and $\mathrm{Var}(Y)$ by the variance of $\psi(Z)$. If moreover $\mathcal{F}_0 = \emptyset$, i.e., G has non-degenerate marginals, then the upper bound reduces to the bound in Remark 2 in Giessing (2023a).*

Remark 2. *By Theorem 1.1 in Bobkov and Chistyakov (2015), it holds for arbitrary continuous random variables X*

$$\sup_{t\in\mathbb{R}} \mathrm{P}(t \leq X \leq t+\varepsilon) \geq \frac{\varepsilon/\sqrt{12}}{\sqrt{\mathrm{Var}(X) + \varepsilon^2/12}}.$$

This shows that the bound in Theorem 14 is of the right order as $\varepsilon \to 0$ when $\mathcal{F}_0 = \emptyset$.

Remark 3. *If $Z \sim \mathcal{N}(\mu, \sigma^2)$, then for any $t \in \mathbb{R}$ and $\varepsilon > 0$,*

$$\mathrm{P}(t \leq Z \leq t + \varepsilon) \leq \frac{\varepsilon}{\sqrt{\sigma^2 + \varepsilon^2/2}}.$$

For a derivation, see Remark 3 in Giessing (2023a). Thus, as in Giessing (2023a), sublinear transformations of Gaussian random vectors possess essentially the same anti-concentration properties as a Gaussian random variable with variance $\mathrm{Var}(\psi(Z))$.

In applications, the functional ψ and the vector Z typically restrict the underlying vector space, but there is some degree of freedom in the choice of the norm or the choice of the Banach space. This choice faces a trade-off. On the one hand, while the sublinearity of ψ only depends on the vector space structure, the continuity of ψ is directly affected by the choice of the norm and the stronger the norm, the easier to achieve continuity. On the other hand, the choice of the norm affects the tightness of Z and implies in particular that $\mathrm{E}[\|Z\|^k] < \infty$ for any $k \in \mathbb{N}$. Therefore, a stronger norm imposes stronger integrability properties on Z. In order to illustrate this, consider for example $Z = (\lambda_n g_n)_{n\in\mathbb{N}}$, where $(g_n)_{n\in\mathbb{N}}$ denotes an i.i.d. sequence of standard normally distributed random variables and λ_n is a sequence of nonnegative real numbers. If $(\lambda_n) \in \ell_p$ for $p \in (1, \infty)$ but not in ℓ_1, then $\mathrm{E}[\|Z\|_p^p] < \infty$ but $\mathrm{E}[\|Z\|_1] = \infty$.

The bound in Theorem 14 is dimension-free and only imposes weak requirements on the marginal distributions of Z. As they are dimension-free, they apply equally to finite as well as infinite dimensional settings. In comparison to other proposals in the literature (cf. Chernozhukov et al. (2014b), Chernozhukov et al. (2015a), Giessing (2023a)), we do not require that the marginal variances are bounded away from zero.[2] This allows for settings where G has degenerate marginals. For example, consider $\mathbb{D} = \mathbb{R}$ and $x \mapsto \psi(x) = (x)_+$ and $Z \sim \mathcal{N}(0, 1)$. In this case, $G = \{Z, 0\}$ includes a degenerate normal with variance 0. Further, this allows for settings where marginal variances decay to zero. Examples include settings such as in PCA, count data or functional data analysis as demonstrated in Lopes et al. (2020). Further, we also do not require Z to be centered. This can be of interest in the construction of tests. For example, in the setting of Example 8, this allows us to derive a distributional approximation of the test statistics $T_{n,\ell}$ both under null hypothesis and the

[2]Strictly speaking, we consider a more specialized setting than the aforementioned papers. However, one may compare our bound to these papers if one interprets our bound as an anti-concentration bound for the supremum of the Gaussian process G.

alternative as we will show in the following section. Therefore, this property can be useful in the power analysis of tests as noted in Chernozhukov et al. (2016). Finally, our bound does not require the Gaussian process G to be separable. This allows to apply our bound to the space of continuous functions or L_1.

2.3.1 Lower bounds on the variance

The proposed bounds are only operational when we can compute the variance of $\psi(Z)$ or at least bound it from below. For instance, for norms in high-dimensional statistics, the variance often converges to zero as the dimension increases, see for example Biau and Mason (2015) for p-norms. In such situations, it is crucial to understand at which rate the variance decreases. In the case of suprema of separable Gaussian processes, lower bounds on the variance have been derived by Giessing (2023a) and lower bounds on the variance of p-norms of Gaussian k-vectors have been derived in Giessing and Fan (2023). See also the further discussion and references in Giessing (2023a) and Giessing and Fan (2023).

Before presenting our lower bound on the variance, we apply the lower bound in Proposition 1 in Giessing (2023a) to our anti-concentration bound.[3]

Corollary 7. *Suppose the assumptions of Theorem 14 hold. Furthermore, suppose that Z is centered and $\mathrm{Var}(f(Z)) \geq \underline{\sigma} > 0$ for all $f \notin \mathcal{F}_0$. Then, $0 \leq \mathrm{E}[Y] < \infty$ and*

$$\begin{aligned}\mathrm{P}(t \leq \psi(Z) \leq t + \varepsilon) \leq & 15\sqrt{12}\frac{\varepsilon}{\underline{\sigma}^2}(\mathrm{E}[Y] + \underline{\sigma}) \\ & + \mathrm{P}(\psi(Z) = \bar{\mu})\mathbb{1}(\bar{\mu} \in [t, t + \varepsilon]).\end{aligned}$$

The resulting anti-concentration bound is similar to the bounds in Chernozhukov et al. (2014b) and Chernozhukov et al. (2015a) although with a worse constant. In particular, besides the lower bound on the marginal variances, it only depends on Z through the expected value of the supremum Y. This simplifies the applicability of this bound considerably as there is a wide array of techniques to bound the expected value of suprema of Gaussian processes in the literature.

Further, we propose a lower bound on the variance of the supremum of a separable Gaussian process which does not rely on a lower bound of the marginal variances

[3]While Proposition 1 in Giessing (2023a) requires G to be separable, this result can be relaxed to only require that we consider a supremum over a countable index set. See also the comment to Lemma 45 below.

but instead only uses an upper bound. Thus, this lower bound allows for example for variance decay as in Lopes et al. (2020).

Lemma 45. *Let $X = \{X_u : u \in U\}$, $U \neq \emptyset$, be a separable and centered Gaussian process such that $\mathrm{Var}(X_u) > 0$ for all $u \in U$. Set $Y = \sup_{u \in U} X_u$ and assume that $Y < \infty$ a.s. Then, $0 \leq \mathrm{E}[Y] < \infty$ and*

$$\mathrm{Var}(Y) \geq \begin{cases} \bar{\sigma}^2/32 & , \text{ if } \mathrm{E}[Y] < \bar{\sigma}/2 \\ \frac{1}{5}\bar{\sigma}^2(2\bar{\sigma}^2 + (\mathrm{E}[Y]+1)^2)^{-1} & , \text{ if } \mathrm{E}[Y] \geq \bar{\sigma}/2. \end{cases}$$

Here, $\bar{\sigma} = \sup_{u \in U} \sqrt{\mathrm{Var}(X_u)}$ denotes the weak variance of X.

The assumption that the Gaussian process X is separable is not essential for this result to hold. In particular, it applies to $G = \{f(Z) : f \in \mathcal{F}\}$ in our setting as follows from Lemma 49 in the Appendix. Heuristically, this follows since we only need that the supremum of G has a separable index set and the latter follows by tightness of Z together with special properties of $\mathcal{F}$.

As an application of this lower bound, we obtain the following Corollary to Theorem 14.

Corollary 8. *Suppose the assumptions of Theorem 14 hold. Furthermore, suppose that Z is centered and that there exists some $f \in \mathcal{F}$ so that $\mathrm{Var}(f(Z)) \geq c > 0$ and that $\bar{\sigma} = \sup_{\|f\|^* = 1} \sqrt{\mathrm{Var}[f(Z)]} < \infty$. Then, if $\mathrm{E}[Y] \geq L\bar{\sigma}/2$ for some $L \geq 1$,*

$$\begin{aligned} \mathrm{P}(t \leq \psi(Z) \leq t + \varepsilon) \leq & \frac{5\sqrt{12}}{c}\varepsilon\sqrt{(\mathrm{E}[Y]+1)^2 + 2L^2\bar{\sigma}^2} \\ & + \mathrm{P}(\psi(Z) = \bar{\mu})\mathbb{1}(\bar{\mu} \in [t, t+\varepsilon]). \end{aligned}$$

2.4 Applications

In the following, we will apply our anti-concentration bound to derive Kolmogorov bounds for Gaussian approximation of sums of independent high-dimensional random vectors in section 2.4.1 and of kernel-type estimators in section 2.4.2.

2.4.1 Sums of high-dimensional random vectors

In this section, we propose a bound on the Kolmogorov distance for transformations ψ of sums of high-dimensional random vectors. Our main result is the following.

Proposition 2. *Let $p \geq 1$ and $\psi : \mathbb{R}^k \to \mathbb{R}$ be sublinear and L-Lipschitz continuous with respect to the ℓ_p norm. Suppose $X_1, \ldots, X_n$ are independent random k-vectors, and let*

$$\beta_p = \sum_{i=1}^{n} \mathrm{E}[\|X_i - \mathrm{E}[X_i]\|_2^2 \|X_i - \mathrm{E}[X_i]\|_p] + \sum_{i=1}^{n} \mathrm{E}[\|g_i\|_2^2 \|g_i\|_p]$$

be finite, where $g_1, \ldots, g_n$ is a sequence of independent random k-vectors such that $g_i \sim \mathcal{N}(0, \mathrm{Var}(X_i))$ for all $i = 1, \ldots, n$. Further, denote by $\phi_p(k) = \sqrt{pk^{2/p}}$ for $p \in [1, \infty)$ and $\phi_\infty(k) = \sqrt{2 \log 2k}$ and let $c_p(k) = k^{\frac{2-p}{2p}}$ if $p \in [1, 2)$ and $c_p(k) = 1$ if $p \in [2, \infty]$. Lastly, set $S_n = X_1 + \cdots + X_n$. Then, for $Z_n \sim \mathcal{N}(\mathrm{E}[S_n], \mathrm{Var}(S_n))$ and if $\mathrm{Var}(\psi(Z_n)) > 0$, for all $\tau \in \mathbb{R}$,

$$\begin{aligned} \sup_{t \geq \tau} |\mathrm{P}(\psi(S_n) \leq t) - \mathrm{P}(\psi(Z_n) \leq t)| \leq & C \Delta_n \log\left(\frac{\phi_p(k)^3}{c_p(k) \Delta_n} \vee e \right) \\ & + \mathrm{P}(\psi(Z_n) = \bar{\mu}) \mathbb{1}(\bar{\mu} \geq \tau) \end{aligned}$$

for some constant C independent of n, k, p and the distribution of X and where

$$\Delta_n = \left(\frac{L^3 \beta_p \phi_p(k)^2}{(\mathrm{Var}(\psi(Z_n)))^{3/2}} \right)^{1/4}.$$

The proof relies on a Yurinskii Coupling due to Belloni et al. (2019b). The independence assumption can be weakened to Martingale assumptions by using results from Cattaneo et al. (2022). The assumption that ψ is L-Lipschitz is already implied if ψ is sublinear and continuous. In this sense, it is here only required that the Lipschitz constant (or an upper bound) is known. Further, as this result is non-asymptotic, ψ and therefore also $\mathcal{F}$ and L are allowed to depend on n. This is for example needed in the context of Example 8 when the dimension is allowed to increase with the sample size. Finally, note that the result does not assume that the X_i are centered.

In the derivation of the bound, we were agnostic about the cardinality of $\mathcal{F}$.

When $\mathcal{F}$ only has finite cardinality, the results may be improved by using the Yurinskii Coupling in Chernozhukov et al. (2014a). This is because we use a coupling for the whole vector S_n and rely on the Lipschitz continuity of ψ to obtain a coupling for $\psi(S_n)$. Chernozhukov et al. (2014a) in comparison, can be used to derive couplings for $\psi(S_n)$ directly which might lead to improvements when $\mathcal{F}$ has a relatively low complexity.

Example 10. Consider for example $\psi(x) = \|x\|_2$ and let $Y_1, \dots, Y_n$ be i.i.d. and centered with values in $\mathbb{R}^k$. For simplicity, suppose that $\mathrm{Var}(Y_i) = I_k$ and take $X_i = Y_i/\sqrt{n}$. Then, we can apply Proposition 2 with $p = 2$, $L = 1$, $\mathrm{Var}(\psi(Z_n)) \approx 1$[4] and

$$\beta_2 = \frac{\mathrm{E}[\|Y_i\|_2^3]}{\sqrt{n}} + \frac{\mathrm{E}[\|g_i\|_2^3]}{\sqrt{n}}.$$

Suppose that $\max_{1 \le j \le k} \mathrm{E}[|Y_{ij}|^3] \le \beta$ for all i, j, then $\beta_2 = O(\sqrt{k^3/n})$. Thus,

$$\sup_{t \in \mathbb{R}} |\mathrm{P}(\psi(S_n) \le t) - \mathrm{P}(\psi(Z_n) \le t)| = O\left\{ \left(\frac{k^5}{n} \right)^{1/8} \log(nk) \right\}.$$

The Kolmogorov distance converges to zero whenever $k^5/n = o(\log^8 n)$, thus having similar restrictions as a Yurinskii coupling for S_n and Z_n with respect to the Euclidean distance.

Example 11. The dependence on k in the above example can be improved, when the Y_i are highly correlated across coordinates and have light tails. Consider again the setup as in Example 10 with the exception that now Y_i has covariance matrix Σ with ordered eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_k \ge 0$ and associated eigenvectors $e_1, \dots, e_k$. Say, for simplicity, the $e_j^\intercal X_i$ are uniformly bounded or sub-Gaussian. Then, we show in Lemma 53 in the Appendix that

$$\beta_2 = O\left(\sqrt{\frac{\mathrm{tr}(\Sigma)\lambda_1 \log k}{n}} \right).$$

Further, by the variance bound in Lemma 45,

$$\mathrm{Var}(\psi(Z)) \ge O\left(\frac{\lambda_1}{\lambda_1 + (\mathrm{tr}(\Sigma)\lambda_1 \log k)^2} \right)$$

[4] See Theorem A in Lytova and Tikhomirov (2019).

and therefore if $\mathrm{P}(\psi(Z_n) = 0) = 0$,

$$\sup_{t\in\mathbb{R}} |\mathrm{P}(\psi(S_n) \le t) - \mathrm{P}(\psi(Z_n) \le t)| = O\left\{ \left(\frac{k^2 \operatorname{tr}(\Sigma)^8 \lambda_1^2}{n} \right)^{1/8} \log^2(nk) \right\}.$$

Thus, if the trace of Σ is uniformly bounded, the requirements on k can be reduced to $k^2/n = o(\log^{16} n)$.

Example 12 (Testing many inequalities). The same arguments as in Examples 10 and 11 apply to the framework of testing many inequalities studied in Example 8. Here we only study the test statistic $T_{n,2}$ and $S_{n,2}$; the sup-based test statistics can be analyzed similarly. Suppose for simplicity, that the variances $\sigma_j^2 = \mathrm{Var}(X_{ij}) > 0$ are known. Then, $S_{n,2}$ can be written as

$$S_{n,2} = \psi(\sqrt{n}(\hat{\mu} - \mu)/\sigma)$$

and can be approximated by the distribution of $\psi(Z_n)$ for some centered Gaussian vector Z_n. If the covariance matrix of X_i is nonsingular, $\mathcal{F}_0 = \emptyset$ and therefore

$$\sup_{t>0} |\mathrm{P}(S_{n,2} \le t) - \mathrm{P}(\psi(Z_n) \le t)| \to 0$$

under the conditions in Examples 10 or 11.

We can use this result in order to approximate the $(1-\alpha)$-quantile of $S_{n,2}$ for $\alpha \in (0, 1/2)$. Let $c_{n,1-\alpha}$ denote the $(1-\alpha)$-quantile of $\psi(Z_n)$. Since Z_n is centered, $c_{n,1-\alpha} > 0$ and does not depend on the particular choice of μ. Therefore,

$$|\mathrm{P}(S_{n,2} > c_{n,1-\alpha}) - (1-\alpha)| \le \sup_{t>0} |\mathrm{P}(S_{n,2} \le t) - \mathrm{P}(\psi(Z_n) \le t)| \to 0.$$

Moreover, we can use Proposition 2 to derive an approximation of the quantiles of $T_{n,2}$ directly. This is because Proposition 2 does not require the approximating Gaussian vector to be centered. In particular, we can approximate $T_{n,2}$ by $\psi(Y_n)$ for some Gaussian vector Y_n with mean $\mu = (\sqrt{n}\mu_j/\sigma_j)_{j=1}^k$. Thus, we obtain in this setting

$$\sup_{t>0} |\mathrm{P}(T_{n,2} \le t) - \mathrm{P}(\psi(Y_n) \le t)| \to 0$$

and since $c_{n,1-\alpha} > 0$ also

$$|\mathrm{P}(T_{n,2} \leq c_{n,1-\alpha}) - \mathrm{P}(\psi(Y_n) \leq c_{n,1-\alpha})| \to 0$$

both under the null hypothesis and the alternative. This allows for the study of size and power properties of a test based on the critical values $c_{n,1-\alpha}$. In particular, size control follows by construction of $S_{n,2}$

$$\sup_{\mu \leq 0} \mathrm{P}(T_{n,2} > c_{n,1-\alpha}) \leq \mathrm{P}(S_{n,2} > c_{n,1-\alpha}) \to \alpha.$$

For power on the other hand, we have for any μ_n with $\psi(\mu_n) \geq c_{n,1-\alpha}$

$$\begin{aligned}\mathrm{P}(\psi(Z_n + \mu_n) > c_{n,1-\alpha}) \geq& \sup_{f \in \mathcal{F}} \mathrm{P}(f(Z_n) > c_{n,1-\alpha} - f(\mu_n)) \\ \geq& 1 - \Phi\left(\frac{c_{n,1-\alpha} - \psi(\mu_n)}{\bar{\sigma}}\right) \geq \frac{1}{2},\end{aligned}$$

where Φ denotes the cdf of the standard normal distribution and $\bar{\sigma}^2$ denotes the weak variance, i.e. $\bar{\sigma}^2 = \sup_{f \in \mathcal{F}} \mathrm{Var}(f(Z_n))$. Thus, by Markov's inequality, this test is consistent against any alternative satisfying $\psi(\mu_n)/\mathrm{E}[\psi(Z_n)] \to \infty$.[5]

In practice, one often faces estimators $\hat{\theta}_n$ for some parameter of interest θ which cannot be written as a sum of independent random vectors X_i, but which are asymptotically linear in the sense that

$$\sqrt{n}(\hat{\theta}_n - \theta_n) = \sum_{i=1}^{n} X_i + o_p(\delta_n),$$

for some independent random vectors $X_i \in \mathbb{R}^k$ and some sequence of real numbers δ_n. In particular, such linear expansions often hold for high-dimensional M-estimators such as maximum likelihood estimators or least squares estimators. If the linearization error converges sufficiently fast to zero, we can also obtain a bound on the Kolmogorov distance between the distributions of $\sqrt{n}\psi(\hat{\theta}_n - \theta_n)$ and $\psi(Z_n)$.

[5]Note that this power argument does not rely on any properties specific to this example but applies analogously for any sublinear and continuous ψ and tight Gaussian random vector Z_n. In particular, it also applies to infinite-dimensional settings.

Corollary 9. *Consider the setup in Proposition 2. Suppose further that*

$$\left\| \sqrt{n}(\hat{\theta}_n - \theta_n) - \sum_{i=1}^{n} X_i \right\|_p = o_p(\delta_n),$$

for some δ_n *satisfying* $\delta_n = O(\sqrt{\mathrm{Var}(\psi(Z_n))})$. *If in addition,* $\mathrm{P}(\psi(Z_n) = \bar{\mu}) \to 0$ *and*

$$\Delta_n \log\left(\frac{\phi_p(k)^3}{c_p(k)\Delta_n} \vee e \right) \to 0,$$

then

$$\sup_{t \in \mathbb{R}} |\mathrm{P}(\sqrt{n}\psi(\hat{\theta}_n - \theta_n) \leq t) - \mathrm{P}(\psi(Z_n) \leq t)| \to 0, \qquad \text{as } n \to \infty.$$

To the best of my knowledge, there is no similar bound for the Kolmogorov distance for sublinear and continuous functionals in the literature. In particular, there does not seem to be any result applicable to the test statistic $T_{n,2}$ in Example 8 when the number of inequalities is allowed to increase with the sample size. On the other hand, there is a growing literature on Gaussian approximation of ℓ_p norms for sums of high-dimensional random vectors. In particular, for the case of the Euclidean distance, Bentkus (2003), Bentkus (2005) and Raič (2019) obtained a bound of order $k^{1/4} \sum_{i=1}^{n} \mathrm{E}[\|\Sigma^{-1} X_i\|_2^3]$ and the $k^{1/4}$ term can be removed when $\Sigma = I_k$. In the setting of Example 10, this reduces to an upper bound of order $(k^3/n)^{1/2}$ which is better than our bound. Fang and Koike (2024) derived in the same setting, under bounded 4th moments, a bound of the form $n^{-1/8} + (k/n)^{1/6}$. For the case of the maximum norm, Chernozhukov et al. (2013a), Chernozhukov et al. (2017) and Chernozhuokov et al. (2022) derived bounds for sub-exponential X_i with non-degenerate marginal variances only requiring $(\log^5(nk)/n)^{1/4}$ thus allowing for very high-dimensional vectors. Fang and Koike (2021) derive a bound of order $(\log k \log^2 n/n)^{1/2}$ under the assumption that the X_i have a log concave density and under a lower bound on the smallest eigenvalue. Lopes et al. (2020) assume decaying marginal variances and establish a bound that is independent of k and only slightly slower than $n^{-1/2}$. Moreover, Giessing and Fan (2020), Giessing (2023b) and Giessing and Fan (2023) establish bounds for general ℓ_p norms which are dimension-free in that they only depend on the moments of $\|X_i\|_p$. In summary, while there are many results in the literature which yield better rates than Propo-

sition 2 for specific norms, our result also applies to settings where ψ is not a norm and therefore complements the results in the literature.

2.4.2 Kernel-type Estimators

Consider again Example 9. There, the functional ψ depends on a nonparametric estimator $\hat{\theta}_n$ of an unknown function θ_0. In practice, one may estimate θ_0 using kernel or series methods. These methods can be approximated by local or series empirical processes. Both local and series empirical processes can be characterized as empirical processes whose classes change with n and their complexity diverges as $n \to \infty$. In particular, these empirical processes do not have tight limits. Still, we can obtain distributional approximations in Kolmogorov distance using couplings due to Chernozhukov et al. (2013b) and Chernozhukov et al. (2014a) together with our anti-concentration bound in Theorem 14. Here, we only consider kernel methods, although similar results can be derived for series estimators.

Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be i.i.d. random variables taking values in the product space $\mathcal{Y} \times \mathbb{R}^d$, where $(\mathcal{Y}, \mathcal{A}_\mathcal{Y})$ is an arbitrary measurable space. Suppose that there is a measurable function $g_n : \mathcal{Y} \times \mathcal{I}_y \to \mathbb{R}$ for a compact subset $\mathcal{I}_y$ of $\mathbb{R}^{d_1}$. Let $k(\cdot)$ be a kernel function on $\mathbb{R}^d$, that is, $k(\cdot)$ is integrable with respect to the Lebesgue measure on $\mathbb{R}^d$ and its integral on $\mathbb{R}^d$ is normalized to be 1, but we do not assume $k(\cdot)$ to be nonnegative, that is, higher order kernels are allowed. Let h_n be a sequence of positive constants such that $h_n \to 0$ as $n \to \infty$, and let $\mathcal{I}_x$ be a compact subset of $\mathbb{R}^d$. Consider the kernel-type statistics

$$S_n(x, y) = \frac{1}{n h_n^d p(x)} \sum_{i=1}^{n} g_n(Y_i, y) k\left(\frac{X_i - x}{h_n}\right), \qquad (x, y) \in \mathcal{I}_x \times \mathcal{I}_y,$$

where $p(\cdot)$ denotes a Lebesgue density of the distribution of X_1.

Typically, under suitable regularity conditions, $S_n(x, y)$ will be a consistent estimator of $\mathrm{E}[g_n(Y_1, y) \,|\, X_1 = x]$. For example, when $g = 1$, $S_n(x, y)$ consistently estimates the marginal density of X, $p(x)$; when $\mathcal{Y} = \mathbb{R}$ and $g(y, \cdot) = y$, $S_n(x, y)$ consistently estimates the conditional mean function of Y given X, $\mathrm{E}[Y_1 \,|\, X_1 = x]$; and when $\mathcal{Y} = \mathbb{R}$ and $g_n(\cdot, y) = h_n^{-1} k(h_n^{-1}(\cdot - y)), y \in \mathbb{R}$, $S_n(x, y)$ will be a consistent estimator of $p_{Y|X}(y, x)$, the conditional density of Y_1 at y given $X_1 = x$.

By our representation result, we can express $\psi(S_n - \mathrm{E}[S_n])$ as a supremum of an

empirical process. For this purpose, let

$$\mathcal{F}_n = \{(Y_i, X_i) \mapsto f(p(\cdot)^{-1} g_n(Y_i, \cdot) k(h_n^{-1}(X_i - \cdot))) : f \in \mathcal{F}\}$$

and denote by $\mathbb{P}_n$ the empirical measure that puts mass $1/n$ at each observation. Formally, $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{(Y_i, X_i)}$, where $\delta_{(y,x)}$ denotes the Dirac measure at the point $(y, x) \in \mathcal{Y} \times \mathbb{R}^d$. Then, we can write

$$\sqrt{nh_n^d}\psi(S_n - \mathrm{E}[S_n]) = \sup_{f \in \mathcal{F}_n} \sqrt{\frac{n}{h_n^d}} \int f d(\mathbb{P}_n - P) =: \sup_{f \in \mathcal{F}_n} h_n^{-d/2} \mathbb{G}_n(f).$$

Motivated by this representation, we will approximate the distribution of $\psi(S_n - \mathrm{E}[S_n])$ by the distribution of a supremum of some centered Gaussian process Z_n indexed by $h_n^{-d/2}\mathcal{F}_n$ with covariance function

$$\Sigma_n(f, \tilde{f}) := h_n^{-d} \mathrm{E}[Z_n(f) Z_n(\tilde{f})] = h_n^{-d} \mathrm{Cov}[f(Y_1, X_1), \tilde{f}(Y_1, X_1)], \qquad f, \tilde{f} \in \mathcal{F}_n. \tag{2.2}$$

That is, Z_n has the same covariance process as the empirical process $h_n^{-d/2}\mathbb{G}_n$. Thus, we approximate the distribution of $\sqrt{nh_n^d}(S_n - \mathrm{E}[Z_n])$ by a Gaussian that matches the first two moments. In particular, since the covariance function depends on the bandwidth h_n, such an approximation may better capture the dependence on the choice of the tuning parameter than a distributional approximation in the limit where $h_n \to 0$.

For our approximation results, we distinguish two cases according to the complexity of $\mathcal{F}_n$. Following Chernozhukov et al. (2014a), we will measure the complexity using the concept of a VC type class. In order to introduce this concept, we need some further notation. For $\varepsilon > 0$, an *ε-net* of a semimetric space (T, d) is a subset T_ε of T such that for every $t \in T$ there exists a point $t_\varepsilon \in T_\varepsilon$ with $d(t, t_\varepsilon) < \varepsilon$. The *$\varepsilon$-covering number* $N(\varepsilon, T, d)$ of T is the infimum of the cardinality of ε-nets of T, that is $N(T, d, \varepsilon) := \inf\{\mathrm{Card}(T_\varepsilon) : T_\varepsilon \text{ is an } \varepsilon\text{-net of } T\}$. For a class of measurable functions $\mathcal{F}$ on some measurable space $(S, \mathcal{S})$, a function F is an envelope of $\mathcal{F}$ if $\sup_{f \in \mathcal{F}} |f(x)| \leq F(x)$ for all $x \in S$.

Definition 4 (VC type class)**.** *Let $\mathcal{F}$ be a class of measurable functions on a measurable space $(S, \mathcal{S})$, to which a measurable envelope F is attached. We say that $\mathcal{F}$ is VC type with envelope F if there are constants $A, v > 0$ such that* $\sup_Q N(\varepsilon \|F\|_{Q,2}, \mathcal{F},$

$L_2(Q)) \leq (A/\varepsilon)^v$ *for all* $0 < \varepsilon \leq 1$*, where the supremum is taken over all finitely discrete probability measures on* $(S, \mathcal{S})$*.*

As noted by Chernozhukov et al. (2014a), the VC type class is a wider concept than VC subgraph class (van der Vaart and Wellner (1996), Chapter 2.6). The VC type property is stable under summation, product, or more generally Lipschitz-type transformations, which makes it much easier to check whether a function class is VC type (see Lemma A.6 in the supplementary material of Chernozhukov et al. (2014a)).

We derive two different approximation results. Our first result in Proposition 3 requires that $\mathcal{F}_n$ is a VC type class uniformly in n. This assumption allows us to directly couple the supremum of the empirical process and results in weak restrictions on the bandwidth sequence. Our second result in Proposition 4 does not require that $\mathcal{F}_n$ is a VC type class. This larger applicability comes at the price of slightly stronger restrictions on the choice of the bandwidth sequence. The VC type class assumption is for example satisfied in Example 9 for the sup-based test statistic while it fails for the L_2-based statistic.

For our first result, we make the following assumptions. These assumptions are essentially the same as in Chernozhukov et al. (2014a).

Assumption 14. *(B1)* $\mathcal{F}_n$ *is a VC type class uniformly in* n*.*

(B2) $\mathcal{G}$ *is a pointwise measurable class of functions* $\mathcal{Y} \to \mathbb{R}$ *which is continuous in* y *and uniformly bounded by some constant* B *which does not depend on* n*.*

(B3) $k(\cdot)$ *is a bounded and continuous kernel function on* $\mathbb{R}^d$*.*

(B4) The distribution of X_1 *has a Lebesgue density* $p(\cdot)$ *which is bounded away from zero on* $\mathcal{I}_x$*, that is,* $p(x) \geq \underline{p} > 0$ *for all* $x \in \mathcal{I}$*.*

(B5) $h_n \to 0$*,* $\log(1/h_n) = O(\log n)$ *and* $\operatorname{Var}(\psi(Z_n)) = O((nh_n^d) \log^8 n)$ *as* $n \to \infty$*.*

Assumptions (B3)-(B5) are rather standard for kernel-type estimators and rather mild. In particular, we are silent about the order of the kernel k. Only the assumption that $\mathcal{G}$ is uniformly bounded is restrictive when we are for example interested in the estimation of a conditional moment such as the conditional mean. However, this assumption can be weakened to moment restrictions on the envelope function of $\mathcal{G}$ using similar arguments as in Proposition 3.2 in Chernozhukov et al. (2014a).

Proposition 3. *Suppose that* $\psi : \mathcal{C}(\mathcal{I}_x \times \mathcal{I}_y) \to \mathbb{R}$ *is sublinear and L-Lipschitz continuous. Further, suppose that assumptions (B1)-(B5) are satisfied. Then, there*

is a sequence of tight Gaussian random variable Z_n in $\mathcal{C}(\mathcal{I}_x \times \mathcal{I}_y)$ with mean zero and covariance function Σ_n given in (2.2) such that

$$\begin{aligned}&\sup_{t\in\mathbb{R}}|\mathrm{P}(\sqrt{nh_n^d}\psi(S_n - \mathrm{E}[S_n]) \leq t) - \mathrm{P}(\psi(Z_n) \leq t)| \\ \leq & C\left\{\frac{1}{(\mathrm{Var}(\psi(Z_n)))^{1/8}}\left(\frac{\log^4 n}{nh_n^d}\right)^{3/8} + \frac{\log n}{n}\right\} + \mathrm{P}(\psi(Z_n) = 0),\end{aligned}$$

where $C > 0$, is a constant that only depends on $\|g\|_\infty$, $\|k\|_\infty$, $\underline{p}$ and L.

Proposition 3 only imposes mild restrictions on the choice of the bandwidth sequence. Indeed, if $\mathrm{P}(\psi(Z_n) = 0) = 0$, then the Kolmogorov distance in Proposition 3 converges to zero if $\log^{12} n/(\mathrm{Var}(\psi(Z_n))(nh_n^d)^3) \to 0$. By the variance lower bound in Lemma 45 and by standard arguments, this is satisfied whenever $\log^{25/6} n/(nh_n^d) \to 0$. Thus, Proposition 3 requires only slightly more than $\log n/(nh_n^d) \to 0$, which is needed for consistent estimation of $\mathrm{E}[S_n]$ in the sup-norm.

Proposition 3 is based on the coupling in Corollary 2.2 in Chernozhukov et al. (2014a). Proposition 3 is an extension of the results in Proposition 3.1 in Chernozhukov et al. (2014a) to allow for other types of suprema of local empirical processes. Further, our anti-concentration bound together with the lower bound on the variance does not require a lower bound on the marginal variances.

In practice, one is often not interested on the expected value of S_n since this is a smoothed version of the parameter of interest. Moreover, the density of X_i is usually unknown and therefore has to be estimated. We can also allow for such estimators as long as they can be asymptotically approximated by S_n. This is the content of the following Corollary.

Corollary 10. *Consider the setup in Proposition 3. Suppose further that the kernel estimator $(x, y) \mapsto \hat{\theta}_n(x, y)$ of some target function $(x, y) \mapsto \theta_n(x, y)$ has an asymptotic linear expansion uniformly in $(x, y) \in \mathcal{I}_x \times \mathcal{I}_y$*

$$\sqrt{nh_n^d}(\hat{\theta}_n(x, y) - \theta_n(x, y)) = \sqrt{nh_n^d}(S_n(x, y) - \mathrm{E}[S_n(x, y)]) + o_P(\delta_n),$$

for some δ_n satisfying $\delta_n = O(\sqrt{\mathrm{Var}(\psi(Z_n))})$. If in addition, $\mathrm{P}(\psi(Z_n) = 0) \to 0$ and

$$\frac{\log^{12} n}{(nh_n^d)^3 \mathrm{Var}(\psi(Z_n))} \to 0,$$

then

$$\sup_{t\in\mathbb{R}}|\mathrm{P}(\sqrt{nh_n^d}\psi(\hat{\theta}_n-\theta_n)\leq t)-\mathrm{P}(\psi(Z_n)\leq t)|\to 0, \qquad \text{as } n\to\infty.$$

Such uniform asymptotic linear expansions have been derived by Masry (1996) and Kong et al. (2010) for Nadaraya-Watson estimators and local polynomial estimators under mild assumptions. Implicit in the above expansion is the assumption that the bias is asymptotically negligible. This can be achieved for example by undersmoothing, that is, choosing the bandwidth smaller than the rate-optimal bandwidth.

For our second result, we rely on a Rio-Massart coupling for kernel estimators in Chernozhukov et al. (2013b). In comparison to the setup above, we restrict our attention here to univariate $Y_i \in \mathbb{R}$ and a single transformation g_n which is allowed to change with n. This still allows for the estimation of conditional moments but precludes for example an application to conditional density estimation. This restriction may be weakened when the class $\{g_{n,y} : y \in \mathcal{I}_y\}$ depends smoothly on y.

We make the following assumptions. These are a modified version of Condition R in Chernozhukov et al. (2013b).

Assumption 15. *(C1) The random vectors (Y_i, X_i) have bounded support with joint density bounded from above and below by some constants $\bar{f}$ and $\underline{f}$.*

(C2) g_n is bounded by some constant B_1 and continuously differentiable such that $\|g_n'\|_\infty \leq B_2$ for some constant B_2 uniformly over n.

(C3) The kernel function K is twice continuously differentiable product kernel function with support on $[-1,1]^d$.

(C4) $h_n \to 0$, $\log(1/h_n) = O(\log n)$ and $\Delta_n/\sqrt{\mathrm{Var}(\psi(Z_n))} = O(n^{-\xi})$ for some $\xi > 0$, where

$$\Delta_n = \left\{\sqrt{\frac{1}{n^{1/(d+1)}h_n}} + \sqrt{\frac{\log n}{nh_n^d}}\right\}.$$

In comparison to Assumption 14, we require here more smoothness on the kernel and the function g_n. Moreover, we directly assume boundedness of the Y_i which is restrictive in comparison to what is needed in order to establish pointwise convergence in distribution for kernel estimators.

Our second result is the following.

Proposition 4. *Suppose that $\psi : \mathcal{C}(\mathcal{I}_x) \to \mathbb{R}$ is sublinear and L-Lipschitz continuous. Further, suppose that assumptions (C1)-(C4) are satisfied. Then, there is a sequence of tight Gaussian random variable Z_n in $\mathcal{C}(\mathcal{I}_x)$ with mean zero and covariance function Σ_n given in (2.2) such that*

$$\begin{aligned}&\sup_{t\in\mathbb{R}}|\mathrm{P}(\sqrt{nh_n^d}\psi(S_n - \mathrm{E}[S_n]) \leq t) - \mathrm{P}(\psi(Z_n) \leq t)| \\ \leq & C\frac{\Delta_n}{\sqrt{\mathrm{Var}(\psi(Z_n))}}\left\{1 + \log\left(\frac{\sqrt{\mathrm{Var}(\psi(Z_n))}}{\Delta_n}\right)\right\} + \mathrm{P}(\psi(Z_n) = 0),\end{aligned}$$

where

$$\Delta_n = \left\{\sqrt{\frac{1}{n^{1/(d+1)}h_n}} + \sqrt{\frac{\log n}{nh_n^d}}\right\}.$$

Proposition 4 only imposes mild restrictions on the choice of the bandwidth sequence. Indeed, if $\mathrm{P}(\psi(Z_n) = 0) = 0$, then the Kolmogorov distance in Proposition 3 converges to zero if $\log^{2d+1}(\mathrm{Var}(\psi(Z_n))nh_n^{d+1})/(nh_n^{d+1}\,\mathrm{Var}(\psi(Z_n))^{d+1}) \to 0$. By the variance lower bound in Lemma 45 and by standard arguments, this is satisfied whenever $\log^{3d+2} n/(nh_n^{d+1}) \to 0$. In comparison to Proposition 3, Proposition 4 requires stronger conditions on the bandwidth sequence but still not much more than $\log n/(nh_n^d) \to 0$, which is needed for consistent estimation of $\mathrm{E}[S_n]$ in the sup-norm. It might be the case that this stronger requirement is a proof artifact and can be weakened by using a different coupling for the empirical process.

As for Proposition 3, we also have a result for kernel estimators which can be uniformly approximated by S_n. The same comments as for Corollary 10 apply.

Corollary 11. *Consider the setup in Proposition 4. Suppose further that the kernel estimator $x \mapsto \hat{\theta}_n(x)$ of some target function $x \mapsto \theta_n(x)$ has an asymptotic linear expansion uniformly in $x \in \mathcal{I}_x$.*

$$\sqrt{nh_n^d}(\hat{\theta}_n(x) - \theta_n(x)) = \sqrt{nh_n^d}(S_n(x) - \mathrm{E}[S_n(x)]) + o_P(\delta_n),$$

for some δ_n satisfying $\delta_n = O(\sqrt{\mathrm{Var}(\psi(Z_n))})$. If in addition, $\mathrm{P}(\psi(Z_n) = 0) \to 0$ and

$$\frac{\Delta_n}{\sqrt{\mathrm{Var}(\psi(Z_n))}}\left\{1 + \log\left(\frac{\sqrt{\mathrm{Var}(\psi(Z_n))}}{\Delta_n}\right)\right\} \to 0,$$

then

$$\sup_{t\in\mathbb{R}}|\mathrm{P}(\sqrt{nh_n^d}\psi(\hat{\theta}_n-\theta_n)\leq t)-\mathrm{P}(\psi(Z_n)\leq t)|\to 0, \qquad \text{as } n\to\infty.$$

2.5 Conclusion

In this paper, we establish an anti-concentration bound for sublinear and continuous functionals of tight Gaussian random vectors in real-valued Banach spaces. The bound is dimension-free and therefore equally applies to high-dimensional as well as nonparametric statistical settings. Its proof relies on a translation of the problem to the study of the anti-concentration behavior of suprema of tight Gaussian processes. In comparison to other anti-concentration bounds for Gaussian processes, we assume that the Gaussian process is tight instead of separable and therefore our bounds can e.g. be applied to the space of continuous or absolutely integrable functions for which the related Gaussian process is necessarily non-separable. Further, we impose only weak restrictions on the covariance structure of the Gaussian vectors and therefore allow for applications to arbitrary continuous and sublinear functionals.

As an illustration of the usefulness of our anti-concentration bound, we apply our bound to derive Berry-Esseen type bounds for continuous and sublinear functionals of high-dimen-sional mean vectors and kernel-type estimators. These bounds can be applied in high-dimensional or non-parametric one-sided testing problems which arise e.g. in moment inequality models or functional inequality testing.

Appendix

2.A Anti-Concentration Bound for $\psi(Z)$

2.A.1 Representation as a supremum

For the proof of our representation result, we use that continuous sublinear functionals are Lipschitz continuous. This is the content of the following Lemma whose proof is included for completeness.

Lemma 46. *Let $\psi : \mathbb{D} \to \mathbb{R}$ be sublinear. If ψ is continuous at zero, then ψ is Lipschitz continuous with Lipschitz constant $\sup_{\|x\|=1}|\psi(x)| < \infty$.*

Proof. We divide the proof into two steps. First we show that for sublinear functionals ψ the Lipschitz satisfies

$$\sup_{x \neq y} \frac{|\psi(x) - \psi(y)|}{\|x - y\|} = \sup_{x \neq 0} \frac{|\psi(x)|}{\|x\|}. \tag{2.3}$$

Secondly, we show that sublinear functionals which are continuous at zero satisfy $\sup_{\|x\|=1}|\psi(x)| < \infty$ and are therefore Lipschitz continuous by the first part.

Regarding the first part, by subadditivity of ψ, we have for any $x, y \in \mathbb{D} : x \neq y$

$$\psi(x) \leq \psi(x - y) + \psi(y) \qquad \text{and} \qquad \psi(y) \leq \psi(y - x) + \psi(x)$$

implying

$$-\psi(y - x) \leq \psi(x) - \psi(y) \leq \psi(x - y).$$

Therefore,

$$|\psi(x) - \psi(y)| \leq \max\{|\psi(y - x)|, |\psi(x - y)|\}$$

implying

$$\frac{|\psi(x)-\psi(y)|}{\|x-y\|} \leq \max\left\{\frac{|\psi(y-x)|}{\|x-y\|}, \frac{|\psi(x-y)|}{\|x-y\|}\right\}.$$

By positive homogeneity of ψ, this further implies

$$\sup_{x\neq y} \frac{|\psi(x)-\psi(y)|}{\|x-y\|} \leq \sup_{\|x\|=1} |\psi(x)|.$$

Further, by positive homogeneity of ψ, $\psi(0)=0$ and therefore

$$\sup_{x\neq 0} \frac{|\psi(x)|}{\|x\|} = \sup_{x\neq 0} \frac{|\psi(x)-\psi(0)|}{\|x-0\|} \leq \sup_{x\neq y} \frac{|\psi(x)-\psi(y)|}{\|x-y\|}.$$

This implies (2.3).

For the second step, suppose that ψ is continuous at zero. We will first show that ψ is then uniformly continuous. By continuity at zero, for any $\varepsilon > 0$, there exists a $\delta > 0$ such that $|\psi(x)| < \varepsilon$ for all x satisfying $\|x\| < \delta$. Now, for any $x, y \in \mathbb{D}$ s.t. $\|x-y\| < \delta$, we have by the above arguments

$$|\psi(x)-\psi(y)| \leq \max\{|\psi(y-x)|, |\psi(x-y)|\} < \varepsilon$$

and therefore ψ is uniformly continuous. By uniform continuity, we know that the modulus of continuity

$$w(\delta) = \sup\{|\psi(x)-\psi(y)| : \|x-y\| \leq \delta\}, \qquad \delta \geq 0,$$

exists and if finite for all $\delta \geq 0$. Hence, for any $x \neq 0$,

$$\frac{|\psi(x)|}{\|x\|} = \left|\psi\left(\frac{x}{\|x\|}\right)\right| \leq w(1) < \infty.$$

The claim follows. □

Now, the representation result follows by the Hahn-Banach dominated extension theorem.

Proof of Lemma 44: For any $x \in \mathbb{D}$, let $A = \{\lambda x : \lambda \in \mathbb{R}\}$ and define $f : A \to \mathbb{R}$ by $f(\lambda x) = \lambda\psi(x)$, $\lambda x \in A$. For $\lambda \geq 0$, we have by positive homogeneity $f(\lambda x) =$

$\psi(\lambda x)$, and for $\lambda < 0$ we have

$$f(\lambda x) = -f(|\lambda| x) = -\psi(|\lambda| x) \leq \psi(\lambda x),$$

where we have used that $0 = \psi(0) \leq \psi(\lambda x) + \psi(|\lambda| x)$ by sublinearity of ψ. Thus, $f(y) \leq \psi(y)$ for all $y \in A$.

By the Hahn-Banach dominated extension theorem we can extend f to a linear functional on the whole of $\mathbb{D}$. More rigorously, there exists a linear functional $g : \mathbb{D} \to \mathbb{R}$ such that $g(y) = f(y)$ for all $y \in A$ and

$$-\psi(-x) \leq g(x) \leq \psi(x), \qquad x \in \mathbb{D}.$$

Moreover, since ψ is continuous and sublinear it is Lipschitz continuous by Lemma 46 with Lipschitz constant $\sup_{\|x\|=1} |\psi(x)|$. Hence,

$$\|g\|^* = \sup_{\|x\|=1} |g(x)| \leq \sup_{\|x\|=1} |\psi(x)| < \infty$$

and therefore g is continuous.

Since $x \in \mathbb{D}$ in the construction of f was chosen arbitrarily, we have thus shown that for any $x \in \mathbb{D}$, there exists an $f \in \mathbb{D}^*$ such that $f \leq \psi$ and $f(x) = \psi(x)$. As for all other $g \in \mathbb{D}^*$ with $g \leq \psi$ it holds $g(x) \leq \psi(x)$, we have

$$\psi(x) = \sup\{f(x) \,|\, f \in \mathbb{D}^* : f \leq \psi\}$$

and the claim follows. □

2.A.2 Anti-Concentration Bound for $\psi(Z)$

Our proof relies on Theorem 1 in Giessing (2023a), which we reproduce here for convenience. Actually, this is only an implication of the theorem. The authors deal more generally with order statistics. In particular, note that X is not assumed to be centered.

Theorem 15 (Giessing (2023a) Thm 1:)**.** *Let* $X = \{X_i : 1 \leq i \leq n\}$ *be multivariate Gaussian such that* $\mathrm{Var}(X_i) > 0$ *and* $\mathrm{Cor}(X_i, X_j) < 1$ *for all* $1 \leq i \neq j \leq n$*. Then,*

for all $\varepsilon > 0$ *and any* $t \in \mathbb{R}$,

$$\mathrm{P}\left(t \leq \max_{1\leq i\leq n} X_i \leq t+\varepsilon\right) \leq \frac{\varepsilon\sqrt{12}}{\sqrt{\mathrm{Var}(\max_i X_i)+\varepsilon^2/12}}.$$

Proof of Theorem 14: This proof is only a slight extension of Remark 2 in Giessing (2023a) to allow for $\inf_{f\in\mathcal{F}} \mathrm{Var}(f(Z)) = 0$ and not necessarily separable Gaussian processes. We therefore follow closely his arguments.

First, we consider the case that $\mathrm{Var}(f(Z)) > 0$, for all $f \in \mathcal{F}$, that is, $\mathcal{F}_0 = \emptyset$. Without loss of generality, we may assume that $\mathrm{Var}(\psi(Z)) > 0$ as otherwise the upper bound is trivial. By Lemma 50 (vi), there exists a sequence of finite sets $\mathcal{F}_n \subset \mathcal{F}$ such that $Y_n := \max_{f\in\mathcal{F}_n} f(Z)$ converges to $Y := \sup_{f\in\mathcal{F}} f(Z)$ in probability as $n \to \infty$; hence $Y_n \xrightarrow{d} Y$. Moreover, by Lemma 49, we can replace $\mathcal{F}$ by a countable subset $F = \{f_n : n \in \mathbb{N}\} \subset \mathcal{F}$ so that $Y = \sup_n f_n(Z)$ a.s. This approximation is useful as we can bound $\mathrm{P}(t \leq Y_n \leq t+\varepsilon)$ using the anti-concentration bound in Theorem 1 in Giessing (2023a). This theorem applies since by assumption $\mathrm{Var}(f(Z)) > 0, f \in \mathcal{F}$, and without loss of generality $\mathrm{Cor}(f,g) < 1$, $f, g \in \mathcal{F}$ by Lemma 47. Moreover, by the reverse equivalence of moments for suprema of Gaussian processes[6] and the bounded convergence theorem, $\mathrm{Var}(Y_n) \to \mathrm{Var}(Y)$ as $n \to \infty$. Hence, by Theorem 1 in Giessing (2023a),

$$\lim_{n\to\infty} \mathrm{P}(t \leq Y_n \leq t+\varepsilon) \leq \lim_{n\to\infty} \frac{\varepsilon\sqrt{12}}{\sqrt{\mathrm{Var}(Y_n)+\varepsilon^2/12}} = \frac{\varepsilon\sqrt{12}}{\sqrt{\mathrm{Var}(Y)+\varepsilon^2/12}}. \tag{2.4}$$

It remains to show that $\mathrm{P}(t \leq Y_n \leq t+\varepsilon)$ converges to the corresponding probability with respect to Y. By the reverse triangle inequality, for all $t \in \mathbb{R}$ and every $\varepsilon > 0$,

$$|\mathrm{P}(t \leq Y_n \leq t+\varepsilon) - \mathrm{P}(t \leq Y \leq t+\varepsilon)| \leq 2 \sup_{t\in\mathbb{R}} |\mathrm{P}(Y_n \leq t) - \mathrm{P}(Y \leq t)|. \tag{2.5}$$

The latter converges to zero by Le Cams's inequality in 43, (2.4), $\mathrm{Var}(\psi(Z)) > 0$ and since $Y_n = Y + o_p(1)$. The claimed bound now follows by combining (2.4) and (2.5).

Next, suppose that $\mathcal{F}_0 \neq \emptyset$. By Lemma 48, there exists some $f_0 \in \mathcal{F}_0$ so that $\mathrm{E}[f_0(Z)] = \bar{\mu} < \infty$. Therefore, $\psi(Z) = Y\mathbb{1}(Y > \bar{\mu}) + \bar{\mu}\mathbb{1}(Y \leq \bar{\mu})$ with

[6] See Lemma 50(iii) below.

$Y := \sup_{f \in \mathcal{F} \setminus \mathcal{F}_0} f(Z)$. Hence, for any $t < \bar{\mu}$ and ε sufficiently small,

$$\mathrm{P}(t \leq \psi(Z) \leq t + \varepsilon) = 0 \tag{2.6}$$

which is smaller than the upper bound. For $t > \bar{\mu}$ and any $\varepsilon > 0$, $\psi(Z) = Y$ on the event $\{t \leq \psi(Z) \leq t + \varepsilon\}$. Thus, by the arguments in the first part of the proof,

$$\mathrm{P}(t \leq \psi(Z) \leq t + \varepsilon) \leq \frac{\varepsilon\sqrt{12}}{\sqrt{\mathrm{Var}(Y) + \varepsilon^2/12}}.$$

For $t = \bar{\mu}$, we have to be careful as $\mathrm{P}(\psi(Z) = \bar{\mu})$ might be positive. Note that since $\psi(Z) = Y\mathbb{1}(Y > \bar{\mu}) + \bar{\mu}\mathbb{1}(Y \leq \bar{\mu})$, we have for all $x > \bar{\mu}$

$$\mathrm{P}(\psi(Z) \leq x) = \mathrm{P}(Y \leq x \,|\, Y > \bar{\mu})\mathrm{P}(Y > \bar{\mu}) + \mathrm{P}(\psi(Z) = \bar{\mu}).$$

Since Y has some density f_Y by the first part of the proof, we have

$$\frac{d}{dx}\mathrm{P}(Y \leq x \,|\, Y > \bar{\mu}) = \frac{1}{\mathrm{P}(Y > \bar{\mu})}\frac{d}{dx}\int_{\bar{\mu}}^{x} f_Y(y)dy = \frac{f_Y(x)}{\mathrm{P}(Y > \bar{\mu})}$$

and therefore

$$\begin{aligned}\mathrm{P}(\bar{\mu} \leq \psi(Z) \leq \bar{\mu} + \varepsilon) &= \int_{\bar{\mu}}^{\bar{\mu}\varepsilon} f_Y(y)dy + \mathrm{P}(\psi(Z) = \bar{\mu}) \\ &= \varepsilon f_{Y+\varepsilon U}(\bar{\mu}) + \mathrm{P}(\psi(Z) = \bar{\mu}),\end{aligned}$$

where $f_{\psi(Z)+\varepsilon U}$ denotes the density of the convolution of the absolutely continuous part of $\psi(Z)$ with εU for $U \sim \mathrm{Unif}[0,1]$. Since Y is a supremum of a Gaussian process with non-degenerate marginals, the upper bound from the first part applies:

$$f_{Y+\varepsilon U}(\bar{\mu}) \leq \operatorname*{ess\,sup}_{t \in \mathbb{R}} f_{Y+\varepsilon U}(t) \leq \frac{\sqrt{12}}{\sqrt{\mathrm{Var}(Y) + \varepsilon^2/12}},$$

and implies

$$\mathrm{P}(\bar{\mu} \leq \psi(Z) \leq \bar{\mu} + \varepsilon) \leq \frac{\varepsilon\sqrt{12}}{\sqrt{\mathrm{Var}(Y) + \varepsilon^2/12}} + \mathrm{P}(\psi(Z) = \bar{\mu}).$$

For $t < \bar{\mu}$ and $t + \varepsilon > \bar{\mu}$, by (2.6) and the same argument as in the case $t = \bar{\mu}$,

$$\begin{aligned}&\mathrm{P}(t \leq \psi(Z) \leq t + \varepsilon) = \mathrm{P}(\bar{\mu} \leq \psi(Z) \leq \bar{\mu} + (t + \varepsilon - \bar{\mu}))\\ \leq &\frac{(\varepsilon + t - \bar{\mu})\sqrt{12}}{\sqrt{\mathrm{Var}(Y) + (\varepsilon + t - \bar{\mu})^2/12}} + \mathrm{P}(\psi(Z) = \bar{\mu}).\end{aligned}$$

By monotonicity of the right-hand side in $\varepsilon + t - \bar{\mu}$, we obtain

$$\mathrm{P}(t \leq \psi(Z) \leq t + \varepsilon) \leq \frac{\varepsilon\sqrt{12}}{\sqrt{\mathrm{Var}(Y) + \varepsilon^2/12}} + \mathrm{P}(\psi(Z) = \bar{\mu})$$

finishing the proof.

Note that for $\bar{\mu} = 0$, $\psi(Z) = Y\mathbb{1}(Y > 0)$ and therefore $\mathrm{Var}(\psi(Z)) \leq \mathrm{Var}(Y)$. Thus, in this case, the variance of Y in the bound can be replaced by the variance of $\psi(Z)$. □

Lemma 47. *Assume the setting of Theorem 14. For any $f, g \in \mathcal{F} \setminus \mathcal{F}_0$, we can assume $Cor(f, g) < 1$.*

Proof. Suppose there are $f, g \in \mathcal{F} \setminus \mathcal{F}_0$ so that $\mathrm{Cor}(f, g) = 1$. Then there exist $a > 0$ and $b \in \mathbb{R}$ so that $g(Z) = af(Z) + b$. Since $g \in \mathcal{F}$, we have $af(x) + b \leq \psi(x)$, $x \in \mathbb{D}$. In particular, this implies $b = 0$ since $g(0) = af(0) + b = b = \psi(0) = 0$. Moreover, we can assume that $a > 1$. As otherwise, if $a < 1$, we can switch the roles of f and g. If $a = 1$, then f and g coincide, and we can discard one of them.

Note that we can discard all f for which there does not exist an $x \in \mathbb{D} \setminus \{0\}$ with $f(x) = \psi(x)$. Indeed, if there would be such an f, then we can find for any $x \in \mathbb{D} \setminus \{0\}$, an $f_x \in \mathcal{F}$ with $f_x(x) = \psi(x)$ (take the one constructed in Lemma 44). Thus, for any x there exists some dominating function, and therefore we can discard f for the evaluation of the supremum.

Now, suppose that there exists some $x \in \mathbb{D} \setminus \{0\}$ such that $f(x) = \psi(x) > 0$. Since $a > 1$, g dominates f for all x for which $f(x) > 0$. But since there is some x such that $f(x) = \psi(x)$ this would imply that there is some x so that $g(x) > \psi(x)$. Hence, a cannot be larger than one.

If there does not exist $x \in \mathbb{D} \setminus \{0\}$ such that $f(x) = \psi(x) > 0$, but there exist such an x so that $g(x) = \psi(x) > 0$, then g dominates f whenever g is a maximizer. If this is also not the case, then we can assume without loss of generality that there exist $x \in \mathbb{D} \setminus \{0\}$ such that $f(x) = \psi(x) < 0$. Since $a > 1$, f dominates g for all x for which $f(x) < 0$. But since there is some x such that $f(x) = \psi(x)$ this would

imply that there is some x so that $g(x) < \psi(x)$. We distinguish two cases. Either both f and g are dominated for all x so that $\psi(x) < 0$, or there is some x so that at least one of them is a maximizer. In the first case we can discard both of them on this event. In the second case, we can discard g since $g(x) = af(x) < f(x) = \psi(x)$. The claim follows. □

Lemma 48. *Assume the setting of Theorem 14. If $\mathcal{F}_0 \neq \emptyset$, there exists some $f_0 \in \mathcal{F}_0$ such that $\mathrm{E}[f_0(Z)] = \bar{\mu} < \infty$.*

Proof. We will show that $\mathcal{F}_0$ is compact and that $f \mapsto \mathrm{E}[f(Z)]$ is continuous with respect to weak*-convergence. This then implies the result.

We start by recalling some results from functional analysis. A sequence $\{f_n : n \in \mathbb{N}\} \subset \mathbb{D}^*$ converges to $f \in \mathbb{D}^*$ in the weak*-topology if

$$\lim_{n\to\infty} f_n(x) = f(x), \qquad x \in \mathbb{D}.$$

Continuity: Let $\{f_n : n \in \mathbb{N}\} \subset \mathcal{F}$ be a weak*-convergent sequence with limit $f \in \mathbb{D}^*$. Since for any $f \in \mathcal{F}$, $|f(x)| \leq \sup_{f\in\mathcal{F}}|f(x)|$, $x \in \mathbb{D}$ and $\mathrm{E}[\sup_{f\in\mathcal{F}}|f(Z)|] \leq L\,\mathrm{E}[\|Z\|] < \infty$, by the dominated convergence theorem,

$$\lim_{n\to\infty} \mathrm{E}[f_n(Z)] = \mathrm{E}[f(Z)],$$

that is, $f \mapsto \mathrm{E}[f(Z)]$ is weak*-continuous on $\mathcal{F}$.

$\mathcal{F}_0$ is closed: Since $\mathbb{D}$ is separable, the weak*-topology is metrizable and therefore a set is closed if the limit of any weak*-convergent sequence in F lies in F. Let f_n be a convergent sequence in $\mathcal{F}$, i.e., for each $x \in \mathbb{D}$, we have $\lim_{n\to\infty} f_n(x) = f(x)$ for some $f \in \mathbb{D}^*$. Since $f_n(x) \leq \psi(x)$ for all x and all n, $f(x) \leq \psi(x)$ and therefore $\mathcal{F}$ is closed with respect to the weak*-topology. Moreover, by the same argument as above, one can show that $f \mapsto \mathrm{Var}(f(Z))$ is weak*-continuous. Thus, for any weak*-convergent sequence f_n with $\mathrm{Var}(f_n(Z)) = 0$ and limit f, we have $\mathrm{Var}(f(Z)) = 0$. Thus, $\mathcal{F}_0$ is closed in the weak*-topology.

Now, since $\mathcal{F}_0$ is a closed subset of $\{f \in \mathbb{D}^* : \|f\|^* \leq L\}$ which is compact in the weak*-topology, $\mathcal{F}_0$ itself is compact with respect to this topology. The claim follows. □

2.A.3 Lower bound for the variances

Proof of Lemma 67: The proof idea is based on the proof of Proposition 1 in Giessing (2023a).

Observe that, under the stated assumptions, $\mathrm{E}[Z] \geq \mathrm{E}[X_u] = 0$ and $\mathrm{E}[Z] < \infty$. We distinguish two cases. First, consider the case $\mathrm{E}[Z/\bar{\sigma}] < 1/2$. By Chebyshev's inequality,

$$\mathrm{Var}(Z/\bar{\sigma}) \geq \mathrm{P}(Z/\bar{\sigma} \geq 1)(1 - \mathrm{E}[Z/\bar{\sigma}])^2 \geq \frac{1}{4}\mathrm{P}(Z/\bar{\sigma} \geq 1) \geq \frac{1}{4}\sup_{u \in U} \mathrm{P}(X_u \geq \bar{\sigma}).$$

Since X is a Gaussian process

$$\sup_{u \in U} \mathrm{P}(X_u \geq \bar{\sigma}) = \sup_{u \in U}(1 - \Phi(\bar{\sigma}/\sigma_u)) = 1 - \Phi(1) \geq \frac{1}{8}.$$

Hence, if $\mathrm{E}[Z/\bar{\sigma}] \leq 1/2$,

$$\mathrm{Var}(Z) \geq \frac{\bar{\sigma}^2}{32}.$$

Next, consider the case $\mathrm{E}[Z/\bar{\sigma}] \geq 1/2$. By Cantelli's inequality,

$$\mathrm{P}(Z \geq \mathrm{E}[Z] + 1) \leq \frac{\mathrm{Var}(Z)}{\mathrm{Var}(Z) + 1}$$

or equivalently,

$$\mathrm{Var}(Z) \geq \frac{\mathrm{P}(Z \geq \mathrm{E}[Z] + 1)}{1 - \mathrm{P}(Z \geq \mathrm{E}[Z] + 1)}.$$

Now, by definition of Z as a supremum over the X_u,

$$\begin{aligned}\mathrm{P}(Z \geq \mathrm{E}[Z] + 1) \geq & \sup_{u \in U} \mathrm{P}(X_u \geq \mathrm{E}[Z] + 1) \\ = & \sup_{u \in U} 1 - \Phi\left(\frac{\mathrm{E}[Z] + 1}{\sigma_u}\right) = 1 - \Phi\left(\frac{\mathrm{E}[Z] + 1}{\bar{\sigma}}\right).\end{aligned}$$

Thus,

$$\frac{\mathrm{P}(Z \geq \mathrm{E}[Z] + 1)}{1 - \mathrm{P}(Z \geq \mathrm{E}[Z] + 1)} \geq \frac{1 - \Phi(\frac{\mathrm{E}[Z]+1}{\bar{\sigma}})}{\Phi(\frac{\mathrm{E}[Z]+1}{\bar{\sigma}})}.$$

Note that $t \leq \mathrm{erf}(t)\sqrt{1+t^2} \leq \sqrt{2}t$ for all $t \geq 0$ where $\mathrm{erf}(t) = \frac{2}{\sqrt{\pi}} \int_0^t \exp(-x^2)dx$. Moreover, since $\Phi(t) = \frac{1}{2}(1 + \mathrm{erf}(x/\sqrt{2}))$, we have for any $t \geq 0$,

$$\Phi(t) \leq \frac{1}{2}\left(1 + \frac{t}{\sqrt{1+t^2/2}}\right) \qquad \text{and} \qquad 1 - \Phi(t) \geq \frac{1}{2}\left(1 - \frac{t/\sqrt{2}}{\sqrt{1+t^2/2}}\right)$$

and therefore

$$\frac{1-\Phi(t)}{\Phi(t)} \geq \frac{\sqrt{1+t^2/2} - t/\sqrt{2}}{\sqrt{1+t^2/2}+t} = \frac{\sqrt{2+t^2}-t}{\sqrt{2+t^2}+\sqrt{2}t}.$$

This implies

$$\mathrm{Var}(Z) \geq \frac{\sqrt{2\bar{\sigma}^2 + (\mathrm{E}[Z]+1)^2} - (\mathrm{E}[Z]+1)}{\sqrt{2\bar{\sigma}^2 + (\mathrm{E}[Z]+1)^2} + \sqrt{2}(\mathrm{E}[Z]+1)}.$$

Consider the function $f(x) = (1-x)/(1+\sqrt{2}x)$ and note that the right-hand side equals $f((\mathrm{E}[Z]+1)/\sqrt{2\bar{\sigma}^2 + (\mathrm{E}[Z]+1)^2})$. It holds

$$(\mathrm{E}[Z]+1)/\sqrt{2\bar{\sigma}^2 + (\mathrm{E}[Z]+1)^2} = \frac{1}{\sqrt{1+1/E^2}}$$

where $E = (\mathrm{E}[Z]+1)/\sqrt{2}\bar{\sigma}$. Since $\mathrm{E}[Z]/(\sqrt{2}\bar{\sigma}) \geq 1/2$, $E \geq (1/2 + 1/\bar{\sigma})/\sqrt{2}$ and therefore

$$\sqrt{1+\frac{1}{E^2}} \leq \sqrt{1 + \frac{8\bar{\sigma}^2}{(\bar{\sigma}+2)^2}} = \sqrt{\frac{8\bar{\sigma}^2 + (\bar{\sigma}+2)^2}{(\bar{\sigma}+2)^2}} \leq \sqrt{\frac{8(\bar{\sigma}+2)^2 + (\bar{\sigma}+2)^2}{(\bar{\sigma}+2)^2}} = 3.$$

This implies

$$\frac{1 - \frac{1}{\sqrt{1+1/E^2}}}{1 + \sqrt{2}\frac{1}{\sqrt{1+1/E^2}}} = \frac{\sqrt{1+1/E^2}-1}{\sqrt{1+1/E^2}+\sqrt{2}} \geq \frac{1}{5}\left(\sqrt{1+\frac{1}{E^2}} - 1\right).$$

Note that the function $x \mapsto (\sqrt{1+1/x^2}-1)(1+x^2)$ is monotonically decreasing for $x > 0$ with limit $1/2$ as $|x| \to \infty$. Thus,

$$\sqrt{1+1/x^2} - 1 \geq \frac{1}{2(1+x^2)}$$

implying

$$\mathrm{Var}(Z) \geq \frac{1}{10}\frac{1}{1+E^2} = \frac{1}{5}\frac{\bar{\sigma}^2}{2\bar{\sigma}^2 + (\mathrm{E}[Z]+1)^2}$$

The claim follows. □

2.A.4 Regularity of the Gaussian process G

Lemma 49. *Let $E \subset \mathbb{D}$ be a separable closed linear subspace of a real Banach space $\mathbb{D}$ and let $\mathcal{F} = \{f \in \mathbb{D}^* | \forall x \in \mathbb{D} : f(x) \leq \psi(x)\}$ for a continuous and sublinear functional ψ. Then, there exists a countable subset F of $\mathcal{F}$ so that*

$$\psi(x) = \sup_{f\in\mathcal{F}} f(x) = \sup_{f\in F} f(x), \qquad \forall x \in E.$$

Proof. The first equality follows by our representation Lemma 44. For the second equality, let $\{x_n : n \in \mathbb{N}\}$ be a countable dense subset of E. By the construction in our representation result Lemma 44, for any n, there exists an $f_n \in \mathcal{F}$ such that $f_n(x_n) = \psi(x_n)$. Take $F = \{f_n : n \in \mathbb{N}\}$. Indeed, for any $x \in E$ and any $\varepsilon > 0$, we can find an n so that $\|x - x_n\| < \varepsilon$ and therefore

$$\psi(x) \leq \sup_{f\in\mathcal{F}} f(x_n) + \sup_{f\in\mathcal{F}} f(x - x_n) \leq f_n(x_n) + L\varepsilon = \sup_{m\in\mathbb{N}} f_m(x_n) + L\varepsilon,$$

where we have used that $\|f\|^* \leq L$ for all $f \in \mathcal{F}$. There is a sequence $(x_k) \subset F$ so that $\|x - x_k\| \leq 1/k$ and

$$\psi(x) \leq \sup_{m\in\mathbb{N}} f_m(x_k) + \frac{L}{k}$$

Now, since $\mathcal{F}$ is uniformly bounded,

$$\begin{aligned}\sup_{m\in\mathbb{N}} f_m(x_k) &\leq \sup_{m\in\mathbb{N}}\{f_m(x)\} + \sup_{m\in\mathbb{N}}\{f_m(x_k - x)\} \\ &\leq \sup_{m\in\mathbb{N}}\{f_m(x)\} + \sup_{m\in\mathbb{N}}\|f_m\|^*\|x - x_k\| = \sup_{m\in\mathbb{N}}\{f_m(x)\} + \frac{L}{k}.\end{aligned}$$

Further, since $F \subset \mathcal{F}$, $\sup_{f\in F} f(x) \leq \sup_{f\in\mathcal{F}} f(x)$ and therefore

$$\left|\psi(x) - \sup_{f\in F} f(x)\right| \leq 2\frac{L}{k}.$$

Taking $k \to \infty$ proves the claim. □

In the following lemma, we collect some regularity properties of Banach space valued Gaussian random vectors and for their corresponding Gaussian process on the dual space $\mathbb{D}^*$. These results follow easily from well established results in the literature on probability in Banach spaces, as e.g. in Ledoux and Talagrand (1991), and are included for completeness.

Lemma 50. *Let $\mathbb{D}$ be a real Banach space and let $\mathcal{F} = \{f \in \mathbb{D}^* | \forall x \in \mathbb{D} : f(x) \leq \psi(x)\}$ for a continuous and sublinear functional ψ. Further, let Z be a $\mathbb{D}$-valued tight Gaussian random vector. Denote by $G = \{f(Z) : f \in \mathcal{F}\}$. Then,*

(i) $\|Z\|$ has a finite unique median M.

(ii) The mean function $\mu(f) = \mathrm{E}[f(Z)]$ is bounded and Lipschitz continuous, that is, $|\mu(f) - \mu(g)| \leq \mathrm{E}[\|Z\|]\|f - g\|^*$, $f, g \in \mathbb{D}^*$.

(iii) All moments of $\|Z\|$ are finite. Moreover, the weak variance satisfies $\sigma^2 = \sup_{\|f\|^* \leq 1} \mathrm{Var}(f(Z)) < \infty$.

(iv) G has almost surely bounded sample paths.

(v) G has almost surely Lipschitz continuous sample paths.

(vi) There is a sequence $(f_n)_{n \in \mathbb{N}} \subset \mathcal{F}$ so that

$$\max_{1 \leq i \leq n} f_i(Z) \overset{a.s.}{\to} \sup_{f \in \mathcal{F}} f(Z).$$

Proof. (i) Since Z is tight, there exists a compact set $K \subset \mathbb{D}$ such that $\mathrm{P}(Z \notin K) < 1/2$. Compact sets are necessarily bounded and therefore, there exists a constant C such that $\mathrm{P}(\|Z\| \geq C) \leq \mathrm{P}(Z \notin K)$. By definition of a median M of $\|Z\|$, $\mathrm{P}(\|Z\| \geq M) \geq 1/2$ implying $\mathrm{P}(|Z\| \geq C) \leq \mathrm{P}(\|Z\| \geq M)$. In turn, this implies that $M \leq C$. Since M was arbitrary, it follows that any median is bounded. The uniqueness follows by Chapter 3.1 in Ledoux and Talagrand (1991).

(ii) The boundedness follows by $|\mu(f)| \leq \mathrm{E}[\sup_{f \in F} |f(Z)|] \leq L\,\mathrm{E}[\|Z\|]$, where we used that $\|f\|^* \leq L$ for all $f \in \mathcal{F}$. Existence of $\mathrm{E}[\|Z\|]$ follows by the triangle inequality and existence of $\mathrm{E}[\|Z - \mathrm{E}[Z]\|]$ as shown in Chapter 3.1 in Ledoux and Talagrand (1991). The Lipschitz continuity follows by definition of the norm on the dual space.

(iii) By the triangle inequality, we can centralize the moments of $\|Z\|$, i.e.,

$$\|\|Z\|\|_{L_p} \leq \|\|Z - \mathrm{E}[Z]\|\|_{L_p} + \|\mathrm{E}[Z]\|.$$

By tightness of Z and (ii), $Z - \mathrm{E}[Z]$ is tight, and its median is finite by the same arguments in (i). Therefore, Corollary 3.2 in Ledoux and Talagrand (1991) implies that there exist constants $0 < c_p \leq C_p < \infty$ such that $c_p m \leq \|\|Z - \mathrm{E}[Z]\|\|_{L_p} \leq C_p m$, where m denotes the median of $\|Z - \mathrm{E}[Z]\|$. Finiteness of the weak variance follows by the bound $\sigma^2 \leq 2m$ derived in Chapter 3.1 in Ledoux and Talagrand (1991).

(iv) It is sufficient to show that

$$\mathrm{P}(|\psi(Z)| = \infty) = 0.$$

It holds

$$\begin{aligned}\mathrm{P}(|\psi(Z)| = \infty) =& \mathrm{P}(\forall a : |\psi(Z)| > a) = \mathrm{P}(\forall n \in \mathbb{N} : |\psi(Z)| > a_n) \\ \leq& \mathrm{P}(\forall n \in \mathbb{N} : \exists i \geq n : |\psi(Z)| > a_i)\end{aligned}$$

where the latter statement holds for any sequence $a_n \to \infty$ as $n \to \infty$. The latter event can be written as $\bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} A_i$, where $A_i = \{|\psi(Z)| > a_i\}$. Thus, the Borel-Cantelli Lemma applies. Since $\mathcal{F} \subset \{f \in \mathbb{D}^* : \|f\|^* \leq L\} =: B_L$, it holds $|\psi(Z)| \leq \sup_{f \in \mathcal{F}} |f(Z)| \leq \sup_{f \in B_L} |f(Z)| = L\|Z\|$. As argued above, we further have $\mathrm{E}[\|Z\|] < \infty$ and hence by Markov's inequality for $a_n = 2^n \mathrm{E}[\|Z\|]L$, we have $\mathrm{P}(L\|Z\| > a_n) \leq 2^{-n}$ for all $n \in \mathbb{N}$. Together with $\mathrm{P}(|\psi(Z)| > t) \leq \mathrm{P}(L\|Z\| > t)$, $t \in \mathbb{R}$, this implies $\sum_{n \in \mathbb{N}} \mathrm{P}(|\psi(Z)| > a_n) < \infty$ and hence by Borel-Cantelli and the above inequalities $\mathrm{P}(|\psi(Z)| = \infty) = 0$ implying $|\psi(Z)| < \infty$ a.s. and $\|G\|_{\mathcal{F}} < \infty$ a.s. where $\|G\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |f(Z)|$.

(v) We need to show that

$$\mathrm{P}\left(\exists L < \infty : \forall f, g \in \mathcal{F} : |f(Z) - g(Z)| \leq L\|f - g\|\right) = 1.$$

By definition of the dual norm, $|f(Z) - g(Z)| \leq \|Z\|\|f - g\|^*$ and therefore, on $\{\|Z\| < \infty\}$, we can choose $L = \|Z\|$. Hence,

$$\mathrm{P}\left(\exists L < \infty : \forall f, g \in \mathcal{F} : |f(Z) - g(Z)| \leq L\|f - g\|\right) \geq \mathrm{P}(\|Z\| < \infty) = 1,$$

where the last equality follows by (iv).

(vi) By tightness of Z, for any $\eta > 0$, there is a compact set $K_\eta \subset \mathbb{D}$ so that $\mathrm{P}(Z \notin K_\eta) < \eta$. Since K_η is compact, for any $\varepsilon > 0$, there are $x_1, \ldots, x_N \in K_\eta$, with $N < \infty$, so that for any $x \in K_\eta$, there is some j so that $\|x - x_j\| \leq \varepsilon$. On the event $\{Z \in K_\eta\}$,

$$\sup_{f \in \mathcal{F}} f(Z) \leq \sup_{f \in \mathcal{F}} f(x_j) + \sup_{f \in \mathcal{F}} f(Z - x_j).$$

Since $\mathcal{F}$ is uniformly bounded, i.e., $\|f\|^* \leq L$ for all $f \in \mathcal{F}$ and some $L < \infty$,

$$\sup_{f \in \mathcal{F}} f(Z - x_j) \leq \|f\|^* \|Z - x_j\| \leq L\varepsilon$$

Moreover, by the same arguments as in the representation Lemma 44, there exists for each x_j, an $f_j \in \mathcal{F}$ such that $f_j(x_j) = \psi(x_j) = \sup_{f \in \mathcal{F}} f(x_j)$ and therefore

$$\sup_{f \in \mathcal{F}} f(x_j) = \max_{1 \leq i \leq N} f_i(x_j).$$

Moreover, by the same arguments as above,

$$\max_{1 \leq i \leq N} f_i(x_j) \leq \max_{1 \leq i \leq N} f_i(Z) + L\varepsilon$$

Together with $\sup_{f \in \mathcal{F}} f(Z) \geq \max_{1 \leq i \leq N} f_i(Z)$, this implies

$$\left| \sup_{f \in \mathcal{F}} f(Z) - \max_{1 \leq i \leq N} f_i(Z) \right| \leq 2L\varepsilon.$$

Thus,

$$\begin{aligned} &\mathrm{P}\left(\left| \sup_{f \in \mathcal{F}} f(Z) - \max_{1 \leq i \leq N} f_i(Z) \right| > 2L\varepsilon \right) \\ \leq &\mathrm{P}\left(\left| \sup_{f \in \mathcal{F}} f(Z) - \max_{1 \leq i \leq N} f_i(Z) \right| > 2L\varepsilon, Z \in K \right) + \mathrm{P}(Z \notin K) \leq \eta. \end{aligned}$$

Now, for the construction of the sequence (f_n), let $\varepsilon_n = \eta_n = 2^{-n}$, $n \in \mathbb{N}$ and repeat for each n the construction from above to obtain $x_1^{(n)}, \ldots, x_{N_n}^{(n)}$ and corresponding

$f_1^{(n)}, \ldots, f_{N_n}^{(n)}$ satisfying, for all $n \in \mathbb{N}$,

$$\mathrm{P}\left(\left|\sup_{f\in\mathcal{F}} f(Z) - \max_{1\leq i\leq N_n} f_i^{(n)}(Z)\right| > 2^{-n+1}L\right) \leq 2^{-n}.$$

Construct the sequence (f_n) by stacking the finite sequences $f_1^{(n)}, \ldots, f_{N_n}^{(n)}$ on each other. The resulting sequence (f_n) shares the same approximation properties with the finite sequences. Indeed, it still holds $\max_{1\leq i\leq n} f_i(Z) \leq \sup_{f\in\mathcal{F}} f(Z)$ as well as for some N_n, $\sup_{f\in\mathcal{F}} f(Z) \leq \max_{1\leq i\leq N_n} f_i(Z) + 2^{-n+1}L$. Therefore, for all n,

$$\mathrm{P}\left(\left|\sup_{f\in\mathcal{F}} f(Z) - \max_{1\leq i\leq N_n} f_i(Z)\right| > 2^{-n+1}L\right) \leq 2^{-n}.$$

This implies together with the Borel Cantelli Lemma,

$$\max_{1\leq i\leq n} f_i(Z) \overset{a.s.}{\to} \sup_{f\in\mathcal{F}} f(Z)$$

and the claim follows. □

2.B Proofs for applications

Our proofs for the Applications are based on a slight refinement of Le Cam's inequality as presented in Lemma 43. This refinement allows us to circumvent the potential point mass in the approximate distribution by restricting the approximation of the cdfs to a subset of $\mathbb{R}$ only.

Lemma 51. *For $X, Z \in \mathbb{R}$ arbitrary random variables, $\tau \in \mathbb{R}$ and $\lambda > 0$,*

$$\sup_{t\geq\tau} |\mathrm{P}(X \leq t) - \mathrm{P}(Z \leq t)| \leq \mathrm{P}(|X - Z| > \lambda) + \zeta_\lambda(X) \wedge \zeta_\lambda(Z),$$

where $\zeta_\lambda(V) = \sup_{t\geq\tau} \mathrm{P}(t \leq V \leq t + \lambda)$ for real-valued $V \in \mathbb{R}$.

Proof of Lemma 51: For any $t \in \mathbb{R}$ and any $\lambda > 0$,

$$\begin{aligned}
\mathrm{P}(X \leq t) \leq& \mathrm{P}(X \leq t, |X - Z| \leq \lambda) + \mathrm{P}(|X - Z| > \lambda) \\
\leq& \mathrm{P}(Z \leq t + \lambda) + \mathrm{P}(|X - Z| > \lambda) \\
\leq& \mathrm{P}(Z \leq t) + \mathrm{P}(t \leq Z \leq t + \lambda) + \mathrm{P}(|X - Z| > \lambda).
\end{aligned}$$

Moreover, it holds

$$\begin{aligned}\mathrm{P}(Z \le t) \le& \mathrm{P}(Z \le t-\lambda) + \mathrm{P}(t-\lambda \le Z \le t)\\ \le& \mathrm{P}(Z \le t, |X-Z| \le \lambda) + \mathrm{P}(|X-Z| > \lambda) + \mathrm{P}(t-\lambda \le Z \le t)\\ \le& \mathrm{P}(X \le t) + \mathrm{P}(|X-Z| > \lambda) + \mathrm{P}(t-\lambda \le Z \le t).\end{aligned}$$

Thus, we have shown

$$\sup_{t\ge\tau} |\mathrm{P}(X \le t) - \mathrm{P}(Z \le t)| \le \mathrm{P}(|X-Z| > \lambda) + \sup_{t\ge\tau} \mathrm{P}(t \le Z \le t+\lambda).$$

By changing the roles of X and Z, we also obtain

$$\sup_{t\ge\tau} |\mathrm{P}(X \le t) - \mathrm{P}(Z \le t)| \le \mathrm{P}(|X-Z| > \lambda) + \sup_{t\ge\tau} \mathrm{P}(t \le X \le t+\lambda).$$

The claim follows. □

2.B.1 Proof of the HD-CLT

Proof of Proposition 2: By the representation result $\psi(x) = \sup_{f\in\mathcal{F}} f(x)$ for some $\mathcal{F} \subset (\mathbb{R}^k)^*$. Let q satisfy $\frac{1}{p} + \frac{1}{q} = 1$. In the following, we identify $(\mathbb{R}^k)^*$ as $\mathbb{R}^k$ endowed with the ℓ_q-norm. By ℓ_p Lipschitz continuity of ψ and the representation result, $\mathcal{F} \subset \{f \in \mathbb{R}^k : \|f\|_q \le L\}$.

Further, let $T_n = \sum_{i=1}^n (X_i - \mathrm{E}[X_i]) = S_n - \mu_n$, where $\mu_n = \sum_{i=1}^n \mathrm{E}[X_i]$. By Lemma 38 in Belloni et al. (2019b), for any $\delta > 0$, there exists $Z_n \sim \mathcal{N}(\mu_n, \Sigma_n)$ for $\Sigma_n = \mathrm{E}[T_n T_n^\intercal]$ so that

$$\mathrm{P}(\|S_n - Z_n\|_p > 3\delta) \le \min_{t\ge 0}\left\{\frac{\beta_p}{\delta^3}t^2 + 2\mathrm{P}(\|Z\|_p > t)\right\},$$

where $Z \sim \mathcal{N}(0, I)$. By Lipschitz continuity of ψ,

$$\mathrm{P}\Big(|\psi(S_n) - \psi(Z_n)| > 3\delta\Big) \le \min_{t\ge 0}\left\{\frac{L^3\beta_p}{\delta^3}t^2 + 2\mathrm{P}(\|Z\|_p > t)\right\}.$$

By Lemma 52 and Markov's inequality,

$$\mathrm{P}(\|Z\|_p > t) \le \mathrm{P}\left(\exp\left(\frac{\|Z\|_p^2}{4\sigma^2}\right) > \left(\frac{t^2}{4\sigma^2}\right)\right) \le C\left(\frac{M}{\sigma} + 1\right)\exp\left(\frac{M^2}{4\sigma^2}\right)\exp\left(-\frac{t^2}{4\sigma^2}\right).$$

The right-hand side is monotonely increasing in M and therefore M can be replaced by any upper bound. In particular, by Lemma A.4 in Cattaneo et al. (2022), $\mathrm{E}[\|Z\|] \leq \phi_p(k)$, thus $M \leq 2\phi_p(k)$. Further, since $Z \sim \mathcal{N}(0, I_k)$, for $q \geq 0$ such that $\frac{1}{p} + \frac{1}{q} = 1$,

$$\sigma^2 = \sup_{\|y\|_q=1} \mathrm{E}[(y^\intercal Z)^2] = \sup_{\|y\|_q=1} \|y\|_2^2 = c_p(k)^2,$$

where $c_p(k) = 1$ if $p \in [2, \infty]$ and $c_p(k) = k^{\frac{1}{2}-\frac{p-1}{p}} = k^{\frac{2-p}{2p}}$ if $p \in [1, 2)$. This implies

$$\mathrm{P}(\|Z\|_p > t) \leq C\left(\frac{2\phi_p(k)}{c_p(k)} + 1\right) \exp\left(\frac{\phi_p(k)^2}{c_p(k)^2}\right) \exp\left(-\frac{t^2}{4c_p(k)^2}\right)$$

Optimizing over t yields for $t^2 = C\phi_p(k)^2 \log(\phi_p(k)/c_p(k)\delta^3/L^3/\beta_p)$

$$\mathrm{P}\Big(|\psi(S_n) - \psi(Z_n)| > 3\delta\Big) \leq C\frac{L^3\beta_p}{\delta^3}\left\{\phi_p(k)^2 \log\left(\frac{\phi_p(k)\delta^3}{c_p(k)L^3\beta_p}\right) + 1\right\}.$$

Combining Le Cam's Lemma and our anti-concentration bound in Theorem 14 implies

$$\begin{aligned}
&\sup_{t\in\mathbb{R}} |\mathrm{P}(\psi(S_n) \leq t) - \mathrm{P}(\psi(Z_n) \leq t)| \\
&\leq \min_{\delta>0}\left\{C\frac{L^3\beta_p}{\delta^3}\left\{\phi_p(k)^2 \log\left(\frac{\phi_p(k)\delta^3}{c_p(k)L^3\beta_p}\right) + 1\right\} + \frac{\delta\sqrt{12}}{\sqrt{\mathrm{Var}(\psi(Z)) + \delta^2(12)}}\right\} \\
&\qquad + \mathrm{P}(\psi(Z_n) = \bar{\mu}) \\
&\leq \min_{\delta>0}\left\{C\frac{L^3\beta_p\phi_p(k)^2}{\delta^3} \log\left(\frac{\phi_p(k)\delta^3}{c_p(k)L^3\beta_p} \vee e\right) + \frac{\delta\sqrt{12}}{\sqrt{\mathrm{Var}(\psi(Z))}}\right\} + \mathrm{P}(\psi(Z_n) = \bar{\mu}).
\end{aligned}$$

Choose

$$\delta = \left(L^3\beta_p\phi_p(k)^2\sqrt{\mathrm{Var}(\psi(Z_n))}\right)^{1/4}$$

and let

$$\Delta_n = \left(\frac{L^3\beta_p\phi_p(k)^2}{(\mathrm{Var}(\psi(Z_n)))^{3/2}}\right)^{1/4}.$$

Then,

$$\frac{L^3\beta_p\phi_p(k)^2}{\delta^3}\log\left(\frac{\phi_p(k)\delta^3}{c_p(k)L^3\beta_p}\vee e\right)=\Delta_n\log\left(\frac{\phi_p(k)^3}{c_p(k)\Delta_n}\vee e\right)$$

and therefore

$$\sup_{t\in\mathbb{R}}|\mathrm{P}(\psi(S_n)\le t)-\mathrm{P}(\psi(Z_n)\le t)|\le C\Delta_n\log\left(\frac{\phi_p(k)^3}{c_p(k)\Delta_n}\vee e\right)+\mathrm{P}(\psi(Z_n)=\bar{\mu}).$$

If, furthermore,

$$\frac{\phi_p(k)^3}{c_p(k)\Delta_n}\to 0,$$

then the upper bound can be simplified to

$$\sup_{t\in\mathbb{R}}|\mathrm{P}(\psi(S_n)\le t)-\mathrm{P}(\psi(Z_n)\le t)|\le C\left(\frac{L^3\beta_p\phi_p(k)^2}{(\mathrm{Var}(\psi(Z_n)))^{3/2}}\right)^{1/4}+\mathrm{P}(\psi(Z_n)=\bar{\mu})$$

for n sufficiently large. □

Proof of Corollary 9: By the triangle inequality,

$$\begin{aligned}|\sqrt{n}\psi(\hat{\theta}_n-\theta_n)-\psi(Z_n)|\le&|\sqrt{n}\psi(\hat{\theta}_n-\theta_n)-\psi(S_n)|\\&+|\psi(S_n)-\psi(Z_n)|\end{aligned}$$

If $A\ge a$ and $A\le B+C$, then $B\ge a-\varepsilon$ or $C\ge\varepsilon$. As suppose that both $B<a-\varepsilon$ and $C<\varepsilon$, then $A<a$ and by taking complements the claim follows. Thus,

$$\begin{aligned}&\mathrm{P}\big(|\sqrt{n}\psi(\hat{\theta}_n-\theta_n)-\psi(Z_n)|\ge 2\sqrt{\mathrm{Var}(\psi(Z_n))}\varepsilon\big)\\\le&\mathrm{P}\big(|\psi(S_n)-\psi(Z_n)|\ge\sqrt{\mathrm{Var}(\psi(Z_n))}\varepsilon\big)\\&+\mathrm{P}\big(|\sqrt{n}\psi(\hat{\theta}_n-\theta_n)-\psi(S_n)|\ge\sqrt{\mathrm{Var}(\psi(Z_n))}\varepsilon\big)\end{aligned}\tag{2.7}$$

By $\delta_n/\sqrt{\mathrm{Var}(\psi(Z_n))}=O(1)$, $\sqrt{\mathrm{Var}(\psi(Z_n))}\ge c\delta_n$ eventually for some constant c, and therefore

$$\begin{aligned}&\mathrm{P}\big(|\sqrt{n}\psi(\hat{\theta}_n-\theta_n)-\psi(S_n)|\ge\sqrt{\mathrm{Var}(\psi(Z_n))}\varepsilon\big)\\\le&\mathrm{P}\big(|\sqrt{n}\psi(\hat{\theta}_n-\theta_n)-\psi(S_n)|\ge c\delta_n\varepsilon\big)\to 0,\qquad\text{as } n\to\infty.\end{aligned}$$

Regarding (2.7), note that

$$\delta = \left(L^3 \beta_p \sqrt{\mathrm{Var}(\psi(Z_n))} \left\{ \phi_p(k)^2 \log\left(\frac{\phi_p(k)}{c_p(k) L^3 \beta_p} \right) + 1 \right\} \right)^{1/4}$$

in the proof of Proposition 2. Moreover, by assumption, $\delta / \sqrt{\mathrm{Var}(\psi(Z))} \to 0$ and therefore $\delta \leq \sqrt{\mathrm{Var}(\psi(Z))}\varepsilon$ eventually. Thus,

$$\begin{aligned} &\mathrm{P}\big(|\psi(S_n) - \psi(Z_n)| \geq \sqrt{\mathrm{Var}(\psi(Z_n))}\varepsilon\big) \\ \leq &\mathrm{P}\big(|\psi(S_n) - \psi(Z_n)| \geq \delta\big) \to 0. \end{aligned}$$

Finally, by applying Le Cam's Lemma 51 with $\eta = \sqrt{\mathrm{Var}(\psi(Z_n))}\varepsilon$

$$\sup_{t \in \mathbb{R}} |\mathrm{P}(\sqrt{n h_n^d}\psi(\hat{\theta}_n - \theta_n) \leq t) - \mathrm{P}(\psi(Z_n) \leq t)| \leq \varepsilon + o(1),$$

proving the claim. □

The following lemma is essentially a part of Corollary 3.2 in Ledoux and Talagrand (1991) but with an explicit upper bound.

Lemma 52. *Let $Z \sim \mathcal{N}(0, I_k)$, M denote the median of $\|Z\|_p$ and $\sigma^2 = \sup_{\|y\|_q \leq 1} \|y\|_2^2$. Then, there is some absolute constant C, so that*

$$\mathrm{E}\left[\exp\left(\frac{\|Z\|_p^2}{4\sigma^2} \right) \right] \leq C \left(\frac{M}{\sigma} + 1 \right) \exp\left(\frac{M^2}{4\sigma^2} \right).$$

Proof. Let $a^2 = 1/(4\sigma^2)$. It holds,

$$\begin{aligned} \mathrm{E}[\exp(a^2 \|Z\|_p^2)] &= \int_0^\infty \mathrm{P}(\exp(a^2\|Z\|_p^2) \geq t) dt = \int_0^\infty \mathrm{P}\left(\|Z\|_p \geq \frac{1}{a}\sqrt{\log t} \right) dt \\ &= \int_0^{\exp(a^2M^2)} \mathrm{P}\left(\|Z\|_p \geq \frac{1}{a}\sqrt{\log t} \right) dt \\ &\quad + \int_{\exp(a^2M^2)}^\infty \mathrm{P}\left(\|Z\|_p \geq \frac{1}{a}\sqrt{\log t} \right) dt \\ &\leq \exp(a^2M^2) + \int_{\exp(a^2M^2)}^\infty \mathrm{P}\left(\|Z\|_p \geq \frac{1}{a}\sqrt{\log t} \right) dt. \end{aligned}$$

Substitute $M + \sigma u = 1/a\sqrt{\log t}$, i.e., $u = \frac{1}{\sigma}(-M + 1/a\sqrt{\log t})$, with derivative

$du/dt = \frac{1}{\sigma a}\frac{1}{t\sqrt{\log t}}$ i.e.,

$$\begin{aligned}
&\int_M^\infty \mathrm{P}\left(\|Z\|_p \geq \frac{1}{a}\sqrt{\log t}\right)dt \\
&= \int \sigma a t\sqrt{\log t}\mathrm{P}(\|Z\|_p \geq M + \sigma u)du \\
&= \int \sigma a^2(M+\sigma u)\exp(a^2(M+\sigma u)^2)\mathrm{P}(\|Z\|_p \geq M + \sigma u)du \\
&\leq \frac{1}{2}\int \sigma a^2(M+\sigma u)\exp(a^2(M+\sigma u)^2)\exp\left(-\frac{u^2}{2}\right)du
\end{aligned}$$

where we used that by Borel's concentration inequality,

$$\mathrm{P}(\|Z\|_p > t) \leq \frac{1}{2}\exp\left(-\frac{(t-M)^2}{2\sigma^2}\right).$$

Next, consider a quadratic expansion

$$a^2(M+\sigma u)^2 - \frac{1}{2}u^2 = a^2M^2 + (a^2\sigma^2 - 0.5)u^2 + 2a^2M\sigma u$$

take $a^2 = 1/(4\sigma^2)$, then this reduces to

$$\begin{aligned}
a^2M^2 + (a^2\sigma^2 - 0.5)u^2 + 2a^2M\sigma u &= \frac{M^2}{4\sigma^2} - 0.25u^2 + 2\frac{M}{2\sigma}\frac{u}{2} - \frac{M^2}{4\sigma^2} + \frac{M^2}{4\sigma^2} \\
&= \frac{M^2}{2\sigma^2} - (u/2 - M/(2\sigma))^2
\end{aligned}$$

and therefore

$$\begin{aligned}
&\int \sigma\frac{1}{4\sigma^2}(M+\sigma u)\exp\left(\frac{M^2}{2\sigma^2}\right)\exp\left(-\frac{(u-M/\sigma)^2}{4}\right)du \\
=&\frac{M}{4\sigma}\exp(M^2/4\sigma^2)\int \exp(-(u-M/\sigma)^2/4)du \\
&+ \frac{1}{4}\exp(M^2/4\sigma^2)\int u\exp(-(u-M/\sigma)^2/4)du
\end{aligned}$$

The lower boundary of integration is $\exp(a^2M^2) = \exp(M^2/4\sigma^2)$

$$\int_{\exp(M^2/4\sigma^2)}^\infty \exp\left(-\frac{(u-M/\sigma)^2}{4}\right)du \leq 2\sqrt{\pi}$$

and similarly

$$\int u \exp(-(u - M/\sigma)^2/4) du \leq C(M/\sigma + 1)$$

So the overall upper bound for this second part is of the form

$$C\left(\frac{M}{\sigma} + 1\right) \exp\left(\frac{M^2}{4\sigma^2}\right)$$

Since the first part is also of this form, we have

$$\mathrm{E}\left[\exp\left(\frac{\|Z\|_p^2}{4\sigma^2}\right)\right] \leq C\left(\frac{M}{\sigma} + 1\right) \exp\left(\frac{M^2}{4\sigma^2}\right).$$

This proves the claim. □

Lemma 53. *Let $Z \sim \mathcal{N}(0, \Sigma)$. The covariance matrix Σ has ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0$ and associated unit eigenvectors $e_1, \ldots, e_k$. Then, there is some absolute constant C, so that, for $k \geq 2$,*

$$\mathrm{E}\left[\max_{1 \leq j \leq k} |e_j Z|^3\right] \leq C \operatorname{tr}(\Sigma) \sqrt{\lambda_1 \log k}.$$

Moreover, for any Vector $X \in \mathbb{R}^k$,

$$\mathrm{E}[\|X\|_2^3] \leq (\operatorname{tr}(\Sigma))^{3/2} \mathrm{E}\big[\max_{1 \leq j \leq d} |e_j^\intercal X|^3\big].$$

Proof. By the Cauchy-Schwarz inequality,

$$\mathrm{E}\left[\max_{1 \leq j \leq k} |e_j^\intercal Z|^3\right] \leq \mathrm{E}\left[\|Z\|_2^2 \max_{1 \leq j \leq k} |e_j^\intercal g_i|\right] \leq \sqrt{\mathrm{E}[\|Z\|_2^4]} \sqrt{\mathrm{E}\big[\max_{1 \leq j \leq k} |e_j^\intercal Z|^2\big]}.$$

By the rotational symmetry of the Euclidean distance, and since the $e_j^\intercal Z$ are mutually independent,

$$\begin{aligned}
\mathrm{E}[\|Z\|_2^4] &= \mathrm{E}\left[\left(\sum_{j=1}^k |e_j^\intercal Z|^2\right)^2\right] = \sum_{i=1}^k \sum_{j=1}^k \mathrm{E}[|e_i^\intercal Z|^2 |e_j^\intercal Z|^2] \\
&= \sum_{j=1}^k \mathrm{E}[|e_j^\intercal Z|^4] + \sum_{i=1}^k \sum_{j \neq i} \mathrm{E}[|e_i^\intercal Z|^2] \, \mathrm{E}[|e_j^\intercal Z|^2]
\end{aligned}$$

$$= \sum_{j=1}^{k} 3\lambda_j^2 + \sum_{i=1}^{k} \sum_{j \neq i} \lambda_i \lambda_j \leq 3 \sum_{i=1}^{k} \sum_{j=1}^{k} \lambda_i \lambda_j = 3(\mathrm{tr}(\Sigma))^2.$$

By Jensen's inequality,

$$\begin{aligned} \mathrm{E}\Big[\max_{1 \leq j \leq k} |e_j^\intercal Z|^2\Big] \leq & 4\lambda_1 \, \mathrm{E}\Big[\max_{1 \leq j \leq k} \frac{|e_j^\intercal Z|^2}{4\lambda_j}\Big] \leq 4\lambda_1 \log\Big(\mathrm{E}\Big[\max_{1 \leq j \leq k} \exp\Big(\frac{|e_j^\intercal Z|^2}{4\lambda_j}\Big)\Big]\Big) \\ \leq & 4\lambda_1 \log\Big(\sum_{j=1}^{k} \mathrm{E}\Big[\exp\Big(\frac{|e_j^\intercal Z|^2}{4\lambda_j}\Big)\Big]\Big) \leq 4\lambda_1 \log\big(\sqrt{2}k\big), \end{aligned}$$

where we used in the last inequality that for standard normal random variables X, $\mathrm{E}[\exp(X^2/4)] \leq \sqrt{2}$. The first part of the claim follows by combining the above bounds.

For the second part, note that by rotational symmetry of the Euclidean distance, we can bound

$$\|X_i\|_2^2 = \sum_{j=1}^{k} \lambda_j (e_j^\intercal X_i)^2 \leq \mathrm{tr}(\Sigma) \max_{1 \leq j \leq k} (e_j^\intercal X_i)^2$$

implying

$$\mathrm{E}[\|X_i\|_2^3] \leq (\mathrm{tr}(\Sigma))^{3/2} \, \mathrm{E}\big[\max_{1 \leq j \leq d} |e_j^\intercal X_i|^3\big].$$

□

2.B.2 Proofs for kernel-type estimators

Lemma 54. *Under assumptions (B1)-(B5), Z_n in Propositions 3 ans 4 has a version with values in $\mathcal{C}(\mathcal{I} \times \mathcal{G})$. In particular, Z_n can be realized as the canonical process on $(\mathcal{C}(\mathcal{G} \times \mathcal{I}), \mathcal{B}(\mathcal{C}(\mathcal{G} \times \mathcal{I})), \mathrm{P}_{Z_n})$, where P_{Z_n} denotes the distribution of Z_n.*

Proof. By Dudley's inequality (Theorem 11.17 in Ledoux and Talagrand (1991)),

$$\mathrm{E}\Big[\sup_{(y,x),(\tilde{y},\tilde{x}) \in \mathcal{G} \times \mathcal{I}} |Z_n(y,x) - Z_n(\tilde{y},\tilde{x})|\Big] \leq C \int_0^\infty \sqrt{\log N(\varepsilon, \mathcal{G} \times \mathcal{I}, L_2(P))} d\varepsilon.$$

Since $\mathcal{G} \times \mathcal{I}$ is a compact set of a finite dimensional Banach space, the entropy integral on the right-hand side is finite and therefore Z_n has almost surely uniformly continuous sample paths by the same arguments as in Lemma 50. Thus, there is

a set of measure zero N such that outside of this set, every sample path is uniformly continuous. If we change all sample paths in N to be just constantly 0 (or any continuous function), then the resulting process $\tilde{Z}_n$ coincides with Z_n for each $(y,x) \in \mathcal{G} \times \mathcal{I}$ $P-a.s.$, that is $\tilde{Z}_{n,y,x}(\omega) = Z_{n,y,x}(\omega)$ for all $\omega \notin N$. Furthermore, the existence of a version with continuous sample paths implies that one can realize Z_n as the canonical process on $(\mathcal{C}(\mathcal{G}\times\mathcal{I}), \mathcal{B}(\mathcal{C}(\mathcal{G}\times\mathcal{I})), \mathrm{P}_{Z_n})$, where P_{Z_n} denotes the distribution of Z_n. □

Proof of Proposition 3: In order to prove the Kolmogorov bound, we will use Le Cam's Lemma 51. In the first step, we will construct a coupled Gaussian process Z_n using Corollary 2.2 in Chernozhukov et al. (2010) in order to bound $\mathrm{P}(|\sqrt{nh_n^d}\psi(S_n) - \psi(Z_n)| > \eta)$. Thereafter, we will use our anti-concentration bound in Theorem 14 to upper bound the Kolmogorov distance.

Regarding the coupling construction, note that for any $f_n \in \mathcal{F}_n$ and any $k = 2, 3$,

$$\mathrm{E}[|f_n(Y,X) - \mathrm{E}[f_n(Y,X)]|^k] \leq C h_n^d \Sigma_n(f_n, f_n) \left(\frac{L\|g\|_\infty \|k\|_\infty}{\underline{p}} \right)^{k-2},$$

where Σ_n denotes the covariance function of $h_n^{-d/2}\mathbb{G}_n$. Let Γ_n denote the covariance function of the kernel estimator

$$\Gamma_n(x,y,\tilde{x},\tilde{y}) = \mathrm{Cov}(p(x)^{-1}g(Y,y)k(h_n^{-1}(X-x)), p(\tilde{x})^{-1}g(Y,\tilde{y})k(h_n^{-1}(X-\tilde{x})))$$

Since $f \in \mathbb{D}^*$, there is a regular Borel measure μ and by Fubini's Theorem,

$$\Sigma_n(f,f) = \iint \Gamma_n(x,y,\tilde{x},\tilde{y}) \mu(d(x,y)) \mu(d(\tilde{x},\tilde{y})).$$

Since $\sup_{(x,y,\tilde{x},\tilde{y})} \Gamma_n(x,y) = O(1)$, we have $\sigma_n^2 := \sup_{f\in\mathcal{F}_n} h_n^d \Sigma_n(f,f) = O(h_n^d)$. Further, let

$$b = \frac{L\|g\|_\infty \|k\|_\infty}{\underline{p}}.$$

Then, b is a valid envelope function of $\mathcal{F}_n$ and

$$\sup_{f\in\mathcal{F}_n} E[|f(Y,X)|^k] \leq \sigma_n^2 b^{k-2}, \qquad k = 2, 3.$$

Since $\mathcal{F}_n$ is a VC type class uniformly over n, A and v can be chosen independent

of n. Now, Corollary 2.2 in Chernozhukov et al. (2014a) applies and implies that for any $\gamma \in (0,1)$, there exists a sequence of random variables $W_n(\gamma)$ such that $W_n \stackrel{d}{=} \psi(Z_n)$ and

$$\begin{aligned}&\mathrm{P}\bigg(|\psi(S_n) - W_n| > C\bigg(\frac{\log n}{\gamma^{1/2}(nh_n^d)^{1/2}} + \frac{\log^{3/4} n}{\gamma^{1/2}(nh_n^d)^{1/4}} + \frac{\log^{2/3} n}{\gamma^{1/3}(nh_n^d)^{1/6}}\bigg)\bigg)\\ \leq& C\bigg(\gamma + \frac{\log n}{n}\bigg).\end{aligned}$$

Here C is a constant that only depends on $\|g\|_\infty$, $\|k\|_\infty$ and $\underline{p}$.

Note that

$$\begin{aligned}\frac{1}{\sqrt{\gamma}}\sqrt{\frac{\log^2 n}{nh_n^d}} \leq& \frac{1}{\gamma^{1/3}}\bigg(\frac{\log^4 n}{nh_n^d}\bigg)^{1/6}\\ \gamma^{1/6} = \gamma^{1/2-1/3} \geq& \bigg(\frac{\log^6 n}{(nh_n^d)^3}\frac{nh_n^d}{\log^4 n}\bigg)^{1/6} = \bigg(\frac{\log n}{nh_n^d}\bigg)^{1/3}\\ \gamma \geq& \bigg(\frac{\log n}{nh_n^d}\bigg)^2\end{aligned}$$

as well as

$$\begin{aligned}\frac{1}{\gamma^{1/2}}\bigg(\frac{\log^3 n}{nh_n^d}\bigg)^{1/4} \leq& \frac{1}{\gamma^{1/3}}\bigg(\frac{\log^4 n}{nh_n^d}\bigg)^{1/6}\\ \gamma^{1/6} \geq& \bigg(\frac{\log^9 n}{(nh_n^d)^3}\frac{(nh_n^d)^2}{\log^8 n}\bigg)^{1/12}\\ \gamma \geq& \sqrt{\frac{\log n}{nh_n^d}}.\end{aligned}$$

Thus, if $\gamma \geq \sqrt{\log n/(nh_n^d)}$, then

$$\mathrm{P}\bigg(|\sqrt{nh_n^d}\psi(S_n - \mathrm{E}[S_n]) - W_n| > C\frac{\log^{2/3} n}{\gamma^{1/3}(nh_n^d)^{1/6}}\bigg) \leq C\bigg(\gamma + \frac{\log n}{n}\bigg).$$

Take $\eta = C\frac{\log^{2/3} n}{\gamma^{1/3}(nh_n^d)^{1/6}}$, i.e.,

$$\gamma = \sqrt{\bigg(\frac{C}{\eta}\bigg)^{1/3}\sqrt{\frac{\log^4 n}{nh_n^d}}}$$

implying

$$\mathrm{P}(|\sqrt{nh_n^d}\psi(S_n - \mathrm{E}[S_n]) - W_n| > \eta) \leq C\left(\frac{1}{\eta^{1/3}}\sqrt{\frac{\log^4 n}{nh_n^d}} + \frac{\log n}{n}\right).$$

Together with Le Cam's Lemma 51 and the anti-concentration bound in Theorem 14, this implies[7]

$$\begin{aligned}
&\sup_{t\in\mathbb{R}}|\mathrm{P}(\sqrt{nh_n^d}\psi(S_n - \mathrm{E}[S_n]) \leq t) - \mathrm{P}(W_n \leq t)| \\
\leq &\inf_{\eta>0}\left\{C\left(\frac{1}{\eta^{1/3}}\sqrt{\frac{\log^4 n}{nh_n^d}} + \frac{\log n}{n} + \frac{\eta}{\sqrt{\mathrm{Var}(\psi(Z_n))}}\right)\right\} \\
\leq &C\left\{\frac{1}{(\mathrm{Var}(\psi(Z_n)))^{1/8}}\left(\frac{\log^4 n}{nh_n^d}\right)^{3/8} + \frac{\log n}{n}\right\},
\end{aligned}$$

where we chose $\eta = (\mathrm{Var}(\psi(Z_n))\log^4 n/(nh_n^d))^{3/8}$. It remains to show that the so chosen γ satisfies $\gamma \geq \sqrt{\log n/(nh_n^d)}$. Indeed, since

$$\begin{aligned}
\gamma = \frac{1}{(\mathrm{Var}(\psi(Z_n)))^{1/8}}\left(\frac{\log^4 n}{nh_n^d}\right)^{3/8} \geq &\sqrt{\frac{\log n}{nh_n^d}} \\
\mathrm{Var}(\psi(Z_n)) \leq &(nh_n^d)\log^8 n
\end{aligned}$$

which holds by assumption. □

Proof of Corollary 10: By the triangle inequality,

$$\begin{aligned}
|\sqrt{nh_n^d}\psi(\hat{\theta}_n - \theta_n) - \psi(Z_n)| \leq &|\sqrt{nh_n^d}\psi(\hat{\theta}_n - \theta_n) - \sqrt{nh_n^d}\psi(S_n - \mathrm{E}[S_n])| \\
&+ |\sqrt{nh_n^d}\psi(S_n - \mathrm{E}[S_n]) - \psi(Z_n)|
\end{aligned}$$

If $A \geq a$ and $A \leq B + C$, then $B \geq a - \varepsilon$ or $C \geq \varepsilon$. As suppose that both $B < a - \varepsilon$ and $C < \varepsilon$, then $A < a$ and by taking complements the claim follows. Thus,

$$\begin{aligned}
&\mathrm{P}\big(|\sqrt{nh_n^d}\psi(\hat{\theta}_n - \theta_n) - \psi(Z_n)| \geq 2\sqrt{\mathrm{Var}(\psi(Z_n))}\varepsilon\big) \\
\leq &\mathrm{P}\big(|\sqrt{nh_n^d}\psi(S_n - \mathrm{E}[S_n]) - \psi(Z_n)| \geq \sqrt{\mathrm{Var}(\psi(Z_n))}\varepsilon\big) \\
&+ \mathrm{P}\big(|\sqrt{nh_n^d}\psi(\hat{\theta}_n - \theta_n) - \sqrt{nh_n^d}\psi(S_n - \mathrm{E}[S_n])| \geq \sqrt{\mathrm{Var}(\psi(Z_n))}\varepsilon\big)
\end{aligned} \tag{2.8}$$

[7]We show in Lemma 54 that Z_n can be chosen so that it satisfies the regularity conditions imposed by our anti-concentration in Theorem 14.

By $\delta_n/\sqrt{\mathrm{Var}(\psi(Z_n))} = O(1)$, $\sqrt{\mathrm{Var}(\psi(Z_n))} \geq c\delta_n$ eventually for some constant c, and therefore

$$\begin{aligned}&\mathrm{P}\big(|\sqrt{nh_n^d}\psi(\hat{\theta}_n - \theta_n) - \sqrt{nh_n^d}\psi(S_n - \mathrm{E}[S_n])| \geq \sqrt{\mathrm{Var}(\psi(Z_n))}\varepsilon\big)\\ \leq&\mathrm{P}\big(|\sqrt{nh_n^d}\psi(\hat{\theta}_n - \theta_n) - \sqrt{nh_n^d}\psi(S_n - \mathrm{E}[S_n])| \geq c\delta_n\varepsilon\big) \to 0, \qquad \text{as } n \to \infty.\end{aligned}$$

Regarding (2.8), note that $\eta = (\mathrm{Var}(\psi(Z_n))\log^4 n/(nh_n^d))^{3/8}$ so that

$$\mathrm{P}(|\sqrt{nh_n^d}\psi(S_n - \mathrm{E}[S_n]) - W_n| > \eta) \leq C\bigg(\frac{1}{(\mathrm{Var}(\psi(Z_n)))^{1/8}}\bigg(\frac{\log^4 n}{nh_n^d}\bigg)^{3/8} + \frac{\log n}{n}\bigg).$$

Moreover, by assumption, $\eta/\sqrt{\mathrm{Var}(\psi(Z))}$

$$\frac{\eta}{\sqrt{\mathrm{Var}(\psi(Z_n))}} = \frac{1}{(\mathrm{Var}(\psi(Z_n)))^{1/8}}\bigg(\frac{\log^4 n}{nh_n^d}\bigg)^{3/8} \to 0$$

and therefore $\eta \leq C\sqrt{\mathrm{Var}(\psi(Z))}$ eventually. Thus,

$$\begin{aligned}&\mathrm{P}\big(|\sqrt{nh_n^d}\psi(S_n - \mathrm{E}[S_n]) - \psi(Z_n)| \geq \sqrt{\mathrm{Var}(\psi(Z_n))}\varepsilon\big)\\ \leq&\mathrm{P}\big(|\sqrt{nh_n^d}\psi(S_n - \mathrm{E}[S_n]) - \psi(Z_n)| \geq \eta\big) \to 0, \qquad \text{as } n \to \infty.\end{aligned}$$

Finally, by applying Le Cam's Lemma 51 with $\eta = \sqrt{\mathrm{Var}(\psi(Z_n))}\varepsilon$

$$\sup_{t\in\mathbb{R}}|\mathrm{P}(\sqrt{nh_n^d}\psi(\hat{\theta}_n - \theta_n) \leq t) - \mathrm{P}(\psi(Z_n) \leq t)| \leq \varepsilon + o(1),$$

proving the claim. □

Proof of Proposition 4: Under Assumption 15, the Rio-Massart Coupling in Theorem 8 in Chernozhukov et al. (2010) constructs a sequence of P-Gaussian Bridges Z_n (on a possibly enriched probability space) such that for any $t \geq C\log n$,

$$\mathrm{P}\bigg(\sup_{v\in V}|h_n^{-d/2}\mathbb{G}_n(g_v) - Z_n(g_v)| \geq C\bigg\{\sqrt{\frac{t}{n^{1/(d_1+d)}h_n}} + t\sqrt{\frac{\log n}{nh_n^d}}\bigg\}\bigg) \leq \exp(-t).$$

Moreover, the paths of $\nu \mapsto Z_n(g_v)$ can be chosen to be continuous a.s. and therefore Z_n implies a Gaussian $\mathcal{C}(\mathcal{V})$-valued random vector.

By Lipschitz continuity of ψ, for any $\eta > 0$,

$$\mathrm{P}(|\psi(\sqrt{nh_n^d}(S_n - \mathrm{E}[S_n])) - \psi(Z_n)| > \eta) \leq \mathrm{P}(\|\sqrt{nh_n^d}(S_n - \mathrm{E}[S_n]) - Z_n\| \geq \eta/L)$$

and since L does not depend on n, the Rio-Massart Coupling from above implies

$$\mathrm{P}\bigg(|\psi(\sqrt{nh_n^d}(S_n - \mathrm{E}[S_n])) - \psi(Z_n)| \geq C\bigg\{\sqrt{\frac{t}{n^{1/(d_1+d)}h_n}} + t\sqrt{\frac{\log n}{nh_n^d}}\bigg\}\bigg) \leq \exp(-t).$$

Thus, for the Kolmogorov distance, set

$$\eta = tC\bigg\{\sqrt{\frac{1}{n^{1/(d_1+d)}h_n}} + \sqrt{\frac{\log n}{nh_n^d}}\bigg\} \geq C\bigg\{\sqrt{\frac{t}{n^{1/(d_1+d)}h_n}} + t\sqrt{\frac{\log n}{nh_n^d}}\bigg\}$$

i.e., take t as

$$t = C\eta\bigg\{\sqrt{\frac{1}{n^{1/(d_1+d)}h_n}} + \sqrt{\frac{\log n}{nh_n^d}}\bigg\}^{-1}$$

implying for our Kolmogorov distance minimization problem:

$$\min_t \exp(-C\eta\bigg\{\sqrt{\frac{1}{n^{1/(d_1+d)}h_n}} + \sqrt{\frac{\log n}{nh_n^d}}\bigg\}^{-1}) + \frac{\eta}{\sqrt{\mathrm{Var}(\psi(Z_n))}}$$

Let $\Delta_n = \bigg\{\sqrt{\frac{1}{n^{1/(d_1+d)}h_n}} + \sqrt{\frac{\log n}{nh_n^d}}\bigg\}$. Then, the first order condition of the above problems can be written as

$$\begin{aligned}
\Delta_n^{-1}\exp(-\eta\Delta_n^{-1}) &= \frac{1}{\sqrt{\mathrm{Var}(\psi(Z_n))}} \\
\iff \qquad -\eta\Delta_n^{-1} &= \log\bigg(\frac{\Delta_n}{\sqrt{\mathrm{Var}(\psi(Z_n))}}\bigg) \\
\iff \qquad \eta &= \Delta_n \log\bigg(\frac{\sqrt{\mathrm{Var}(\psi(Z_n))}}{\Delta_n}\bigg).
\end{aligned}$$

This implies

$$t = C\log\bigg(\frac{\sqrt{\mathrm{Var}(\psi(Z_n))}}{\Delta_n}\bigg)$$

$$\begin{aligned}&=C\{\log(\mathrm{Var}(\psi(Z_n))) - \log(\Delta_n)\}\\&=C\{\log(\mathrm{Var}(\psi(Z_n))) + \log(n) + o(\log(n))\}\end{aligned}$$

which is larger than $C \log n$ since $\Delta_n/\sqrt{\mathrm{Var}(\psi(Z_n))} = O(n^{-\xi})$ for some $\xi > 0$ and therefore the above choice of η is valid. This implies the following rate for the Kolmogorov bound

$$C\frac{\Delta_n}{\sqrt{\mathrm{Var}(\psi(Z_n))}}\left\{1 + \log\left(\frac{\sqrt{\mathrm{Var}(\psi(Z_n))}}{\Delta_n}\right)\right\}$$

and the claim follows. □

Proof of Corollary 11: By the triangle inequality,

$$\begin{aligned}|\sqrt{nh_n^d}\psi(\hat{\theta}_n - \theta_n) - \psi(Z_n)| \leq &|\sqrt{nh_n^d}\psi(\hat{\theta}_n - \theta_n) - \sqrt{nh_n^d}\psi(S_n - \mathrm{E}[S_n])|\\&+ |\sqrt{nh_n^d}\psi(S_n - \mathrm{E}[S_n]) - \psi(Z_n)|\end{aligned}$$

If $A \geq a$ and $A \leq B + C$, then $B \geq a - \varepsilon$ or $C \geq \varepsilon$. As suppose that both $B < a - \varepsilon$ and $C < \varepsilon$, then $A < a$ and by taking complements the claim follows. Thus,

$$\begin{aligned}&\mathrm{P}\big(|\sqrt{nh_n^d}\psi(\hat{\theta}_n - \theta_n) - \psi(Z_n)| \geq 2\sqrt{\mathrm{Var}(\psi(Z_n))}\varepsilon\big)\\\leq&\mathrm{P}\big(|\sqrt{nh_n^d}\psi(S_n - \mathrm{E}[S_n]) - \psi(Z_n)| \geq \sqrt{\mathrm{Var}(\psi(Z_n))}\varepsilon\big)\\&+ \mathrm{P}\big(|\sqrt{nh_n^d}\psi(\hat{\theta}_n - \theta_n) - \sqrt{nh_n^d}\psi(S_n - \mathrm{E}[S_n])| \geq \sqrt{\mathrm{Var}(\psi(Z_n))}\varepsilon\big)\end{aligned} \tag{2.9}$$

By $\delta_n/\sqrt{\mathrm{Var}(\psi(Z_n))} = O(1)$, $\sqrt{\mathrm{Var}(\psi(Z_n))} \geq c\delta_n$ eventually for some constant c, and therefore

$$\begin{aligned}&\mathrm{P}\big(|\sqrt{nh_n^d}\psi(\hat{\theta}_n - \theta_n) - \sqrt{nh_n^d}\psi(S_n - \mathrm{E}[S_n])| \geq \sqrt{\mathrm{Var}(\psi(Z_n))}\varepsilon\big)\\\leq&\mathrm{P}\big(|\sqrt{nh_n^d}\psi(\hat{\theta}_n - \theta_n) - \sqrt{nh_n^d}\psi(S_n - \mathrm{E}[S_n])| \geq c\delta_n\varepsilon\big) \to 0, \qquad \text{as } n \to \infty.\end{aligned}$$

Regarding (2.9), note that

$$\eta = \Delta_n \log\left(\frac{\sqrt{\mathrm{Var}(\psi_n)}}{\Delta_n}\right)$$

in the proof of Proposition 4. Moreover, by assumption, $\eta/\sqrt{\mathrm{Var}(\psi(Z))}$ converges

to zero and therefore $\eta c \leq C\sqrt{\mathrm{Var}(\psi(Z))}$ eventually. Therefore,

$$\begin{aligned}&\mathrm{P}\big(|\sqrt{nh_n^d}\psi(S_n - \mathrm{E}[S_n]) - \psi(Z_n)| \geq \sqrt{\mathrm{Var}(\psi(Z_n))}\varepsilon\big)\\ \leq&\mathrm{P}\big(|\sqrt{nh_n^d}\psi(S_n - \mathrm{E}[S_n]) - \psi(Z_n)| \geq \eta\big) \leq C\frac{\eta}{\sqrt{\mathrm{Var}(\psi(Z_n))}} \to 0, \qquad \text{as } n \to \infty.\end{aligned}$$

Finally, by applying Le Cam's Lemma 51 with $\eta = \sqrt{\mathrm{Var}(\psi(Z_n))}\varepsilon$

$$\sup_{t\in\mathbb{R}}|\mathrm{P}(\sqrt{nh_n^d}\psi(\hat{\theta}_n - \theta_n) \leq t) - \mathrm{P}(\psi(Z_n) \leq t)| \leq \varepsilon + o(1),$$

proving the claim. □

Chapter 3

Inference in the High-Dimensional Partially Linear Model using Orthogonalized Kernels

3.1 Introduction

High-dimensional linear models have become increasingly popular in various areas in economics. Examples include estimation of impulse response functions of macroeconomic time-series using local projections (Jordà, 2005), optimal portfolio estimation with many underlying assets (Fan et al., 2011), demand estimation with many generates controls and instruments (Chernozhukov et al., 2015b), textual analysis of patents in the context of growth estimation (Kelly et al., 2021) and the analysis of job training programs with many generated controls Spiess et al. (2023). Inference in the high-dimensional linear model has been extensively studied over the last decades (Meinshausen et al. (2009), Meinshausen and Bühlmann (2010), Wasserman and Roeder (2009), Shah and Samworth (2013), Belloni et al. (2014), van de Geer et al. (2014) and Zhang and Zhang (2014)).

In this paper, we study inference in the high-dimensional partially linear model (HD PLM). This model is a semiparametric extension of the high-dimensional linear model and allows variables to enter nonlinearly into the regression equation. It therefore maintains the flexibility of nonparametric models in some of the variables while avoiding the curse of dimensionality of a fully nonparametric model. In the low-dimensional setting, the partially linear model has a long history in the econo-

metrics and semiparametric statistics literature and has been studied for example in Engle et al. (1986), Robinson (1988), Chen (1988) and Härdle et al. (2000). Our model extends this model in that we allow the dimension of the linear part to increase with the sample size and in particular allow the number of regressors to be (potentially) larger than the sample size.

Recently, there has been a growing interest in the high-dimensional partially linear model. Much progress has been made to understand the estimation properties of various proposals (Müller and van de Geer (2015), Ma and Huang (2016), Zhu (2017), Yu et al. (2019)) and inference on the linear part has been studied by Zhu et al. (2019). We complement this literature in that we focus on inference on the nonlinear part.

More rigorously, we consider the following model: Let $\{(Y_i, T_i, X_i) : i = 1, \dots, n\}$ be a sample of i.i.d data, where Y_i and T_i are real-valued random variables and $X_i = (X_{i1}, \dots, X_{ip})^T$ is a p-dimensional random vector with p potentially very large, in particular, $p \gg n$. Suppose that the data satisfy the model

$$Y_i = \delta + m(T_i) + X_i^T \beta + \varepsilon_i \tag{3.1}$$

where m is smooth unknown function, $\beta = (\beta_1, \dots, \beta_p)^T$ is an unobserved sparse parameter vector with s nonzero entries and ε_i is an error term satisfying $\mathrm{E}[\varepsilon_i \,|\, T_i, X_i] = 0$ as well as $\mathrm{E}[\varepsilon_i^2 \,|\, T_i, X_i] = \sigma^2$. m is normalized so that $\mathrm{E}[m(T_i)] = 0$ in order to achieve identifiability and the X_i are assumed to be centered in order to shorten the notation. We assume that T_i has compact support, which w.l.o.g. is normalized to be the unit interval $[0, 1]$.

Estimation of the nonlinear part in the above model is challenging due to the high-dimensional linear part as we illustrate in the following. As suggested by Robinson (1988), estimation of the nonlinear part in a partially linear model can be based on the identity

$$m(T) = \mathrm{E}[Y - \delta - X^\intercal \beta \,|\, T].$$

Therefore, given a preliminary estimator $(\hat{\delta}, \hat{\beta})$ of (δ, β), one can estimate m using arbitrary nonparametric estimators of univariate conditional mean functions. We focus on Nadaraya-Watson estimation using the uniform kernel, but it might also be interesting to extend our proposed method to other kernels or local polynomial

estimators. This Nadaraya-Watson estimator is given by

$$\tilde{m}(t) = \frac{\sum_{i=1}^n \mathbb{1}(|T_i - t| \leq h)(Y_i - \hat{\delta} - X_i^\intercal \hat{\beta})}{\sum_{i=1}^n \mathbb{1}(|T_i - t| \leq h)},$$

where $h > 0$ denotes a bandwidth chosen by the researcher. This estimator can be decomposed into three parts

$$\begin{aligned}\tilde{m}(t) =& \frac{\sum_{i=1}^n \mathbb{1}(|T_i - t| \leq h)(Y_i - \delta - X_i^\intercal \beta)}{\sum_{i=1}^n \mathbb{1}(|T_i - t| \leq h)} \\ &+ \frac{\sum_{i=1}^n \mathbb{1}(|T_i - t| \leq h) X_i^\intercal (\beta - \hat{\beta})}{\sum_{i=1}^n \mathbb{1}(|T_i - t| \leq h)} + (\delta - \hat{\delta}).\end{aligned}$$

The first term corresponds to the infeasible Nadaraya-Watson estimator which knows the true values of δ and β. The second and the third term capture the influence of estimation of (δ, β) on the estimator. While $\hat{\delta}$ converges sufficiently fast to zero, showing that the second term is negligible requires fairly strong conditions on the estimation properties of $\hat{\beta}$. This problem is absent in the low-dimensional setting when p is fixed or much smaller than n. The second term therefore resembles a bias due to the high-dimensionality of the problem, and we refer to it as the HD bias in the following.

In order to deal with the HD bias, we modify the uniform kernel so that it is nearly orthogonal to X and by this we reduce the impact of the estimation error of β. More precisely, we construct orthogonalization parameters $\hat{\gamma}_t$ which depend only on (T_i, X_i), $i = 1, \ldots, n$, and base our test statistic on the orthogonalized Nadaraya-Watson estimator

$$\hat{m}(t) = \frac{\sum_{i=1}^n \mathbb{1}(|T_i - t| \leq h)(1 - X_i^\intercal \hat{\gamma}_t)(Y_i - \hat{\delta} - X_i^\intercal \hat{\beta})}{\sum_{i=1}^n \mathbb{1}(|T_i - t| \leq h)(1 - X_i^\intercal \hat{\gamma}_t)}.$$

These orthogonalization parameters are chosen to ensure that

$$\max_{1 \leq j \leq p} \left| \frac{1}{nh} \sum_{i=1}^n \mathbb{1}(|T_i - t| \leq h)(1 - X_i^\intercal \hat{\gamma}_t) X_{ij} \right| \leq \mu, \tag{3.2}$$

where μ is a tuning parameter which can be chosen by the researcher.

We are interested in inference on the nonlinear part m and want to test the

global hypothesis

$$H_0 : \forall t : m(t) = 0 \quad \text{vs.} \quad H_1 : \exists t : m(t) \neq 0.$$

Under model (3.1), the null hypothesis implies that the conditional mean of Y given T and X does not depend on T. In this sense, this test problem generalizes the idea of significance testing frequently used in applied research by allowing T to affect Y in a general nonlinear way. Based on the orthogonalized Nadaraya-Watson estimator, we can construct a test statistic for the global null by constructing a test for the equivalent hypotheses

$$H_0' : \sup_{t\in[0,1]} |m(t)| = 0 \quad \text{vs.} \quad H_1' : \sup_{t\in[0,1]} |m(t)| > 0.$$

Such a sup-norm test can also be interpreted as a joint test of the local hypotheses

$$H_{0,t} : m(t) = 0, \qquad t \in [0,1]$$

against the alternatives

$$H_{+,t} : m(t) > 0 \quad \text{and} \quad H_{-,t} : m(t) < 0.$$

A rejection of the global null therefore allows us to infer the locations and signs of the deviations from the null. This is in contrast to other equivalent formulations of the global null as e.g. formulations based on L_p-norms.

However, the evaluation of the sup-norm would require us to compute a continuum of orthogonalization parameters $\hat{\gamma}_t$ which is both theoretically and computationally challenging. We therefore discretize the support of T and consider only a grid of locations $\mathcal{T}_L := \{t_\ell : \ell = 1, \dots, L\} \subset [0,1]$ and the surrogate global hypotheses

$$H_{0,L}' : \max_{t_\ell\in\mathcal{T}_L} |m(t_\ell)| = 0 \quad \text{vs.} \quad H_1' : \max_{t_\ell\in\mathcal{T}_L} |m(t_\ell)| > 0,$$

and define the local hypotheses $H_{0,\ell} : m(t_\ell) = 0, H_{+,\ell} : m(t_\ell) > 0$ and $H_{-,\ell} : m(t_\ell) < 0$, $\ell = 1, \dots, L$ analogously. Due to the smoothness of m, the impact of the discretization on the power of the test is negligible when the grid is chosen sufficiently dense.

We study our proposed test procedure using commonly imposed assumptions from the literature on estimation in the HD PLM. We propose a first stage Lasso estimator of β and study its estimation properties. In particular, we show that it has the same rate of convergence for its estimation and prediction error as the infeasible Lasso estimator which knows the true conditional means of Y and X given T. Further, we study the bias and asymptotic variance of the orthogonalized Nadaraya-Watson estimator and find that its estimation properties are of the same order as those of the infeasible Nadaraya-Watson estimator introduced above. Additionally, we provide a consistent multiplier bootstrap for the critical values of the test statistic and show uniform consistency of the proposed test against local Hölder balls. Our results are based on a mild sparsity assumption of $s = o(\sqrt{n})$ (up to log terms) which is only mildly stronger than what is needed for ℓ_1-consistency of the Lasso in the high-dimensional linear model and a growth restriction to population counterparts of the orthogonalization parameters $\hat{\gamma}_t$ which can be shown to be satisfied when T and X are only weakly dependent.

We study the finite sample performance of our test in a Monte Carlo simulation. We implement our method for different data adaptive tuning parameter choices and compare its size and power to a test based on the infeasible Nadaraya-Watson estimator as well as a Nadaraya-Watson estimator which does not use an orthogonalized kernel. The simulation results show that without orthogonalization the Nadaraya-Watson estimator is heavily size-distorted and that the orthogonalization yields size control. The power of our proposed method is comparable to the power of the infeasible estimator although smaller.

Related Literature: Our estimation results for the first stage Lasso estimator and the orthogonalized Nadaraya-Watson estimator contribute to the literature on estimation in the HD PLM (Müller and van de Geer (2015), Ma and Huang (2016), Zhu (2017), Yu et al. (2019)). While our first stage Lasso estimator shares the rates with the proposal in the literature, the orthogonalized Nadaraya-Watson estimator yields the same rates as other proposals under weakened sparsity assumptions. We collect the sparsity requirements of the aforementioned papers in Table 3.1.1. These requirements are computed to ensure that the nonlinear part can be estimated at the optimal nonparametric rate.

Our proposal is also related to the literature on inference in the high-dimensional / sparse additive model or spAM for short (Kozbur (2020), Lu et al. (2020), Gregory et al. (2021) and Guo et al. (2019)). In this literature, the conditional mean is

Paper	Model	sparsity requirements	sample splitting
Müller and van de Geer (2015)	HD PLM	$s = o(n^{1/5})$	No
Ma and Huang (2016)	HD PLM	$s = o(n^{-3/10})$	No
Yu et al. (2019)	HD PLM	$s = o(n^{1/5})$	No
Gregory et al. (2021)	spAM	$s = o(n^{3/10})$	No
Lu et al. (2020)	spAM	$s = O(n^{1/6})$	No
Guo et al. (2019)	spAM	$s = o(n^{4/5})$	Yes
Our proposal	HD PLM	$s = o(\sqrt{n})$	No

Table 3.1.1: Sparsity requirements from related methods in HD PLM and additive models. These rates were computed to ensure that the nonlinear part can be estimated at the optimal nonparametric rate. The sparsity requirements are only correct up to powers of log terms.

modeled as

$$Y_i = \delta_0 + m(T_i) + \sum_{j=1}^{d} g_j(X_{ij}) + \varepsilon_i,$$

where g_j are unknown smooth functions satisfying $\mathrm{E}[g_j(X_{ij})] = 0$ and the residual ε_i satisfies $\mathrm{E}[\varepsilon_i|T, X] = 0$. This model is closely related to the HD PLM if one approximates the unknown functions g_j by a series expansion

$$g_j(x) = \sum_{k=1}^{m} \psi_k(x)\beta_{0,kj} + r_{mj}(x)$$

where the ψ_k are known basis functions and r_{mj} is an approximation error. Then,

$$Y_i = \delta_0 + m(T_i) + \sum_{j=1}^{d}\sum_{k=1}^{m} \psi_k(X_{ij})\beta_{0,kj} + \sum_{j=1}^{d} r_j(X_{ij}) + \varepsilon_i \tag{3.3}$$

Up to the approximation error, this is an HD PLM model and in this sense our results are related to this literature.

Kozbur (2020) propose the Post-Nonparametric Double Selection method which is an extension of the Post Double Selection method developed in Belloni et al. (2014) to the spAM. Heuristically, they also linearize m in the additive model with

expansion

$$m(t) = \sum_{k=1}^{m} \varphi_k(t)\gamma_{0,k} + r_{m0}(t)$$

for some basis functions φ_k and approximation error r_{m0}. For each basis function φ_k and for Y they use a Lasso-type procedure to select the important regressors in a linear regression of φ_k (or Y) on $\{(\psi_k(X_{ij})) : k = 1, \dots, m, j = 1, \dots, d\}$. Then, in a second stage, they perform OLS on the union of all the selected regressors. Under regularity conditions, they show that this two-stage procedure has a normal limiting distribution when the number of selected variables is not too large. However, it is not clear under what type of assumptions overselection can be avoided. For example, suppose that the test functions form a basis of the space of square-integrable functions. Assuming that most of the technical regressors φ_k on $\{(\psi_k(X_{ij})) : k = 1, \dots, m, j = 1, \dots, d\}$ do not affect any of the φ_k is close to assuming that T is independent of most of the included variables. In this sense, it seems that they impose a stronger form of weak dependence than what we need for our orthogonalization approach.

Gregory et al. (2021) propose a three stage pre- and re-smoothing estimator. In the first step, they construct an initial estimator of m and β which they debiase in the second step by using an extension of the construction in van de Geer et al. (2014). In the third step, they re-smooth the resulting estimator using nonparametric smoothing. They study the performance of their estimator in comparison to an oracle estimator which knows the nuisance functions g_j. They show that their pre- and re-smoothing estimator has the same properties as the oracle estimator if the sparsity of β_0, satisfies $s \ll n^{3/10}$ which is more restrictive than our corresponding assumption.

Lu et al. (2020) propose the Kernel-Sieve Method. In this two-stage procedure, they first estimate m and β at some location $t \in [0, 1]$ by minimizing

$$\min_{a,b} \sum_{i=1}^{n} k_h(T_i - t)\left(Y_i - \bar{Y} - a - \sum_{j=1}^{d}\sum_{k=1}^{m} \psi_k(X_{ij})b_{jk}\right)^2 + \lambda\mathcal{P}(a, b)$$

Here k is a kernel function, λ a tuning parameter and $\mathcal{P}$ a penalty term that induces an ℓ_1 penalty on a and a Group Lasso penalty on b. In the second stage, the debiase their estimator of m locally using a similar construction as in

Javanmard and Montanari (2014). Their estimator allows for changes in β depending on the location t. That is, their estimator is suitable in a varying coefficient extension of model (3.3):

$$Y_i = \delta + m(T_i) + \sum_{j=1}^{d} \sum_{k=1}^{m} \phi_k(Z_{ij}) \beta_{kj}(T_i) + \varepsilon_i$$

This also explains why they obtain comparably slow rates of convergence for their estimator of m. Further, their sparsity requirements as listed in Table 3.1.1 are stronger than what we need for our approach, and they impose a weak dependence assumption on the joint distribution of T and X which is stronger than our weak dependence requirements.

Guo et al. (2019) propose a decorrelated local linear estimator and study pointwise inference on the derivative of m. While their idea of decorrelating the kernel is similar to our proposal, their construction of the orthogonalization weights relies on assuming that the conditional mean of T given X follows a high-dimensional sparse model and therefore also relies on a kind of weak dependence assumption between T and X. Their sparsity requirements as displayed in Table 3.1.1 are best-case bounds from Section (A.4) in their Supplementary Material. The actual requirements depend on the sparsity of the conditional mean of T given X and therefore hard to compare to our sparsity requirements especially since they are interested in inference on the derivative instead of level of m.

Also related to our proposal is the literature on inference in high-dimensional linear models with structured or group sparsity. Mitra and Zhang (2016), van de Geer and Stucky (2016) and Stucky and van de Geer (2018) are interested in joint inference on multiple coefficients of the linear model. They allow the number of coefficients of interest to grow as the sample size increases and therefore such an approach can be applied to a linearized model, where m is expanded according to some basis. However, they are interested in chi-squared type inference and therefore cannot test the local hypotheses $H_{0,\ell}$.

Agenda: The paper is organized as follows. In section 3.2, we introduce the test statistic and present a Bootstrap to obtain critical values. In section 3.3, we present asymptotic properties of our proposed estimators and test procedure. Finally, we present results of a Monte Carlos study in section 3.4. We outline the proofs in the Appendix and give detailed proofs of technical results in the Supplementary

Material.

Notation: Let $\|\cdot\|_q$ for $1 \leq q \leq \infty$ denote the ℓ_q-norm on $\mathbb{R}^p$. Denote by $Y = (Y_1, \ldots, Y_n)^\intercal$, $X^(j) = (X_{1j}, \ldots, X_{nj})^\intercal$, for $j = 1, \ldots, p$, $X = (X^{(1)}, \ldots, X^{(p)})$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\intercal$. Further, for a vector $v = (v_1, \ldots, v_n)^\intercal \in \mathbb{R}^n$, denote by $\bar{v} = \frac{1}{n}\sum_{i=1}^n v_i$ and by $\dot{v}_i = v_i - \bar{v}$. Finally, c and C without any subscript denote constants whose value might change at each occurrence.

3.2 Estimation and Test Procedure

Our test procedure can be interpreted as a two-stage procedure. In the first stage, estimators for β as well as the orthogonalization parameters $\hat{\gamma}$ are constructed which are then used in the second step to build the orthogonalized Nadaraya-Watson estimator $\hat{m}$.

In order to motivate our estimator for β, note that in the partially linear model m is given by

$$m(t) = \mathrm{E}[Y_i - \delta - X_i^\intercal \beta \,|\, T_i = t],$$

which implies for the model (3.1)

$$Y_i - \mathrm{E}[Y_i \,|\, T_i] = (X_i - \mathrm{E}[X_i \,|\, T_i])^\intercal \beta + \varepsilon_i.$$

This is a high-dimensional linear model in the partial residuals $Y_i - \mathrm{E}[Y_i \,|\, T_i]$ and $X_i - \mathrm{E}[X_i \,|\, T_i]$. These partial residuals can be estimated using the Nadaraya-Watson estimator as

$$\tilde{Y}_i = Y_i - \frac{\sum_{j=1}^n k_g(T_j - T_i) Y_j}{\sum_{j=1}^n k_g(T_j - T_i)} \quad \text{and} \quad \tilde{X}_i = X_i - \frac{\sum_{j=1}^n k_g(T_j - T_i) X_j}{\sum_{j=1}^n k_g(T_j - T_i)}.$$

Here, k denotes a kernel function, g a bandwidth and $k_g(v) = k(v/g)$. Now, β can be estimated using any estimator for high-dimensional linear regression on the dataset $(\tilde{Y}_i, \tilde{X}_i)$, $i = 1, \ldots, n$. We use the Lasso

$$\hat{\beta}_\lambda = \operatorname*{argmin}_{b \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^n \left(\tilde{Y}_i - \tilde{X}_i^T b\right)^2 + \lambda \|b\|_1 \right\}$$

and call the resulting estimator $\hat{\beta}_\lambda$ the profile Lasso due to its similar construction

as used in profile likelihood methods.

In order to achieve the near orthogonality property (3.2), we construct the orthogonalization parameters using an ℓ_1-penalized approach. Let $\mu > 0$ and denote by $\dot{X}_i$ the demeaned control variables $\dot{X}_i = X_i - \bar{X}$, where $\bar{X}$ denotes the sample average of the X_i. Take $\hat{\gamma}_t$ be a minimizer of

$$\min_{c \in \mathbb{R}^p} \left\{ \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}(|T_i - t| \leq h)(1 - \dot{X}_i^\intercal c)^2 + \mu \|c\|_1 \right\} \tag{3.4}$$

The tuning parameter μ governs the degree of orthogonality to the estimation error of the profile Lasso $\hat{\beta}_\lambda$. More precisely, the KKT conditions assure that the near orthogonality property (3.2) holds uniformly over $j = 1, \ldots, p$.

Given the profile Lasso $\hat{\beta}_\lambda$ and the orthogonalization parameters $\hat{\gamma}_t$, we construct the orthogonalized Nadaraya-Watson estimator $\hat{m}(t)$ as

$$\hat{m}(t) = \frac{\sum_{i=1}^{n} \mathbb{1}(|T_i - t| \leq h)(1 - \dot{X}_i^\intercal \hat{\gamma}_t) R_i}{\sum_{i=1}^{n} \mathbb{1}(|T_i - t| \leq h)(1 - \dot{X}_i^\intercal \hat{\gamma}_t)}, \tag{3.5}$$

where $R_i = \dot{Y}_i - \dot{X}_i^\intercal \hat{\beta}_\lambda$ denote the partial residuals from a linear regression of Y_i on X_i and a constant using the profile Lasso. This estimator can be interpreted as a Nadaraya-Watson estimator of the conditional mean of R_i given T_i using the orthogonalized kernel $\frac{1}{h}\mathbb{1}(|T_i - t| \leq h)(1 - \dot{X}_i^\intercal \hat{\gamma}_t)$. Further, the denominator in (3.5) is nonzero as long as $\mu > 0$ and some T_i lie in $[t_\ell - h, t_\ell + h]$ regardless of the dimension p.[1]

For the construction of the test statistic, we discretize the support of t into the set of locations $\mathcal{T}_L = \{t_\ell : \ell = 1, \ldots, L\}$ and for each $\ell = 1, \ldots, L$ set

$$\hat{\mathcal{S}}_{n,\ell} = \frac{\sqrt{nh}}{\hat{\sigma}_{n,\ell}} \hat{m}(t_\ell), \tag{3.6}$$

where $\hat{\sigma}_{n,\ell}$ is an estimator of the asymptotic variance of $\hat{\mathcal{S}}_{n,\ell}$ given by

$$\begin{aligned} \hat{\sigma}_{n,\ell}^2 &= \frac{\frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}(|T_i - t_\ell| \leq h)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})^2}{\left(\frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}(|T_i - t_\ell| \leq h)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell}) \right)^2} \hat{\sigma}_n^2 \\ \hat{\sigma}_n^2 &= \frac{1}{n} \sum_{i=1}^{n} (\tilde{Y}_i - \tilde{X}_i^\intercal \hat{\beta}_\lambda)^2. \end{aligned} \tag{3.7}$$

[1] Compare Lemma 70 in the Supplementary Material.

The test statistic for the global null hypothesis is given by

$$\hat{T}_n = \max_{1\leq \ell \leq L} |\hat{\mathcal{S}}_{n,\ell}|.$$

In the next section, we will show that under certain conditions, the distribution of $\hat{T}_n$ can be approximated by the distribution of the maximum of a multivariate Gaussian random vector Z with mean zero and covariance matrix related to the covariance matrix of the vector $(\hat{\mathcal{S}}_{n,\ell})_{\ell=1,\dots,L}$. However, the latter distribution is not known in practice, and therefore we use a Gaussian multiplier bootstrap to estimate the critical value of our test statistic. In order to describe this bootstrap, set $\hat{\varepsilon}_i = \tilde{Y}_i - \tilde{X}_i^\intercal \hat{\beta}_\lambda$ and let $\{e_i : i = 1, \dots, n\}$ be i.i.d. standard normal random variables independent of the data $\mathcal{D} := \{(Y_i, T_i, X_i) : i = 1, \dots, n\}$. Further, denote by

$$\tilde{W}_{i,\ell} = \frac{1}{\sqrt{h}\hat{\sigma}_{n,\ell}} \frac{\mathbb{1}(|T_i - t_\ell| \leq h)(1 - \dot{X}_i^\intercal \hat{\gamma}_{\ell,\mu})\hat{\varepsilon}_i}{\frac{1}{nh}\sum_{i=1}^n \mathbb{1}(|T_i - t_\ell| \leq h)(1 - \dot{X}_i^\intercal \hat{\gamma}_{\ell,\mu})} \tag{3.8}$$

$$\tilde{\overline{W}}_\ell = \frac{1}{n}\sum_{i=1}^n W_{i,\ell} \tag{3.9}$$

and define the bootstrapped values of $\hat{S}_{n,\ell}$ and $\hat{T}_n$ as

$$\hat{S}^*_{n,\ell} = \frac{1}{\sqrt{n}}\sum_{i=1}^n (\tilde{W}_{i,\ell} - \tilde{\overline{W}}_\ell)e_i$$
$$\hat{T}^*_n = \max_{\ell=1,\dots,L} |\hat{S}^*_{n,\ell}|.$$

Using repeated draws of the e_i, we simulate the conditional distribution of $\hat{T}^*_n$

$$\mathrm{P}(\hat{T}^*_n \leq q \,|\, \mathcal{D})$$

and set the critical value $\hat{q}_{1-\alpha}$ as the $(1-\alpha)$-quantile of this distribution.

Given these critical values, we can use the test statistics $\hat{\mathcal{S}}_{n,\ell}$ to either test the global null hypothesis $H_0 : m(t) = 0$ for all $t \in [0,1]$ or the local hypotheses $H_{0,\ell} : m(t_\ell) = 0$. By this, we can not only infer whether m deviates from zero, but also the locations where it differs. Moreover, we can use the test for alternatives $H_{-1,\ell} : m(t_\ell) < 0$ and $H_{1,\ell} : m(t_\ell) > 0$ and thus infer also the sign of the deviation from zero. Concretely, we can reject the global null H_0 if $\hat{T}_n > \hat{q}_{1-\alpha}$ and reject the

local hypothesis $H_{0,\ell}$ in favor of $H_{1,\ell}$ whenever $\hat{\mathcal{S}}_{n,\ell} > \hat{q}_{1-\alpha}$ or in favor of $H_{-1,\ell}$ when $\hat{\mathcal{S}}_{n,\ell} < -\hat{q}_{1-\alpha}$.

3.3 Theoretical Properties

In this section, we study the theoretical properties of the estimation and inference procedures presented in Section 3.2. Here, we will first introduce some general assumptions on the data-generating process which will be used across all the upcoming results which will be refined in the following sections. In Section 3.3.1, we will analyze the estimation properties of the profile Lasso $\hat{\beta}_\lambda$ and describe the bias and variance properties of the orthogonalized Nadaraya-Watson estimator in Section 3.3.2. In Section 3.3.3, we derive a distributional approximation of the test statistic and show consistency of the bootstrap in Section 3.3.4. Finally, we analyze the power properties of our test in Section 3.3.5.

Our asymptotic results rely on row-wise i.i.d. triangular array data $\{(Y_{i,n}, T_{i,n}, X_{i,n}) : i = 1, \ldots, n\}$ whose distribution is allowed to change in n. In particular, all the parameters that characterize the distribution of $\{(Y_{i,n}, T_{i,n}, X_{i,n})\}$ such as the conditional means of $X_{i,n}$ given $T_{i,n}$, m, the dimension of p and the sparsity of β are allowed to change with n. We omit the dependence on n where possible in order to increase readability. We use such triangular array asymptotics to better capture finite-sample phenomena which occur when the number of covariates is large compared to the sample size. A positive side effect of triangular asymptotics is that our asymptotic results hold uniformly over all sequences of data generating processes which satisfy the imposed assumptions below.

Our assumptions on the data generating process make use of the following notion of smoothness: The Hölder class $\mathcal{H}(\eta, M)$ on $[0,1]$ is the class of $m = \lfloor \eta \rfloor$ times differentiable functions $f : [0,1] \to \mathbb{R}$ satisfying

$$\sup_{t \in [0,1]} |f(t)| + \sup_{t \neq s} \frac{|f^{(m)}(t) - f^{(m)}(s)|}{|t-s|^{\eta - m}} \leq M.$$

Here, $\lfloor \eta \rfloor$ denotes the largest integer smaller or equal to η and $f^{(m)}$ the mth derivative of f.

We assume on the data-generating process:

Assumption 16. *(i) The marginal density of T_i, f_T, satisfies $f_T \in \mathcal{H}(\eta_f, M_f)$ for*

some constants $\eta_f \in (2,3]$ *and* $M_f < \infty$ *and there exists a constant* $0 < c < \infty$ *such that* $f_T(t) \geq c$ *for all* $t \in [0,1]$.

(ii) $X_i \in [-1,1]^p$ *and the conditional means* $m_{X,j}(T_i) := \mathrm{E}[X_{ij} \,|\, T_i]$, $j = 1, \ldots, p$, *satisfy* $m_{X_j} \in \mathcal{H}(\eta_X, M_X)$ *for some constants* $\eta_X \in (2,3]$ *and* $M_X < \infty$. *The matrix* $\mathrm{E}[u_i u_i^\intercal]$, *where* $u_i = X_i - \mathrm{E}[X_i \,|\, T_i]$, *is nonsingular. The number of covariates grows at most polynomially in* n, *i.e.,* $p = O(n^{\vartheta_p})$ *for some constant* ϑ_p.

(iii) The residuals ε_i *are conditionally sub-Gaussian given* (T_i, X_i) *in the sense that there exists a constant* K_ε *such that* $\mathrm{E}[\exp(\varepsilon_i^2 / K_\varepsilon^2) \,|\, T_i, X_i] \leq 2$ *a.s.*

(iv) The nonlinear part, m, *satisfies* $m \in \mathcal{H}(\eta_m, M_m)$ *for some constants* $\eta_m \in (2,3]$ *and* $M_m < \infty$.

(v) Let $S = \{j : \beta_j \neq 0\}$. *The sparsity of* β, $s = \#S$, *satisfies* $s = o(\sqrt{n} / \log^3 p_n)$.

(vi) The constants $\eta_f, M_f, c, \eta_X, M_X, \vartheta_p, K_\varepsilon, \eta_m$ *and* M_m *are independent of* n.

Assumption 16 is mild and imposes common restrictions from the literatures on nonparametric and high-dimensional linear regression. For instance, the assumptions on the marginal density of T and the smoothness assumptions on m and $m_{X,j}$ are standard in the nonparametric regression literature and ensure that the estimation error of the partial residuals converges sufficiently fast to zero. The nonsingularity of $\mathrm{E}[u_i u_i^\intercal]$ is needed for identifiability of β_0 (Robinson, 1988). Heuristically, no covariate is allowed to be a deterministic transformation of T and particularly no constant is allowed in X. Moreover, the boundedness of the covariates, the growth requirement on the number of covariates and Assumption (iii) and the sparsity requirement in (v) are common in the literature on high-dimensional linear regression. Specifically, the sparsity condition on β is only slightly stronger than what is needed for ℓ_1-consistency in the Lasso literature. As discussed in the literature review, the sparsity condition is weaker than in Gregory et al. (2021), Lu et al. (2020) and Kozbur (2020) and potentially weaker than in Guo et al. (2019).

3.3.1 Analysis of the Profile Lasso

For the estimation properties of the Profile Lasso, we further impose:

Assumption 17. *(i) For* $b = (b_1, \dots, b_p)^\intercal \in \mathbb{R}^p$*, let* $b_S = (b_j \mathbb{1}(j \in S))_{j=1}^p$ *and* $b_{S^c} = (b_j \mathbb{1}(j \notin S))_{j=1}^p$*. The matrix* $\mathrm{E}[u_i u_i^\intercal]$*, where* $u_i = X_i - \mathrm{E}[X_i \,|\, T_i]$*, satisfies the compatibility condition*

$$\phi_0^2 := \min\left\{ s\,\mathrm{E}\left[(u_i^\intercal b)^2\right] : \|b_S\|_1 = 1,\ \|b_{S^c}\|_1 \leq 6 \right\} \geq c$$

for some constant $c > 0$ *which is independent of* n*.*

(ii) The kernel k *is a Lipschitz continuous second order kernel with boundary correction. By this, we mean that*

$$k_g(x,t) = \begin{cases} k_{low}^{(t/g)}\left(\frac{x-t}{h}\right) & , \text{ if } t \in [0, g) \\ k_{int}\left(\frac{x-t}{g}\right) & , \text{ if } t \in [g, 1-g] \\ k_{up}^{(1-t/g)}\left(\frac{x-t}{g}\right) & , \text{ if } t \in (1-g, 1] \end{cases}$$

where $k_{low}^{(q)}$*,* $k_{up}^{(q)}$ *and* k_{int} *are Lipschitz-continuous second order kernels with support* $[-q, 1]$*,* $[-1, q]$ *and* $[-1, 1]$ *respectively. Further,* $k_g(x, \cdot)$ *satisfies*

$$\sup_{(x,g)} \sup_{t \neq s} g \frac{|k_g(x,s) - k_g(x,t)|}{|s-t|} < \infty.$$

(iii) The Lasso penalty λ *satisfies* $\lambda = C_\lambda \sqrt{\log p_n / n}$ *for some sufficiently large constant* C_λ *and the bandwidths* g *is chosen as* $g = C_g n^{-1/5}$ *for some constant* C_g*. The constants* C_λ *and* C_g *are independent of* n*.*

Assumption (i) is a compatibility condition on the covariance matrix of the partial residuals u_i. It can be seen as an extension of the usual compatibility condition used in the Lasso literature to the HD PLM. Similar conditions are made in the literature on estimation in the HD PLM (cf. Müller and van de Geer (2015), Zhu (2017) and Yu et al. (2019)). It is implied by lower bounded eigenvalues of $\mathrm{E}[u_i u_i^\intercal]$ and the latter is slightly stronger than what is needed for identification of β_0. The growth conditions on the bandwidth g and the Lasso tuning parameter λ are standard in the literature. These growth conditions guarantee that the effective noise of the profile Lasso is under control and that a sample version of the compatibility condition in (i) holds with high probability.

We use boundary corrected kernels since the estimation properties of our esti-

mator $\hat{\beta}_\lambda$ depend on the kernel k through the partial residuals $\tilde{X}_i$ and

$$\frac{1}{n}\sum_{i=1}^{n}\big(\hat{m}_k^*(T_i) - m(T_i)\big)^2 \tag{3.10}$$

where $\hat{m}_k^*(T_i)$ denotes the infeasible Nadaraya-Watson Estimator

$$\hat{m}_k^*(t) = \frac{\sum_{j=1}^n k_g(T_j - t)(m(T_i) + \varepsilon_i)}{\sum_{j=1}^n k_g(T_j - t)}.$$

The use of boundary-corrected kernels guarantees that the partial residuals $\tilde{X}_i$ are sufficiently well-behaved and that (3.10) converges sufficiently fast to zero.

Given these assumptions, we can derive upper bounds on the rate of convergence of the prediction and estimation error of the profile Lasso $\hat{\beta}_\lambda$.

Theorem 16. *Suppose that Assumptions 16 and 17 hold. Then, the prediction and estimation error of $\hat{\beta}_\lambda$ satisfy*

$$\frac{1}{n}\|\tilde{X}(\hat{\beta}_\lambda - \beta_0)\|_2^2 = O_p\left(s\frac{\log p}{n}\right) \quad \textit{and} \quad \|\hat{\beta}_\lambda - \beta_0\|_1 = O_p\left(s\sqrt{\frac{\log p}{n}}\right).$$

This result shows that the rates of the estimation error of the profile Lasso are not affected by the estimation of the partial residuals. Heuristically, this can be explained as follows: The contribution of the nonparametric estimation to the effective noise in the Lasso estimation is of the form

$$\frac{2}{n}\|(\hat{m}_k^*(T) - m(T))^\intercal \tilde{X}\|_\infty,$$

where $\hat{m}_k^*$ denotes the infeasible Nadaraya-Watson estimator of m defined above. Asymptotically, this term is of smaller order than $\frac{2}{n}\|\varepsilon^\intercal \tilde{X}\|_\infty$ and therefore any choice of λ which controls $\frac{2}{n}\|\varepsilon^\intercal \tilde{X}\|_\infty$ also controls the nonparametric error.

The proof of Theorem 16 relies on an in-sample version of the compatibility condition in Assumption 17(i) to hold. We justify this by showing that the in-sample compatibility condition converges to its population counterpart under the sparsity requirement in Assumption 16(v). If we would instead directly assume that this in-sample compatibility condition holds, Theorem 16 would hold true without any requirement on the sparsity of β_0.

The rates given in Theorem 16 are comparable to the results in the literature on

the HD PLM. Müller and van de Geer (2015) and Yu et al. (2019) study a doubly penalized least squares estimator which has besides an ℓ_1-penalty on β a smoothness penalty on m. Zhu (2017) use a two-stage procedure, where they estimate the univariate conditional means $\mathrm{E}[Y_i \,|\, T_i]$ and $\mathrm{E}[X_i \,|\, T_i]$ using nonparametric (penalized) least squares and in the second stage, apply Lasso to the partial residuals $\tilde{Y}_i, \tilde{X}_i$ as an estimator for β_0. Ma and Huang (2016) use a similar estimator to our approach with the difference that they estimate the partial residuals using splines instead of the Nadaraya-Watson estimator. All these papers find that the rates of the prediction and estimation error coincide with the usual rates obtained in the Lasso literature.

3.3.2 Properties of the Orthogonalized Nadaraya-Watson Estimator

In order to analyze the orthogonalized kernels, we introduce a population counterpart of $\hat{\gamma}_{t_\ell}$. Let γ_{t_ℓ}, $\ell = 1, \ldots, L$, be given by

$$\gamma_{t_\ell} \in \operatorname*{argmin}_{\gamma \in \mathbb{R}^p} \mathrm{E}[\mathbb{1}(|T_i - t_\ell| \leq h)(1 - X_i^\intercal \gamma)^2]. \tag{3.11}$$

In addition to Assumptions 16 and 17, we impose:

Assumption 18. *(i) The conditional density of T_i given X_i, $f_{T|X}$, satisfies $f_{T|X=x} \in \mathcal{H}(\eta_{f_{T|X}}, M_{f_{T|X}})$ for some constants $\eta_{f_{T|X}} \in (2, 3]$ and $M_{f_{T|X}} < \infty$ uniformly over x. There exists a constant $c > 0$ such that $f_{T|X}(t, x) \geq c$ for all t and x.*

(ii) The number of locations grows at most polynomially in n, i.e., $L = O(n^{\vartheta_L})$ for some ϑ_L and $\mathcal{T}_L \subseteq [h, 1-h]$.

(iii) $\{\gamma_{t_\ell} : \ell = 1, \ldots, L\}$ satisfies the growth-requirement $\max_{\ell=1,\ldots,L} \|\gamma_{t_\ell}\|_1 = o\left(\sqrt{\frac{nh}{\log^9 Lp}}\right)$.

(iv) The penalty μ satisfies $\mu = C_\mu \sqrt{\log L_n p_n / (nh_n)}$ for some sufficiently large constant C_μ.

(v) The bandwidth h satisfies $nh \to \infty$ and $nh^5 = O(1)$.

(vi) The constants $\eta_{f_{T|X}}, M_{f_{T|X}}, c, \vartheta_L$ and C_μ are independent of n.

The assumptions on the conditional density are rather mild and restrict the dependence between T and X. The lower bound on the conditional density may be justified by restricting the support of X and only considering values of X for which T varies sufficiently. The lower bound implies together with Assumption 16(ii) that the matrices $\Sigma_{n,\ell} := \mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell) X_i X_i^\intercal]$, $\ell = 1, \ldots, L_n$, are positive definite and therefore γ_{t_ℓ} identified. Further, it implies that the orthogonalized kernels can be bounded by the properties of the rectangular kernel. More precisely, the lower bound on the conditional density implies that there exists a constant $C \in (0, 1)$ such that

$$C \,\mathrm{E}[\mathbb{1}(|T_i - t_\ell| \leq h)] \leq \mathrm{E}[\mathbb{1}(|T_i - t_\ell| \leq h)(1 - X_i^\intercal \gamma_{t_\ell})] \leq \mathrm{E}[\mathbb{1}(|T_i - t_\ell| \leq h)]. \tag{3.12}$$

This property is very useful both for the analysis of the smoothing bias and the variance of the orthogonalized Nadaraya-Watson estimator. We could replace the assumption on the lower bound by bounds on the eigenvalues of $\Sigma_{n,\ell}$ and $\mathrm{E}[X_i X_i^\intercal]$ and by assuming the lower bound in (3.12) instead.

The required growth conditions on μ, h and L are rather mild. The restrictions on the bandwidth allow for example for the optimal rate $h \propto n^{-1/5}$ and undersmoothing choices satisfying $h \ll n^{-1/5}$. The choice of the penalty term μ is imposed to guarantee that the effective noise in the computation of the orthogonalization parameters $\hat{\gamma}_{t_\ell}$ is under control. The restriction on the number of locations is very mild and should not be binding in any practical application. We restrict the locations to lie in $[h, 1 - h]$ in order to circumvent boundary issues of the Nadaraya-Watson estimator.

The rate requirements on the size of γ_{t_ℓ} are rather abstract. These rates are only slightly stronger than what would be needed for prediction consistency of the Lasso on a sample with nh observations without assuming any type of compatibility condition. The rate requirements on γ_{t_ℓ} are for example satisfied when there is only a relatively small subset $S_\gamma \subset \{1, \ldots, p\}$ of regressors which affect T in the sense that $(T, X_{S_\gamma}) \perp\!\!\!\perp X_{S_\gamma^c}$. This independence implies that the γ_{t_ℓ} are sparse for all ℓ and their number of non-zero elements is controlled by the cardinality of S_γ. Alternatively, the rate requirements on γ_{t_ℓ} are implied by a nonparametric weak dependence assumption as in Assumption (A6) in Lu et al. (2020): Suppose there

exists a sufficiently small constant c such that

$$\sum_{j=1}^{p} \|f_{T,X_j} - f_T f_{X_j}\|_2 \leq c \quad \text{and} \quad \max_{1\leq k\leq p} \sum_{j=1}^{p} \|f_{T,X_j,X_k} - f_T f_{X_j} f_{X_k}\|_2 \leq c,$$

where f_{T,X_j,X_k}, f_{T,X_j} and f_{X_j} denote the joint (marginal) densities of T, X_{ij} and X_{ik} for $j,k = 1,\ldots,p$. This assumption implies that $\|\gamma_{t_\ell}\|_1 = O(1)$ uniformly over ℓ. Heuristically, this can be seen as follows. For each ℓ, $\gamma_{t_\ell} = \Sigma_\ell^{-1} \mathrm{E}[\mathbb{1}(|T_i - t_\ell| \leq h)X_i]$. The nonparametric weak dependence assumption allows comparing the ℓ_1-norms of $\mathrm{E}[\mathbb{1}(|T_i - t_\ell| \leq h)X_i]$ and Σ_ℓ^{-1} to the corresponding terms which would result when all X_j and T would be independent. In particular, it implies that the ℓ_1-norms of $\mathrm{E}[\mathbb{1}(|T_i - t_\ell| \leq h)X_i]$ and Σ_ℓ^{-1} are uniformly bounded.

Given these assumptions, we prove in the Appendix:

Theorem 17. *Suppose that Assumptions 16 - 18 hold. Then, for any $\ell = 1,\ldots,L$*

$$\sqrt{nh}\frac{\hat{m}_n(t_\ell) - m(t_\ell) - h^2 B_n(t_\ell)}{\sigma_{n,\ell}} \xrightarrow{d} \mathcal{N}(0,1),$$

where $\sigma_n(t_\ell)$ and $B_n(t_\ell)$ are given by

$$\begin{aligned}\sigma_{n,\ell}^2 =& \frac{\sigma^2}{2f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})} \\ B_n(t_\ell) =& \frac{1}{3}m''(t_\ell) + \frac{1}{3}m'(t_\ell)\frac{\frac{\partial f_T(t_\ell)(1-m_X(t_\ell)^\intercal\gamma_{t_\ell})}{\partial t}}{f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})}.\end{aligned}$$

Theorem 17 shows that the orthogonalized Nadaraya-Watson estimator behaves asymptotically comparable to the infeasible Nadaraya-Watson estimator which knows the true value of β_0. The leading term of the bias of $\hat{m}_N$ is captured by B_n and resembles the usual bias for the Nadaraya-Watson estimator up to the orthogonalization term $(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})$. Similarly, the asymptotic variance of $\hat{m}_n$ differs from the asymptotic variance of the infeasible estimator only by the same factor. The lower bound on the conditional density of T_i given X_i implies that this term is bounded away from zero and infinity uniformly over p and therefore both the bias and variance are of the same order of magnitude as the corresponding moments of the infeasible estimator. The orthogonalization factor $(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})$ can be interpreted as a price to pay for the reduction of the HD bias. This price reduces to zero when X_i and T_i are independent.

Theorem 17 implies that the orthogonalized Nadaraya-Watson estimator converges at the familiar rate

$$\hat{m}_n(t_\ell) - m(t_\ell) = O_p\bigg(\frac{1}{\sqrt{nh}} + h^2\bigg).$$

This rate coincides with other proposals in the literature on estimation in the HD PLM. As discussed in the literature review, our approach relies on weaker sparsity assumptions than the other proposals. This can be attributed to the orthogonalization of the kernels and the resulting reduction of the HD bias.

3.3.3 Asymptotic Distribution of the Test Statistic

Regarding the asymptotic distribution of the test statistic $\hat{T}_n$, we have

Theorem 18. *Let $\mathcal{I}_\ell = \{t : |t - t_\ell| \leq h\}$, $\ell = 1, \dots, L$, and $Z_1, \dots, Z_n$ be independent random vectors in $\mathbb{R}^L$ with $Z_i \sim \mathcal{N}\big(0, V\big)$, where*

$$V_{\ell k} = \frac{\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell \cap \mathcal{I}_k)\{1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}\}\{1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_k\}]}{\sqrt{\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)\{1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}\}^2]\,\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_k)\{1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_k\}^2]}}$$

for $\ell, k = 1, \dots, L$. Under $H_0 : m = 0$ and Assumptions 16 - 18, we have

$$\sup_{q \in \mathbb{R}} \bigg| \mathrm{P}(\hat{T}_n \leq q) - \mathrm{P}\bigg(\max_{\ell = 1, \dots, L} \bigg| \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_{i\ell} \bigg| \leq q \bigg) \bigg| = o(1).$$

The Theorem states that we can approximate the distribution of $\hat{T}_n$ by the distribution of a maximum of a multivariate normal distribution and can be thought of as an extension of Theorem 17 to a uniform distributional approximation. The core argument behind the theorem can be explained through the following decomposition of the test statistic: For $\ell = 1, \dots, L$, it holds that $\hat{\mathcal{S}}_{n,\ell} = \hat{\mathcal{S}}_{m,\ell} + \hat{\mathcal{S}}_{X,\ell} + \hat{\mathcal{S}}_{\varepsilon,\ell}$ with

$$\hat{\mathcal{S}}_{m,\ell} = \sqrt{nh} \frac{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell}) \left(m(T_i) - \frac{1}{n}\sum_{i=1}^n m(T_i)\right)}{\hat{\sigma}_\ell \sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})} \tag{3.13}$$

$$\hat{\mathcal{S}}_{X,\ell} = \sqrt{nh} \frac{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell}) \dot{X}_i^\intercal (\beta - \hat{\beta}_\lambda)}{\hat{\sigma}_\ell \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})} \tag{3.14}$$

$$\hat{\mathcal{S}}_{\varepsilon,\ell} = \sqrt{nh} \frac{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell}) \left(\varepsilon_i - \frac{1}{n}\sum_{i=1}^n \varepsilon_i\right)}{\hat{\sigma}_\ell \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})}. \tag{3.15}$$

$\hat{\mathcal{S}}_{m,\ell}$ is the only term that depends on m and is the term that drives the power of the test statistic. Under the null hypothesis, this term is zero and hence not important for Theorem 18. We will come back to this term in section 3.5 when we discuss the power properties of the test.

$\hat{\mathcal{S}}_{\varepsilon,\ell}$ governs the asymptotic distribution of the test statistic. We can show that

$$\hat{\mathcal{S}}_{\varepsilon,\ell} = \frac{1}{\sqrt{nh}\sigma_\ell} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)\{1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}\}\varepsilon_i + o_p(1)$$

uniformly in ℓ. Here, σ_ℓ denotes a population counterpart of $\hat{\sigma}_\ell$. Using a high-dimensional Berry-Esseen inequality given in Chernozhukov et al. (2017), we can show that this term converges to the distribution described in Theorem 18.

$\hat{\mathcal{S}}_{X,\ell}$ is the term through which the test statistic is affected by the estimation error of $\hat{\beta}_\lambda$. We have introduced the auxiliary Lasso problem leading to $\hat{\gamma}_{\ell,\mu}$ in order for this term to converge to zero. The near orthogonality property (3.2) of $\hat{\gamma}_\ell$ implies that the numerator of $\hat{\mathcal{S}}_{X,\ell}$ is less or equal to $\sqrt{nh}\mu\|\hat{\beta}_\lambda - \beta_0\|_1$. This implies together with the rate requirement on μ that $\hat{\mathcal{S}}_{X,\ell}$ converges to zero whenever the profile Lasso converges slightly faster than logarithmically to β_0. Without the orthogonalization parameters, this bias term would be of order $\sqrt{nh}\|\hat{\beta}_\lambda - \beta_0\|_1$ and one needs much stronger assumptions on the sparsity of β_0 for this term to converge to zero.

3.3.4 Consistency of the Bootstrap

As the quantiles of the asymptotic distribution given in Theorem 18 depend on the unknown covariance matrix V, we use the multiplier bootstrap described earlier to estimate the critical values. The next theorem shows that the bootstrap is consistent and therefore implies consistency of the estimated critical values.

Theorem 19. *Under the same assumptions as in Theorem 18, it holds*

$$\sup_{q\in\mathbb{R}} \big|\mathrm{P}^*(\hat{T}_n^* \leq q) - \mathrm{P}\big(\max_{\ell=1,\dots,L} |Z_\ell| \leq q\big)\big| = o_p(1),$$

where P^* *denotes the probability measure conditional on* $\{(Y_i, T_i, X_i) : 1 \leq i \leq n\}$.

Note that in the construction of the $\hat{\mathcal{S}}_{n,\ell}^*$, we have replaced the residuals $\hat{R}_i$ by $\hat{\varepsilon}_i$. This is in order to guarantee that the bootstrapped distribution is consistent also under the alternative. While the multipliers ensure that the test statistic is correctly

centered, the variance of the bootstrapped distribution increases in the size of m. While this should not affect the size of the resulting test, it should decrease its power.

3.3.5 Power Properties

In this section, we show that our test is uniformly consistent for testing the global null over a local smoothness class under strengthened conditions on the set of locations as well as the bandwidth h. In Addition, we derive an asymptotic power function for testing the local hypotheses $H_{0,\ell}$.

As for our alternatives, we consider a subset of the Hölder class $\mathcal{H}(R,\xi)$ which shrinks to zero. For this purpose, let $r_n \to 0$ such that $\sqrt{nh/\log L}r_n \to \infty$ and define

$$\mathcal{M}(R,\xi,r_n) = \big\{m \in \mathcal{H}(R,\xi) : \sup_{t\in[h,1-h]} |m(t)| = r_n\big\}.$$

This collection of alternatives is more general than considering Pitman alternatives of the form $m_n = r_n m$ for some fixed functions $m \in \mathcal{H}(R,\xi)$ with $\|m\|_\infty = 1$. Such Pitman sequences also imply that the size of m_n converges to zero but furthermore imply that the derivatives of m also converge to zero. Our considered collection of alternatives on the other hand allow the alternative to only shrink in size and keep the size of the derivatives constant.

Theorem 20. *Suppose the locations are chosen as a uniform grid of $[h,1-h]$ with $t_1 = h < t_2 < \cdots < t_L = 1-h$ and $L^2 r_n \to \infty$ as well as $nh^5 = o(1)$. Under Assumptions 16 - 18, we have*

$$\lim_{n\to\infty} \inf_{m\in\mathcal{M}(R,\xi,r_n)} \mathrm{P}(\hat{T}_n > \hat{q}_{1-\alpha}) = 1$$

for all $\alpha \in (0,1)$, where $\sqrt{nh/\log L}r_n \to \infty$.

This Theorem shows that our test is uniformly consistent against any m in $\mathcal{M}(R,\xi,r_n)$. This is a stronger statement than consistency against Pitman alternatives of the form $m_n(t) = r_n m(t)$ for some function $m \in \mathcal{M}(R,\xi,1)$. Heuristically, Pitman alternatives are only allowed to change in size, while uniform consistency allows the alternatives also to change in shape. For example, a Pitman alternative would allow for $m_n(t) = r_n \cos(ft)$ for some fixed frequency f. $\mathcal{M}(R,\xi,r_n)$, on the

other hand, allows the frequency to increase with n so that the alternative becomes more volatile as the sample size increases.

The requirements on the set of locations and the bandwidth can be explained using the decomposition of the $\hat{\mathcal{S}}_{n,\ell}$ in (3.13). As mentioned before, the term that drives the power is $\hat{\mathcal{S}}_{m,\ell}$. For the latter, we can show that uniformly in ℓ

$$\hat{\mathcal{S}}_{m,\ell} = \sqrt{nh}\,\frac{m(t_\ell)}{\sigma_{n,\ell}} + O_p\big(\sqrt{nh^5}\big) \tag{3.16}$$

where $\sigma_{n,\ell}$ was given in Theorem 17. The undersmoothing assumption $nh^5 = o(1)$ together with the results on the asymptotic distribution of $\hat{T}_n$ under H_0, imply that the test statistic can be bounded from below by

$$\hat{T}_n \geq \sqrt{nh} \max_{\ell=1,\dots,L} \frac{|m(t_\ell)|}{\sigma_{n,\ell}} - O_p\big(\sqrt{\log L}\big).$$

We choose a uniform grid for the locations t_ℓ in order to guarantee that $\max_\ell |m(t_\ell)|$ is close to $\sup_t |m(t)|$. For this purpose, the choice of the locations plays two roles: First, choosing a uniform grid ensures that any location can be approximated by a point in the grid. Secondly, $L^2 r_n \to \infty$, assures that the approximation error converges to zero sufficiently fast.

Additionally, we derive asymptotic power functions for tests of the local hypotheses $H_{0,\ell}$:

Theorem 21. *Suppose that the assumptions of Theorem 20 hold. Then, it holds for all $\ell = 1, \dots, L$ and $\alpha \in (0,1)$*

$$\begin{aligned}
\mathrm{P}\big(\hat{\mathcal{S}}_{n,\ell} > \hat{q}_{1-\alpha}\big) &= 1 - \Phi\left(q_{1-\alpha} - \sqrt{nh}\frac{m(t_\ell)}{\sigma_{n,\ell}}\right) + o(1) \\
\mathrm{P}\big(\hat{\mathcal{S}}_{n,\ell} < -\hat{q}_{1-\alpha}\big) &= \Phi\left(-q_{1-\alpha} - \sqrt{nh}\frac{m(t_\ell)}{\sigma_{n,\ell}}\right) + o(1)
\end{aligned}$$

where $q_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of the asymptotic distribution specified in Theorem 18.

By standard arguments on the distribution of the maximum of multivariate Gaussian random variables, one can show $q_{1-\alpha} = O(\sqrt{\log L})$. Moreover, by Assumption 18, σ_ℓ can be bounded from above and below by a positive constant. Thus, the asymptotic power functions in Theorem 21 converge to one whenever

$m(t_\ell)\sqrt{nh/\log L} \to \infty$ or $m(t_\ell)\sqrt{nh/\log L} \to -\infty$ respectively.

3.4 Numerical Results

In this section we present the results of a small Monte Carlo study to illustrate the finite-sample performance of our proposed test. Besides our proposed test, we implement an infeasible test based on a Nadaraya-Watson estimator which knows the true β and a Nadaraya-Watson estimator which does not deal with the HD bias. We compare the procedures according to their size and power properties.

3.4.1 Data-Generating Process

Our data-generating process follows the model

$$Y_i = m(T_i) + X_i^\intercal \beta_0 + \varepsilon_i.$$

The X_i are generated as follows: Let V_{ij} be i.i.d. uniformly distributed random variables on $[-1, 1]$ independent of W_i which themselves are i.i.d. uniform on $[-1, 1]$. Further, let $\theta > 0$ and set

$$X_{ij} = \frac{1}{1+\theta} V_{ij} + \frac{\theta}{1+\theta} W_i$$

These X_i have mean zero and are equicorrelated with correlation given by $\theta^2/(1+\theta^2)$. Given the vector X_i, we generate T_i according to

$$T_i = \frac{\exp(X_i^\intercal \tau_0 + \eta_i)}{1 + \exp(X_i^\intercal \tau_0 + \eta_i)}$$

where $\eta_i \sim \mathcal{N}(0, 1)$ is independent of X_i. Further, we choose $\varepsilon_i \sim \mathcal{N}(0, 1)$ independent of (T_i, X_i) and $m(t) = \cos(2\pi t) - \mathrm{E}[\cos(2\pi T_i)]$. β_0 (τ_0) is sparse with the first s_1 (s_2) elements non-zero which all take the value $\bar{\beta}$ ($\bar{\tau}$) and the remaining elements all zero.

Based on this model, we simulate 6 DGPs which use the parameter constellations as described in Table 3.4.1. These DGPs vary in their sample size, number of covariates and sparsity both in β_0 and τ_0. We choose $\bar{\beta}$ and $\bar{\tau}$ in order to ensure that the signal-to-noise ratio $\mathrm{Var}(X_i^\intercal \beta_0)/\mathrm{Var}(\varepsilon) \approx \mathrm{Var}(X_i^\intercal \tau_0)/\mathrm{Var}(\eta) \approx 1$. This serves two purposes. First, the constant signal-to-noise ratio in τ_0 ensures that the

DGP	n	p	s_1	s_2	$\bar{\beta}$	$\bar{\tau}$
DGP1	100	10	5	5	0.658	0.658
DGP2	200	100	5	5	0.658	0.658
DGP3	200	100	10	10	0.353	0.353
DGP4	500	200	5	5	0.658	0.658
DGP5	500	200	10	10	0.353	0.353
DGP6	500	200	20	20	0.182	0.182

Table 3.4.1: Definition of the different DGPs.

marginal distribution of T_i does not change too much as the sparsity of τ_0 varies and secondly the constant signal-to-noise ratio in β_0 ensures that the signal strength of the nonlinear part remains comparable to the linear part across different DGPs. Moreover, we choose $\theta = 1$ across all DGPs which results in a correlation of 0.5 within the covariates X_i. The locations are chosen as a uniform grid ranging from h to $1 - h$ with $L = 100$ locations.

We have chosen this model in order to investigate how the orthogonalized Nadaraya-Watson estimator performs for a DGP with relatively strong dependence between T_i and X_i. Due to the common component W_i in the construction of X_i, T_i depends on all the X_{ij}. Further, the constant signal-to-noise ratio in τ_0 ensures that the dependence between T_i and X_{ij}, $j = 1, \dots, p$, remains comparable across the different DGPs. Together with the symmetry across j, this implies that the γ_{t_ℓ} are quickly growing in p and do not satisfy the rate requirements for large p. Alternatively, a DGP which satisfies the growth condition on the γ_{t_ℓ} can be obtained by generating the first s_2 X_{ij} independently of the remaining X_{ij} so that T and the first s_2 X_{ij} are jointly independent of the remaining X_{ij}. In this alternative DGP, the size of the γ_{t_ℓ} would be directly related to s_2.

3.4.2 Implementation Details

We estimate the partial residuals using the Epanechnikov kernel for k_{int} and for $k_{low}^{(q)}$ we use $k_{low}^{(q)}(x) = (a(q) + b(q)x)k_{int}(x)$ where a and be are given by

$$a(q) = \left(\frac{3}{4}(1+q) - \frac{1+q^3}{20} - \frac{\left(\frac{3}{8}(1-q^2) - \frac{3}{80}(1-q^4)\right)^2}{\frac{1}{4}(1+q^3) - \frac{3}{100}(1+q^5)} \right)^{-1}$$
$$b(q) = -\frac{\frac{3}{8}(1-q^2) - \frac{3}{80}(1-q^4)}{\frac{1}{4}(1+q^3) - \frac{3}{100}(1+q^5)} a(q).$$

Finally, for $k_{up}^{(-q)}$ we use $k_{up}^{(-q)}(x) = (a(q) - b(q)x)k_{int}(x)$ for the same functions a and b. We show in Section B.2 in the Supplementary Material that this kernel satisfies Assumption 17. The bandwidth g is chosen for each partial residual individually based on leave-one-out cross-validation. Further, the penalty of the profile Lasso is chosen by 10-fold cross-validation.

The biased estimator applies the Nadaraya-Watson estimator based on the rectangular kernel to the residuals $R_i = \dot{Y}_i - \dot{X}_i^\intercal \hat{\beta}_\lambda$ without using the orthogonalized kernel. We choose its bandwidth using leave-one-out cross-validation and multiply it by a factor $n^{1/5-2/7}$ in order to ensure undersmoothing. Similarly, we implement the infeasible estimator as a Nadaraya-Watson estimator based on the rectangular kernel applied to $m(T_i) + \varepsilon_i$. As for the biased estimator, we choose the bandwidth of the infeasible estimator using leave-one-out cross-validation with the same undersmoothing factor. We compute the critical values of both methods as for the orthogonalized Nadaraya-Watson estimator using a Gaussian multiplier bootstrap with adapted choices for the standard errors and 1000 repetitions.

We implement the orthogonalization parameters $\hat{\gamma}_{t_\ell}$ using four different methods. While a theoretical investigation of the tuning parameter choice is interesting, it is out of the scope of the current paper. The following implementations are therefore not guided by asymptotic arguments but instead should be interpreted as experimental approaches. Our first approach, denoted by μ_{CV}, chooses the penalty μ for each location t_ℓ separately using 10-fold cross-validation resulting in a curve of estimated penalty parameters $\hat{\mu}_{t_\ell,CV}$, $\ell = 1, \dots, L$. In our simulations, the curve $\hat{\mu}_{t_\ell,CV}$ varied considerably across locations and was highly non-smooth. On the contrary, both the in-sample and the population problem should vary rather smoothly across locations. The other three methods are therefore based on a smoothed version of the cross-validated μ curve. For all of these, we first smooth the cross-validated μ using a moving-average with bandwidth 5 and multiply the resulting penalty term by a shrinkage factor $\kappa \in \{0.25, 0.5, 1\}$. The shrinkage factor is applied to improve control over the HD bias. In the following, we denote these smoothed and shrunken methods by $\mu_{\mathrm{MA},\kappa}$. For all of these four methods, we select the same bandwidth which results from the biased Nadaraya-Watson estimator and implement the bootstrap using 1000 repetitions.

DGP		μ_{CV}	$\mu_{\text{MA},1}$	$\mu_{\text{MA},0.5}$	$\mu_{\text{MA},0.25}$	infeasible	biased
DGP1	size	0.044	0.034	0.028	0.026	0.032	0.128
	power	0.908	0.85	0.802	0.732	0.968	0.984
DGP2	size	0.084	0.046	0.028	0.03	0.03	0.30
	power	0.988	0.95	0.922	0.908	0.996	1
DGP3	size	0.09	0.056	0.056	0.058	0.03	0.312
	power	0.982	0.946	0.94	0.93	1	1
DGP4	size	0.064	0.054	0.046	0.044	0.03	0.33
	power	1	0.998	1	0.996	1	1
DGP5	size	0.076	0.054	0.052	0.048	0.02	0.384
	power	1	0.998	0.996	0.992	1	1
DGP6	size	0.062	0.05	0.046	0.034	0.022	0.35
	power	1	0.998	0.998	0.996	1	1

Table 3.4.2: Results of Monte Carlo experiments for 500 independent repetitions.

3.4.3 Simulation Results

Our results are displayed in Table 3.4.2. The debiased estimator is severely oversized across all DGPs with size estimates ranging between 12.8% and 38.4% while the infeasible estimator has a size in the range of 2%-3.2%. This can be attributed to the HD bias since the only difference between the two estimator is the estimation of the high-dimensional part of the model. In comparison, the orthogonalized Nadaraya-Watson estimators can effectively reduce the size problem over all the studied DGPs. The cross-validated orthogonalized Nadaraya-Watson estimator μ_{CV} is the only orthogonalized method which also faces size problems with size between 6.2% and 9% in DGPs 2-6. The smoothed and shrunken methods $\mu_{\text{MA},\kappa}$ on the other hand all approximately control size with slight overcoverage for $\kappa = 1$ and in DGP 3. Moreover, the smaller the shrinkage factor κ, the smaller the size. This improvement of the orthogonalization in size control is accompanied by a reduction in power. As for the size control, the reduction in power is larger the smaller the value of the shrinkage factor κ and ranges from only a mild power reduction for μ_{CV} and $\mu_{\text{MA},1}$ to a reduction of up to 24% for $\mu_{\text{MA},0.25}$. Thus, the penalty choice for the orthogonalization parameters introduces a trade-off between size and power and needs further investigation.

The effect of the HD bias on the test statistics can also be seen in the local power and size properties. Figures 3.4.1 and 3.4.2 display the size and power estimates of

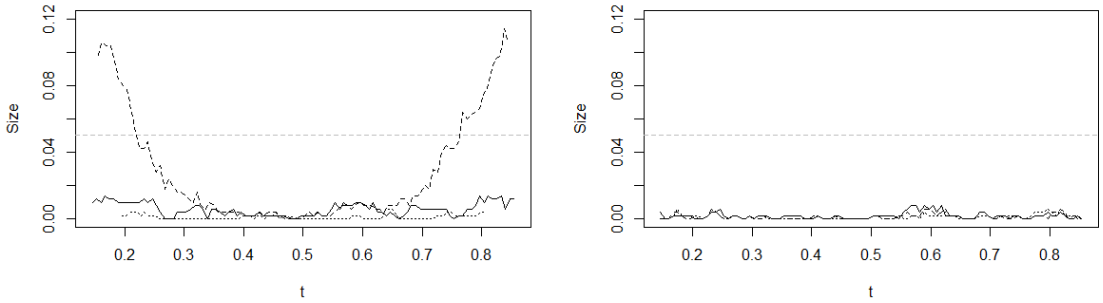

Figure 3.4.1: Monte Carlo results on local size. In the left panel, the line corresponds to μ_{CV}, the dotted line to the infeasible estimator and the dashed line to the biased estimator. In the right panel, the line corresponds to $\mu_{\text{MA},1}$, the dotted line to $\mu_{\text{MA},0.5}$ and the crossed line to $\mu_{\text{MA},0.25}$.

the studied methods for DGP3. It can be seen in the size of the biased estimator that it mostly rejects for locations close to the boundary, while the size of the infeasible estimator and the orthogonalized kernel methods is not altered at the boundary. This suggests that the HD bias is larger at the boundary than at the center. This is also supported by the power estimates. While the biased estimator has similar power properties near the boundary as in the center, the orthogonalized kernel methods possess almost no power at near the boundary (except for the cross-validated tuning parameter choice).

3.5 Discussion

In this paper, we study estimation and inference in the high-dimensional partially linear model $Y = \delta + m(T) + X^\intercal\beta + \varepsilon$, where m is a smooth unknown function measuring the potentially nonlinear effect of T on Y and β a sparse unknown regression parameter. In contrast to the classical partially linear model, the dimension of the covariates X is allowed to increase with the sample size and in particular is allowed to be larger than the sample size. We propose a Lasso-type estimator of β which attains the same rates an infeasible Lasso estimator which knows the unknown conditional mean functions $\mathrm{E}[Y_i|T_i]$ and $\mathrm{E}[X_i|T_i]$. Further, we show that ad-hoc estimators of m might be biased due to the estimation of the high-dimensional parameter β and In order to deal with this high-dimensional bias, we propose an orthogonalized Nadaraya-Watson estimator of m which is based on a

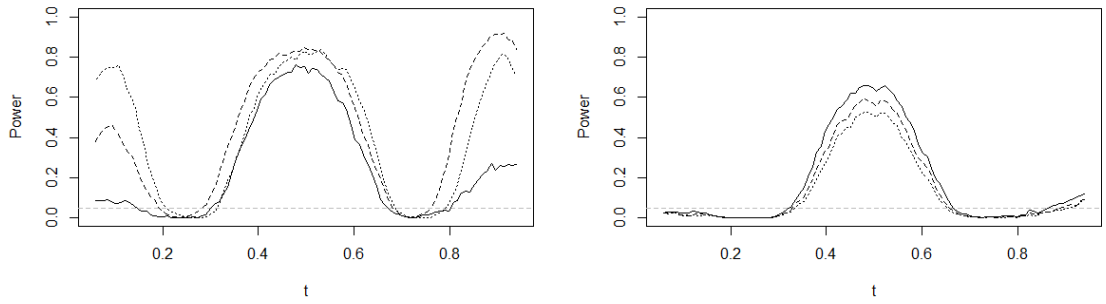

Figure 3.4.2: Monte Carlo results on local power. In the left panel, the line corresponds to μ_{CV}, the dotted line to the infeasible estimator and the dashed line to the biased estimator. In the right panel, the line corresponds to $\mu_{\text{MA},1}$, the dotted line to $\mu_{\text{MA},0.5}$ and the crossed line to $\mu_{\text{MA},0.25}$.

near-orthogonalization of the kernel with respect to the influence of the covariates X and show that this orthogonalization effectively decreases this high-dimensional bias. We further show that this orthogonalization does not affect the rate of the orthogonalized Nadaraya-Watson estimator, and in particular the orthogonalized Nadaraya-Watson estimator is shown to converge at the same rates as an infeasible Nadaraya-Watson estimator which knows the true values of δ and β.

We propose a test for the hypothesis that T has no effect on Y, i.e. $m = 0$, based on the orthogonalized Nadaraya-Watson estimator. This test generalizes the idea of significance testing in linear models by allowing for general smooth nonlinear effects of T on Y. We derive a distributional approximation of our test statistic and propose a consistent multiplier bootstrap to set its critical values. The proposed test is shown to be uniformly consistent against local Hölder balls. We study the finite sample performance of our proposed test in a simulation study and demonstrate its good debiasing and power properties. Within this simulation study, we implement our test statistic using different data-adaptive tuning parameter choices and find reasonable performance of these methods.

Appendix

3.A Proofs of the Main Theorems

This Appendix contains the proofs of the main theorems. Further technical results can be found in the Supplementary Material.

3.A.1 Results on the Profile Lasso

The proof of Theorem 16 can be easily seen from the following three lemmas and is omitted for brevity.

Lemma 55. *Suppose the compatibility condition*

$$0 < \phi^2 = \min\left\{ sb^\intercal \frac{1}{n}\tilde{X}^\intercal \tilde{X} b \,:\, \|b_S\|_1 = 1,\, \|b_{S^c}\|_1 \leq 3 \right\} \tag{3.17}$$

where $S = \{j : \beta_{j,0} \neq 0\}$ *is satisfied. Further, let*

$$\mathcal{T} = \left\{ \frac{2}{n} \|(\hat{m}_k^*(T) - m(T) + \varepsilon)^\intercal \tilde{X}\|_\infty \leq \lambda_0 \right\}$$

Then, on $\mathcal{T}$ *and for* $\lambda \geq 2\lambda_0$*, it holds*

$$\frac{1}{n}\|\tilde{X}(\hat{\beta}_\lambda - \beta_0)\|_2^2 + \lambda \|\hat{\beta}_\lambda - \beta_0\|_1 \leq 4\frac{s\lambda^2}{\phi^2}.$$

Proof. The proof is only a slight adaption of the corresponding arguments for the Lasso as in Bühlmann and van de Geer (2011) chapter 6. The only part that changes is the derivation of the basic inequality and the effective noise term given in the definition of the event $\mathcal{T}$. For the basic inequality, note that by definition of

$\hat{\beta}_\lambda$ it holds

$$\frac{1}{n}\|\tilde{Y} - \tilde{X}\hat{\beta}_\lambda\|_2^2 + \lambda\|\hat{\beta}_\lambda\|_1 \leq \frac{1}{n}\|\tilde{Y} - \tilde{X}\beta_0\|_2^2 + \lambda\|\beta_0\|_1.$$

Further, we have

$$\begin{aligned}\|\tilde{Y} - \tilde{X}\hat{\beta}_\lambda\|_2^2 =& \|\tilde{X}(\beta_0 - \hat{\beta}_\lambda)\|_2^2 + 2\{m(T) - \hat{m}_k^*(T) + \varepsilon\}^\intercal \tilde{X}(\beta_0 - \hat{\beta}_\lambda) \\ &+ \|m(T) - \hat{m}_k^*(T) + \varepsilon\|_2^2\end{aligned}$$

as well as $\|\tilde{Y} - \tilde{X}\beta_0\|_2^2 = \|m(T) - \hat{m}_k^*(T) + \varepsilon\|_2^2$. Combination of the last results yields together with Hölder's inequality the basic inequality

$$\begin{aligned}\frac{1}{n}\|\tilde{X}(\beta_0 - \hat{\beta}_\lambda)\|_2^2 + \lambda\|\hat{\beta}_\lambda\|_1 &\leq \frac{2}{n}\{m(T) - \hat{m}_k^*(T) + \varepsilon\}^\intercal \tilde{X}(\hat{\beta}_\lambda - \beta_0) + \lambda\|\beta_0\|_1 \\ &\leq \frac{2}{n}\|\{m(T) - \hat{m}_k^*(T) + \varepsilon\}^\intercal \tilde{X}\|_\infty \|\hat{\beta}_\lambda - \beta_0\|_1 + \lambda\|\beta_0\|_1.\end{aligned}$$

On $\mathcal{T}$ and for $\lambda \geq 2\lambda_0$, the basic inequality implies

$$\begin{aligned}\frac{1}{n}\|\tilde{X}(\hat{\beta}_\lambda - \beta_0)\|_2^2 \leq& \frac{2}{n}\|\tilde{X}^\intercal \varepsilon\|_\infty \|\hat{\beta}_\lambda - \beta_0\|_1 + \lambda(\|\beta_0\|_1 - \|\hat{\beta}_\lambda\|_1) \\ \leq& \frac{\lambda}{2}\|\hat{\beta}_\lambda - \beta_0\|_1 + \lambda(\|\beta\|_1 - \|\hat{\beta}_\lambda\|_1)\end{aligned}$$

Writing $\delta = \hat{\beta}_\lambda - \beta_0$ and letting δ_A denote the vector with entries δ_j for $j \in A$, we further get

$$\frac{1}{n}\|\tilde{X}\delta\|_2^2 \leq \frac{\lambda}{2}\|\delta_S\|_1 + \frac{\lambda}{2}\|\delta_{S^c}\|_1 + \lambda\|\delta_S\|_1 + \underbrace{\lambda\|\hat{\beta}_{\lambda,S}\|_1 - \lambda\|\hat{\beta}_\lambda\|_1}_{=-\lambda\|\delta_{S^c}\|_1}$$

that is,

$$\frac{1}{n}\|\tilde{X}\delta\|_2^2 \leq \frac{3}{2}\lambda\|\delta_S\|_1 - \frac{\lambda}{2}\|\delta_{S^c}\|_1. \tag{3.18}$$

This implies

$$\begin{aligned}\frac{2}{n}\|\tilde{X}\delta\|_2^2 + \lambda\|\delta\|_1 =& \frac{2}{n}\|\tilde{X}\delta\|_2^2 + \lambda\|\delta_S\|_1 + \lambda\|\delta_{S^c}\|_1 \\ \leq& 3\lambda\|\delta_S\|_1 - \lambda\|\delta_{S^c}\|_1 + \lambda\|\delta_S\|_1 + \lambda\|\delta_{S^c}\|_1 = 4\lambda\|\delta_S\|_1.\end{aligned}$$

Note that (3.18) implies $\|\delta_{S^c}\|_1 \leq 3\|\delta_S\|_1$. We can thus use the compatibility condition (3.17) to obtain that

$$\begin{aligned} 4\lambda\|\delta_S\|_1 &\leq 4\frac{\sqrt{s}\lambda}{\phi}\frac{1}{\sqrt{n}}\|\tilde{X}\delta\|_2 \\ &\leq \frac{1}{n}\|\tilde{X}\delta\|_2^2 + 4\frac{s\lambda^2}{\phi^2}, \end{aligned}$$

where the last inequality uses $4ab \leq a^2 + 4b^2$. The result follows. □

The following lemmas give probability bounds on the event $\mathcal{T}$ as well as that the compatibility condition (3.17) holds.

Lemma 56. *Suppose that Assumptions 16 and 17 hold. Then,* $\mathrm{P}(\mathcal{T}) \to 1$ *as* $n \to \infty$ *for*

$$\lambda_0 = 2K_\varepsilon\sqrt{\frac{\log 2p}{n}}.$$

Proof. By the triangle inequality,

$$\frac{2}{n}\|(\hat{m}_k^*(T) - m(T) + \varepsilon)^\intercal\tilde{X}\|_\infty \leq \frac{2}{n}\|\varepsilon^\intercal\tilde{X}\|_\infty + \frac{2}{n}\|(\hat{m}_k^*(T) - m(T))^\intercal\tilde{X}\|_\infty$$

By using standard arguments, one can show for the first term that

$$\lim_{n\to\infty}\mathrm{P}\left(\frac{2}{n}\|\varepsilon^\intercal\tilde{X}\|_\infty \leq \lambda_0\right) = 1.$$

For the second term, we have

$$\begin{aligned} \frac{1}{n}\|(\hat{m}_k^*(T) - m(T))^\intercal\tilde{X}\|_\infty \leq &\frac{1}{n}\|(\hat{m}_k^*(T) - m(T))^\intercal u\|_\infty \\ &+ \frac{1}{n}\|(\hat{m}_k^*(T) - m(T))^\intercal(\hat{m}_X(T) - m_X(T))\|_\infty, \end{aligned}$$

where we have used that $\tilde{X}_{ij} = u_{ij} - (m_{X,j}(T_i) - \hat{m}_{X,j}(T_i))$. We show in Lemma 64 in the Supplementary Material that

$$\frac{1}{n}\|(\hat{m}_k^*(T) - m(T))^\intercal u\|_\infty = O_p\left(\sqrt{\frac{\log p}{n}}\left\{\frac{1}{\sqrt{ng}} + g^2\right\}\right).$$

Regarding the second term on the right-hand side, we have by the Cauchy-Schwarz

inequality

$$\begin{aligned}&\frac{1}{n}\|(\hat{m}_k^*(T)-m(T))^\intercal(\hat{m}_X(T)-m_X(T))\|_\infty\\ \leq&\frac{1}{\sqrt{n}}\|\hat{m}_k^*(T)-m(T)\|_2\max_{j=1,\dots p}\frac{1}{\sqrt{n}}\|\hat{m}_X(T)-m_X(T)\|_2\end{aligned}$$

We show in Lemmas 61 and 65 in the Supplementary Material

$$\begin{aligned}\frac{1}{n}\|\hat{m}_k^*(T)-m(T)\|_2^2&=O_p\left(\frac{1}{ng}+g^4\right)\\ \max_{j=1,\dots p}\frac{1}{n}\|\hat{m}_X(T)-m_X(T)\|_2^2&=O_p\left(\frac{\sqrt{\log p}}{ng}+g^4\right).\end{aligned}$$

Hence,

$$\frac{1}{n}\|(\hat{m}_k^*(T)-m(T))^\intercal\tilde{X}\|_\infty=o_p\left(\sqrt{\frac{\log p}{n}}\right)$$

and the claim follows. □

The next result shows that the compatibility condition in (3.17) holds at least asymptotically.

Lemma 57. *Suppose Assumptions 16 and 17 hold. Then*

$$\begin{aligned}&\min\left\{sb^\intercal\frac{\tilde{X}^\intercal\tilde{X}}{n}b\,:\,\|b_S\|_1=1,\|b_{S^c}\|_1\leq 6\right\}\\ \geq&\min\left\{sb^\intercal\,\mathrm{E}[u_iu_i^\intercal]b\,:\,\|b_S\|_1=1,\|b_{S^c}\|_1\leq 6\right\}-\varrho_n,\end{aligned}$$

where

$$\varrho_n=O_p\left(s\frac{\log(pn)}{ng}+sg^4+s\sqrt{\frac{\log p}{n}}\right).$$

Proof. Denote by $\mathcal{B}$ the set of feasible vectors b, that is,

$$\mathcal{B}=\{b\in\mathbb{R}^p:\|b_S\|_2=1,\,\|b_{S^c}\|_1\leq 6\}$$

We can bound the compatibility condition (3.17) from below by

$$\begin{aligned}\min_{b\in\mathcal{B}} s\frac{1}{n}\sum_{i=1}^{n}(\tilde{X}_i^\intercal b)^2 \geq \min_{b\in\mathcal{B}} s\,\mathrm{E}[(u_i^\intercal b)^2] \\ - \max_{b\in\mathcal{B}} s\left|\frac{1}{n}\sum_{i=1}^{n}(u_i^\intercal b)^2 - \mathrm{E}[(u_i^\intercal b)^2]\right| \qquad (3.19)\\ - \max_{b\in\mathcal{B}} s\frac{1}{n}\sum_{i=1}^{n}\big(\{\hat{m}_X(T_i) - m_X(T_i)\}^\intercal b\big)^2 \qquad (3.20)\\ - \max_{b\in\mathcal{B}} s\left|\frac{2}{n}\sum_{i=1}^{n} b^\intercal u_i\{\hat{m}_X(T_i) - m_X(T_i)\}^\intercal b\right|, \qquad (3.21)\end{aligned}$$

where $\hat{m}_X = (\hat{m}_{X,1}, \hat{m}_{X,2}, \ldots, \hat{m}_{X,p})^\intercal$ and $\hat{m}_{X,j}$ is the Nadaraya-Watson estimator of $m_{X,j}$ given by

$$\hat{m}_{X,j}(t) = \frac{\sum_{i=1}^n k_g(T_i,t)X_{ij}}{\sum_{i=1}^n k_g(T_i,t)}.$$

We can bound (3.19) by

$$\left|\frac{1}{n}\sum_{i=1}^{n}(u_i^\intercal b)^2 - \mathrm{E}[(u_i^\intercal b)^2]\right| \leq \left\|\frac{1}{n}\sum_{i=1}^{n} u_i u_i^\intercal - \mathrm{E}[u_i u_i^\intercal]\right\|_\infty \|b\|_1^2 = O_p\left(\sqrt{\frac{\log p}{n}}\right),$$

where we have used that $\|b\|_1 \leq 7$ for all $b \in \mathcal{B}$ and the probability bound follows by applying Hoeffding's inequality together with a union bound. Hence,

$$\max_{b\in B} s\left|\frac{1}{n}\sum_{i=1}^{n}(u_i^\intercal b) - \mathrm{E}[(u_i^\intercal b)^2]\right| = O_p\left(s\sqrt{\frac{\log p}{n}}\right).$$

Regarding (3.20), we can bound

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\big(\{\hat{m}_X(T_i) - m_X(T_i)\}^\intercal b\big)^2\\ \leq&\frac{2}{n}\sum_{i=1}^{n}\left(\left\{\frac{\frac{1}{ng}\sum_{j=1}^n k_g(T_j,T_i)X_i}{f_T(T_i)} - m_X(T_i)\right\}^\intercal b\right)^2 \qquad (3.22)\\ &+\frac{2}{n}\sum_{i=1}^{n}\left(\frac{1}{\hat{f}_T(T_i)} - \frac{1}{f_T(T_i)}\right)^2\left(\left\{\frac{1}{ng}\sum_{j=1}^n k_g(T_j,T_i)X_i\right\}^\intercal b\right)^2 \qquad (3.23)\end{aligned}$$

where $\hat{f}_T$ denotes the kernel density estimator

$$\hat{f}_T(t) = \frac{1}{ng} \sum_{i=1}^{n} k_g(T_i, t).$$

For (3.23), it holds

$$\begin{aligned}
&\frac{2}{n} \sum_{i=1}^{n} \left(\frac{1}{\hat{f}_T(T_i)} - \frac{1}{f_T(T_i)} \right)^2 \left(\left\{ \frac{1}{ng} \sum_{j=1}^{n} k_g(T_j, T_i) X_i \right\}^\intercal b \right)^2 \\
\leq &\left(\sup_{t \in [0,1]} \left| \frac{1}{\hat{f}_T(t)} - \frac{1}{f_T(t)} \right| \right)^2 \frac{2}{n} \sum_{i=1}^{n} \left(\left\{ \frac{1}{ng} \sum_{j=1}^{n} k_g(T_j - T_i) X_j \right\}^\intercal b \right)^2
\end{aligned}$$

By Lemma 60 in the Supplementary Material and boundedness of f_T it holds that

$$\sup_{t \in [0,1]} \left| \frac{1}{\hat{f}_T(t)} - \frac{1}{f_T(t)} \right| = O_p\left(\sqrt{\frac{\log n}{ng}} + g^2 \right) \tag{3.24}$$

The second factor on the right-hand side is bounded in probability, which can be seen as follows

$$\begin{aligned}
&\frac{2}{n} \sum_{i=1}^{n} \left(\left\{ \frac{1}{ng} \sum_{j=1}^{n} k_g(T_j, T_i) X_j \right\}^\intercal b \right)^2 \\
\leq &\frac{2}{n} \sum_{i=1}^{n} \Bigg(\underbrace{\left\| \frac{1}{ng} \sum_{j=1}^{n} k_g(T_j, T_i) X_j \right\|_\infty}_{\leq 2\hat{f}_T(T_i)} \underbrace{\|b\|_1}_{\leq 7} \Bigg)^2 \\
\leq &\underbrace{\frac{784}{n} \sum_{i=1}^{n} \left(\hat{f}_T(T_i) - f_T(T_i) \right)^2}_{=o_p(1)} + \underbrace{\frac{784}{n} \sum_{i=1}^{n} f_T(T_i)^2}_{=O(1)} = O_p(1).
\end{aligned}$$

This implies for (3.23)

$$\frac{2}{n} \sum_{i=1}^{n} \left(\frac{1}{\hat{f}_T(T_i)} - \frac{1}{f_T(T_i)} \right)^2 \left(\left\{ \frac{1}{ng} \sum_{j=1}^{n} k_g(T_j, T_i) X_i \right\}^\intercal b \right)^2 = O_p\left(\frac{\log n}{ng} + g^4 \right).$$

Turning to (3.22), we have

$$\begin{aligned}
&\frac{2}{n}\sum_{i=1}^{n}\left(\left\{\frac{\frac{1}{ng}\sum_{j=1}^{n}k_g(T_j,T_i)X_i}{f_T(T_i)}-m_X(T_i)\right\}^\intercal b\right)^2\\
\leq&\frac{1}{\min_{t\in[0,1]}f_T(t)^2}\frac{2}{n}\sum_{i=1}^{n}\left(\left\{\frac{1}{ng}\sum_{j=1}^{n}k_g(T_j,T_i)X_i-m_X(T_i)f_T(T_i)\right\}^\intercal b\right)^2,
\end{aligned}$$

and as f_T is bounded away from zero by Assumption 16, we can focus on the second factor on the right-hand side. Decompose this term as follows

$$\begin{aligned}
&\frac{2}{n}\sum_{i=1}^{n}\left(\left\{\frac{1}{ng}\sum_{j=1}^{n}k_g(T_j,T_i)X_i-m_X(T_i)f_T(T_i)\right\}^\intercal b\right)^2\\
\leq&\frac{6}{n}\sum_{i=1}^{n}\left(\left\{\frac{1}{ng}\sum_{j=1}^{n}k_g(T_j,T_i)u_i\right\}^\intercal b\right)^2 &(3.25)\\
&+\frac{6}{n}\sum_{i=1}^{n}\left(\left\{\frac{1}{ng}\sum_{j=1}^{n}k_g(T_j,T_i)m_X(T_j)-\frac{1}{g}\operatorname{E}[k_g(T_j,T_i)m_X(T_j)\,|\,T_i]\right\}^\intercal b\right)^2 &(3.26)\\
&+\frac{6}{n}\sum_{i=1}^{n}\left(\left\{\frac{1}{g}\operatorname{E}[k_g(T_j,T_i)m_X(T_j)\,|\,T_i]-m_X(T_i)f_T(T_i)\right\}^\intercal b\right)^2, &(3.27)
\end{aligned}$$

where we have used that $(a+b+c)^2\leq 3a^2+3b^2+3c^2$ for any real numbers a,b and c.

For the last term on the right side, i.e., (3.27):

$$\begin{aligned}
&\frac{6}{n}\sum_{i=1}^{n}\left(\left\{\frac{1}{g}\operatorname{E}[k_g(T_j,T_i)m_X(T_j)\,|\,T_i]-m_X(T_i)f_T(T_i)\right\}^\intercal b\right)^2\\
\leq&6\left(\max_{j=1,\dots,p}\sup_{t\in[0,1]}\left|\frac{1}{g}\operatorname{E}[k_g(T_i-t)m_X(T_i)]-m_X(t)f_T(t)\right|\right)^2\|b\|_1^2
\end{aligned}$$

By the usual arguments in the kernel regression literature, one can show

$$\begin{aligned}
&\max_{j=1,\dots,p}\sup_{t\in[0,1]}\left|\frac{1}{g}\operatorname{E}[k_g(T_i-t)m_X(T_i)]-m_X(t)f_T(t)\right|\\
\leq&\frac{g^2}{2}\Bigg(\sup_{t\in[0,1]}|f_T(t)|\max_{\ell=1,\dots,p}\sup_{t\in[0,1]}|m''_{X.\ell}(t)|\\
&\quad+\sup_{t\in[0,1]}|f''_t(t)|\max_{\ell=1,\dots,p}\sup_{t\in[0,1]}|m_{X,j}(t)|
\end{aligned}$$

$$+ \sup_{t\in[0,1]} |f_T'(t)| \max_{\ell=1,\ldots,p} \sup_{t\in[0,1]} |m_{X,\ell}'(t)|\Big) \int k(v)v^2 dv + o(g^2) = O(g^2)$$

implying for (3.27)

$$\max_{b\in\mathcal{B}} \frac{6}{n} \sum_{i=1}^{n} \left(\left\{ \frac{1}{g} \mathrm{E}[k_g(T_j, T_i) m_X(T_j)] - m_X(T_i) f_T(T_i) \right\}^\intercal b \right)^2 = O(g^4).$$

Regarding (3.26), we can argue analogously

$$\begin{aligned}
&\frac{6}{n} \sum_{i=1}^{n} \left(\left\{ \frac{1}{ng} \sum_{j=1}^{n} k_g(T_j, T_i) m_X(T_j) - \frac{1}{g} \mathrm{E}[k_g(T_j, T_i) m_X(T_j) \,|\, T_i] \right\}^\intercal b \right)^2 \\
\leq &\frac{6}{n} \sum_{i=1}^{n} \left(\left\| \frac{1}{ng} \sum_{j=1}^{n} k_g(T_j, T_i) m_X(T_j) - \frac{1}{g} \mathrm{E}[k_g(T_j, T_i) m_X(T_j) \,|\, T_i] \right\|_\infty \underbrace{\|b\|_1}_{\leq 7} \right)^2 \\
\leq &294 \left(\max_{j=1,\ldots,p} \sup_{t\in[0,1]} \left| \frac{1}{ng} \sum_{i=1}^{n} k_g(T_i - t) m_{X,j}(T_i) - \frac{1}{g} \mathrm{E}[k_g(T_i - t) m_{X,j}(T_i)] \right| \right)^2.
\end{aligned}$$

By Lemma 62 in the Supplementary Material, we have

$$\max_{j=1,\ldots,p} \sup_{t\in[0,1]} \left| \frac{1}{ng} \sum_{i=1}^{n} k_g(T_i, t) m_{X,j}(T_i) - \frac{1}{g} \mathrm{E}[k_g(T_i, t) m_{X,j}(T_i)] \right| = O_p\left(\sqrt{\frac{\log(pn)}{ng}} \right)$$

implying

$$\begin{aligned}
&\max_{b\in\mathcal{B}} s \frac{6}{n} \sum_{i=1}^{n} \left(\left\{ \frac{1}{ng} \sum_{j=1}^{n} k_g(T_j, T_i) m_X(T_j) - \frac{1}{g} \mathrm{E}[k_g(T_j, T_i) m_X(T_j) \,|\, T_i] \right\}^\intercal b \right)^2 \\
= &O_p\left(s \frac{\log(pn)}{ng} \right).
\end{aligned}$$

For (3.25),

$$\frac{6}{n} \sum_{i=1}^{n} \left(\left\{ \frac{1}{ng} \sum_{j=1}^{n} k_g(T_j, T_i) u_i \right\}^\intercal b \right)^2 \leq 294 \left(\max_{j=1,\ldots,p} \sup_{t\in[0,1]} \left| \frac{1}{ng} \sum_{i=1}^{n} k_g(T_i, t) u_{ij} \right| \right)^2$$

and using standard kernel regression arguments, we show in Lemma 63 in the Sup-

plementary Material that

$$\max_{j=1,\dots,p} \sup_{t\in[0,1]} \left| \frac{1}{ng} \sum_{i=1}^{n} k_g(T_i, t) u_{ij} \right| = O_p\left(\sqrt{\frac{\log(pn)}{ng}} \right).$$

Hence, (3.25) satisfies

$$\max_{b\in\mathcal{B}} s\frac{6}{n} \sum_{i=1}^{n} \left(\left\{ \frac{1}{ng} \sum_{j=1}^{n} k_g(T_j, T_i) u_i \right\}^\intercal b \right)^2 = O_p\left(s\frac{\log(np)}{ng} \right).$$

Finally, (3.21) is of smaller order than (3.19) or (3.20). In order to see this, decompose

$$\begin{aligned}
&\frac{2}{n} \sum_{i=1}^{n} b^\intercal u_i \{\hat{m}_X(T_i) - m_X(T_i)\}^\intercal b \\
=&\frac{2}{n} \sum_{i=1}^{n} b^\intercal u_i \frac{1}{f_T(T_i)} \frac{1}{ng} \sum_{j=1}^{n} k_g(T_j, T_i) u_j^\intercal b && (3.28) \\
&+ \frac{2}{n} \sum_{i=1}^{n} b^\intercal u_i \left(\frac{1}{\hat{f}_T(T_i)} - \frac{1}{f_T(T_i)} \right) \frac{1}{ng} \sum_{j=1}^{n} k_g(T_j, T_i) u_j^\intercal b && (3.29) \\
&+ \frac{2}{n} \sum_{i=1}^{n} b^\intercal u_i \frac{1}{f_T(T_i)} \left(\frac{1}{ng} \sum_{j=1}^{n} k_g(T_j, T_i) m_X(T_j) - m_X(T_i) f_T(T_i) \right)^\intercal b && (3.30) \\
&+ \frac{2}{n} \sum_{i=1}^{n} b^\intercal u_i \left(\frac{1}{\hat{f}_T(T_i)} - \frac{1}{f_T(T_i)} \right) \left(\frac{1}{ng} \sum_{j=1}^{n} k_g(T_j, T_i) m_X(T_j) - m_X(T_i) f_T(T_i) \right)^\intercal b && (3.31)
\end{aligned}$$

Observe that (3.29) and (3.31) are of smaller order than (3.28) and (3.30) by (3.24). For (3.28), we can bound

$$\begin{aligned}
&\left| \frac{2}{n} \sum_{i=1}^{n} b^\intercal u_i \frac{1}{f_T(T_i)} \frac{1}{ng} \sum_{j=1}^{n} k_g(T_j, T_i) u_j^\intercal b \right| \\
\leq &\left\| \frac{2}{n} \sum_{i=1}^{n} u_i \frac{1}{f_T(T_i)} \frac{1}{ng} \sum_{j=1}^{n} k_g(T_j, T_i) u_j^\intercal - 2\,\mathrm{E}\left[u_i \frac{1}{f_T(T_i)} \frac{1}{ng} \sum_{j=1}^{n} k_g(T_j, T_i) u_j^\intercal \,\middle|\, T \right] \right\|_\infty \|b\|_1^2 && (3.32)
\end{aligned}$$

$$+ \underbrace{\left\| 2\,\mathrm{E}\left[u_i \frac{1}{f_T(T_i)} \frac{1}{ng} \sum_{j=1}^{n} k_g(T_j, T_i) u_j^\intercal \,\middle|\, T \right] \right\|_\infty}_{=O(1/(ng))} \|b\|_1^2$$

We can deal with the first term on the right-hand side by the Hanson-Wright inequality in Theorem 1.1 in Rudelson and Vershynin (2013). Even though this Hanson-Wright inequality is not directly applicable to the off-diagonal elements of this matrix, a closer inspection of its proof reveals that the same result applies also to the off-diagonal elements at least if the random vectors have bounded entries. In particular, one can adapt the Theorem to:

Theorem 22. *Let $X_1 = (X_{11}, \ldots, X_{n1})^\intercal$ and $X_2 = (X_{12}, \ldots, X_{n2})^\intercal$ be random vectors with independent components X_{ij} which satisfy $\mathrm{E}[X_{ij}] = 0$ and $|X_{ij}| \leq K$. Let A be an $n \times n$ matrix. Then, for every $t \geq 0$,*

$$\mathrm{P}\big(|X_1^\intercal A X_2 - \mathrm{E}[X_1^\intercal A X_2]| > t\big) \leq 2 \exp\left(-c \min\left\{ \frac{t^2}{K^4 \|A\|_{HS}^2}, \frac{t}{K^2 \|A\|} \right\} \right),$$

where c is an absolute constant, $\|A\|_{HS}^2 = \sum_{i,j} a_{ij}^2$ denotes the Hilbert-Schmidt norm and $\|A\| = \max_{\|x\|_2=1} \|Ax\|_2$ denotes the operator norm.

Apply Theorem 22 to (3.32) conditionally on $\{T_i, i = 1, \ldots, n\}$. For this, take $X_1 = (u_{1,\ell}, \ldots, u_{n,\ell})^\intercal$, $X_2 = (u_{1,k}, \ldots, u_{n,k})^\intercal$) and let A be the matrix with entries $a_{i,j}$ given by

$$a_{ij} = \frac{1}{f_T(T_i)} \frac{1}{n^2 g} k_g(T_j, T_i).$$

Note that $\|A\|^2 \leq \|A\|_{HS}^2 = O(1/(ng)^2)$. Hence, together with a union bound, we obtain for (3.32)

$$\left\| \frac{2}{n} \sum_{i=1}^{n} u_i \frac{1}{f_T(T_i)} \frac{1}{ng} \sum_{j=1}^{n} k_g(T_j, T_i) u_j^\intercal - 2\,\mathrm{E}\left[u_i \frac{1}{f_T(T_i)} \frac{1}{ng} \sum_{j=1}^{n} k_g(T_j, T_i) u_j^\intercal \,\middle|\, T \right] \right\|_\infty$$
$$= O_p\left(\frac{\log p}{ng} \right)$$

For (3.30),

$$\begin{aligned}&\left|\frac{2}{n}\sum_{i=1}^{n} b^\intercal u_i \frac{1}{f_T(T_i)}\left(\frac{1}{ng}\sum_{j=1}^{n} k_g(T_j,T_i)m_X(T_j) - m_X(T_i)f_T(T_i)\right)^\intercal b\right| \\ \leq&\left\|\frac{2}{n}\sum_{i=1}^{n} u_i \frac{1}{f_T(T_i)}\left(\frac{1}{ng}\sum_{j=1}^{n} k_g(T_j,T_i)m_X(T_j) - m_X(T_i)f_T(T_i)\right)^\intercal\right\|_\infty \|b\|_1^2\end{aligned}$$

By Hoeffding's inequality conditional on $\{T_1,\dots,T_n\}$ and using that

$$\left\|\frac{1}{f_T(T_i)}\frac{1}{ng}\sum_{j=1}^{n} k_g(T_j,T_i)m_X(T_j) - m_X(T_i)f_T(T_i)\right\|_\infty^2 = O_p\left(\frac{\log p}{ng} + g^4\right)$$

one can show that

$$\begin{aligned}&\left\|\frac{2}{n}\sum_{i=1}^{n} u_i \frac{1}{f_T(T_i)}\left(\frac{1}{ng}\sum_{j=1}^{n} k_g(T_j,T_i)m_X(T_j) - m_X(T_i)f_T(T_i)\right)^\intercal\right\|_\infty \\ =&O_p\left(\frac{\sqrt{\log p}}{n\sqrt{g}} + \frac{g^2}{\sqrt{n}}\right).\end{aligned}$$

Hence, (3.30) is of smaller order than (3.19) and (3.20). This completes the proof. □

3.A.2 Results on the Orthogonalized Nadaraya-Watson Estimator

Proof of Theorem 17: Let $\ell = 1,\dots,L$ be arbitrary. Decompose the estimator as follows

$$\hat{m}_k^*(t_\ell) = \frac{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})\dot{m}(T_i)}{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})} \tag{3.33}$$

$$- \frac{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})\dot{X}_i^\intercal(\hat{\beta}_\lambda - \beta)}{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})} \tag{3.34}$$

$$+ \frac{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})\dot{\varepsilon}_i}{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})}. \tag{3.35}$$

Here, (3.33) captures the estimate of m, (3.34) the high-dimensional bias due to estimation of β and (3.35) captures the influence of the residual. Consider the first

term on the right-hand side. It holds

$$\begin{aligned}&\frac{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})\dot{m}(T_i)}{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})}\\ =&\frac{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})m(T_i)}{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})} - \frac{1}{n}\sum_{i=1}^n m(T_i)\end{aligned}$$

and the second term on the right-hand side converges at $\sqrt{n}$-rate and therefore is negligible in the following. We can deal with the first term using a second order Taylor expansion around t_ℓ, i.e., there exist intermediate values θ_i, $i = 1, \ldots, n$ such that

$$\begin{aligned}&\frac{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})m(T_i)}{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})}\\ =&m(t_\ell) + m'(t_\ell)\frac{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})(T_i - t_\ell)}{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})}\\ &+ \frac{m''(t_\ell)}{2}\frac{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})(T_i - t_\ell)^2}{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})}\\ &+ \frac{1}{2}\frac{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})\{m''(\theta_i) - m''(t_\ell)\}(T_i - t_\ell)^2}{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})}.\end{aligned}$$

In Lemma 75 below, we show

$$\begin{aligned}&\frac{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})(T_i - t_\ell)}{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})}\\ =&h^2\frac{1}{3}\frac{\frac{\partial f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})}{\partial t}}{f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})} + o_p\left(\frac{1}{\sqrt{nh}} + h^2\right)\end{aligned}$$

and by similar arguments one can show that

$$\begin{aligned}\frac{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})(T_i - t_\ell)^2}{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})} =&\frac{1}{3}h^2 + o_p\left(\frac{1}{\sqrt{nh}} + h^2\right)\\ \frac{1}{2}\frac{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})\{m''(\theta_i) - m''(t_\ell)\}(T_i - t_\ell)^2}{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})} =&o_p\left(\frac{1}{\sqrt{nh}} + h^2\right)\end{aligned}$$

and therefore

$$\frac{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell}) m(T_i)}{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})} = m(t_\ell) + h^2 B(t_\ell) + o(h^2).$$

Next, consider (3.34). By Lemma 77, the denominator is asymptotically bounded away from zero and hence it suffices to consider the numerator. For the latter, we have by the KKT conditions of the auxiliary Lasso

$$\left| \frac{1}{nh} \sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell}) \dot{X}_i^\intercal (\hat{\beta}_\lambda - \beta) \right| \leq \frac{\mu}{2} \|\hat{\beta}_\lambda - \beta\|_1 = o_p\left(\frac{1}{\sqrt{nh}}\right)$$

by Assumption 18 and Theorem 16.

Finally, (3.35) satisfies

$$\begin{aligned}
&\frac{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell}) \dot{\varepsilon}_i}{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})} \\
=&\frac{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell}) \varepsilon_i}{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})} + o_p\left(\frac{1}{\sqrt{n}}\right) \\
=&\frac{\frac{1}{nh}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell}) \varepsilon_i}{f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})}(1 + o_p(1)) + o_p\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}$$

where the second equality follows by Lemma 77 together with Lemma 74. Moreover, we show in Lemma 76 that

$$\frac{1}{nh} \sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell}) \varepsilon_i = \frac{1}{nh} \sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - X_i^\intercal \gamma_{t_\ell}) \varepsilon_i + o_p\left(\frac{1}{\sqrt{nh}} + h^2\right)$$

which implies

$$\begin{aligned}
&\frac{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell}) \dot{\varepsilon}_i}{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})} \\
=&\frac{\frac{1}{nh}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - X_i^\intercal \gamma_{t_\ell}) \varepsilon_i}{f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})}(1 + o_p(1)) + o_p\left(\frac{1}{\sqrt{nh}} + h^2\right)
\end{aligned}$$

where we have used that the denominator is bounded away from zero by Lemma 66 and 74. Thus, we can approximate (3.35) by a sum of i.i.d. mean zero random

variables with variance

$$\begin{aligned}\sigma_{n,\ell}^2 &= \sigma^2 \frac{\frac{1}{h}\,\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - X_i^\intercal \gamma_{t_\ell})^2]}{f_T(t_\ell)^2(1 - m_X(t)^\intercal \gamma_{t_\ell})^2} \\ &= \frac{\sigma^2}{f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})} + o(1)\end{aligned}$$

where we have again used Lemmas 66 and 74. Further, it holds

$$\begin{aligned}\frac{\mathrm{E}[|\frac{1}{h}\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - X_i^\intercal \gamma_{t_\ell})\varepsilon_i|^3]}{\sqrt{n}(\sigma_{n,\ell}/\sqrt{h})^3} \leq& \frac{C}{\sqrt{nh}} \frac{1}{h}\,\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)|1 - X_i^\intercal \gamma_{t_\ell}|^3 |\varepsilon_i|^3] \\ \leq& \frac{C}{\sqrt{nh}} \frac{1}{h}\,\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)|1 - X_i^\intercal \gamma_{t_\ell}|^3] \\ \leq& \frac{C(1 + \|\gamma_{t_\ell}\|_1)}{\sqrt{nh}} \frac{1}{h}\,\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)|1 - X_i^\intercal \gamma_{t_\ell}|^2] \to 0,\end{aligned}$$

where the first inequality follows by Lemmas 66 and 74, the second inequality as ε_i is conditionally sub-Gaussian and the last inequality by boundedness of X_i, Hölder's inequality and Lemma 66. This implies by the Berry-Esseen bound

$$\frac{\frac{1}{nh}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - X_i^\intercal \gamma_{t_\ell})\varepsilon_i}{\sigma_{n,\ell} f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})} \xrightarrow{d} \mathcal{N}(0,1)$$

which proves the claim. □

3.A.3 On the Asymptotic Distribution of $\hat{T}_n$

Proof of Theorem 18: Under H_0, we have, for $\ell = 1, \ldots, L$,

$$\begin{aligned}\hat{\mathcal{S}}_{n,\ell} &= \sqrt{nh} \frac{\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{\ell,\mu})\hat{R}_i}{\hat{\sigma}_\ell \sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{\ell,\mu})} \\ &= \mathcal{S}_{\varepsilon,\ell} + \hat{\mathcal{S}}_{X,\ell} + \{\hat{\mathcal{S}}_{\varepsilon,\ell} - \mathcal{S}_{\varepsilon,\ell}\}\end{aligned}$$

with $\hat{\mathcal{S}}_{X,\ell}, \hat{\mathcal{S}}_{\varepsilon,\ell}$ as defined in (3.14) and (3.15) and

$$\mathcal{S}_{\varepsilon,\ell} = \frac{1}{\sqrt{nh}\sigma_\ell} \sum_{i=1}^n \frac{\mathbb{1}(T_i \in \mathcal{I}_\ell)\{1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}\}\varepsilon_i}{f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})}.$$

We rewrite $\hat{T}_n$ as follows: Define $\hat{W}_i = (\hat{W}_{i,1}, \ldots, \hat{W}_{i,2L})^T$ with

$$\hat{W}_{i,2\ell-1} = \frac{1}{\sqrt{h}\hat{\sigma}_\ell} \frac{\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})\hat{R}_i}{\frac{1}{nh}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{\ell,\mu})}$$
$$\hat{W}_{i,2\ell} = -\frac{1}{\sqrt{h}\hat{\sigma}_\ell} \frac{\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})\hat{R}_i}{\frac{1}{nh}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{\ell,\mu})},$$

which allows us to write

$$\begin{aligned}\hat{T}_n &= \max_{\ell=1,\ldots,L} |\hat{\mathcal{S}}_{n,\ell}| \\ &= \max_{\ell=1,\ldots,2L} \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{W}_{i,\ell}.\end{aligned}$$

Consider the (unobserved) statistic

$$T_n = \max_{\ell=1,\ldots,L} |\mathcal{S}_{\varepsilon,\ell}|,$$

which can be rewritten as

$$T_n = \max_{\ell=1,\ldots,2L} \frac{1}{\sqrt{n}} \sum_{i=1}^n W_{i,\ell},$$

where $W_i = (W_{i,1}, \ldots, W_{i,2L})^T$ with

$$W_{i,2\ell-1} = \frac{1}{\sqrt{h}\sigma_\ell} \frac{\mathbb{1}(T_i \in \mathcal{I}_\ell)\{1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}\}\varepsilon_i}{f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})} \tag{3.36}$$
$$W_{i,2\ell} = -\frac{1}{\sqrt{h}\sigma_\ell} \frac{\mathbb{1}(T_i \in \mathcal{I}_\ell)\{1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}\}\varepsilon_i}{f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})}. \tag{3.37}$$

For any $q \in \mathbb{R}$,

$$\begin{aligned}\mathrm{P}(T_n \le q) &= \mathrm{P}\left(\max_{\ell=1,\ldots,2L} \frac{1}{\sqrt{n}} \sum_{i=1}^n W_{i,\ell} \le q\right) \\ &= \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n W_{i,\ell} \le q \,\forall \ell\right) \\ &= \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n W_i \le q\vec{1}\right),\end{aligned}$$

where $\vec{1} = (1, 1, \ldots, 1)^\intercal$ and $\leq$ in the last line is to be understood elementwise. We prove in the Supplementary Material using Prop 2.1 in Chernozhukov et al. (2017):

Lemma 58. *Let W_i be the random vectors defined in (3.36) and (3.37) and let $Z_1, \ldots, Z_n$ be independent random vectors in $\mathbb{R}^{2L}$ with $Z_i \sim \mathcal{N}\big(0, \mathrm{E}\big[W_i W_i^T\big]\big)$. Further, denote by $\mathcal{A}$ the class of all hyperrectangles A in $\mathbb{R}^{2L}$. Then, under Assumptions 16 - 18,*

$$\sup_{A \in \mathcal{A}} \left| \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i \in A \right) - \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \in A \right) \right| = o(1).$$

This high-dimensional CLT implies in particular that we can approximate the unobserved test statistic T_n by a corresponding Gaussian test statistic, i.e.,

$$\sup_{q \in \mathbb{R}} \Bigg| \underbrace{\mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i \leq q \right)}_{= \mathrm{P}(T_n \leq q)} - \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \leq q \right) \Bigg| = o(1).$$

It remains to bound the approximation of $\hat{T}_n$ by T_n. For this purpose, write

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{W}_i = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Delta_i.$$

We show in Lemma 80 in the Supplementary Material

$$\max_{\ell = 1, \ldots, 2L} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Delta_{i,\ell} \right| = o_p(\varrho_n) = o_p\left(\frac{1}{\log L} \right). \tag{3.38}$$

Thus, for any $q \in \mathbb{R}$,

$$\begin{aligned}
\mathrm{P}(\hat{T}_n \leq q) &= \mathrm{P}\left(\max_{\ell = 1, \ldots, 2L} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{W}_{i\ell} \leq q \right) \\
&= \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{W}_{i\ell} \leq q \, \forall \ell \right) \\
&= \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_{i\ell} \leq q - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Delta_{i\ell} \, \forall \ell \right) \\
&= \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i \leq q\vec{1} - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Delta_i \right)
\end{aligned}$$

$$\begin{cases} \leq \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i \leq (q + \varrho_n)\vec{1} \right) + o(1) \\ \geq \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i \leq (q - \varrho_n)\vec{1} \right) + o(1) \end{cases}$$

where the last inequalities follow by (3.38). This implies together with Lemma 58

$$\begin{aligned} &\Bigg| \underbrace{\mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{W}_i \leq q\vec{1} \right)}_{=\mathrm{P}(\hat{T}_n \leq q)} - \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \leq q\vec{1} \right) \Bigg| \\ \leq \max\Bigg\{ &\left| \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i \leq (q + \varrho_n)\vec{1} \right) - \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \leq q\vec{1} \right) \right|, \\ &\left| \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i \leq (q - \varrho_n)\vec{1} \right) - \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \leq q\vec{1} \right) \right| \Bigg\} + o(1) \end{aligned}$$

and

$$\begin{aligned} &\left| \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i \leq (q + \varrho_n)\vec{1} \right) - \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \leq q\vec{1} \right) \right| \\ \leq &\underbrace{\left| \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i \leq (q + \varrho_n)\vec{1} \right) - \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \leq (q + \varrho_n)\vec{1} \right) \right|}_{=o(1) \text{ by Lemma 58}} \\ &+ \left| \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \leq (q + \varrho_n)\vec{1} \right) - \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \leq q\vec{1} \right) \right|. \end{aligned}$$

The remaining term on the right-hand side can be dealt with using Nazarov's inequality (Chernozhukov et al. (2017) Lemma A.1):

Lemma (Nazarov's inequality)**.** *Let $V = (V_1, \ldots, V_p)^T$ be a centered Gaussian random vector in $\mathbb{R}^p$ with $\mathrm{E}[V_j^2] \geq c > 0$ for all j. Then for every $v \in \mathbb{R}^p$ and every $\phi > 0$,*

$$\mathrm{P}(V \leq v + \phi\vec{1}) - \mathrm{P}(V \leq v) \leq C\phi\sqrt{\log p}$$

with C only depending on c.

It implies

$$\left|\mathrm{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i \leq (q+\varrho_n)\vec{1}\right) - \mathrm{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i \leq q\vec{1}\right)\right| \leq C\varrho_n\sqrt{\log 2L} = o(1).$$

As a result,

$$\sup_{q\in\mathbb{R}}\Bigg|\underbrace{\mathrm{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{W}_i \leq q\vec{1}\right)}_{=\mathrm{P}(\hat{T}_n\leq q)} - \mathrm{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i \leq q\vec{1}\right)\Bigg| = o(1)$$

proving the claim. □

3.A.4 Consistency of the Bootstrap

Proof of Theorem 19: We can proceed similarly as in Theorem 18. Let $e_1, \ldots, e_n \sim \mathcal{N}(0,1)$ i.i.d. and independent of the data $\{(Y_i, T_i, X_i) : i = 1, \ldots, n\}$. Consider

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\tilde{W}_i - \tilde{\overline{W}})e_i = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(W_i - \overline{W})e_i + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\Delta_i - \bar{\Delta})e_i$$

with $\tilde{\overline{W}} = (\tilde{\overline{W}}_1, \ldots, \tilde{\overline{W}}_{2L})^T$ and $\tilde{\overline{W}}_\ell = \frac{1}{n}\sum_{i=1}^{n}\tilde{W}_{i\ell}$. We prove in Lemma 82 in the Supplementary Material

$$\mathrm{P}^*\left(\max_{\ell=1,\ldots,2L}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\Delta_{i\ell} - \bar{\Delta}_\ell)e_i\right| > \frac{1}{\log L}\right) = o_p(1), \tag{3.39}$$

Further, we show in Lemma 81 in the Supplementary Material by using Theorem 4.1 and Remark 4.1 in Chernozhukov et al. (2017)

$$\sup_{q\in\mathbb{R}}\left|\mathrm{P}^*\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(W_i - \overline{W})e_i \leq q\right) - \mathrm{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i \leq q\right)\right| = o_p(1). \tag{3.40}$$

Here P^* denotes the probability measure conditional on $(Y_1, , T_1, X_1), \ldots, (Y_n, T_n, X_n)$. Given these two ingredients and Nazarov's inequality, the proof follows as in Theorem 18. □

3.A.5 Power Properties

Proof of Theorem 20: We can decompose the test statistic $\hat{T}_n$ into

$$\hat{T}_n \geq \max_{\ell=1,\dots,L} |\hat{\mathcal{S}}_{m,\ell}| - \max_{\ell=1,\dots,L} |\hat{\mathcal{S}}_{n,\ell} - \hat{\mathcal{S}}_{m,\ell}|$$

where $\hat{\mathcal{S}}_{n,\ell}$ and $\hat{\mathcal{S}}_{m,\ell}$ where defined in (3.6) and (3.13). The second term on the right-hand side behaves as the test statistic under H_0, whereas the first term drives the power. In particular, we have for the second term

$$\max_{\ell=1,\dots,L} |\hat{\mathcal{S}}_{n,\ell} - \hat{\mathcal{S}}_{m,\ell}| = O_p\big(\sqrt{\log L}\big)$$

by Theorem 18 and standard arguments on the distribution of the maximum of multivariate Gaussian distributions. By similar arguments as in the proof of Theorem 17, one can show that

$$\max_{\ell=1,\dots,L} \left| \hat{\mathcal{S}}_{m,\ell} - \sqrt{nh} \frac{m(t_\ell)}{\sigma_\ell} \right| = O_p\big(\sqrt{nh^5}\big).$$

Moreover, by Assumptions 16 and 18, σ_ℓ is bounded away from zero uniformly over ℓ, i.e., there exists a constant $c > 0$ such that $\sigma_\ell \geq c$ for all ℓ. This implies, together with $nh^5 = o(1)$, that

$$\hat{T}_n \geq \frac{1}{c} \max_{\ell=1,\dots,L} \sqrt{nh} |m(t_\ell)| - O_p(\sqrt{\log L}).$$

It remains to assess the influence of the discretization of the support of T on the power of the test. One may suspect that the maximum of m is attained outside the grid t_ℓ, $\ell = 1, \dots, L$, and that m is substantially smaller on the grid. We will show in the following that this is no problem for a sufficiently closed-meshed grid. For this purpose, let t^* be such that $|m(t^*)| = r_n$. We assume without loss of generality that $m(t^*) > 0$ as the case in which $m(t^*) < 0$ can be dealt with by the same argument. Since $t^* \in [h, 1-h]$, necessarily $m'(t^*) = 0$ and by a Taylor expansion of $m(t)$ around t^*, there exists some θ such that

$$m(t) = m(t^*) + \frac{1}{2} m''(\theta)(t - t^*)^2$$

and by the Landau-Kolmogorov inequalities there exists a constant M such that for

any $m \in \mathcal{M}(R, \xi, r_n)$

$$|m(t) - m(t^*)| \leq M(t - t^*)^2.$$

In particular, there is an $\ell = 1, \ldots, L$ satisfying $|t_\ell - t^*| \leq \frac{1}{L}$ and therefore

$$|m(t_\ell) - m(t^*)| \leq \frac{M}{L^2}.$$

This implies

$$m(t_\ell) \geq m(t^*) - |m(t_\ell) - m(t^*)| \geq r_n - \frac{M}{L^2} = r_n(1 + o(1))$$

and hence

$$\hat{T}_n \geq \sqrt{nh} r_n \frac{1}{c} - O_p(\sqrt{\log L}).$$

Finally, as $\sqrt{nh/\log L} r_n \to \infty$, the claim follows. □

3.B Supplementary Material

We first prove some technical lemmas whose main purpose is to introduce a reference in order to avoid recurring arguments.

Lemma 59. *Let $\{Z_{ij}\colon i = 1, \dots, n, j = 1, \dots, J_n\}$, $J_n \to \infty$ as $n \to \infty$, be a collection of uniformly bounded random variables with mean zero and bound $|Z_{ij}| \leq M_n$ for some deterministic sequence M_n. Further, suppose that the Z_{ij} are independent across i.*

(i) If $M_n = O(1/\sqrt{\log J_n})$ and $\max_j \operatorname{Var}\left(\sum_{i=1}^n Z_{ij}\right) = O(1)$, then

$$\max_{j=1,\dots,J_n} \left|\sum_{i=1}^n Z_{ij}\right| = O_p\big(\sqrt{\log J_n}\big).$$

(ii) If $M_n = o(1/\sqrt{\log J_n})$ and $\max_j \operatorname{Var}\left(\sum_{i=1}^n Z_{ij}\right) = o(1)$, then

$$\max_{j=1,\dots,J_n} \left|\sum_{i=1}^n Z_{ij}\right| = o_p\big(\sqrt{\log J_n}\big).$$

Proof. By a union bound, we have, for $\eta > 0$,

$$\begin{aligned}\mathrm{P}\left(\max_{j=1,\dots,J_n} \left|\sum_{i=1}^n Z_{ij}\right| \geq \sqrt{\log J_n}\eta\right) &\leq \sum_{j=1}^{J_n} \mathrm{P}\left(\left|\sum_{i=1}^n Z_{ij}\right| \geq \sqrt{\log J_n}\eta\right) \\ &\leq J_n \max_{j=1,\dots,J_n} \mathrm{P}\left(\left|\sum_{i=1}^n Z_{ij}\right| \geq \sqrt{\log J_n}\eta\right).\end{aligned}$$

(i) As $M_n = O(1/\sqrt{\log J_n})$ and $\max_j \operatorname{Var}\left(\sum_{i=1}^n Z_{ij}\right) = O(1)$, there are constants C_1 and C_2 such that for sufficiently large n,

$$M_n \leq \frac{C_1}{\sqrt{\log J_n}} \quad \text{and} \quad \operatorname{Var}\left(\sum_{i=1}^n Z_{ij}\right) \leq C_2.$$

Thus, by Bernstein's inequality, it holds

$$\mathrm{P}\left(\left|\sum_{i=1}^n Z_{ij}\right| \geq \sqrt{\log J_n}\eta\right) \leq 2\exp\left(-\frac{1}{2}\frac{\log(J_n)\eta^2}{C_2 + C_1\eta/3}\right)$$

implying

$$\mathrm{P}\left(\max_{j=1,\dots,J_n}\left|\sum_{i=1}^{n} Z_{ij}\right| \geq \sqrt{\log J_n}\eta\right) \leq 2\exp\left(\log J_n - \frac{1}{2}\frac{\log(J_n)\eta^2}{C_2 + C_1\eta/3}\right).$$

For η sufficiently large, the exponent becomes negative and the right-hand side converges to zero as $J_n \to \infty$. The claim follows.

(ii) By Bernstein's inequality, it holds

$$\mathrm{P}\left(\left|\sum_{i=1}^{n} Z_{ij}\right| \geq \sqrt{\log J_n}\eta\right) \leq 2\exp\left(-\frac{1}{2}\frac{\log(J_n)\eta^2}{\mathrm{Var}\left(\sum_{i=1}^{n} Z_{ij}\right) + M_n\sqrt{\log J_n}\eta/3}\right)$$

implying

$$\begin{aligned}&\mathrm{P}\left(\max_{j=1,\dots,J_n}\left|\sum_{i=1}^{n} Z_{ij}\right| \geq \sqrt{\log J_n}\eta\right)\\ \leq& 2\exp\left(\log J_n - \frac{1}{2}\frac{\log(J_n)\eta^2}{\mathrm{Var}\left(\sum_{i=1}^{n} Z_{ij}\right) + M_n\sqrt{\log J_n}\eta/3}\right).\end{aligned}$$

As $M_n = o(1/\sqrt{\log J_n})$ and $\max_j \mathrm{Var}\left(\sum_{i=1}^{n} Z_{ij}\right) = o(1)$, the denominator converges to zero for any fixed η and hence the right-hand side converges to zero as $n \to \infty$.

□

As a simple Corollary, we include a result which deals with the case when the Z_{ij} do not have mean zero:

Corollary 12. *Let $\{Z_{ij}\colon i = 1,\dots,n, j = 1,\dots,J_n\}$, $J_n \to \infty$ as $n \to \infty$, be a collection of uniformly bounded random variables with bound $|Z_{ij}| \leq M_n$ for some deterministic sequence M_n. Further, suppose that the Z_{ij} are independent across i.*

If $M_n = O(1/\sqrt{\log J_n})$, $\max_j \mathrm{Var}\left(\sum_{i=1}^{n} Z_{ij}\right) = O(1)$ and $\max_{i,j}|\mathrm{E}[Z_{ij}]| = O(\rho_n)$, then

$$\max_{j=1,\dots,J_n}\left|\sum_{i=1}^{n} Z_{ij}\right| = O_p\big(\sqrt{\log J_n} \vee (n\rho_n)\big).$$

Proof. The claim is a straightforward consequence of the triangle inequality and

Lemma 59:

$$\max_{j=1,\dots,J_n}\left|\sum_{i=1}^{n} Z_{ij}\right| \leq \underbrace{\max_{j=1,\dots,J_n}\left|\sum_{i=1}^{n}(Z_{ij}-\mathrm{E}[Z_{ij}])\right|}_{=O_p(\sqrt{\log J_n})} + \underbrace{n \max_{\substack{i=1,\dots,n,\\ j=1,\dots,J_n}} |\mathrm{E}[Z_{ij}]|}_{=O(n\rho_n)}.$$

□

3.C Results on Boundary Corrected Kernels

In this section, we derive rates of several terms related to the Nadaraya-Watson estimator in the definition of the estimated partial residuals $\tilde{Y}_i$ and $\tilde{X}_i$. Even though the results seem pretty standard, they are not in that we employ boundary corrected kernels and need some results on the joint behavior of a vector of Nadaraya-Watson estimators. The first subsection is devoted to the rate results while in the second we construct the boundary corrected kernel used in our simulation. The idea underlying the construction of this kernel builds upon Gasser et al. (1985) but with the difference that we do not impose that the kernel needs to equal zero at the boundary of its support.

3.C.1 Technical Lemmas

Through all results of this section, we assume that Assumptions 16 - 18 hold.

Lemma 60. *It holds*

$$\sup_{t\in[-1,1]} |\hat{f}_T(t) - f_T(t)| = O_p\left(\sqrt{\frac{\log n}{ng}} + g^2\right).$$

Proof. Decompose into bias and variance part

$$\hat{f}_T(t) - f_T(t) = \hat{f}_T(t) - \frac{1}{g}\mathrm{E}[k_g(T_i,t)] + \frac{1}{g}\mathrm{E}[k_g(T_i,t)] - f_T(t).$$

For the bias part, we have, by the standard argument for second order interior kernels, for $t \in [g, 1-g]$

$$\frac{1}{g}\mathrm{E}[k_g(T_i - t)] = \frac{1}{g}\int_0^1 k\left(\frac{T_i - t}{g}\right) f_T(T_i) dT_i$$

$$
\begin{aligned}
&= \int_{-1}^{1} k(v) f_T(t+gv) dv \\
&= f_T(t) \int_{-1}^{1} k(v) dv + g f_T'(t) \int_{-1}^{1} v k(v) dv \\
&\quad + \frac{g^2}{2} f_T''(t) \int_{-1}^{1} v^2 k(v) dv + o(g^2) \\
&= f_T(t) + \frac{g^2}{2} f_T''(t) \int_{-1}^{1} v^2 k(v) dv + o(g^2)
\end{aligned}
$$

For $t \in [0, g]$ or $t \in [1-g, 1]$, the same argument applies due to the support constraints as well as the moment conditions of k_{low} and k_{up} and hence

$$
\sup_{t \in [0,1]} |\mathrm{E}[k_g(T_i - t)]/g - f_T(t)| \leq \frac{g^2}{2} \sup_{t \in [0,1]} |f_T''(t)| \left| \int_{-1}^{1} v^2 k(v) dv \right| + o(g^2).
$$

For the variance part, we show in the following

$$
\sup_{t \in [0,1]} \left| \frac{1}{ng} \sum_{i=1}^{n} (k_g(T_i, t) - \mathrm{E}[k_g(T_i, t)]) \right| = O_p\left(\sqrt{\frac{\log n}{ng}} \right).
$$

We first discretize the supremum. For that purpose, let $t_0 = 0$, $t_M = 1$ and $t_j = j/M$. For any $t \in [0,1]$, there is a $j(t)$ in $0, 1, \ldots, M$ such that $|t - j(t)| \leq 1/M$. Moreover, $\frac{1}{ng} \sum_{i=1}^{n} (k_g(T_i, t) - \mathrm{E}[k_g(T_i, t)])$ is Lipschitz continuous in t as for $s, t \in [0,1]$

$$
\begin{aligned}
&\left| \frac{1}{ng} \sum_{i=1}^{n} (k_g(T_i, s) - \mathrm{E}[k_g(T_i, s)]) - \frac{1}{ng} \sum_{i=1}^{n} (k_g(T_i, t) - \mathrm{E}[k_g(T_i, t)]) \right| \\
\leq &\left| \frac{1}{ng} \sum_{i=1}^{n} (k_g(T_i, s) - k_g(T_i, t) - \mathrm{E}[k_g(T_i, s) - k_g(T_i, t)]) \right| \\
\leq &\frac{1}{ng} \sum_{i=1}^{n} (|k_g(T_i, s) - k_g(T_i, t)| + \mathrm{E}[|k_g(T_i, s) - k_g(T_i, t)|]) \\
\leq &\frac{C}{g^2} |s - t|
\end{aligned}
$$

and therefore

$$
\sup_{t \in [0,1]} \left| \frac{1}{ng} \sum_{i=1}^{n} (k_g(T_i - t) - \mathrm{E}[k_g(T_i - t)]) \right|
$$

$$\begin{aligned}
&\leq \sup_{t\in[0,1]}\left|\frac{1}{ng}\sum_{i=1}^{n}(k_g(T_i-t)-\mathrm{E}[k_g(T_i-t)])\right| \\
&\quad+\max_{j=0,\dots,M}\left|\frac{1}{ng}\sum_{i=1}^{n}(k_g(T_i-t_j)-\mathrm{E}[k_g(T_i-t_j)])\right| \\
&\leq \max_{j=0,\dots,M}\left|\frac{1}{ng}\sum_{i=1}^{n}(k_g(T_i-t_j)-\mathrm{E}[k_g(T_i-t_j)])\right|+\frac{C}{Mg^2}.
\end{aligned}$$

Since the kernel is bounded, the first term on the right-hand side is a sum of mean-zero, independent and bounded random variables. Therefore, by Hoeffding's inequality together with a union bound,

$$\max_{j=0,\dots,M}\left|\frac{1}{ng}\sum_{i=1}^{n}(k_g(T_i-t_j)-\mathrm{E}[k_g(T_i-t_j)])\right| = O_p\left(\sqrt{\frac{\log M}{ng}}\right)$$

Taking $M = n^2$, the claim follows. □

Lemma 61. *It holds*

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{m}_k^*(T_i)-m(T_i))^2 = O_p\left(\frac{1}{ng}+g^4\right).$$

Proof. Let $\hat{m}_{k,-i}^*(t)$ denote the leave-one-out estimator which is defined by

$$\hat{m}_{k,-i}^*(T_i) = \frac{\sum_{j\neq i}k_g(T_j,T_i)(m(T_j)+\varepsilon_j)}{\sum_{j\neq i}k_g(T_j,T_i)}$$

Note that

$$\begin{aligned}
\hat{m}_k^*(T_i) =& \hat{m}_{k,-i}^*(T_i) - \frac{k(0)}{k(0)+\sum_{j\neq i}k_g(T_j,T_i)}(\hat{m}_{k,-i}^*(T_i)-m(T_i)) \\
&+\frac{k(0)}{k(0)+\sum_{j\neq i}k_g(T_j,T_i)}\varepsilon_i
\end{aligned}$$

This implies

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\{\hat{m}_k^*(T_i)-m(T_i)\}^2 \leq& \left(\frac{\sum_{j\neq i}k_g(T_j,T_i)}{k(0)+\sum_{j\neq i}k_g(T_j,T_i)}\right)^2\frac{2}{n}\sum_{i=1}^{n}\{\hat{m}_{k,-i}^*(T_i)-m(T_i)\}^2 \\
&+\left(\frac{k(0)}{k(0)+\sum_{j\neq i}k_g(T_j,T_i)}\right)\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i^2
\end{aligned}$$

By Lemma 60,

$$\left(\frac{\sum_{j\neq i} k_g(T_j,T_i)}{k(0)+\sum_{j\neq i} k_g(T_j,T_i)}\right)^2 = 1+o_p(1)$$
$$\left(\frac{k(0)}{k(0)+\sum_{j\neq i} k_g(T_j,T_i)}\right) = O_p\left(\frac{1}{ng}\right).$$

Hence,

$$\left(\frac{k(0)}{k(0)+\sum_{j\neq i} k_g(T_j,T_i)}\right)\frac{2}{n}\sum_{i=1}^n \varepsilon_i^2 = O_p\left(\frac{1}{ng}\right).$$

Next, we decompose

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \{\hat{m}^*_{k,-i}(T_i)-m(T_i)\}^2 \\
\leq &\frac{3}{n}\sum_{i=1}^n \left\{\frac{1}{\hat{f}_{T,-i}(T_i)}\frac{1}{ng}\sum_{j\neq i} k_g(T_j,T_i)\varepsilon_j\right\}^2 && (3.41)\\
&+\frac{3}{n}\sum_{i=1}^n \left\{\frac{1}{\hat{f}_{T,-i}(T_i)}\frac{1}{ng}\sum_{j\neq i} k_g(T_j,T_i)m(T_j)-\mathrm{E}[k_g(T_j,T_i)m(T_j)\,|\,T_i]\right\}^2 && (3.42)\\
&+\frac{3}{n}\sum_{i=1}^n \left\{\frac{1}{g\hat{f}_{T,-i}(T_i)}\mathrm{E}[k_g(T_j,T_i)m(T_j)\,|\,T_i]-m(T_i)\right\}^2 && (3.43)
\end{aligned}$$

Regarding (3.41),

$$\left\{\frac{1}{\hat{f}_{T,-i}(T_i)}\frac{1}{ng}\sum_{j\neq i} k_g(T_j,T_i)\varepsilon_j\right\}^2 = \frac{1}{\hat{f}_{T,-i}(T_i)^2}\frac{1}{(ng)^2}\sum_{j\neq i}\sum_{\ell\neq i} k_g(T_j,T_i)k_g(T_\ell,T_i)\varepsilon_j\varepsilon_\ell$$

the latter has conditional mean (conditionally on $T_1,\ldots,T_n$)

$$\begin{aligned}
&\frac{1}{\hat{f}_{T,-i}(T_i)^2}\frac{1}{(ng)^2}\sum_{j\neq i} k_g(T_j,T_i)^2\sigma^2 \\
\leq &\frac{\sigma^2}{ng}\frac{C}{\hat{f}_{T,-i}(T_i)} \\
\leq &\frac{\sigma^2}{ng}\frac{C}{\min_t f_T(t)-\sup_t|\hat{f}_{T,-i}(t)-f_T(t)|} = O_p\left(\frac{1}{ng}\right).
\end{aligned}$$

Thus, by Markov's inequality applied to the conditional distribution (again condi-

tional on $T_1, \ldots, T_n$)

$$\frac{3}{n}\sum_{i=1}^{n}\left\{\frac{1}{\hat{f}_{T,-i}(T_i)}\frac{1}{ng}\sum_{j\neq i}k_g(T_j,T_i)\varepsilon_j\right\}^2 = O_p\left(\frac{1}{ng}\right).$$

For (3.42), we first apply Lemma 60 in order to deal with the kernel density estimator

$$\begin{aligned}&\frac{3}{n}\sum_{i=1}^{n}\left\{\frac{1}{\hat{f}_{T,-i}(T_i)}\frac{1}{ng}\sum_{j\neq i}k_g(T_j,T_i)m(T_j) - \mathrm{E}[k_g(T_j,T_i)m(T_j)\,|\,T_i]\right\}^2\\ \leq&\frac{3}{n}\sum_{i=1}^{n}\left\{\frac{1}{\min_t f_T(t)}\frac{1}{ng}\sum_{j\neq i}k_g(T_j,T_i)m(T_j) - \mathrm{E}[k_g(T_j,T_i)m(T_j)\,|\,T_i]\right\}^2(1+o_p(1)).\end{aligned}$$

Let

$$\xi_{ij} := \frac{k_g(T_j,T_i)m(T_j) - \mathrm{E}[k_g(T_j,T_i)m(T_j)\,|\,T_i]}{g},$$

and note that the ξ_{ij} are i.i.d. across j for $j \neq i$ and with mean zero and variance bounded by

$$\mathrm{E}[\xi_{ij}^2] \leq \frac{1}{g}\int k(v)^2 dv \sup_{t\in[0,1]}|m(t)|^2.$$

Now, (3.42) can be written as

$$\begin{aligned}\frac{1}{\min_t f_T(t)}\frac{3}{n}\sum_{i=1}^{n}\left\{\frac{1}{n}\sum_{j\neq i}\xi_{ij}\right\}^2 =&\frac{1}{\min_t f_T(t)}\frac{3}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{\ell\neq i}\xi_{ij}\xi_{i\ell}\\ =&\frac{1}{\min_t f_T(t)}\left(\frac{3}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\xi_{ij}^2 + \frac{3}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{\ell\neq i,j}\xi_{ij}\xi_{i\ell}\right)\end{aligned}$$

The mean of the first term on the right-hand side satisfies

$$\frac{3}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\mathrm{E}\big[\xi_{ij}^2\big] \leq \frac{3}{ng}\int k(v)^2 dv \sup_{t\in[0,1]}|m(t)|^2$$

and hence, by Markov's inequality,

$$\frac{1}{\min_t f_T(t)}\frac{3}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\xi_{ij}^2 = O_p\left(\frac{1}{ng}\right).$$

Further, note that the mean of second term on the right-hand side is zero as the ξ_{ij} are mean zero and independent across j conditionally on T_i. Thus, by Markov's inequality, (3.42) is of order

$$\frac{3}{n}\sum_{i=1}^{n}\left\{\frac{1}{\hat{f}_{T,-i}(T_i)}\frac{1}{ng}\sum_{j\neq i}k_g(T_j,T_i)m(T_j)-\mathrm{E}[k_g(T_j,T_i)m(T_j)\,|\,T_i]\right\}^2=O_p\left(\frac{1}{ng}\right).$$

Finally, for (3.43), we have

$$\begin{aligned}&\frac{2}{n}\sum_{i=1}^{n}\left\{\frac{1}{g\hat{f}_{T,-i}(T_i)}\,\mathrm{E}[k_g(T_j,T_i)m(T_j)\,|\,T_i]-m(T_i)\right\}^2\\ \leq&2\sup_{t\in[0,1]}\left\{\frac{1}{gf_T(T_i)}\,\mathrm{E}[k_g(T_j,T_i)m(T_j)\,|\,T_i]-m(T_i)\right\}^2(1+o_p(1))=O_p(g^4),\end{aligned}$$

where we have used that $1/\hat{f}_{T,-i}(t)=1/f_T(t)+o_p(1)$ uniformly in t. Thus, it holds

$$\left(\frac{\sum_{j\neq i}k_g(T_j,T_i)}{k(0)+\sum_{j\neq i}k_g(T_j,T_i)}\right)^2\frac{2}{n}\sum_{i=1}^{n}\{\hat{m}^*_{k,-i}(T_i)-m(T_i)\}^2=O_p\left(\frac{1}{ng}+g^4\right)$$

and the claim follows. □

Lemma 62. *It holds*

$$\max_{j=1,\ldots,p}\sup_{t\in[-1,1]}\left|\frac{1}{ng}\sum_{i=1}^{n}k_g(T_i,t)m_{X,j}(T_i)-\frac{1}{g}\,\mathrm{E}[k_g(T_i,t)m_{X,j}(T_i)]\right|=O_p\left(\sqrt{\frac{\log(pn)}{ng}}\right)$$

Proof. We first discretize the supremum. For that purpose, let $t_0=0$, $t_M=1$ and $t_j=j/M$. For any $t\in[0,1]$, there is a $j(t)$ in $0,1,\ldots,M$ such that $|t-j(t)|\leq 1/M$. Moreover, by similar arguments as in the proof of Lemma 60, $\frac{1}{ng}\sum_{i=1}^{n}(k_g(T_i,t)m_{X,j}(T_i)-\mathrm{E}[k_g(T_i,t)m_{X,j}(T_i)])$ is Lipschitz-continuous in t and therefore

$$\begin{aligned}&\sup_{t\in[0,1]}\left|\frac{1}{ng}\sum_{i=1}^{n}(k_g(T_i,t)m_{X,j}(T_i)-\mathrm{E}[k_g(T_i,t)m_{X,j}(T_i)])\right|\\ \leq&\sup_{t\in[0,1]}\left|\frac{1}{ng}\sum_{i=1}^{n}(k_g(T_i,t)m_{X,j}(T_i)-\mathrm{E}[k_g(T_i,t)m_{X,j}(T_i)])\right|\\ &+\max_{j=0,\ldots,M}\left|\frac{1}{ng}\sum_{i=1}^{n}(k_g(T_i,t_j)m_{X,j}(T_i)-\mathrm{E}[k_g(T_i,t_j)m_{X,j}(T_i)])\right|\end{aligned}$$

$$\leq \max_{j=0,\dots,M} \left| \frac{1}{ng} \sum_{i=1}^{n} (k_g(T_i, t_j) m_{X,j}(T_i) - \mathrm{E}[k_g(T_i, t_j) m_{X,j}(T_i)]) \right| + \frac{C}{Mg^2}.$$

Since the $m_{X,j}$ and the kernel are uniformly bounded, the first term on the right-hand side is a sum of mean-zero, independent and bounded random variables. Therefore, by Hoeffding's inequality together with a union bound,

$$\max_{j=0,\dots,M} \left| \frac{1}{ng} \sum_{i=1}^{n} (k_g(T_i, t_j) m_{X,j}(T_i) - \mathrm{E}[k_g(T_i, t_j) m_{X,j}(T_i)]) \right| = O_p\left(\sqrt{\frac{\log pM}{ng}} \right)$$

Taking $M = n^2$, the claim follows. □

Lemma 63. *It holds*

$$\max_{j=1,\dots,p} \sup_{t \in [-1,1]} \left| \frac{1}{ng} \sum_{i=1}^{n} k_g(T_i, t) u_{ij} \right| = O_p\left(\sqrt{\frac{\log(pn)}{ng}} \right)$$

Proof. For any $s, t \in [0, 1]$, we have

$$\begin{aligned} &\left| \frac{1}{ng} \sum_{i=1}^{n} k_g(T_i, s) u_{ij} - \frac{1}{ng} \sum_{i=1}^{n} k_g(T_i, t) u_{ij} \right| \\ \leq &\frac{1}{ng} \sum_{i=1}^{n} |k_g(T_i, s) - k_g(T_i, t)| |u_{ij}| \\ \leq &\frac{C}{g^2} |s - t|, \end{aligned}$$

i.e., $\frac{1}{ng} \sum_{i=1}^{n} k_g(T_i, s) u_{ij}$ is Lipschitz-continuous in t. Hence, by an analogous discretization argument as in Lemma 62, one can show that

$$\sup_{t \in [-1,1]} \left| \frac{1}{ng} \sum_{i=1}^{n} k_g(T_i, t) u_{ij} \right| \leq \max_{\ell=0,\dots,M} \left| \frac{1}{ng} \sum_{i=1}^{n} k_g(T_i, t_\ell) u_{ij} \right| + \frac{C}{Mg^2}$$

and therefore by the same arguments as in Lemma 62

$$\max_{j=1,\dots,p} \sup_{t \in [-1,1]} \left| \frac{1}{ng} \sum_{i=1}^{n} k_g(T_i, t) u_{ij} \right| = O_p\left(\sqrt{\frac{\log(pn)}{ng}} + \frac{1}{Mg^2} \right).$$

Taking $M = n^2$ yields the result. □

Lemma 64. *It holds*

$$\frac{1}{n}\|(\hat{m}_k^*(T) - m(T))^\intercal u\|_\infty = O_p\left(\sqrt{\frac{\log p}{n}}\left\{\frac{1}{\sqrt{ng}} + g^2\right\}\right)$$

Proof. By the triangle inequality,

$$\begin{aligned}\frac{1}{n}\|(\hat{m}_k^*(T) - m(T))^\intercal u\|_\infty \leq & \max_{j=1,\ldots,p}\left|\frac{1}{n}\sum_{i=1}^n \frac{\sum_{\ell=1,\ldots,n} k_g(T_\ell, T_i)\varepsilon_\ell}{\sum_{\ell=1,\ldots,n} k_g(T_\ell, T_i)} u_{ij}\right| \\ & + \max_{j=1,\ldots,p}\left|\frac{1}{n}\sum_{i=1}^n \frac{\sum_{\ell=1,\ldots,n} k_g(T_\ell, T_i)\{m(T_\ell) - m(T_i)\}}{\sum_{\ell=1,\ldots,n} k_g(T_\ell, T_i)} u_{ij}\right|\end{aligned}$$

Denote by $w_{i\ell}$ the weights

$$w_{i\ell} = \frac{k_g(T_\ell, T_i)}{\sum_{\ell=1,\ldots,n} k_g(T_\ell, T_i)}.$$

With these weights, we can rewrite the first term as

$$\frac{1}{n}\sum_{i=1}^n \frac{\sum_{\ell=1,\ldots,n} k_g(T_\ell, T_i)\varepsilon_\ell}{\sum_{\ell=1,\ldots,n} k_g(T_\ell, T_i)} u_{ij} = \frac{1}{n}\sum_{\ell=1}^n \sum_{i=1}^n w_{i\ell} u_{ij} \varepsilon_\ell.$$

Conditionally on $\{(T_i, X_i), i = 1, \ldots, n\}$ this is a weighted sum of the ε_i with weights $\tilde{w}_\ell$ given by

$$\tilde{w}_\ell = \frac{1}{n}\sum_{i=1}^n w_{i\ell} u_{ij}.$$

The weights satisfy

$$\begin{aligned}\mathrm{E}[\tilde{w}_\ell^2 \,|\, T_i, i = 1, \ldots, n] &= \frac{1}{n^2}\sum_{i=1}^n w_{i\ell}^2 \underbrace{\mathrm{E}[u_{ij}^2 \,|\, T_i]}_{\leq 4} = O\left(\frac{1}{n^3 g}\right), \\ &\leq \frac{4}{\min_t f_T(t) + o_p(1)} \frac{1}{n^4 g^2}\sum_{i=1}^n k_g(T_\ell, T_i)^2 \\ &= \frac{1}{n^3 g}\frac{4}{\min_t f_T(t) + o_p(1)}\left(\frac{1}{g}\mathrm{E}[k_g(T_j, T_i)] + o_p(1)\right) \\ &= \frac{1}{n^3 g}\frac{4}{\min_t f_T(t)}\int f_T(t)^2 dt (1 + o_p(1)),\end{aligned}$$

where we have used Lemma 60 and that

$$\begin{aligned}
\mathrm{E}[k_g(T_j,T_i)] &= \iint k_g(T_j,T_i) f_T(T_j) f_T(T_i) dT_j dT_i \\
&= \int_{[0,g]} \int k_{low}^{(T_i/g)}\left(\frac{T_j - T_i}{g}\right) f_T(T_j) f_T(T_i) dT_j dT_i \\
&\quad + \int_{[g,1-g]} \int k_{int}\left(\frac{T_j - T_i}{g}\right) f_T(T_j) f_T(T_i) dT_j dT_i \\
&\quad + \int_{[1-g,1]} \int k_{up}^{(1-T_i/g)}\left(\frac{T_j - T_i}{g}\right) f_T(T_j) f_T(T_i) dT_j dT_i
\end{aligned}$$

By substitution with $v = (T_j - T_i)/g$ and a Taylor expansion of f_T, we have

$$\begin{aligned}
&\int_{[0,g]} \int k_{low}^{(T_i/g)}\left(\frac{T_j - T_i}{g}\right) f_T(T_j) f_T(T_i) dT_j dT_i \\
=& g \int_{[0,g]} \int_{-T_i/g}^{1} k_{low}^{(T_i/g)}(v) f_T(T_i + gv) f_T(T_i) dv dT_i \\
=& g \int_{[0,g]} \underbrace{\int_{-T_i/g}^{1} k_{low}^{(T_i/g)}(v) dv}_{=1} f_T(T_i)^2 dT_i \\
&+ g^2 \int_{[0,g]} \underbrace{\int_{-T_i/g}^{1} k_{low}^{(T_i/g)}(v) v dv}_{=0} f_T(T_i) f_T'(T_i) dT_i \\
&+ g^3 \int_{[0,g]} \int_{-T_i/g}^{1} k_{low}^{(T_i/g)}(v) v^2 dv f_T(T_i) f_T''(T_i) dT_i \\
&+ g^3 \int_{[0,g]} \int_{-T_i/g}^{1} k_{low}^{(T_i/g)}(v) v^2 \{f_T''(T_i + \xi g v) d - f_T''(T_i)\} v f_T(T_i) dT_i \\
=& g \int_{[0,g]} f_T(t)^2 dt + O(g^3)
\end{aligned}$$

By the same arguments, one can show for the other two terms,

$$\begin{aligned}
\int_{[g,1-g]} \int k_{int}\left(\frac{T_j - T_i}{g}\right) f_T(T_j) f_T(T_i) dT_j dT_i &= g \int_{[g,1-g]} f_T(t)^2 dt + O(g^3) \\
\int_{[1-g,1]} \int k_{up}^{(1-T_i/g)}\left(\frac{T_j - T_i}{g}\right) f_T(T_j) f_T(T_i) dT_j dT_i &= g \int_{[1-g,1]} f_T(t)^2 dt + O(g^3)
\end{aligned}$$

and hence,

$$\mathrm{E}[k_g(T_j, T_i)] = g \int f_T(t)^2 dt + O(g^3).$$

This implies by Markov's inequality and the law of iterated expectations

$$\sum_{\ell=1}^{n} \tilde{w}_\ell^2 = O_p\left(\frac{1}{n^2 g}\right).$$

Thus, the probability of the event A_M given by

$$A_M = \left\{ \sum_{\ell=1}^{n} \tilde{w}_\ell \leq \frac{M}{n^2 g} \right\}$$

can be made arbitrarily close to one by choosing M sufficiently large. Now, we can bound the first term. For any $K > 0$, we have

$$\mathrm{P}\left(\max_{j=1,\ldots,p} \left| \sum_{\ell=1}^{n} \tilde{w}_\ell \varepsilon_\ell \right| \geq K\rho_n \right) \leq \mathrm{P}\left(\max_{j=1,\ldots,p} \left| \sum_{\ell=1}^{n} \tilde{w}_\ell \varepsilon_\ell \right| \geq K\rho_n, A_M \right) + \mathrm{P}(A_M^c),$$

where ρ_n denotes

$$\rho_n = \sqrt{\frac{\log p}{n}} \left\{ \frac{1}{\sqrt{ng}} + g^2 \right\}.$$

By conditional sub-Gaussianity of the ε_i, we have

$$\mathrm{P}\left(\left| \sum_{\ell=1}^{n} \tilde{w}_\ell \varepsilon_\ell \right| \geq K, A_M \,\middle|\, (T_i, X_i), i = 1, \ldots, n \right) \leq 2 \exp\left(-C \frac{n^2 g K^2}{M} \right) \quad \text{a.s.}$$

where C denotes some absolute constant which only depends on K_ε. By the law of iterated expectations and a union bound, we thus obtain

$$\begin{aligned} \mathrm{P}\left(\left| \sum_{\ell=1}^{n} \tilde{w}_\ell \varepsilon_\ell \right| \geq K\rho_n, A_M \right) &\leq 2 \exp\left(\log p - C \frac{n^2 g \rho_n^2 K^2}{M} \right) \\ &\leq 2 \exp\left(\log p - C \frac{(1 + \sqrt{ng^5})^2 K^2 \log p}{M} \right). \end{aligned}$$

Choosing M and K sufficiently large, $\mathrm{P}\left(\max_j |\sum_{\ell=1}^{n} \tilde{w}_\ell \varepsilon_\ell| \geq K\rho_n\right)$ becomes arbitrar-

ily close to zero, implying

$$\max_{j=1,\dots,p}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{\ell=1,\dots,n}k_g(T_\ell,T_i)\varepsilon_\ell}{\sum_{\ell=1,\dots,n}k_g(T_\ell,T_i)}u_{ij}\right| = O_p(\rho_n).$$

Regarding the second term in our initial decomposition, we can interpret this as a weighted sum of the u_{ij} with weights

$$v_i = \frac{1}{n}\frac{\sum_{\ell=1,\dots,n}k_g(T_\ell,T_i)\{m(T_\ell)-m(T_i)\}}{\sum_{\ell=1,\dots,n}k_g(T_\ell,T_i)}$$

satisfying

$$\sum_{i=1}^{n}v_i^2 = O_p\left(\frac{1}{n}\left\{\frac{1}{ng}+g^4\right\}\right).$$

Thus, by similar arguments as for the first term, one can show

$$\max_{j=1,\dots,p}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{\ell=1,\dots,n}k_g(T_\ell,T_i)\{m(T_\ell)-m(T_i)\}}{\sum_{\ell=1,\dots,n}k_g(T_\ell,T_i)}u_{ij}\right| = O_p(\rho_n).$$

The claim follows. □

By the same arguments as in Lemma 61 and 62 one can show:

Lemma 65. *It holds*

$$\max_{j=1,\dots p}\frac{1}{n}\|\hat{m}_X(T)-m_X(T)\|_2^2 = O_p\left(\frac{\sqrt{\log p}}{ng}+g^4\right).$$

3.C.2 Construction of Boundary Corrected Kernels

Motivated by Gasser and Rosenblatt (1979), we construct the auxiliary kernels $k_{low}^{(q)}$ and $k_{up}^{(q)}$ by a polynomial weighting of some ordinary second order kernel k. We set

$$k_{low}^{(q)}(x) = (a(q)+b(q)x)k(x)$$

and $a(q)$, $b(q)$ are chosen so that

$$\int k_{low}^{(q)}(x) = 1 \qquad \text{and} \qquad \int xk_{low}^{(q)}(x) = 0,$$

or equivalently,

$$a(q)\int_{-q}^{1} k(x)dx + b(q)\int_{-q}^{1} xk(x)dx = 1$$
$$a(q)\int_{-q}^{1} xk(x)dx + b(q)\int_{-q}^{1} x^2k(x)dx = 0.$$

Solve the second equality for $a(q)$

$$b(q) = -\frac{\int_{-q}^{1} xk(x)dx}{\int_{-q}^{1} x^2k(x)dx}a(q)$$

and insert this into the first to get

$$a(q) = \left(\int_{-q}^{1} k(x)dx - \frac{(\int_{-q}^{1} xk(x)dx)^2}{\int_{-q}^{1} x^2k(x)dx}\right)^{-1}.$$

In particular, for k equal to the Epanechnikov kernel, we have

$$\int_{-q}^{1} \frac{3}{4}(1-x^2)dx = \frac{3}{4}(1+q) - \frac{1}{4}(1+q^3)$$
$$\int_{-q}^{1} x\frac{3}{4}(1-x^2)dx = \frac{3}{8}(1-q^2) - \frac{3}{16}(1-q^4)$$
$$\int_{-q}^{1} x^2\frac{3}{4}(1-x^2)dx = \frac{1}{4}(1+q^3) - \frac{3}{20}(1+q^5).$$

The kernel can be constructed with the given information.

The derivation of $a(q)$ and $b(q)$ for $k_{up}^{(q)}$ follow analogously. They only differ in the support of the moments of k:

$$a(q) = \left(\int_{-1}^{q} k(x)dx - \frac{(\int_{-1}^{q} xk(x)dx)^2}{\int_{-1}^{q} x^2k(x)dx}\right)^{-1}$$
$$b(q) = -\frac{\int_{-1}^{q} xk(x)dx}{\int_{-1}^{q} x^2k(x)dx}a(q).$$

For k equal to the Epanechnikov kernel, it holds

$$\int_{-1}^{q} \frac{3}{4}(1-x^2)dx = \frac{3}{4}(1+q) - \frac{1}{4}(1+q^3)$$

$$\int_{-1}^{q} x\frac{3}{4}(1-x^2)dx = \frac{3}{8}(q^2-1) - \frac{3}{16}(q^4-1)$$
$$\int_{-1}^{q} x^2\frac{3}{4}(1-x^2)dx = \frac{1}{4}(1+q^3) - \frac{3}{20}(1+q^5)$$

and the auxiliary kernel can be constructed from this information.

3.D Results on the Orthogonalization Parameters

In the first part of this section, we discuss properties of the orthogonalization parameters γ_{t_ℓ} which we use in the second part of this section to study the prediction properties of the auxiliary Lasso.

3.D.1 Properties of γ_{t_ℓ}

The main goal of this section is to prove Lemma 69. We start off with some auxiliary results which are needed repeatedly.

Lemma 66. *Let γ_{t_ℓ} satisfy (3.11). Then, uniformly over ℓ and p,*

(i) $\mathrm{E}\big[\mathbb{1}(T_i \in \mathcal{I}_\ell)\big\{1-(X_i-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell}\big\}^2\big] = \mathrm{E}\big[\mathbb{1}(T_i \in \mathcal{I}_\ell)\big\{1-(X_i-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell}\big\}\big]$

(ii) $\mathrm{E}\big[\mathbb{1}(T_i \in \mathcal{I}_\ell)\big\{1-(X_i-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell}\big\}^2\big] \leq \mathrm{P}(T_i \in \mathcal{I}_\ell)$

(iii) $\mathrm{E}\big[\big(1-(X_i-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell}\big)^2 \,|\, T_i \in \mathcal{I}_\ell\big] \leq 1$

(iv) $\mathrm{E}[\mathrm{Var}((X_i - m_X(T_i))^\intercal\gamma_{t_\ell} \,|\, T_i) \,|\, T_i \in \mathcal{I}_\ell] \leq 1$

(v) $\mathrm{E}\big[\big|1-(X_i-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell}\big| \,|\, T_i \in \mathcal{I}_\ell\big] \leq 2.$

Proof. For brevity of notation, let $\mathrm{E}[X_i] = 0$.

(i): By the first order conditions of (3.11),

$$\begin{aligned}
&\mathrm{E}\big[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1-X_i^\intercal\gamma_{t_\ell})^2\big] \\
=&\mathrm{P}(T_i \in \mathcal{I}_\ell) - 2\,\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)X_i^\intercal]\gamma_{t_\ell} + \gamma_{t_\ell}^\intercal\,\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)X_iX_i^\intercal]\gamma_{t_\ell} \\
=&\mathrm{P}(T_i \in \mathcal{I}_\ell) - \mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)X_i^\intercal]\gamma_{t_\ell} \\
&- \big\{\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)X_i^\intercal] - \gamma_{t_\ell}^\intercal\,\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)X_iX_i^\intercal]\big\}\gamma_{t_\ell} \\
=&\mathrm{P}(T_i \in \mathcal{I}_\ell) - \mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)X_i^\intercal]\gamma_{t_\ell} \\
=&\,\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1-X_i^\intercal\gamma_{t_\ell})]
\end{aligned}$$

(ii): By definition of γ_{t_ℓ} in (3.11) and as $\gamma = 0$ is a feasible solution of this problem, it holds

$$\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - X_i^\intercal \gamma_{t_\ell})^2] \leq \mathrm{P}(T_i \in \mathcal{I}_\ell).$$

(iii): Follows directly by rearranging (ii).

(iv): It holds

$$\begin{aligned}
&\mathrm{E}\big[\mathbb{1}(T_i \in \mathcal{I}_\ell)\big\{1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}\big\}^2\big] \\
\geq &\min_{g \in L_2(P_T)} \mathrm{E}\big[\mathbb{1}(T_i \in \mathcal{I}_\ell)\big\{1 - X_i^\intercal \gamma_{t_\ell} - g(T_i)\big\}^2\big] \\
= &\mathrm{E}\big[\mathbb{1}(T_i \in \mathcal{I}_\ell)\big\{1 - X_i^\intercal \gamma_{t_\ell} - \mathrm{E}[1 - X_i^\intercal \gamma_{t_\ell} \,|\, T_i]\big\}^2\big] \\
= &\mathrm{E}\big[\mathbb{1}(T_i \in \mathcal{I}_\ell)\big\{(X_i - m_X(T_i))^\intercal \gamma_{t_\ell}\big\}^2\big]
\end{aligned}$$

the claim now follows by (ii) and the law of iterated expectations.

(v): By (i) and (ii),

$$\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - X_i^\intercal \gamma_{t_\ell})] = \mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - X_i^\intercal \gamma_{t_\ell})^2] \leq \mathrm{P}(T_i \in \mathcal{I}_\ell).$$

Moreover, for any $\lambda > 1$,

$$\begin{aligned}
\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - X_i^\intercal \gamma_{t_\ell})^2] \geq &\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - X_i^\intercal \gamma_{t_\ell})^2 \mathbb{1}(|1 - X_i^\intercal \gamma_{t_\ell}| > \lambda)] \\
\geq &\lambda^2 \mathrm{P}(|1 - X_i^\intercal \gamma_{t_\ell}| > \lambda, T_i \in \mathcal{I}_\ell).
\end{aligned}$$

Thus,

$$\mathrm{P}(|1 - X_i^\intercal \gamma_{t_\ell}| > \lambda \,|\, T_i \in \mathcal{I}_\ell) \leq \frac{1}{\lambda^2}$$

which implies

$$\begin{aligned}
\mathrm{E}[|1 - X_i^\intercal \gamma_{t_\ell}| \,|\, T_i \in \mathcal{I}_\ell] &= \int_0^\infty \mathrm{P}(|1 - X_i^\intercal \gamma_{t_\ell}| > \lambda \,|\, T_i \in \mathcal{I}_\ell) d\lambda \\
&\leq \int_0^1 d\lambda + \int_1^\infty \frac{1}{\lambda^2} d\lambda = 2
\end{aligned}$$

and the claim follows. □

Lemma 67. *Suppose that the conditional density of T_i given X_i satisfies*

$$f_{T|X}(t, X_i) \geq c \quad a.s.$$

for some constant $c > 0$. Then, there exist constants $C_1, C_2, C_3 \in (0, \infty)$ satisfying uniformly over ℓ and p

(i) $\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - X_i^\intercal \gamma_{t_\ell})^2] \geq C_1 h$

(ii) $\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - X_i^\intercal \gamma_{t_\ell})] \in [0, C_2 h]$

(iii) $\gamma_{t_\ell}^\intercal \mathrm{E}[X_i X_i^\intercal] \gamma_{t_\ell} \leq C_3$

(iv) $\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell) X_i X_i^\intercal]$, $\ell = 1, \ldots, L$ *are positive definite.*

Proof. (i) It holds

$$\begin{aligned}
\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - X_i^\intercal \gamma_{t_\ell})^2] &\geq \mathrm{E}[\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)](1 - X_i^\intercal \gamma_{t_\ell})^2] \\
&\geq 2ch\, \mathrm{E}[(1 - X_i^\intercal \gamma_{t_\ell})^2] \\
&\geq 2ch \min_{\gamma \in \mathbb{R}^p} \mathrm{E}[(1 - X_i^\intercal \gamma)^2] = 2ch,
\end{aligned}$$

where we have used that

$$\min_{\gamma \in \mathbb{R}^p} \mathrm{E}[(1 - X_i^\intercal \gamma)^2] = \min_{\gamma \in \mathbb{R}^p} 1 + \mathrm{E}[(X_i^\intercal \gamma)^2] - 2\, \mathrm{E}[X_i]^\intercal \gamma = 0$$

since $\mathrm{E}[X_i] = 0$.

(ii) This follows by Lemma 66(i), (ii) and part (i) of this Lemma.

(iii) By definition of γ_{t_ℓ}, it holds

$$\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - X_i^\intercal \gamma_{t_\ell})^2] = \mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)] - \gamma_{t_\ell} \mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell) X_i X_i^\intercal] \gamma_{t_\ell}.$$

By Lemma 66(i), this implies

$$\gamma_{t_\ell} \mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell) X_i X_i^\intercal] \gamma_{t_\ell} = \mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell) X_i]^\intercal \gamma_{t_\ell}.$$

Further, by the same argument as in part (i)

$$\gamma_{t_\ell}^\intercal \mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell) X_i X_i^\intercal] \gamma_{t_\ell} \geq 2ch \gamma_{t_\ell}^\intercal \mathrm{E}[X_i X_i^\intercal] \gamma_{t_\ell}$$

and the claim follows by part (ii).

(iv) For any $\gamma \in \mathbb{R}^p$ and $\ell = 1, \dots, L$,

$$\gamma^\intercal \operatorname{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell) X_i X_i^\intercal]\gamma \geq 2ch\gamma^\intercal \operatorname{E}[X_i X_i^\intercal]\gamma \geq 2ch\gamma^\intercal \operatorname{E}[u_i u_i^\intercal]\gamma > 0.$$

The claim follows.

□

Lemma 68. *Suppose that Assumptions 16 and 18 hold. Then, there exists a constant which is independent of ℓ and p such that*

(i) $\operatorname{E}[|1 - X_i^\intercal \gamma_{t_\ell}|] \leq C$

(ii) $|1 - m_X(t_\ell)^\intercal \gamma_{t_\ell}| \leq C$

(iii) $|m_X'(t_\ell)^\intercal \gamma_{t_\ell}| \leq C$

(iv) $|m_X''(t_\ell)^\intercal \gamma_{t_\ell}| \leq C$.

Proof. (i) By the triangle inequality and Jensen's inequality

$$\operatorname{E}[|1 - X_i^\intercal \gamma_{t_\ell}|] \leq \operatorname{E}[1] + \operatorname{E}[|X_i^\intercal \gamma_{t_\ell}|] \leq 1 + \sqrt{\operatorname{E}[(X_i^\intercal \gamma_{t_\ell})^2]}$$

The claim follows by Lemma 67.

(ii) Note that

$$\begin{aligned}\int f_{T|X}(t_\ell, x)(1 - x^\intercal \gamma_{t_\ell}) f_X(x) dx =& f_T(t_\ell) \operatorname{E}[(1 - X_i^\intercal \gamma_{t_\ell}) \,|\, T_i = t_\ell] \\ =& f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})\end{aligned}$$

Thus,

$$\begin{aligned}|f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})| \leq& \int |f_{T|X}(t_\ell, x)||1 - x^\intercal \gamma_{t_\ell}| f_X(x) dx \\ \leq& \|f_{T|X}\|_\infty \operatorname{E}[|1 - X_i^\intercal \gamma_{t_\ell}|]\end{aligned}$$

and the claim follows by (i) and boundedness of f_T.

(iii) The argument follows similarly to (ii). Note that

$$\int f_{T|X}'(t_\ell, x)(1 - x^\intercal \gamma_{t_\ell}) f_X(x) dx = \frac{\partial}{\partial t} \int (1 - x^\intercal \gamma_{t_\ell}) f_{T|X}(t, x) f_X(x) dx \Big|_{t = t_\ell}$$

$$
\begin{aligned}
&= \frac{\partial}{\partial t} f_T(t)(1 - m_X(t)^\intercal \gamma_{t_\ell})\big|_{t=t_\ell} \\
&= f_T'(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell}) + f_T(t_\ell) m_X'(t_\ell)^\intercal \gamma_{t_\ell}
\end{aligned}
$$

where taking out the derivative is justified since the integrand is continuous with continuous partial derivative with respect to t. Hence,

$$
|f_T'(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell}) + f_T(t_\ell) m_X'(t_\ell)^\intercal \gamma_{t_\ell}| \leq \|f_{T|X}'\|_\infty \, \mathrm{E}[|1 - X_i^\intercal \gamma_{t_\ell}|] \leq C
$$

and by the reverse triangle inequality

$$
\begin{aligned}
&|f_T'(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell}) + f_T(t_\ell) m_X'(t_\ell)^\intercal \gamma_{t_\ell}| \\
\geq &|f_T(t_\ell) m_X'(t_\ell)^\intercal \gamma_{t_\ell}| - |f_T'(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})|,
\end{aligned}
$$

i.e.,

$$
|f_T(t_\ell) m_X'(t_\ell)^\intercal \gamma_{t_\ell}| \leq 2\|f_{T|X}'\|_\infty \, \mathrm{E}[|1 - X_i^\intercal \gamma_{t_\ell}|]
$$

and the claim follows by (i) and boundedness of f_T.

(iv) The final claim follows analogously as (iii) and is omitted for brevity. □

Lemma 69. *Let*

$$
\begin{aligned}
\beta^{(j,\ell)} &= \operatorname*{argmin}_{b \in \mathbb{R}^{p-1}} \mathrm{E}\left[\mathbb{1}(|T_i - t_\ell| \leq h) \left(X_{ij} - \sum_{k \neq j} X_{ik} b_k \right)^2 \right] \\
\eta_{j,\ell}^2 &= \mathrm{E}\left[\mathbb{1}(|T_i - t_\ell| \leq h) \left(X_{ij} - \sum_{k \neq j} X_{ik} \beta_k^{(j,\ell)} \right)^2 \right].
\end{aligned}
$$

Suppose that there exist sequences $r_{1,n}$, $r_{2,n}$ and $r_{3,n}$ such that $\max_{\ell,j} \|\beta^{(j,\ell)}\|_1 = O(r_{1,n})$, $\min_{\ell,j} \eta_{j,\ell}^2 = O(r_{2,n}^{-1})$ and

$$
\max_{\ell=1,\ldots,L} \sum_{j=1}^{p} \big| \mathrm{Cov}\big(\mathrm{P}(|T_i - t_\ell| \leq h \,|\, X_i), X_i \big) \big| = O(r_{3,n}).
$$

If $r_{1,n} r_{3,n} / r_{2,n} = o\left(\sqrt{\frac{nh}{\log^9(Lp)}} \right)$, then $\max_{\ell=1,\ldots,L} \|\gamma_{t_\ell}\|_1 = o\left(\sqrt{\frac{nh}{\log^9(Lp)}} \right)$.

Proof of Lemma 69. Let $\Sigma_\ell = \mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(X_i - \mathrm{E}[X_i])(X_i - \mathrm{E}[X_i])^\intercal]$ and denote by $\sigma_{ij,\ell}$ the ijth element of Σ_ℓ. Further, let $(\Sigma_\ell^{-1})^{(j)}$ the jth column (or row) of Σ_ℓ^{-1}.

Then

$$\|\Sigma_\ell^{-1}\|_1 = \max_{j=1,\dots,p} \|(\Sigma_\ell^{-1})^{(j)}\|_1.$$

In order to bound the ℓ_1-norm of $(\Sigma_\ell^{-1})^{(j)}$, we can assume without loss of generality that $j = 1$. Now, partition Σ_ℓ as follows

$$\Sigma_\ell = \begin{pmatrix} \sigma_{11,\ell} & b^\intercal \\ b & C \end{pmatrix}$$

where $b = (\sigma_{21,\ell}, \sigma_{31,\ell}, \dots, \sigma_{p1,\ell})^\intercal$ and $C = (\sigma_{ij,\ell})_{i,j=2,\dots,p}$. The inverse of Σ_ℓ can now be written as

$$\Sigma_\ell^{-1} = \begin{pmatrix} (\sigma_{11,\ell} - b^\intercal C^{-1} b)^{-1} & -(\sigma_{11,\ell} - b^\intercal C^{-1} b)^{-1} b^\intercal C^{-1} \\ -(\sigma_{11,\ell} - b^\intercal C^{-1} b)^{-1} C^{-1} b & (C - b\sigma_{11,\ell}^{-1} b^\intercal)^{-1} \end{pmatrix}.$$

Note that $\beta^{(1,\ell)} = C^{-1}b$ and $\eta_{1,\ell}^2 = \sigma_{11,\ell} - b^\intercal C^{-1} b$. Hence, $(\Sigma_\ell^{-1})^{(1)} = (1, -\beta^{(1,\ell)})^\intercal \eta_{1,\ell}^{-2}$ and therefore

$$\|\Sigma_\ell^{-1}\|_1 = \max_{j=1,\dots,p} (1 + \|\beta^{(j,\ell)}\|_1)\eta_{j,\ell}^{-2}$$

which is of order $r_{1,n}/r_{2,n}$ uniformly in ℓ. Next, by the law of iterated expectations,

$$\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(X_{ij} - \mathrm{E}[X_{ij}])] = \mathrm{Cov}\big(\mathrm{P}(T_i \in \mathcal{I}_\ell \,|\, X_i), X_i\big)$$

and thus

$$\max_{\ell=1,\dots,L} |\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(X_{ij} - \mathrm{E}[X_{ij}])]\|_1 = O(r_{3,n}).$$

Now,

$$\begin{aligned} \max_{\ell=1,\dots,L} \|\gamma_{t_\ell}\|_1 &= \max_{\ell=1,\dots,L} \|\Sigma_\ell^{-1} \,\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(X_{ij} - \mathrm{E}[X_{ij}])]\|_1 \\ &\le \max_{\ell=1,\dots,L} \|\Sigma_\ell^{-1}\|_1 \max_{\ell=1,\dots,L} \|\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(X_{ij} - \mathrm{E}[X_{ij}])]\|_1 = O(r_{1,n} r_{3,n}/r_{2,n}) \end{aligned}$$

and the claim follows. □

3.D.2 On Prediction Properties of $\hat{\gamma}_{\ell,\mu}$

We begin by showing that Lasso does not overfit under essentially no assumptions on the design.

Lemma 70. *Let $Y \in \mathbb{R}^n, Y \neq 0$ and $X \in \mathbb{R}^{n\times p}$. Then, no solution $\hat{\beta}$ of*

$$\operatorname*{argmin}_{b\in\mathbb{R}^p} \frac{1}{n}\|Y - Xb\|_2^2 + \lambda\|b\|_1$$

satisfies $X\hat{\beta} = Y$ as long as $\lambda > 0$.

Proof. If Y is not in the span of X, the claim is obvious. Thus, suppose that Y is contained in the span of X. We prove the claim by contradiction. So suppose that $\lambda > 0$ and that there exists a solution $\hat{\beta}$ satisfying $X\hat{\beta} = Y$. It is sufficient to construct a vector $\beta \in \mathbb{R}^p$ which leads to a strictly smaller value of the criterion function than $\hat{\beta}$. For this purpose, note that, as $Y \neq 0$, there exists some $j = 1, \ldots, p$ such that $\hat{\beta}_j \neq 0$ and the column $X^{(j)}$ of X is not the zero vector. Further, let

$$\varepsilon = \frac{1}{2}\min\left\{\frac{\sqrt{n\lambda}}{\|X^{(j)}\|_2}, \frac{|\hat{\beta}_j|}{\sqrt{\lambda}}\right\}$$

and $h \in \mathbb{R}^p$ such that $h_k = 0$ for all $k \neq j$ and $h_j = -\operatorname{sgn}(\hat{\beta}_j)\sqrt{\lambda}\varepsilon$. Set $\beta = \hat{\beta} + h$. Then the value of the criterion function evaluated at β is given by

$$\begin{aligned}
\frac{1}{n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 &= \frac{1}{n}\|Y - X\hat{\beta} - Xh\|_2^2 + \lambda\|\hat{\beta} + h\|_1 \\
&= \frac{1}{n}\|Xh\|_2^2 + \lambda\|\hat{\beta} + h\|_1 \\
&= \frac{1}{n}\|X^{(j)}\|_2^2\lambda\varepsilon^2 + \lambda\|\hat{\beta} + h\|_1 \\
&= \frac{1}{n}\|X^{(j)}\|_2^2\lambda\varepsilon^2 + \lambda\big(\|\hat{\beta}\|_1 - \sqrt{\lambda}\varepsilon\big)
\end{aligned}$$

where we have used in the last equality that $\|\beta\|_1 = \|\hat{\beta}\|_1 - \sqrt{\lambda}\varepsilon$ as $\varepsilon < |\hat{\beta}_j|/\sqrt{\lambda}$.

Thus, β leads to a strictly smaller value of the criterion function than $\hat{\beta}$ iff

$$\frac{1}{n}\|X^{(j)}\|_2^2\lambda\varepsilon^2 - \lambda^{3/2}\varepsilon < 0$$

which is satisfied since $\varepsilon < \sqrt{n\lambda}/\|X^{(j)}\|_2$. Hence, $\hat{\beta}$ cannot be a solution to the Lasso problem. □

The remainder of this section is devoted to results on the prediction error of $\hat{\gamma}_{t_\ell}$. In particular, we are only interested in the so-called "slow rates" for prediction. These have the advantage that they do not rely on restrictive conditions on the design such as restricted eigenvalue or compatibility conditions. These slow rates for $\hat{\gamma}_{\ell,\mu}$ can be derived by the same arguments as for the usual Lasso. We only have to take a closer look at the stochastic properties of the corresponding "effective noise term".

Lemma 71. *On the event*

$$\mathcal{T} := \left\{ \max_{j=1,\dots,p} \left| \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \gamma_{t_\ell}) \dot{X}_{ij} \right| \leq \mu_0 \right\} \tag{3.44}$$

it holds for all $\mu \geq 2\mu_0$

$$\frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell) \left\{ \dot{X}_i^\intercal (\hat{\gamma}_{t_\ell} - \gamma_{t_\ell}) \right\}^2 \leq \frac{3}{2} \mu \|\gamma_{t_\ell}\|_1.$$

The proof of this statement follows along the same lines as the corresponding statement for the Lasso and is omitted for brevity. For a text book treatment, see Bühlmann and van de Geer (2011) section 6.

Regarding the probability of the event $\mathcal{T}$:

Lemma 72. *Suppose that Assumptions 16 and 18 hold. Then, we have*

$$\max_{\ell=1,\dots,L} \max_{j=1,\dots,p} \left| \frac{2}{nh} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \gamma_{t_\ell}) \dot{X}_{ij} \right| = O_p\left(\sqrt{\frac{\log Lp}{nh}} \right).$$

Proof. For $\ell = 1, \dots, L$, $j = 1, \dots, p$, we have

$$\begin{aligned}
&\frac{2}{\sqrt{nh}} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \gamma_{t_\ell}) \dot{X}_{ij} \\
=&\frac{2}{\sqrt{nh}} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)\{1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}\}(X_{ij} - \mathrm{E}[X_{ij}]) && (3.45) \\
&+ \frac{2}{\sqrt{nh}} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)\{1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}\}(\mathrm{E}[X_{ij}] - \bar{X}_j) && (3.46) \\
&+ \frac{2}{\sqrt{nh}} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)(X_{ij} - \mathrm{E}[X_{ij}])(\bar{X} - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell} && (3.47)
\end{aligned}$$

$$+\frac{2}{\sqrt{nh}}\sum_{i=1}^{n}\mathbb{1}(T_i\in\mathcal{I}_\ell)(\bar{X}-\mathrm{E}[X_{ij}])(\bar{X}-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell}. \tag{3.48}$$

Note that (3.45) is a sum of independent bounded and mean zero random variables with bound

$$\begin{aligned}|Z_{ij}| &:= \frac{2}{\sqrt{nh}}|\mathbb{1}(T_i\in\mathcal{I}_\ell)\{1-(X_i-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell}\}(X_{ij}-\mathrm{E}[X_{ij}])|\\ &\le \frac{2}{\sqrt{nh}}(1+\max_\ell\|\gamma_{t_\ell}\|_1)=o\bigg(\frac{1}{\sqrt{\log Lp}}\bigg)\end{aligned}$$

and variance

$$\begin{aligned}\mathrm{Var}(Z_{ij}) &\le \frac{4}{nh}\,\mathrm{E}[\mathbb{1}(T_i\in\mathcal{I}_\ell)\{1-(X_i-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell}\}^2]\\ &\le \frac{4}{nh}\,\mathrm{E}[\mathbb{1}(T_i\in\mathcal{I}_\ell)]=O\bigg(\frac{1}{n}\bigg)\end{aligned}$$

where we have used that f_T is bounded from above. Hence, Lemma 59(i) implies

$$\max_{\ell,j}\bigg|\frac{2}{\sqrt{nh}}\sum_{i=1}^{n}\mathbb{1}(T_i\in\mathcal{I}_\ell)\{1-(X_i-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell}\}(X_{ij}-\mathrm{E}[X_{ij}])\bigg|=O_p(\sqrt{\log Lp}).$$

For (3.46), we have

$$\begin{aligned}&\bigg|\frac{2}{\sqrt{nh}}\sum_{i=1}^{n}\mathbb{1}(T_i\in\mathcal{I}_\ell)\{1-(X_i-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell}\}(\mathrm{E}[X_{ij}]-\bar{X}_j)\bigg|\\ \le&\underbrace{\bigg|\frac{2}{\sqrt{nh}}\sum_{i=1}^{n}\mathbb{1}(T_i\in\mathcal{I}_\ell)\{1-(X_i-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell}\}\bigg|}_{=O_p(\sqrt{nh})}\underbrace{\big\|\bar{X}-\mathrm{E}[X_i]\big\|_\infty}_{=O_p(\sqrt{\log p/n})}=o_p\big(\sqrt{\log Lp}\big).\end{aligned}$$

For (3.47), we have

$$\begin{aligned}&\bigg|\frac{2}{\sqrt{nh}}\sum_{i=1}^{n}\mathbb{1}(T_i\in\mathcal{I}_\ell)(X_{ij}-\mathrm{E}[X_{ij}])(\bar{X}-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell}\bigg|\\ \le&\underbrace{\max_\ell\frac{2}{\sqrt{nh}}\sum_{i=1}^{n}\mathbb{1}(T_i\in\mathcal{I}_\ell)}_{=O_p(\sqrt{nh})}\max_\ell|(\bar{X}-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell}|\\ =&O_p\big(\sqrt{h\log Lp}\big)=o_p\big(\sqrt{\log Lp}\big).\end{aligned}$$

Here, we have used that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}$$

is a sum of independent, bounded and mean zero random variables with bound

$$\frac{1}{\sqrt{n}}|(X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}| \leq \frac{1}{\sqrt{n}} \max_\ell \|\gamma_{t_\ell}\|_1 = o\bigg(\frac{1}{\sqrt{\log(Lp)}}\bigg)$$

and variance (see Lemma 67)

$$\mathrm{Var}((X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}) = O(1)$$

Hence, Lemma 59(i) implies that

$$\max_{\ell,j}|(\bar{X} - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}| = O_p\bigg(\sqrt{\frac{\log Lp}{n}}\bigg). \tag{3.49}$$

For (3.48), we have analogously

$$\begin{aligned}
&\max_{\ell,j}\bigg|\frac{2}{\sqrt{nh}}\sum_{i=1}^{n}\mathbb{1}(T_i \in \mathcal{I}_\ell)(\bar{X} - \mathrm{E}[X_{ij}])(\bar{X} - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}\bigg| \\
\leq &\underbrace{\max_\ell \frac{2}{\sqrt{nh}}\sum_{i=1}^{n}\mathbb{1}(T_i \in \mathcal{I}_\ell)}_{=O_p(\sqrt{nh})} \underbrace{\|\bar{X} - \mathrm{E}[X_i]\|_\infty^2}_{=O_p(\log p/n)} \max_\ell \|\gamma_{t_\ell}\|_1 = o_p\big(\sqrt{\log Lp}\big).
\end{aligned}$$

The claim follows. □

Lemma (71) and (72) imply:

Corollary 13. *Under Assumptions 16 - 18, we have*

$$\max_{\ell=1,\ldots,L}\frac{1}{nh}\sum_{i=1}^{n}\mathbb{1}(T_i \in \mathcal{I}_\ell)\big\{\dot{X}_i^\intercal(\hat{\gamma}_{t_\ell} - \gamma_{t_\ell})\big\}^2 = O_p\bigg(\sqrt{\frac{\log Lp}{nh}}\max_{\ell=1,\ldots,L}\|\gamma_{t_\ell}\|_1\bigg).$$

The next result shows that the prediction error of $\hat{\gamma}_{t_\ell}$ also converges in expectation to zero for μ chosen as a sufficiently large multiple of $\sqrt{\frac{\log Lp}{nh}}$.

Lemma 73. *Suppose that Assumptions 16 - 18 hold. Then*

$$\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(T_i \in \mathcal{I}_\ell)(\dot{X}_i^\intercal(\hat{\gamma}_{t_\ell} - \gamma_{t_\ell}))^2\right] \to 0$$

and

$$\mathrm{E}[(\dot{X}_i^\intercal(\hat{\gamma}_{t_\ell} - \gamma_{t_\ell}))^2 \,|\, T_i \in \mathcal{I}_\ell] \to 0.$$

Proof. The basic inequality for the Lasso, i.e.,

$$\begin{aligned}&\frac{1}{nh}\sum_{i=1}^{n}\mathbb{1}(T_i \in \mathcal{I}_\ell)(X_i^\intercal(\hat{\gamma_{t_\ell}} - \gamma_{t_\ell}))^2 + \mu\|\hat{\gamma_{t_\ell}}\|_1\\ \leq&\frac{2}{n}\sum_{i=1}^{n}(1 - X_i^\intercal\gamma_{t_\ell})X_i^\intercal(\hat{\gamma}_{t_\ell} - \gamma_{t_\ell}) + \lambda\|\gamma_{t_\ell}\|_1\end{aligned}$$

implies together with Hölder's inequality

$$\begin{aligned}\frac{1}{nh}\sum_{i=1}^{n}\mathbb{1}(T_i \in \mathcal{I}_\ell)(X_i^\intercal(\hat{\gamma_{t_\ell}} - \gamma_{t_\ell}))^2 + \mu\|\hat{\gamma_{t_\ell}}\|_1 \leq&M_n\|\hat{\gamma}_{t_\ell} - \gamma_{t_\ell}\|_1 + \mu\|\gamma_{t_\ell}\|_1\\ \leq&M_n\|\hat{\gamma}_{t_\ell}\|_1 + (M_n + \mu)\|\gamma_{t_\ell}\|_1\end{aligned}$$

where

$$M_n = \left\|\frac{2}{n}\sum_{i=1}^{n}(1 - X_i^\intercal\gamma_{t_\ell})X_i^\intercal\right\|_\infty.$$

On the event $\{M_n \leq \mu\}$, this implies

$$\begin{aligned}\frac{1}{nh}\sum_{i=1}^{n}\mathbb{1}(T_i \in \mathcal{I}_\ell)(X_i^\intercal(\hat{\gamma_{t_\ell}} - \gamma_{t_\ell}))^2 \leq&(M_n - \mu)\|\hat{\gamma}_{t_\ell}\|_1 + (M_n + \mu)\|\gamma_{t_\ell}\|_1\\ \leq&2\mu\|\gamma_{t_\ell}\|\end{aligned}$$

and therefore

$$\mathrm{E}\left[\frac{1}{nh}\sum_{i=1}^{n}\mathbb{1}(T_i \in \mathcal{I}_\ell)(X_i^\intercal(\hat{\gamma_{t_\ell}} - \gamma_{t_\ell}))^2\mathbb{1}(M_n \leq \mu)\right] \leq 2\mu\|\gamma_{t_\ell}\|_1.$$

On the alternative $\{M_n > \mu\}$, we have by definition of the Lasso and feasibility of

the zero vector

$$\mu\|\hat{\gamma}_{t_\ell}\|_1 \leq \frac{1}{nh}\sum_{i=1}^{n}\mathbb{1}(T_i \in \mathcal{I}_\ell).$$

This implies together with our basic inequality above

$$\begin{aligned}&\mathrm{E}\bigg[\frac{1}{nh}\sum_{i=1}^{n}\mathbb{1}(T_i \in \mathcal{I}_\ell)(X_i^\intercal(\hat{\gamma}_{t_\ell} - \gamma_{t_\ell}))^2\mathbb{1}(M_n \leq \mu)\bigg]\\ \leq&\frac{\mathrm{E}[M_n\frac{1}{nh}\sum_{i=1}^{n}\mathbb{1}(T_i \in \mathcal{I}_\ell)\mathbb{1}(M_n > \mu)]}{\mu} + (\mathrm{E}[M_n] + \mu)\|\gamma_{t_\ell}\|_1.\end{aligned}$$

By choosing the constant in μ in Assumption 18 sufficiently large, the second part on the right-hand side is of the same order as $\mu\|\gamma_{t_\ell}\|_1$ and therefore converges to zero. Further, we can bound the first part using the Cauchy-Schwarz inequality

$$\begin{aligned}&\frac{\mathrm{E}[M_n\frac{1}{nh}\sum_{i=1}^{n}\mathbb{1}(T_i \in \mathcal{I}_\ell)\mathbb{1}(M_n > \mu)]}{\mu}\\ \leq&\sqrt{\frac{\mathrm{E}[M_n^2/\mu^2\mathbb{1}(M_n > \mu)]\,\mathrm{Var}(\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\mathbb{1}(T_i \in \mathcal{I}_\ell))}{nh}}\\ &+\mathrm{E}[M_n/\mu\mathbb{1}(M_n > \mu)]\frac{\mathrm{P}(T_i \in \mathcal{I}_\ell)}{h}\end{aligned}$$

By the arguments in the proof of Lemma 72, M_n (or actually its upper bound which also maximizes over the locations ℓ) is sub-Gaussian and satisfies for all $t > 0$

$$\mathrm{P}(M_n - \mathrm{E}[M_n] > t) \leq C\exp(-Cnht^2)$$

for some suitably chosen constants C. In particular, also

$$\mathrm{P}\bigg(\frac{M_n - \mathrm{E}[M_n]}{\mathrm{E}[M_n]} > t\bigg) \leq C\exp(-Cnh\,\mathrm{E}[M_n]^2t^2) \quad \text{for all } t > 0.$$

Since $\mathrm{E}[M_n] = O(\sqrt{\frac{\log Lp}{nh}})$ and μ is a sufficiently large multiple of $\mathrm{E}[M_n]$, we have $\mathrm{P}(M_n > \mu) \to 0$. Moreover, by the sub-Gaussian concentration inequality, one can show using analogous arguments as in the proof of Corollary 3.2 in Ledoux and Talagrand (2013) that $\mathrm{E}[((M_n - \mathrm{E}[M_n])/\mathrm{E}[M_n])^2]$ is bounded by some constant which is independent of ℓ and p. Hence, by the bounded convergence theorem $\mathrm{E}[M_n/\mu\mathbb{1}(M_n > \mu)]$ and $\mathrm{E}[(M_n/\mu)^2\mathbb{1}(M_n > \mu)]$ both converge to zero which proves

the claim. □

3.E Technical Results on the Orthogonalized NWE

Lemma 74. *Let Assumption 16 hold. Then, for any $g \in \mathcal{H}(\eta, M)$ with $\eta \in (2, 3]$ and any ℓ such that $t_\ell \in [h, 1-h]$, it holds*

$$\begin{aligned}
&\frac{1}{h}\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell})] \\
=&2f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell}) + \frac{h^2}{3}\frac{\partial^2 f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})}{\partial t^2} + o(h^2) \qquad (3.50) \\
&\frac{1}{h}\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell})(T_i - t_\ell)] \\
=&\frac{2}{3}h^2\frac{\partial f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})}{\partial t} + o(h^2) \qquad (3.51) \\
&\frac{1}{h}\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell})(T_i - t_\ell)^2] \\
=&\frac{2}{3}h^2 f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell}) + o(h^2) \qquad (3.52)
\end{aligned}$$

Proof. Regarding (3.50), by Fubini's theorem

$$\begin{aligned}
&\frac{1}{h}\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell})] \\
=&\int \frac{1}{h}\int \mathbb{1}(t \in \mathcal{I}_\ell) f_{T|X}(t, x)dt(1 - x^\intercal \gamma_{t_\ell}) f_X(x)dx.
\end{aligned}$$

For the inner integral, standard arguments imply the existence of some intermediate values $\tau(v)$ between t_ℓ and $t_\ell + hv$ so that

$$\frac{1}{h}\int \mathbb{1}(t \in \mathcal{I}_\ell) f_{T|X}(t, x)dt = 2f_{T|X}(t_\ell, x) + \frac{h^2}{2} f''_{T|X}(t_\ell, x)\int_{-1}^{1} v^2 dv + R_\ell(x),$$

where R_ℓ is given by

$$R_\ell(x) = \frac{h^2}{2}\int_{-1}^{1}\{f''_{T|X}(\tau, x) - f''_{T|X}(t_\ell, x)\}v^2 dv.$$

Since $f_{T|X}(\cdot, x) \in \mathcal{H}(\eta_X, M_X)$ uniformly over x for some $\eta_X \in (2, 3]$, we can bound

R_ℓ by

$$|R_\ell(x)| \leq \frac{M_X h^{\eta_X}}{2} \int_{-1}^{1} |v|^{\eta_X} dv = \frac{M_X h^{\eta_X}}{\eta_X + 1}.$$

Thus,

$$\begin{aligned}
&\frac{1}{h} \mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell})] \\
=&2 \int (1 - x^\intercal \gamma_{t_\ell}) f_{T|X}(t_\ell, x) f_X(x) dx + \frac{h^2}{3} \int f''_{T|X}(t_\ell, x)(1 - x^\intercal \gamma_{t_\ell}) f_X(x) dx \\
&+ \int R_\ell(x)(1 - x^\intercal \gamma_{t_\ell}) f_X(x) dx.
\end{aligned}$$

For the first term on the right-hand side, it holds

$$\begin{aligned}
&\int (1 - x^\intercal \gamma_{t_\ell}) f_{T|X}(t_\ell, x) f_X(x) dx \\
=&f_T(t_\ell) \, \mathrm{E}[(1 - X_i^\intercal \gamma_{t_\ell}) \,|\, T_i = t_\ell] = f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})
\end{aligned}$$

and for the last term

$$\left| \int R_\ell(x)(1 - x^\intercal \gamma_{t_\ell}) f_X(x) dx \right| \leq \frac{M_X h^{\eta_X}}{\eta_X + 1} \mathrm{E}[|1 - X_i^\intercal \gamma_{t_\ell}|] = o(h^2),$$

where the last bound follows from Lemma 68. The integral in the second term can be rewritten as

$$\begin{aligned}
\int f''_{T|X}(t_\ell, x)(1 - x^\intercal \gamma_{t_\ell}) f_X(x) dx =& \frac{\partial^2}{\partial t^2} \int f_{T|X}(t, x)(1 - x^\intercal \gamma_{t_\ell}) f_X(x) dx \Big|_{t=t_\ell} \\
=& \frac{\partial^2 f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})}{\partial t^2}
\end{aligned}$$

where we can take out the derivative by Leibniz' rule. Hence (3.50) follows.

The other two results follow similarly. In particular, we have for (3.51)

$$\begin{aligned}
&\frac{1}{h} \mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell})(T_i - t_\ell)] \\
=& \int \frac{1}{h} \int \mathbb{1}(t \in \mathcal{I}_\ell)(t - t_\ell) f_{T|X}(t, x) dt (1 - x^\intercal \gamma_{t_\ell}) f_X(x) dx.
\end{aligned}$$

The inner integral satisfies

$$\frac{1}{h}\int \mathbb{1}(t \in \mathcal{I}_\ell)(t-t_\ell) f_{T|X}(t,x)dt = h^2 f'_{T|X}(t_\ell, x)\int_{-1}^{1} v^2 dv + R_{\ell,2}(x),$$

where

$$R_{\ell,2}(x) = h^2 \int \{f'_{T|X}(\tau_2(v), x) - f'_{T|X}(t_\ell, x)\} v^2 dv$$

for some intermediate value $\tau_2(v)$ between t_ℓ and $t_\ell + hv$. By similar arguments as in the first part of the proof, one can show that $|R_{\ell,2}(x)| \leq Ch^3$ uniformly over x. Hence, we have for the whole expression

$$\begin{aligned}&\frac{1}{h}\,\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell})(T_i - t_\ell)] \\ =&\frac{2}{3}h^2 \int f'_{T|X}(t_\ell, x)(1 - x^\intercal \gamma_{t_\ell}) f_X(x)dx + o(h^2),\end{aligned}$$

where the bound on the remainder term follows as in the first part of the proof. The integral on the right-hand side can be written as

$$\begin{aligned}\int f'_{T|X}(t_\ell, x)(1 - x^\intercal \gamma_{t_\ell}) f_X(x)dx =& \frac{\partial}{\partial t}\int (1 - x^\intercal \gamma_{t_\ell}) f_{T|X}(t,x) f_X(x)dx\Big|_{t=t_\ell} \\ =& \frac{\partial}{\partial t} f_T(t)(1 - m_X(t)^\intercal \gamma_{t_\ell})\big|_{t=t_\ell} \\ =& f'_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell}) + f_T(t_\ell) m'_X(t_\ell)^\intercal \gamma_{t_\ell}\end{aligned}$$

where we can take out the derivative since the integrand is continuous with continuous partial derivative with respect to t. This proves (3.51).

Finally, for (3.52), we repeat the above argument:

$$\begin{aligned}&\frac{1}{h}\,\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell})(T_i - t_\ell)^2] \\ =&\int \frac{1}{h}\int \mathbb{1}(T_i \in \mathcal{I}_\ell)(t - t_\ell)^2 f_{T|X}(t,x)dt(1 - x^\intercal \gamma_{t_\ell}) f_X(x)dx\end{aligned}$$

and the inner integral satisfies

$$\frac{1}{h}\int \mathbb{1}(T_i \in \mathcal{I}_\ell)(t-t_\ell)^2 f_{T|X}(t,x)dt = h^2 f_{T|X}(t_\ell, x)\int_{-1}^{1} v^2 dv + R_{\ell,3}(x)$$

where $R_{\ell,3}(x)$ satisfies $|R_{\ell,3}(x)| \leq Ch^3$ uniformly over x for some constant C. Plugging this into the above expression yields

$$\begin{aligned}&\frac{1}{h}\,\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1-(X_i-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell})(T_i-t_\ell)^2]\\ =&\frac{2}{3}h^2\int(1-x^\intercal\gamma_{t_\ell})f_{T|X}(t_\ell,x)f_X(x)dx+o(h^2),\end{aligned}$$

where the remainder term can be dealt with by the same arguments as for the other two expressions. The leading integral can be written as

$$\begin{aligned}&\int(1-x^\intercal\gamma_{t_\ell})f_{T|X}(t_\ell,x)f_X(x)dx\\ =&f_T(t_\ell)\,\mathrm{E}[(1-X_i^\intercal\gamma_{t_\ell})\,|\,T_i=t_\ell]=f_T(t_\ell)(1-m_X(t_\ell)^\intercal\gamma_{t_\ell})\end{aligned}$$

which proves (3.52). □

Lemma 75. *Under the same assumptions as in Theorem 17, it holds*

$$\begin{aligned}&\frac{\sum_{i=1}^n\mathbb{1}(T_i\in\mathcal{I}_\ell)(1-\dot{X}_i^\intercal\hat{\gamma}_{t_\ell})(T_i-t_\ell)}{\sum_{i=1}^n\mathbb{1}(T_i\in\mathcal{I}_\ell)(1-\dot{X}_i^\intercal\hat{\gamma}_{t_\ell})}\\ =&h^2\frac{1}{3}\frac{\frac{\partial f_T(t_\ell)(1-m_X(t_\ell)^\intercal\gamma_{t_\ell})}{\partial t}}{f_T(t_\ell)(1-m_X(t_\ell)^\intercal\gamma_{t_\ell})}+o_p\bigg(\frac{1}{\sqrt{nh}}+h^2\bigg).\end{aligned}$$

Proof. It is sufficient to only consider the numerator. The denominator is asymptotically bounded away from zero by Lemma 77 and Assumption 18. Decompose the numerator as follows

$$\begin{aligned}&\frac{1}{nh}\sum_{i=1}^n\mathbb{1}(T_i\in\mathcal{I}_\ell)(1-\dot{X}_i^\intercal\hat{\gamma}_{t_\ell})(T_i-t_\ell)\\ =&\frac{1}{h}\,\mathrm{E}[\mathbb{1}(T_i\in\mathcal{I}_\ell)(1-X_i^\intercal\gamma_{t_\ell})(T_i-t_\ell)]\\ &+\frac{1}{nh}\sum_{i=1}^n\{\mathbb{1}(T_i\in\mathcal{I}_\ell)(1-X_i^\intercal\gamma_{t_\ell})(T_i-t_\ell)-\mathrm{E}[\mathbb{1}(T_i\in\mathcal{I}_\ell)(1-X_i^\intercal\gamma_{t_\ell})(T_i-t_\ell)]\}\\ &+\frac{1}{nh}\sum_{i=1}^n\mathbb{1}(T_i\in\mathcal{I}_\ell)(T_i-t_\ell)\bar{X}_n^\intercal\gamma_{t_\ell}\\ &+\frac{1}{nh}\sum_{i=1}^n\mathbb{1}(T_i\in\mathcal{I}_\ell)(T_i-t_\ell)\dot{X}_i^\intercal(\gamma_{t_\ell}-\hat{\gamma}_{t_\ell}).\end{aligned}$$

The first term satisfies by Lemma 74

$$\frac{1}{h}\operatorname{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - X_i^\intercal \gamma_{t_\ell})(T_i - t_\ell)] = \frac{2}{3}h^2 \frac{\partial f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})}{\partial t} + o(h^2).$$

The second term is of smaller order than $(nh)^{-1/2}$ by a variance bound, and the third term is of smaller order by Lemma 59 and as $\bar{X}_n^\intercal \gamma_{t_\ell} = o_p(1)$.

It remains to bound the fourth term. Here, we have to be careful as the summands are not independent. Luckily, $\hat{\gamma}_{t_\ell}$ does only depend on $X = (X_1, \ldots, X_n)$ and $\mathbb{1}_\ell = (\mathbb{1}(T_1 \in \mathcal{I}_\ell), \ldots, \mathbb{1}(T_n \in \mathcal{I}_\ell))$, while the T_i are still independent across i conditionally on X and $\mathbb{1}_\ell$. Thus, center the fourth term

$$\begin{aligned}
S_n =& \frac{1}{nh}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(T_i - t_\ell)\dot{X}_i^\intercal(\gamma_{t_\ell} - \hat{\gamma}_{t_\ell}) \\
=& \frac{1}{nh}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)\{(T_i - t_\ell) - \operatorname{E}[(T_i - t_\ell) \mid X, \mathbb{1}_\ell]\}\dot{X}_i^\intercal(\gamma_{t_\ell} - \hat{\gamma}_{t_\ell}) \\
&+ \frac{1}{nh}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)\operatorname{E}[(T_i - t_\ell) \mid X, \mathbb{1}_\ell]\dot{X}_i^\intercal(\gamma_{t_\ell} - \hat{\gamma}_{t_\ell}).
\end{aligned}$$

The centered sum on the right-hand side satisfies

$$\begin{aligned}
\operatorname{Var}(S_n \mid X, \mathbb{1}_\ell) =& \frac{1}{(nh)^2}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)\operatorname{Var}((T_i - t_\ell) \mid X, \mathbb{1}_\ell)(\dot{X}_i^\intercal(\hat{\gamma}_{t_\ell} - \gamma_{t_\ell}))^2 \\
\leq& \frac{1}{n^2}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(\dot{X}_i^\intercal(\hat{\gamma}_{t_\ell} - \gamma_{t_\ell}))^2,
\end{aligned}$$

where we have used that $\operatorname{Var}((T_i - t_\ell) \mid X, \mathbb{1}_\ell) \leq \operatorname{E}[(T_i - t_\ell)^2 \mid X, \mathbb{1}_\ell] \leq Ch^2$. Thus, the expectation of the conditional variance of the centered sum S_n can be bounded by

$$\operatorname{E}[\operatorname{Var}(S_n \mid X, \mathbb{1}_\ell)] = \frac{h}{n}\operatorname{E}\left[\frac{1}{nh}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(\dot{X}_i^\intercal(\hat{\gamma}_{t_\ell} - \gamma_{t_\ell}))^2\right] = o\left(\frac{1}{nh}\right).$$

This implies by the Chebychev inequality on the conditional distribution and the law of iterated expectations that $\sqrt{nh}S_n = o_p(1)$ unconditionally. Next, we deal with the conditional mean of S_n. By independence of our data and standard arguments

$$\frac{1}{h}\operatorname{E}[(T_i - t_\ell) \mid X, \mathbb{1}_\ell] = \frac{2}{3}h^2 f'_{T|X}(t_\ell, X_i) + R(X_i)$$

where the remainder R satisfies $\|R\|_\infty \leq Ch^{\eta_X} = o(h^2)$. Hence,

$$\begin{aligned}|\mathrm{E}[S_n \,|\, X, \mathbb{1}_\ell]| \leq & h^3 \left| \frac{1}{nh} \sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell) \big(2/3 f'_{T|X}(t_\ell, X_i) + R(X_i)\big) \dot{X}_i^\intercal (\gamma_{t_\ell} - \hat{\gamma}_{t_\ell}) \right| \\ \leq & h^3 \sqrt{\frac{1}{nh} \sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell) \big(2/3 f'_{T|X}(t_\ell, X_i) + R(X_i)\big)^2} \\ & \times \sqrt{\frac{1}{nh} \sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell) (\dot{X}_i^\intercal (\gamma_{t_\ell} - \hat{\gamma}_{t_\ell}))^2}\end{aligned}$$

where the second inequality follows by the Cauchy-Schwarz inequality. This implies for the unconditional mean by a further application of Cauchy-Schwarz

$$\begin{aligned}|\mathrm{E}[S_n]| \leq & h^3 \underbrace{\sqrt{\mathrm{E}\left[\frac{1}{nh} \sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell) \big(2/3 f'_{T|X}(t_\ell, X_i) + R(X_i)\big)^2\right]}}_{=O(1)} \\ & \times \underbrace{\sqrt{\mathrm{E}\left[\frac{1}{nh} \sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell) (\dot{X}_i^\intercal (\gamma_{t_\ell} - \hat{\gamma}_{t_\ell}))^2\right]}}_{=o(1)}\end{aligned}$$

and therefore $\sqrt{nh}\,\mathrm{E}[S_n] = o_p(1)$. The claim follows. □

Lemma 76. *Suppose the Assumptions in Theorem 17 hold. Then*

$$\begin{aligned}& \frac{1}{nh} \sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell}) \varepsilon_i \\ = & \frac{1}{nh} \sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - X_i^\intercal \gamma_{t_\ell}) \varepsilon_i + o_p\left(\frac{1}{\sqrt{nh}} + h^2\right).\end{aligned}$$

Proof. We can decompose

$$\begin{aligned}\frac{1}{\sqrt{nh}} \sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell}) \varepsilon_i = & \frac{1}{\sqrt{nh}} \sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - X_i^\intercal \gamma_{t_\ell}) \varepsilon_i \\ & + \frac{1}{\sqrt{nh}} \sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell) \dot{X}_i^\intercal (\gamma_{t_\ell} - \hat{\gamma}_{t_\ell}) \varepsilon_i \quad (3.53)\end{aligned}$$

$$+ \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell) \varepsilon_i \big(\bar{X} - \mathrm{E}[X_i]\big)^\intercal \gamma_{t_\ell} \tag{3.54}$$

First, consider (3.53). Denote by A the event

$$A = \left\{ \max_{\ell=1,\dots,L} \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell) \big\{ \dot{X}_i^\intercal (\hat{\gamma}_{t_\ell} - \gamma_{t_\ell}) \big\}^2 \leq \frac{3}{2} \mu \max_{\ell=1,\dots,L} \|\gamma_{t_\ell}\|_1 \right\}.$$

On A, we have by conditional sub-Gaussianity of ε_i for any $t > 0$

$$\begin{aligned} &\mathrm{P}\left(\left| \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell) \dot{X}_i^\intercal (\hat{\gamma}_{t_\ell} - \gamma_{t_\ell}) \varepsilon_i \right| \geq t \,\middle|\, T, X \right) \\ \leq & 2 \exp\left(- \frac{t^2}{K_\varepsilon^2 \frac{3}{2} \mu \max_{\ell=1} \|\gamma_{t_\ell}\|_1} \right). \end{aligned}$$

The probability of the complement of A is well-behaved by Lemma 72 and hence by a union bound and the law of iterated expectations

$$\max_{\ell=1,\dots,L} \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell) \dot{X}_i^\intercal (\gamma_{t_\ell} - \hat{\gamma}_{t_\ell}) \varepsilon_i = o_p\left(\frac{1}{\log Lp} \right).$$

For (3.54), it follows by Lemma 59

$$\max_{\ell=1,\dots,L} \left| \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell) \varepsilon_i \right| = O_p(\sqrt{\log L}).$$

Furthermore, by (3.49), we have

$$\max_{\ell=1,\dots,L} |\big(\bar{X} - \mathrm{E}[X_i]\big)^\intercal \gamma_{t_\ell}| = O_p\left(\sqrt{\frac{\log Lp}{n}} \right).$$

Hence, (3.54) satisfies

$$\max_{\ell=1,\dots,L} \left| \frac{1}{\sqrt{nh}\sigma_\ell} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell) \varepsilon_i \big(\bar{X} - \mathrm{E}[X_i]\big)^\intercal \gamma_{t_\ell} \right| = O_p\left(\frac{\log Lp}{\sqrt{n}} \right) = o_p\left(\frac{1}{\log Lp} \right).$$

This proves the claimed result. □

3.F Technical Results on the Asymptotic Variance Estimator

In this section, we will show that our estimator of the asymptotic variance of the orthogonalized Nadaraya-Watson estimator is consistent as we capture in the following proposition.

Proposition 1. *Suppose that Assumptions 16 - 18 hold. Then*

$$\max_{\ell=1,\dots,L} \frac{|\hat{\sigma}_{n,\ell} - \sigma_{n,\ell}|}{\sigma_{n,\ell}} = o_p(1).$$

Before proving this result, we need some further technical results.

Lemma 77. *Suppose that Assumptions 16 and 18 hold. Then*

$$\max_{\ell=1,\dots,L} \Bigg| \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell}) \\ - \mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)\{1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}\}]/h \Bigg| = O_p(\rho_n),$$

where

$$\rho_n = \sqrt{\sqrt{\frac{\log Lp}{nh}} \max_{\ell=1,\dots,L} \|\gamma_{t_\ell}\|_1}.$$

Further, it holds that

$$\max_{\ell=1,\dots,L} \left| \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell}) - \mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)\{1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}\}]}{\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell})]} \right| \\ = o_p\left(\frac{1}{\log^2 Lp}\right).$$

Proof. Decompose

$$\frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell}) = \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}) \\ + \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)(\bar{X}_n - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}$$

$$+ \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell) \dot{X}_i^\intercal (\hat{\gamma}_{t_\ell} - \gamma_{t_\ell}).$$

By the same arguments as in the proof of Lemma 72, we have for the first term on the right-hand side

$$\max_{\ell=1,\dots,L} \left| \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}) \right.$$
$$\left. - \mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)\{1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}\}]/h \right| = O_p\left(\sqrt{\frac{\log Lp}{nh}} \right)$$

for the second term

$$\underbrace{\frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)}_{=O_p(1)} \underbrace{(\bar{X}_n - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}}_{=O_p\left(\sqrt{\frac{\log p}{n}}\right)} = O_p\left(\sqrt{\frac{\log p}{n}} \max_\ell \|\gamma_{t_\ell}\|_1 \right)$$

and for the third by the Cauchy-Schwarz inequality

$$\begin{aligned}
&\left| \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell) \dot{X}_i^\intercal (\hat{\gamma}_{t_\ell} - \gamma_{t_\ell}) \right|^2 \\
\leq &\underbrace{\frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)}_{=O_p(1)} \underbrace{\frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell) \dot{X}_i^\intercal (\hat{\gamma}_{t_\ell} - \gamma_{t_\ell})^2}_{=O_p(\mu \max_\ell \|\gamma_{t_\ell}\|_1)} \\
= &O_p\left(\mu \max_{\ell=1,\dots,L} \|\gamma_{t_\ell}\|_1 \right).
\end{aligned}$$

This proves the first part of the claim. The second part follows by noting that by Lemma 66 and 67

$$\frac{1}{h} \mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell})] = \frac{1}{h} \mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell})^2] \geq 2c$$

which is bounded away from zero uniformly over ℓ and p. Finally, by the growth conditions on μ and $\|\gamma_{t_\ell}\|_1$, the claimed rate follows. □

Lemma 78. *Suppose that Assumptions 16 and 18 hold. Then*

$$\max_{\ell=1,\dots,L} \left| \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})^2 - \mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)\{1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}\}^2]/h \right|$$

$=O_p(\rho_n)$,

where

$$\rho_n = \sqrt{\sqrt{\frac{\log Lp}{nh}} \max_{\ell=1,\ldots,L} \|\gamma_{t_\ell}\|_1}.$$

Proof. We have

$$\begin{aligned}
&\frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{\ell,\mu})^2 && (3.55)\\
=&\frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \gamma_{t_\ell})^2 \\
&+ \frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)(\dot{X}_i^\intercal \{\gamma_{t_\ell} - \hat{\gamma}_{\ell,\mu}\})^2 \\
&+ \frac{2}{nh}\sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \gamma_{t_\ell})(\dot{X}_i^\intercal \{\gamma_{t_\ell} - \hat{\gamma}_{\ell,\mu}\}). && (3.56)
\end{aligned}$$

By Corollary 13, the second term on the right-hand side satisfies,

$$\max_{\ell=1,\ldots,L} \frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)\big\{\dot{X}_i^\intercal(\hat{\gamma}_{t_\ell} - \gamma_{t_\ell})\big\}^2 = O_p\bigg(\sqrt{\frac{\log Lp}{nh}} \max_{\ell=1,\ldots,L} \|\gamma_{t_\ell}\|_1\bigg)$$

and regarding the first term, we further split

$$\begin{aligned}
&\frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \gamma_{t_\ell})^2 \\
=&\frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \{X_i - \mathrm{E}[X_i]\}^\intercal \gamma_{t_\ell})^2 && (3.57)\\
&+ \frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)(\{\bar{X} - \mathrm{E}[X_i]\}^\intercal \gamma_{t_\ell})^2 && (3.58)\\
&+ \frac{2}{nh}\sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \{X_i - \mathrm{E}[X_i]\}^\intercal \gamma_{t_\ell})\{\bar{X} - \mathrm{E}[X_i]\}^\intercal \gamma_{t_\ell}. && (3.59)
\end{aligned}$$

Demean (3.57)

$$\begin{aligned}
&\frac{1}{nh}\sum_{i=1}^{n}\mathbb{1}(T_i \in \mathcal{I}_\ell)(1-\{X_i-\mathrm{E}[X_i]\}^\intercal\gamma_{t_\ell})^2 \\
=&\,\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)\{1-(X_i-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell}\}^2]/h \\
&+\frac{1}{nh}\sum_{i=1}^{n}\big\{\mathbb{1}(T_i \in \mathcal{I}_\ell)(1-\{X_i-\mathrm{E}[X_i]\}^\intercal\gamma_{t_\ell})^2 \\
&\qquad -\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)\{1-(X_i-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell}\}^2]\big\}.
\end{aligned}$$

The second term is a sum of independent, bounded, mean zero random variables $Z_{i,\ell}$ given by

$$\begin{aligned}
Z_{i,\ell} =&\frac{1}{nh}\big\{\mathbb{1}(T_i \in \mathcal{I}_\ell)(1-\{X_i-\mathrm{E}[X_i]\}^\intercal\gamma_{t_\ell})^2 \\
&\qquad -\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)\{1-(X_i-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell}\}^2]\big\}
\end{aligned}$$

with bound

$$|Z_{i,\ell}| = O\bigg(\frac{1}{nh}\max_\ell\|\gamma_{t_\ell}\|_1^2\bigg)$$

and variance

$$\begin{aligned}
\mathrm{Var}(Z_{i,\ell}) \le&\,\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)\{1-(X_i-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell}\}^2]/h\frac{1}{n^2h}\max_\ell\|\gamma_{t_\ell}\|_1^2 \\
=&O\bigg(\frac{1}{n^2h}\max_\ell\|\gamma_{t_\ell}\|_1^2\bigg).
\end{aligned}$$

Hence, by Bernstein's inequality and a union bound

$$\max_{\ell=1,\ldots,L}\bigg|\sum_{i=1}^{n}Z_{i,\ell}\bigg| = O_p\bigg(\frac{\log L}{nh}\max_\ell\|\gamma_{t_\ell}\|_1^2\bigg).$$

For (3.58) and (3.59), by the same arguments as in the proof of Lemma 72

$$\max_{\ell=1,\ldots,L}\underbrace{\frac{1}{nh}\sum_{i=1}^{n}\mathbb{1}(T_i \in \mathcal{I}_\ell)}_{=O_p(1)}\underbrace{(\{\bar{X}-\mathrm{E}[X_i]\}^\intercal\gamma_{t_\ell})^2}_{=O_p(\log(Lp)/n)} = o_p(\rho_n)$$

$$\max_{\ell=1,\dots,L} \underbrace{\frac{2}{nh}\sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \{X_i - \mathrm{E}[X_i]\}^\intercal \gamma_{t_\ell})}_{=O_p(1)} \underbrace{\{\bar{X} - \mathrm{E}[X_i]\}^\intercal \gamma_{t_\ell}}_{=O_p(\sqrt{\log Lp/n})} = o_p(\rho_n).$$

Finally, (3.56) satisfies

$$\max_{\ell} \left| \frac{2}{nh}\sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \gamma_{t_\ell})(\dot{X}_i^\intercal \{\gamma_{t_\ell} - \hat{\gamma}_{\ell,\mu}\}) \right| = O_p\left(\sqrt{\sqrt{\frac{\log Lp}{nh}} \max_{\ell} \|\gamma_{t_\ell}\|_1} \right)$$

by the Cauchy-Schwarz inequality. The assertion follows. □

Lemma 79. *Suppose that Assumptions 16 - 18 hold. Then*

$$\hat{\sigma}^2 = \sigma^2 + O_p\left(\sqrt{\left(s\frac{\log p}{n} \right) \vee \left(\frac{1}{ng} + g^4 \right)} \right).$$

Proof. Decompose

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} (\tilde{Y}_i - \tilde{X}_i^\intercal \hat{\beta}_\lambda)^2 =& \sigma^2 \\
&+ \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 - \sigma^2 && (3.60)\\
&+ \frac{1}{n}\sum_{i=1}^{n} (\hat{m}_k^*(T_i) - m(T_i))^2 && (3.61)\\
&+ \frac{1}{n}\sum_{i=1}^{n} \left\{ \tilde{X}_i^\intercal (\hat{\beta}_\lambda - \beta_0) \right\}^2 && (3.62)\\
&- \frac{2}{n}\sum_{i=1}^{n} (\hat{m}_k^* - m(T_i)) \left\{ \tilde{X}_i^\intercal (\hat{\beta}_\lambda - \beta_0) \right\} && (3.63)\\
&- \frac{2}{n}\sum_{i=1}^{n} (\hat{m}_k^*(T_i) - m(T_i))\varepsilon_i && (3.64)\\
&- \frac{2}{n}\sum_{i=1}^{n} \left\{ \tilde{X}_i^\intercal (\hat{\beta}_\lambda - \beta_0) \right\} \varepsilon_i, && (3.65)
\end{aligned}$$

where $\hat{m}_k^*(T_i)$ denotes the infeasible Nadaraya-Watson estimator

$$\hat{m}_k^*(t) = \frac{\sum_{i=1}^{n} k_g(T_i - t)(m(T_i) + \varepsilon_i)}{\sum_{i=1}^{n} k_g(T_i + t)}.$$

By the CLT, (3.60) satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 - \sigma^2 = O_p\left(\frac{1}{\sqrt{n}}\right).$$

(3.61) satisfies

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{m}_k^*(T_i) - m(T_i))^2 = O_p\left(\frac{1}{ng} + g^4\right)$$

by the usual kernel regression arguments.

(3.62) satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\tilde{X}_i^\intercal(\hat{\beta}_\lambda - \beta_0)\right\}^2 = O_p\left(s\frac{\log p}{n}\right)$$

by Theorem 16.

(3.63) is of smaller order than the maximum of (3.61) and (3.62) by the Cauchy-Schwarz inequality.

Again, by the Cauchy-Schwarz inequality, (3.64) and (3.65) satisfy

$$\frac{2}{n}\sum_{i=1}^{n}(\hat{m}_k^*(T_i) - m(T_i))\varepsilon_i = O_p\left(\sqrt{\frac{1}{ng} + g^4}\right)$$
$$\frac{2}{n}\sum_{i=1}^{n}\left\{\tilde{X}_i^\intercal(\hat{\beta}_\lambda - \beta_0)\right\}\varepsilon_i = O_p\left(\sqrt{s\frac{\log p}{n}}\right).$$

The claim follows. □

Proof of Propostion 1. We actually show the stronger result

$$\max_{\ell=1,\ldots,L}\frac{|\hat{\sigma}_{n,\ell} - \sigma_{n,\ell}|}{\sigma_{n,\ell}} = o_p\left(\frac{1}{\log^3 Lp}\right),$$

which we need in our derivation of the uniform limiting distribution and the bootstrap.

It holds by Lemma 79

$$\hat{\sigma}_{n,\ell}^2 = \frac{\frac{1}{nh}\sum_{i=1}^{n}\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal\hat{\gamma}_{t_\ell})^2}{(\frac{1}{nh}\sum_{i=1}^{n}\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal\hat{\gamma}_{t_\ell}))^2}\hat{\sigma}_n^2$$

$$= \frac{\frac{1}{nh}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})^2}{(\frac{1}{nh}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell}))^2} \sigma^2(1 + o_p(1)),$$

and therefore it is sufficient to consider the fraction only. For the denominator, we have by Lemmas 66, 77 and 74

$$\max_{\ell=1,\dots,L} \left| \frac{\frac{1}{nh}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell}) - f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})}{f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})} \right| = o_p(1).$$

Thus, by a Delta method type argument

$$\hat{\sigma}^2_{n,\ell} = \frac{\frac{1}{nh}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})^2}{f_T(t_\ell)^2(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})^2} \sigma^2(1 + o_p(1)).$$

For the numerator, we have by Lemmas 66, 78 and 74

$$\max_{\ell=1,\dots,L} \left| \frac{\frac{1}{nh}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})^2 - f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})}{f_T(t_\ell)(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})} \right| = o_p\left(\frac{1}{\log^3 Lp}\right)$$

and therefore

$$\begin{aligned}\hat{\sigma}^2_{n,\ell} =& \frac{\sigma^2}{f_T(t_\ell)^2(1 - m_X(t_\ell)^\intercal \gamma_{t_\ell})}\left(1 + o_p\left(\frac{1}{\log^3 Lp}\right)\right) \\ =& \sigma^2_{n,\ell}\left(1 + o_p\left(\frac{1}{\log^3 Lp}\right)\right)\end{aligned}$$

uniformly with respect to ℓ. This implies on the one hand that asymptotically $\hat{\sigma}^2_{n,\ell}$ is close to $\sigma^2_{n,\ell}$ and bounded away from zero with probability converging to one. In particular, the event $\{\hat{\sigma}^2_{n,\ell} \geq \sigma^2_{n,\ell}/2\}$ has probability converging to one. The map $f(x) = \sqrt{x}$ restricted to $[\min_\ell \sigma^2_{n,\ell}, \infty)$ is continuously differentiable with bounded second derivative. Hence, by a Delta method type argument,

$$\hat{\sigma}_{n,\ell} = \sigma_{n,\ell}\left(1 + o_p\left(\frac{1}{\log^3 Lp}\right)\right)$$

and the claim follows. □

3.G Technical Results on the Test Statistic

3.G.1 Technical Results on the Asymptotic Distribution

Proof of Lemma 58: By Proposition 2.1 in Chernozhukov et al. (2017), it is sufficient to show that there exists a constant $b > 0$ and a sequence $B_n \geq 1$ such that

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[W_{i\ell}^2] \geq b \tag{3.66}$$

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[|W_{i\ell}|^{2+k}] \leq B_n^k \tag{3.67}$$

for $k = 1, 2$ and

$$\mathrm{E}[\exp(|W_{i\ell}|/B_n)] \leq 2 \tag{3.68}$$

as well as

$$B_n = o\left(\sqrt{\frac{n}{\log^7(Ln)}}\right). \tag{3.69}$$

Take $b = 1$ and for some sufficiently large constant B

$$B_n = B\frac{1 + \max_\ell \|\gamma_{t_\ell}\|_1}{\sqrt{h}\min_\ell \sigma_\ell} \vee 1,$$

then (3.66) holds with equality and B_n satisfies both $B_n \geq 1$ and (3.69) by Assumptions 16 - 18. Furthermore, (3.67) follows since

$$\begin{aligned}\mathrm{E}[|W_{i\ell}|^{2+k}] &\leq C\frac{(1 + \|\gamma_{t_\ell}\|_1)^k}{(\sqrt{h}\sigma_\ell)^k}\frac{\mathrm{E}[\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - (X_i - \mathrm{E}[X_i])^\intercal\gamma_{t_\ell})^2|\varepsilon_i|^{2+k}]}{h\sigma_\ell^2(f_T(t\ell)(1 - m_X(t_\ell)^\intercal\gamma_{t_\ell})^{2+k})} \\ &\leq C\frac{(1 + \|\gamma_{t_\ell}\|_1)^k}{(\sqrt{h}\sigma_\ell)^k}3(2+k)\left(\frac{2+k}{2}\right)^{2+k/2}K_\varepsilon^{2+k} \leq B_n^k\end{aligned}$$

where we have used that conditional sub-Gaussianity of the ε_i implies for all $p \geq 1$[2]

$$\mathrm{E}[|\varepsilon_i|^p \,|\, T_i, X_i] \leq 3p\left(\frac{p}{2}\right)^{p/2}K_\varepsilon^p \tag{3.70}$$

[2]See Proposition 2.5.2 in Vershynin (2018).

and finally

$$|W_{i\ell}| \leq \frac{1 + \|\gamma_{t_\ell}\|_1}{\sqrt{h}\sigma_\ell}|\varepsilon_i| \leq \frac{B_n}{48(K_\varepsilon^4 \vee 1)}|\varepsilon_i|$$

implying

$$\mathrm{E}\left[\exp\left(\frac{|W_{i\ell}|}{B_n}\right)\right] \leq \mathrm{E}\left[\exp\left(\frac{|\varepsilon_i|}{C}\right)\right]$$

for $C = 48(K_\varepsilon^4 \vee 1)$. By sub-Gaussianity of the ε_i (cf. Assumption 16), the right-hand side can be bounded by

$$\begin{aligned}
\mathrm{E}\left[\exp\left(\frac{|\varepsilon_i|}{C}\right) - 1\right] &= \int_0^\infty \frac{1}{C}\exp\left(\frac{t}{C}\right)\mathrm{P}(|\varepsilon_i| \geq t)dt \\
&\leq \int_0^\infty \frac{2}{C}\exp\left(-\frac{t^2}{K_\varepsilon^2} + \frac{t}{C}\right)dt \\
&= \frac{2}{C}\exp\left(\frac{K_\varepsilon}{4C^2}\right)\int_0^\infty \exp\left(-\frac{(t - K_\varepsilon/(2C))^2}{K_\varepsilon^2}\right)dt \\
&\leq \frac{2\sqrt{\pi}K_\varepsilon}{C}\exp\left(\frac{K_\varepsilon}{4C^2}\right) < 1
\end{aligned}$$

and hence, (3.68) is satisfied. □

Lemma 80. *Under Assumptions 16 - 18, it holds*

$$\max_{\ell=1,\ldots,L}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \Delta_{i,\ell}\right| = o_p\left(\frac{1}{\log Lp}\right).$$

Proof. By the triangle inequality, we can bound

$$\max_{\ell=1,\ldots,L}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \Delta_{i,\ell}\right| \leq \max_{\ell=1,\ldots,L}|\hat{\mathcal{S}}_{X,\ell}| \tag{3.71}$$

$$+ \max_{\ell=1,\ldots,L}|\hat{\mathcal{S}}_{\varepsilon,\ell} - \mathcal{S}_{\varepsilon,\ell}|. \tag{3.72}$$

As shown in the proof of Theorem 17, the approximation of the denominator converges sufficiently fast to zero when measured by the relative error and hence it is sufficient in the following to consider only the numerators.

For (3.71), the KKT conditions of (3.4) imply uniformly over ℓ

$$\max_{1\le j\le p}\left|\frac{2}{nh}\sum_{i=1}^{n}\mathbb{1}(T_i\in\mathcal{I}_\ell)\big(1-\dot{X}_i^\intercal\hat{\gamma}_\ell\big)\dot{X}_{ij}\right|=\mu$$

with $\mu=O(\sqrt{\log Lp/nh})$ by Assumption 18. Moreover, by Theorem 16, $\|\hat{\beta}_\lambda-\beta\|_1=O_p(s\sqrt{\log p/n})$. This implies the bound

$$|\hat{\mathcal{S}}_{X,\ell}|\le C\sqrt{nh}\max_{1\le j\le p}\left|\frac{2}{nh}\sum_{i=1}^{n}\mathbb{1}(T_i\in\mathcal{I}_\ell)\big(1-\dot{X}_i^\intercal\hat{\gamma}_\ell\big)\dot{X}_{ij}\right|\|\hat{\beta}_\lambda-\beta\|_1=O_p\bigg(\frac{s\log Lp}{\sqrt{n}}\bigg).$$

For (3.72), we can decompose $\hat{\mathcal{S}}_{\varepsilon,\ell}$ into

$$\begin{aligned}\hat{\mathcal{S}}_{\varepsilon,\ell}=&\mathcal{S}_{\varepsilon,\ell}\\&+\frac{\sqrt{nh}}{\hat{\sigma}_\ell}\frac{\sum_{i=1}^n\mathbb{1}(T_i\in\mathcal{I}_\ell)\dot{X}_i^\intercal(\gamma_{t_\ell}-\hat{\gamma}_{t_\ell})\varepsilon_i}{\sum_{i=1}^n\mathbb{1}(T_i\in\mathcal{I}_\ell)(1-\dot{X}_i^\intercal\hat{\gamma}_\ell)}&(3.73)\\&+\frac{\sqrt{nh}}{\hat{\sigma}_\ell}\frac{\sum_{i=1}^n\mathbb{1}(T_i\in\mathcal{I}_\ell)\varepsilon_i\big(\bar{X}-\mathrm{E}[X_i]\big)^\intercal\gamma_{t_\ell}}{\sum_{i=1}^n\mathbb{1}(T_i\in\mathcal{I}_\ell)(1-\dot{X}_i^\intercal\hat{\gamma}_\ell)}&(3.74)\\&+\frac{\sqrt{nh}}{\hat{\sigma}_\ell}\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j&(3.75)\\&+\left(\frac{\sigma_\ell}{\hat{\sigma}_\ell}-1\right)S_{\varepsilon,\ell}&(3.76)\end{aligned}$$

In order to deal with (3.73), let A denote the event

$$A=\bigg\{\max_{\ell=1,\dots,L}\frac{1}{nh}\sum_{i=1}^{n}\mathbb{1}(T_i\in\mathcal{I}_\ell)\big\{\dot{X}_i^\intercal(\hat{\gamma}_{t_\ell}-\gamma_{t_\ell})\big\}^2\le\frac{3}{2}\mu\max_{\ell=1,\dots,L}\|\gamma_{t_\ell}\|_1\bigg\}.$$

On A, we have by conditional sub-Gaussianity of ε_i for any $t>0$

$$\begin{aligned}&\mathrm{P}\bigg(\bigg|\frac{1}{\sqrt{nh}\sigma_\ell}\sum_{i=1}^{n}\mathbb{1}(T_i\in\mathcal{I}_\ell)\dot{X}_i^\intercal(\hat{\gamma}_{t_\ell}-\gamma_{t_\ell})\varepsilon_i\bigg|\ge t\,\bigg|\,T,X\bigg)\\\le&2\exp\bigg(-\frac{\sigma_\ell^2t^2}{K_\varepsilon^2\frac{3}{2}\mu\max_{\ell=1}\|\gamma_{t_\ell}\|_1}\bigg).\end{aligned}$$

The probability of the complement of A is well-behaved by Lemma 72 and hence by

a union bound

$$\max_{\ell=1,\dots,L} \frac{1}{\sqrt{nh}\hat{\sigma}_\ell} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell) \dot{X}_i^\intercal (\gamma_{t_\ell} - \hat{\gamma}_{t_\ell}) \varepsilon_i = o_p\left(\frac{1}{\log Lp}\right).$$

For (3.74), one can show that

$$\max_{\ell=1,\dots,L} \left| \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell) \varepsilon_i \right| = O_p(\sqrt{\log L})$$

by using sub-Gaussianity of ε combined with a union bound. Furthermore, by (3.49), we have

$$\max_{\ell=1,\dots,L} |\big(\bar{X} - \mathrm{E}[X_i]\big)^\intercal \gamma_{t_\ell}| = O_p\left(\sqrt{\frac{\log Lp}{n}}\right).$$

Hence, (3.74) satisfies

$$\max_{\ell=1,\dots,L} \left| \frac{\sqrt{nh}}{\hat{\sigma}_\ell} \frac{\sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell) \varepsilon_i \big(\bar{X} - \mathrm{E}[X_i]\big)^\intercal \gamma_{t_\ell}}{\sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_\ell)} \right| = O_p\left(\frac{\log Lp}{\sqrt{n}}\right) = o_p\left(\frac{1}{\log Lp}\right).$$

(3.75) is of smaller order by the CLT and (3.76) is of smaller order by the bound derived in the proof of Proposition 1. □

3.G.2 Technical Results on the Bootstrap

Lemma 81. *Under Assumptions 16 - 18, it holds*

$$\sup_{q \in \mathbb{R}} \left| \mathrm{P}^*\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (W_i - \overline{W}) e_i \leq q \right) - \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \leq q \right) \right| = o_p(1).$$

Proof. Denote by $\hat{\Sigma}$ the matrix with (ℓ, k)th element

$$\hat{\Sigma}_{\ell,k} = \frac{1}{n} \sum_{i=1}^{n} (W_{i\ell} - \overline{W}_\ell)(W_{ik} - \overline{W}_k),$$

$\Sigma = \mathrm{E}[W_i W_i^\intercal]$ and let $\Delta_{n,r}$ denote $\Delta_{n,r} = \max_{\ell,k} |\hat{\Sigma} - \Sigma|$.

Theorem 4.1 and Remark 4.1 in Chernozhukov et al. (2017) imply that, on the

event $\Delta_{n,r} \leq \bar{\Delta}$, where $\bar{\Delta}$ can be chosen freely,

$$\sup_{q\in\mathbb{R}}\left|\mathrm{P}^*\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(W_i-\overline{W})e_i \leq q\right) - \mathrm{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i \leq q\right)\right| \leq C\bar{\Delta}_n^{1/3}\log^{2/3} 2L.$$

In the following, we will show that

$$\max_{\ell,k}|\hat{\Sigma}_{\ell,k} - \Sigma_{\ell,k}| = o_p\left(\frac{1}{\log^3 L}\right)$$

which is sufficient to prove the claim.

Decompose

$$\hat{\Sigma}_{\ell,k} = \frac{1}{n}\sum_{i=1}^{n} W_{i\ell}W_{ik} - \overline{W}_\ell\overline{W}_k$$

and therefore

$$\max_{\ell,k}|\hat{\Sigma} - \Sigma| \leq \max_{\ell,k}\left|\frac{1}{n}\sum_{i=1}^{n} W_{i\ell}W_{ik} - \mathrm{E}[W_{i\ell}W_{ik}]\right| + \max_{\ell}|\overline{W}_\ell|^2 \tag{3.77}$$

In order to increase readability, let $w_{i\ell}$ denote

$$w_{i\ell} = \frac{1}{\sigma_\ell}\frac{\mathbb{1}(T_i \in \mathcal{I}_\ell)\{1-(X_i-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell}\}}{f_T(t_\ell)(1-m_X(t_\ell)^\intercal\gamma_{t_\ell})}$$

Then we can write (3.77) as

$$\max_{\ell,k}\left|\frac{1}{nh}\sum_{i=1}^{n}\left\{w_{i\ell}w_{ik}\varepsilon_i^2 - \mathrm{E}[w_{i\ell}w_{ik}\varepsilon_i^2]\right\}\right|.$$

By the triangle inequality,

$$\max_{\ell,k}\left|\frac{1}{nh}\sum_{i=1}^{n}\left\{w_{i\ell}w_{ik}\varepsilon_i^2 - \mathrm{E}[w_{i\ell}w_{ik}\varepsilon_i^2]\right\}\right| \leq \max_{\ell,k}\left|\frac{1}{nh}\sum_{i=1}^{n} w_{i\ell}w_{ik}(\varepsilon_i^2-\sigma^2)\right| \tag{3.78}$$

$$+ \max_{\ell,k}\left|\frac{\sigma^2}{nh}\sum_{i=1}^{n}\left\{w_{i\ell}w_{ik} - \mathrm{E}[w_{i\ell}w_{ik}]\right\}\right| \tag{3.79}$$

In order to deal with the first term on the right-hand side, we will use a Bernstein

inequality for weighted sums of subexponentially distributed random variables as given in Theorem 2.8.2 in Vershynin (2018) which we reproduce here for the convenience of the reader.

Theorem 23. *Let $X_1, \dots, X_n$ be independent, mean zero, sub-exponential random variables, and $a = (a_1, \dots, a_n) \in \mathbb{R}^n$. Then, for every $t \geq 0$, we have*

$$\mathrm{P}\bigg(\Big|\sum_{i=1}^n a_i X_i\Big| \geq t\bigg) \leq 2\exp\bigg(-c\min\bigg\{\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\bigg\}\bigg)$$

where $K = \max_i \|X_i\|_{\psi_1}$.

Here, $\|X\|_{\psi_1}$ denotes the smallest number such that $\mathrm{E}[\exp(\|X\|/K)] \leq 2$. If no such number exists, we set $\|X\|_{\psi_1} = \infty$.

We are going to apply Theorem 23 to the conditional distribution, P^*, conditionally on $\{(T_i, X_i) : i = 1, \dots, n\}$. Therefore, $\|\cdot\|_{\psi_1}$ also is to be interpreted with respect to the corresponding conditional expectation. Take $X_i = (\varepsilon_i^2 - \sigma^2)$ and $a_i = \frac{\log^4 L}{nh} w_{i\ell} w_{ik}$ for $i = 1, \dots, n$. Then, $K = \max_i \|\varepsilon_i^2 - \sigma^2\|_{\psi_1} \leq C K_\varepsilon^2$, by Assumption 16, Exercise 2.7.10 in Vershynin (2018) and Lemma 2.7.6 in Vershynin (2018). Furthermore,

$$\|a\|_\infty = O\bigg(\frac{\log^4 L}{nh} \max_\ell \|\gamma_{t_\ell}\|_1^2\bigg) = o\bigg(\frac{1}{\log L}\bigg)$$

and

$$\|a\|_2^2 = O\bigg(\frac{1}{nh}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)\big(1 + (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}\big)^2 \frac{\log^8 L}{nh} \max_\ell \|\gamma_{t_\ell}\|_1^2\bigg).$$

Denote by A_K the event

$$A_K = \bigg\{\max_\ell \frac{1}{nh}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)\big(1 + (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}\big)^2 \leq K\sqrt{\log L}\bigg\},$$

for some constant $K > 0$. On A_K, it holds

$$\|a\|_2^2 \leq C\frac{\log^{8.5} L}{nh} \max_\ell \|\gamma_{t_\ell}\|_1^2.$$

Thus, by a union bound together with Theorem 23, we have on A_K for any $t > 0$

$$\mathrm{P}^*\left(\max_{\ell,k}\left|\frac{1}{nh}\sum_{i=1}^{n} w_{i\ell}w_{ik}(\varepsilon_i^2 - \sigma^2)\right| \geq \frac{t}{\log^3 L}\right) \to 0 \quad \text{a.s.}$$

This implies together with the law of iterated expectations and the dominated convergence theorem

$$\mathrm{P}\left(\max_{\ell,k}\left|\frac{1}{nh}\sum_{i=1}^{n} w_{i\ell}w_{ik}(\varepsilon_i^2 - \sigma^2)\right| \geq \frac{t}{\log^3 L},\, A_C\right) \to 0.$$

By the same arguments as in Lemma 78 in (3.57), we have that $\lim_n \mathrm{P}(A_K^c) = 0$ when K is chosen sufficiently large and therefore

$$\max_{\ell,k}\left|\frac{1}{nh}\sum_{i=1}^{n} w_{i\ell}w_{ik}(\varepsilon_i^2 - \sigma^2)\right| = o_p\left(\frac{1}{\log^3 L}\right).$$

In order to bound (3.79), let

$$Z_{i\ell k} = \frac{\log^4 L}{nh}\{w_{i\ell}w_{ik} - \mathrm{E}[w_{i\ell}w_{ik}]\}.$$

The $Z_{i\ell k}$ have mean zero and are bounded by

$$|Z_{i\ell k}| \leq C\frac{\log^4 L}{nh\sigma_\ell\sigma_k}\big(1 + \max_\ell \|\gamma_{t_\ell}\|_1\big)^2 = O\left(\frac{\log^4 L}{nh}\max_\ell \|\gamma_{t_\ell}\|_1^2\right)$$

which is sufficient for Lemma 59(ii) to be applicable. Their variances can be bounded by

$$\begin{aligned}
&\mathrm{Var}(Z_{i\ell k}) \\
\leq& C\frac{\log^8 L}{nh\sigma_\ell^2\sigma_k^2}\,\mathrm{E}\big[\mathbb{1}(T_i \in \mathcal{I}_\ell)\mathbb{1}(T_i \in \mathcal{I}_k)\big(1 - (X_i - \mathrm{E}[X_i])^\intercal\gamma_{t_\ell}\big)^2\big(1 - (X_i - \mathrm{E}[X_i])^\intercal\gamma_{t_k}\big)^2\big] \\
\leq& C\frac{1}{n\sigma_k^2}\underbrace{\frac{1}{h\sigma_\ell^2}\,\mathrm{E}\big[\mathbb{1}(T_i \in \mathcal{I}_\ell)\big(1 - (X_i - \mathrm{E}[X_i])^\intercal\gamma_{t_\ell}\big)^2\big]}_{=O(1)}\underbrace{\frac{\log^8 L}{nh}\big(2\max_\ell\|\gamma_{t_\ell}\|_1^2 + 2\big)}_{=o(\log^{-1} Lp)}
\end{aligned}$$

which converges sufficiently fast to zero. Hence, by Lemma 12(ii),

$$\max_{\ell,k}\left|\frac{\sigma^2}{nh}\sum_{i=1}^{n}\{w_{i\ell}w_{ik}-\mathrm{E}[w_{i\ell}w_{ik}]\}\right| = o_p\left(\frac{1}{\log^3 L}\right),$$

and

$$\max_{\ell,k}\left|\frac{1}{n}\sum_{i=1}^{n}W_{i\ell}W_{ik}-\mathrm{E}[W_{i\ell}W_{ik}]\right| = o_p\left(\frac{1}{\log^3 L}\right).$$

For the second term on the right-hand side in (3.77), we have by conditional sub-Gaussianity of ε_i

$$\mathrm{P}\big(\sqrt{n}|\overline{W}_\ell| \geq t \,|\, T, X\big) \leq \exp\left(-\frac{t^2}{K_\varepsilon^2 \frac{1}{nh\sigma_\ell^2}\sum_{i=1}^{n}\mathbb{1}(T_i\in\mathcal{I}_\ell)(1-(X_i-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell})^2}\right)$$

By the same arguments as in the proof of Lemma 78, we have

$$\max_{\ell=1,\ldots,L}\frac{1}{nh\sigma_\ell^2}\sum_{i=1}^{n}\mathbb{1}(T_i\in\mathcal{I}_\ell)(1-(X_i-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell})^2 = O_p(\sqrt{\log L}).$$

These statements imply $\max_{\ell,k}|\overline{W}_\ell|^2 = o_p(\log^{-3} L)$ and the claim follows. □

Lemma 82. *Under Assumptions 16 - 18, it holds*

$$\mathrm{P}^*\left(\max_{\ell=1,\ldots,2L}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\Delta_{i\ell}-\bar{\Delta}_\ell)e_i\right| > \varrho_n\right) = o_p(1)$$

with $\varrho_n = 1/\log L$.

Proof. Note that $\hat{\varepsilon}_i = m(T_i) - \hat{m}_k^*(T_i) + \tilde{X}_i^\intercal(\beta_0 - \hat{\beta}_\lambda) + \varepsilon_i$ and decompose

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\Delta_{i\ell}-\bar{\Delta}_\ell)e_i \\
=&\frac{1}{\sqrt{nh}\hat{\sigma}_\ell a_\ell}\sum_{i=1}^{n}\Bigg(\mathbb{1}(T_i\in\mathcal{I}_\ell)(1-\dot{X}_i^\intercal\hat{\gamma}_{t_\ell})\{m(T_i)-\hat{m}_k^*(T_i)\} \\
&\qquad -\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}(T_j\in\mathcal{I}_\ell)(1-\dot{X}_j^\intercal\hat{\gamma}_{t_\ell})\{m(T_j)-\hat{m}_k^*(T_j)\}\Bigg)e_i \qquad (3.80)\\
&+\frac{1}{\sqrt{nh}\hat{\sigma}_\ell a_\ell}\sum_{i=1}^{n}\Bigg(\mathbb{1}(T_i\in\mathcal{I}_\ell)(1-\dot{X}_i^\intercal\hat{\gamma}_{t_\ell})\tilde{X}_i^\intercal(\beta_0-\hat{\beta}_\lambda)
\end{aligned}$$

$$- \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}(T_j \in \mathcal{I}_\ell)(1 - \dot{X}_j^\intercal \hat{\gamma}_{t_\ell})\tilde{X}_j^\intercal(\beta_0 - \hat{\beta}_\lambda)\Bigg) e_i \tag{3.81}$$

$$+ \frac{1}{\sqrt{nh}\hat{\sigma}_\ell a_\ell}\sum_{i=1}^{n}\Bigg(\mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell})\varepsilon_i$$

$$- \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}(T_j \in \mathcal{I}_\ell)(1 - \dot{X}_j^\intercal \hat{\gamma}_{t_\ell})\varepsilon_j\Bigg) e_i \tag{3.82}$$

$$- \frac{1}{\sqrt{nh}\sigma_\ell a_\ell}\sum_{i=1}^{n}\Bigg(\mathbb{1}(T_i \in \mathcal{I}_\ell)\{1 - (X_i - \mathrm{E}[X_i])^\intercal \gamma_{t_\ell}\}\varepsilon_i$$

$$- \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}(T_j \in \mathcal{I}_\ell)\{1 - (X_j - \mathrm{E}[X_j])^\intercal \gamma_{t_\ell}\}\varepsilon_j\Bigg) e_i, \tag{3.83}$$

where a_ℓ is given by

$$a_\ell = \frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}(T_i \in \mathcal{I}_\ell)(1 - \dot{X}_i^\intercal \hat{\gamma}_{t_\ell}).$$

We can derive the rates of the terms (3.80) - (3.83) by the same recurring argument: Note that the terms all are of the form

$$\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}(Z_{i\ell} - \bar{Z}_\ell)e_i$$

where the $Z_{i\ell}$ only depend on $\mathcal{D} := (Y_i, T_i, X_i)_{i=1}^n$ and $\bar{Z}_\ell$ denotes $\bar{Z}_\ell = \frac{1}{n}\sum_{i=1}^n Z_{i\ell}$. Since the e_i are standard normal and independent of $\mathcal{D}$, these sums are (conditionally on $\mathcal{D}$) normally distributed with mean zero and variance

$$\frac{1}{nh}\sum_{i=1}^{n}(Z_{i\ell} - \bar{Z}_\ell)^2$$

Then, by a union bound together with sub-Gaussian tail bounds,

$$\mathrm{P}^*\left(\max_{\ell=1,\dots,L}\left|\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}(Z_{i\ell} - \bar{Z}_\ell)e_i\right| \geq \rho_n\right) \leq 2\exp\left(\log L - \frac{\rho_n^2}{\max_\ell \frac{2}{nh}\sum_{i=1}^n (Z_{i\ell} - \bar{Z}_\ell)^2}\right)$$

Further, we can derive rates for an upper bound of the variance, say,

$$\max_\ell \frac{1}{nh}\sum_{i=1}^{n}(Z_{i\ell} - \bar{Z}_\ell)^2 \leq \max_\ell \frac{1}{nh}\sum_{i=1}^{n} Z_{i\ell}^2 = O_p(r_n) = o_p(\rho_n^3).$$

Here the oh-p statements are to be understood unconditionally. Such a bound is sufficient for the right-hand side to converge to zero in probability. Therefore, we focus in the following on deriving rates for the conditional variances.

An upper bound on the variance of (3.80) is

$$\frac{1}{nh\hat{\sigma}_\ell^2 a_\ell^2}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1-\dot{X}_i^\intercal \hat{\gamma}_{t_\ell})^2\{m(T_i)-\hat{m}_k^*(T_i)\}^2$$

By the usual arguments in kernel regression, one can show

$$\sup_{t\in[0,1]}|\hat{m}_k^*(t)-m(t)| = O_p\left(\sqrt{\frac{\log n}{ng}}+g^2\right).$$

Moreover, by definition of $\hat{\sigma}_\ell^2$,

$$\frac{1}{nh\hat{\sigma}_\ell^2 a_\ell^2}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1-\dot{X}_i^\intercal \hat{\gamma}_{t_\ell})^2 = \frac{1}{\hat{\sigma}^2} = O_p(1)$$

and therefore

$$\max_\ell \frac{1}{nh\hat{\sigma}_\ell^2 a_\ell^2}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1-\dot{X}_i^\intercal \hat{\gamma}_{t_\ell})^2\{m(T_i)-\hat{m}_k^*(T_i)\}^2 = o_p(\rho_n^3).$$

Note that $|\tilde{X}_{ij}| \leq 2$ and therefore the variance of (3.81) can be bounded by

$$\begin{aligned}\max_{\ell=1,\ldots,L}\frac{1}{nh\hat{\sigma}_\ell^2 a_\ell^2}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)(1-\dot{X}_i^\intercal \hat{\gamma}_{t_\ell})^2 \underbrace{\{\tilde{X}_i^\intercal(\hat{\beta}_\lambda-\beta_0)\}^2}_{\leq 4\|\hat{\beta}_\lambda-\beta_0\|_1^2} \\ \leq \frac{4}{\hat{\sigma}^2}\|\hat{\beta}_\lambda-\beta_0\|_1^2 = O_p\left(s^2\frac{\log p}{n}\right) = o_p(\rho_n^3)\end{aligned}$$

where we have used Theorem 16 and Lemma 79.

The third term, (3.82) can be decomposed as follows

$$\begin{aligned}\frac{1}{\sqrt{nh}\hat{\sigma}_\ell a_\ell}\sum_{i=1}^n\Bigg(&\mathbb{1}(T_i \in \mathcal{I}_\ell)(1-\dot{X}_i^\intercal \hat{\gamma}_{t_\ell})\varepsilon_i \\ &-\frac{1}{n}\sum_{j=1}^n \mathbb{1}(T_j \in \mathcal{I}_\ell)(1-\dot{X}_j^\intercal \hat{\gamma}_{t_\ell})\varepsilon_j\Bigg)e_i\end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{\sqrt{nh}\sigma_\ell a_\ell}\sum_{i=1}^{n}\Bigg(\mathbb{1}(T_i\in\mathcal{I}_\ell)\{1-(X_i-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell}\}\varepsilon_i \\
&\qquad-\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}(T_j\in\mathcal{I}_\ell)\{1-(X_j-\mathrm{E}[X_j])^\intercal\gamma_{t_\ell}\}\varepsilon_j\Bigg)e_i \qquad (3.84)\\
&+\left(\frac{1}{\hat\sigma_\ell}-\frac{1}{\sigma_\ell}\right)\frac{1}{\sqrt{nh}a_\ell}\sum_{i=1}^{n}\Bigg(\mathbb{1}(T_i\in\mathcal{I}_\ell)\{1-(X_i-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell}\}\varepsilon_i \\
&\qquad-\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}(T_j\in\mathcal{I}_\ell)\{1-(X_j-\mathrm{E}[X_j])^\intercal\gamma_{t_\ell}\}\varepsilon_j\Bigg)e_i \qquad (3.85)\\
&-\frac{1}{\sqrt{nh}\hat\sigma_\ell a_\ell}\sum_{i=1}^{n}\Bigg(\mathbb{1}(T_i\in\mathcal{I}_\ell)\varepsilon_i-\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}(T_j\in\mathcal{I}_\ell)\varepsilon_j\Bigg)e_i\big((\bar X-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell}\big) \qquad (3.86)\\
&-\frac{1}{\sqrt{nh}\hat\sigma_\ell a_\ell}\sum_{i=1}^{n}\Bigg(\mathbb{1}(T_i\in\mathcal{I}_\ell)\{\dot X_i^\intercal(\hat\gamma_{t_\ell}-\gamma_{t_\ell})\}\varepsilon_i \\
&\qquad-\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}(T_j\in\mathcal{I}_\ell)\{\dot X_i^\intercal(\hat\gamma_{t_\ell}-\gamma_{t_\ell})\}\varepsilon_j\Bigg)e_i \qquad (3.87)
\end{aligned}
$$

The first term, (3.84), cancels with (3.83). For (3.85), we have by Lemma 81

$$
\begin{aligned}
\max_{\ell=1,\dots,L}\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\Bigg(&\mathbb{1}(T_i\in\mathcal{I}_\ell)\{1-(X_i-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell}\}\varepsilon_i \\
&-\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}(T_j\in\mathcal{I}_\ell)\{1-(X_j-\mathrm{E}[X_j])^\intercal\gamma_{t_\ell}\}\varepsilon_j\Bigg)e_i=O_p(\sqrt{\log L})
\end{aligned}
$$

and rate in the proof of Proposition 1 implies

$$
\max_{\ell=1,\dots,L}\left(\frac{1}{\hat\sigma_\ell}-\frac{1}{\sigma_\ell}\right)=O_p(\rho_n^2).
$$

Thus, (3.85) converges sufficiently fast to zero.

An upper bound on the variance of (3.86) is given by

$$
\underbrace{\max_{\ell=1,\dots,L}\frac{1}{nh\hat\sigma_\ell^2a_\ell^2}\sum_{i=1}^{n}\mathbb{1}(T_i\in\mathcal{I}_\ell)\varepsilon_i^2}_{=O_p(1)}\underbrace{\max_{\ell=1,\dots,L}|(\bar X-\mathrm{E}[X_i])^\intercal\gamma_{t_\ell}|^2}_{=O_p(\log Lp/n)}=o_p(\rho_n^3)
$$

where we have used (3.49). For the final term, (3.87), an upper bound for the

variance is given by

$$\frac{1}{nh\hat{\sigma}_\ell^2 a_\ell^2}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)\{\dot{X}_i^\intercal(\hat{\gamma}_{t_\ell} - \gamma_{t_\ell})\}^2\varepsilon_i^2.$$

In order to deal with this term, we will use some further notation: Denote by $\mathcal{D}' := \{(T_i, X_i) : i = 1, \dots, n\}$ and let $\|Z\|_{\psi_1}$, for some random variable Z, denote the smallest constant satisfying

$$\mathrm{E}\left[\exp\left(\frac{|Z|}{\|Z\|_{\psi_1}}\right)\,\middle|\,\mathcal{D}'\right] \leq 2.$$

If no finite constant satisfies this inequality, set $\|Z\|_{\psi_1} = \infty$. By Problem 8 in section 2.2 in van der Vaart and Wellner (2000), it holds for any finite collection of random variables $Z_1, \dots, Z_m$

$$\mathrm{E}\big[\max_{i=1,\dots,m}|Z_i|\,\big|\,\mathcal{D}'\big] \leq \log(1+m)\max_{i=1,\dots,m}\|Z_i\|_{\psi_1}.$$

Now, let Z_ℓ be given by

$$Z_\ell := \frac{1}{nh\hat{\sigma}_\ell^2 a_\ell^2}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)\{\dot{X}_i^\intercal(\hat{\gamma}_{t_\ell} - \gamma_{t_\ell})\}^2\varepsilon_i^2.$$

We show in Lemma 83

$$\|Z_\ell\|_{\psi_1} \leq K_\varepsilon^2\frac{1}{nh\hat{\sigma}_\ell^2 a_\ell^2}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)\{\dot{X}_i^\intercal(\hat{\gamma}_{t_\ell} - \gamma_{t_\ell})\}^2$$

and thus

$$\begin{aligned}
&\mathrm{E}\left[\max_{\ell=1,\dots,L}\frac{1}{nh\hat{\sigma}_\ell^2 a_\ell^2}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)\{\dot{X}_i^\intercal(\hat{\gamma}_{t_\ell} - \gamma_{t_\ell})\}^2\varepsilon_i^2\,\middle|\,\mathcal{D}'\right]\\
\leq& K_\varepsilon^2\log(1+L)\max_{\ell=1,\dots,L}\frac{1}{nh\hat{\sigma}_\ell^2 a_\ell^2}\sum_{i=1}^n \mathbb{1}(T_i \in \mathcal{I}_\ell)\{\dot{X}_i^\intercal(\hat{\gamma}_{t_\ell} - \gamma_{t_\ell})\}^2\\
=& O_p\big(\log(L)\mu\max_\ell\|\gamma_{t_\ell}\|_1\big) = o_p(\rho_n^3)
\end{aligned}$$

where we have used Lemma 13 and Assumption 18. Hence, also this term converges sufficiently fast to zero and the claim follows. □

Lemma 83. *Under Assumption 16, it holds*

$$\begin{aligned}&\left\|\frac{1}{nh\hat{\sigma}_\ell^2}\sum_{i=1}^n \mathbb{1}(T_i\in\mathcal{I}_\ell)\{\dot{X}_i^\intercal(\hat{\gamma}_{t_\ell}-\gamma_{t_\ell})\}^2\varepsilon_i^2\right\|_{\psi_1}\\ \leq& K_\varepsilon^2\frac{1}{nh\hat{\sigma}_\ell^2 a_\ell^2}\sum_{i=1}^n \mathbb{1}(T_i\in\mathcal{I}_\ell)\{\dot{X}_i^\intercal(\hat{\gamma}_{t_\ell}-\gamma_{t_\ell})\}^2.\end{aligned}$$

Proof. The argument relies on the following form of Young's inequality: if $p_i\in[0,1]$ with $\sum_i p_i=1$, then

$$\prod_{i=1}^n a_i^{p_i}\leq\sum_{i=1}^n p_i a_i.$$

Take $a_i=\exp(\varepsilon_i^2/K_\varepsilon^2)$ and

$$p_i=\frac{\mathbb{1}(T_i\in\mathcal{I}_\ell)\{\dot{X}_i^\intercal(\hat{\gamma}_{t_\ell}-\gamma_{t_\ell})\}^2}{\sum_{i=1}^n\mathbb{1}(T_i\in\mathcal{I}_\ell)\{\dot{X}_i^\intercal(\hat{\gamma}_{t_\ell}-\gamma_{t_\ell})\}^2}.$$

These satisfy $p_i\in[0,1]$ as well as $\sum_i p_i=1$ and so Young's inequality is applicable. Therefore,

$$\begin{aligned}&\mathrm{E}\left[\exp\left(\frac{\sum_{i=1}^n\mathbb{1}(T_i\in\mathcal{I}_\ell)\{\dot{X}_i^\intercal(\hat{\gamma}_{t_\ell}-\gamma_{t_\ell})\}^2\varepsilon_i^2}{\sum_{i=1}^n\mathbb{1}(T_i\in\mathcal{I}_\ell)\{\dot{X}_i^\intercal(\hat{\gamma}_{t_\ell}-\gamma_{t_\ell})\}^2K_\varepsilon^2}\right)\,\middle|\,\mathcal{D}'\right]\\ =&\,\mathrm{E}\left[\prod_{i=1}^n a_i^{p_i}\,\middle|\,\mathcal{D}'\right]\leq\sum_{i=1}^n p_i\,\mathrm{E}\left[\exp\left(\frac{\varepsilon_i^2}{K_\varepsilon^2}\right)\,\middle|\,\mathcal{D}'\right]\leq 2,\end{aligned}$$

where we have used in the last inequality that the ε_i^2 satisfy $\max_i\|\varepsilon_i^2\|_{\psi_1}\leq K_\varepsilon^2$ by Assumption 16. □

Bibliography

D. W. Andrews and P. Guggenberger. Hybrid and size-corrected subsampling methods. *Econometrica*, 77(3):721–762, 2009a.

D. W. Andrews and P. Guggenberger. Validity of subsampling and "plug-in asymptotic" inference for parameters defined by moment inequalities. *Econometric Theory*, 25(3):669–709, 2009b.

D. W. Andrews and P. Guggenberger. Asymptotic size and a problem with subsampling and with the m out of n bootstrap. *Econometric Theory*, 26(2):426–468, 2010.

D. W. Andrews and X. Shi. Inference based on conditional moment inequalities. *Econometrica*, 81(2):609–666, 2013.

D. W. Andrews and X. Shi. Nonparametric inference based on conditional moment inequalities. *Journal of Econometrics*, 179(1):31–45, 2014.

Y. Bai, A. Santos, and A. Shaikh. A practical method for testing many moment inequalities. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2019-116), 2019.

P. Bajari, C. L. Benkard, and J. Levin. Estimating dynamic models of imperfect competition. *Econometrica*, 75(5):1331–1370, 2007.

K. Ball. The reverse isoperimetric problem for gaussian measure. *Discrete & Computational Geometry*, 10(4):411–420, 1993.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81 (2):608–650, 2014.

A. Belloni, V. Chernozhukov, D. Chetverikov, and K. Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366, 2015.

A. Belloni, V. Chernozhukov, I. Fernandez-Val, and C. Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

A. Belloni, F. A. Bugni, and V. Chernozhukov. Subvector inference in pi models with many moment inequalities. Technical report, cemmap working paper, 2019a.

A. Belloni, V. Chernozhukov, D. Chetverikov, and I. Fernández-Val. Conditional quantile processes based on series or many regressors. *Journal of Econometrics*, 213(1):4–29, 2019b.

V. Bentkus. On the dependence of the berry–esseen bound on dimension. *Journal of Statistical Planning and Inference*, 113(2):385–402, 2003. ISSN 0378-3758. doi: https://doi.org/10.1016/S0378-3758(02)00094-0. URL `https://www.sciencedirect.com/science/article/pii/S0378375802000940`.

V. Bentkus. A lyapunov-type bound in rd. *Theory of Probability & Its Applications*, 49(2):311–323, 2005.

E. Beutner and H. Zähle. A modified functional delta method and its application to the estimation of risk functionals. *Journal of Multivariate Analysis*, 101(10): 2452–2463, 2010.

G. Biau and D. M. Mason. High-dimensional p p-norms. *Mathematical Statistics and Limit Theorems: Festschrift in Honour of Paul Deheuvels*, pages 21–40, 2015.

P. Billingsley. *Convergence of Probability Measures*. Wiley Series in Probability and Statistics. Wiley, 1999. ISBN 9780471197454. URL `https://books.google.de/books?id=QY06uAAACAAJ`.

S. G. Bobkov and G. P. Chistyakov. On concentration functions of random variables. *Journal of Theoretical Probability*, 28(3):976–988, 2015.

H. Bong, A. K. Kuchibhotla, and A. Rinaldo. Dual induction clt for high-dimensional m-dependent data. *arXiv preprint arXiv:2306.14299*, 2023.

P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

I. Canay and A. Shaikh. Practical and theoretical advances for inferences in partially identified models, n in advances in economics and econometric: Eleventh world congress, vol. 2, 271o306, ed. by b. honoré, a. pakes, m. piazzesi, and l samuelson, 2016.

M. D. Cattaneo, M. Jansson, K. Nagasawa, et al. Bootstrap-based inference for cube root consistent estimators. *arXiv preprint arXiv:1704.08066*, 2017.

M. D. Cattaneo, R. P. Masini, and W. G. Underwood. Yurinskii's coupling for martingales, 2022. URL `https://arxiv.org/abs/2210.00362`.

M. D. Cattaneo, R. Chandak, M. Jansson, and X. Ma. Boundary adaptive local polynomial conditional density estimators. *Bernoulli*, 30(4):3193–3223, 2024.

H. Chen. Convergence rates for parametric components in a partly linear model. *The Annals of Statistics*, pages 136–146, 1988.

Q. Chen and Z. Fang. Inference on functionals under first order degeneracy. *Journal of Econometrics*, 210(2):459–481, 2019.

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

X. Chen and T. M. Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, 9 (1):39–84, 2018.

X. Chen and Z. Liao. Sieve semiparametric two-step gmm under weak dependence. *Journal of Econometrics*, 189(1):163–186, 2015.

X. Chen and D. Pouzo. Sieve wald and qlr inferences on semi/nonparametric conditional moment models. *Econometrica*, 83(3):1013–1079, 2015.

X. Chen and X. Shen. Sieve extremum estimates for weakly dependent data. *Econometrica*, pages 289–314, 1998.

X. Chen, O. Linton, and I. Van Keilegom. Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71(5):1591–1608, 2003.

G. Cheng and Y.-C. Chen. Nonparametric inference via bootstrapping the debiased estimator. *Electronic Journal of Statistics*, 13(1):2194 – 2256, 2019. doi: 10.1214/19-EJS1575. URL `https://doi.org/10.1214/19-EJS1575`.

V. Chernozhukov, I. Fernández-Val, and A. Galichon. Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125, 2010.

V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786 – 2819, 2013a. doi: 10.1214/13-AOS1161. URL `https://doi.org/10.1214/13-AOS1161`.

V. Chernozhukov, S. Lee, and A. M. Rosen. Intersection bounds: Estimation and inference. *Econometrica*, 81(2):667–737, 2013b.

V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564–1597, 2014a.

V. Chernozhukov, D. Chetverikov, and K. Kato. Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 42(5):1787 – 1818, 2014b. doi: 10.1214/14-AOS1235. URL `https://doi.org/10.1214/14-AOS1235`.

V. Chernozhukov, D. Chetverikov, and K. Kato. Comparison and anti-concentration bounds for maxima of gaussian random vectors. *Probability Theory and Related Fields*, 162:47–70, 2015a.

V. Chernozhukov, C. Hansen, and M. Spindler. Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review*, 105(5):486–490, 2015b.

V. Chernozhukov, D. Chetverikov, and K. Kato. Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related gaussian couplings. *Stochastic Processes and their Applications*, 126(12):3632–3651, 2016.

V. Chernozhukov, D. Chetverikov, and K. Kato. Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4):2309–2352, 2017.

V. Chernozhukov, D. Chetverikov, and K. Kato. Inference on causal and structural parameters using many moment inequalities. *The Review of Economic Studies*, 86(5):1867–1900, 2019.

V. Chernozhuokov, D. Chetverikov, K. Kato, and Y. Koike. Improved central limit theorem and bootstrap approximations in high dimensions. *The Annals of Statistics*, 50(5):2562–2586, 2022.

A. Chesher, A. M. Rosen, and K. Smolinski. An instrumental variable model of multiple discrete choice. *Quantitative Economics*, 4(2):157–196, 2013.

D. Chetverikov, A. Santos, and A. M. Shaikh. The econometrics of shape restrictions. *Annual Review of Economics*, 10(1):31–63, 2018.

F. Ciliberto and E. Tamer. Market structure and multiple equilibria in airline markets. *Econometrica*, 77(6):1791–1828, 2009.

H. Deng and C.-H. Zhang. Beyond Gaussian approximation: Bootstrap for maxima of sums of independent random vectors. *The Annals of Statistics*, 48(6): 3643 – 3671, 2020. doi: 10.1214/20-AOS1946. URL `https://doi.org/10.1214/20-AOS1946`.

R. A. DeVore and G. G. Lorentz. *Constructive approximation*, volume 303. Springer Science & Business Media, 1993.

R. M. Dudley. Frechet Differentiability, p-Variation and Uniform Donsker Classes. *The Annals of Probability*, 20(4):1968 – 1982, 1992. doi: 10.1214/aop/1176989537. URL `https://doi.org/10.1214/aop/1176989537`.

R. M. Dudley. The Order of the Remainder in Derivatives of Composition and Inverse Operators for p-Variation Norms. *The Annals of Statistics*, 22(1):1 – 20, 1994.

R. M. Dudley and R. Norvaiša. *Differentiability of six operators on nonsmooth functions and p-variation*. Springer, 1999.

R. M. Dudley and W. Philipp. Invariance principles for sums of banach space valued random elements and empirical processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 62(4):509–552, 1983.

L. Dümbgen. On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields*, 95(1):125–140, 1993.

R. F. Engle, C. W. Granger, J. Rice, and A. Weiss. Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American statistical Association*, 81(394):310–320, 1986.

J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*, volume 66. CRC Press, 1996.

J. Fan, J. Lv, and L. Qi. Sparse high-dimensional models in economics. *Annu. Rev. Econ.*, 3(1):291–317, 2011.

X. Fang and Y. Koike. High-dimensional central limit theorems by Stein's method. *The Annals of Applied Probability*, 31(4):1660 – 1686, 2021. doi: 10.1214/20-AAP1629. URL `https://doi.org/10.1214/20-AAP1629`.

X. Fang and Y. Koike. Large-dimensional central limit theorem with fourth-moment error bounds on convex sets and balls. *The Annals of Applied Probability*, 34(2): 2065 – 2106, 2024. doi: 10.1214/23-AAP2014. URL `https://doi.org/10.1214/23-AAP2014`.

Z. Fang and A. Santos. Inference on directionally differentiable functions. *The Review of Economic Studies*, 86(1):377–412, 09 2018. ISSN 0034-6527.

S. Firpo, A. F. Galvao, and T. Parker. Uniform inference for value functions. *arXiv preprint arXiv:1911.10215*, 2019.

J. Freyberger and B. Larsen. How well does bargaining work in consumer markets? a robust bounds approach. Technical report, National Bureau of Economic Research, 2021.

T. Gasser and M. Rosenblatt. *Smoothing techniques for curve estimation*. Springer, 1979.

T. Gasser, H.-G. Muller, and V. Mammitzsch. Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47 (2):238–252, 1985. ISSN 00359246.

S. Ghosal, A. Sen, and A. W. Van Der Vaart. Testing monotonicity of regression. *Annals of statistics*, pages 1054–1082, 2000.

A. Giessing. Anti-concentration of suprema of gaussian processes and gaussian order statistics. *arXiv preprint arXiv:2310.12119*, 2023a.

A. Giessing. Gaussian and bootstrap approximations for suprema of empirical processes. *arXiv preprint arXiv:2309.01307*, 2023b.

A. Giessing and J. Fan. Bootstrapping ℓ_p-statistics in high dimensions. *arXiv preprint arXiv:2006.13099*, 2020.

A. Giessing and J. Fan. A bootstrap hypothesis test for high-dimensional mean vectors. *arXiv preprint arXiv:2309.01254*, 2023.

F. Götze, A. Naumov, V. Spokoiny, and V. Ulyanov. Large ball probabilities, Gaussian comparison and anti-concentration. *Bernoulli*, 25(4A):2538 – 2563, 2019. doi: 10.3150/18-BEJ1062. URL `https://doi.org/10.3150/18-BEJ1062`.

K. Gregory, E. Mammen, and M. Wahl. Statistical inference in sparse high-dimensional additive models. *The Annals of Statistics*, 49(3):1514 – 1536, 2021.

Z. Guo, W. Yuan, and C.-H. Zhang. Decorrelated local linear estimator: Inference for non-linear effects in high-dimensional additive models, 2019.

W. Härdle, H. Liang, and J. Gao. *Partially linear models*. Springer Science & Business Media, 2000.

K. Hirano and J. R. Porter. Impossibility results for nondifferentiable functionals. *Econometrica*, 80(4):1769–1790, 2012.

H. Hong and J. Li. The numerical delta method. *Journal of Econometrics*, 206(2): 379–394, 2018.

H. Hong and J. Li. The numerical bootstrap. *The Annals of Statistics*, 48(1): 397–412, 2020.

Y.-W. Hsieh, X. Shi, and M. Shum. Inference on estimators defined by mathematical programming. *Journal of Econometrics*, 226(2):248–268, 2022.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1): 2869–2909, 2014.

Ò. Jordà. Estimation and inference of impulse responses by local projections. *American economic review*, 95(1):161–182, 2005.

M. Kasy. Uniformity and the delta method. *Journal of Econometric Methods*, 2018 2018.

B. Kelly, D. Papanikolaou, A. Seru, and M. Taddy. Measuring technological innovation over the long run. *American Economic Review: Insights*, 3(3):303–320, 2021.

A. R. Klivans, R. O'Donnell, and R. A. Servedio. Learning geometric concepts via gaussian surface area. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 541–550. IEEE, 2008.

V. Koltchinskii. Estimation of smooth functionals in high-dimensional models: Bootstrap chains and Gaussian approximation. *The Annals of Statistics*, 50(4): 2386 – 2415, 2022. doi: 10.1214/22-AOS2197. URL `https://doi.org/10.1214/22-AOS2197`.

V. I. Koltchinskii. Komlós-major-tusnády approximation for the general empirical process and haar expansions of classes of functions. *Journal of Theoretical Probability*, 7:73–118, 1994.

E. Kong, O. Linton, and Y. Xia. Uniform bahadur representation for local polynomial estimates of m-regression and its application to the additive model. *Econometric Theory*, 26(5):1529–1564, 2010.

D. Kozbur. Inference in additively separable models with a high-dimensional set of conditioning variables. *Forthcoming in Journal of Business & Economic Statistics*, 2020.

D. Kozbur. Dimension-free anticoncentration bounds for gaussian order statistics with discussion of applications to multiple testing. *arXiv preprint arXiv:2107.10766*, 2021.

I. Krinsky and A. L. Robb. On approximating the statistical properties of elasticities. *The review of economics and statistics*, pages 715–719, 1986.

I. Krinsky and A. L. Robb. On approximating the statistical properties of elasticities: A correction. *Review of Economics & Statistics*, 72(1):189–190, 1990.

A. K. Kuchibhotla and A. Rinaldo. High-dimensional clt for sums of non-degenerate random vectors: $n^{\{} - 1/2\}$-rate. *arXiv preprint arXiv:2009.13673*, 2020.

C. Kwon and E. Mbakop. Estimation of the number of components of nonparametric multivariate finite mixture models. *The Annals of Statistics*, 49(4):2178–2205, 2021.

L. Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 1991.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

S. Lee, K. Song, and Y.-J. Whang. Testing functional inequalities. *Journal of Econometrics*, 172(1):14–32, 2013.

M. E. Lopes. Central limit theorem and bootstrap approximation in high dimensions: Near 1/n rates via implicit smoothing. *The Annals of Statistics*, 50(5):2492–2513, 2022.

M. E. Lopes, Z. Lin, and H.-G. Müller. Bootstrapping max statistics in high dimensions: Near-parametric rates under weak variance decay and application to functional and multinomial data. *The Annals of Statistics*, 48(2):1214 – 1229, 2020. doi: 10.1214/19-AOS1844. URL `https://doi.org/10.1214/19-AOS1844`.

J. Lu, M. Kolar, and H. Liu. Kernel meets sieve: Post-regularization confidence bands for sparse additive model. *Journal of the American Statistical Association*, 115:2084–2099, 2020.

A. Lytova and K. Tikhomirov. The variance of the ℓ_- {p} -norm of the gaussian vector, and dvoretzky's theorem. *St. Petersburg Mathematical Journal*, 30(4): 699–722, 2019.

C. Ma and J. Huang. Asymptotic properties of lasso in high-dimensional partially linear models. *Science China Mathematics*, 59(4):769–788, 2016.

C. F. Manski. Monotone treatment response. *Econometrica: Journal of the Econometric Society*, pages 1311–1334, 1997.

C. F. Manski and J. V. Pepper. Monotone instrumental variables: With an application to the returns to schooling. *Econometrica*, 68(4):997–1010, 2000. ISSN 00129682, 14680262.

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

P. Massart. Strong approximation for multivariate empirical and related processes, via kmt constructions. *The Annals of probability*, pages 266–291, 1989.

M. A. Masten, A. Poirier, and L. Zhang. Assessing sensitivity to unconfoundedness: Estimation and inference. *arXiv preprint arXiv:2012.15716*, 2020.

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

N. Meinshausen, L. Meier, and P. Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

R. Mitra and C.-H. Zhang. The benefit of group sparsity in group inference with de-biased scaled group Lasso. *Electronic Journal of Statistics*, 10(2):1829 – 1873, 2016.

F. Molinari. Microeconometrics with partial identification. *Handbook of econometrics*, 7:355–486, 2020.

H.-G. Müller. Kernel estimators of zeros and of location and size of extrema of regression functions. *Scandinavian journal of statistics*, pages 221–232, 1985.

P. Müller and S. van de Geer. The partial linear model in high dimensions. *Scandinavian Journal of Statistics*, 42(2):580–608, 2015. ISSN 03036898, 14679469.

F. Nazarov. On the maximal perimeter of a convex set in r^n with respect to a gaussian measure. In *Geometric Aspects of Functional Analysis: Israel Seminar 2001-2002*, pages 169–187. Springer, 2003.

W. K. Newey. Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, 10(2):1–21, 1994a.

W. K. Newey. The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382, 1994b.

W. K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of econometrics*, 79(1):147–168, 1997.

G. Peccati and N. Turchi. The discrepancy between min–max statistics of gaussian and gaussian-subordinated matrices. *Stochastic Processes and their Applications*, 158:315–341, 2023.

P. C. Phillips. Folklore theorems, implicit maps, and indirect inference. *Econometrica*, 80(1):425–454, 2012.

D. Pollard. *A User's Guide to Measure Theoretic Probability*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2001. doi: 10.1017/CBO9780511811555.

M. Raič. A multivariate Berry–Esseen theorem with explicit constants. *Bernoulli*, 25(4A):2824 – 2853, 2019. doi: 10.3150/18-BEJ1072. URL `https://doi.org/10.3150/18-BEJ1072`.

J. Reeds. *On the Definition of Von Mises Functionals*. Harvard University, 1976.

E. Rio. Local invariance principles and their application to density estimation. *Probability Theory and Related Fields*, 98(1):21–45, 1994.

P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4): 931–954, 1988. ISSN 00129682, 14680262.

J. P. Romano and A. M. Shaikh. Inference for identifiable parameters in partially identified econometric models. *Journal of Statistical Planning and Inference*, 138 (9):2786–2807, 2008.

J. P. Romano and A. M. Shaikh. Inference for the identified set in partially identified econometric models. *Econometrica*, 78(1):169–211, 2010.

M. Rudelson and R. Vershynin. The littlewood–offord problem and invertibility of random matrices. *Advances in Mathematics*, 218(2):600–633, 2008.

M. Rudelson and R. Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 62(12):1707–1739, 2009.

M. Rudelson and R. Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:9 pp., 2013. doi: 10.1214/ECP.v18-2865.

D. E. Runkle. Vector autoregressions and reality. *Journal of Business & Economic Statistics*, 5(4):437–442, 1987.

J. Scherer. An anti-concentration bound for sublinear and continuous functionals of gaussian random vectors. unpublished, 2024.

V. Semenova. Adaptive estimation of intersection bounds: a classification approach. *arXiv preprint arXiv:2303.00982*, 2023.

R. D. Shah and R. J. Samworth. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):55–80, 2013.

A. Shapiro. On concepts of directional differentiability. *Journal of optimization theory and applications*, 66(3):477–487, 1990.

A. Shapiro. Asymptotic analysis of stochastic programs. *Annals of Operations Research*, 30(1):169–186, 1991.

X. Shen. On methods of sieves and penalization. *The Annals of Statistics*, 25(6): 2555–2591, 1997.

R. Singh and S. Vijaykumar. Kernel ridge regression inference. *arXiv preprint arXiv:2302.06578*, 2023.

M. Sommerfeld and A. Munk. Inference for empirical wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):219–238, 2018.

J. Spiess, A. Venugopal, et al. Double and single descent in causal inference with an application to high-dimensional synthetic control. *Advances in Neural Information Processing Systems*, 36:63642–63659, 2023.

B. Stucky and S. van de Geer. Asymptotic confidence regions for high-dimensional structured sparsity. *IEEE Transactions on Signal Processing*, 66(8):2178–2190, 2018.

S. van de Geer and B. Stucky. χ^2-confidence sets in high-dimensional regression. In *Statistical analysis for high-dimensional data*, pages 279–306. Springer, 2016.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166 – 1202, 2014. doi: 10.1214/14-AOS1221.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996. ISBN 9780387946405.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer New York, 2000. ISBN 9781475725452.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

L. Wasserman and K. Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

T. Woutersen and J. Ham. Confidence sets for continuous and discontinuous functions of parameters. Technical report, Working paper.[331], 2014.

S. Yamamuro. *Differential calculus in topological linear spaces*, volume 374. Springer, 2006.

Z. Yu, M. Levine, and G. Cheng. Minimax optimal estimation in partially linear additive models under high dimension. *Bernoulli*, 25(2):1289 – 1325, 2019.

A. Y. Zaitsev. The accuracy of strong gaussian approximation for sums of independent random vectors. *Russian Mathematical Surveys*, 68(4):721, 2013.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 76(1):217–242, 2014. ISSN 13697412, 14679868.

Y. Zhu. Nonasymptotic analysis of semiparametric regression models with high-dimensional parametric coefficients. *The Annals of Statistics*, 45(5):2274 – 2298, 2017.

Y. Zhu, Z. Yu, and G. Cheng. High dimensional inference in partially linear models. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2760–2769. PMLR, 16–18 Apr 2019.

K. Ziegler. On nonparametric kernel estimation of the mode of the regression function in the random design model. *Journal of Nonparametric Statistics*, 14(6): 749–774, 2002.