universität wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation / Title of the Doctoral Thesis

„Assumption-lean conditional predictive inference via the Jackknife“

verfasst von / submitted by

Dipl.-Ing. Nicolai David Amann, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien, 2023 / Vienna, 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 794 370 136
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Statistik und Operations Research
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Hannes Leeb Ass.-Prof. Mag. Lukas Steinberger Bakk. BA, PhD

Dedicated to Walchsee
and all the wonderful people
who spent their time with me there

Acknowledgements

Achieving your doctoral degree is, like finishing a marathon, the result of steady progress and continuous work. But sometimes even the PhD studies may feel like training for a marathon: Once you feel comfortable with the current speed or distance, you accelerate to reach the next level. When approaching the finish line it is important to look back in gratitude for all the people who have brought you there, helped you on your path or just brightened up your way through their companionship. So here we are. And I would like to start with thanking Mathias Pohl, whose thesis reminded me that there is no need to write more than a few lines of acknowledgements *unless* one really feels grateful (he wrote three pages of acknowledgements).

It has been for a while, but I want to express my appreciation for my former math teachers in high school, Martin Hölbling and Walter Kastanek. My gratitude is not restricted to their mathematical foundation they gave me, but rather *the way* they taught me. In particular, I would like to thank Martin Hölbling for letting us discuss for one hour how $(a + b)^2$ can be expressed alternatively, resulting in a far better understanding than any other lesson could have done together with the conviction that the third [sic!] binomial formula is highly underrated. I also want to thank Walter Kastanek, who, once he found out that I programmed a football manager game on the calculator, gave me an even better calculator instead of prohibiting me from sharing the game with my class mates.

On my academical way to the doctoral degree, three persons played a crucial role: Firstly, I want to thank Ulrike Schneider, who I challenged several times with the clutter of my master thesis, for supporting me nonetheless. I hope in this thesis I can explain better why mathematical correct statements indeed hold true from an intuitive point of view. Most importantly, I want to express my gratitude to my supervisors Hannes Leeb and Lukas Steinberger for their guidance and the patience they had with me. Without them, this thesis would not have been possible.

I also want to thank all attendants of the weekly seminar of our research group and highlight Benedikt Pötscher for his helpful comments on my research. A warm thanks goes to my colleagues at the department: Gianluca Finoccio for sharing the same passion in discussing fundamental questions in statistics even if they go far beyond our original research fields. Thomas Stark, who has by far become more a friend than an office partner to me, for going through all the ups and downs with me during our PhD studies. Together with Georg Köstenberger he completes our small conference group informally called *Speedboater*. Besides all the fruitful discussions on statistical topics I had with them, I also want to highlight our discovery that trains, apart from their obvious ecological benefits, may not always be the most restful means of transport. Although not being at our department any more, I would like to name some former colleagues, who brightened

up my days at the department. Foremost, I enjoyed the conversations with Kory Johnson and the time we spent together at the conference in Salzburg a lot. I also want to mention Corina Birghila and Daniela Escobar, who welcomed me warmly during my first year as PhD. Conversely, special thanks go to Nina Dörnemann for motivating me during the last year of my thesis to finish it.

In a different way the non-academic staff at our department have eased my life through my PhD studies: I am grateful to Svetlana Mihajlovic and Birgit Ewald for helping me with thousand little things, Manuela Nicham-Zorn for all her stories during lunch at our favourite restaurant *Plain Vienna* and last, but not least Lisa Carli for her steady support and writing me more mails than I could have ever asked for.

I am especially grateful for my dear friends, who accompanied me through the last years without naming them all. I would like to highlight our joint holidays in Walchsee and all the memories associated with it. I cannot imagine a better place to reload my batteries together with my friends and enjoy the famous *Vinschgerl* of the local bakery. Nevertheless, I want to point out Philipp Salat, who has always been there for me whenever I needed him most, Kenneth Duringer for being the first person visiting Walchsee, Dağcan Mermi for inspiring me in so many ways and Alena Chalupka for thousand things. Being one of the least important yet representative things among all, I want to thank her for taking me on a 12-hours hiking trip to the *Pyramidenspitze*, a challenge I dreamed of mastering for more than a decade, but never actually tried before.

Last, but certainly not least I want to thank my beloved family for making all this possible: My dear grandfather Jochen, to whom I look upon for his *joie de vivre* and the kind way he treated other people, and my grandmother Anneli, who lovingly cared for him. I am grateful beyond words for my father Lars, my sister Mattea and my brother Tobias for their support, not solely during the past years but throughout my whole life. And I am thankful for the precious time I was given to share with my mother Vroni.

Contents

1. Introduction

In practice, a statistician is often given training data T_n containing real-valued observations of a variable of interest y in order to predict a future outcome y_0. For a given model – whether misspecified or not – and a predictor associated with it one is often interested in the distribution of the prediction error for a broad range of applications. If the decision on the predictor is already fixed, then an accurate estimation of the prediction error's distribution allows to create prediction intervals, whose actual coverage probability is close to the nominal level. Another reason for the interest in the prediction error's distribution would be the evaluation of the quality of the model together with the predictor in terms of its mean-squared prediction error or mean-absolute deviation in order to compare different models or predictors against each other.

From the perspective of application, the statistician often has only one set of training data to which the methods shall be applied. Hence, she would be interested in the statistical properties conditional on her specific training set rather than the expected behavior averaged over all possible values, which is the reason why this thesis deals with methods of statistical inference conditional on the training data.

The present work extends the results of Steinberger and Leeb (2023), who showed that a Jackknife-approach may provide asymptotically valid prediction intervals under some assumptions: One of them directly addresses the fact that the Jackknife-approach intrinsically relies on the stability of the underlying prediction algorithm in the sense that the exclusion of one observation should not change the resulting prediction too much (on average). In that paper, it was shown that under additional assumptions many predictors fulfill this stability condition if the underlying model is linear. Furthermore, Steinberger and Leeb (2023) assumed that the conditional distribution of the response variable given its regressor is absolutely continuous. The main contributions of this thesis are tripartite: Firstly, we generalize the results of Steinberger and Leeb (2023) to the non-continuous case by removing the assumption on the conditional distribution of the response variable. For example, our results may be applied to classification algorithms and show that the Jackknife-approach may yield an asymptotically valid estimation of the misclassification error. Secondly, we show that the stability condition derived in the present work (which is closely related to their stability condition) is also fulfilled for a selection of predictors used in practice *without* assuming any connection between the response variable and the regressor. In particular, the data-generating process need not be a linear model or correctly specified in any sense. Thirdly, our results are not restricted to the creation of prediction intervals, but are also applicable to the consistent estimation of a whole class of functions of the prediction error including the mean-squared prediction error.

Both this thesis and the results of Steinberger and Leeb (2023) can be used in a setting, where the number of regressors grows linearly with the number of observations. This

setting has seen an increasing interest in statistical research, as in the last decade the acquisition of huge data sets has been supported by the rapid development of mass storage and computing power. While in the classical setting, where the number of observations tends to infinity for a fixed number of regressors, the empirical cumulative distribution function (ecdf) of the residuals might give an asymptotically accurate estimation of the distribution of the prediction error, this approach fails in the setting, where the number of parameters grows proportionally with the number of observations (cf. Mammen 1996). Furthermore, also the bootstrap will fail in that setting (see Bickel and Freedman 1983 and El Karoui and Purdom 2018).

An alternative approach of estimating the prediction error's distribution is based on the use of the leave-one-out residuals, also known as the Jackknife, which will be examined in the present work. While the original Jackknife dates back to the Fifties of the last century (cf. Quenouille 1956), only little is known when it is applied in high-dimensional statistics. One recent result is the paper Steinberger and Leeb (2023), which shows that the Jackknife-approach gives asymptotically valid prediction intervals *conditionally on the training data* if the distribution of the response y_0 conditional on the new regressor x_0 is absolutely continuous. The fact that the coverage probability of the prediction intervals is close to the desired level even if one conditions on the training data is in line with corresponding results in a large-sample framework with a fixed number of regressors (cf. Butler and Rothman 1980).

Recently, a remarkable modification, the so-called Jackknife+, has been established in Barber et al. (2021a) and was shown not to rely on the stability of the predictor. While prediction intervals with nominal coverage probability of $1 - \alpha$ created by the Jackknife+ guarantee an actual unconditional coverage probability of at least $1 - 2\alpha$ for symmetric predictors, Bian and Barber (2023) showed that the Jackknife+ (alongside with full conformal prediction) fails to give a similar guarantee for the conditional coverage probability if the distribution of the regressors is nonatomic.

We would also like to emphasize that we condition on the training data rather than conditioning on the new feature as the latter comes at a high price: Prediction intervals satisfying the so-called object conditional validity, that is, providing a valid coverage probability conditional on the new feature, possess an infinitely large expected length for nonatomic distributions (cf. Vovk 2012 and Lei and Wasserman 2014). Moreover, Barber et al. (2021b) showed that a relaxed version of object conditional validity can be reached by an adaptation of split conformal predictive intervals that essentially cannot be outperformed in terms of the interval length by another procedure. However, the split conformal prediction framework – although enjoying approximate conditional validity too – suffers from the fact that the predictor loses accuracy through only using a subset of the full data on which it is trained, while the interval length is determined on the holdout set (see Steinberger and Leeb 2023 for a discussion on the interval length). Lastly, Barber et al. (2021a) proposed a K-fold cross-validation procedure based on the Jackknife+, whose conditional coverage probability satisfies a PAC-type inequality, in the sense that its conditional coverage probability is larger than $1 - 2\alpha - \mathcal{O}(1)$ with high probability if the learning algorithm is symmetric and the (equally large) size of each fold is large compared to the number of folds (cf. Bian and Barber 2023). Notice that the Jackknife,

which we investigate here, corresponds to n-fold cross-validation, that is, there are n folds each of size 1.

The remaining of the thesis is organized as follows: Chapter 2 introduces the notation and the setting, while Chapter 3 shows that in the non-continuous case the approach of Steinberger and Leeb (2023) will fail and subsequently discusses different measurements for the accuracy of the prediction error's distribution and its usefulness regarding the coverage probability of prediction intervals. Chapter 4 presents a Jackknife-approach to estimate the prediction error's distribution and contains results both for the finite sample and the asymptotic setting, showing that the approach hinges on the stability of the predictor. Correspondingly, Chapter 5 discusses this property and shows that in a broad setting a range of different predictors fulfill the stability assumption and hence can be used for (asymptotically) valid prediction intervals. Chapter 6 contains a discussion on various topics including the necessity of our assumptions. All proofs of our results are deferred to the Appendix.

2. Setting

The data

Assume that we are given training data T_n of n i.i.d. data points $(y_i, x_i')_{i=1}^n$, where y_i is a real-valued random variable and x_i is a random vector in $\mathbb{R}^p$. We are interested in predicting another real-valued random variable y_0, where we are additionally given a new regressor x_0. We assume that (y_0, x_0') is independent of the training data and has the same distribution as (y_1, x_1'). Denoting our prediction based on the training data T_n and on x_0 with $\hat{y}_0$, our main goal will be the creation of prediction intervals for y_0 and establishing guarantees for its conditional coverage probability. In principle, our arguments can be also generalized to the case where we are given no regressors at all $(p = 0)$ and to the case of randomized prediction algorithms.

The algorithm

A learning algorithm is a procedure, which uses the training data T_n together with a new feature vector x_0 to predict the response variable associated with that feature vector. To formalize this, we fix $p \in \mathbb{N}, n \in \mathbb{N}$ and define $\mathcal{A}_{p,n}$ to be a Borel measurable function from $\mathbb{R}^{(p+1)n} \times \mathbb{R}^p$ to $\mathbb{R}$ and call $\hat{y}_0 := \mathcal{A}_{p,n}(T_n, x_0)$ the prediction for y_0 given x_0 and the training data T_n. If we want to allow for randomized predictors, we replace the domain of the learning algorithm by $\mathbb{R}^{(p+1)n} \times \mathbb{R}^p \times \Omega$, where Ω is the sample space of a probability space $(\Omega, \mathcal{F}, P)$. In the special case $p = 0$ we define the algorithm $\mathcal{A}_{0,n}$ to be a function only of the response variables $(y_i)_{i=1}^n$ or, in the randomized setting, a random function from $\mathbb{R}^n$ to $\mathbb{R}$. A learning algorithm $\mathcal{A}_{p,n}$ is called symmetric if the learning algorithm does not change by a permutation of the training data, i.e., for all $x_0 \in \mathbb{R}^p$, any permutation $\pi : \{1, \dots, n\} \to \{1, \dots, n\}$ and training data $T_n = (y_i, x_i')_{i=1}^n$ the learning algorithm satisfies $\mathcal{A}_{p,n}(T_n, x_0) = \mathcal{A}_{p,n}((y_{\pi(i)}, x'_{\pi(i)})_{i=1}^n, x_0)$. Being somewhat imprecise, we call a predictor $\hat{y}_0$ symmetric if the underlying learning algorithm $\mathcal{A}_{p,n}$ is symmetric for all $n \in \mathbb{N}$ and $p \in \mathbb{N}$.

Notation

For a measurable function $f : \mathbb{R} \to \mathbb{R}$ we will denote its supremum norm with $\|f\|_\infty := \sup_{x \in \mathbb{R}} |f(x)|$, its essential supremum (with respect to the Lebesgue-measure) with $\|f\|_{\mathcal{L}_\infty}$ and the $\mathcal{L}_p$ norm with $\|f\|_{\mathcal{L}_p} := (\int_{\mathbb{R}} |f(x)|^p d\lambda(x))^{1/p}$ for $p \in [1, \infty)$, where λ denotes the Lebesgue-measure. Furthermore, we extend this definition also to the case where $p \in (0, 1)$ although $\|f\|_{\mathcal{L}_p}$ is then no longer a norm. For a matrix $A \in \mathbb{R}^{a \times b}$ we write A' for its transpose, $\|A\|_F = \sqrt{\operatorname{tr} A'A}$ for its Frobenius norm and denote the $k \times k$ identity matrix with I_k. Furthermore, e_i denotes the i-th canonical basis vector, $\|\cdot\|_2$ the Euclidean

norm and $K_\varepsilon(t)$ the closed ball with respect to the Euclidean metric with radius ε and center t.

For a matrix $A \in \mathbb{R}^{a\times b}$ we will denote its Moore-Penrose pseudoinverse with $A^\dagger$ and the $a \times a$ dimensional orthogonal projection matrix onto the column space of A with $P_A := AA^\dagger$. The smallest and the largest eigenvalue of a symmetric matrix $A \in \mathbb{R}^{a\times a}$ will be denoted by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively. We will use σ instead of λ if we refer to the smallest and the largest singular value of a matrix $A \in \mathbb{R}^{a\times b}$. For a symmetric matrix $A \in \mathbb{R}^{a\times a}$ we write $A > 0$ if A is positive definite and $A \geq 0$ to denote a positive semidefinite matrix. Moreover, for a symmetric matrix $B \in \mathbb{R}^{a\times a}$ we write $A \geq B$ if $A - B$ is positive semidefinite and $A > B$ if $A - B$ is positive definite. If $A \in \mathbb{R}^{a\times a}$ is a positive semidefinite matrix, we denote the positive semidefinite square root of A with $A^{1/2}$.

For a set $S \subseteq \mathbb{R}^k$ we denote with $|S|$ its cardinality and $\mathbb{1}_S : \mathbb{R}^k \to \mathbb{R}$ the indicator function on the set S. For a real number x the expression $\lceil x \rceil$ describes the smallest integer larger than or equal to x and $\lfloor x \rfloor = -\lceil -x \rceil$. The expression $\lim_{x \searrow t}$ will denote the limit from above (in contrast to the usual limit $\lim_{x\to t}$). For a sequence of random variables X_n we will abbreviate convergence in probability to a random variable Y by $X_n \xrightarrow{p} Y$. Furthermore, we will write $X_n \sim \mathcal{O}_p(1)$ if the sequence $(X_n)_{n\in\mathbb{N}}$ is bounded in probability and $X_n \sim \mathcal{O}_p(n^\alpha)$ with $\alpha \in \mathbb{R}$ if the sequence $(X_n n^{-\alpha})_{n\in\mathbb{N}}$ is bounded in probability. Moreover, for a sequence $(v_n)_{n\in\mathbb{N}}$ we will write $X_n \sim o_p(v_n)$ if the sequence $(X_n v_n^{-1})_{n\in\mathbb{N}}$ converges to 0 in probability. For two distribution functions $F : \mathbb{R} \to \mathbb{R}$ and $G : \mathbb{R} \to \mathbb{R}$, the expression

$$L(F, G) := \inf\{\varepsilon \geq 0 : F(t - \varepsilon) - \varepsilon \leq G(t) \leq F(t + \varepsilon) + \varepsilon \text{ for all } t \in \mathbb{R}\}$$

will denote the Lévy metric between F and G. Moreover, we define $F^-(x) := \lim_{\delta \searrow 0} F(x - \delta)$ to be the limit from the left of F at the point $x \in \mathbb{R}$.

3. Accuracy of prediction intervals

In this chapter we will link the coverage probability of prediction intervals to the estimation of the prediction error's distribution. Although some intuitive approaches arise, it will not be immediately clear how we would like to measure the accuracy of the latter.

We start with the following definition: Let $\mathrm{F}_n(t) := \mathbb{P}\left(y_0 - \hat{y}_0 \leq t \| T_n\right)$ denote the cumulative distribution function of the prediction error conditional on the training data T_n and let $\widehat{\mathrm{F}}_n : \mathbb{R} \to [0,1]$ another distribution function based on the training data, which will be used to estimate F_n. Now assume we have found two points $\hat{q}_{\alpha_i}$ with $i \in \{1,2\}$, such that $\widehat{\mathrm{F}}_n(\hat{q}_{\alpha_i}) = \alpha_i$ for some $0 \leq \alpha_1 \leq \alpha_2 \leq 1$. We then can define a prediction interval for y_0 as $\hat{y}_0 + (\hat{q}_{\alpha_1}, \hat{q}_{\alpha_2}]$, whose coverage probability conditional on the training data satisfies

$$\left|\mathbb{P}\left(y_0 \in \hat{y}_0 + (\hat{q}_{\alpha_1}, \hat{q}_{\alpha_2}] \| T_n\right) - (\alpha_2 - \alpha_1)\right| \leq 2\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_\infty. \tag{3.1}$$

To see that equation (3.1) holds true we start with the following equality:

$$\begin{aligned}
&\mathbb{P}\left(y_0 \in \hat{y}_0 + (\hat{q}_{\alpha_1}, \hat{q}_{\alpha_2}] \| T_n\right) - (\alpha_2 - \alpha_1) \\
&= \mathrm{F}_n(\hat{q}_{\alpha_2}) - \mathrm{F}_n(\hat{q}_{\alpha_1}) - \left(\widehat{\mathrm{F}}_n(\hat{q}_{\alpha_2}) - \widehat{\mathrm{F}}_n(\hat{q}_{\alpha_1})\right).
\end{aligned}$$

Now, using the fact that $|\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(x)| \leq \|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_\infty$ for all $x \in \mathbb{R}$ together with the triangle inequality yields

$$\begin{aligned}
\left|\mathbb{P}\left(y_0 \in \hat{y}_0 + (\hat{q}_{\alpha_1}, \hat{q}_{\alpha_2}] \| T_n\right) - (\alpha_2 - \alpha_1)\right| &\leq \left|\mathrm{F}_n(\hat{q}_{\alpha_2}) - \widehat{\mathrm{F}}_n(\hat{q}_{\alpha_2})\right| + \left|\mathrm{F}_n(\hat{q}_{\alpha_1}) - \widehat{\mathrm{F}}_n(\hat{q}_{\alpha_1})\right| \\
&\leq 2\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_\infty,
\end{aligned}$$

which proves equation (3.1). Equation (3.1) was used in Steinberger and Leeb (2023), who restricted their analysis to continuous distribution functions F_n. Besides the fact that for non-continuous functions $\widehat{\mathrm{F}}_n$ it is not always possible to find points $\hat{q}_{\alpha_i}$ with $\widehat{\mathrm{F}}_n(\hat{q}_{\alpha_i}) = \alpha_i$, an even larger problem can occur if the original function F_n is not continuous: While the inequality above is correct, it might be useless in practice if the discontinuity points of F_n are close to the edges of the prediction interval: For simplicity, we consider the extreme case where $\mathrm{F}_n = \mathbb{1}_{[s,\infty)}$ is a Dirac distribution at point s and we are able to estimate s by some $\hat{s}$ with very high accuracy (but not perfectly) in the sense that $0 < |s - \hat{s}| \leq \varepsilon$ almost surely. Now, a reasonable approach would be to estimate F_n by the Dirac distribution $\widehat{\mathrm{F}}_n = \mathbb{1}_{[\hat{s},\infty)}$. However, the Kolmogorov distance between F_n and $\widehat{\mathrm{F}}_n$ equals 1 and the corresponding prediction interval $\{\hat{y}_0 + \hat{s}\}$ with nominal level 1 has an actual coverage

probability of 0.[1]

However, this problem can be solved if we allow our prediction intervals to be larger by the small amount of 2ε: For the prediction interval $\hat{y}_0 + [\hat{s} - \varepsilon, \hat{s} + \varepsilon]$ we have the guarantee that its conditional coverage probability equals 1, while the price was only an increase of length 2ε. Moreover, the length of $\hat{y}_0 + [\hat{s} - \varepsilon, \hat{s} + \varepsilon]$ is only larger by 2ε than the best (infeasible) interval $\{\hat{y}_0 + s\}$.

3.1. The epsilon-variational divergence

The solution above hinges on the knowledge of ε and the fact that by an enlargement of our original prediction interval by the length of 2ε the conditional coverage probability is guaranteed to equal (or exceed) our desired value. While in practice these two properties may not be fulfilled, this approach can be generalized to more realistic cases. However, for the ε-inflated intervals we cannot use equation (3.1) any more. To tackle this problem, we start with the following definition:

Definition 3.1. Let $F : \mathbb{R} \to [0, 1]$ and $G : \mathbb{R} \to [0, 1]$ be cumulative distribution functions and $\varepsilon \geq 0$. Then, we define the ε-variational divergence between F and G with diffusion parameter $\varepsilon \geq 0$ as

$$\ell_\varepsilon(F, G) := \sup_{t \in \mathbb{R}} \inf_{x,y \in K_{\varepsilon/2}(t)} |F(x) - G(y)|,$$

where $K_{\varepsilon/2}(t)$ denotes the closed ball with radius $\varepsilon/2$ and center t.

As indicated before, it may be useful to inflate the length of prediction intervals by an additional amount of ε. For this, let q_1, q_2 and $\hat{y}_0$ be real numbers and $\varepsilon \geq 0$. We then define two prediction intervals as follows:

$$PI^+(\hat{y}_0, q_1, q_2, \varepsilon) := \hat{y}_0 + [q_1 - \varepsilon, q_2 + \varepsilon] \tag{3.2}$$

whenever $q_2 \geq q_1 - 2\varepsilon$ and $PI^+(\hat{y}_0, q_1, q_2, \varepsilon) := \emptyset$ if $q_2 < q_1 - 2\varepsilon$. Analogously, we define

$$PI^-(\hat{y}_0, q_1, q_2, \varepsilon) := \hat{y}_0 + (q_1 + \varepsilon, q_2 - \varepsilon) \tag{3.3}$$

whenever $q_2 > q_1 + 2\varepsilon$ and $PI^-(\hat{y}_0, q_1, q_2, \varepsilon) := \emptyset$ else.

Equipped with these definitions, we are able to state the following result:

Lemma 3.2. *Let $F : \mathbb{R} \to [0, 1]$ and $G : \mathbb{R} \to [0, 1]$ be cumulative distribution functions and $\varepsilon \geq 0$. We then have*

$$F(t - \varepsilon) - \ell_\varepsilon(F, G) \leq G(t) \leq F(t + \varepsilon) + \ell_\varepsilon(F, G) \text{ for all } t \in \mathbb{R}. \tag{3.4}$$

[1] Note that here the distribution of the prediction error does not allow any non-randomized prediction interval to possess a coverage probability differing from 0 or 1. We would like to point out that this example is an extreme case to illustrate the problem when facing a non-continuous distribution function F_n.

Furthermore, let $\mathrm{F}_n(t) = \mathbb{P}\left(y_0 - \hat{y}_0 \leq t \| T_n\right)$ *denote the conditional distribution function of the prediction error* $y_0 - \hat{y}_0$ *given the training data* T_n *and* $\widehat{\mathrm{F}}_n : \mathbb{R} \to [0,1]$ *another distribution function. Then, for every* $q_2 \geq q_1 - 2\varepsilon$ *the following holds true almost surely:*

$$\begin{aligned} \widehat{\mathrm{F}}_n(q_2) - \widehat{\mathrm{F}}_n^-(q_1) - 2\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n) &\leq \mathbb{P}\left(y_0 \in PI^+(\hat{y}_0, q_1, q_2, \varepsilon) \| T_n\right) \\ &\leq \widehat{\mathrm{F}}_n(q_2 + 2\varepsilon) - \widehat{\mathrm{F}}_n^-(q_1 - 2\varepsilon) + 2\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n). \end{aligned} \tag{3.5}$$

Alternatively, the following inequality holds true for every $q_2 > q_1 + 2\varepsilon$ *almost surely:*

$$\begin{aligned} \widehat{\mathrm{F}}_n^-(q_2) - \widehat{\mathrm{F}}_n(q_1) + 2\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n) &\geq \mathbb{P}\left(y_0 \in PI^-(\hat{y}_0, q_1, q_2, \varepsilon) \| T_n\right) \\ &\geq \widehat{\mathrm{F}}_n^-(q_2 - 2\varepsilon) - \widehat{\mathrm{F}}_n(q_1 + 2\varepsilon) - 2\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n). \end{aligned} \tag{3.6}$$

Starting with a distribution function $\widehat{\mathrm{F}}_n$, which serves as an approximation for F_n, Lemma 3.2 provides an easy way to create prediction intervals for y_0 whose actual conditional coverage probability is bounded from below by $(\alpha_2 - \alpha_1) - 2\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n)$ almost surely. To do so, we only have to choose two values q_2 and q_1 such that $\widehat{\mathrm{F}}_n(q_2) \geq \alpha_2$ and $\widehat{\mathrm{F}}_n(x) \leq \alpha_1$ for all $x < q_1$, which is always possible for a distribution function $\widehat{\mathrm{F}}_n$ as long as $0 < \alpha_1 \leq \alpha_2 < 1$ holds true. The guarantee for the coverage probability comes at the price of increasing the length of the interval by an additional amount of 2ε compared to the intervals of Steinberger and Leeb (2023). As shown at the beginning of this section, this increase may be necessary if the true distribution F_n is non-continuous even if we are able to estimate the discontinuity points well.

We also would like to point out that the values $\widehat{\mathrm{F}}_n(q_i \pm 2\varepsilon)$ are easy to compute in practice if the function $\widehat{\mathrm{F}}_n$ is given explicitly. Thus, in order to calculate the bounds for the actual (conditional) coverage probability we are left with the task of estimating the value of $\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n)$, which will typically be unknown. Hence, in order to use the coverage probabilities of Lemma 3.2, one would be interested in controlling $\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n)$, which will be the subject of the present work.

3.2. Applications beyond prediction intervals

Let F_X and F_Y be the distribution functions of the random variables X and Y, respectively. Then the usability of equation (3.4) based on $\ell_\varepsilon(F_X, F_Y)$ is not restricted to prediction intervals. Rather, we can use it to approximate the expected value of $f(X)$ by $f(Y)$ well for certain functions f:

Lemma 3.3. *Let* $f : \mathbb{R} \to [M_1, M_2]$ *be a non-decreasing function and let* X *and* Y *be real random variables. We then have for every* $\varepsilon \geq 0$

$$\mathbb{E}(f(X)) \geq \mathbb{E}(f(Y - \varepsilon)) - (M_2 - M_1)\ell_\varepsilon(F_X, F_Y) \text{ as well as} \tag{3.7}$$

$$\mathbb{E}(f(X)) \leq \mathbb{E}(f(Y + \varepsilon)) + (M_2 - M_1)\ell_\varepsilon(F_X, F_Y), \tag{3.8}$$

where F_X *and* F_Y *denote the distribution functions of* X *and* Y*, respectively. If, addi-*

tionally, f is Lipschitz continuous with constant L this yields the inequality

$$|\,\mathbb{E}(f(X)) - \mathbb{E}(f(Y))| \leq L\varepsilon + (M_2 - M_1)\ell_\varepsilon(F_X, F_Y). \tag{3.9}$$

Furthermore, if $g : \mathbb{R} \to \mathbb{R}$ is a function of finite total variation $V_{-K}^{K}(g)$ on any interval $[-K, K]$ with $K > 0$, we also get

$$|\,\mathbb{E}(g(X)) - \mathbb{E}(g(Y))| \leq \sup_{K>0} V_{-K}^{K}(g) \|F_X - F_Y\|_\infty \tag{3.10}$$

whenever $\sup_{K>0} V_{-K}^{K}(g)$ is finite.

Lemma 3.3 allows us to approximate $\mathbb{E}(f(X))$ by $\mathbb{E}(f(Y \pm \varepsilon))$ well for some functions f if $\ell_\varepsilon(F_X, F_Y)$ is small. In Chapter 4 we will use this result with X being the prediction error $y_0 - \hat{y}_0$ conditional on the training data and Y being a random variable distributed on a set of n points (the leave-one-out errors). In that case equation (3.10) yields a link between the generalization error and the leave-one-out error, which is closely related to Koksma's inequality (cf. Kuipers and Niederreiter 1974). For a more detailed discussion on the relationship of the generalization error and the leave-one-out error we refer to Bousquet and Elisseeff (2002). We would like to emphasize that the result of Lemma 3.3 stays valid for general random variables X and Y and also gives a link to the ε-variational divergence between F_X and F_Y.

3.3. Relation to other metrics

As we will see, the ε-variational divergence is closely related to the Kolmogorov distance and the Lévy metric: By definition, the Kolmogorov distance between two distribution functions F and G coincides with $\ell_0(F, G)$. In particular, equation (3.4) also entails the well-known inequality

$$F(t) - \|F - G\|_\infty \leq G(t) \leq F(t) + \|F - G\|_\infty \text{ for all } t \in \mathbb{R} \tag{3.11}$$

as a special case. Furthermore, we have the inequality $\ell_\varepsilon(F, G) \leq \ell_0(F, G) = \|F - G\|_\infty$ since $\ell_\varepsilon(F, G)$ is non-increasing in ε. We even get the following result:

Lemma 3.4. *Let $F : \mathbb{R} \to [0, 1]$ and $G : \mathbb{R} \to [0, 1]$ be two distribution functions. Then the function $\varepsilon \mapsto \ell_\varepsilon(F, G)$ is continuous from the right on the interval $[0, \infty)$ in the sense that*

$$\lim_{\varepsilon \searrow \delta} \ell_\varepsilon(F, G) = \ell_\delta(F, G) \quad \forall \delta \in [0, \infty).$$

In particular, the Kolmogorov distance $\|F - G\|_\infty = \ell_0(F, G)$ is the continuous extension of $\ell_\varepsilon(F, G)$ at the point 0 in the sense that $\lim_{\varepsilon \searrow 0} \ell_\varepsilon(F, G) = \ell_0(F, G)$.

As explained at the beginning of that section, the use of $\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n)$ can be much more helpful if the distribution F_n is not continuous. However, if both F_n and $\widehat{\mathrm{F}}_n$ are a_i-Hölder continuous with constants C_i ($i \in 1, 2$) respectively, then the two measurements

of distance are closely related in the sense that

$$\|\mathrm{F}_n - \widehat{\mathrm{F}}_n\|_\infty - C_1(\varepsilon/2)^{a_1} - C_2(\varepsilon/2)^{a_2} \leq \ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n) \leq \|\mathrm{F}_n - \widehat{\mathrm{F}}_n\|_\infty.$$

To see this, we start with an arbitrary $\delta > 0$ and a $t \in \mathbb{R}$ such that $\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_\infty \leq |\widehat{\mathrm{F}}_n(t) - \mathrm{F}_n(t)| + \delta$ holds true. We then have

$$\begin{aligned}|\widehat{\mathrm{F}}_n(t) - \mathrm{F}_n(t)| &\leq |\widehat{\mathrm{F}}_n(t) - \widehat{\mathrm{F}}_n(x)| + |\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(y)| + |\mathrm{F}_n(y) - \mathrm{F}_n(t)| \\ &\leq |\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(y)| + C_1(\varepsilon/2)^{a_1} + C_2(\varepsilon/2)^{a_2}\end{aligned}$$

for all $x, y \in K_{\varepsilon/2}(t)$, which implies

$$\begin{aligned}\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_\infty &\leq \delta + \inf_{x,y \in K_{\varepsilon/2}(t)} |\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(y)| + C_1(\varepsilon/2)^{a_1} + C_2(\varepsilon/2)^{a_2} \\ &\leq \delta + \ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n) + C_1(\varepsilon/2)^{a_1} + C_2(\varepsilon/2)^{a_2}.\end{aligned}$$

Since $\delta > 0$ can be made arbitrarily small, the statement follows. Furthermore, even without the assumption of Hölder continuity $\|\mathrm{F}_n - \widehat{\mathrm{F}}_n\|_\infty$ will be close to $\ell_\varepsilon(\mathrm{F}_n, \widehat{\mathrm{F}}_n)$ for small ε as Lemma 3.4 shows.

Equation (3.4) strongly suggests a relationship between the ε-variational divergence $\ell_\varepsilon(F, G)$ between F and G and the Lévy metric $L(F, G)$. By the monotonicity of F we get

$$\begin{aligned}F(t - \max(\varepsilon, \ell_\varepsilon(F,G))) - \max(\varepsilon, \ell_\varepsilon(F,G)) \leq G(t) \\ \leq F(t + \max(\varepsilon, \ell_\varepsilon(F,G))) + \max(\varepsilon, \ell_\varepsilon(F,G))\end{aligned}$$

for all $t \in \mathbb{R}$. Recalling the definition of the Lévy metric $L(F, G)$

$$L(F,G) := \inf\{\varepsilon \geq 0 : F(t-\varepsilon) - \varepsilon \leq G(t) \leq F(t+\varepsilon) + \varepsilon \text{ for all } t \in \mathbb{R}\} \tag{3.12}$$

we immediately conclude $L(F, G) \leq \max(\varepsilon, \ell_\varepsilon(F, G))$ for all $\varepsilon \geq 0$. By the continuity from the right of F and G we have

$$F(t - L(F,G)) - L(F,G) \leq G(t) \leq F(t + L(F,G)) + L(F,G) \text{ for all } t \in \mathbb{R}. \tag{3.13}$$

Using the approach of Lemma 3.2 together with equation (3.13) instead of equation (3.4), we could create a prediction interval

$$PI^+_{levy}(\hat{y}_0, q_1, q_2) := \hat{y}_0 + [q_1 - L(\mathrm{F}_n, \widehat{\mathrm{F}}_n), q_2 + L(\mathrm{F}_n, \widehat{\mathrm{F}}_n)]. \tag{3.14}$$

Replacing equation (3.4) by equation (3.13) in the proof of Lemma 3.2 we get the following guarantees:

$$\begin{aligned}\widehat{\mathrm{F}}_n(q_2) - \widehat{\mathrm{F}}^-_n(q_1) - 2L(\widehat{\mathrm{F}}_n, \mathrm{F}_n) \leq \mathbb{P}\left(y_0 \in PI^+_{levy}(\hat{y}_0, q_1, q_2) \| T_n\right) \\ \leq \widehat{\mathrm{F}}_n(q_2 + 2L(\widehat{\mathrm{F}}_n, \mathrm{F}_n)) - \widehat{\mathrm{F}}^-_n(q_1 - 2L(\widehat{\mathrm{F}}_n, \mathrm{F}_n)) + 2L(\widehat{\mathrm{F}}_n, \mathrm{F}_n) \text{ a.s.}\end{aligned} \tag{3.15}$$

However, using equation (3.14) together with equation (3.15) based on the Lévy metric instead of equation (3.2) and equation (3.5) has three major disadvantages: Firstly, we had to enlarge our intervals by an unknown and random quantity $L(\mathrm{F}_n, \widehat{\mathrm{F}}_n)$, which is not feasible in practice. To put it in other words: If F_n and $\widehat{\mathrm{F}}_n$ are distribution functions depending on the training data – as they will be in the remainder of the thesis – then we would have to increase the length of our prediction intervals by an unknown random length if we use the prediction interval given in equation (3.14). In contrast, $\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n)$ allows a statistician to determine a length ε she is willing to pay in advance (in the sense of an increasing interval length) and bounds the loss in coverage probability. More precisely, if $\mathbb{E}(\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n))$ is small or converges to 0 asymptotically, the actual conditional coverage probability is guaranteed to be close to or larger than its nominal level for most of the training data.

Secondly, equation (3.15) fails to give a guarantee for the coverage probability if our chosen ε-inflation is smaller than $L(\mathrm{F}_n, \widehat{\mathrm{F}}_n)$: In particular, if we estimate the unknown quantity $L(\mathrm{F}_n, \widehat{\mathrm{F}}_n)$ by some $\widehat{L}$, which is smaller than $L(\mathrm{F}_n, \widehat{\mathrm{F}}_n)$, equation (3.15) is not able to give a lower bound for the conditional coverage probability of $\hat{y}_0 + [q_1 - \widehat{L}, q_2 + \widehat{L}]$ even if $\widehat{L}$ is close to $L(\mathrm{F}_n, \widehat{\mathrm{F}}_n)$. In contrast, equation (3.5) gives guarantees for the coverage probability for every ε-inflation with $\varepsilon \geq 0$.

Thirdly, the Lévy metric is not able to deal with a scaling of our variables: For simplicity consider the case where we have found a prediction interval $[L, U]$ for y_0 with a guaranteed conditional coverage probability of at least $\alpha_2 - \alpha_1$. If we now change the units of our response variable resulting in a scaling of y_0 by some constant $c > 0$, then the prediction interval $[cL, cU]$ will have the same coverage probability for cy_0. Hence, it seems reasonable to ask for a procedure whose prediction intervals are scaling proportionally with the scaling of the response variable – at least under some conditions including the predictor $\hat{y}_0$ to be scale equivariant. Moreover, one could expect the bound for the coverage probability of the prediction interval $[cL, cU]$ for cy_0 to be independent of the scaling factor c.

However, the Lévy metric is not able to reflect the change of units appropriately. Denoting the function $t \mapsto \widehat{\mathrm{F}}_n(\frac{t}{c})$ with $\widehat{\mathrm{F}}_n(\frac{\cdot}{c})$, the Lévy metric does not fulfill the equation $L(\widehat{\mathrm{F}}_n(\frac{\cdot}{c}), \mathrm{F}_n(\frac{\cdot}{c})) = cL(\widehat{\mathrm{F}}_n, \mathrm{F}_n)$ for all $c > 0$ (unless $\widehat{\mathrm{F}}_n$ equals F_n) because the Lévy metric between two distribution functions is bounded by 1. Moreover, the prediction interval given in equation (3.14) will typically not be scale equivariant for every $c > 0$: Assume our prediction algorithm is scale equivariant in the sense that our prediction $\widehat{cy_0}$ for cy_0 corresponds with $c\hat{y}_0$ and we replace $q_1, q_2, \widehat{\mathrm{F}}_n$ and F_n in the definition of equation (3.14) by $cq_1, cq_2, \widehat{\mathrm{F}}_n(\frac{\cdot}{c})$ and $\mathrm{F}_n(\frac{\cdot}{c})$. Then the resulting prediction interval $\widetilde{PI}^+_{levy}(c\hat{y}_0, cq_1, cq_2)$ for cy_0 will not correspond with $cPI^+_{levy}(\hat{y}_0, q_1, q_2)$ for all $c > 0$. Furthermore, the bounds for the coverage probability $\mathbb{P}\left(cy_0 \in \widetilde{PI}^+_{levy}(c\hat{y}_0, cq_1, cq_2) \| T_n\right)$ given in equation (3.15) will depend on our choice of c, which is not surprising as the corresponding prediction interval does not scale appropriately.

In contrast, the ε-variational divergence $\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n)$ fulfills $\ell_{c\varepsilon}(\widehat{\mathrm{F}}_n(\frac{\cdot}{c}), \mathrm{F}_n(\frac{\cdot}{c})) = \ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n)$. Furthermore, we have $PI^+(c\hat{y}_0, cq_1, cq_2, c\varepsilon) = cPI^+(\hat{y}_0, q_1, q_2, \varepsilon)$ for every $c > 0$. That is, the prediction intervals given in equation (3.2) are scaling appropriately if the correspond-

ing values $\hat{y}_0, q_1, q_2$ and ε are replaced by their scaled analogues. Since the ε-variational divergence is scale equivariant in the sense that $\ell_{c\varepsilon}(\widehat{\mathrm{F}}_n(\frac{\cdot}{c}), \mathrm{F}_n(\frac{\cdot}{c})) = \ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n)$, equation (3.5) is also able to deal with a scaling well and will give the same coverage guarantees if we used the prediction interval $cPI^+(\hat{y}_0, q_1, q_2, \varepsilon)$ for cy_0. A more detailed discussion, why $\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n)$ is a more suitable choice for our prediction intervals than the $\mathcal{L}_p$-norms can be found in Chapter 6.

We would like to point out that $\ell_\varepsilon(\cdot, \cdot)$ is not even a pseudo-metric as it does not provide a triangle inequality: To see this, let F_0 be a distribution function and define $F_1(x) = F_0(x + \varepsilon)$ and $F_2(x) = F_1(x + \varepsilon) = F_0(x + 2\varepsilon)$ for all $x \in \mathbb{R}$. We then have $\ell_\varepsilon(F_0, F_1) = \ell_\varepsilon(F_1, F_2) = 0$, while at the same time $\ell_\varepsilon(F_0, F_2) = \|F_0 - F_1\|_\infty > 0$ holds true. Nevertheless, in the classical case of deterministic distribution functions it is closely related to the topology of weak convergence, which yields another remarkable link to the Lévy metric:

Lemma 3.5. *Let $F : \mathbb{R} \to [0, 1]$ be a distribution function and $(F_n)_{n\in\mathbb{N}}$ a sequence of distribution functions on $\mathbb{R}$. Then F_n converges weakly to F if, and only if $\lim_{n\to\infty} \ell_\varepsilon(F_n, F) = 0$ for all $\varepsilon > 0$.*

Since $L(F_n, F) \leq \max(\varepsilon, \ell_\varepsilon(F_n, F))$ holds true, the reverse direction of the proof is trivial. The more interesting statement of Lemma 3.5 is that $\ell_\varepsilon(F_n, F)$ converges to 0 if F_n converges weakly to F.

4. The Jackknife for prediction intervals

4.1. Preliminaries

After the theoretical discussion on distribution functions and their connection to prediction intervals in Chapter 3, we will head back to the case where we are given n i.i.d. data points (y_i, x_i') and a new regressor x_0 to predict y_0 by some predictor $\hat{y}_0$. For the remaining of the thesis we will always implicitly assume $n \geq 2$ as otherwise the leave-one-out approach will not provide a meaningful procedure. In the current chapter we will present a Jackknife-approach to estimate the conditional distribution function F_n of the prediction error $y_0 - \hat{y}_0$. Instead of finding or analysing some prediction algorithm, which performs well in some sense, we take the prediction algorithm as given and try to estimate the distribution function of the prediction error $y_0 - \hat{y}_0$ related to that algorithm conditional on the training data. Hence, our approach includes both accurate and ill-suited prediction algorithms. For this, let $1 \leq i \leq n$ and denote $T_{n-1}^{[-i]}$ the training data, where the i-th pair (y_i, x_i') is removed. For $0 \leq j \leq n$ and $1 \leq i \leq n$ let be $\tilde{y}_j^{[-i]}$ the prediction for y_j based on the regressor x_j and the training data $T_{n-1}^{[-i]}$. We then define the *leave-one-out* residuals $\hat{u}_i = y_i - \tilde{y}_i^{[-i]}$ and estimate F_n via the empirical distribution function $\widehat{\mathrm{F}}_n$ of the leave-one-out residuals:

$$\widehat{\mathrm{F}}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[\hat{u}_i, \infty)}(t).$$

Since $T_{n-1}^{[-i]}$ is independent from x_i and y_i, the leave-one-out approach is able to capture the inherent nature of independence of the algorithm fitted to the training data T_n from the new observation y_0. However, the approach may fail if the inclusion of one observation to the predictor results in a highly different prediction. We will refer to the property that a new observation does not change the prediction too much as the *stability* of a predictor and formalize this in Definition 4.4. For example, $\hat{y}_0$ could be the predictor $x_0' \widehat{\beta}_{LS}$ based on the OLS estimator $\widehat{\beta}_{LS}$ which uses all training data T_n and $\tilde{y}_i^{[-i]}$ would coincide with $x_i' \widetilde{\beta}_{LS}^{[-i]}$, where $\widetilde{\beta}_{LS}^{[-i]}$ only uses $T_{n-1}^{[-i]}$. In principle, our results can be applied to any combination of predictors $\hat{y}_0$ and $(\tilde{y}_1^{[-1]}, \ldots, \tilde{y}_n^{[-n]})$, where $\hat{y}_0$ uses all data and $\tilde{y}_i^{[-i]}$ excludes the i-th observation $(1 \leq i \leq n)$. However, in order to estimate F_n accurately by $\widehat{\mathrm{F}}_n$ we will need $\tilde{y}_i^{[-i]}$ to be close to $\hat{y}_0$ in some sense - which restricts the choice of possible predictors drastically.

To make the link to prediction intervals, we define the α-quantile $\hat{q}_\alpha$ for $0 \leq \alpha \leq 1$ as

$$\hat{q}_\alpha = \begin{cases} \hat{u}_{(\lceil \alpha n \rceil)} & \text{if } \alpha > 0 \\ \hat{u}_{(1)} - e^{-n} & \text{else,} \end{cases}$$

where $\hat{u}_{(1)} \leq \hat{u}_{(2)} \leq \ldots \leq \hat{u}_{(n)}$ are the ordered leave-one-out residuals. For $\alpha > 0$, this definition coincides with the empirical α-quantile of $\widehat{\mathrm{F}}_n$. We then define a prediction interval with nominal coverage probability of $\alpha_2 - \alpha_1$ for y_0 as follows:

$$PI_{\alpha_1,\alpha_2} = \hat{y}_0 + (\hat{q}_{\alpha_1}, \hat{q}_{\alpha_2}], \tag{4.1}$$

with $0 \leq \alpha_1 \leq \alpha_2 \leq 1$. As in the non-continuous case there possibly is no non-randomized prediction interval with coverage probability of $\alpha_2 - \alpha_1$, we define two prediction intervals as follows:

$$PI^+_{\alpha_1,\alpha_2}(\varepsilon) = \hat{y}_0 + [\hat{q}_{\alpha_1} - \varepsilon, \hat{q}_{\alpha_2} + \varepsilon], \tag{4.2}$$

$$PI^-_{\alpha_1,\alpha_2}(\varepsilon) = \hat{y}_0 + (\hat{q}_{\alpha_1} + \varepsilon, \hat{q}_{\alpha_2} - \varepsilon), \tag{4.3}$$

with $\varepsilon \geq 0$ and the convention that $(a, b) = \emptyset$ if $a \geq b$ and $[a, b] = \emptyset$ if $a > b$. Recalling equation (3.2) and equation (3.3), this entails $PI^+_{\alpha_1,\alpha_2}(\varepsilon) = PI^+(\hat{y}_0, \hat{q}_{\alpha_1}, \hat{q}_{\alpha_2}, \varepsilon)$ as well as $PI^-_{\alpha_1,\alpha_2}(\varepsilon) = PI^-(\hat{y}_0, \hat{q}_{\alpha_1}, \hat{q}_{\alpha_2}, \varepsilon)$. Furthermore, we will use the following abbreviation:

$$\ell_n(\varepsilon) := \ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n).$$

We then have the following result, which formalizes our approach of Chapter 3 of coverage probabilities for prediction intervals.

Proposition 4.1. *Let $\varepsilon \geq 0$, $0 \leq \alpha_1 \leq \alpha_2 \leq 1$ and denote $PI^+_{\alpha_1,\alpha_2}(\varepsilon)$ and $PI^-_{\alpha_1,\alpha_2}(\varepsilon)$ the prediction intervals given in* (4.2) *and* (4.3)*. We then have the following guarantees for the coverage probability:*

$$\mathbb{P}\left(y_0 \in PI^+_{\alpha_1,\alpha_2}(\varepsilon) \| T_n\right) \geq (\alpha_2 - \alpha_1) - 2\ell_n(\varepsilon) \ \text{a.s.} \tag{4.4}$$

$$\mathbb{P}\left(y_0 \in PI^-_{\alpha_1,\alpha_2}(\varepsilon) \| T_n\right) \leq (\alpha_2 - \alpha_1) + 2\ell_n(\varepsilon) \ \text{a.s.} \tag{4.5}$$

If y_0 conditional on x_0 is a continuous random variable almost surely, combining (4.4) *with* (4.5) *implies*

$$\left|\mathbb{P}\left(y_0 \in PI_{\alpha_1,\alpha_2} \| T_n\right) - (\alpha_2 - \alpha_1)\right| \leq 2\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_\infty \ \text{a.s.} \tag{4.6}$$

We would like to point out that the third statement cannot be derived from the first two results without the continuity assumption as $PI^+_{\alpha_1,\alpha_2}(0) = PI^-_{\alpha_1,\alpha_2}(0) \cup \{\hat{q}_{\alpha_1}, \hat{q}_{\alpha_1}\}$ is larger than $PI^-_{\alpha_1,\alpha_2}(0)$. Moreover, equation (3.1) cannot be applied in the general case as it requires to find points such that $\widehat{\mathrm{F}}_n(\hat{q}_\alpha) = \alpha$, which will typically not be the case for all values of α. Thus, in the case where y_0 given x_0 is continuous equation (4.6) is a generalization of equation (3.1). An inequality similar to equation (4.6) can be found in

Steinberger and Leeb (2023) (cf. Proposition 2.1 therein), where we were able to improve their factor 4 on the right-hand side to the factor 2 by a different proof. While the first two statements of Proposition 4.1 are a direct consequence of the definition of $\ell_n(\varepsilon)$, equation (4.6) draws a link between the actual and the nominal coverage probability for *every* level $0 \leq \alpha_2 - \alpha_1 \leq 1$. While it seems disappointing to have no such result for non-continuous distributions, it is no surprise that we are not able to achieve every coverage probability with a non-randomized prediction interval if our target function F_n is non-continuous. As we will see, many things will be easier in the continuous case. Thus, we will distinguish between the *general case* and the *continuous case* in our results by using the following (somewhat stronger) definition:

Definition 4.2 (**CC1** Assumption)**.** We say the Continuous Case Assumption 1 (in short **CC1**) is fulfilled if for almost every x the random variable y_0 conditional on $x_0 = x$ is absolutely continuous and the supremum norm of its density $f_{y_0 \| x_0 = x}$ is finite.

4.2. Finite sample results

Proposition 4.1 bounds the difference of the conditional (actual) coverage probability to some desired level in terms of $\ell_n(\varepsilon) = \ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n)$, which is unknown in practice. One might be interested in its expectation for several reasons. First of all, it allows to control the marginal coverage probability to some prescribed level. Secondly, one can use the expectation together with Markov's inequality to control the probability of $\ell_n(\varepsilon)$ getting too large, yielding a PAC-type inequality for the conditional coverage probability. Besides this, it might be of interest in view of Lemma 3.3.

Theorem 4.3. *The expected ε-variational divergence $\ell_n(\varepsilon) = \ell_\varepsilon(\mathrm{F}_n, \widehat{\mathrm{F}}_n)$ can be bounded from above as follows:*

(i) For every $\varepsilon > 0$, $\delta > 0$, $K \in \mathbb{N}$ and $\mu \in \mathbb{R}$ we have

$$\mathbb{E}\left(\ell_n(\varepsilon)\right) \leq \mathbb{P}\left(|y_0 - \hat{y}_0 - \mu| \geq \frac{(K-2)\varepsilon}{4}\right) + \left(\frac{K}{4(n-1)} + \frac{20\delta}{\varepsilon} + \frac{5K}{n}\sum_{i=1}^{n} \mathbb{P}(|\hat{y}_0 - \tilde{y}_0^{[-i]}| > \delta)\right)^{\frac{1}{2}}.$$

*(ii) If Assumption **CC1** is fulfilled, we get the following bound for $\varepsilon = 0$, i.e., the Kolmogorov distance:*

$$\mathbb{E}\left(\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_\infty\right) \leq \mathbb{P}(|y_0 - \hat{y}_0 - \mu| \geq 2L) + \mathbb{E}(\min(1, \delta \|f_{y_0 \| x_0}\|_\infty)) + \left[\left(\frac{4L}{\delta} + 2\right)\left(\frac{1}{4(n-1)} + \frac{5}{n}\sum_{i=1}^{n} \mathbb{E}(\min(1, \|f_{y_0 \| x_0}\|_\infty |\hat{y}_0 - \tilde{y}_0^{[-i]}|))\right)\right]^{\frac{1}{2}},$$

where $\delta > 0$, $\mu \in \mathbb{R}$ and $L > 0$ can be chosen arbitrarily.

In the continuous case, a similar statement can be found in Steinberger and Leeb (2023). While Theorem 2.6. of Steinberger and Leeb (2023) is stated for general k-fold cross validation, we focus on the Jackknife approach (coinciding with n-fold cross validation) with the minor improvement of replacing the term $\mathbb{P}(|y_0 - g(x_0)| \geq L_1) + \mathbb{P}(|g(x_0) - \hat{y}_0| > L_2)$ by $\mathbb{P}(|y_0 - \hat{y}_0 - \mu| \geq 2L)$, where g is a measurable function and $2L = L_1 + L_2$. However, the significance of Theorem 4.3 is its generalization to the non-continuous case. In addition, Theorem 4.3 shows that the trade-off between the stability of a predictor and the boundedness of its prediction error found in Steinberger and Leeb (2023) extends to the general case: A more stable predictor may be less precise in prediction and still get the same guarantees as a less stable, but more accurate predictor.

While Theorem 4.3 comes in a rather technical form, it can be read as follows: The conditional distribution of the prediction error can be estimated well (up to a dilatation of size ε) if the following two conditions are met: Firstly, the prediction should not change too much if one data point is excluded from the training data as otherwise the leave-one-out prediction can give a misleading picture of the full prediction. Secondly, the prediction error should not vary too much around a point μ, which implies that a systematic bias of the prediction does not impair the quality of its distribution's estimation. The latter is not really surprising as – given the stability of the predictor – a systematic bias should also be reflected by the leave-one-out predictors well and thus will be taken into account in its distribution's estimation. In the continuous case we additionally assume that $\|f_{y_0\|x_0}\|_\infty$ is finite for almost all x_0 to control the effects of a horizontal shift in F_n by some ε.

While the results of Theorem 4.3 are stated for a finite sample, the expressions will typically be hard to evaluate as they depend on the distribution of the leave-one-out predictor and the prediction error itself. However, we can hope that at least for large data sets the estimation of the prediction error's distribution will be accurate regardless of the underlying distribution. For this, we proceed as follows: In the following section we give sufficient conditions for consistent estimation of the prediction error's distribution, while in Chapter 5 we show that these conditions are fulfilled for a large class of data generating distributions.

4.3. Asymptotics

From now on, we are dealing with the following setting: For each $n \in \mathbb{N}$ there is a distribution $\mathcal{P}_n$, such that $(y^{(n)}, (x^{(n)})')$ follows the distribution $\mathcal{P}_n$, where $y^{(n)}$ is a (one-dimensional, real-valued) random variable and $x^{(n)}$ is a p_n-dimensional random vector. Furthermore, we are given n i.i.d. training data points (y_i, x_i'), distributed like $(y^{(n)}, (x^{(n)})')$, and a new regressor x_0, where (y_0, x_0') is again distributed like $(y^{(n)}, (x^{(n)})')$ and independent from the training data. Thus, all quantities considered now depend on n. For the sake of readability we will suppress this dependence whenever it is clear from the context. For example, if we assume the prediction error to be bounded in probability, we mean that the sequence of random variables $(y_0^{(n)} - \hat{y}_0^{(n)})_{n\in\mathbb{N}}$ is bounded in probability,

in the sense that for each $\varepsilon > 0$ there exists an $M > 0$, such that

$$\sup_{n\in\mathbb{N}} \mathbb{P}\left(|y_0^{(n)} - \hat{y}_0^{(n)}| \geq M\right) \leq \varepsilon,$$

where the probability is taken with respect to the training data T_n and $(y_0^{(n)}, (x_0^{(n)})')$.

As Theorem 4.3 shows, the influence of one single data point on the predictor should vanish for an increasing sample size if we want the Jackknife-approach to give a consistent estimation of the prediction error's distribution. We formalize this in the following definition:

Definition 4.4. We say a predictor $\hat{y}_0$ is asymptotically stable with respect to its *leave-one-out* analogues $(\tilde{y}_0^{[-1]}, \ldots, \tilde{y}_0^{[-n]})$ if for every $\varepsilon > 0$ the term $\frac{1}{n}\sum_{i=1}^n \mathbb{P}(|\hat{y}_0 - \tilde{y}_0^{[-i]}| > \varepsilon)$ converges to 0.

Sometimes we will abbreviate the definition and say $\hat{y}_0$ is asymptotically stable if its leave-one-out analogues are clear from the context. Moreover, we will denote $\tilde{y}_0^{[-n]}$ with $\tilde{y}_0$. Although the leave-one-out prediction $\tilde{y}_0^{[-i]}$ has the same distribution as $\tilde{y}_0$ for all $1 \leq i \leq n$ because the data (y_i, x_i') are independent and identically distributed, the term $\hat{y}_0 - \tilde{y}_0^{[-i]}$ need not have the same distribution for all $1 \leq i \leq n$. However, we will typically deal with symmetric predictors, which implies that the distributions of $\hat{y}_0 - \tilde{y}_0^{[-i]}$ and $\hat{y}_0 - \tilde{y}_0$ coincide for all $1 \leq i \leq n$. In that case the definition above reduces to the requirement that $\hat{y}_0 - \tilde{y}_0$ has to converge to 0 in probability.

As the following theorem shows, the conclusions drawn from Theorem 4.3 naturally transfer into the asymptotic setting:

Theorem 4.5 (Consistent estimation of F_n)**.** *Assume there exists a sequence $(v_n)_{n\in\mathbb{N}}$ of positive numbers, such that $(y_0 - \hat{y}_0)/v_n$ is bounded in probability and the scaled predictor y_0/v_n is asymptotically stable with respect to its scaled leave-one-out analogues $(\tilde{y}_0^{[-1]}/v_n, \ldots, \tilde{y}_0^{[-n]}/v_n)$.*

(i) General case: *For every $\varepsilon > 0$ we have $\lim_{n\to\infty} \mathbb{E}(\ell_n(\varepsilon v_n)) = 0$. In particular, there exists a null-sequence $(\varepsilon_n)_{n\in\mathbb{N}}$, such that $\lim_{n\to\infty} \mathbb{E}(\ell_n(\varepsilon_n v_n)) = 0$.*

(ii) Continuous Case: *If Assumption **CC1** is fulfilled and $v_n \|f_{y_0\|x_0}\|_\infty$ is bounded in probability, we get the following result for the case $\varepsilon = 0$: $\lim_{n\to\infty} \mathbb{E}(\ell_n(0)) = \lim_{n\to\infty} \mathbb{E}(\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_\infty) = 0$.*

Theorem 4.5 gives sufficient conditions for an asymptotically accurate estimation of the prediction error and extends existing results of Steinberger and Leeb (2023) to the non-continuous case. As mentioned in Chapter 3, we have $\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n) = \ell_{v_n\varepsilon}(\widehat{\mathrm{F}}_n(\frac{\cdot}{v_n}), \mathrm{F}_n(\frac{\cdot}{v_n}))$, which allows for an arbitrary scaling by some sequence $(v_n)_{n\in\mathbb{N}}$ at the price that we have to widen our prediction intervals by $v_n\varepsilon$ instead of ε. For example, we could choose v_n to be the standard deviation σ_y of y_0 and interpret Theorem 4.5 as follows: If the scaled prediction error $(y_0 - \hat{y}_0)/\sigma_y$ is bounded in probability and the scaled predictor

is asymptotically stable, we can estimate the conditional distribution F_n of the original prediction error $y_0 - \hat{y}_0$ by $\widehat{\mathrm{F}}_n$ well up to a variation of no more than $\mathcal{O}_p(\sigma_y)$.

The intuition behind this observation is simple: If we can estimate the distribution of $(y_0 - \hat{y}_0)/v_n$ well up to some shifting of $(y_0 - \hat{y}_0)/v_n$ by no more than ε, then we can estimate the distribution of $y_0 - \hat{y}_0$ well if we allow for a shifting by no more than $v_n\varepsilon$. Moreover, we also know that there is a null-sequence ε_n, such that the shift ε needed to deal with some discontinuity is allowed to vanish asymptotically (up to the scaling v_n). Thus, we do not lose much in terms of interval length compared to the approach of Steinberger and Leeb (2023) as our shifts are allowed to become arbitrarily small asymptotically.

In the continuous case, the factor v_n comes into play through the additional assumption on $v_n\|f_{y_0\|x_0}\|_\infty$ instead of shifting by some εv_n. Here, we also see the invariance to scaling: if all our data are scaled by some $1/v_n > 0$, then $\|f_{y_0/v_n\|x_0/v_n}\|_\infty$ equals $v_n\|f_{y_0\|x_0}\|_\infty$. Thus, a change of units or some other scaling will not have any influence in the estimation of the prediction error's distribution, which coincides with the intuition. We also want to mention that for all sequences of random variables $(X_n)_{n\in\mathbb{N}}$, one can find a scaling $(v_n)_{n\in\mathbb{N}}$, such that the scaled sequence of random variables $(X_n/v_n)_{n\in\mathbb{N}}$ is bounded (or converges to 0) in probability.[1] Thus, the challenge for applying Theorem 4.5 does not lie in finding *some* sequence $(v_n)_{n\in\mathbb{N}}$, such that the assumptions are met, but rather in finding the smallest such sequence (or at least a good approximation of it).

In many cases it seems natural to assume the data to be scaled a priori such that they are bounded over n in probability. From now on, we focus on the special case $v_n = 1$ for all $n \in \mathbb{N}$ for simplicity although our results hold in a more general setting with minor adaptations. We start with the following statement, which is a direct consequence of Theorem 4.5 combined with Proposition 4.1:

Corollary 4.6 (Prediction intervals via the Jackknife)**.** *Assume $y_0 - \hat{y}_0$ is bounded in probability and $\hat{y}_0$ is asymptotically stable. Then, for every $0 \le \alpha_1 \le \alpha_2 \le 1$ the following statements hold true:*

(i) General case: *For every $\varepsilon > 0$ and $\delta > 0$, we have*

$$\lim_{n\to\infty} \mathbb{P}\left(\mathbb{P}\left(y_0 \in PI^+_{\alpha_1,\alpha_2}(\varepsilon)\|T_n\right) \le \alpha_2 - \alpha_1 - \delta\right) = 0,$$
$$\lim_{n\to\infty} \mathbb{P}\left(\mathbb{P}\left(y_0 \in PI^-_{\alpha_1,\alpha_2}(\varepsilon)\|T_n\right) \ge \alpha_2 - \alpha_1 + \delta\right) = 0$$

and, in particular, this also entails the marginal coverage guarantees

$$\lim_{n\to\infty} \mathbb{P}\left(y_0 \in PI^-_{\alpha_1,\alpha_2}(\varepsilon)\right) \le \alpha_2 - \alpha_1 \le \lim_{n\to\infty} \mathbb{P}\left(y_0 \in PI^+_{\alpha_1,\alpha_2}(\varepsilon)\right).$$

(ii) Continuous Case: *If Assumption **CC1** is fulfilled and $\|f_{y_0\|x_0}\|_\infty$ is bounded in probability, we get the following result:*

$$\lim_{n\to\infty} \mathbb{E}(|\mathbb{P}\left(y_0 \in PI_{\alpha_1,\alpha_2}\|T_n\right) - (\alpha_2 - \alpha_1)|) = 0.$$

[1] For details see Lemma A.4 in the Appendix.

Again, Corollary 4.6 is an extension of results of Steinberger and Leeb (2023) to the non-continuous case. In the continuous case the statement of Corollary 4.6 is a consequence of Theorem 2.4. of Steinberger and Leeb (2023) with the minor generalization that we only need to guarantee the prediction error $y_0 - \hat{y}_0$ to be bounded in probability instead of bounding $y_0 - g(x_0)$ and $g(x_0) - \hat{y}_0$ for a measurable function g. Furthermore, the second statement in the general case is no surprise in view of Theorem 5 of Barber et al. (2021a), who are dealing with the *marginal* coverage probability, as in the case of a symmetric predictor the asymptotic stability can be linked to the *out-of-sample stability* property in their paper easily. In an ensuing remark Barber et al. (2021a) mention that in the case where $\|f_{y_0\|x_0}\|_\infty$ is bounded the *marginal* coverage probability of their prediction intervals will not undershoot $\alpha_2 - \alpha_1 - \delta_n$ for a small δ_n (converging to 0 for $n \to \infty$).[2] However, Corollary 4.6 gives a much more refined statement: Firstly, we are comparing the *conditional* coverage probability to its nominal level which indeed implies a similar result for the *marginal* coverage probability. Secondly, we do not require symmetric predictors. Furthermore, in the continuous case we also avoid overshooting the nominal coverage probability asymptotically. To sum it up, the important novelty of Corollary 4.6 is to provide a coverage guarantee for the ε-inflated prediction intervals in the general case, which is stated as an asymptotic version of a PAC-type inequality for prediction intervals.

The fact that we are only in the continuous case able to guarantee to meet the prescribed level $\alpha_2 - \alpha_1$ asymptotically is – again – a consequence of the fact that we use non-randomized prediction intervals. In contrast, it may not be possible to achieve every desired coverage probability in the non-continuous case with non-randomized prediction intervals.

In the continuous case the additional assumption on the boundedness of $\|f_{y_0\|x_0}\|_\infty$ is needed to ensure that the distribution is "sufficiently" smooth. Without that assumption a continuous distribution can be arbitrarily close to a non-continuous one, which can lead to the same problems as if the underlying distribution actually were non-continuous. To solve the problem, the following statement can be useful in the practical application: For all $\varepsilon > 0$ and $\delta > 0$ the following statement holds true:

$$\lim_{n\to\infty} \mathbb{P}\left(\widehat{\mathrm{F}}_n(\hat{q}_{\alpha_2} + 2\varepsilon) - \widehat{\mathrm{F}}_n^-(\hat{q}_{\alpha_1} - 2\varepsilon) + \delta \geq \mathbb{P}\left(y_0 \in PI^+_{\alpha_1,\alpha_2}(\varepsilon) \| T_n\right) \geq \alpha_2 - \alpha_1 - \delta\right) = 1. \tag{4.7}$$

To see this, we can use equation (3.5) of Lemma 3.2 with $q_2 = \hat{q}_{\alpha_2}$ and $q_1 = \hat{q}_{\alpha_1}$ to get

$$\begin{aligned}\widehat{\mathrm{F}}_n(\hat{q}_{\alpha_2} + 2\varepsilon) - \widehat{\mathrm{F}}_n^-(\hat{q}_{\alpha_1} - 2\varepsilon) + 2\ell_n(\varepsilon) &\geq \mathbb{P}\left(y_0 \in PI^+_{\alpha_1,\alpha_2}(\varepsilon) \| T_n\right) \\ &\geq \widehat{\mathrm{F}}_n(\hat{q}_{\alpha_2}) - \widehat{\mathrm{F}}_n^-(\hat{q}_{\alpha_1}) - 2\ell_n(\varepsilon) \text{ a.s.}\end{aligned}$$

Now, equation (4.7) follows since $\ell_n(\varepsilon) = \ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n)$ converges to 0 in probability by

[2] The Jackknife prediction intervals considered in Barber et al. (2021a) differ slightly from our ones in the following way: Firstly, they are using a symmetrized version in the sense that they are replacing the leave-one-out residuals by their absolute values. Furthermore, the definition of the quantiles does not coincide with our definition. For example, they are using $\lceil \alpha(n+1) \rceil$ instead of $\lceil \alpha n \rceil$.

Theorem 4.5 and $\widehat{\mathrm{F}}_n(\hat{q}_{\alpha_2}) - \widehat{\mathrm{F}}_n^-(\hat{q}_{\alpha_1}) \geq \alpha_2 - \alpha_1$. The crucial point of equation (4.7) is that $\widehat{\mathrm{F}}_n(\hat{q}_{\alpha_2}+2\varepsilon) - \widehat{\mathrm{F}}_n^-(\hat{q}_{\alpha_1}-2\varepsilon)$ is computable by the statistician. Hence, in applications one can compare the value $\widehat{\mathrm{F}}_n(\hat{q}_{\alpha_2}+2\varepsilon) - \widehat{\mathrm{F}}_n^-(\hat{q}_{\alpha_1}-2\varepsilon)$ with $\alpha_2 - \alpha_1$. If the both values are close, one can hope to meet the prescribed target well. Otherwise, there are many leave-one-out residuals close to $\hat{u}_{(\lceil n\alpha_i \rceil)}$ (for $i \in \{1,2\}$), which could indicate a high density or even a point mass of the prediction error's conditional distribution located at the boundaries of the prediction interval. In that case, the prediction interval has potentially a larger conditional coverage probability than desired while paying a price of a larger prediction interval by only 2ε.

However, the scope of this thesis does not end with the construction of prediction intervals. Lemma 3.3 even gives a full range of possible applications. To give an impression of the usability, we state the following proposition, which combines Lemma 3.3 with Theorem 4.5 appropriately:

Proposition 4.7 (Consistent estimation of the MSE and MAE)**.** *Let $k \in \mathbb{N}$ and both $|y_0 - \hat{y}_0|^k$ and $|y_0 - \tilde{y}_0|^k$ be uniformly integrable. If, additionally, the predictor $\hat{y}_0$ is asymptotically stable we have*

$$\lim_{n\to\infty} \mathbb{E}\left| \mathbb{E}\left(|y_0 - \hat{y}_0|^k \| T_n \right) - \frac{1}{n}\sum_{i=1}^{n} |\hat{u}_i|^k \right| = 0.$$

To put it in other words: if we want to estimate the mean-squared prediction error (or, alternatively, the mean-absolute prediction error), we can estimate it by the empirical second (or first) moment of the leave-one-out residuals accurately if the predictor is stable with respect to the exclusion of one data point. In fact, this conclusion is no surprise in view of the results of Bousquet and Elisseeff (2002) and can be seen as an asymptotic consequence of Lemma 9 therein. Furthermore, we need the uniform integrability of the prediction error as Lemma 3.3 only considers bounded functions and we have to deal with the tails separately. However, we do not lose much by the assumption of uniform integrability: If we want to estimate the second moment of the prediction error precisely in large samples, one would be willing to restrict only to the case, where the $(2+\varepsilon)$-th moment is uniformly bounded over n. As uniformly bounded $(2+\varepsilon)$-th moments imply uniform integrability of $|y_0 - \hat{y}_0|^2$, we do not lose much by posing this additional assumption.

We also want to point out that we do not distinguish between the general and the continuous case here. The reason for this is very simple: in the non-continuous case we can approximate the conditional MSE by $\frac{1}{n}\sum_{i=1}^{n}(|\hat{u}_i| \pm \varepsilon)^2$ from above and below. As we can make $\varepsilon > 0$ arbitrarily small, we achieve the original target.

We would like to emphasize another interpretation of Proposition 4.7: If we are dealing with binary classification, then Proposition 4.7 allows us to estimate the misclassification error well if the predictor $\hat{y}_0$ is asymptotically stable: That is, the probability of the original predictor assigning an element to the same group as its leave-one-out analogue should converge to 1.

5. Stability of prediction algorithms

In Chapter 4 we derived sufficient conditions for the Jackknife-approach to give asymptotically valid prediction intervals and accurate estimation of the mean-squared prediction error. We here show that for a wide class of distributions and some common prediction algorithms these conditions are fulfilled. In contrast to the results of Steinberger and Leeb (2023) (cf. Section 3 therein) we do not restrict our analysis to models which are linear or "close" to a linear model in some sense (cf. Theorem 3.1. therein). Furthermore, we do not assume the existence of an (additive) error term being independent of x_0. In fact, we do not pose any restrictions on the connection between y_0 and x_0. As we will see, the stability is an inherent property of the algorithms given some conditions on y_0 and x_0, but can be proven without any assumption on the connection between y_0 and x_0. Moreover, we allow y_0 to have a non-continuous conditional distribution given x_0. For example, in the case of binary classification y_0 could be a discrete variable with the values $\{0, 1\}$ only. Then asymptotic stability corresponds to the property that the probability that a predictor $\hat{y}_0$ assigns a different class than its leave-one-out analogue converges to 0 asymptotically.

To show that the sufficient conditions for the Jackknife-approach are fulfilled, we will distinguish between the high-dimensional setting and the case where $p < n$ asymptotically and derive results for both cases. Throughout the current section we will additionally assume that the limit of p/n exists and denote it with ρ:

$$\rho := \lim_{n\to\infty} \frac{p}{n} \in [0, \infty].$$

This does not pose a restriction, as we could always consider converging subsequences of $(p/n)_{n\in\mathbb{N}}$ and apply the following results to them.

Firstly, we will state our assumptions in the low-dimensional setting:

Definition 5.1 (Low-dimensional setting)**.** We say Assumption **LD** is fulfilled if the following statements hold true:

i) $\lim_{n\to\infty} p/n = \rho \in [0, 1)$.

ii) The second moment of y_0 is bounded by some constant S_y independent of n.

iii) For all $n \in \mathbb{N}$ the regressor x_0 has mean-zero and a positive definite covariance matrix $\Sigma > 0$ (which is allowed to vary over n).

and (at least) *one* of the following assumptions on $z_0 = \Sigma^{-\frac{1}{2}} x_0$ is fulfilled:

L1) The components of z_0 are independent with finite $(2+\delta)$-th moments uniformly bounded over n for some $\delta > 0$, i.e., $\mathbb{E}(|z_{0,i}|^{2+\delta}) \leq C$ for some δ and C independent of i and n.

L2) The components of $(z_0^{(n)})_{n\in\mathbb{N}}$ are independent and distributed like ξ, where ξ is not allowed to vary over n.

L3) The distribution of z_0 is log-concave for all $n \in \mathbb{N}$.

In contrast, we impose the following assumptions for the high-dimensional setting:

Definition 5.2 (High-dimensional setting)**.** We say Assumption **HD** is fulfilled if the following statements hold true:

i) $\lim_{n\to\infty} p/n = \rho \in (1, \infty]$.

ii) The second moment of y_0 is bounded by some constant S_y independent of n.

iii) For all $n \in \mathbb{N}$ the regressor x_0 has mean-zero and a positive definite covariance matrix $\Sigma > 0$ (which is allowed to vary over n).

iv) The vector $z_0 = \Sigma^{-\frac{1}{2}} x_0$ consists of *i.i.d.* components distributed like ξ_n, where ξ_n is allowed to vary over n.

and (at least) *one* of the following assumptions on ξ_n is fulfilled:

H1) ξ_n does not vary over n, i.e., $\xi_n = \xi$ for all $n \in \mathbb{N}$.

H2) The $(2+\delta)$-th moments of ξ_n are uniformly bounded, i.e., $\sup_{n\in\mathbb{N}} \mathbb{E}(|\xi_n|^{2+\delta}) \leq C$ for some $\delta > 0$.

H3) The distribution of ξ_n is log-concave for every $n \in \mathbb{N}$.

The condition on ρ separates the high-dimensional case from the case where $p < n$ asymptotically. The moment assumption on y_0 ensures that the mass of $\frac{1}{n}\sum_{i=1}^n y_i^2$ does not escape to ∞ for large n. Moreover, it even guarantees the boundedness in probability of $\frac{1}{n}\sum_{i=1}^n y_i^2$. On the other hand, the assumptions on x_0 and $\Sigma^{-\frac{1}{2}} x_0$ allow us to control the smallest eigenvalue of $\Sigma^{-1/2} X' X \Sigma^{-1/2}/n$ in the low-dimensional case and the smallest eigenvalue of $X\Sigma^{-1}X'/n$ in the high-dimensional setting as the ensuing lemma shows. However, it would suffice to pose the additional restriction on some $\tilde{z_0} = Rx_0$, such that $R\Sigma R' = I_p$. We decided to choose $R = \Sigma^{-1/2}$ for the sake of simplicity.

In the low-dimensional case this condition can even be relaxed to the assumption that $x_0 = S' z_0$, where z_0 is a centered random vector of $d \geq p$ independent components with unit variance and $S'S = \Sigma$ (cf. Lemma A.16 in the appendix). To put it in other words, x_0 can also be the linear function of a random vector z_0 whose dimension d is much larger than p, allowing for a more complicated structure. In particular, this includes misspecified models in the sense that y_0 can be a function of a d-dimensional random vector z_0, while we only observe some lower-dimensional linear combination of it. In that

case our predictor will use only the observable components rather than the full vector z_0. Furthermore, this also includes the case where we have decided to willingly use just some components x_0 of the original vector z_0 for example as a consequence of a model selection procedure.[1] Thus, in contrast to Steinberger and Leeb (2023), our results also include the case of misspecified models. For example, consider the linear model $y_0 = \theta' z_0 + u_0$, where z_0 is independent of u_0 and θ is an unknown parameter in $\mathbb{R}^d$. However, we only use $x_i = S' z_i$ instead of z_i in the training data ($1 \leq i \leq n$), where S is a non-random matrix of dimension $d \times p$. Then, the result of Lemma 5.3 for the low-dimensional case stays valid if the assumption L1), L2) or L3) is fulfilled for the d-dimensional random vector z_0 rather than for the p-dimensional random vector $\Sigma^{-1/2} x_0 = (S'S)^{-1/2} S' z_0$.

In fact, the assumptions on x_0 and $\Sigma^{-1/2} x_0$ are only used to guarantee that in the low-dimensional case the smallest eigenvalue of $\Sigma^{-1/2} X' X \Sigma^{-1/2}/n$ is asymptotically bounded away from 0 in the sense that $\lim_{n\to\infty} \mathbb{P}(\lambda_{\min}\left(\Sigma^{-1/2} X' X \Sigma^{-1/2}/n\right) \geq \delta) = 1$ for some $\delta > 0$. Analogously, in the high-dimensional case we have to control the smallest eigenvalue of $X\Sigma^{-1}X'/n$ (which coincides with the n-th largest eigenvalue of $\Sigma^{-1/2} X' X \Sigma^{-1/2}/n$). Hence, we can replace the assumptions on x_0 by any other condition, which allows to control the smallest eigenvalue.

Lemma 5.3 (Controlling the singular values of $X\Sigma^{-1/2}$)**.** *Let X be the random matrix consisting of n i.i.d. rows distributed like x_0. We then have*

1. ***Low-dimensional case:*** *If Assumption **LD** is fulfilled, then the smallest eigenvalue of $\Sigma^{-1/2} X' X \Sigma^{-1/2}/n$ converges to $(1 - \sqrt{\rho})^2$ in probability.*

2. ***High-dimensional case:*** *If Assumption **HD** is fulfilled, then the smallest eigenvalue of $X\Sigma^{-1}X'/n$ converges to $(1 - \sqrt{\rho})^2$ in probability for $\rho < \infty$. In the case $\rho = \infty$ the smallest eigenvalue of $X\Sigma^{-1}X'/n$ tends to ∞ in the sense that*

$$\lim_{n\to\infty} \mathbb{P}\left(\lambda_{\min}\left(X\Sigma^{-1}X'/n\right) \leq M\right) = 0$$

   for all $M > 0$.

The statements hold true if we replace the matrix X by its leave-one-out analogue $\tilde{X}$, consisting of $n-1$ i.i.d. rows distributed like x_0.

Lemma 5.3 fundamentally relies on the results of Chafaï and Tikhomirov (2018) with some minor extensions. In contrast to the classical Bai-Yin theorem we allow for dependence in the low-dimensional case and, in general, for a distribution which changes over n. However, we would like to stress that Lemma 5.3 only proves convergence in probability, while the statement in Bai and Yin (1993) yields almost sure convergence.

As we will see, the control of the eigenvalues is essential for some predictors to show that the Jackknife-approach can be used to estimate the prediction error's distribution consistently. In the following subsections we will present some predictors and show that they fulfill the stability assumption and that their prediction errors are bounded in

[1] However, the matrix S has to be deterministic. In order to fulfill this requirement we could apply the model selection to a different data set and condition on the resulting model described by the matrix S.

probability in order to meet the prerequisites of Theorem 4.5 and Corollary 4.6. Hence, in the following cases the Jackknife-approach gives asymptotically valid prediction intervals.

In the following sections we will deal with predictors which are linear with respect to the regressor x_0 in the sense that $\hat{y}_0 = x_0'\widehat{\beta}$ and $\tilde{y}_0^{[-i]} = x_0'\widetilde{\beta}^{[-i]}$ for $\widehat{\beta}$ and $\widetilde{\beta}^{[-i]}$ being a function of T_n and $T_{n-1}^{[-i]}$, respectively. Furthermore, the following predictors are symmetric, i.e., they are invariant to any permutation of the order of the training data. We consider this to be the prevailing case as the data are assumed to be i.i.d. and there is no need to treat them differently. Thus, for each $i \in \{1, \ldots, n\}$ the term $x_0'(\widehat{\beta} - \widetilde{\beta}^{[-i]})$ has the same distribution and the asymptotic stability reduces to the requirement of $x_0'(\widehat{\beta} - \widetilde{\beta}^{[-n]})$ converging to 0 in probability. As here the distribution is not affected by the index i, we will simplify our notation and write $\widetilde{\beta}$ instead of $\widetilde{\beta}^{[-n]}$. Moreover, we will use the notation $X = (x_1, \ldots, x_n)'$, $Y = (y_1, \ldots, y_n)'$ and its leave-one-out analogues $\tilde{X} = (x_1, \ldots, x_{n-1})'$ and $\tilde{Y} = (y_1, \ldots, y_{n-1})'$. Furthermore, we will write T_{n-1} for $T_{n-1}^{[-n]}$.

5.1. The Ridge

We start with the ridge estimator as we do not require any additional assumptions on the data generating process and only need an appropriate choice of the tuning parameter $c_{n,p}$. Furthermore, the ridge estimator is asymptotically stable regardless of the value of ρ. Moreover, we even allow the tuning parameter to be data-dependent. To model this, we start with the formal definition:

Definition 5.4. Let $c_{n,p}$ denote a T_n-measurable, random variable with values in $[0, \infty)$ almost surely which is symmetric with respect to the training data. Then we define the ridge estimator $\widehat{\beta}_R(c_{n,p})$ with a (possibly data-depending) tuning parameter $c_{n,p}$ as

$$\widehat{\beta}_R(c_{n,p}) = (X'X + c_{n,p}I_p)^\dagger X'Y,$$

where the Moore-Penrose pseudoinverse is required to cover the cases where $c_{n,p} = 0$. Moreover, we define $c_{n-1,p}^{[-i]}$ to be the leave-one-out analogue of $c_{n,p}$ based on the training data $T_{n-1}^{[-i]}$ and abbreviate $c_{n-1,p}^{[-n]}$ with $c_{n-1,p}$.

Here, the symmetry of $c_{n,p}$ is assumed for simplicity. Without this the Ridge would not be symmetric any more and we would have to use the original definition of asymptotic stability, implying a small change in the assumptions of Theorem 5.5, Proposition 5.6 and Proposition 5.7: For example, the requirement $\lambda_{\max}(\Sigma^{-1})|c_{n,p} - c_{n-1,p}|/n \xrightarrow{p} 0$ would change to the assumption that for every $\varepsilon > 0$ the expression $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}(\lambda_{\max}(\Sigma^{-1})|c_{n,p} - c_{n-1,p}^{[-i]}|/n > \varepsilon) = 0$. Moreover, we could also use a randomized tuning parameter, in the sense that it suffices to ask for $c_{n,p}$ being independent from (y_0, x_0') and the leave-one-out analogue $c_{n-1,p}^{[-i]}$ being independent from (y_i, x_i') and (y_0, x_0'). Furthermore, as the tuning parameter $c_{n,p}$ is allowed to depend on the training data, it can be chosen, for example, by GCV or cross-validation.

Equipped with the definition we are now ready to state the following proposition:

Theorem 5.5 (Assumption-lean stability and boundedness of the Ridge)**.** *Suppose the second moments of x_0 exist for all $n \in \mathbb{N}$ and denote $\mathbb{E}(x_0 x_0')$ with Σ. Then the prediction error $y_0 - x_0' \widehat{\beta}_R(c_{n,p})$ is bounded in probability and the predictor $x_0' \widehat{\beta}_R(c_{n,p})$ is asymptotically stable in the sense that $x_0'(\widehat{\beta}_R(c_{n,p}) - \widetilde{\beta}_R(c_{n-1,p})) \xrightarrow{p} 0$ if the following conditions are met:*

1. *$c_{n,p}$ is positive asymptotically: $\lim_{n\to\infty} \mathbb{P}(c_{n,p} = 0) = 0$.*
2. *The ratio of the tuning parameters $c_{n-1,p}/c_{n,p}$ converges in probability to 1.*
3. *$\lambda_{\max}(\Sigma) n / c_{n,p}$ is bounded in probability.*
4. *The second moments of y_0 are uniformly bounded in n.*

In this Section, we will present two more propositions for the Ridge estimator. However, the strength of Theorem 5.5 lies in its universality: It works regardless of the matrix X and the value of ρ and can even be applied for the critical case $\rho = 1$ as long as the tuning parameter $c_{n,p}$ is of the same magnitude as (or larger than) $n\lambda_{\max}(\Sigma)$. Moreover, we emphasize that it does not require Assumption **LD** or **HD** to hold true. As pointed out before, this can be especially of interest in the cases where ρ is very close (or equal) to 1 as in these cases the OLS or minimum-norm interpolator will typically have a bad performance.[2] We also stress the fact that condition 1 and condition 2 in Theorem 5.5 are automatically fulfilled if the tuning parameters are deterministic and chosen such that $c_{n-1,p} = c_{n,p} > 0$ holds true.

In order to analyze the stability of the OLS estimator and the minimum-norm interpolator we present two propositions, which allow the tuning parameter $c_{n,p}$ to be 0.

Proposition 5.6 (The Ridge in low dimensions)**.** *Let Assumption **LD** be fulfilled. If $\lambda_{\max}(\Sigma^{-1})|c_{n,p} - c_{n-1,p}|/n$ converges in probability to 0, then the prediction error $y_0 - x_0' \widehat{\beta}_R(c_{n,p})$ is bounded in probability and the predictor $x_0' \widehat{\beta}_R(c_{n,p})$ is asymptotically stable.*

While the assumptions of Theorem 5.5 are easier to fulfill the larger $c_{n,p}$ is, the assumptions of Proposition 5.6 favor a small difference $c_{n,p} - c_{n-1,p}$. The reason for this is the following: In both proofs we need to control the spectral norm of $\Sigma^{1/2}(X'X + c_{n,p} I_p)^{\dagger} X'$. In Theorem 5.5 this is achieved by bounding it from above by $1/(2\sqrt{c_{n,p}})$ and therefore needs large values of $c_{n,p}$. In Proposition 5.6 we can additionally control the smallest eigenvalue of $\Sigma^{-1/2} X'X \Sigma^{-1/2}/n$ due to Assumption **LD**, which suffices to control the spectral norm of $\Sigma^{1/2}(X'X + c_{n,p} I_p)^{\dagger} X'$. We only have to bound the difference between $c_{n,p}$ and $c_{n-1,p}$ in order to guarantee that the two predictors $x_0' \widehat{\beta}_R(c_{n,p})$ and $x_0' \widetilde{\beta}_R(c_{n-1,p})$ do not differ too much.

We also want to point out that the condition on $\lambda_{\max}(\Sigma^{-1})|c_{n,p} - c_{n-1,p}|/n$ is automatically fulfilled if the tuning parameter is independent of the data and the statistician uses

[2] For a more detailed discussion why the Ridge outperforms the OLS near $\rho = 1$ for large penalty parameters we refer to Hastie, Montanari, et al. (2022).

the same tuning parameter for the prediction as for the calculation of the *leave-one-out* residuals, i.e., $c_{n,p} = c_{n-1,p}$. Conversely, if a symmetric tuning parameter $c_{n,p}$ does not depend on (y_n, x_n') and fulfills $c_{n-1,p} = c_{n,p}$ almost surely, Lemma D.6 in Steinberger and Leeb (2023) shows that $c_{n,p}$ cannot depend on the training data at all in the sense that $c_{n,p}$ is constant almost surely.[3] However, if $c_{n,p}$ does depend on (y_n, x_n') (e.g., in case $c_{n,p}$ is chosen by cross-validation) we cannot set $c_{n-1,p} := c_{n,p}$ as then (y_n, x_n') and $c_{n-1,p}$ would be no longer independent. Next, we present the high-dimensional analogue to Proposition 5.6, which additionally needs to bound the condition number of the matrix Σ.

Proposition 5.7 (The Ridge in high dimensions)**.** *The prediction error $y_0 - x_0'\widehat{\beta}_R(c_{n,p})$ is bounded in probability and the predictor $x_0'\widehat{\beta}_R(c_{n,p})$ is asymptotically stable in the sense that $x_0'(\widehat{\beta}_R(c_{n,p}) - \widetilde{\beta}_R(c_{n-1,p})) \xrightarrow{p} 0$ if Assumption **HD** is fulfilled and the following two properties are fulfilled:*

1. $\lambda_{\max}(\Sigma^{-1})\dfrac{c_{n,p} - c_{n-1,p}}{n} \xrightarrow{p} 0.$
2. *The condition number of Σ is bounded over n, i.e.,*

$$\limsup_{n\to\infty} \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} < \infty.$$

To sum it up, the Ridge is an intrinsically stable predictor in the following sense: For a large penalty parameter $c_{n,p}$ the regularization is strong enough to stabilize the ridge estimator regardless of the number of regressor variables. On the other hand, if the singular values of the regressor matrix are only affected a little by the removal of one row (which is guaranteed by Assumption **LD** or Assumption **HD**), the Ridge needs no regularization to stabilize and hence can be used for our approach even for small penalty parameters. Thus, in any of the cases above the Jackknife-approach performs well for prediction intervals.

Our results fit well with existing results addressing the stability of M-estimators: El Karoui (2018) also showed the boundedness in probability of the prediction error and the asymptotic stability of M-estimators under some assumptions (cf. Theorem 2.1 and Theorem 2.2. therein). However, we neither assume a linear model nor the existence of an independent error term (in fact, we made no assumptions on the connection between y_0 and x_0 at all) at the cost of a bounded second moment of y_0. Moreover, El Karoui (2018) only considered the case where $c_{n,p} = c_{n-1,p} = \tau n$ for a fixed $\tau > 0$ (independent of n) and $\Sigma = I_p$.

[3] To see this, we apply Lemma D.6 in Steinberger and Leeb (2023) to the predictor $\hat{\mu}_n := c_{n,p}$ and note that by the symmetry of $c_{n,p}$ the distributions of $c_{n,p} - c_{n-1,p}^{[-i]}$ and $c_{n,p} - c_{n-1,p}$ coincide for all $1 \le i \le n$, which implies that the stability coefficient of Lemma D.6 equals 0.

5.2. OLS

As mentioned before, Assumption **LD** allows us to use the Ridge even for very small regularization parameters. This observation can be put to its extreme, which leads to the OLS estimator: By setting $c_{n,p} = 0$, Proposition 5.6 allows us to directly derive a result for the least-squares estimator:

Corollary 5.8. *Let $\widehat{\beta}_{LS} = X^{\dagger}Y$ be the OLS estimator.*[4] *If Assumption **LD** is fulfilled, then the prediction error $y_0 - x_0'\widehat{\beta}_{LS}$ is bounded in probability and $x_0'(\widehat{\beta}_{LS} - \widetilde{\beta}_{LS})$ converges to 0 in probability.*

All we need for Corollary 5.8 is to ensure that the smallest eigenvalue of the matrix $\Sigma^{-1/2}X'X\Sigma^{-1/2}/n$ stays away from 0 asymptotically and that $\|Y\|_2^2/n$ is bounded in probability, which are both guaranteed by Assumption **LD**. However, this eigenvalue decreases the larger $\rho = \lim_{n\to\infty} p/n$ gets. In the extreme case of $\rho = 1$, the OLS will fail to meet the assumptions of Theorem 4.5 as the smallest eigenvalues can be arbitrarily close to 0 in that case. However, the case $\rho = 1$ is excluded by Assumption **LD**. We would like to point out that in this case the Ridge can still perform well in our approach if the regularization through $c_{n,p}$ is large enough as can be seen from Theorem 5.5.

5.3. Minimum-norm interpolator

As for ρ larger than one the non-zero eigenvalues of $\Sigma^{-1/2}X'X\Sigma^{-1/2}/n$ can be bounded away from 0, one could ask whether in the high-dimensional case the minimum-norm interpolator can also perform well. From Proposition 5.7 we can immediately conclude by setting $c_{n,p} = c_{n-1,p} = 0$ the following result for the minimum-norm interpolator:

Corollary 5.9. *Let $\widehat{\beta}_{MN} = X^{\dagger}Y$ denote the minimum norm interpolator. If Assumption **HD** is fulfilled and the condition number of the matrix Σ is bounded in n, then the prediction error $y_0 - x_0'\widehat{\beta}_{MN}$ is bounded in probability and the predictor is asymptotically stable.*

Thus, the Jackknife-approach can be used together with the minimum-norm interpolator in the high-dimensional setting.

5.4. James-Stein estimator

In the current subsection we will consider a James-Stein type estimator $\widehat{\beta}_{JS}(c_{n,p})$, which we define as follows:

Definition 5.10. Let $c_{n,p}$ denote a measurable function of T_n which is symmetric with respect to the training data and lies in $[0, 1]$ almost surely. Then we define the James-Stein

[4] We extend the definition of the OLS to the case, where the matrix X does not posses full rank p. However, the matrix $X'X$ will be invertible with asymptotic probability 1 under Assumption **LD**.

estimator $\widehat{\beta}_{JS}(c_{n,p})$ with a (possibly data-depending) tuning parameter $c_{n,p}$ as

$$\widehat{\beta}_{JS}(c_{n,p}) = \begin{cases} \max\left(0, \left(1 - \frac{c_{n,p} p \hat{\sigma}^2}{\widehat{\beta}'_{LS} X' X \widehat{\beta}_{LS}}\right)\right) \widehat{\beta}_{LS} & \text{if } 0 < \widehat{\beta}'_{LS} X' X \widehat{\beta}_{LS} \\ 0 & \text{else,} \end{cases}$$

where $\hat{\sigma}^2 = \|Y - X\widehat{\beta}_{LS}\|_2^2/(n-p)$.

The definition of $\widehat{\beta}_{JS}(c_{n,p})$ is closely related to Baranchik (1973) and – in contrast to the original definition in James and Stein (1961) – takes not only $\|X\widehat{\beta}_{LS}\|_2$, but also $\hat{\sigma}$ into account. The use of the James-Stein estimator may be of interest as it can perform comparable to the Maximum-Likelihood estimator in terms of the *out-of-sample* risk in a linear regression model (cf. Huber and Leeb 2013).[5]

Proposition 5.11. *If Assumption **LD** is fulfilled, then the prediction error of the James-Stein estimator $y_0 - x_0'\widehat{\beta}_{JS}(c_{n,p})$ is bounded in probability. Furthermore, the James-Stein estimator is asymptotically stable if, additionally, $c_{n,p} - c_{n-1,p}$ converges to 0 in probability.*

Similar to the OLS estimator, we only need to control the smallest eigenvalue of the matrix $\Sigma^{-1/2} X' X \Sigma^{-1/2}/n$ and the tails of $\|Y\|_2^2/n$ to ensure the prediction error based on the James-Stein estimator to be bounded in probability. For the stability we additionally have to ensure that $c_{n,p}$ is close to $c_{n-1,p}$ as otherwise the leave-one-out residuals are reflecting the behavior of a different James-Stein estimator. Again, in the case where the tuning parameter does not depend on the training data, this can easily be achieved by setting $c_{n-1,p} = c_{n,p}$.

5.5. Binary classification

In this section we present two results for binary classification. Since here y_0 can only take the values -1 or 1, its conditional distribution given x_0 is discrete. We would like to point out, that in the case of binary classification estimating F_n accurately would allow us to estimate the misclassification error (conditional on our training data) of our method with high precision. Assuming that a predictor in a binary classification setting is also restricted to the values -1 or 1, we have $\mathbb{E}\left(|y_0 - \hat{y}_0| \| T_n\right) = 2\mathbb{P}\left(y_0 \neq \hat{y}_0 \| T_n\right)$. Thus, estimating the conditional misclassification error coincides with estimating the conditional mean-absolute error (up to the factor 2). Since the prediction error is bounded, Proposition 4.7 guarantees a consistent estimation of the misclassification error as long as the predictor is asymptotically stable.

We start with the following definitions:

[5] However, Huber and Leeb also state that this conclusion may change dramatically if one considers a worst-case scenario instead of averaging over all possible cases.

Definition 5.12. We define $\widetilde{\operatorname{sgn}} : \mathbb{R} \to \{-1, 1\}$ as

$$\widetilde{\operatorname{sgn}}(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0. \end{cases}$$

Definition 5.13. Let $\widehat{\beta}$ be a real-valued, measurable function of the training data T_n, which is symmetric in the sense that it does not change by a permutation of the training data. We then define the predictor for y_0 based on the training data T_n and a new feature vector x_0 as $\hat{y}_0 := \widetilde{\operatorname{sgn}}(x_0'\widehat{\beta})$. We will refer to $\widehat{\beta}$ as the classifier corresponding to $\hat{y}_0$.

The following proposition links the stability of the predictor to the stability of the corresponding classifier:

Proposition 5.14. *Let $\hat{y}_0$ be a predictor as in Definition 5.13 and $\widehat{\beta}$ be the corresponding classifier. Furthermore, let $\tilde{y}_0$ and $\widetilde{\beta}$ denote the leave-one-out analogues of $\hat{y}_0$ and $\widehat{\beta}$, respectively. Then $\hat{y}_0$ is asymptotically stable if the two assumptions are met:*

1. *The term $x_0'\widehat{\beta}$ is asymptotically stable in the sense that $x_0'(\widehat{\beta} - \widetilde{\beta})$ converges to 0 in probability.*
2. *The term $x_0'\widehat{\beta}$ is bounded away from 0 asymptotically in the sense that*

$$\lim_{\delta \searrow 0} \limsup_{n \to \infty} \mathbb{P}(|x_0'\widehat{\beta}| \leq \delta) = 0. \tag{5.1}$$

If $y_0 \in \{-1, 1\}$, then the prediction error trivially fulfills $|y_0 - \hat{y}_0| \leq 2$ almost surely and is therefore bounded in probability.

Proposition 5.14 shows that the asymptotic stability of the underlying classifier transfers to the predictor if the classifier does not concentrate at 0 asymptotically as without that condition an arbitrarily small change in the classifier could result in a different prediction since the function $\widetilde{\operatorname{sgn}}$ is discontinuous at 0. We would like to point out that we already showed that condition 1 holds true (under some assumptions) for several classifiers $\widehat{\beta}$ including the Ridge estimator.

While Proposition 5.14 yields sufficient conditions for a class of predictors in the binary classification case, we also present an example, where we prove the asymptotic stability directly.

Definition 5.15. Let $n \in \mathbb{N}, p \in \mathbb{N}$ and assume we are given training data $(y_i, x_i')_{i=1}^n$, such that $y_i \in \{-1, 1\}$ and $x_i \in \mathbb{R}^p$ for every $1 \leq i \leq n$. If there exists a vector $(\alpha, \gamma')' \in \mathbb{R}^{p+1}$ such that $y_i(x_i'\gamma + \alpha) \geq 1$ for all $1 \leq i \leq n$, then any solution $(\widehat{\alpha}, \widehat{\gamma}')'$ of the optimization problem

$$\begin{aligned} &\min_{(\alpha,\gamma')' \in \mathbb{R}^{p+1}} \|\gamma\|_2 \\ &\text{s.t. } y_i(x_i'\gamma + \alpha) \geq 1 \text{ for all } 1 \leq i \leq n \end{aligned} \tag{5.2}$$

is called a (linear hard-margin) support vector classifier or the optimal separating hyperplane (cf. Burges 1998, Vapnik 1999 and Hastie, Tibshirani, et al. 2009) with respect to the data $(y_i, x_i')_{i=1}^n$.

For the support vector classifier we assume the following model:

Definition 5.16 (Linearly separable binary model)**.** For every $n \in \mathbb{N}$ let ξ_n be a p_n-dimensional absolutely continuous random vector (with respect to the p_n-dimensional Lebesgue-measure), b_n a vector in $\mathbb{R}^{p_n}$, $a_n \in \mathbb{R}$ and $\eta_n = \widetilde{\operatorname{sgn}}(b_n' \xi_n + a_n)$. Furthermore, for every $n \in \mathbb{N}$ the feature-response pairs $(y_1, x_1'), \ldots, (y_n, x_n')$ of the training data are independent and identically distributed like (η_n, ξ_n').

The linearly separable binary model guarantees the existence of a linear hard-margin support vector classifier (cf. Lemma A.21 in the Appendix). We stress the fact that the predictor based on the support vector classifier is unique even if the corresponding classifier is not.[6] Since the support vector classifier $(\widehat{\alpha}, \widehat{\gamma}')'$ is symmetric by definition and the corresponding predictor $\hat{y}_0 = \widetilde{\operatorname{sgn}}(\widehat{\gamma}' x_0 + \widehat{\alpha})$ can only take two values, the predictor $\hat{y}_0$ is asymptotically stable if $\lim_{n \to \infty} \mathbb{P}(\hat{y}_0 \neq \tilde{y}_0) = 0$, where $\tilde{y}_0 = \widetilde{\operatorname{sgn}}(\widetilde{\gamma}' x_0 + \widetilde{\alpha})$ denotes its leave-one-out analogue based on the leave-one-out classifier $(\widetilde{\alpha}, \widetilde{\gamma}')'$. As the following result shows, the linear hard-margin support vector classifier is asymptotically stable in the linearly separable binary model as long as the number of observations n grows faster than the dimension p_n:

Proposition 5.17. *Assume a linearly separable binary model as in Definition 5.16. Let $\hat{y}_0 = \widetilde{\operatorname{sgn}}(\widehat{\gamma}' x_0 + \widehat{\alpha})$ be the prediction based on the support vector classifier $(\widehat{\alpha}, \widehat{\gamma})$ derived from the training data T_n and $\tilde{y}_0$ its leave-one-out analogue. We then have*

$$\mathbb{P}(\hat{y}_0 \neq \tilde{y}_0) \leq \frac{p_n + 1}{n}.$$

In particular, the predictor $\hat{y}_0$ based on the support vector classifier is asymptotically stable if $\lim_{n \to \infty} p_n / n = 0$. Furthermore, its prediction error is bounded by definition.

Proposition 5.17 shows that the support vector classifier is indeed asymptotically stable in the linearly separable binary model as long as the number of observations grows faster than the dimension of the underlying feature space. Furthermore, it even bounds the probability that a predictor based on the support vector classifier differs from its leave-one-out analogue in finite samples.

[6] This counterintuitive situation can be seen as follows: The support vector classifier is unique unless all observations y_i belong to the same class, that is $y_i = y_1$ for all $1 \leq i \leq n$. However, in the case of only one class every predictor based on a support vector classifier assigns a new observation to the same class as y_1 and hence is unique (cf. Lemma A.21 in the Appendix).

6. Discussion

6.1. On the choice of the distance measure

The main advantage of the ε-variational divergence is given by equation (3.4). Although equation (3.13) gives a similar inequality for the Lévy metric, we would refrain from using the Lévy metric due to the disadvantages explained in Chapter 3. However, one could consider the $\mathcal{L}_p$-norm $\|\cdot\|_{\mathcal{L}_p}$ for $p \in [1, \infty)$ and ask, whether a similar result to equation (3.4) can be found. As the following lemma shows, there is indeed a connection between the ε-variational divergence $\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n)$ and $\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_p}$:

Lemma 6.1. *For every $\varepsilon > 0$ and $p \in (0, \infty)$ we have*

$$\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n) \leq \frac{\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_p}}{\varepsilon^{\frac{1}{p}}} \ \text{a.s.}$$

Furthermore, this naturally extends to the case $p = \infty$:

$$\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n) \leq \|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_\infty} \ \text{a.s.}$$

*If Assumption **CC1** is fulfilled, we have for every $\varepsilon > 0$ and $p \in (0, \infty)$*

$$\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_\infty \leq \mathbb{E}(\min(1, \varepsilon \|f_{y_0\|x_0}\|_\infty)) + \frac{\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_p}}{\varepsilon^{\frac{1}{p}}} \ \text{a.s.}$$

Hence, whenever $\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_p}$ converges to 0, the same holds true for the ε-variational divergence for every $\varepsilon > 0$. If, additionally, Assumption **CC1** is fulfilled and $\|f_{y_0\|x_0}\|_\infty$ is bounded in probability, also the Kolmogorov distance vanishes asymptotically. Furthermore, combining Lemma 6.1 with equation (4.4) immediately gives the following result for all $\varepsilon > 0$ and $p \in (0, \infty)$:

$$\mathbb{P}\left(y_0 \in PI^+_{\alpha_1,\alpha_2}(\varepsilon) \| T_n\right) \geq (\alpha_2 - \alpha_1) - \frac{2\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_p}}{\varepsilon^{1/p}} \ \text{a.s.}$$

To put it in other words, convergence of the $\mathcal{L}_p$-norm is sufficient for the creation of asymptotically valid prediction intervals as it implies the convergence of $\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n)$. This offers another, but different, approach to deal with prediction intervals based on the Jackknife. For the sake of simplicity, we present the following results for symmetric predictors only:

Proposition 6.2. *Let Assumption **CC1** be fulfilled, $\hat{y}_0$ be a symmetric predictor and assume the first moments of $y_0 - \hat{y}_0$ and $y_0 - \tilde{y}_0$ exist. Then, for any $M > 0$ we have*

$$\mathbb{E}(\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_1}) \leq 2M \left(\frac{1}{4(n-1)} + 5\,\mathbb{E}(\min(1, \|f_{y_0\|x_0}\|_\infty |\hat{y}_0 - \tilde{y}_0|)) \right)^{\frac{1}{2}} + \mathbb{E}(|y_0 - \hat{y}_0| \mathbb{1}_{[M,\infty)}(|y_0 - \hat{y}_0|)) + \mathbb{E}(|y_0 - \tilde{y}_0| \mathbb{1}_{[M,\infty)}(|y_0 - \tilde{y}_0|)).$$

As $\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_\infty} \leq 1$, we can use the foregoing proposition to trivially extend the asymptotic consequences to the $\mathcal{L}_p$-norm for $p \geq 1$.

Proposition 6.3. *Let Assumption **CC1** be fulfilled, $\hat{y}_0$ be a symmetric predictor and assume $\|f_{y_0\|x_0}\|_\infty$ is bounded in probability. Furthermore, assume $y_0 - \hat{y}_0$ and $y_0 - \tilde{y}_0$ are uniformly integrable and $\hat{y}_0 - \tilde{y}_0$ converges to 0 in probability. Then, for any $p \geq 1$ we have*

$$\lim_{n\to\infty} \mathbb{E}(\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_p}) = 0.$$

The drawback of using $\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_p}$ instead of $\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n)$ is that we need stronger assumptions in order to ensure $\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_p}$ to be small: Here we need the existence of the first moment of the prediction error. This might seem to be only a small difference; However, this would require for our examples in Chapter 5 to control the expectation of the *inverse* of the smallest eigenvalue of a random matrix, for which there are less results available in the literature.

To sum it up, we can control $\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_p}$ if we replace boundedness in probability by the stronger assumption of uniform integrability. Moreover, since in the case of real-valued random variables the $\mathcal{L}_1$ distance between two distribution functions coincides with the Wasserstein distance $\mathcal{W}_1$ (cf. Panaretos and Zemel 2019), Proposition 6.2 and Proposition 6.3 even yield statements concerning the Wasserstein distance.

Lastly, one could propose the following scale-adapting divergence $\delta(\mathrm{F}_n, \widehat{\mathrm{F}}_n) := \inf\{\varepsilon \geq 0 : \ell_\varepsilon(\mathrm{F}_n, \widehat{\mathrm{F}}_n) = 0\}$ instead of ℓ_ε, where we define the infimum of the empty set as $+\infty$.[1] Combining equation (3.5) of Lemma 3.2 with the fact that $\delta(\mathrm{F}_n, \widehat{\mathrm{F}}_n)$ is a measurable function of the training data T_n this yields

$$\mathbb{P}\left(y_0 \in \hat{y}_0 + [\hat{q}_{\alpha_1} - \delta(\mathrm{F}_n, \widehat{\mathrm{F}}_n), \hat{q}_{\alpha_2} + \delta(\mathrm{F}_n, \widehat{\mathrm{F}}_n)] \| T_n\right) \geq \alpha_2 - \alpha_1 \text{ a.s.}$$

for every $1 \geq \alpha_2 \geq \alpha_1 \geq 0$. However, $\delta(\mathrm{F}_n, \widehat{\mathrm{F}}_n)$ comes with a price we do not want to pay: assume that F_n is a distribution function with $\mathrm{F}_n(x) < 1$ for all x almost surely and we estimate it via an empirical distribution function $\widehat{\mathrm{F}}_n$ based on n observations. Then, for all x large enough, $\widehat{\mathrm{F}}_n(x)$ equals 1 implying $\delta(\mathrm{F}_n, \widehat{\mathrm{F}}_n) = \infty$ even for good choices of $\widehat{\mathrm{F}}_n$ in the sense that $\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_\infty$ is small. Moreover, in that case the prediction interval

[1] By Lemma 3.4 the function $\varepsilon \mapsto \ell_\varepsilon(\mathrm{F}_n, \widehat{\mathrm{F}}_n)$ is continuous from the right. Thus, the infimum in the definition of $\delta(\mathrm{F}_n, \widehat{\mathrm{F}}_n)$ can be replaced by a minimum whenever the set $\{\varepsilon \geq 0 : \ell_\varepsilon(\mathrm{F}_n, \widehat{\mathrm{F}}_n) = 0\}$ is not empty. In particular, this yields $\ell_{\delta(\mathrm{F}_n, \widehat{\mathrm{F}}_n)}(\mathrm{F}_n, \widehat{\mathrm{F}}_n) = 0$ whenever $\delta(\mathrm{F}_n, \widehat{\mathrm{F}}_n)$ is finite.

$\hat{y}_0 + [\hat{q}_{\alpha_1} - \delta(\mathrm{F}_n, \widehat{\mathrm{F}}_n), \hat{q}_{\alpha_2} + \delta(\mathrm{F}_n, \widehat{\mathrm{F}}_n)]$ coincides with $\mathbb{R}$ almost surely, which makes it impractical for usage. To put it in other words, $\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_\infty$ measures the distance on the range of the functions, while $\delta(\mathrm{F}_n, \widehat{\mathrm{F}}_n)$ measures the distance on the domain. Using $\ell_\varepsilon(\mathrm{F}_n, \widehat{\mathrm{F}}_n)$ gives a divergence which takes variations both on the range and on the domain into account while allowing the user to choose how the both scales should be weighted.

6.2. On the assumptions of Theorem 4.5

As we will see our Jackknife-approach may fail if the corresponding prediction algorithm is not stable. Now one could ask whether the statements of Theorem 4.5 and Corollary 4.6 also hold true without the assumption on the boundedness in probability of the scaled prediction error. As it turns out, it can be replaced by other assumptions. However, we consider the new assumptions to be harder to fulfill.

Lemma 6.4. *Let $\hat{y}_0$ be a symmetric predictor. If the first moment of $\hat{y}_0 - \tilde{y}_0$ and of $|\hat{u}_1 - \hat{u}_2|$ exists, we have*

$$\mathbb{E}\left(\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_2}^2\right) \leq 5\,\mathbb{E}(|\hat{y}_0 - \tilde{y}_0|) + \frac{1}{2n}\,\mathbb{E}(|\hat{u}_1 - \hat{u}_2|).$$

Lemma 6.1 shows that whenever $\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_p}$ converges to 0, the same holds true for $\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n)$, which implies the following corollary based on Lemma 6.4.

Corollary 6.5. *Let $\hat{y}_0$ be a symmetric predictor. If $\lim_{n\to\infty} \mathbb{E}(|\hat{y}_0 - \tilde{y}_0|) = 0$ and $\lim_{n\to\infty} \mathbb{E}(|\hat{u}_1 - \hat{u}_2|)/n = 0$, then for any $p \in [2, \infty)$ we have*

$$\lim_{n\to\infty} \mathbb{E}\left(\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_p}\right) = 0.$$

Moreover, there exists a null-sequence ε_n, such that

$$\lim_{n\to\infty} \mathbb{E}\left(\ell_{\varepsilon_n}(\widehat{\mathrm{F}}_n, \mathrm{F}_n)\right) = 0.$$

*If, additionally, Assumption **CC1** is fulfilled and $\|f_{y_0\|x_0}\|_\infty$ is bounded in probability, we have*

$$\lim_{n\to\infty} \mathbb{E}\left(\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_\infty\right) = 0.$$

Corollary 6.5 indeed shows that we do not need the scaled prediction error to be bounded in probability. However, the stability assumption is replaced by a stronger notion (convergence in $\mathcal{L}_1$ instead of convergence in probability) and the expected difference between two different leave-one-out residuals has to grow slower than with rate n. However, the stronger notion of stability can indeed pose a problem for the predictors considered in Chapter 5 as in that case one needs to control the inverse of the smallest eigenvalue of $\Sigma^{-1/2}X'X\Sigma^{-1/2}/n$ in expectation rather than in probability.

Nevertheless, Corollary 6.5 gives another set of conditions where the Jackknife-approach provides asymptotically valid prediction intervals.

While it may be intuitive that a Jackknife-approach crucially relies on the stability of the predictors, we present an example showing that the lack of stability cannot be fixed by a simple enlargement of the prediction intervals:

Lemma 6.6. *For any $q \in [0,1]$, $n \geq 2$, $0 \leq \alpha_1 \leq \alpha_2 \leq 1$ and any $\varepsilon > 0$ there exists a predictor $\hat{y}_0$ and a distribution $\mathcal{P}$ of (y_0, x_0'), such that the conditional coverage probability of the prediction interval $PI^+_{\alpha_1,\alpha_2}(\varepsilon)$ is less or equal to q almost surely. Moreover, the prediction error can be bounded by a constant C_ε independent of n almost surely and Assumption **CC1** is fulfilled.*

Lemma 6.6 shows that indeed the stability of the prediction algorithm is a crucial requirement as all the other assumptions are fulfilled and the Jackknife-approach fails nevertheless.

6.3. On exact conditional coverage probability

Corollary 4.6 shows that in the continuous case – under some assumptions – the actual coverage probability of the prediction interval derived from a Jackknife-approach converges in probability to its nominal level. Loosely speaking this means that for a large amount of training data the conditional coverage probability of the prediction interval is close to its nominal level if the number of observations is large. Furthermore, this statement is valid for a large class of distributions as long as the assumptions of Corollary 4.6 are fulfilled. Now one can ask whether it is possible to create a prediction interval whose conditional coverage probability equals its nominal level in finite samples for all training data. To answer this question we present the following results:

Lemma 6.7. *Fix a learning algorithm $\mathcal{A}_{p,n}$ and training data T_n, such that the class $\mathcal{F}(T_n)$ of all possible distributions of $y_0 - \hat{y}_0$ conditional on T_n contains (at least) two distribution functions $F_1 \neq F_2$ fulfilling the following property:*

- *There is a $\lambda > 0$, such that $F_1(t) = F_2(\lambda t)$ for all $t \in \mathbb{R}$.*
- *F_1 is a strictly increasing function from $\mathbb{R}$ to $(0,1)$.*

Furthermore, assume that $0 < \alpha < \min(F_1(0), 1 - F_1(0))$. Then, it is not possible to find a (non-randomized) prediction interval of the form $\hat{y}_0 + (\hat{L}, \hat{U}]$ based on the training data T_n, such that the conditional coverage probability for y_0 of the prediction interval given the training data T_n equals $1 - \alpha$ under both distributions F_1 and F_2.

As Lemma 6.7 shows, asking for exact conditional coverage probability in finite samples for all training data is too demanding if this should hold true for two distributions F_1 and F_2, fulfilling the following two properties: Firstly, the distributions should coincide up to a scaling (for example if the distributions only differ in their standard deviation). Secondly, the distributions should be strictly increasing, implying that the set of all possible values of $y_0 - \hat{y}_0$ conditional on T_n is unbounded.

However, the result above hinges on the fact that we are only considering non-randomized prediction intervals. To see why we cannot remove this assumption one could consider the following prediction interval PI^{rand}: Let $PI^{rand} = \mathbb{R}$ with probability $1 - \alpha$ and $PI^{rand} = \emptyset$ with probability α. Now, this trivial but in practice useless prediction interval provides an exact conditional coverage probability of $1 - \alpha$ for every (real-valued) random variable.

6.4. Asymptotically valid prediction intervals in the non-continuous case

While the results of Corollary 4.6 provide a prediction interval whose actual conditional coverage probability converges to the nominal level of $1 - \alpha$ under assumption **CC1**, we do not get a similar result for the non-continuous case as Corollary 4.6 only provides lower bounds for the actual conditional coverage probability in the non-continuous case. To solve this problem theoretically, we could use a randomized prediction interval, as the following lemma shows.

Lemma 6.8. *Let $\alpha \in [0, 1], \varepsilon \geq 0$ and PI^r be the following randomized prediction interval based on the leave-one-out residuals $(\hat{u}_i)_{i=1}^n$: With probability $(1-\alpha)$ the prediction interval coincides with $\hat{y}_0 + [\min_{1 \leq i \leq n}(\hat{u}_i) - \varepsilon, \max_{1 \leq i \leq n}(\hat{u}_i) + \varepsilon]$ and with probability α it coincides with the empty set. We then have the following result for the actual conditional coverage probability:*

$$|\mathbb{P}\left(y_0 \in PI^r \| T_n\right) - (1 - \alpha)| \leq 2\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n) \text{ a.s.}$$

Thus, Lemma 6.8 yields a comparable result to the statement of Proposition 4.1 without using the continuity of y_0 given x_0. However, the practicability of this approach is debatable.

6.5. On the notion of stability

As mentioned before, the definition of asymptotic stability reduces to the condition of $\hat{y}_0 - \tilde{y}_0^{[-n]} \xrightarrow{p} 0$ in the case of a symmetric predictor. Moreover, we would like to point out that our notion of stability requires the predictor to be stable on average (in large samples) rather than dealing with a worst-case scenario like the uniform stability. Hence, the *no-free-lunch* theorem for sparse algorithms of Xu et al. (2012) does not apply here. Our notion of stability is closely related to the *hypothesis stability* defined in Bousquet and Elisseeff (2002) and in spirit of the original definition in Kearns and Ron (1999), the latter of which is stated in terms of probability rather than expectation. For a comparison of different notions of stability we refer to the paper Bousquet and Elisseeff (2002) and the references therein.

6.6. The Weak Tail Projection property

Assumption **LD** is used to guarantee that the smallest eigenvalue of $(\Sigma^{-1/2}X'X\Sigma^{-1/2})/n$ is bounded away from 0 asymptotically. This is done by applying the results of Chafaï and Tikhomirov (2018) which use the so-called *Weak Tail Projection* property introduced therein. For the sake of understanding, we state it here:[2]

Definition 6.9 (Weak Tail Projection property (WTP) of Chafaï and Tikhomirov 2018)**.** Let $(v_p)_{p\in\mathbb{N}}$ be a sequence of random vectors, where for each $p \in \mathbb{N}$ the random vector v_p takes values in $\mathbb{R}^p$ and is centered with unit covariance (isotropy). We say that the *WTP* property holds when the following is true:

1. The family $((v_p'y)^2)_{p\in\mathbb{N},y\in S^{p-1}}$ is uniformly integrable, in other words

$$\lim_{M\to\infty} \sup_{p\in\mathbb{N},y\in S^{p-1}} \mathbb{E}\left((v_p'y)^2 \mathbb{1}_{\{(v_p'y)^2\geq M\}}\right) = 0, \tag{WTP-a}$$

   where $S^{p-1} = \{y \in \mathbb{R}^p : \|y\|_2 = 1\}$ denotes the unit sphere of $\mathbb{R}^p$.

2. There exist two functions $f : \mathbb{N} \to [0, 1]$ and $g : \mathbb{N} \to \mathbb{R}_+$ such that $f(r) \to 0$ and $g(r) \to 0$ as $r \to \infty$ and for every $p \in \mathbb{N}$ and any orthogonal projection $P : \mathbb{R}^p \to \mathbb{R}^p$ with $P \neq 0$,

$$\mathbb{P}\left(\frac{\|Pv_p\|_2^2}{\operatorname{rank}(P)} - 1 \geq f(\operatorname{rank}(P))\right) \leq g(\operatorname{rank}(P)). \tag{WTP-b}$$

To show the importance of Assumption **LD** we present an example, where the *Weak Tail Projection* property is not fulfilled and the smallest eigenvalue of $(\Sigma^{-1/2}X'X\Sigma^{-1/2})/n$ equals 0 with probability 1 asymptotically: For each $1 \leq i \leq n$ we define a_i to be 0 with probability α and $\pm(1-\alpha)^{-1/2}$ with probability $(1-\alpha)/2$ each, where $\alpha \in (0, 1)$ will be defined later. Furthermore, for $1 \leq j \leq p$ let $b_{i,j}$ be independent Rademacher random variables being independent from $(a_k)_{k=1}^n$. Now we define $x_{i,j} = a_i b_{i,j}$ for all $1 \leq i \leq n$ and $1 \leq j \leq p$. Then, the $(x_{i,j})_{1\leq i\leq n,1\leq j\leq p}$ have mean-zero, unit variance and the vectors $x_i = (x_{i,1}, \ldots, x_{i,p})', 1 \leq i \leq n$ are isotropic and independent of each other. As the length $\|x_i\|_2$ of x_i is either 0 or $\sqrt{p/(1-\alpha)}$, the Weak Tail Projection property cannot be fulfilled, as it implies $\|x_i\|_2$ to concentrate within the centered ball of radius $\sqrt{p}$. However, for the smallest eigenvalue of $X'X/n$ we have

$$\begin{aligned}
\mathbb{P}\left(\lambda_{\min}(X'X/n) = 0\right) \geq& \mathbb{P}\left(\operatorname{rank}(X'X) < p\right) \geq \mathbb{P}\left(\operatorname{rank}(\sum_{i=1}^n x_i x_i') < p\right) \\
\geq& \mathbb{P}\left(|\{i : x_i x_i' = 0_{d\times d}\}| > n - p\right) \\
=& \mathbb{P}\left(|\{i : a_i = 0\}| > n - p\right).
\end{aligned}$$

[2] In order to avoid confusion, we have adapted the notation slightly.

However, $|\{i : a_i = 0\}|$ is a binomial distributed random variable with parameters n and α. If $\rho > 0$, we can choose α such that $1 > \alpha > 1 - \rho$ holds true, which yields $\lim_{n\to\infty} \mathbb{P}\left(|\{i : a_i = 0\}| > n - p\right) = 1$.

6.7. On the convergence of p/n

In Chapter 5 we assumed the existence of the limit of p/n. However, all our proofs are still valid if we replace the requirement $\rho = \lim_{n\to\infty} p/n = [0, 1)$ in Assumption **LD** by $\limsup_{n\to\infty} p/n < 1$ as we only need the assumption on ρ to ensure that the smallest eigenvalue of $\Sigma^{-1/2} X'X \Sigma^{-1/2}/n$ stays away from 0 asymptotically. Similarly, in Assumption **HD** it is enough to assume $\liminf_{n\to\infty} p/n > 1$. This can be seen by a similar argument as in the proof of the case $\rho = 0$ of Lemma 5.3, which crucially relies on the fact that the limit of the smallest non-zero singular value is decreasing in ρ for $\rho < 1$ and increasing for $\rho > 1$.

7. Conclusion

The present thesis deals with the construction of prediction intervals in a setting where the number of regressors is not negligible compared to the number of observations. Our Jackknife-approach is based on the *leave-one-out* residuals and therefore crucially relies on the stability of the predictor with respect to the exclusion of one data point. We derive bounds for the difference between the actual and the nominal coverage probability conditionally on the training data in finite samples which converge to 0 asymptotically. From these results we conclude that our approach will give (asymptotically) valid prediction intervals if the following two conditions are met: Firstly, the prediction error should be bounded in probability and secondly the average influence of one single data point on the prediction should vanish if the number of observations tends to infinity. This generalizes existing results of Steinberger and Leeb (2023) to the non-continuous case.

Our results are stated in the general and in the continuous case, where for the latter we additionally assume the response y_0 conditionally on the regressor x_0 to be an absolutely continuous random variable. In the continuous case we measure the accuracy of the estimation of the prediction error's distribution in terms of the Kolmogorov distance. In the general (non-continuous) case we show that this is not a good choice and present another measurement $\ell_\varepsilon(\cdot,\cdot)$ for the estimation's accuracy, which is related to the Lévy metric and is, in fact, a generalization of the Kolmogorov distance. However, in the non-continuous case we have to enlarge our prediction intervals by the (small) amount of ε to account for possible discontinuities, where ε can be made arbitrarily small asymptotically.

Furthermore, we show that in a broad setting the OLS, the Ridge, the James-Stein estimator and the minimum-norm interpolator fulfill the requirements that guarantee asymptotically valid prediction intervals based on our Jackknife-approach. Moreover, we show that the regularization of the Ridge ensures its stability under very mild assumptions regardless of the number of regressors as long as the regularization does not vanish asymptotically. Additionally, we present an example in the case of binary classification where the predictor based on the support vector classifier is also stable.

To avoid the high computational costs of the classical Jackknife, one could consider a k-fold cross-validation approach based on the Jackknife as it is done in Steinberger and Leeb (2023) in the continuous case. The extension of their k-fold cross validation results to the non-continuous case is the subject of ongoing research.

References

Bai, Z. D. and Y. Q. Yin (1993). "Limit of the Smallest Eigenvalue of a Large Dimensional Sample Covariance Matrix". *The Annals of Probability* 21.3, pp. 1275–1294.

Baranchik, A. J. (1973). "Inadmissibility of maximum likelihood estimators in some multiple regression problems with three or more independent variables". *The Annals of Statistics* 1.2, pp. 312–321.

Barber, R. F., E. J. Candes, A. Ramdas, and R. J. Tibshirani (2021a). "Predictive inference with the jackknife+". *The Annals of Statistics* 49.1, pp. 486–507.

— (2021b). "The limits of distribution-free conditional predictive inference". *Information and Inference: A Journal of the IMA* 10.2, pp. 455–482.

Bian, M. and R. F. Barber (2023). "Training-conditional coverage for distribution-free predictive inference". *arXiv preprint arXiv:2205.03647v2*.

Bickel, P. J. and D. A. Freedman (1983). "Bootstrapping regression models with many parameters". In: *A Festschrift for Erich L. Lehmann*. Ed. by P. J. Bickel, K. A. Dokum, and J. L. Hodges. Wadsworth Inc., pp. 28–48.

Billingsley, P. (1995). *Probability and Measure*. Third edition. A Wiley-Interscience publication. New York: John Wiley & Sons, Inc.

Bousquet, O. and A. Elisseeff (2002). "Stability and Generalization". *Journal of Machine Learning Research* 2, pp. 499–526.

Burges, C. J. C. (1998). "A tutorial on support vector machines for pattern recognition". *Data mining and knowledge discovery* 2.2, pp. 121–167.

Burkholder, D. L. (1966). "Martingale transforms". *The Annals of Mathematical Statistics* 37.6, pp. 1494–1504.

Butler, R. and E. D. Rothman (1980). "Predictive intervals based on reuse of the sample". *Journal of the American Statistical Association* 75.372, pp. 881–889.

Chafaï, D. and K. Tikhomirov (2018). "On the convergence of the extremal eigenvalues of empirical covariance matrices with dependence". *Probability Theory and Related Fields* 170.3-4, pp. 847–889.

El Karoui, N. (2018). "On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators". *Probability Theory and Related Fields* 170, pp. 95–175.

El Karoui, N. and E. Purdom (2018). "Can we trust the bootstrap in high-dimensions? The case of linear models". *The Journal of Machine Learning Research* 19.1, pp. 170–235.

Hastie, T., A. Montanari, S. Rosset, and R. J. Tibshirani (2022). "Surprises in high-dimensional ridgeless least squares interpolation". *The Annals of Statistics* 50.2, pp. 949–986.

Hastie, T., R. Tibshirani, J. H. Friedman, and J. H. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Second edition. New York: Springer.

Higham, N. J. and S. H. Cheng (1998). "Modifying the inertia of matrices arising in optimization". *Linear Algebra and its Applications* 275, pp. 261–279.

Huber, N. and H. Leeb (2013). "Shrinkage estimators for prediction out-of-sample: Conditional performance". *Communications in Statistics-Theory and Methods* 42.7, pp. 1246–1264.

James, W. and C. Stein (1961). "Estimation with quadratic loss". *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* 1, pp. 361–379.

Kearns, M. and D. Ron (1999). "Algorithmic stability and sanity-check bounds for leave-one-out crossvaildation". *Neural Computation* 11.6, pp. 1427–1453.

Kuipers, L. and H. Niederreiter (1974). *Uniform distribution of sequences*. A Wiley-Interscience publication. New York: John Wiley & Sons, Inc.

Lei, J. and L. A. Wasserman (2014). "Distribution-free prediction bands for non-parametric regression". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76, pp. 71–96.

Mammen, E. (1996). "Empirical process of residuals for high-dimensional linear models". *The Annals of Statistics* 24.1, pp. 307–335.

Panaretos, V. M. and Y. Zemel (2019). "Statistical Aspects of Wasserstein Distances". *Annual Review of Statistics and Its Application* 6, pp. 405–431.

Penrose, R. (1955). "A generalized inverse for matrices". *Mathematical Proceedings of the Cambridge Philosophical Society* 51.3, pp. 406–413.

Quenouille, M. H. (1956). "Notes on bias in estimation". *Biometrika* 43.3/4, pp. 353–360.

Saumard, A. and J. A. Wellner (2014). "Log-concavity and strong log-concavity: a review". *Statistics Surveys* 8, pp. 45–114.

Srivastava, N. and R. Vershynin (2013). "Covariance estimation for distributions with $2 + \varepsilon$ moments". *The Annals of Probability* 41.5, pp. 3081–3111.

Steinberger, L. and H. Leeb (2023). "Conditional predictive inference for stable algorithms". *The Annals of Statistics* 51.1, pp. 290–311.

Vapnik, V. N. (1999). *The Nature of Statistical Learning Theory*. Second edition. New York: Springer.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vovk, V. (2012). "Conditional validity of inductive conformal predictors". In: *Asian conference on machine learning*. PMLR, pp. 475–490.

Wu, W. B. and X. Shao (2007). "A Limit Theorem for Quadratic Forms and Its Applications". *Econometric Theory* 23.5, pp. 930–951.

Xu, H., C. Caramanis, and S. Mannor (2012). "Sparse Algorithms Are Not Stable: A No-Free-Lunch Theorem". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.1, pp. 187–193.

A. Appendix

In our proofs we will use several well-known properties of the Moore-Penrose pseudoinverse, which we collect in the following lemma:

Lemma A.1. *Let $A \in \mathbb{R}^{a\times b}$ and let $A^\dagger$ denote its Moore-Penrose pseudoinverse. Write $P_A \in \mathbb{R}^{a\times a}$ for the orthogonal projection onto the column space of A. We then have:*

1) *Let (one version of) the singular value decomposition of A be given by $A = USV'$, where U and V are orthogonal matrices and $S \in \mathbb{R}^{a\times b}$ is a matrix whose diagonal entries coincide with the singular values of A. Then, a singular value decomposition of $A^\dagger$ is given by $VS^\dagger U'$.*

2) *$A^\dagger$ can equivalently be written as $(A'A)^\dagger A'$ or $A'(AA')^\dagger$.*

3) *$P_A = AA^\dagger = A(A'A)^\dagger A' = (AA')^\dagger AA' = (A')^\dagger A'$*

4) *$P_A A = A$, $A' = A'P_A$ and $A^\dagger P_A = A^\dagger$.*

5) *If A has full column rank b, then $A^\dagger = (A'A)^{-1}A'$.*

6) *The Moore-Penrose pseudoinverse of an orthogonal projection P_A is P_A itself.*

Proof. 1) is a special case of Lemma 1.6 of Penrose (1955), 2) a direct combination of Lemma 1.5 of Penrose (1955) and equation (10) therein.

The equations $AA^\dagger = A(A'A)^\dagger A'$ and $(AA')^\dagger AA' = (A')^\dagger A'$ are a consequence of 2). By Lemma 2.1 of Penrose (1955) $AA^\dagger$ is symmetric and idempotent, hence an orthogonal projection matrix. By symmetry we also have $AA^\dagger = (AA^\dagger)' = (A')^\dagger A'$. It remains to show that $AA^\dagger$ projects onto the column space of A. For this, we show that $AA^\dagger v = 0 \in \mathbb{R}^a$ if and only if $v \in \ker(A')$: If $A'v = 0$, then $AA^\dagger v = A(A'A)^\dagger A'v = 0$. If, conversely, $v \notin \ker(A')$ we have $0 \neq A'v = A'(A')^\dagger A'v = A'AA^\dagger v$, which implies that $AA^\dagger v \neq 0$. Hence, $AA^\dagger$ is the orthogonal projection onto the column space of A.

4) follows from $AA^\dagger A = A$ and $A^\dagger AA^\dagger = A^\dagger$ combined with 3) and the fact that $(P_A A)' = A'P_A$.

If A has full column rank b, then $A'A$ is invertible. Thus, by Lemma 1.3 of Penrose (1955) we have $(A'A)^\dagger = (A'A)^{-1}$. Together with 2) this proves 5).

Recalling the fact that a matrix $P \in \mathbb{R}^{a\times a}$ is an orthogonal projection if and only if it is symmetric and idempotent, 6) is given by Lemma 2.2 of Penrose (1955).

□

A.1. Additional results

This subsection provides some general results used in this work. We start with two purely algebraic lemmas:

Lemma A.2. *Fix a vector $Y = (Y_1, \dots, Y_n)' \in \mathbb{R}^n$, any matrix $X \in \mathbb{R}^{n \times p}$ and a constant $c \geq 0$. Define $\tilde{Y} = (Y_1, \dots, Y_{n-1})' \in \mathbb{R}^{n-1}$, $\tilde{X} \in \mathbb{R}^{(n-1)\times p}$ the matrix X without the last row and $\bar{u}_n = e_n'(I_n - P_X)Y = y_n - x_n' X^\dagger Y$. Then the following equation holds true:*

$$X^\dagger Y - \tilde{X}^\dagger \tilde{Y} = X^\dagger e_n (y_n - x_n' \tilde{X}^\dagger \tilde{Y}). \tag{A.1}$$

If $\tilde{X}$ has full column rank p, we also have the identity

$$X^\dagger Y - \tilde{X}^\dagger \tilde{Y} = (\tilde{X}'\tilde{X})^{-1} x_n \bar{u}_n. \tag{A.2}$$

Define $A_c := X'X + cI_p$, $\widehat{\beta}_R(c) = A_c^\dagger X'Y$ and $\widetilde{\beta}_R(c) = (\tilde{X}'\tilde{X} + cI_p)^\dagger \tilde{X}'\tilde{Y}$. The following formula holds true even if $\tilde{X}$ does not possess full rank p:

$$\widehat{\beta}_R(c) - \widetilde{\beta}_R(c) = A_c^\dagger x_n (y_n - x_n' \widetilde{\beta}_R(c)). \tag{A.3}$$

Proof. The row space of $\tilde{X}$ is a subspace of the row space of X, which implies

$$P_{X'} \tilde{X}^\dagger = P_{X'} \tilde{X}' (\tilde{X}\tilde{X}')^\dagger = \tilde{X}' (\tilde{X}\tilde{X}')^\dagger = \tilde{X}^\dagger.$$

Using the fact that $X'X = \tilde{X}'\tilde{X} + x_n x_n'$ together with $X'Y = \tilde{X}'\tilde{Y} + x_n y_n$ yields

$$\begin{aligned} X^\dagger Y = (X'X)^\dagger X'Y &= (X'X)^\dagger \left(\tilde{X}'\tilde{Y} + x_n y_n \right) = (X'X)^\dagger \left(\tilde{X}'\tilde{X}\tilde{X}^\dagger \tilde{Y} + x_n y_n \right) \\ &= (X'X)^\dagger \left(X'X\tilde{X}^\dagger \tilde{Y} + x_n (y_n - x_n' \tilde{X}^\dagger \tilde{Y}) \right). \end{aligned} \tag{A.4}$$

Using the fact that $x_n = X'e_n$ and the property $P_{X'} = (X'X)^\dagger X'X$, the last line of equation (A.4) can be equivalently written as

$$\begin{aligned} &P_{X'} \tilde{X}^\dagger \tilde{Y} + (X'X)^\dagger X' e_n (y_n - x_n' \tilde{X}^\dagger \tilde{Y}) \\ &\qquad = \tilde{X}^\dagger \tilde{Y} + X^\dagger e_n (y_n - x_n' \tilde{X}^\dagger \tilde{Y}). \end{aligned}$$

We prove the second statement with a similar argument:

$$\begin{aligned} \tilde{X}'\tilde{X} X^\dagger Y &= X'X X^\dagger Y - x_n x_n' X^\dagger Y = X'Y - x_n x_n' X^\dagger Y \\ &= \tilde{X}'\tilde{Y} + x_n (y_n - x_n' X^\dagger Y) = \tilde{X}'\tilde{X}\tilde{X}^\dagger \tilde{Y} + x_n \bar{u}_n. \end{aligned}$$

As $\tilde{X}'\tilde{X}$ is invertible, the claim follows.

We now prove the third statement. The case $c = 0$ is already treated by the first statement as $\widehat{\beta}_R(0) = X^\dagger Y$. Thus, it suffices to show the third statement for the case

$c > 0$. Hence, we can assume the regularity of A_c and of $\tilde{X}'\tilde{X} + cI_p$ to get

$$\begin{aligned} A_c\left(\widehat{\beta}_R(c) - \widetilde{\beta}_R(c)\right) &= X'Y - (\tilde{X}'\tilde{X} + cI_p)\widetilde{\beta}_R(c) - x_n x_n'\widetilde{\beta}_R(c) \\ &= X'Y - \tilde{X}'\tilde{Y} - x_n x_n'\widetilde{\beta}_R(c) = x_n(y_n - x_n'\widetilde{\beta}_R(c)). \end{aligned}$$

□

Lemma A.3. *Fix a vector $Y = (Y_1, \ldots, Y_n)' \in \mathbb{R}^n$ and any matrix $X \in \mathbb{R}^{n\times p}$. Define $\tilde{Y} = (Y_1, \ldots, Y_{n-1})' \in \mathbb{R}^{n-1}$, $\tilde{X} \in \mathbb{R}^{(n-1)\times p}$ the matrix X without the last row and $\bar{u}_n = e_n'(I_n - P_X)Y = y_n - x_n'\widehat{\beta}_{LS}$. If $\tilde{X}$ has full column rank p, we have*

$$Y'P_X Y - \tilde{Y}'P_{\tilde{X}}\tilde{Y} = \bar{u}_n^2 x_n'(\tilde{X}'\tilde{X})^{-1}x_n + (e_n'P_X Y)^2 + 2\bar{u}_n x_n'\tilde{X}^\dagger\tilde{Y}.$$

Furthermore, we have

$$Y'(I_n - P_X)Y - \tilde{Y}'(I_{n-1} - P_{\tilde{X}})\tilde{Y} = \bar{u}_n^2\left(x_n'(\tilde{X}'\tilde{X})^{-1}x_n + 1\right).$$

Proof. As $P_X = XX^\dagger$ and $X'X = \tilde{X}'\tilde{X} + x_n x_n'$, we can rewrite $Y'P_X Y$ as follows:

$$Y'P_X Y = \|XX^\dagger Y\|_2^2 = \|\tilde{X}X^\dagger Y\|_2^2 + (x_n'X^\dagger Y)^2.$$

Using the rank-one update formula from Lemma A.2 this equals

$$\|\tilde{X}\tilde{X}^\dagger\tilde{Y} + \tilde{X}(\tilde{X}'\tilde{X})^{-1}x_n\bar{u}_n\|_2^2 + (e_n'XX^\dagger Y)^2,$$

which can be reformulated as

$$\begin{aligned} &\|P_{\tilde{X}}\tilde{Y} + (\tilde{X}^\dagger)'x_n\bar{u}_n\|_2^2 + (e_n'P_X Y)^2 \\ &= \tilde{Y}'P_{\tilde{X}}\tilde{Y} + \bar{u}_n^2 x_n'(\tilde{X}'\tilde{X})^{-1}x_n + 2\bar{u}_n x_n'\tilde{X}^\dagger\tilde{Y} + (e_n'P_X Y)^2, \end{aligned}$$

which proves the first part. For the second part we start with the observation

$$\tilde{Y}'(I_{n-1} - P_{\tilde{X}})\tilde{Y} - Y'(I_n - P_X)Y = Y'P_X Y - \tilde{Y}'P_{\tilde{X}}\tilde{Y} - y_n^2,$$

which equals

$$\bar{u}_n^2 x_n'(\tilde{X}'\tilde{X})^{-1}x_n + 2\bar{u}_n x_n'\tilde{X}^\dagger\tilde{Y} + (e_n'P_X Y)^2 - y_n^2.$$

Using the equality

$$y_n^2 = (e_n'Y)^2 = (\bar{u}_n + e_n'P_X Y)^2 = \bar{u}_n^2 + (e_n'P_X Y)^2 + 2\bar{u}_n e_n'P_X Y$$

yields

$$\begin{aligned}
&\tilde{Y}'(I_{n-1} - P_{\tilde{X}})\tilde{Y} - Y'(I_n - P_X)Y \\
&= \bar{u}_n^2 x_n'(\tilde{X}'\tilde{X})^{-1}x_n + 2\bar{u}_n x_n'\tilde{X}^\dagger\tilde{Y} - \bar{u}_n^2 - 2\bar{u}_n e_n' P_X Y \\
&= \bar{u}_n^2\left(x_n'(\tilde{X}'\tilde{X})^{-1}x_n - 1\right) + 2\bar{u}_n x_n'(\tilde{X}^\dagger\tilde{Y} - X^\dagger Y).
\end{aligned}$$

Inserting the rank-one update formula from Lemma A.2 allows us to rewrite the last line as

$$\begin{aligned}
&\bar{u}_n^2\left(x_n'(\tilde{X}'\tilde{X})^{-1}x_n - 1\right) - 2\bar{u}_n^2 x_n'(\tilde{X}'\tilde{X})^{-1}x_n \\
&= -\bar{u}_n^2\left(x_n'(\tilde{X}'\tilde{X})^{-1}x_n + 1\right).
\end{aligned}$$

□

Lemma A.4. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of real-valued random variables. Then there exists a (non-stochastic) sequence $(c_n)_{n\in\mathbb{N}}$ with $c_n \in (0,\infty)$ for every n, such that the sequence $(Y_n)_{n\in\mathbb{N}}$, defined as $Y_n := X_n/c_n$ for $n \in \mathbb{N}$, is bounded in probability.*

Proof. If we additionally assume the existence of any moment, the statement is trivial by using Markov's inequality and a suitable choice of c_n. However, the statement remains true without any moment assumption as the following proof shows:

We choose $c_n := 1 + F_n^\dagger(1 - 1/n)$ and for an arbitrary $\varepsilon \in (0,1)$ we define $M_\varepsilon = 1 + \max_{1\le i\le 1/\varepsilon} F_i^\dagger(1-\varepsilon)$, where $F_i(x) := \mathbb{P}(|X_i| \le x)$ is the distribution function of the absolute value of X_i and $F_i^\dagger(p) = \inf\{x \ge 0 : F_i(x) \ge p\}$ the p-quantile of $|X_i|$. We then have

$$\sup_{n\in\mathbb{N}} \mathbb{P}(|X_n|/c_n \ge M_\varepsilon) \le \max\left(\max_{1\le i\le 1/\varepsilon} \mathbb{P}(|X_i| \ge M_\varepsilon), \sup_{i>1/\varepsilon} \mathbb{P}(|X_i| \ge c_i)\right) \le \varepsilon.$$

□

The next lemma is a generalization of the following statement: Let $(X_n)_{n\in\mathbb{N}}$ a sequence of real-valued random variables converging to 0 in probability. Then there exists a (non-stochastic) null sequence $(c_n)_{n\in\mathbb{N}}$ with $c_n \in (0,1)$ for every n, such that the sequence of random variables $(Y_n)_{n\in\mathbb{N}}$, defined as $Y_n := X_n/c_n$ for $n \in \mathbb{N}$, also converges to 0 in probability. This statement can be directly derived from Lemma A.5 by setting $S_n(\varepsilon) = \mathbb{P}(|X_n| \ge \varepsilon)$ for all $n \in \mathbb{N}$.

Lemma A.5. *Let $(S_n)_{n\in\mathbb{N}}$ be a sequence of non-increasing functions from $[0,\infty)$ to $[0,1]$. If for every fixed $\varepsilon > 0$ we have*

$$\lim_{n\to\infty} S_n(\varepsilon) = 0,$$

then there even exists a (non-stochastic) null-sequence $(\varepsilon_n)_{n\in\mathbb{N}}$ *with* $\varepsilon_n \in (0,1)$ *for every* $n \in \mathbb{N}$*, such that the following holds true:*

$$\lim_{n\to\infty} S_n(\varepsilon_n) = 0.$$

Proof. Define $p_n(k) = S_n(1/k)$. Now, fix an arbitrary $k \in \mathbb{N}$. As $\lim_{n\to\infty} p_n(k) = 0$, we can find an $\tilde{N}_k$, such that we have

$$p_n(k) \leq 1/k \text{ for all } n \geq \tilde{N}_k.$$

However, we will define N_k to be the smallest such $\tilde{N}_k$:

$$\begin{aligned} N_k &= \min\{l \in \mathbb{N} : p_n(k) \leq 1/k \text{ for all } n \geq l\} \\ &= 1 + \max\{n \in \mathbb{N} : p_n(k) > 1/k\}, \end{aligned}$$

where the equality in the second line of the previous display holds true whenever the set $\{n \in \mathbb{N} : p_n(k) > 1/k\}$ is not empty.[1] We would like to point out that the set in the second definition is a subset of $\{1, \ldots, \tilde{N}_k - 1\}$ and hence is finite. Moreover, note that N_1 equals 1 and N_k is non-decreasing in k. Now, by definition we have

$$p_n(k) \leq 1/k \text{ for all } n \geq N_k. \tag{A.5}$$

For any $n \in \mathbb{N}$ we now define the index $k(n)$ as

$$k(n) := \max\{k \in \{1, \ldots, n\} : N_k \leq n\}.$$

As $N_1 = 1 \leq n$, the set in the definition is non-empty and therefore the maximum is well defined. Moreover, $k(n)$ is non-decreasing in n, $N_{k(n)}$ is bounded by n and we have the property

$$N_{k(n)+1} > n \text{ or } k(n) = n.$$

Thus, the inequality $N_{k(n)} \leq n$ together with inequality (A.5) allows us to derive the following statement without restriction to n:

$$p_n(k(n)) \leq 1/k(n) \text{ for all } n \in \mathbb{N}.$$

If we can show that $\lim_{n\to\infty} k(n) = \infty$, we could define a null-sequence $\varepsilon_n = (k(n))^{-1}$ to get

$$0 \leq \lim_{n\to\infty} S_n(\varepsilon_n) = \lim_{n\to\infty} S_n\left(\frac{1}{k(n)}\right) = \lim_{n\to\infty} p_n(k(n)) \leq \lim_{n\to\infty} \frac{1}{k(n)} = 0.$$

Thus, it only remains to show that $\lim_{n\to\infty} k(n) = \infty$. We now assume this is not the case and try to find a contradiction. As by definition $k(n)$ is non-decreasing in n and

[1] However, the equality holds true if we define $\max\{l \in \mathbb{N} : l \in \emptyset\} := 0$.

$k(n) \in \mathbb{N}$, this would imply the existence of $M > 0$, such that $k(n) \leq M$ for all $n \in \mathbb{N}$. However, as stated before, we have for any $n \in \mathbb{N}$ the properties that either $N_{k(n)+1} > n$ or $k(n) = n$ holds true. For $n > M$ the second statement cannot be fulfilled. Moreover, by the monotonicity $N_{k(n)+1} \leq N_{M+1} \in \mathbb{N}$ is bounded over $n \in N$, while n is not.[2] Thus, we have $\lim_{n\to\infty} k(n) = \infty$. □

Lemma A.6. *Let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be sequences of p_n-dimensional random vectors, such that a_n is independent of b_n and $\mathbb{E}(a_n a_n') = \Sigma_{a_n}$ for all $n \in \mathbb{N}$. We then have*

- *If $(\|\Sigma_{a_n}^{1/2} b_n\|_2)_{n\in\mathbb{N}}$ is bounded in probability, then so is $(a_n' b_n)_{n\in\mathbb{N}}$.*
- *If $(\|\Sigma_{a_n}^{1/2} b_n\|_2)_{n\in\mathbb{N}}$ converges to 0 in probability, then so does $(a_n' b_n)_{n\in\mathbb{N}}$.*

Furthermore, suppose $(C_n)_{n\in\mathbb{N}}$ is a sequence of $(p_n \times p_n)$-dimensional random, nonnegative definite matrices, such that C_n is independent of a_n for every $n \in \mathbb{N}$. Then $(a_n' C_n a_n)_{n\in\mathbb{N}}$ is bounded in probability whenever $(\mathrm{tr}(\Sigma_{a_n} C_n))_{n\in\mathbb{N}}$ is.

Proof. For every $\varepsilon > 0$ and $L > 0$ we have

$$\begin{aligned}\mathbb{P}\left(|a_n' b_n| \geq L\right) &\leq \mathbb{E}\left(\mathbb{P}\left(|a_n' b_n| \geq L \| b_n\right)\right) \leq \mathbb{E}\left(\min\left(1, \mathbb{E}\left(\frac{\mathbb{E}\left(|a_n' b_n| \| b_n\right)}{L}\right)\right)\right) \\ &\leq \mathbb{P}\left(\frac{\mathbb{E}\left(|a_n' b_n| \| b_n\right)}{\varepsilon} \geq L\right) + \varepsilon.\end{aligned}$$

Now, Jensen's inequality implies

$$\mathbb{E}\left(|a_n' b_n| \| b_n\right) \leq \left(\mathbb{E}\left((a_n' b_n)^2 \| b_n\right)\right)^{\frac{1}{2}} = \left(b_n' \mathbb{E}\left(a_n a_n' \| b_n\right) b_n\right)^{\frac{1}{2}}.$$

Since a_n and b_n are independent and $\mathbb{E}(a_n a_n') = \Sigma_{a_n}$ we have

$$\left(b_n' \mathbb{E}\left(a_n a_n' \| b_n\right) b_n\right)^{1/2} = \left(b_n' \Sigma_{a_n} b_n\right)^{1/2} = \|\Sigma_{a_n}^{1/2} b_n\|_2.$$

Putting the pieces together, we end up with

$$\mathbb{P}\left(|a_n' b_n| \geq L\right) \leq \mathbb{P}\left(\|\Sigma_{a_n}^{1/2} b_n\|_2 \geq L\varepsilon\right) + \varepsilon. \tag{A.6}$$

Starting with an arbitrary $\varepsilon > 0$, the boundedness in probability of $\|\Sigma_{a_n}^{1/2} b_n\|_2$ implies the boundedness in probability of $\|\Sigma_{a_n}^{1/2} b_n\|_2/\varepsilon$. Thus, there is an $L_\varepsilon > 0$, such that $\sup_{n\in\mathbb{N}} \mathbb{P}\left(\frac{\|\Sigma_{a_n}^{1/2} b_n\|_2}{\varepsilon} \geq L_\varepsilon\right) \leq \varepsilon$. Hence, equation (A.6) yields $\sup_{n\in\mathbb{N}} \mathbb{P}\left(|a_n' b_n| \geq L_\varepsilon\right) \leq 2\varepsilon$, which proves the boundedness in probability of $(a_n' b_n)_{n\in\mathbb{N}}$.

[2] We could also argue that $\lim_{n\to\infty} k(n) < \infty$ implies that $k(n) = M$ for all $n \geq L$ for some L. In that case we would not need the monotonicity of N_k in our argument

For the second statement we start with an arbitrary $\varepsilon > 0$ and apply equation (A.6) with $L = \varepsilon$, which yields

$$\limsup_{n\in\mathbb{N}} \mathbb{P}\left(|a_n' b_n| \geq \varepsilon\right) \leq \limsup_{n\in\mathbb{N}} \mathbb{P}\left(\|\Sigma_{a_n}^{1/2} b_n\|_2 \geq \varepsilon^2\right) + \varepsilon = \varepsilon,$$

where we used the fact that $\|\Sigma_{a_n}^{1/2} b_n\|_2$ converges to 0 in probability. As $\varepsilon > 0$ was arbitrary, the claim follows.

With the same argument as above one can show that for every $\varepsilon > 0$ and $L > 0$ the following statement holds true:

$$\mathbb{P}\left(a_n' C_n a_n \geq L\right) \leq \mathbb{P}\left(\frac{\mathbb{E}\left(a_n' C_n a_n \| C_n\right)}{\varepsilon} \geq L\right) + \varepsilon.$$

Since a_n is independent of C_n, we have

$$\mathbb{E}\left(a_n' C_n a_n \| C_n\right) = \operatorname{tr}(\mathbb{E}\left(C_n a_n a_n' \| C_n\right)) = \operatorname{tr}(C_n \mathbb{E}\left(a_n a_n' \| C_n\right)) = \operatorname{tr}(C_n \Sigma_{a_n}).$$

Now, with the same argument as for $a_n' b_n$ the boundedness in probability of $\operatorname{tr}(C_n \Sigma_{a_n})$ implies the boundedness in probability of $a_n' C_n a_n$. □

Lemma A.7. *Let A and B be random variables. We then have*[3]

$$\int_{\mathbb{R}} \mathbb{P}(A \leq t < B) + \mathbb{P}(B \leq t < A) d\lambda(t) = \mathbb{E}(|B - A|).$$

The statement remains valid if we replace any " $<$ " sign by " $\leq$ " or the other way round.

Proof. We define $C := \max(0, B - A)$. Then it suffices to show that

$$\int_{\mathbb{R}} \mathbb{P}(A \leq t < A + C) d\lambda(t) = \mathbb{E}(C), \tag{A.7}$$

as the statement can be derived by using the fact that $\mathbb{E}(|B - A|) = \max(0, B - A) + \max(0, A - B)$ and exchanging the roles of A and B in equation (A.7).

Rewriting the probability and using Tonelli's theorem gives

$$\begin{aligned}\int_{\mathbb{R}} \mathbb{P}(A \leq t < A + C) d\lambda(t) &= \int_{\mathbb{R}} \mathbb{E}\left(\mathbb{1}\{A \leq t < A + C\}\right) d\lambda(t) \\ &= \mathbb{E}\left(\int_{\mathbb{R}} \mathbb{1}\{A \leq t < A + C\} d\lambda(t)\right) = \mathbb{E}(C).\end{aligned}$$

The last statement of Lemma A.7 can be seen by the fact that any single point t has

[3] We would like to emphasize that the statement also remains true if one of the two sides is infinitely large in the sense that the right-hand side is equal to ∞ if and only if the left-hand side is.

Lebesgue-measure 0, which implies

$$\int_{\mathbb{R}} \mathbb{1}\{A \leq t < A + C\} d\lambda(t) = \int_{\mathbb{R}} \mathbb{1}\{A \leq t \leq A + C\} d\lambda(t) = \int_{\mathbb{R}} \mathbb{1}\{A < t < A + C\} d\lambda(t).$$

□

Lemma A.8. *Let $f : \mathbb{R} \to [0, \infty)$ be a Lebesgue-integrable function. We then have*

$$\inf_{c \in [0,\varepsilon)} \varepsilon \sum_{j \in \mathbb{Z}} f(j\varepsilon + c) \leq \int_{\mathbb{R}} f(s) d\lambda(s).$$

Proof. Defining $g : \mathbb{R} \to [0, \infty]$ as $g(x) = \sum_{j \in \mathbb{Z}} f(j\varepsilon + x)$, we have

$$\inf_{c \in [0,\varepsilon)} \varepsilon g(c) \leq \int_{[0,\varepsilon)} g(x) d\lambda(x) = \int_{[0,\varepsilon)} \sum_{j \in \mathbb{Z}} f(j\varepsilon + x) d\lambda(x).$$

As f is nonnegative, we get (using the monotone convergence theorem)

$$\int_{[0,\varepsilon)} \sum_{j \in \mathbb{Z}} f(j\varepsilon + x) d\lambda(x) = \sum_{j \in \mathbb{Z}} \int_{[0,\varepsilon)} f(j\varepsilon + x) d\lambda(x).$$

Now, rewriting the expression above yields

$$\sum_{j \in \mathbb{Z}} \int_{[j\varepsilon,(j+1)\varepsilon)} f(x) d\lambda(x) = \int_{\mathbb{R}} f(x) d\lambda(x).$$

□

The statement of Lemma A.8 would be trivial if we replace $\inf_{c \in [0,\varepsilon)} \varepsilon \sum_{j \in \mathbb{Z}} f(j\varepsilon + c)$ by $\sum_{j \in \mathbb{Z}} \inf_{c \in [0,\varepsilon)} \varepsilon f(j\varepsilon + c)$. The interesting part lies in the fact that it even holds true if we are allowed to optimize only once.

The following lemma is the generalization of the following observation to an asymptotic setting: Let $\mu := \mathbb{E}(X) < \infty$. If $\mathbb{P}(X < \mu) = 0$, then $X = \mu$ almost surely.

Lemma A.9. *Let $(X_n)_{n \in \mathbb{N}}$ be a sequence of nonnegative random variables with means $\mu_n \in [0, K]$ for some $K \geq 0$. If for every $\varepsilon > 0$ we have*

$$\lim_{n \to \infty} \mathbb{P}(X_n \leq K - \varepsilon) = 0,$$

then X_n converges to K in $\mathcal{L}_1$ (and therefore in probability as well). Moreover, the limit of μ_n exists and equals K.

Proof. We start proving that $X_n \xrightarrow{p} K$ and show the convergence in $\mathcal{L}_1$ afterwards. First we point out that also $\lim_{n \to \infty} \mathbb{P}(X_n \leq \mu_n - \varepsilon) = 0$ holds true for every $\varepsilon > 0$ as

$\mu_n \leq K$. For every $n \in \mathbb{N}$ and $\varepsilon > 0$ we have

$$\mathbb{E}(X_n) \geq (\mu_n + \varepsilon)\mathbb{P}(X_n > \mu_n + \varepsilon) + (\mu_n - \varepsilon^2)\mathbb{P}(X_n \in [\mu_n - \varepsilon^2, \mu_n + \varepsilon]).$$

Now, define $\Delta_n := \mathbb{P}(X_n > \mu_n + \varepsilon)$ and $\delta_n := \mathbb{P}(X_n < \mu_n - \varepsilon^2)$. We then have

$$\mathbb{E}(X_n) \geq \Delta_n(\mu_n + \varepsilon) + (\mu_n - \varepsilon^2)(1 - \Delta_n - \delta_n) \geq \Delta_n(\varepsilon + \varepsilon^2) + (\mu_n - \varepsilon^2) - \delta_n\mu_n,$$

which yields

$$\limsup_{n\to\infty} \Delta_n \leq \limsup_{n\to\infty} \frac{\varepsilon^2 + \delta_n\mu_n + \mathbb{E}(X_n) - \mu_n}{\varepsilon^2 + \varepsilon} \leq \varepsilon,$$

where we used $\limsup_{n\to\infty} \delta_n\mu_n \leq \limsup_{n\to\infty} K\mathbb{P}(X_n \leq K - \varepsilon^2) = 0$. Now, fix an $\varepsilon' > 0$. Then, for any $0 < \varepsilon < \varepsilon'$ we have

$$\limsup_{n\to\infty} \mathbb{P}(X_n > \mu_n + \varepsilon') \leq \limsup_{n\to\infty} \mathbb{P}(X_n > \mu_n + \varepsilon) \leq \varepsilon.$$

As we can make ε arbitrarily small, we have

$$\lim_{n\to\infty} \mathbb{P}(X_n > \mu_n + \varepsilon') = 0 \text{ for all } \varepsilon' > 0.$$

Thus, $X_n - \mu_n$ converges to 0 in probability. Furthermore, the fact that μ_n is bounded by K also entails that

$$\lim_{n\to\infty} \mathbb{P}(X_n > K + \varepsilon') \leq \lim_{n\to\infty} \mathbb{P}(X_n > \mu_n + \varepsilon') = 0 \text{ for all } \varepsilon' > 0,$$

which also proves the convergence in probability to K and, by the uniqueness of the limit, $\lim_{n\to\infty} \mu_n = K$. To show the convergence in $\mathcal{L}_1$ we need to control the term $\tau_n := \mathbb{E}(X_n \mathbb{1}_{(\mu_n+\varepsilon,\infty)}(X_n))$ for an arbitrarily small ε, which can be done as follows:

$$\begin{aligned}
\mu_n = \mathbb{E}(X_n) &\geq \mathbb{E}(X_n \mathbb{1}_{(\mu_n-\varepsilon,\mu_n+\varepsilon)}(X_n)) + \mathbb{E}(X_n \mathbb{1}_{(\mu_n+\varepsilon,\infty)}(X_n)) \\
&\geq (\mu_n - \varepsilon)\mathbb{P}(X_n \in (\mu_n - \varepsilon, \mu_n + \varepsilon)) + \tau_n \\
&\geq \mu_n\mathbb{P}(X_n \in (\mu_n - \varepsilon, \mu_n + \varepsilon)) - \varepsilon + \tau_n.
\end{aligned}$$

Rearranging the terms and taking the limits yields

$$\limsup_{n\to\infty} \tau_n \leq \varepsilon + \limsup_{n\to\infty} \mu_n\mathbb{P}(|X_n - \mu_n| \geq \varepsilon) = \varepsilon.$$

Altogether we have

$$\begin{aligned}
\mathbb{E}(|X_n - \mu_n|) &\leq \mathbb{E}((\mu_n - X_n)\mathbb{1}_{[0,\mu_n-\varepsilon)}(X_n)) + \mathbb{E}(|\mu_n - X_n|\mathbb{1}_{[\mu_n-\varepsilon,\mu_n+\varepsilon]}(X_n)) + \tau_n \\
&\leq K\mathbb{P}(X_n \leq \mu_n - \varepsilon) + \varepsilon + \tau_n.
\end{aligned}$$

Taking the limits gives

$$\limsup_{n\to\infty} \mathbb{E}(|X_n - \mu_n|) \leq 2\varepsilon.$$

As $\varepsilon > 0$ can be made arbitrarily small, the result follows. □

We would like to emphasize that Lemma A.9 not only deals with the limit of $X_n - \mu_n$, but also guarantees the existence of $\lim_{n\to\infty} \mu_n$ and connects it to K.

Lemma A.10. *Let $\hat{\alpha}, \hat{\delta}, \tilde{\alpha}$ and $\tilde{\delta}$ be nonnegative numbers. We then have the following inequality:*

$$\left|\min(1, \hat{\alpha}\hat{\delta}) - \min(1, \tilde{\alpha}\tilde{\delta})\right| \leq \min\left(1, \hat{\delta}|\hat{\alpha} - \tilde{\alpha}|\right) + \min\left(1, \tilde{\alpha}\left|\hat{\delta} - \tilde{\delta}\right|\right). \tag{A.8}$$

Proof. We start with the observation that it suffices to show that

$$\left|\min(1, \hat{\alpha}\hat{\delta}) - \min(1, \tilde{\alpha}\tilde{\delta})\right| \leq \hat{\delta}\,|\hat{\alpha} - \tilde{\alpha}| + \tilde{\alpha}\left|\hat{\delta} - \tilde{\delta}\right|. \tag{A.9}$$

To see this, we point out that $|\min(1, \hat{\alpha}\hat{\delta}) - \min(1, \tilde{\alpha}\tilde{\delta})| \leq 1$. Therefore, equation (A.9) implies

$$\left|\min(1, \hat{\alpha}\hat{\delta}) - \min(1, \tilde{\alpha}\tilde{\delta})\right| \leq \min\left(1, \hat{\delta}\,|\hat{\alpha} - \tilde{\alpha}| + \tilde{\alpha}\left|\hat{\delta} - \tilde{\delta}\right|\right),$$

where the right-hand side of the preceding display can be bounded from above by

$$\min\left(1, \hat{\delta}|\hat{\alpha} - \tilde{\alpha}|\right) + \min\left(1, \tilde{\alpha}\left|\hat{\delta} - \tilde{\delta}\right|\right)$$

as $\hat{\delta}$ and $\tilde{\alpha}$ are nonnegative.

We now distinguish four cases:

If $\hat{\alpha}\hat{\delta} \geq 1$ and $\tilde{\alpha}\tilde{\delta} \geq 1$, the left-hand side of (A.8) equals 0. Recalling that $\hat{\delta}$ and $\tilde{\alpha}$ are nonnegative, the statement holds true as the right-hand side cannot be negative.

If $\hat{\alpha}\hat{\delta} < 1$ and $\tilde{\alpha}\tilde{\delta} < 1$, we have

$$\left|\min(1, \hat{\alpha}\hat{\delta}) - \min(1, \tilde{\alpha}\tilde{\delta})\right| = \left|\hat{\alpha}\hat{\delta} - \tilde{\alpha}\tilde{\delta}\right| \leq \left|\hat{\alpha}\hat{\delta} - \tilde{\alpha}\hat{\delta}\right| + \left|\tilde{\alpha}\hat{\delta} - \tilde{\alpha}\tilde{\delta}\right|.$$

Now the nonnegativity of $\hat{\delta}$ and $\tilde{\alpha}$ implies

$$\left|\hat{\alpha}\hat{\delta} - \tilde{\alpha}\hat{\delta}\right| + \left|\tilde{\alpha}\hat{\delta} - \tilde{\alpha}\tilde{\delta}\right| = \hat{\delta}\,|\hat{\alpha} - \tilde{\alpha}| + \tilde{\alpha}\left|\hat{\delta} - \tilde{\delta}\right|,$$

which proves this case.

If $\hat{\alpha}\hat{\delta} \geq 1 > \tilde{\alpha}\tilde{\delta}$, we have

$$\left|\min(1, \hat{\alpha}\hat{\delta}) - \min(1, \tilde{\alpha}\tilde{\delta})\right| = 1 - \tilde{\alpha}\tilde{\delta}.$$

Now, we have

$$1 - \tilde{\alpha}\tilde{\delta} \leq \hat{\alpha}\hat{\delta} - \tilde{\alpha}\tilde{\delta} = \hat{\delta}\left(\hat{\alpha} - \tilde{\alpha}\right) + \tilde{\alpha}\left(\hat{\delta} - \tilde{\delta}\right) \leq \hat{\delta}\left|\hat{\alpha} - \tilde{\alpha}\right| + \tilde{\alpha}\left|\hat{\delta} - \tilde{\delta}\right|,$$

which proves this case.

The case $\hat{\alpha}\hat{\delta} < 1 \leq \tilde{\alpha}\tilde{\delta}$ can be proven analogously:

$$1 - \hat{\alpha}\hat{\delta} \leq \tilde{\alpha}\tilde{\delta} - \hat{\alpha}\hat{\delta} = \hat{\delta}\left(\tilde{\alpha} - \hat{\alpha}\right) + \tilde{\alpha}\left(\tilde{\delta} - \hat{\delta}\right) \leq \hat{\delta}\left|\hat{\alpha} - \tilde{\alpha}\right| + \tilde{\alpha}\left|\hat{\delta} - \tilde{\delta}\right|.$$

□

We will make use of the following Burkholder-type inequality given in Wu and Shao (2007). Since they do not provide a proof, we will show that this is a simple consequence of Burkholder (1966).

Lemma A.11 (Lemma 1 of Wu and Shao 2007)**.** *Let $D_1, \ldots, D_n$ be a martingale difference sequence for which $\mathbb{E}(|D_i|^p) < \infty$, $p > 1$. Let $p' = \min(2, p)$. Then*

$$\left[\mathbb{E}\left|\sum_{i=1}^{n} D_i\right|^p\right]^{\frac{p'}{p}} \leq C_p^{p'} \sum_{i=1}^{n} \left[\mathbb{E}|D_i|^p\right]^{\frac{p'}{p}},$$

where C_p is a constant only depending on p.[4]

Proof. We start with the Burkholder-type inequality given by Theorem 9 of Burkholder (1966):

$$M_p\,\mathbb{E}\left(\sum_{i=1}^{n} D_i^2\right)^{\frac{p}{2}} \leq \mathbb{E}\left|\sum_{i=1}^{n} D_i\right|^p \leq N_p\,\mathbb{E}\left(\sum_{i=1}^{n} D_i^2\right)^{\frac{p}{2}}, \tag{A.10}$$

where N_p and M_p are positive real numbers only depending on $1 < p < \infty$. We distinguish two cases: Case $1 < p < 2$: As $p < 2$ we have for any vector $\|x\|_p \geq \|x\|_2$, implying

$$\left(\sum_{i=1}^{n} D_i^2\right)^{\frac{p}{2}} \leq \left(\sum_{i=1}^{n} |D_i|^p\right).$$

Putting the pieces together, we end up with

$$\mathbb{E}\left|\sum_{i=1}^{n} D_i\right|^p \leq N_p\,\mathbb{E}\left(\sum_{i=1}^{n} |D_i|^p\right),$$

which proves the first part.

[4] In Lemma 1 of Wu and Shao (2007) the constant is explicitly given as $C_p = 18p^{3/2}/(p-1)^{1/2}$. However, we will only prove the existence of such a constant without its exact value.

Case $p \geq 2$: We start using equation (A.10) to get

$$\left[\mathbb{E} \left| \sum_{i=1}^{n} D_i \right|^p \right]^{\frac{2}{p}} \leq N_p^{\frac{2}{p}} \left[\mathbb{E} \left(\sum_{i=1}^{n} D_i^2 \right)^{\frac{p}{2}} \right]^{\frac{2}{p}}.$$

Since $\|X\|_{L_q} := [\mathbb{E}(|X|^q)]^{1/q}$ is a norm for $q \geq 1$ the triangle inequality yields

$$\left[\mathbb{E} \left(\sum_{i=1}^{n} D_i^2 \right)^{\frac{p}{2}} \right]^{\frac{2}{p}} \leq \sum_{i=1}^{n} \left[\mathbb{E} \left(D_i^2 \right)^{\frac{p}{2}} \right]^{\frac{2}{p}} = \sum_{i=1}^{n} \left[\mathbb{E} |D_i|^p \right]^{\frac{2}{p}}.$$

□

A.2. Proofs of Chapter 3

Proof of Lemma 3.2: We start proving equation (3.4): To do so, we fix an arbitrary $t \in \mathbb{R}$. Then, for any $s \geq t$ and every $u \in \mathbb{R}$ we have

$$G(t) \leq G(s) \leq F(u) + |G(s) - F(u)|.$$

Thus, we have

$$\begin{aligned} G(t) &\leq \inf_{u,s \in K_{\varepsilon/2}(t+\varepsilon/2)} F(u) + |G(s) - F(u)| \leq F(t+\varepsilon) + \inf_{u,s \in K_{\varepsilon/2}(t+\varepsilon/2)} |G(s) - F(u)| \\ &\leq F(t+\varepsilon) + \ell_\varepsilon(F, G). \end{aligned}$$

With a similar argument one can show that also $G(t) \geq F(t-\varepsilon) - \ell_\varepsilon(F, G)$ holds true.

Equation (3.5) can be shown by using the fact that

$$\mathbb{P}\left(y_0 \in \hat{y}_0 + [q_1 - \varepsilon, q_2 + \varepsilon] \| T_n\right) = \mathrm{F}_n(q_2 + \varepsilon) - \lim_{\delta \searrow 0} \mathrm{F}_n(q_1 - \delta - \varepsilon)$$

and using equation (3.4) twice to get

$$\begin{aligned} &\widehat{\mathrm{F}}_n(q_2 + 2\varepsilon) - \lim_{\delta \searrow 0} \widehat{\mathrm{F}}_n(q_1 - \delta - 2\varepsilon) + 2\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n) \\ \geq &\mathrm{F}_n(q_2 + \varepsilon) - \lim_{\delta \searrow 0} \mathrm{F}_n(q_1 - \delta - \varepsilon) \\ \geq &\widehat{\mathrm{F}}_n(q_2) - \lim_{\delta \searrow 0} \widehat{\mathrm{F}}_n(q_1 - \delta) - 2\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n), \end{aligned}$$

which proves equation (3.5). For equation (3.6) we start with the observation that for every $q_2 > q_1 + 2\varepsilon$ we have

$$\mathbb{P}\left(y_0 \in \hat{y}_0 + (q_1 + \varepsilon, q_2 - \varepsilon) \| T_n\right) = \lim_{\delta \searrow 0} \mathrm{F}_n(q_2 - \varepsilon - \delta) - \mathrm{F}_n(q_1 + \varepsilon).$$

Now the claim follows with the same argument as for equation (3.5). □

In order to proof Lemma 3.3 we need the following result:

Lemma A.12. *Let $f : \mathbb{R} \to [M_1, M_2]$ be a non-decreasing function and let X and Y be real random variables. Then, for every $\varepsilon \geq 0$ and every $t \in \mathbb{R}$ we have*

$$\mathbb{P}(f(X) > t) - \mathbb{P}(f(Y + \varepsilon) > t) \leq \ell_\varepsilon(F_X, F_Y), \tag{A.11}$$

where F_X and F_Y denote the distribution functions of X and Y, respectively.

Proof. We start with the definition

$$f^{-1}(t) := \sup\{a \in \mathbb{R} : f(a) \leq t\} \in [-\infty, \infty],$$

where we use the convention $\sup \emptyset = -\infty$, and distinguish four cases:

Case $f^{-1}(t) = -\infty$: Here, $f(x) > t$ holds true for all $x \in \mathbb{R}$. Thus, $\mathbb{P}(f(X) > t) = 1 = \mathbb{P}(f(Y + \varepsilon) > t)$ holds true. Now the claim follows as $\ell_\varepsilon(F_X, F_Y) \geq 0$.

Case $f^{-1}(t) = \infty$. Since f is non-decreasing $f(x) \leq t$ holds true for all $x \in \mathbb{R}$. Thus, we have $\mathbb{P}(f(X) > t) = 0 = \mathbb{P}(f(Y + \varepsilon) > t)$, which proves equation (A.11).

Case $|f^{-1}(t)| < \infty$ and $f(f^{-1}(t)) \leq t$: We claim $a > f^{-1}(t) \Leftrightarrow f(a) > t$ holds true or equivalently $a \leq f^{-1}(t) \Leftrightarrow f(a) \leq t$. To prove this, we start with an $a \leq f^{-1}(t)$. By monotonicity, we have $f(a) \leq f(f^{-1}(t))$, where the right-hand side is not larger than t in this case. For the reverse direction we start with an a fulfilling $f(a) \leq t$. Thus, $a \in \{x \in \mathbb{R} : f(x) \leq t\}$ and therefore $a \leq \sup\{x \in \mathbb{R} : f(x) \leq t\} = f^{-1}(t)$. Hence, we have

$$\begin{aligned}\mathbb{P}(f(X) > t) - \mathbb{P}(f(Y + \varepsilon) > t) &= \mathbb{P}(X > f^{-1}(t)) - \mathbb{P}(Y > f^{-1}(t) - \varepsilon) \\ &= F_Y(f^{-1}(t) - \varepsilon) - F_X(f^{-1}(t)) \leq \ell_\varepsilon(F_X, F_Y),\end{aligned}$$

where the last inequality is given by equation (3.4).

Case $|f^{-1}(t)| < \infty$ and $f(f^{-1}(t)) > t$: We claim that $a \geq f^{-1}(t) \Leftrightarrow f(a) > t$. By monotonicity, $a \geq f^{-1}(t)$ implies $f(a) \geq f(f^{-1}(t))$, which in this case is strictly larger than t. For the other direction, we start with an a fulfilling $f(a) > t$. Thus, a is not contained in the set $\{x \in \mathbb{R} : f(x) \leq t\}$. Recalling that f is non-decreasing, we conclude $a \geq \sup\{x \in \mathbb{R} : f(x) \leq t\}$. Hence, we have

$$\begin{aligned}\mathbb{P}(f(X) > t) - \mathbb{P}(f(Y + \varepsilon) > t) &= \mathbb{P}(X \geq f^{-1}(t)) - \mathbb{P}(Y \geq f^{-1}(t) - \varepsilon) \\ &= \lim_{\delta \searrow 0} \left[F_Y(f^{-1}(t) - \varepsilon - \delta) - F_X(f^{-1}(t) - \delta)\right] \\ &\leq \ell_\varepsilon(F_X, F_Y),\end{aligned}$$

where, again, the last inequality is given by equation (3.4). □

Proof of Lemma 3.3: We start showing equation (3.8). W.l.o.g. we assume that $M_1 = 0$ as otherwise we could shift f by $-M_1$ and let M denote $M_2 - M_1$. As f is bounded and

nonnegative, we get

$$\begin{aligned}\mathbb{E}(f(X)) - \mathbb{E}(f(Y+\varepsilon)) &= \int_0^M \mathbb{P}(f(X) > t) - \mathbb{P}(f(Y+\varepsilon) > t)dt \\ &\leq \int_0^M \ell_\varepsilon(F_X, F_Y)dt = M\ell_\varepsilon(F_X, F_Y),\end{aligned}$$

where the inequality in the preceding display is given by Lemma A.12. With the same arguments as before (or by applying the results above to $\tilde{X} = Y - \varepsilon$ and $\tilde{Y} = X - \varepsilon$) we also get $\mathbb{E}(f(X)) - \mathbb{E}(f(Y-\varepsilon)) \geq -M\ell_\varepsilon(F_X, F_Y)$, proving equation (3.7).

Equation (3.9) can be directly derived from equation (3.7) and equation (3.8) as the Lipschitz-continuity implies $\mathbb{E}(f(Y-\varepsilon)) \geq \mathbb{E}(f(Y)) - L\varepsilon$ and $\mathbb{E}(f(Y+\varepsilon)) \leq \mathbb{E}(f(Y)) + L\varepsilon$.

In order to prove equation (3.10) we would like to point out that we already showed $|E(f(X)) - \mathbb{E}(f(Y))| \leq M\ell_0(F_X, F_Y) = M\|F_X - F_Y\|_\infty$ to hold true for non-decreasing functions f. Since $\sup_{L>0} V_{-L}^L(g) < \infty$ we can find for every $\delta > 0$ a $K > 0$, such that the total variation $\sup_{L>0} V_{-L}^L(g) \leq V_{-K}^K(g) + \delta$. Thus, we have $|g(x) - g(K)| \leq \delta$ for every $x \geq K$ and $|g(x) - g(-K)| \leq \delta$ for every $x \leq -K$. We define a function $g_K(x) : \mathbb{R} \to \mathbb{R}$ as

$$g_K(x) = \begin{cases} g(x) & \text{if } x \in [-K, K] \\ g(-K) & \text{if } x < -K \\ g(K) & \text{if } x > K. \end{cases}$$

We then have $\|g_K - g\|_\infty \leq \delta$, implying $|\,\mathbb{E}(g(X)) - \mathbb{E}(g_K(X))| \leq \delta$ as well as $|\,\mathbb{E}(g(Y)) - \mathbb{E}(g_K(Y))| \leq \delta$. Moreover, we have $V_{-K}^K(g_K) = V_{-K}^K(g) \leq \sup_{L>0} V_{-L}^L(g)$.

Now, as the total variation of g_K on $[-K, K]$ is finite, we can split it into two non-decreasing functions g^+ and g^-, such that $g_K(x) = g^+(x) - g^-(x)$ for all x in $[-K, K]$ (cf. Section 31 of Chapter 6 in Billingsley 1995). By setting $g^+(x) = g^+(K)$ for all $x > K$, $g^+(x) = g^+(-K)$ for all $x < -K$ and defining g^- outside of $[-K, K]$ analogously, we see that $g_K(x) = g^+(x) - g^-(x)$ for all $x \in \mathbb{R}$. As the total variation of g_K is finite, the functions g^+ and g^- are also bounded. Furthermore, we have $V_{-K}^K(g_K) = g^+(K) - g^+(-K) + g^-(K) - g^-(-K)$. By the monotonicity of g^+ and g^- the latter term can be expressed equivalently as $\sup_{x\in\mathbb{R}} g^+(x) - \inf_{x\in\mathbb{R}} g^+(x) + \sup_{x\in\mathbb{R}} g^-(x) - \inf_{x\in\mathbb{R}} g^-(x)$. Applying equation (3.7) and equation (3.8) with $\varepsilon = 0$ to the functions g^+ and g^- we conclude

$$\begin{aligned}|\,\mathbb{E}(g_K(X)) - \mathbb{E}(g_K(Y))| &\leq \big|\,\mathbb{E}(g^+(X)) - \mathbb{E}(g^+(Y))\big| + \big|\,\mathbb{E}(g^-(X)) - \mathbb{E}(g^-(Y))\big| \\ &\leq \|F_X - F_Y\|_\infty \left(\sup_{x\in\mathbb{R}} g^+(x) - \inf_{x\in\mathbb{R}} g^+(x) + \sup_{x\in\mathbb{R}} g^-(x) - \inf_{x\in\mathbb{R}} g^-(x)\right) \\ &= V_{-K}^K(g_K)\|F_X - F_Y\|_\infty.\end{aligned}$$

Putting the pieces together, we end up with

$$\begin{aligned}|\,\mathbb{E}(g(X)) - \mathbb{E}(g(Y))| &\leq |\,\mathbb{E}(g_K(X)) - \mathbb{E}(g_K(Y))| + 2\delta \\ &\leq V_{-K}^{K}(g_K)\|F_X - F_Y\|_\infty + 2\delta \\ &\leq \sup_{L>0} V_{-L}^{L}(g)\|F_X - F_Y\|_\infty + 2\delta.\end{aligned}$$

As $\delta > 0$ can be made arbitrarily small, equation (3.10) holds true. □

Proof of Lemma 3.4: Before starting the proof we would like to mention that the limit $\lim_{\varepsilon_n \searrow \varepsilon} \ell_{\varepsilon_n}(F,G)$ exists, as it is a non-increasing, real-valued function bounded from above. Further, we only have to show that $\ell_\varepsilon(F,G) \leq \lim_{k\to\infty} \ell_{\varepsilon+1/k}(F,G)$ as the other inequality follows directly from the fact that $\ell_\varepsilon(F,G) \geq \ell_{\varepsilon+1/k}(F,G)$ for every $\varepsilon \geq 0$.

We distinguish two cases: If $\ell_\varepsilon(F,G) = 0$, the inequality $\ell_\varepsilon(F,G) \leq \lim_{k\to\infty} \ell_{\varepsilon+1/k}(F,G)$ is trivial.

In the case $\ell_\varepsilon(F,G) \in (0,1]$, we take an arbitrary (small) $\delta \in (0, \ell_\varepsilon(F,G))$ and a t^*, such that $\inf_{x,y\in K_{\varepsilon/2}(t^*)} |F(x) - G(y)| \geq \ell_\varepsilon(F,G) - \delta > 0$. The proof will fundamentally rely on the following inequality: For every $x \in \mathbb{R}$ we have

$$\begin{aligned}\inf_{y\in[t^*-\frac{\varepsilon}{2},t^*+\frac{\varepsilon}{2}]} |F(x) - G(y)| \leq &\inf_{y\in[t^*-\frac{\varepsilon}{2},t^*+\frac{\varepsilon}{2}+\frac{2}{k}]} |F(x) - G(y)| \\ &+ \sup_{y\in[t^*+\frac{\varepsilon}{2},t^*+\frac{\varepsilon}{2}+\frac{2}{k}]} \left|G(y) - G\left(t^* + \frac{\varepsilon}{2}\right)\right|.\end{aligned} \tag{A.12}$$

To see that this holds true, we start with the triangle inequality $|F(x) - G(z)| \leq |F(x) - G(y)| + |G(y) - G(z)|$, which implies

$$\inf_{z\in[t^*-\frac{\varepsilon}{2},t^*+\frac{\varepsilon}{2}]} |F(x) - G(z)| \leq |F(x) - G(y)| + \inf_{z\in[t^*-\frac{\varepsilon}{2},t^*+\frac{\varepsilon}{2}]} |G(y) - G(z)|.$$

Taking the supremum and afterwards the infimum over y yields

$$\begin{aligned}\inf_{z\in[t^*-\frac{\varepsilon}{2},t^*+\frac{\varepsilon}{2}]} |F(x) - G(z)| \leq &\inf_{y\in[t^*-\frac{\varepsilon}{2},t^*+\frac{\varepsilon}{2}+\frac{2}{k}]} |F(x) - G(y)| \\ &+ \sup_{y\in[t^*-\frac{\varepsilon}{2},t^*+\frac{\varepsilon}{2}+\frac{2}{k}]} \inf_{z\in[t^*-\frac{\varepsilon}{2},t^*+\frac{\varepsilon}{2}]} |G(y) - G(z)|.\end{aligned}$$

Now, we conclude

$$\sup_{y\in[t^*-\frac{\varepsilon}{2},t^*+\frac{\varepsilon}{2}+\frac{2}{k}]} \inf_{z\in[t^*-\frac{\varepsilon}{2},t^*+\frac{\varepsilon}{2}]} |G(y) - G(z)| \leq \sup_{y\in[t^*+\frac{\varepsilon}{2},t^*+\frac{\varepsilon}{2}+\frac{2}{k}]} \left|G(y) - G\left(t^* + \frac{\varepsilon}{2}\right)\right|,$$

which can be seen by choosing $z = y$ if $y \in K_{\varepsilon/2}(t^*)$ and $z = t^* + \varepsilon/2$ otherwise. Thus, we proved inequality (A.12), which we will use to draw the link between $\ell_\varepsilon(F,G)$ and

$\ell_{\varepsilon+1/k}(F,G)$ as follows:

$$\begin{aligned}&\inf_{x,y\in K_{\frac{\varepsilon}{2}}(t^*)}|F(x)-G(y)| = \inf_{x\in K_{\frac{\varepsilon}{2}}(t^*)}\ \inf_{y\in K_{\frac{\varepsilon}{2}}(t^*)}|F(x)-G(y)|\\ \leq &\inf_{x\in K_{\frac{\varepsilon}{2}}(t^*)}\ \inf_{y\in K_{\frac{\varepsilon}{2}+\frac{1}{k}}(t^*+\frac{1}{k})}|F(x)-G(y)| + \sup_{y\in[t^*+\frac{\varepsilon}{2},t^*+\frac{\varepsilon}{2}+\frac{2}{k}]}\left|G(y)-G\left(t^*+\frac{\varepsilon}{2}\right)\right|.\end{aligned}$$

The last line in the preceding display coincides with

$$\inf_{y\in K_{\frac{\varepsilon}{2}+\frac{1}{k}}(t^*+\frac{1}{k})}\ \inf_{x\in K_{\frac{\varepsilon}{2}}(t^*)}|F(x)-G(y)| + \sup_{y\in[t^*+\frac{\varepsilon}{2},t^*+\frac{\varepsilon}{2}+\frac{2}{k}]}\left|G(y)-G\left(t^*+\frac{\varepsilon}{2}\right)\right|.$$

By changing the roles of F and G in inequality (A.12) we get an upper bound for $\inf_{x\in K_{\varepsilon/2}(t^*)}|F(x)-G(y)|$, which is used to bound the preceding display from above by

$$\begin{aligned}&\inf_{y\in K_{\frac{\varepsilon}{2}+\frac{1}{k}}(t^*+\frac{1}{k})}\ \inf_{x\in K_{\frac{\varepsilon}{2}+\frac{1}{k}}(t^*+\frac{1}{k})}|F(x)-G(y)|\\ &\quad+\sup_{y\in[t^*+\frac{\varepsilon}{2},t^*+\frac{\varepsilon}{2}+\frac{2}{k}]}\left|G(y)-G\left(t^*+\frac{\varepsilon}{2}\right)\right| + \sup_{x\in[t^*+\frac{\varepsilon}{2},t^*+\frac{\varepsilon}{2}+\frac{2}{k}]}\left|F\left(t^*+\frac{\varepsilon}{2}\right)-F(x)\right|\\ &\leq\ \ell_{\varepsilon+\frac{2}{k}}(F,G) + \sup_{y\in[t^*+\frac{\varepsilon}{2},t^*+\frac{\varepsilon}{2}+\frac{2}{k}]}\left|G(y)-G\left(t^*+\frac{\varepsilon}{2}\right)\right|\\ &\quad+\sup_{x\in[t^*+\frac{\varepsilon}{2},t^*+\frac{\varepsilon}{2}+\frac{2}{k}]}\left|F\left(t^*+\frac{\varepsilon}{2}\right)-F(x)\right|.\end{aligned}$$

As F and G are continuous from the right, the expressions containing the suprema vanish as k tends to infinity. Thus, we end up with $\ell_\varepsilon(F,G)\leq\delta+\lim_{k\to\infty}\ell_{\varepsilon+2/k}(F,G)$. As δ can be made arbitrarily small, the result follows. □

We would like to point out that we do not need the monotonicity of F and G. A small modification of the proof including the case where $\ell_\varepsilon(F,G)=\infty$ shows that we do not need the functions to be bounded from above either. Hence, the statement of Lemma 3.4 holds true for all nonnegative functions being continuous from the right. Applying the proof of Lemma 3.4 to the functions $F(-x)$ and $G(-x)$ we can also extend it to the class of all nonnegative functions being continuous from the left. We would like to emphasize that the functions in Lemma 3.4 have to be continuous from the same side. To see this, consider the functions $F=\mathbb{1}_{[0,\infty)}$ and $G=\mathbb{1}_{(\varepsilon,\infty)}$. In that case we have $\ell_\varepsilon(F,G)=1$, while $\ell_{\varepsilon+1/k}(F,G)=0$ for all $k\in\mathbb{N}$. However, in the case $\varepsilon=0$ we have $\lim_{k\to\infty}\ell_{1/k}(F,G)=\|F-G\|_{\mathcal{L}_\infty}$. To see that this does not always need to be the case, we can consider the indicator functions on $\mathbb{R}$ and $\mathbb{Q}$. Then, for all $\varepsilon>0$ we have $\ell_\varepsilon(F,G)=0$, while $\|F-G\|_{\mathcal{L}_\infty}=\|F-G\|_\infty=1$.

Proof of Lemma 3.5. We start with proving the reverse direction, which hinges on the fact that the Lévy metric between F_n and F is bounded from above by $\max(\varepsilon,\ell_\varepsilon(F_n,F))$:

To see this, we use the monotonicity of F together with equation (3.4) to get

$$\begin{aligned}
&F_n(t - \max(\varepsilon, \ell_\varepsilon(F_n, F))) - \max(\varepsilon, \ell_\varepsilon(F_n, F)) \\
&\leq F_n(t - \varepsilon) - \ell_\varepsilon(F_n, F) \leq F(t) \leq F_n(t + \varepsilon) + \ell_\varepsilon(F_n, F) \\
&\leq F_n(t + \max(\varepsilon, \ell_\varepsilon(F_n, F))) + \max(\varepsilon, \ell_\varepsilon(F_n, F))
\end{aligned}$$

for all $t \in \mathbb{R}$. Recalling the definition of the Lévy metric

$$L(F_n, F) = \inf\{\delta \geq 0 : F_n(t - \delta) - \delta \leq F(t) \leq F_n(t + \delta) + \delta \text{ for all } t \in \mathbb{R}\}$$

immediately implies $L(F_n, F) \leq \max(\varepsilon, \ell_\varepsilon(F_n, F))$. Thus, if for every $\varepsilon > 0$ the expression $\ell_\varepsilon(F_n, F)$ converges to 0, then the Lévy metric between F_n and F vanishes as well. Hence, F_n converges weakly to F.

To prove the other direction, we assume F_n to converge weakly to F and show that for every $\varepsilon > 0$ and $\delta > 0$ we have $\limsup_{n\to\infty} \ell_\varepsilon(F_n, F) \leq 2\delta$, which implies for every $\varepsilon > 0$ that $\ell_\varepsilon(F_n, F)$ converges to 0. We start by noticing that weak convergence implies the existence of a dense subset D of $\mathbb{R}$, such that $\lim_{n\to\infty} F_n(x) = F(x)$ for all $x \in D$. We take an arbitrary $\varepsilon > 0$ and $\delta > 0$ and define two points $a_- \in D$ and $a_+ \in D$, such that $F(a_-) \leq \delta$ and $F(a_+) \geq 1 - \delta$. By the pointwise convergence, we can find an $N \in \mathbb{N}$, such that $|F_n(a_-) - F(a_-)| \leq \delta$ and $|F_n(a_+) - F(a_+)| \leq \delta$ for all $n \geq N$. Now the monotonicity of F_n and F implies $|F_n(x) - F(x)| \leq 2\delta$ for all $x \notin [a_-, a_+]$ and $n \geq N$.

As the interval $[a_-, a_+]$ is compact and D is dense, we can find a finite number of points $s_i \in D \cap [a_-, a_+]$ with $1 \leq i \leq l$, such that $\bigcup_{i=1}^{l} K_{\varepsilon/2}(s_i) \supseteq [a_-, a_+]$. We then have

$$\begin{aligned}
&\sup_{t\in[a_-,a_+]} \inf_{x,y\in K_{\varepsilon/2}(t)} |F_n(x) - F(y)| \leq \max_{1\leq i\leq l} \sup_{t\in K_{\varepsilon/2}(s_i)} \inf_{x,y\in K_{\varepsilon/2}(t)} |F_n(x) - F(y)| \\
&\leq \max_{1\leq i\leq l} \sup_{t\in K_{\varepsilon/2}(s_i)} |F_n(s_i) - F(s_i)| = \max_{1\leq i\leq l} |F_n(s_i) - F(s_i)|.
\end{aligned} \tag{A.13}$$

In order to show that $\limsup_{n\to\infty} \ell_\varepsilon(F_n, F) \leq 2\delta$ we start with

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{t\in\mathbb{R}} \inf_{x,y\in K_{\varepsilon/2}(t)} |F_n(x) - F(y)| \\
&\leq \limsup_{n\to\infty} \left(\sup_{t\in[a_-,a_+]} \inf_{x,y\in K_{\varepsilon/2}(t)} |F_n(x) - F(y)| + \sup_{t\notin[a_-,a_+]} \inf_{x,y\in K_{\varepsilon/2}(t)} |F_n(x) - F(y)| \right),
\end{aligned}$$

which can be bounded from above using the property $|F_n(x) - F(x)| \leq 2\delta$ for all $x \notin [a_-, a_+]$ together with equation (A.13) as follows:

$$\begin{aligned}
&\limsup_{n\to\infty} \max_{1\leq i\leq l} |F_n(s_i) - F(s_i)| + \limsup_{n\to\infty} \sup_{t\notin[a_-,a_+]} |F_n(t) - F(t)| \\
&\leq \max_{1\leq i\leq l} \limsup_{n\to\infty} |F_n(s_i) - F(s_i)| + 2\delta = 2\delta,
\end{aligned}$$

where we use the pointwise convergence in the last line as the s_i do not change over n.

Furthermore, the exchange of the maximum and the limit is justified as the number l is fixed over n and thus finite. □

A.3. Proofs of Chapter 4

Proof of Proposition 4.1: Before we start, we would like to point out that $\hat{q}_{\alpha_i}$ (for $i \in \{1,2\}$) depends only on the training data T_n and hence is fixed if we condition on the training data. By the definition of $\hat{q}_\alpha$ we immediately conclude $\widehat{\mathrm{F}}_n^-(\hat{q}_\alpha) < \alpha$ for all $\alpha \in (0,1]$, $\widehat{\mathrm{F}}_n^-(\hat{q}_0) = 0$ and $\alpha \leq \widehat{\mathrm{F}}_n(\hat{q}_\alpha)$ for all $\alpha \in [0,1]$. Thus, we have $\widehat{\mathrm{F}}_n^-(\hat{q}_\alpha) \leq \alpha \leq \widehat{\mathrm{F}}_n(\hat{q}_\alpha)$ for all $\alpha \in [0,1]$. Defining $q_2 = \hat{q}_{\alpha_2}$ and $q_1 = \hat{q}_{\alpha_1}$ we conclude $q_2 \geq q_1 \geq q_1 - 2\varepsilon$, which allows us to apply equation (3.5) of Lemma 3.2 to get

$$\mathbb{P}\left(y_0 \in \hat{y}_0 + [\hat{q}_{\alpha_1} - \varepsilon, \hat{q}_{\alpha_2} + \varepsilon] \| T_n\right) \geq \widehat{\mathrm{F}}_n(\hat{q}_{\alpha_2}) - \widehat{\mathrm{F}}_n^-(\hat{q}_{\alpha_1}) - 2\ell_n(\varepsilon) \geq \alpha_2 - \alpha_1 - 2\ell_n(\varepsilon)$$

almost surely, which proves the first part.

For the second part we distinguish two cases: If $\hat{q}_{\alpha_2} \leq \hat{q}_{\alpha_1} + 2\varepsilon$, then the prediction interval $\hat{y}_0 + (\hat{q}_{\alpha_1} + \varepsilon, \hat{q}_{\alpha_2} - \varepsilon)$ coincides with the empty set possessing a coverage probability of zero. Thus, equation (4.5) is trivially fulfilled as $\ell_n(\varepsilon)$ is nonnegative and $\alpha_2 \geq \alpha_1$. If $\hat{q}_{\alpha_2} > \hat{q}_{\alpha_1} + 2\varepsilon$, we can use equation (3.6) of Lemma 3.2 with $q_2 = \hat{q}_{\alpha_2}$ and $q_1 = \hat{q}_{\alpha_1}$ to get

$$\mathbb{P}\left(y_0 \in \hat{y}_0 + (\hat{q}_{\alpha_1} + \varepsilon, \hat{q}_{\alpha_2} - \varepsilon) \| T_n\right) \leq \widehat{\mathrm{F}}_n^-(\hat{q}_{\alpha_2}) - \widehat{\mathrm{F}}_n(\hat{q}_{\alpha_1}) + 2\ell_n(\varepsilon) \leq (\alpha_2 - \alpha_1) + 2\ell_n(\varepsilon)$$

almost surely.

In order to prove equation (4.6) we use the fact that $PI^+_{\alpha_1,\alpha_2}(0)$, PI_{α_1,α_2} and $PI^-_{\alpha_1,\alpha_2}(0)$ share the same conditional coverage probability given the training data if the conditional distribution of y_0 given x_0 is continuous. This can be seen by the fact that $PI^+_{\alpha_1,\alpha_2}(0) \backslash PI^-_{\alpha_1,\alpha_2}(0) \subseteq \{\hat{q}_{\alpha_1}, \hat{q}_{\alpha_2}\}$, which has Lebesgue-measure zero. Furthermore, we have $PI^-_{\alpha_1,\alpha_2}(0) \subseteq PI_{\alpha_1,\alpha_2} \subseteq PI^+_{\alpha_1,\alpha_2}(0)$. As $\ell_n(0)$ coincides with $\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_\infty$ the statement follows. □

In order to find a uniform bound for the distance of $\widehat{\mathrm{F}}_n$ and F_n for Theorem 4.3, we need a pointwise bound, which is given by the following lemma:

Lemma A.13. *For every $t \in \mathbb{R}$ we have the following pointwise bound:*

$$\begin{aligned}\mathbb{E}((\widehat{\mathrm{F}}_n(t) - \mathrm{F}_n(t))^2) \leq \frac{1}{4(n-1)} + \frac{5}{n}\sum_{i=1}^{n} \mathbb{P}(|\hat{y}_0 - \tilde{y}_0^{[-i]}| > \delta) \\ + 5\mathbb{P}(t - \delta < y_0 - \hat{y}_0 \leq t + \delta),\end{aligned} \tag{A.14}$$

*where $\delta > 0$ can be chosen arbitrarily. Under Assumption **CC1**, we also get the bound*

$$\mathbb{E}((\widehat{\mathrm{F}}_n(t) - \mathrm{F}_n(t))^2) \leq \frac{1}{4(n-1)} + \frac{5}{n}\sum_{i=1}^{n} \mathbb{E}(\min(1, |\hat{y}_0 - \tilde{y}_0^{[-i]}| \| f_{y_0\|x_0}\|_\infty)). \tag{A.15}$$

Proof. By Lemma C.3 of Steinberger and Leeb (2023) we have

$$\mathbb{E}((\widehat{\mathrm{F}}_n(t) - \mathrm{F}_n(t))^2) \leq \frac{1}{4(n-1)} + \frac{5}{n}\sum_{i=1}^{n} \mathbb{E}\left|\mathbb{1}_{(-\infty,t]}(y_0 - \hat{y}_0) - \mathbb{1}_{(-\infty,t]}(y_0 - \tilde{y}_0^{[-i]})\right|.$$

We fix an $i \in \{1, \ldots, n\}$ and rewrite the summand in the last term as

$$\mathbb{P}(y_0 - \hat{y}_0 \leq t < y_0 - \tilde{y}_0^{[-i]}) + \mathbb{P}(y_0 - \tilde{y}_0^{[-i]} \leq t < y_0 - \hat{y}_0),$$

which can be bounded from above by

$$\begin{aligned}&\mathbb{P}(y_0 - \hat{y}_0 \leq t < y_0 - \tilde{y}_0^{[-i]}, 0 \leq \hat{y}_0 - \tilde{y}_0^{[-i]} \leq \delta) + \\ &\mathbb{P}(y_0 - \tilde{y}_0^{[-i]} \leq t < y_0 - \hat{y}_0, 0 \geq \hat{y}_0 - \tilde{y}_0^{[-i]} \geq -\delta) + \mathbb{P}(|\hat{y}_0 - \tilde{y}_0^{[-i]}| > \delta).\end{aligned}$$

The sum of the first two terms in the preceding display can be bounded by

$$\begin{aligned}&\mathbb{P}(y_0 - \hat{y}_0 \leq t < y_0 - \hat{y}_0 + \delta) + \mathbb{P}(y_0 - \hat{y}_0 - \delta \leq t < y_0 - \hat{y}_0) \\ &= \mathbb{P}(t - \delta < y_0 - \hat{y}_0 \leq t) + \mathbb{P}(t < y_0 - \hat{y}_0 \leq t + \delta) = \mathbb{P}(t - \delta < y_0 - \hat{y}_0 \leq t + \delta),\end{aligned}$$

which proves equation (A.14). Equation (A.15) can either be proven by Lemma C.4 of Steinberger and Leeb (2023) or directly by rewriting

$$\mathbb{P}(y_0 - \hat{y}_0 \leq t < y_0 - \tilde{y}_0^{[-i]}) + \mathbb{P}(y_0 - \tilde{y}_0^{[-i]} \leq t < y_0 - \hat{y}_0)$$

as

$$\begin{aligned}&\mathbb{P}(t + \tilde{y}_0^{[-i]} < y_0 \leq t + \hat{y}_0, \tilde{y}_0^{[-i]} < \hat{y}_0) + \mathbb{P}(t + \hat{y}_0 < y_0 \leq t + \tilde{y}_0^{[-i]}, \tilde{y}_0^{[-i]} > \hat{y}_0) \\ &= \mathbb{E}\left(\mathbb{P}\left(t + \tilde{y}_0^{[-i]} < y_0 \leq t + \hat{y}_0, \tilde{y}_0^{[-i]} < \hat{y}_0 \| x_0, T_n\right)\right) \\ &\quad + \mathbb{E}\left(\mathbb{P}\left(t + \hat{y}_0 < y_0 \leq t + \tilde{y}_0^{[-i]}, \tilde{y}_0^{[-i]} > \hat{y}_0 \| x_0, T_n\right)\right),\end{aligned}$$

which can be bounded by

$$\begin{aligned}&\mathbb{E}(\max(0, \min(1, \|f_{y_0\|x_0}\|_\infty(\hat{y}_0 - \tilde{y}_0^{[-i]}))) + \max(0, \min(1, \|f_{y_0\|x_0}\|_\infty(\tilde{y}_0^{[-i]} - \hat{y}_0))) \\ &= \mathbb{E}(\min(1, \|f_{y_0\|x_0}\|_\infty|\hat{y}_0 - \tilde{y}_0^{[-i]}|).\end{aligned}$$

□

We would like to point out that equation (A.15) is a direct consequence of Lemma C.3 and Lemma C.4 in Steinberger and Leeb (2023). However, the novelty of Lemma A.13 is the general case given by equation (A.14). Moreover, the inequalities provided by Lemma A.13 can by changed if we additionally assume the symmetry of the predictor. To be more precise, using Lemma 9 in Bousquet and Elisseeff (2002) instead of Lemma C.3 of Steinberger and Leeb (2023) improves the factor 5 to 3 at the cost of replacing $1/(4(n-1))$ by $1/(2n)$.

In the continuous case a similar statement to Theorem 4.3 can be found in Steinberger and Leeb (2023). However, their proof crucially relies on the Lipschitz-continuity of F_n. The following lemma provides one of the crucial ideas for the proof of Theorem 4.3 in the general case. Together with the definition of $\ell_n(\varepsilon)$ it allows us to draw a link between pointwise and uniform convergence even without Lipschitz-continuity of F_n.

Lemma A.14. *Let $0 < \delta < \varepsilon/4$ and $t_0 < t_1 < \ldots < t_K$, such that $t_j - t_{j-1} = \varepsilon/2$. We then have*

$$\sum_{j=1}^{K} \inf_{t \in [t_{j-1}, t_j]} \mathbb{P}(t - \delta < y_0 - \hat{y}_0 \leq t + \delta) \leq \frac{4\delta}{\varepsilon}.$$

Proof. The proof directly follows after an appropriate application of Lemma A.7 and Lemma A.8. Before using them we need to rewrite the term of interest as follows:

$$\begin{aligned}
&\sum_{j=1}^{K} \inf_{t \in [t_{j-1}, t_j]} \mathbb{P}(t - \delta < y_0 - \hat{y}_0 \leq t + \delta) \\
&= \sum_{j=1}^{K} \inf_{c \in [0, \varepsilon/2]} \mathbb{P}(t_{j-1} + c - \delta < y_0 - \hat{y}_0 \leq t_{j-1} + c + \delta) \\
&\leq \inf_{c \in [0, \varepsilon/2]} \sum_{j=1}^{K} \mathbb{P}(t_{j-1} + c - \delta < y_0 - \hat{y}_0 \leq t_{j-1} + c + \delta).
\end{aligned}$$

Inserting the definition of t_j yields

$$\inf_{c \in [0, \varepsilon/2]} \sum_{j=1}^{K} \mathbb{P}(t_0 + (j-1)\frac{\varepsilon}{2} + c - \delta < y_0 - \hat{y}_0 \leq t_0 + (j-1)\frac{\varepsilon}{2} + c + \delta),$$

which can be bounded from above by

$$\inf_{c \in [0, \varepsilon/2]} \sum_{j \in \mathbb{Z}} \mathbb{P}(t_0 + j\frac{\varepsilon}{2} + c - \delta < y_0 - \hat{y}_0 \leq t_0 + j\frac{\varepsilon}{2} + c + \delta). \tag{A.16}$$

Defining $\gamma := \varepsilon/2$ and $f(x) := \mathbb{P}(t_0 + x - \delta < y_0 - \hat{y}_0 \leq t_0 + x + \delta)$, we can rewrite (A.16) as

$$\inf_{c \in [0, \gamma]} \sum_{j \in \mathbb{Z}} f(j\gamma + c).$$

Now, Lemma A.8 gives the following inequality:

$$\begin{aligned}
&\inf_{c \in [0, \gamma]} \sum_{j \in \mathbb{Z}} f(j\gamma + c) \leq \frac{1}{\gamma} \int_{\mathbb{R}} f(t) d\lambda(t) \\
&= \frac{2}{\varepsilon} \int_{\mathbb{R}} \mathbb{P}(t_0 + t - \delta < y_0 - \hat{y}_0 \leq t_0 + t + \delta) d\lambda(t),
\end{aligned}$$

which can be equivalently written as

$$\frac{2}{\varepsilon}\int_{\mathbb{R}} \mathbb{P}(y_0 - \hat{y}_0 - \delta - t_0 \leq t < y_0 - \hat{y}_0 + \delta - t_0) d\lambda(t).$$

Applying Lemma A.7 with $A := y_0 - \hat{y}_0 - \delta - t_0$ and $B := y_0 - \hat{y}_0 + \delta - t_0$ yields the following upper bound for the last expression:

$$\frac{2}{\varepsilon}\,\mathbb{E}(|B - A|) = \frac{2}{\varepsilon}\,\mathbb{E}(2\delta) = \frac{4\delta}{\varepsilon}.$$

□

Proof of Theorem 4.3: We start proving the continuous case, as the general case uses a similar approach, but comes with some additional technical details. For this, we choose a $\delta > 0$ and define $(t_j)_{j=0}^{K}$ as a set of $K+1 = \lceil \frac{4L}{\delta} \rceil + 1$ points with the following properties: $-2L + \mu = t_0 < t_1 < \ldots < t_K = 2L + \mu$ and $t_j - t_{j-1} \leq \delta$ for all $1 \leq j \leq K$. The Kolmogorov distance $\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_\infty$ between $\widehat{\mathrm{F}}_n$ and F_n may be rewritten as follows:

$$\max\left(\sup_{t\leq t_0}|\widehat{\mathrm{F}}_n(t) - \mathrm{F}_n(t)|, \max_{1\leq j\leq K}\sup_{t_{j-1}\leq t\leq t_j}|\widehat{\mathrm{F}}_n(t) - \mathrm{F}_n(t)|, \sup_{t\geq t_K}|\widehat{\mathrm{F}}_n(t) - \mathrm{F}_n(t)|\right).$$

It is easy to see that $\sup_{t\leq t_0}|\widehat{\mathrm{F}}_n(t) - \mathrm{F}_n(t)| \leq \max(\widehat{\mathrm{F}}_n(t_0), \mathrm{F}_n(t_0))$, which again may be bounded by $\mathrm{F}_n(t_0) + |\widehat{\mathrm{F}}_n(t_0) - \mathrm{F}_n(t_0)|$. Using the equality $|\widehat{\mathrm{F}}_n(t) - \mathrm{F}_n(t)| = |(1 - \mathrm{F}_n(t)) - (1 - \widehat{\mathrm{F}}_n(t))|$ we get with the same argument as before

$$\sup_{t\geq t_K}|\widehat{\mathrm{F}}_n(t) - \mathrm{F}_n(t)| \leq 1 - \mathrm{F}_n(t_K) + |\widehat{\mathrm{F}}_n(t_K) - \mathrm{F}_n(t_K)|.$$

In order to control the supremum inside the interval $[-2L + \mu, 2L + \mu]$, we get the following for the subintervals $[t_{j-1}, t_j]$, where $1 \leq j \leq K$:

$$\begin{aligned}
&\sup_{t_{j-1}\leq t\leq t_j}|\widehat{\mathrm{F}}_n(t) - \mathrm{F}_n(t)| \leq \max(\widehat{\mathrm{F}}_n(t_j) - \mathrm{F}_n(t_{j-1}), \mathrm{F}_n(t_j) - \widehat{\mathrm{F}}_n(t_{j-1}))\\
&= \max(\widehat{\mathrm{F}}_n(t_j) - \mathrm{F}_n(t_j), \mathrm{F}_n(t_{j-1}) - \widehat{\mathrm{F}}_n(t_{j-1})) + \mathrm{F}_n(t_j) - \mathrm{F}_n(t_{j-1})\\
&\leq \max(|\widehat{\mathrm{F}}_n(t_j) - \mathrm{F}_n(t_j)|, |\widehat{\mathrm{F}}_n(t_{j-1}) - \mathrm{F}_n(t_{j-1})|) + \mathrm{F}_n(t_j) - \mathrm{F}_n(t_{j-1}).
\end{aligned}$$

Using the fact that $t_j - t_{j-1} \leq \delta$, we have

$$\begin{aligned}
\mathrm{F}_n(t_j) - \mathrm{F}_n(t_{j-1}) &= \mathbb{P}\left(t_{j-1} < y_0 - \hat{y}_0 \leq t_j \| T_n\right)\\
&= \mathbb{E}\left(\min(1, \mathbb{P}\left(t_{j-1} < y_0 - \hat{y}_0 \leq t_j \| T_n, x_0\right)) \| T_n\right) \text{ a.s.}\\
&\leq \mathbb{E}(\min(1, \delta \| f_{y_0 \| x_0} \|_\infty)) \text{ a.s.}
\end{aligned} \tag{A.17}$$

Putting the pieces together, we end with the following upper bound for $\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_\infty$:

$$1 - \mathrm{F}_n(t_K) + \mathrm{F}_n(t_0) + \max_{0\le j\le K} |\widehat{\mathrm{F}}_n(t_j) - \mathrm{F}_n(t_j)| + \mathbb{E}(\min(1, \delta\|f_{y_0\|x_0}\|_\infty)).$$

As $\|x\|_\infty \le \|x\|_2 \le \|x\|_1$ for any $x \in \mathbb{R}^{K+1}$, we may bound the last display by

$$\mathbb{E}(\min(1, \delta\|f_{y_0\|x_0}\|_\infty)) + 1 - (\mathrm{F}_n(t_K) - \mathrm{F}_n(t_0)) + \sqrt{\sum_{j=0}^{K} |\widehat{\mathrm{F}}_n(t_j) - \mathrm{F}_n(t_j)|^2}.$$

Taking the expected value, using Jensen's inequality and applying Lemma A.13 for the pointwise bound yields

$$\begin{aligned}\mathbb{E}(\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_\infty) \le\ & \mathbb{E}(\min(1, \delta\|f_{y_0\|x_0}\|_\infty)) + \mathbb{P}(|y_0 - \hat{y}_0 - \mu| \ge 2L) \\ & + \sqrt{(K+1)\left(\frac{1}{4(n-1)} + \frac{5}{n}\sum_{i=1}^{n} \mathbb{E}\left(\min(1, \|f_{y_0\|x_0}\|_\infty |\hat{y}_0 - \tilde{y}_0^{[-i]}|\right)\right)}.\end{aligned}$$

Now the claim for the continuous case follows by noticing that $K + 1 < 4L/\delta + 2$.

We now prove the general case. W.l.o.g. we may assume that $\delta < \varepsilon/20$ because otherwise the right-hand side is larger than 1 and the inequality becomes trivial as $\ell_n(\varepsilon) \le 1$. Next, we define $t_j = \mu + (2j - K)\varepsilon/4$ for $0 \le j \le K$, implying $t_0 = \mu - K\varepsilon/4$, $t_K = \mu + K\varepsilon/4$ and $t_j - t_{j-1} = \varepsilon/2$. The reason for the specific choice of the distance between the t_j is the following property for every $1 \le j \le K$: For every $t \in [t_{j-1}, t_j]$ the closed ball with radius $\varepsilon/2$ and center t is a superset of $[t_{j-1}, t_j]$, which yields the following crucial property for the general case:

$$\begin{aligned}\sup_{t\in[t_{j-1},t_j]} \inf_{x,y\in K_{\varepsilon/2}(t)} |\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(y)| &\le \sup_{t\in[t_{j-1},t_j]} \inf_{x\in K_{\varepsilon/2}(t)} |\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(x)| \\ &\le \inf_{t\in[t_{j-1},t_j]} |\widehat{\mathrm{F}}_n(t) - \mathrm{F}_n(t)|.\end{aligned} \tag{A.18}$$

This inequality allows us to proceed even for non-continuous functions F_n yielding a comparable statement to (A.17) which allows us to draw a link between uniform and pointwise bounds. Furthermore, for every $x, y \le s$ we have

$$\begin{aligned}|\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(y)| \le \max(\widehat{\mathrm{F}}_n(x), \mathrm{F}_n(y)) &\le \inf_{t\in[s,s+\varepsilon/2]} \max(\widehat{\mathrm{F}}_n(t), \mathrm{F}_n(t)) \\ &\le \inf_{t\in[s,s+\varepsilon/2]} \left(\mathrm{F}_n(t) + |\widehat{\mathrm{F}}_n(t) - \mathrm{F}_n(t)|\right).\end{aligned} \tag{A.19}$$

Applying the inequality above with $s = t_0$ we get

$$\sup_{t<t_0-\varepsilon/2} \inf_{x,y\in K_{\varepsilon/2}(t)} |\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(y)| \le \inf_{t\in[t_0,t_0+\varepsilon/2]} \mathrm{F}_n(t) + |\widehat{\mathrm{F}}_n(t) - \mathrm{F}_n(t)|.$$

We also get the same upper bound for the following expression:

$$\sup_{t\in[t_0-\varepsilon/2,t_0)} \inf_{x,y\in K_{\varepsilon/2}(t)} |\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(y)| \leq |\widehat{\mathrm{F}}_n(t_0) - \mathrm{F}_n(t_0)| \\ \leq \inf_{t\in[t_0,t_0+\varepsilon/2]} \mathrm{F}_n(t) + |\widehat{\mathrm{F}}_n(t) - \mathrm{F}_n(t)|,$$

where we used equation (A.19) with $s = t_0$ in the last inequality. Putting the pieces together and using the monotonicity of $\widehat{\mathrm{F}}_n$, we have

$$\sup_{t<t_0} \inf_{x,y\in K_{\varepsilon/2}(t)} |\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(y)| \leq \mathrm{F}_n(t_0+\varepsilon/2) + \inf_{t\in[t_0,t_0+\varepsilon/2]} |\widehat{\mathrm{F}}_n(t) - \mathrm{F}_n(t)|. \tag{A.20}$$

With the same argument one can show that

$$\sup_{t>t_K} \inf_{x,y\in K_{\varepsilon/2}(t)} |\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(y)| \leq 1 - \mathrm{F}_n(t_K-\varepsilon/2) + \inf_{t\in[t_K-\varepsilon/2,t_K]} |\widehat{\mathrm{F}}_n(t) - \mathrm{F}_n(t)|. \tag{A.21}$$

By definition, we have

$$\ell_n(\varepsilon) = \max\Bigg[\sup_{t<t_0} \inf_{x,y\in K_{\varepsilon/2}(t)} |\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(y)|, \\ \max_{1\leq j\leq K} \sup_{t\in[t_{j-1},t_j]} \inf_{x,y\in K_{\varepsilon/2}(t)} |\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(y)|, \\ \sup_{t>t_K} \inf_{x,y\in K_{\varepsilon/2}(t)} |\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(y)|\Bigg].$$

Using (A.20) and (A.21), the expression in the preceding display can be bounded by

$$\mathrm{F}_n(t_0+\varepsilon/2) + 1 - \mathrm{F}_n(t_K-\varepsilon/2) + \max_{1\leq j\leq K} \sup_{t\in[t_{j-1},t_j]} \inf_{x,y\in K_{\varepsilon/2}(t)} |\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(y)|.$$

Remembering our crucial property (A.18) from above, the last term can be bounded from above by

$$\max_{1\leq j\leq K} \inf_{t\in[t_{j-1},t_j]} |\widehat{\mathrm{F}}_n(t) - \mathrm{F}_n(t)| \leq \left(\sum_{j=1}^{K} \inf_{t\in[t_{j-1},t_j]} (\widehat{\mathrm{F}}_n(t) - \mathrm{F}_n(t))^2\right)^{\frac{1}{2}}.$$

Taking the expectation with respect to the training data, using Jensen's inequality and the fact that the expectation of the infimum can be bounded by the infimum of the expectation, we end up with

$$\mathbb{E}(\ell_n(\varepsilon)) \leq \mathbb{P}\left(|y_0 - \hat{y}_0 - \mu| \geq \frac{(K-2)\varepsilon}{4}\right) + \left[\sum_{j=1}^{K} \inf_{t\in[t_{j-1},t_j]} \mathbb{E}\left(\widehat{\mathrm{F}}_n(t) - \mathrm{F}_n(t))^2\right)\right]^{\frac{1}{2}}.$$

Using Lemma A.13, the expression inside the square root is bounded by

$$\frac{K}{4(n-1)} + \frac{5K}{n}\sum_{i=1}^{n}\mathbb{P}(|\hat{y}_0 - \tilde{y}_0^{[-i]}| > \delta) + 5\sum_{j=1}^{K}\inf_{t\in[t_{j-1},t_j]}\mathbb{P}(t-\delta < y_0 - \hat{y}_0 \leq t+\delta).$$

Remembering that $\delta < \varepsilon/20 < \varepsilon/4$ holds true, we can apply Lemma A.14, which yields the desired upper bound to complete the first part of the proof. □

Proof of Theorem 4.5: We start with the continuous case. The idea will be the following: For every $n \in \mathbb{N}$ we define an L_n and a suitable δ_n, such that $\lim_{n\to\infty} L_n/v_n = \infty$, $\lim_{n\to\infty} \delta_n/v_n = 0$ and apply the continuous case of Theorem 4.3 with $\mu = 0$. For this, we will denote

$$\max\left(\frac{1}{n}, \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\min(1, \|f_{y_0\|x_0}\|_\infty |\hat{y}_0 - \tilde{y}_0^{[-i]}|))\right)$$

with γ_n. As $v_n\|f_{y_0\|x_0}\|_\infty$ is bounded in probability and $\hat{y}_0/v_n$ is asymptotically stable, we have $\lim_{n\to\infty}\gamma_n = 0$. We define $L_n = v_n\gamma_n^{-1/4}$ and $\delta_n = v_n\gamma_n^{1/4}$, which by dominated convergence implies $\lim_{n\to\infty}\mathbb{P}(|y_0 - \hat{y}_0| > 2v_n\gamma_n^{-1/4}) = 0$ due to the boundedness in probability of $|y_0 - \hat{y}_0|/v_n$. Furthermore, the term $\delta_n\|f_{y_0\|x_0}\|_\infty$ converges in probability to 0, which gives $\lim_{n\to\infty}\mathbb{E}(\min(1, \delta_n\|f_{y_0\|x_0}\|_\infty)) = 0$. To complete the proof we note that $L_n/\delta_n = \gamma_n^{-1/2}$. Thus,

$$\lim_{n\to\infty}\underbrace{\left(\frac{4L_n}{\delta_n} + 2\right)}_{\mathcal{O}_p(\gamma_n^{-1/2})}\underbrace{\left(\frac{1}{4(n-1)} + \frac{5}{n}\sum_{i=1}^{n}\mathbb{E}(\min(1, \|f_{y_0\|x_0}\|_\infty|\hat{y}_0 - \tilde{y}_0^{[-i]}|))\right)}_{\mathcal{O}_p(\gamma_n)} = 0.$$

For the general case, we will define sequences $(\varepsilon_n)_{n\in\mathbb{N}}, (\delta_n)_{n\in\mathbb{N}}$ and $(K_n)_{n\in\mathbb{N}}$ and use the inequality of Theorem 4.3 for each n with $\mu = 0$ and ε replaced by $\varepsilon_n v_n$:

$$\begin{aligned}\mathbb{E}\left(\ell_n(\varepsilon_n v_n)\right) \leq{}& \mathbb{P}\left(\frac{|y_0 - \hat{y}_0|}{v_n} \geq \frac{(K_n - 2)\varepsilon_n}{4}\right)\\ &+ \left(\frac{K_n}{4(n-1)} + \frac{5K_n}{n}\sum_{i=1}^{n}\mathbb{P}\left(|\hat{y}_0 - \tilde{y}_0^{[-i]}| > \delta_n\right) + \frac{20\delta_n}{\varepsilon_n v_n}\right)^{\frac{1}{2}}.\end{aligned}$$

The proof is complete if we can show that every term on the right-hand side converges to 0. For this, we define the function $S_n(x) := \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(|\hat{y}_0 - \tilde{y}_0^{[-i]}|/v_n > x)$. By the asymptotic stability we have $\lim_{n\to\infty} S_n(x) = 0$ for every $x > 0$, which allows us to apply Lemma A.5 to get a null-sequence $(c_n)_{n\in\mathbb{N}}$, such that $\lim_{n\to\infty} q_n = 0$, where

$$q_n := \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(|\hat{y}_0 - \tilde{y}_0^{[-i]}| > c_n v_n) + 1/n.$$

Now we define $\delta_n = c_n v_n$, $K_n = \lceil q_n^{-3/4} \rceil$ and $\varepsilon_n = \max(\sqrt{q_n}, \sqrt{c_n})$. Thus, we have $\varepsilon_n \to 0$, $K_n/n \to 0$ and $\delta_n/(v_n\varepsilon_n) \to 0$. Furthermore, $K_n\varepsilon_n \geq q_n^{-1/4} \to \infty$, which together with the boundedness in probability of $|y_0 - \hat{y}_0|/v_n$ implies

$$\lim_{n\to\infty} \mathbb{P}(|y_0 - \hat{y}_0|/v_n \geq (K_n - 2)\varepsilon_n/4) = 0.$$

The proof is complete by noticing that

$$\begin{aligned}\lim_{n\to\infty} \frac{K_n}{n} \sum_{i=1}^{n} \mathbb{P}(|\hat{y}_0 - \tilde{y}_0^{[-i]}| > \delta_n) &\leq \lim_{n\to\infty} (q_n^{-3/4} + 1) \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{P}(|\hat{y}_0 - \tilde{y}_0^{[-i]}| > c_n v_n) \right) \\ &\leq \lim_{n\to\infty} q_n^{1/4} + q_n = 0.\end{aligned}$$

As $\ell_n(\cdot)$ is non-increasing and $\varepsilon_n \to 0$, we also get that for every fixed $\varepsilon > 0$ the result $\lim_{n\to\infty} \mathbb{E}(\ell_n(\varepsilon v_n)) = 0$ holds true. □

Proof of Corollary 4.6: Define $q_n(T_n) := \mathbb{P}\left(y_0 \in PI^{+}_{\alpha_1,\alpha_2}(\varepsilon) \| T_n\right)$. In view of Proposition 4.1 we conclude that $q_n(T_n) \leq \alpha_2 - \alpha_1 - \delta$ implies $2\ell_n(\varepsilon) \geq \delta$. Since Theorem 4.5 applied with the sequence $v_n \equiv 1$ yields the convergence in $\mathcal{L}_1$ and thus in probability of $\ell_n(\varepsilon)$ to 0, we conclude

$$\lim_{n\to\infty} \mathbb{P}\left(q_n(T_n) \leq \alpha_2 - \alpha_1 - \delta\right) \leq \lim_{n\to\infty} \mathbb{P}\left(2\ell_n(\varepsilon) \geq \delta\right) = 0.$$

The statement for $PI^{-}_{\alpha_1,\alpha_2}(\varepsilon)$ can be proven analogously and the continuous case is also a straightforward combination of Proposition 4.1 and Theorem 4.5. □

Proof of Proposition 4.7: Proposition 4.7 is a consequence of equation (3.9) in Lemma 3.3. However, Lemma 3.3 is only applicable for non-decreasing, bounded, Lipschitz-continuous functions, which prevents us from applying it directly to the function $f(x) = |x|^j$.

To fix this, we start with an $M > 0$ and define two functions $f^{+}(x) = \min(M^j, |x|^j)\mathbb{1}_{[0,\infty)}(x)$ and its analogue on the left half of the real line $f^{-}(x) = -\min(M^j, |x|^j)\mathbb{1}_{(-\infty,0)}(x)$. Thus, we have $f(x) = f^{+}(x) - f^{-}(x) + r_M(x)$, where $r_M(x) = (|x|^j - M^j)\mathbb{1}_{[M,\infty)}(|x|)$. By the uniform integrability we have

$$\begin{aligned}&\lim_{M\to\infty} \sup_{n\in\mathbb{N}} \mathbb{E}(r_M(y_0 - \hat{y}_0)) = 0 \text{ as well as} \\ &\lim_{M\to\infty} \sup_{n\in\mathbb{N}} \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} r_M(y_i - \tilde{y}_i^{[-i]}) \right) = \lim_{M\to\infty} \sup_{n\in\mathbb{N}} \mathbb{E}(r_M(y_0 - \tilde{y}_0)) = 0.\end{aligned}$$

The equation in the last line of the preceding display holds true because the distribution of $y_i - \tilde{y}_i^{[-i]}$ coincides with that one of $y_0 - \tilde{y}_0^{[-i]}$, which itself has the same distribution as $y_0 - \tilde{y}_0$.

Now, the function $|x|^j$ is Lipschitz-continuous with constant jM^{j-1} on the interval

$[-M, M]$, which implies the Lipschitz-continuity for f^+:

$$|f^+(x) - f^+(y)| \leq |x - y| j M^{j-1} \text{ for all } x, y \in \mathbb{R}.$$

With the same argument we also get the Lipschitz-continuity of f^- with the same constant jM^{j-1}. Moreover, f^+ and f^- are non-decreasing and bounded functions, which allows us to apply Lemma 3.3. Now, we fix the training data T_n and let Y denote a random variable uniformly distributed on the points $\{\hat{u}_1, \ldots, \hat{u}_n\}$, in the sense that at every point x it possesses a point mass of $|j : \hat{u}_j = x|/n$. Thus, the distribution function of Y coincides with $\widehat{\mathrm{F}}_n$. Furthermore, let X be a random variable distributed like $y_0 - \hat{y}_0$ conditional on the training data T_n. To put it in other words, the distribution function of X is F_n. Then, for every $\varepsilon > 0$ equation (3.9) in Lemma 3.3 yields

$$\left|\mathbb{E}\left(f^+(y_0 - \hat{y}_0) \| T_n\right) - \mathbb{E}(f^+(Y))\right| \leq \varepsilon j M^{j-1} + M^j \ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n).$$

As $\mathbb{E}(f^+(Y))$ coincides with $\frac{1}{n}\sum_{i=1}^n f^+(\hat{u}_i)$, this gives

$$\left|\mathbb{E}\left(f^+(y_0 - \hat{y}_0) \| T_n\right) - \frac{1}{n}\sum_{i=1}^n f^+(\hat{u}_i)\right| \leq \varepsilon j M^{j-1} + M^j \ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n)$$

and the same holds true if we replace f^+ by f^-.

Putting the pieces together and taking the expectation with respect to the training data, we end up with

$$\begin{aligned}\mathbb{E}\left|\mathbb{E}\left(|y_0 - \hat{y}_0|^j \| T_n\right) - \frac{1}{n}\sum_{i=1}^n |\hat{u}_i|^j\right| \leq\ & 2j\varepsilon M^{j-1} + 2M^j\, \mathbb{E}(\ell_\varepsilon(\mathrm{F}_n, \widehat{\mathrm{F}}_n)) + \\ & \sup_{n\in\mathbb{N}} \mathbb{E}(|r_M(y_0 - \hat{y}_0)|) + \sup_{n\in\mathbb{N}} \mathbb{E}(|r_M(y_0 - \tilde{y}_0)|).\end{aligned} \tag{A.22}$$

To finish the proof we have to find sequences $(\varepsilon_n)_{n\in\mathbb{N}}$ and $(M_n)_{n\in\mathbb{N}}$ such that the right-hand side of equation (A.22) converges to 0 for $n \to \infty$. Now, $|y_0 - \hat{y}_0|$ is bounded in probability because its expectation is uniformly bounded (as it is uniformly integrable). Furthermore, the predictor is asymptotically stable. Hence, the assumptions of Theorem 4.5 are fulfilled and we can find a null-sequence $(\varepsilon_n)_{n\in\mathbb{N}}$, such that $\mathbb{E}(\ell_{\varepsilon_n}(\mathrm{F}_n, \widehat{\mathrm{F}}_n)) \to 0$. Now we can choose

$$M_n = \left(\frac{1}{n} + \varepsilon_n + \mathbb{E}(\ell_{\varepsilon_n}(\widehat{\mathrm{F}}_n, \mathrm{F}_n))\right)^{-\frac{1}{2j}}.$$

Thus, we have $\lim_{n\to\infty} M_n^{j-1}\varepsilon_n = 0$ as well as $\lim_{n\to\infty} \mathbb{E}(\ell_{\varepsilon_n}(\widehat{\mathrm{F}}_n, \mathrm{F}_n))M_n^j = 0$, while at the same time $\lim_{n\to\infty} M_n = \infty$ holds true. Hence, all terms on the right-hand side of equation (A.22) vanish asymptotically. □

A.4. Proofs of Chapter 5

A.4.1. Proofs concerning random matrices

Before we start proving Lemma 5.3 we would like to point out that in fact we are dealing with sequences of random vectors $z_0 = \Sigma^{-1/2}x_0$ changing their distribution and dimension over n. We decided to omit the dependence on n in the notation for an improved readability. As in that proof we have to take this dependence into account, we will adapt the notation for the following subsection and denote the p-dimensional random vector z in the setting of n observations with z_n instead of z.

The proof for the eigenvalues of $X'X$ and XX' mainly relies on the results of Chafaï and Tikhomirov (2018), which use the concept of the *Weak Tail Projection property* given in Definition 6.9.

Chafaï and Tikhomirov (2018) show that a sequence of isotropic random vectors $(v_p)_{p\in\mathbb{N}}$ fulfills the *WTP* property if the v_p are log-concave for all p or consist of i.i.d. components whose distribution does not change with p. The following lemma presents another condition implying the *WTP* property.

Lemma A.15. *Let v_p consist of independent mean-zero components $v_{p,i}$ with unit variance for every $p \in \mathbb{N}$ and $1 \leq i \leq p$ and assume $\sup_{p\in\mathbb{N},1\leq i\leq p} \mathbb{E}(|v_{p,i}|^{2+\varepsilon}) < \infty$ for $\varepsilon > 0$. Then $(v_p)_{p\in\mathbb{N}}$ fulfills the* WTP *property.*

Proof. As uniformly bounded $(1+\varepsilon/2)$-th moments imply uniform integrability of the random variables (WTP-a) will be fulfilled. The proof of (WTP-b) relies on the decomposition of the projection matrix P into a diagonal matrix P_D and the remaining $P_O = P - P_D$ together with decoupling, an idea which can be found for example in the proof of Proposition 1.3. in Srivastava and Vershynin (2013). We will now proof both parts in detail:

Proof of (WTP-a): We start showing that there exists a constant C_1 independent from $p \in \mathbb{N}$ and $y \in S^{p-1}$, such that the $(2+\varepsilon)$-th moment of $v_p'y$ is bounded by C_1. For this we write $y = (y_1, \ldots, y_p)'$ and notice that $\|y\|_2^2 = 1$. We start with the trivial equality

$$\left(\mathbb{E}\left(|v_p'y|^{2+\varepsilon}\right)\right)^{\frac{2}{2+\varepsilon}} = \left\|v_p'y\right\|_{L_{2+\varepsilon}}^2 = \left\|\sum_{i=1}^{p} v_{p,i}y_i\right\|_{L_{2+\varepsilon}}^2.$$

Since $v_{p,i}$ are independent, mean-zero random variables in $L_{2+\varepsilon}$ we can use the Burkholder-type inequality of Lemma A.11 (see also Lemma 1 in Wu and Shao 2007) to bound the expression in the preceding display by

$$C_{2+\varepsilon}^2 \sum_{i=1}^{p} \|y_i v_{p,i}\|_{L_{2+\varepsilon}}^2 \leq C_{2+\varepsilon}^2 K^{\frac{2}{2+\varepsilon}} \sum_{i=1}^{p} y_i^2 = C_{2+\varepsilon}^2 K^{\frac{2}{2+\varepsilon}},$$

where K is an upper bound for $\sup_{p\in\mathbb{N},1\leq i\leq p} \mathbb{E}(|v_{p,i}|^{2+\varepsilon})$ and $C_{2+\varepsilon}$ is a constant only depending on $\varepsilon > 0$. As uniformly bounded $(2+\varepsilon)$-th moments imply uniform integrability, (WTP-a) is fulfilled.

Proof of (WTP-b): While the main idea of the proof originates in Srivastava and Vershynin (2013), we will follow Chafaï and Tikhomirov (2018) (see the proof of Proposition 1.4 therein) closely in this part. Let $P \in \mathbb{R}^{p\times p}$ be an orthogonal projection matrix with rank $r > 0$ and denote $P_D = \operatorname{diag}(P) \in \mathbb{R}^{p\times p}$ its diagonal and $P_O = P - P_D$ the off-diagonal part. To control the tails of $\|Pv_p\|_2^2$ uniformly in p, we will use a decoupling technique from Vershynin (2018) to show that $\mathbb{E}((v_p' P_O v_p)^2) \leq 64r$: For this, let $\tilde{v}_p$ denote an independent copy of v_p. As v_p consists of independent mean-zero components and P_O is a diagonal-free matrix, we have for every convex function F the inequality $\mathbb{E}\left(F(v_p' P_O v_p)\right) \leq \mathbb{E}\left(F(4\tilde{v}_p' P_O v_p)\right)$ (see Theorem 6.1.1. in Vershynin 2018). Applying this to the function $F(x) = x^2$ yields:

$$\mathbb{E}\left((v_p' P_O v_p)^2\right) \leq 16\,\mathbb{E}\left((\tilde{v}_p' P_O v_p)^2\right) = 16\,\mathrm{tr}\,(P_O P_O).$$

We now proceed with

$$\mathrm{tr}\,(P_O P_O) = \|P_O\|_F^2 = \sum_{i=1}^{n}\sum_{j\neq i}(P_{ij})^2 \leq \sum_{i=1}^{n}\sum_{j=1}^{m}(P_{ij})^2 = \|P\|_F^2 = r.$$

Hence, we can use Chebyshev's inequality to bound the tails uniformly by

$$\mathbb{P}\left(v_p' P_O v_p \geq r^{\frac{3}{4}}\right) \leq \mathbb{E}((v_p' P_O v_p)^2) r^{-\frac{3}{2}} \leq \frac{16}{\sqrt{r}}.$$

In the next step we will bound the tails of $v_p' P_D v_p - r$. In order to do so, we rewrite the term of interest as

$$v_p' P_D v_p - r = \sum_{i=1}^{p} P_{ii}(v_{p,i}^2 - 1),$$

where we used the fact that P_D is a diagonal matrix with $\mathrm{tr}\,P_D = \mathrm{tr}\,P = r$. Furthermore, the random vector $(P_{ii}(v_{p,i}^2 - 1))_{i=1}^p$ consists of independent components with mean zero and finite $(1+\varepsilon/2)$-th moments, which allows us to apply the Burkholder-type inequality of Lemma A.11 (see also Lemma 1 in Wu and Shao 2007) to get

$$\begin{aligned}\mathbb{E}\left|\sum_{i=1}^{p} P_{ii}(v_{p,i}^2 - 1)\right|^{1+\varepsilon/2} &\leq C_{1+\varepsilon/2}^{1+\varepsilon/2}\sum_{i=1}^{p}\mathbb{E}\left|P_{ii}(v_{p,i}^2 - 1)\right|^{1+\varepsilon/2} \\ &\leq C_{1+\varepsilon/2}^{1+\varepsilon/2}\sum_{i=1}^{p} P_{ii}^{1+\varepsilon/2}\,\mathbb{E}\left|(v_{p,i}^2 - 1)\right|^{1+\varepsilon/2}.\end{aligned}$$

As $0 \leq P_{ii} \leq 1$ we can bound the last term from above by

$$C_{1+\varepsilon/2}^{1+\varepsilon/2}\sum_{i=1}^{p} P_{ii}\tilde{K} = C_{1+\varepsilon/2}^{1+\varepsilon/2}\tilde{K}r,$$

where $\tilde{K}$ is a uniform upper bound for $\mathbb{E}\left|(v_{p,i}^2-1)\right|^{1+\varepsilon/2}$ and $C_{1+\varepsilon/2}$ is a constant only depending on $\varepsilon > 0$. We stress the fact that $\tilde{K}$ is finite by Minkowski's inequality together with the fact that the $(2+\varepsilon)$-th moments of $v_{p,i}$ are uniformly bounded. Thus, applying Markov's inequality we can uniformly bound the tails:

$$\mathbb{P}\left(v_p' P_D v_p - r \geq r^{\frac{4}{4+\varepsilon}}\right) \leq \mathbb{E}\left|\sum_{i=1}^{p} P_{ii}(v_{p,i}^2-1)\right|^{1+\varepsilon/2} r^{-\frac{4+2\varepsilon}{4+\varepsilon}} \leq C_{1+\varepsilon/2}^{1+\varepsilon/2}\tilde{K} r^{-\frac{\varepsilon}{4+\varepsilon}}.$$

To finish the proof note that $\|Pv_p\|_2^2 = v_p' P_O v_p + v_p' P_D v_p$ and set $f(r) = r^{-1/4} + r^{-\varepsilon/(4+\varepsilon)}$, $g(r) = C_{1+\varepsilon/2}^{1+\varepsilon/2}\tilde{K} r^{-\varepsilon/(4+\varepsilon)} + 64/\sqrt{r}$. □

Proof of Lemma 5.3: The proof relies on the results of Chafaï and Tikhomirov (2018). In order to apply them we have to show that $z_0 = \Sigma^{-1/2}x_0$ fulfills the *Weak Tail Projection* property given in Definition 6.9.

We start with the low-dimensional case: Chafaï and Tikhomirov (2018) showed that a sequence of mean-zero and isotropic random vectors $(z_n)_{n\in\mathbb{N}}$ fulfills the *Weak Tail Projection* property if z_n is log-concave for every n or z_n consists of *i.i.d.* components, where in the latter case the distribution is not allowed to vary over n (cf. Proposition 1.3 and Proposition 1.4 therein). Moreover, we proved in Lemma A.15 that the same holds true if the components are independent with uniformly bounded $(2+\varepsilon)$-th moments.

Hence, the vector $z_0 = \Sigma^{-1/2}x_0$ fulfills the *Weak Tail Projection* property if the assumptions of Definition 5.1 hold true. At this point we distinguish the case $\rho = 0$ and the case $\rho \in (0,1)$: In the latter case, we can apply Theorem 1.6 in Chafaï and Tikhomirov (2018) to the matrix $S_n := \Sigma^{-1/2}X'X\Sigma^{-1/2}/n$ directly, which yields $\lambda_{\min}(S_n) \xrightarrow{p} (1-\sqrt{\rho})^2$.

The case $\rho = 0$ will be proven by an idea which squeezes the smallest eigenvalue between $(1-\sqrt{\varepsilon})^2$ and 1, using the foregoing result in the case $\rho = \varepsilon > 0$. Our main argument will be the following: For every $\varepsilon \in (0,1)$ we can embed S_n in another matrix $S_{n,\varepsilon}$ of dimension $\lceil n\varepsilon\rceil \times \lceil n\varepsilon\rceil$, such that the upper left block of $S_{n,\varepsilon}$ is exactly S_n. Hence, the smallest eigenvalue of $S_{n,\varepsilon}$ is not larger than the one of S_n. Moreover, the matrix $S_{n,\varepsilon}$ will be designed in such a way, that it fulfils the assumptions to ensure that its smallest eigenvalue converges to $(1-\sqrt{\varepsilon})^2$ in probability. Furthermore, by isotropy we know that $\mathbb{E}(\lambda_{\min}(S_n)) \leq 1$. As ε can be made arbitrarily small, we can apply Lemma A.9 with X_n and K replaced by $\lambda_{\min}(S_n)$ and 1 to conclude $\lambda_{\min}(S_n) \xrightarrow{p} 1$.

To make this proof rigorous, we fix an $\varepsilon \in (0,1)$. As $\lim_{n\to\infty} p/n = 0$, we can find an N_ε, such that $n\varepsilon > p$ for all $n \geq N_\varepsilon$. For every such $n \geq N_\varepsilon$, we extend our p-dimensional random vector z to a larger one $(z', z_\varepsilon')'$, where z_ε is an $(\lceil n\varepsilon\rceil - p)$-dimensional random vector independent of z. The components $z_{\varepsilon,i}$ are i.i.d. mean-zero random variables with unit variance and fulfill the same assumptions as z: If the components of z are independent with finite $(2+\delta)$-th moments uniformly bounded by some constant K, then so are the components of $z_{\varepsilon,i}$. If the components are i.i.d. distributed like ξ_n, then so are the components of $z_{\varepsilon,i}$. Lastly, if the vector z is log-concave, then we define the

components of $z_{\varepsilon,i}$ as log-concave random variables as well (e.g. standard Gaussian). We stress the fact that the product of two log-concave probability measures is again log-concave (cf. Saumard and Wellner 2014). Hence, by the independence of z and z_ε (and the independence of the components of z_ε), we get that the stacked vector (z, z_ε) is again log-concave. To sum it up, the vector $(z', z_\varepsilon')'$ fulfils the same properties as z, which implies that it fulfils the *Weak Tail Projection* property (see the beginning of the proof).

Write $S_{n,\varepsilon} := Z_\varepsilon' Z_\varepsilon / n$, where Z_ε is a random matrix consisting of n rows, which are i.i.d copies of $(z', z_\varepsilon')'$. Then the upper left $p \times p$ dimensional block of $S_{n,\varepsilon}$ coincides with $Z'Z/n$, which implies that the smallest eigenvalue of $Z'Z/n$ is bounded from below by the one of $S_{n,\varepsilon}$. This construction can now be done for any n, such that the $\lceil n\varepsilon \rceil \times \lceil n\varepsilon \rceil$ matrix $S_{n,\varepsilon}$ fulfils $\lim_{n\to\infty} \tilde{p}/n = \varepsilon$, where $\tilde{p} = \lceil n\varepsilon \rceil$ is the dimension of $S_{n,\varepsilon}$. Hence, the smallest eigenvalue of $S_{n,\varepsilon}$ converges in probability to $(1 - \sqrt{\varepsilon})^2$. Recalling the block structure of $S_{n,\varepsilon}$ and the fact that $\varepsilon \in (0,1)$ was arbitrary, we conclude for the smallest eigenvalue of $Z'Z/n$:

$$\lim_{n\to\infty} \mathbb{P}\left(\lambda_{\min}\left(\frac{Z'Z}{n} \right) \leq (1 - \sqrt{\varepsilon})^2 - \varepsilon \right) = 0 \text{ for all } \varepsilon \in (0,1).$$

On the other hand, we can bound $\mathbb{E}(\min_{v \in S^{p-1}} v'Z'Zv/n)$ from above by the term $\min_{v \in S^{p-1}} \mathbb{E}(v'Z'Zv/n)$, where S^{p-1} denotes the unit sphere in $\mathbb{R}^p$. Since z is an isotropic vector, we conclude $\min_{v \in S^{p-1}} \mathbb{E}(v'Z'Zv/n) = 1$. Using Lemma A.9 with X_n and K replaced by $\lambda_{\min}(Z'Z/n)$ and 1 together with the fact that ε can be made arbitrarily small, this yields $\lambda_{\min}(Z'Z/n) \xrightarrow{p} 1$.

For the high-dimensional case we are facing the problem that now the *columns* of $X\Sigma^{-1/2}$ have to be independent and identically distributed. However, this is solved by the additional assumption that $\Sigma^{-1/2}x_0$ consists of i.i.d. components. Now, we can proceed similarly to the low-dimensional case if we replace the matrix $X\Sigma^{-1/2}$ by $\Sigma^{-1/2}X'$. It remains to show that the rows of $\Sigma^{-1/2}X'$ fulfill the Weak Tail Projection property. If either ξ_n does not vary over n or the $(2+\delta)$-th moments of the components are uniformly bounded, this is automatically fulfilled (cf. Lemma A.15 in the latter case). If ξ_n is log-concave, then this property transfers to any vector consisting of n i.i.d. components distributed like ξ_n. Thus, in any of the three cases of Definition 5.2 the Weak Tail Projection property is fulfilled. Starting with $\rho \in (1, \infty)$ we can proceed similar to the low-dimensional case with the only difference that p and n change their roles, implying that the smallest eigenvalue of $X\Sigma^{-1}X'/p$ converges to $(1 - \sqrt{r})^2$ in probability, where $r := \lim_{n\to\infty} n/p = 1/\rho$. The case $\rho = \infty$ coincides with $r = 0$ and can be treated analogously to the case $\rho = 0$ in the low-dimensional setting. Hence, we have

$$\lambda_{\min}\left(\frac{X\Sigma^{-1}X'}{p} \right) \xrightarrow{p} \begin{cases} (1 - \rho^{-\frac{1}{2}})^2 & \text{if } \rho \in (1, \infty) \\ 1 & \text{if } \rho = \infty. \end{cases}$$

However, our aim is to deal with the smallest eigenvalue of $X\Sigma^{-1}X'/n = (p/n)(X\Sigma^{-1}X'/p)$

rather than the one of $X\Sigma^{-1}X'/p$. As $\lim_{n\to\infty} p/n = \rho$, we get

$$\lambda_{\min}\left(\frac{X\Sigma^{-1}X'}{n}\right) \xrightarrow{p} (\sqrt{\rho}-1)^2$$

whenever ρ is finite. If $\rho = \infty$, then the smallest eigenvalue of $X\Sigma^{-1}X'/n$ converges to ∞ in the sense that

$$\lim_{n\to\infty} \mathbb{P}\left(\lambda_{\min}\left(\frac{X\Sigma^{-1}X'}{n}\right) \leq M\right) = 0$$

for all $M > 0$. □

Until now, we have put assumptions on $z_0 = \Sigma^{-1/2}x_0$. However, we can extend our results to the general case where $x_0 = R_n z_0$ for a z_0 fulfilling the assumptions in Definition 5.1 or Definition 5.2 and a matrix R_n with $R_n R_n' = \Sigma > 0$. Moreover, the dimension of z_0 can even exceed the dimension x_0 as the following lemma shows:

Lemma A.16. *Assume $(z_n)_{n\in\mathbb{N}}$ is a sequence of d_n-dimensional, mean-zero and isotropic random vectors fulfilling the Weak Tail Projection property. Let $R_n \in \mathbb{R}^{p_n\times d_n}$ such that $p_n \leq d_n$ and $x_n = R_n z_n$ with $R_n R_n' = \Sigma > 0$. Then, $(\Sigma^{-1/2}x_n)_{n\in\mathbb{N}}$ also fulfills the Weak Tail Projection property. Moreover, the smallest eigenvalue of the matrix $\Sigma^{-1/2}X'X\Sigma^{-1/2}/n$ converges to $(1-\sqrt{\rho})^2$ in probability if $\rho := \lim_{n\to\infty} p_n/n \in [0,1)$.*

Proof. The second part directly follows from Theorem 1.6 in Chafaï and Tikhomirov (2018) if $\rho \in (0,1)$. The case $\rho = 0$ can be proven analogously to the proof of Lemma 5.3. The crucial part is only to show that the Weak Tail Projection property transfers from z_0 to $\Sigma^{-1/2}R_n z_0$. Denoting $Q := R_n'\Sigma^{-1/2} \in \mathbb{R}^{d_n\times p_n}$ we start with the following observation: From $Q'Q = I_{p_n}$ we have $\mathbb{E}(Q'z_n) = Q'\,\mathbb{E}(z_n) = 0$, $\mathbb{E}(Q'z_n z_n'Q) = Q'Q = I_{p_n}$. Now, for any orthogonal projection matrix $P \in \mathbb{R}^{p_n\times p_n}$ with rank $r > 0$ we have $\|PQ'z_n\|_2^2 = z_n'QPPQ'z_n = \|QPQ'z_n\|_2^2$. However, $A = QPQ' \in \mathbb{R}^{d_n\times d_n}$ is symmetric and idempotent, hence itself an orthogonal projection. Recalling the fact that $Q'Q = I_{p_n}$, we have

$$\operatorname{rank}(A) = \operatorname{tr}(A) = \operatorname{tr}(QPQ') = \operatorname{tr}(PQ'Q) = \operatorname{tr}(P) = \operatorname{rank}(P) = r.$$

Hence, the (WTP-b) transfers from $(z_n)_{n\in\mathbb{N}}$ to $(Q'z_n)_{n\in\mathbb{N}}$.

For the (WTP-a) we note that for every $y \in S^{p_n-1}$ we have $\|Qy\|_2 = 1$, implying $Qy \in S^{d_n-1}$ and

$$\begin{aligned}&\sup_{n\in\mathbb{N},y\in S^{p_n-1}} \mathbb{E}\left((y'Q'z_n)^2 \mathbb{1}_{\{(y'Q'z_n)^2\geq M\}}\right)\\ &\leq \sup_{n\in\mathbb{N},\bar{y}\in S^{d_n-1}} \mathbb{E}\left((\bar{y}'z_n)^2 \mathbb{1}_{\{(\bar{y}'z_n)^2\geq M\}}\right) \xrightarrow[M\to\infty]{} 0.\end{aligned}$$

□

A.4.2. Proofs concerning the Ridge estimator

The proofs for the Ridge estimator will repeatedly make use of some properties, which we collect in a separate lemma. In the following lemma we are stating one frequently used upper bound:

Lemma A.17. *Fix a matrix $Z \in \mathbb{R}^{n\times p}$, a positive definite matrix $\Sigma \in \mathbb{R}^{p\times p}$, a penalty parameter $c \geq 0$ and a vector $w \in \mathbb{R}^n$. Define $X = Z\Sigma^{1/2}$. We then have*

$$\|(X'X + cI_p)^\dagger X'w\|_2 \leq \sigma_{\max}(Z^\dagger)\frac{\|w\|_2}{\sigma_{\min}(\Sigma^{1/2})}.$$

Furthermore, if X has full column rank p we have

$$\|\Sigma^{1/2}(X'X + cI_p)^\dagger X'w\|_2 \leq \|w\|_2\sigma_{\max}(Z^\dagger).$$

Proof. Let $s_1 \geq s_2 \geq \ldots \geq s_{\min(n,p)}$ denote the singular values of X and $k = \operatorname{rank}(X)$. If $k = 0$, the terms on the left-hand side of the two statements are 0 and hence the statements are true. Thus, we only have to deal with the case $k > 0$. Then the singular values of $(X'X + cI_p)^\dagger X'$ are given by $s_i/(s_i^2 + c)$ if $s_i > 0$ and 0 else, which can be bounded from above by $1/s_k$. However, by Ostrowski's Theorem for eigenvalues (cf. Higham and Cheng 1998) we have

$$s_k^2 = \lambda_k(X'X) \geq \lambda_{\min}(\Sigma^{1/2})\lambda_k(Z'Z)\lambda_{\min}(\Sigma^{1/2}) = \sigma_k^2(Z)\sigma_{\min}^2(\Sigma^{1/2}),$$

where we assumed the singular values of Z to be ordered in the sense that $\sigma_1(Z) \geq \sigma_2(Z) \geq \ldots \geq \sigma_{\min(n,p)}(Z)$. As Σ is positive definite, the rank of Z coincides with k, which entails that $\sigma_k(Z)$ is the smallest non-zero singular value of Z. Thus, we have

$$\frac{1}{s_k} \leq \frac{1}{\sigma_k(Z)}\frac{1}{\sigma_{\min}(\Sigma^{1/2})} = \frac{\sigma_{\max}(Z^\dagger)}{\sigma_{\min}(\Sigma^{1/2})},$$

which proves the first statement.

Furthermore, if X has full column rank p, the matrix $X'X + cI_p$ is invertible. Hence, we have

$$\Sigma^{1/2}(X'X + cI_p)^{-1}X' = (Z'Z + c\Sigma^{-1})^{-1}Z'. \tag{A.23}$$

As Σ is positive definite we conclude that $(Z'Z + c\Sigma^{-1})^{-1}$ is smaller or equal to $(Z'Z)^{-1}$ in the Loewner-ordering. Thus, we have

$$\begin{aligned}
\|\Sigma^{1/2}(X'X + cI_p)^{-1}X'w\|_2^2 &= w'Z(Z'Z + c\Sigma^{-1})^{-2}Z'w \\
&\leq w'Z(Z'Z + c\Sigma^{-1})^{-1}Z'w\lambda_{\max}((Z'Z + c\Sigma^{-1})^{-1}) \\
&\leq w'Z(Z'Z)^{-1}Z'w\lambda_{\max}((Z'Z)^{-1}) \\
&\leq \|w\|_2^2(\sigma_{\max}(Z^\dagger))^2.
\end{aligned}$$

□

The trick in the second part of the foregoing proof is to use the property $(AB)^{-1} = B^{-1}A^{-1}$ in equation (A.23). Without the regularity of A and B one could try to replace the matrix inverse by the Moore-Penrose pseudoinverse. However, the equation $(AB)^\dagger = B^\dagger A^\dagger$ does not always hold true. Thus, we needed a different approach for the first statement of Lemma A.17.

Proof of Theorem 5.5: Before we start, we would like to emphasize that the probability of $c_{n-1,p}$ being 0 vanishes for large n as $c_{n-1,p}/c_{n,p} \xrightarrow{p} 1$. Hence, it will be enough to show that $(y_0 - x_0'\widehat{\beta}_R(c_{n,p}))\mathbb{1}\{c_{n,p} > 0, c_{n-1,p} > 0\}$ is bounded in probability and $(x_0'(\widehat{\beta}_R(c_{n,p}) - \widetilde{\beta}_R(c_{n-1,p})))\mathbb{1}\{c_{n,p} > 0, c_{n-1,p} > 0\}$ converges to 0 in probability. Thus, in the next steps we will implicitly assume that $c_{n,p} > 0, c_{n-1,p} > 0$.

Our arguments substantially rely on the fact that the singular values of $(X'X + c_{n,p}I_p)^\dagger X'$ can be bounded from above by $c_{n,p}^{-1/2}$ for *every* matrix X. To see this, notice that the singular values of $(X'X + c_{n,p}I_p)^\dagger X'$ are given by $s_i/(s_i^2 + c_{n,p})$, where s_i denotes the i-th singular value of X. Now, it is easy to show that $s_i/(s_i^2 + c_{n,p})$ is bounded from above by $(4c_{n,p})^{-1/2}$. Hence, we can bound the singular values of the matrix $(X'X + c_{n,p}I_p)^\dagger X'$ without any assumptions on the singular values of X itself, which yields

$$\|(X'X + c_{n,p}I_p)^\dagger X'w\|_2 \leq \frac{\|w\|_2}{c_{n,p}^{1/2}} \tag{A.24}$$

for every $w \in \mathbb{R}^p$, where we dropped the factor $1/2$ for convenience.

Boundedness of $y_0 - x_0'\widehat{\beta}_R(c_{n,p})$: As the second moments of $y_0^{(n)}$ are bounded over n, the sequence $(y_0^{(n)})_{n\in\mathbb{N}}$ is bounded in probability. Hence, it suffices to bound $\hat{y}_0^2 = (x_0'\widehat{\beta}_R(c_{n,p}))^2$ in probability, which, by Lemma A.6, reduces to the boundedness of its conditional expectation $\|\Sigma^{1/2}\widehat{\beta}_R(c_{n,p})\|_2^2$. Using inequality (A.24) yields

$$\begin{aligned}\|\Sigma^{1/2}\widehat{\beta}_R(c_{n,p})\|_2^2 &= \|\Sigma^{1/2}(X'X + c_{n,p}I_p)^\dagger X'Y\|_2^2 \leq \lambda_{\max}(\Sigma)\|(X'X + c_{n,p}I_p)^\dagger X'Y\|_2^2 \\ &\leq \lambda_{\max}(\Sigma)\frac{\|Y\|_2^2}{c_{n,p}} = \frac{\|Y\|_2^2}{n}\underbrace{\lambda_{\max}(\Sigma)\frac{n}{c_{n,p}}}_{\mathcal{O}_p(1)}.\end{aligned}$$

As the (unconditional) expectation of $\|Y\|_2^2/n$ is bounded over n, it is bounded in probability as well. Thus, $\mathbb{1}\{c_{n,p} > 0, c_{n-1,p} > 0\}\|Y\|_2^2/n$ is also bounded in probability. Now, the boundedness of $\|\Sigma^{1/2}\widehat{\beta}_R(c_{n,p})\|_2^2$ follows as $\lambda_{\max}(\Sigma)\frac{n}{c_{n,p}}$ is bounded by assumption.

To prove the stability we will also need the fact that $y_0 - x_0'\widetilde{\beta}_R(c_{n-1,p})$ is bounded in probability. In order to show this, we use the same argument as before to get

$$\|\Sigma^{1/2}\widetilde{\beta}_R(c_{n-1,p})\|_2^2 \leq \frac{\|\tilde{Y}\|_2^2}{n}\lambda_{\max}(\Sigma)\frac{n}{c_{n-1,p}}.$$

However, as $c_{n,p}/c_{n-1,p}$ converges to 1 in probability, we can conclude that $y_0 -$

$x_0'\widetilde{\beta}_R(c_{n-1,p})$ is bounded in probability.

Convergence of $x_0'(\widehat{\beta}_R(c_{n,p}) - \widetilde{\beta}_R(c_{n-1,p}))$ to 0: For this statement we again use Lemma A.6 and show that $\|\Sigma^{1/2}(\widehat{\beta}_R(c_{n,p}) - \widetilde{\beta}_R(c_{n-1,p}))\|_2$ converges to 0 in probability. For this, we start with the triangle inequality

$$\begin{aligned}\|\Sigma^{1/2}(\widehat{\beta}_R(c_{n,p}) - \widetilde{\beta}_R(c_{n-1,p}))\|_2 \leq \|\Sigma^{1/2}(\widehat{\beta}_R(c_{n,p}) - \widehat{\beta}_R(c_{n-1,p}))\|_2 \\ + \|\Sigma^{1/2}(\widehat{\beta}_R(c_{n-1,p}) - \widetilde{\beta}_R(c_{n-1,p}))\|_2.\end{aligned} \tag{A.25}$$

We will show that each of the two summands converges to 0 in probability and start with the first one

$$\begin{aligned}\widehat{\beta}_R(c_{n,p}) - \widehat{\beta}_R(c_{n-1,p}) &= \left((X'X + c_{n,p}I_p)^\dagger - (X'X + c_{n-1,p}I_p)^\dagger\right) X'Y \\ &= (c_{n-1,p} - c_{n,p})(X'X + c_{n-1,p}I_p)^\dagger(X'X + c_{n,p}I_p)^\dagger X'Y,\end{aligned} \tag{A.26}$$

where the second equality can be derived from the singular value decomposition of X. Thus, we have

$$\begin{aligned}&\|\Sigma^{1/2}(\widehat{\beta}_R(c_{n,p}) - \widehat{\beta}_R(c_{n-1,p}))\|_2 \\ &\leq \lambda_{\max}(\Sigma^{1/2})|c_{n-1,p} - c_{n,p}|\lambda_{\max}((X'X + c_{n-1,p}I_p)^\dagger)\|\widehat{\beta}_R(c_{n,p})\|_2.\end{aligned}$$

As before, inequality (A.24) yields $\|\widehat{\beta}_R(c_{n,p})\|_2^2 \leq \|Y\|_2^2/c_{n,p}$. Furthermore, we trivially have $\lambda_{\max}((X'X + c_{n-1,p}I_p)^\dagger) \leq 1/c_{n-1,p}$, which yields

$$\begin{aligned}\|\Sigma^{1/2}(\widehat{\beta}_R(c_{n,p}) - \widehat{\beta}_R(c_{n-1,p}))\|_2^2 &\leq \frac{(c_{n-1,p} - c_{n,p})^2}{c_{n-1,p}^2}\lambda_{\max}(\Sigma)\frac{\|Y\|_2^2}{c_{n,p}} \\ &= \underbrace{\left(1 - \frac{c_{n,p}}{c_{n-1,p}}\right)^2}_{\xrightarrow{p}0}\underbrace{\lambda_{\max}(\Sigma)\frac{n}{c_{n,p}}}_{\mathcal{O}_p(1)}\underbrace{\frac{\|Y\|_2^2}{n}}_{\mathcal{O}_p(1)} \xrightarrow{p} 0.\end{aligned}$$

To finish the proof, we only have to show that $\|\Sigma^{1/2}(\widehat{\beta}_R(c_{n-1,p}) - \widetilde{\beta}_R(c_{n-1,p}))\|_2 \xrightarrow{p} 0$. For this, we apply Lemma A.2 with $c = c_{n-1,p}$. As we are dealing with the case $c_{n-1,p} > 0$, we have

$$\begin{aligned}&\|\Sigma^{1/2}(\widehat{\beta}_R(c_{n-1,p}) - \widetilde{\beta}_R(c_{n-1,p})\|_2^2 \\ &\leq \lambda_{\max}(\Sigma)\|(X'X + c_{n-1,p})^{-1}X'e_n\|_2^2(y_n - x_n'\widetilde{\beta}_R(c_{n-1,p}))^2 \\ &\leq \lambda_{\max}(\Sigma)\frac{1}{c_{n-1,p}}(y_n - x_n'\widetilde{\beta}_R(c_{n-1,p}))^2,\end{aligned}$$

where we used inequality (A.24) for the last line. Now we want to show that $(y_n - x_n'\widetilde{\beta}_R(c_{n-1,p})\mathbb{1}\{c_{n,p} > 0, c_{n-1,p} > 0\}$ is bounded in probability. For this, it is enough to show that $(y_n - x_n'\widetilde{\beta}_R(c_{n-1,p}))$ is bounded in probability. To do so, we point out that $(y_n, x_n')'$ is independent from $\widetilde{\beta}_R(c_{n-1,p})$ and has the same distribution as $(y_0, x_0')'$.

Thus, $y_n - x_n' \widetilde{\beta}_R(c_{n-1,p})$ is bounded in probability because $y_0 - x_0' \widetilde{\beta}_R(c_{n-1,p})$ is (as shown before). To complete the proof, we notice that

$$\frac{\lambda_{\max}(\Sigma)}{c_{n-1,p}} = \underbrace{\frac{\lambda_{\max}(\Sigma)}{c_{n,p}}}_{\xrightarrow{p} 0} \underbrace{\frac{c_{n,p}}{c_{n-1,p}}}_{\xrightarrow{p} 1}$$

converges to 0 in probability by the assumptions. □

Proof of Proposition 5.6: The proof is a variation of the proof of Theorem 5.5. The two main differences are that now $c_{n,p}$ can be 0 with a positive probability (even asymptotically) and that we replace inequality (A.24) by Lemma A.17. As assumption **LD** is fulfilled, the smallest eigenvalue of $Z'Z/n := \Sigma^{-1/2} X'X \Sigma^{-1/2}/n$ converges to $(1 - \sqrt{\rho})^2$ in probability by Lemma 5.3. As Σ is positive definite, the probability of $X'X$ having full rank p converges to 1 and $\sqrt{n}\sigma_{\max}(Z^\dagger)$ converges to $(1 - \sqrt{\rho})^{-1}$ in probability. We would like to point out that the same holds true if we replace X by its leave-one-out analogue $\tilde{X}$. As the probability of the event where $\tilde{X}$ (and therefore X) has full column rank p converges to 1, it will be enough to show that $(y_0 - x_0' \widehat{\beta}_R(c_{n,p})) \mathbb{1}\{\operatorname{rank}(\tilde{X}) = p\}$ is bounded in probability and $(x_0'(\widehat{\beta}_R(c_{n,p}) - \widetilde{\beta}_R(c_{n-1,p}))) \mathbb{1}\{\operatorname{rank}(\tilde{X}) = p\}$ converges to 0 in probability. Since we are implicitly assuming that both $\tilde{X}$ and X have full column rank p, we are able to apply the second part of Lemma A.17.

Boundedness of $y_0 - x_0' \widehat{\beta}_R(c_{n,p})$: As in the proof of Theorem 5.5, Lemma A.6 implies that it suffices to show that $\|\Sigma^{1/2} \widehat{\beta}_R(c_{n,p})\|_2$ is bounded in probability. By Lemma A.17 we have

$$\|\Sigma^{1/2} \widehat{\beta}_R(c_{n,p})\|_2 \leq \frac{\|Y\|_2}{\sqrt{n}} \sqrt{n} \sigma_{\max}(Z^\dagger),$$

which is bounded in probability as the expected value of $\|Y\|_2^2/n$ is bounded over n. With the same argument one can also show the boundedness in probability of $y_0 - x_0' \widetilde{\beta}_R(c_{n-1,p})$.

Convergence of $x_0'(\widehat{\beta}_R(c_{n,p}) - \widetilde{\beta}_R(c_{n-1,p}))$ to 0: By combining inequality (A.25) with Lemma A.6, it suffices to show that $\|\Sigma^{1/2}(\widehat{\beta}_R(c_{n-1,p}) - \widetilde{\beta}_R(c_{n-1,p}))\|_2$ and $\|\Sigma^{1/2}(\widehat{\beta}_R(c_{n,p}) - \widehat{\beta}_R(c_{n-1,p}))\|_2$ converge to 0 in probability. For the first part we apply Lemma A.2 to $c = c_{n-1,p}$ and get

$$\Sigma^{1/2}(\widehat{\beta}_R(c_{n-1,p}) - \widetilde{\beta}_R(c_{n-1,p})) = \Sigma^{1/2}(X'X + c_{n-1,p} I_p)^{-1} X' e_n (y_n - x_n' \widetilde{\beta}_R(c_{n-1,p})).$$

Thus, Lemma A.17 yields

$$\|\Sigma^{1/2}(X'X + c_{n-1,p} I_p)^{-1} X' e_n (y_n - x_n' \widetilde{\beta}_R(c_{n-1,p}))\|_2 \leq \sigma_{\max}(Z^\dagger) |y_n - x_n' \widetilde{\beta}_R(c_{n-1,p})|,$$

which converges to 0 in probability as the second term is bounded in probability and

$\sigma_{\max}(Z^\dagger) \sim \mathcal{O}_p(n^{-1/2})$. For the second part we again use (A.26) to get

$$\begin{aligned}&\Sigma^{1/2}\left(\widehat{\beta}_R(c_{n,p}) - \widehat{\beta}_R(c_{n-1,p})\right)\\ &= (c_{n-1,p} - c_{n,p})\Sigma^{1/2}(X'X + c_{n-1,p}I_p)^\dagger\Sigma^{1/2}\Sigma^{-1}\Sigma^{1/2}\widehat{\beta}_R(c_{n,p}).\end{aligned}$$

As shown before $\Sigma^{1/2}\widehat{\beta}_R(c_{n,p})$ is bounded in probability. Since $X'X$ has full rank, we have

$$\sigma_{\max}\left(\Sigma^{1/2}(X'X + c_{n-1,p}I_p)^\dagger\Sigma^{1/2}\right) \leq \lambda_{\max}((Z'Z)^{-1}).$$

Altogether, $\|\Sigma^{1/2}(\widehat{\beta}_R(c_{n,p}) - \widehat{\beta}_R(c_{n-1,p}))\|_2$ can be bounded from above by

$$\underbrace{\frac{|c_{n-1,p} - c_{n,p}|}{n}\lambda_{\max}(\Sigma^{-1})}_{\xrightarrow{p}0}\underbrace{\|\Sigma^{1/2}\widehat{\beta}_R(c_{n,p})\|_2}_{\mathcal{O}_p(1)}\underbrace{\lambda_{\max}\left(\left(\frac{Z'Z}{n}\right)^{-1}\right)}_{\mathcal{O}_p(1)},$$

which finishes the proof. □

Proof of Proposition 5.7: We start with the boundedness in probability of $y_0 - x_0'\widehat{\beta}_R(c_{n,p})$. Again, by Lemma A.6 and the boundedness of $\mathbb{E}(y_0^2)$ it suffices to control $\|\Sigma^{1/2}\widehat{\beta}_R(c_{n,p})\|_2$: Using the first part of Lemma A.17 yields

$$\|\Sigma^{1/2}\widehat{\beta}_R(c_{n,p})\|_2 \leq \sqrt{n}\sigma_{\max}(Z^\dagger)\frac{\|Y\|_2}{\sqrt{n}}\frac{\sigma_{\max}(\Sigma^{1/2})}{\sigma_{\min}(\Sigma^{1/2})}.$$

By Lemma 5.3 the smallest eigenvalue of ZZ'/n converges to $(\rho^{1/2} - 1)^2$ in probability or to ∞, if $\rho = \infty$. This implies the convergence of $\sqrt{n}\sigma_{\max}(Z^\dagger)$ to $(\rho^{1/2} - 1)^{-1}$ or 0, respectively. In any case, $\sqrt{n}\sigma_{\max}(Z^\dagger)$ is bounded in probability. Furthermore, $\|Y\|_2/\sqrt{n}$ is bounded in probability and the last term coincides with the square root of the condition number of Σ, which is bounded by assumption. The argument can be repeated to show that $y_0 - x_0'\widetilde{\beta}_R(c_{n-1,p})$ is bounded in probability as well.

For the second part of the proof we again use Lemma A.6 together with the decomposition (A.25). With the same arguments as before we can show that

$$\|\Sigma^{1/2}(\widehat{\beta}_R(c_{n-1,p}) - \widetilde{\beta}_R(c_{n-1,p}))\|_2 = \|\Sigma^{1/2}(X'X + c_{n-1,p}I_p)^\dagger X'e_n(y_n - x_n'\widetilde{\beta}_R(c_{n-1,p})\|_2,$$

which by Lemma A.17 can be bounded from above by

$$\underbrace{\sigma_{\max}(Z^\dagger)}_{\mathcal{O}_p(n^{-1/2})}\underbrace{|y_n - x_n'\widetilde{\beta}_R(c_{n-1,p})|}_{\mathcal{O}_p(1)}\frac{\sigma_{\max}(\Sigma^{1/2})}{\sigma_{\min}(\Sigma^{1/2})},$$

where again the last term is bounded over n by assumption.

To control the term $\Sigma^{1/2}(\widehat{\beta}_R(c_{n,p}) - \widehat{\beta}_R(c_{n-1,p}))$ we use equation (A.26) to get

$$\Sigma^{1/2}(\widehat{\beta}_R(c_{n,p}) - \widehat{\beta}_R(c_{n-1,p})) = (c_{n-1,p} - c_{n,p})\Sigma^{1/2} \underbrace{(X'X + c_{n-1,p}I_p)^\dagger (X'X + c_{n,p}I_p)^\dagger X'}_{=:A} Y.$$

Let $s_1 \geq s_2 \geq \ldots \geq s_n$ denote the singular values of X. Then the i-th singular value of A is given by $s_i/((s_i^2 + c_{n-1,p})(s_i^2 + c_{n,p}))$ if $s_i > 0$ and by 0 otherwise. Recalling the fact that $c_{n,p}$ and $c_{n-1,p}$ are nonnegative, we conclude that $s_i/((s_i^2 + c_{n-1,p})(s_i^2 + c_{n,p}))$ can be bounded from above by s_i^{-3} whenever $s_i > 0$. Hence, we have $\sigma_{\max}(A) \leq (\sigma_{\max}(X^\dagger))^3$, which itself can be bounded from above by $(\sigma_{\max}(Z^\dagger))^3 \sigma_{\max}(\Sigma^{-3/2})$. Putting the pieces together we end up with the following upper bound for $\|\Sigma^{1/2}(\widehat{\beta}_R(c_{n,p}) - \widehat{\beta}_R(c_{n-1,p}))\|_2$:

$$\begin{aligned}
&|c_{n-1,p} - c_{n,p}| \sigma_{\max}(\Sigma^{1/2})(\sigma_{\max}(Z^\dagger))^3 \sigma_{\max}(\Sigma^{-3/2}) \|Y\|_2 \\
&= \underbrace{\frac{|c_{n-1,p} - c_{n,p}|}{n} \lambda_{\max}\left(\Sigma^{-1}\right)}_{\xrightarrow{p} 0} \underbrace{\left(\sqrt{n}\sigma_{\max}(Z^\dagger)\right)^3}_{\mathcal{O}_p(1)} \underbrace{\frac{\|Y\|_2}{\sqrt{n}} \frac{\sigma_{\max}(\Sigma^{1/2})}{\sigma_{\min}(\Sigma^{1/2})}}_{\mathcal{O}_p(1)},
\end{aligned}$$

where we used the fact that $\sigma_{\max}(\Sigma^{-3/2}) = \sigma_{\max}(\Sigma^{-1})\sigma_{\max}(\Sigma^{-1/2}) = \frac{\lambda_{\max}\left(\Sigma^{-1}\right)}{\sigma_{\min}\left(\Sigma^{1/2}\right)}$. □

The main difference between the proofs of Proposition 5.6 and Proposition 5.7 lies in the fact that in the high-dimensional case the matrix $X'X$ is not invertible. Even the fact that XX' has full rank with asymptotic probability 1 together with the alternative representation $\widehat{\beta}_R(c_{n,p}) = X'(XX' + c_{n,p}I_n)^\dagger Y$ cannot fix the problem properly: This is mainly driven by the fact that for $XX' = Z\Sigma Z'$ the matrix Σ is sandwiched between two matrices in contrast to $X'X = \Sigma^{1/2}Z'Z\Sigma^{1/2}$, where the matrix $\Sigma^{1/2}$ can be taken out of the inverse.

A.4.3. Proofs concerning the James-Stein estimator

Before we start proving Proposition 5.11 we need the following definitions: Define

$$\hat{\delta} = \begin{cases} \dfrac{Y'(I_n - P_X)Y}{Y'P_XY} & \text{if } Y'P_XY > 0 \\ 0 & \text{else} \end{cases}$$

and let $\tilde{\delta}$ denote its leave-one-out analogue in the sense that

$$\tilde{\delta} = \begin{cases} \dfrac{\tilde{Y}'(I_{n-1} - P_{\tilde{X}})\tilde{Y}}{\tilde{Y}'P_{\tilde{X}}\tilde{Y}} & \text{if } \tilde{Y}'P_{\tilde{X}}\tilde{Y} > 0 \\ 0 & \text{else.} \end{cases}$$

For convenience, we will use the notation $\hat{\gamma} := \hat{\delta}c_{n,p}p/(n-p)$ and $\tilde{\gamma} := \tilde{\delta}c_{n-1,p}p/(n-p-1)$ and denote $e_n'(I_n - P_X)Y$ with $\bar{u}_n$. In the proof of Proposition 5.11 we will distinguish between the cases whether $\hat{\delta}$ and $\tilde{\delta}$ are zero or not. In the case, where both terms are

positive, we will need some terms to be bounded in probability. The next lemma specifies this statement:

Lemma A.18. *Define* $\hat{t} = \mathbb{1}\{\hat{\delta} > 0, \tilde{\delta} > 0\}$ *and let the assumptions* ***LD*** *be fulfilled. Then the following expressions are bounded in probability:*

1. $n\hat{t}\bar{u}_n^2/(Y'(I_n - P_X)Y)$,
2. $n\hat{t}(e_n'P_XY)^2/(Y'P_XY)$ *and*
3. $\sqrt{n}\hat{t}\dfrac{\|\Sigma^{1/2}\widehat{\beta}_{LS}\|_2}{\|X\widehat{\beta}_{LS}\|_2}$,

where we define all terms to be 0 *if* $\hat{t} = 0$.

Proof. Let $\hat{s}$ denote $\mathbb{1}\{Y'(I_n - P_X)Y = 0\}$. Since $\hat{t} = 1$ implies $\hat{s} = 0$, we can bound $n\hat{t}\bar{u}_n^2/(Y'(I_n - P_X)Y)$ from above by $n\bar{u}_n^2/(\hat{s} + Y'(I_n - P_X)Y)$. As the data $(y_i, x_i')'$ are i.i.d., they are exchangeable as well. Since the term $Y'(I_n - P_X)Y$ and

$$X^\dagger Y = (X'X)^\dagger X'Y = \left(\sum_{i=1}^n x_i x_i'\right)^\dagger \left(\sum_{i=1}^n x_i y_i\right)$$

are invariant under a permutation of the data, the expression

$$\frac{(e_i'(I_n - P_X)Y)^2}{\hat{s} + Y'(I_n - P_X)Y} = \frac{(y_i - x_i'X^\dagger Y)^2}{\hat{s} + Y'(I_n - P_X)Y}$$

has the same distribution as $\bar{u}_n^2/(\hat{s} + Y'(I_n - P_X)Y)$. Hence, we have

$$\begin{aligned}\mathbb{E}\left(n\frac{\bar{u}_n^2}{(\hat{s} + Y'(I_n - P_X)Y}\right) &= \mathbb{E}\left(\frac{\sum_{i=1}^n (e_i'(I_n - P_X)Y)^2}{\hat{s} + Y'(I_n - P_X)Y}\right) \\ &= \mathbb{E}\left(\frac{Y'(I_n - P_X)Y}{\hat{s} + Y'(I_n - P_X)Y}\right) \leq 1.\end{aligned}$$

The boundedness of $n\hat{t}(e_n'P_XY)^2/(Y'P_XY)$ can be shown analogously if we replace $\hat{s}$ by $\hat{r} = \mathbb{1}\{Y'P_XY = 0\}$. For the third statement we point out that $\|X\widehat{\beta}_{LS}\|_2^2 = Y'P_XY > 0$ and therefore the fraction $\frac{\|\Sigma^{1/2}\widehat{\beta}_{LS}\|_2}{\|X\widehat{\beta}_{LS}\|_2}$ is well defined whenever $\hat{t} > 0$. Moreover, by Lemma 5.3 the smallest eigenvalue of $(\Sigma^{-1/2}X'X\Sigma^{-1/2})/n$ converges in probability to $(1 - \sqrt{\rho})^2$. Thus, we can condition on the event where the smallest eigenvalue is positive as it has asymptotic probability 1. On this event, we have

$$\begin{aligned}n\hat{t}^2\frac{\|\Sigma^{1/2}\widehat{\beta}_{LS}\|_2^2}{\|X\widehat{\beta}_{LS}\|_2^2} &\leq n\hat{t}^2\lambda_{\max}\left(\left(\Sigma^{-1/2}X'X\Sigma^{-1/2}\right)^{-1}\right) \\ &\leq \lambda_{\max}\left(\left(\Sigma^{-1/2}X'X\Sigma^{-1/2}/n\right)^{-1}\right) \xrightarrow{p} (1 - \sqrt{\rho})^{-2},\end{aligned}$$

which proves the boundedness in probability. □

Lemma A.19. *Under the assumptions **LD** we have*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{\tilde{\delta}}{\hat{\delta}} - 1\right| > \varepsilon, 0 < \hat{\delta} < \sqrt{n}, 0 < \tilde{\delta}\right) = 0$$

for every $\varepsilon > 0$.

Proof. We start with the observation that

$$\frac{\tilde{\delta}}{\hat{\delta}} = \frac{Y' P_X Y}{\tilde{Y}' P_{\tilde{X}} \tilde{Y}} \frac{\tilde{Y}'(I_{n-1} - P_{\tilde{X}})\tilde{Y}}{Y'(I_n - P_X)Y}$$

whenever $\hat{\delta} > 0, \tilde{\delta} > 0$. Thus, it suffices to prove that both terms converge to 1 in probability on the event $\sqrt{n} > \hat{\delta} > 0, \tilde{\delta} > 0$. For this, we additionally condition on the event where $\tilde{X}'\tilde{X}$ is invertible, which has asymptotic probability 1 due to Lemma 5.3 and the fact that $\Sigma^{1/2}$ is positive definite. By Lemma A.3 we have

$$Y'(I_n - P_X)Y = \tilde{Y}'(I_{n-1} - P_{\tilde{X}})\tilde{Y} + \bar{u}_n^2 \left(x_n'(\tilde{X}'\tilde{X})^{-1}x_n + 1\right).$$

We claim that $x_n'(\tilde{X}'\tilde{X})^{-1}x_n$ is bounded in probability. To see this, we compute the conditional expectation

$$\mathbb{E}\left(x_n'(\tilde{X}'\tilde{X})^{-1}x_n \| T_{n-1}\right) = \operatorname{tr}\left(\Sigma^{1/2}(\tilde{X}'\tilde{X})^{-1}\Sigma^{1/2}\right) \leq \frac{p}{\lambda_{\min}(\Sigma^{-1/2}\tilde{X}'\tilde{X}\Sigma^{-1/2})}.$$

Lemma 5.3 and the fact that $\lim_{n\to\infty} p/n = \rho < \infty$ imply that the latter expression is bounded in probability. Thus, by Lemma A.6 the original term $\left(x_n'(\tilde{X}'\tilde{X})^{-1}x_n\right)$ is bounded in probability as well. Thus, we have

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{\tilde{Y}'(I_{n-1} - P_{\tilde{X}})\tilde{Y}}{Y'(I_n - P_X)Y} - 1\right| \geq \varepsilon, 0 < \hat{\delta} < \sqrt{n}, 0 < \tilde{\delta}\right) \\
&\leq \lim_{n\to\infty} \mathbb{P}\left(\frac{\bar{u}_n^2}{Y'(I_n - P_X)Y}(x_n'(\tilde{X}'\tilde{X})^{-1}x_n + 1) \geq \varepsilon, 0 < \hat{\delta}, 0 < \tilde{\delta}\right) = 0,
\end{aligned}$$

since $\dfrac{\bar{u}_n^2}{Y'(I_n - P_X)Y}\hat{t} \xrightarrow{p} 0$ by Lemma A.18.

Next, we will show that for every $\varepsilon > 0$ the following statement holds true:

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{Y' P_X Y}{\tilde{Y}' P_{\tilde{X}} \tilde{Y}} - 1\right| \geq \varepsilon, 0 < \hat{\delta} < \sqrt{n}, 0 < \tilde{\delta}\right) = 0.$$

Since we are conditioning on the event where $\tilde{X}'\tilde{X}$ is invertible, we can use Lemma A.3 to get

$$Y' P_X Y - \tilde{Y}' P_{\tilde{X}} \tilde{Y} = (e_n' P_X Y)^2 + \bar{u}_n^2 x_n'(\tilde{X}'\tilde{X})^{-1}x_n + 2\bar{u}_n x_n' \tilde{X}^{\dagger}\tilde{Y}.$$

Furthermore, on the event $\hat{\delta} > 0, \tilde{\delta} > 0$ we have $Y'(I_n - P_X)Y > 0, Y'P_XY > 0, \tilde{Y}'P_{\tilde{X}}\tilde{Y} > 0$, which implies

$$\begin{aligned} Y'P_XY - \tilde{Y}'P_{\tilde{X}}\tilde{Y} = Y'P_XY \left[\frac{(e_n'P_XY)^2}{Y'P_XY} + \hat{\delta}\frac{\bar{u}_n^2}{Y'(I-P_X)Y} x_n'(\tilde{X}'\tilde{X})^{-1}x_n \right] \\ + 2\frac{x_n'\tilde{X}^\dagger\tilde{Y}}{\sqrt{\tilde{Y}'P_{\tilde{X}}\tilde{Y}}} \frac{\bar{u}_n}{\sqrt{Y'(I-P_X)Y}} \sqrt{\tilde{Y}'P_{\tilde{X}}\tilde{Y}Y'(I-P_X)Y}. \end{aligned} \tag{A.27}$$

Now, let τ_n denote $[(Y'P_XY)/(\tilde{Y}'P_{\tilde{X}}\tilde{Y})]^{1/2}$, which is well-defined (and positive) on the event $\tilde{\delta} > 0, \hat{\delta} > 0$. Then, dividing the right-hand side of (A.27) by $[\tilde{Y}'P_{\tilde{X}}\tilde{Y}Y'P_XY]^{1/2} > 0$ yields

$$\begin{aligned} \tau_n - \frac{1}{\tau_n} = \tau_n \left[\underbrace{\frac{(e_n'P_XY)^2}{Y'P_XY}}_{=:a_1} + \hat{\delta} \underbrace{\frac{\bar{u}_n^2}{Y'(I-P_X)Y}}_{=:a_2} x_n'(\tilde{X}'\tilde{X})^{-1}x_n \right] \\ + 2 \underbrace{\frac{x_n'\tilde{X}^\dagger\tilde{Y}}{\sqrt{\tilde{Y}'P_{\tilde{X}}\tilde{Y}}}}_{=:a_3} \underbrace{\frac{\bar{u}_n}{\sqrt{Y'(I-P_X)Y}}}_{=:a_4} \sqrt{\hat{\delta}}. \end{aligned} \tag{A.28}$$

On the event $\hat{\delta} > 0, \tilde{\delta} > 0$, Lemma A.18 implies the convergence in probability of a_1 and a_2 to 0 with the rate $1/n$, while a_4 is of magnitude $1/\sqrt{n}$ and $x_n'(\tilde{X}'\tilde{X})^{-1}x_n$ is bounded in probability as shown before. Furthermore, we have

$$\mathbb{E}\left(na_3^2 \| T_{n-1}\right) = n\frac{\|\Sigma^{1/2}\widetilde{\beta}_{LS}\|_2^2}{\|\tilde{X}\widetilde{\beta}_{LS}\|_2^2} \leq \left(\left(\lambda_{\min}\left(\Sigma^{-1/2}\frac{\tilde{X}'\tilde{X}}{n}\Sigma^{-1/2} \right) \right)^{-1} \right),$$

which is bounded in probability. Hence, by Lemma A.6 na_3^2 is bounded in probability, which implies that a_3 converges to 0 in probability with rate $1/\sqrt{n}$. Hence, on the event $\hat{\delta} \leq \sqrt{n}$ the terms $a_1 + \hat{\delta}a_2x_n'(\tilde{X}'\tilde{X})^{-1}x_n$ and $\sqrt{\hat{\delta}}a_3a_4$ converge to 0 in probability. Rewriting equation (A.28) yields

$$\tau_n \underbrace{\left(1 - a_1 - \hat{\delta}a_2x_n'(\tilde{X}'\tilde{X})^{-1}x_n\right)}_{\xrightarrow{p}1} - \frac{1}{\tau_n} = \underbrace{2\sqrt{\hat{\delta}}a_3a_4}_{\xrightarrow{p}0}.$$

Recalling the fact that τ_n is positive, this is only possible if $\tau_n \xrightarrow{p} 1$. Thus, we have proved that for every $\varepsilon > 0$ we have that

$$\lim_{n\to\infty} \mathbb{P}\left(\left| \sqrt{\frac{Y'P_XY}{\tilde{Y}'P_{\tilde{X}}\tilde{Y}}} - 1 \right| \geq \varepsilon, 0 < \hat{\delta} < \sqrt{n}, \tilde{\delta} > 0 \right) = 0.$$

□

Proof of Proposition 5.11: By Lemma 5.3 the smallest eigenvalue of $\Sigma^{-1/2}\tilde{X}'\tilde{X}\Sigma^{-1/2}/n$ converges in probability to $(1-\sqrt{\rho})^2 > 0$. Thus, we will condition on the event where $\tilde{X}$ has full column rank p, which has asymptotic probability of 1. We would like to point out that this even ensures that the matrix X has full column rank p.

Furthermore, we will make use of the fact that $\|\Sigma^{1/2}\widehat{\beta}_{LS}\|_2$ is bounded in probability and $\|\Sigma^{1/2}(\widehat{\beta}_{LS} - \widetilde{\beta}_{LS})\|_2$ converges to 0 in probability, which is shown in the proof of Proposition 5.6. We would like to emphasize that the OLS estimator coincides with the Ridge estimator with parameter $c_{n,p} = 0$. Thus, the assumptions of Proposition 5.6 are fulfilled as in Proposition 5.11 we also supposed Assumption **LD** to hold true.

As $\|\Sigma^{1/2}\widehat{\beta}_{LS}\|_2$ is bounded in probability, the same holds true for $\|\Sigma^{1/2}\widehat{\beta}_{JS}\|_2 = \max(0, 1-\hat{\gamma})\|\Sigma^{1/2}\widehat{\beta}_{LS}\|_2$. Using Lemma A.6 we proved that $x_0'\widehat{\beta}_{JS}$ is bounded in probability. Since the second moment of y_0 is bounded over n, the prediction error $y_0 - x_0'\widehat{\beta}_{JS}$ is bounded in probability as well, which proves the first part of Proposition 5.11.

Now, by Lemma A.6 it remains to show that $\|\Sigma^{1/2}(\widehat{\beta}_{JS} - \widetilde{\beta}_{JS})\|_2$ converges to 0 in probability. Before proving this we introduce some notation to improve the clarity of our arguments: Define $\hat{\alpha} = c_{n,p}p/(n-p)$, $\tilde{\alpha} = c_{n-1,p}p/(n-p-1)$, which yields the identities $\hat{\gamma} = \hat{\alpha}\hat{\delta}$ and $\tilde{\gamma} = \tilde{\alpha}\tilde{\delta}$. In the ensuing proof, we will repeatedly use the fact that for any nonnegative number a we have $\min(1,a)^2 \leq \min(1,a) \leq a$. Moreover, we are using the equality $\max(0, 1-\hat{\gamma}) = 1 - \min(1, \hat{\gamma})$, which can be seen as follows:

$$\max(0, 1-\hat{\gamma}) = \max(1-1, 1-\hat{\gamma}) = 1 + \max(-1, -\hat{\gamma}) = 1 - \min(1, \hat{\gamma}).$$

We bound the expression $\|\Sigma^{1/2}(\widehat{\beta}_{JS} - \widetilde{\beta}_{JS})\|_2$ from above by

$$\|\Sigma^{1/2}(\widehat{\beta}_{LS} - \widetilde{\beta}_{LS})\|_2 + \|\min(1,\hat{\gamma})\Sigma^{1/2}\widehat{\beta}_{LS} - \min(1,\tilde{\gamma})\Sigma^{1/2}\widetilde{\beta}_{LS}\|_2$$

where the first term on the preceding display converges to 0 as explained at the beginning of this proof. The second term can be bounded by

$$\min(1,\tilde{\gamma})\|\Sigma^{1/2}(\widehat{\beta}_{LS} - \widetilde{\beta}_{LS})\|_2 + \underbrace{\|\Sigma^{1/2}\widehat{\beta}_{LS}\|_2\,|\min(1,\hat{\gamma}) - \min(1,\tilde{\gamma})|}_{=:b_0},$$

where, again, the first term converges to 0 in probability. To finish the proof, it remains to show the same for b_0. In order to do this, we need to distinguish between the cases where $\hat{\delta}$ and $\tilde{\delta}$ are zero or not. More precisely, we will use the equation

$$\mathbb{P}(b_0 \geq \varepsilon) \leq \mathbb{P}(b_0 \geq \varepsilon, \hat{\delta} = 0) + \mathbb{P}(b_0 \geq \varepsilon, \hat{\delta} > 0, \tilde{\delta} = 0) + \mathbb{P}(b_0 \geq \varepsilon, \hat{\delta} > 0, \tilde{\delta} > 0)$$

and show that each term on the right-hand side converges to 0 for every $\varepsilon > 0$.

Case $\hat{\delta} = 0$: This can only occur if $Y'(I_n - P_X)Y = 0$ or $Y'P_XY = 0$.

In the subcase $Y'(I_n - P_X)Y = 0$, Lemma A.3 entails that $\tilde{Y}'(I_{n-1} - P_{\tilde{X}})\tilde{Y} \leq Y'(I_n - P_X)Y$, implying that $\tilde{\delta} = 0$ as well. Thus, we have $b_0 = 0$.

In the subcase, where $\|X\widehat{\beta}_{LS}\|_2^2 = Y'P_XY = 0 < Y'(I_n - P_X)Y$, we have $\widehat{\beta}_{LS} = 0$ (as

we assumed X to have full column rank p). Hence, we have $\|\Sigma^{1/2}\widehat{\beta}_{LS}\|_2 = 0$.

Case $\hat{\delta} > 0 = \tilde{\delta}$: We start with the subcase $\tilde{Y}'P_{\tilde{X}}\tilde{Y} = 0$, which implies $\widetilde{\beta}_{LS} = 0$. By Lemma A.2 we have

$$\Sigma^{1/2}\widehat{\beta}_{LS} = \Sigma^{1/2}\widetilde{\beta}_{LS} + \Sigma^{1/2}(\tilde{X}'\tilde{X})^{-1}x_n\bar{u}_n = \Sigma^{1/2}(\tilde{X}'\tilde{X})^{-1}x_n\bar{u}_n,$$

where $\bar{u}_n = e_n'(I_n - P_X)Y$. We now claim, that $\bar{u}_n$ is bounded in probability and $\|\Sigma^{1/2}(\tilde{X}'\tilde{X})^{-1}x_n\|_2$ converges to 0 in probability. Using the fact that $e_i'(I_n - P_X)Y$ has the same distribution as $\bar{u}_n$, this can be seen as follows:

$$\mathbb{E}(\bar{u}_n^2) = \frac{1}{n}\,\mathbb{E}(Y'(I_n - P_X)\sum_{i=1}^{n} e_i e_i'(I_n - P_X)Y)) = \frac{\mathbb{E}(Y'(I_n - P_X)Y)}{n} \leq \mathbb{E}(y_0^2),$$

which is bounded over n. Hence, $\bar{u}_n$ is bounded in probability. Furthermore, we have

$$\begin{aligned}\|\Sigma^{1/2}(\tilde{X}'\tilde{X})^{-1}x_n\|_2 &= \|(\Sigma^{-1/2}\tilde{X}'\tilde{X}\Sigma^{-1/2})^{-1}\Sigma^{-1/2}x_n\|_2 \\ &\leq \frac{1}{\sqrt{n}}\lambda_{\max}\left((\Sigma^{-1/2}X'X\Sigma^{-1/2}/n)^{-1}\right)\frac{\|\Sigma^{-1/2}x_n\|_2}{\sqrt{n}}\end{aligned}$$

which converges to 0 as the largest eigenvalue of $(\Sigma^{-1/2}X'X\Sigma^{-1/2}/n)^{-1}$ is bounded in probability and $\mathbb{E}(x_n'\Sigma^{-1}x_n/n) = p/n$. Together with the fact that $|\min(1,\hat{\gamma})|$ is bounded (in probability), we have $\|\Sigma^{1/2}\widehat{\beta}_{LS}\|_2\,|\min(1,\hat{\gamma}) - \min(1,\tilde{\gamma})| \xrightarrow{p} 0$. Thus, we showed the following: For every $\varepsilon > 0$ we have

$$\lim_{n\to\infty}\mathbb{P}\left(b_0 \geq \varepsilon, \tilde{Y}'P_{\tilde{X}}\tilde{Y} = 0, \hat{\delta} > 0\right) \leq \lim_{n\to\infty}\mathbb{P}\left(\|\Sigma^{1/2}(\tilde{X}'\tilde{X})^{-1}x_n\bar{u}_n\|_2 \geq \varepsilon\right) = 0.$$

The second subcase of $\hat{\delta} > 0 = \tilde{\delta}$ deals with $\tilde{Y}'P_{\tilde{X}}\tilde{Y} > 0 = \tilde{Y}'(I_{n-1} - P_{\tilde{X}})\tilde{Y}$. In that subcase, Lemma A.3 yields

$$\begin{aligned}Y'(I_n - P_X)Y &= \tilde{Y}'(I_{n-1} - P_{\tilde{X}})\tilde{Y} + \bar{u}_n^2\left(x_n'(\tilde{X}'\tilde{X})^{-1}x_n + 1\right) \\ &= \bar{u}_n^2\left(x_n'(\tilde{X}'\tilde{X})^{-1}x_n + 1\right).\end{aligned}$$

As already shown above, $\bar{u}_n$ is bounded in probability. Furthermore, $x_n'(\tilde{X}'\tilde{X})^{-1}x_n$ is also bounded in probability since we can bound it from above as follows:

$$x_n'\Sigma^{-1/2}(\Sigma^{-1/2}\tilde{X}'\tilde{X}\Sigma^{-1/2})^{-1}\Sigma^{-1/2}x_n \leq \frac{x_n'\Sigma^{-1}x_n}{n}\lambda_{\max}\left[\left(\Sigma^{-1/2}\tilde{X}'\tilde{X}\Sigma^{-1/2}/n\right)^{-1}\right].$$

Again, $x_n'\Sigma^{-1}x_n/n$ is bounded in probability. To sum it up, $Y'(I_n - P_X)Y$ is bounded in probability in that subcase. As $\lim_{n\to\infty} p/n = \rho \in [0,1)$, the same holds true for $\hat{\alpha} = c_{n,p}p/(n-p)$. Since we are in the subcase $\tilde{\delta} = 0$ and $\tilde{\gamma} = \tilde{\delta}\tilde{\alpha}$ we have

$$b_0^2 = |\min(1,\hat{\gamma}) - \min(1,\tilde{\gamma})|^2\|\Sigma^{1/2}\widehat{\beta}_{LS}\|_2^2 \leq \hat{\gamma}\|\Sigma^{1/2}\widehat{\beta}_{LS}\|_2^2.$$

The expression above can again be bounded by

$$\hat{\gamma}\|\Sigma^{1/2}\widehat{\beta}_{LS}\|_2^2 = \hat{\alpha}\frac{Y'(I_n - P_X)Y}{Y'P_XY}\|\Sigma^{1/2}\widehat{\beta}_{LS}\|_2^2 = \hat{\alpha}Y'(I_n - P_X)Y\frac{\widehat{\beta}'_{LS}\Sigma\widehat{\beta}_{LS}}{\widehat{\beta}'_{LS}X'X\widehat{\beta}_{LS}}$$
$$\leq \underbrace{\hat{\alpha}}_{\mathcal{O}_p(1)}\underbrace{Y'(I_n - P_X)Y}_{\mathcal{O}_p(1)}\underbrace{\lambda_{\max}\left[\left(\Sigma^{-1/2}X'X\Sigma^{-1/2}\right)^{-1}\right]}_{\mathcal{O}_p(n^{-1})} \xrightarrow{p} 0.$$

To state our argument more precisely, we have shown that for every $\varepsilon > 0$ we have

$$\lim_{n\to\infty}\mathbb{P}\left(\|\Sigma^{1/2}\widehat{\beta}_{LS}\|_2\,|\min(1,\hat{\gamma}) - \min(1,\tilde{\gamma})| \geq \varepsilon, \tilde{Y}'(I_{n-1} - P_{\tilde{X}})\tilde{Y} = 0, \tilde{Y}'P_{\tilde{X}}\tilde{Y} > 0, \hat{\delta} > 0\right)$$
$$\leq \lim_{n\to\infty}\mathbb{P}\left(\hat{\alpha}\bar{u}_n^2\left(x_n'(\tilde{X}'\tilde{X})^{-1}x_n + 1\right)\lambda_{\max}\left(\Sigma^{-1/2}X'X\Sigma^{-1/2}\right)^{-1} \geq \varepsilon^2\right) = 0.$$

We now deal with the case $\hat{\delta} > 0, \tilde{\delta} > 0$. By Lemma A.10 we can bound b_0 by

$$\underbrace{\min\left(1, \hat{\delta}|\hat{\alpha} - \tilde{\alpha}|\right)\|\Sigma^{1/2}\widehat{\beta}_{LS}\|_2}_{=:b_1} + \underbrace{\min\left(1, \tilde{\alpha}|\hat{\delta} - \tilde{\delta}|\right)\|\Sigma^{1/2}\widehat{\beta}_{LS}\|_2}_{=:b_2}.$$

The proof is finished if we can show that each expression converges to 0 in probability. For this, we treat the both terms separately. Starting with b_1, we have

$$b_1^2 \leq \min\left(1, \hat{\delta}|\hat{\alpha} - \tilde{\alpha}|\right)\frac{Y'(I - P_X)Y}{n\hat{\delta}}\frac{n\|\Sigma^{1/2}\widehat{\beta}_{LS}\|_2^2}{\|X\widehat{\beta}_{LS}\|_2^2}$$

where we used the fact that $\hat{\delta}$ and $\|X\widehat{\beta}_{LS}\|_2$ are strictly positive together with the inequality $\min(1,x)^2 \leq \min(1,x)$ for all positive x. Now, we bound the term by

$$|\hat{\alpha} - \tilde{\alpha}|\frac{Y'(I - P_X)Y}{n}\frac{n\|\Sigma^{1/2}\widehat{\beta}_{LS}\|_2^2}{\|X\widehat{\beta}_{LS}\|_2^2}$$
$$\leq |\hat{\alpha} - \tilde{\alpha}|\underbrace{\frac{Y'(I - P_X)Y}{n}}_{\mathcal{O}_p(1)}\underbrace{\lambda_{\max}\left(\left(\Sigma^{-1/2}X'X\Sigma^{-1/2}/n\right)^{-1}\right)}_{\mathcal{O}_p(1)}.$$

Furthermore, $\hat{\alpha} - \tilde{\alpha}$ equals $\frac{p}{n-p-1}(c_{n,p} - c_{n-1,p}) - \frac{c_{n,p}p}{(n-p)(n-p-1)}$. However, $p/(n-p)$ converges to $\rho/(1-\rho)$, $c_{n,p} \in [0,1]$ and $c_{n,p} - c_{n-1,p}$ converges to 0 in probability by assumption. Hence, $\hat{\alpha} - \tilde{\alpha} \xrightarrow{p} 0$.

In order to show that b_2 converges to 0 in probability we split the problem in two parts depending on whether $\hat{\delta}$ is smaller than $\sqrt{n}$ or not:

$$\limsup_{n\to\infty}\mathbb{P}(b_2 > \varepsilon, \hat{\delta} > 0) = \limsup_{n\to\infty}\mathbb{P}(b_2 > \varepsilon, 0 < \hat{\delta} < \sqrt{n}) + \limsup_{n\to\infty}\mathbb{P}(b_2 > \varepsilon, \hat{\delta} \geq \sqrt{n}).$$

On the event $\hat{\delta} \geq \sqrt{n}$ we have

$$\begin{aligned} b_2^2 &\leq \|\Sigma^{1/2}\widehat{\beta}_{LS}\|_2^2 \leq \frac{Y'(I-P_X)Y}{n\hat{\delta}} \frac{n\|\Sigma^{1/2}\widehat{\beta}_{LS}\|_2^2}{\|X\widehat{\beta}_{LS}\|_2^2} \\ &\leq \frac{1}{\sqrt{n}} \underbrace{\frac{Y'(I-P_X)Y}{n}}_{\mathcal{O}_p(1)} \underbrace{\lambda_{\max}\left(\left(\Sigma^{-1/2}X'X\Sigma^{-1/2}/n\right)^{-1}\right)}_{\mathcal{O}_p(1)}, \end{aligned}$$

which proves $\limsup_{n\to\infty} \mathbb{P}(b_2 > \varepsilon, \delta \geq \sqrt{n}) = 0$. Now we treat the case $0 < \hat{\delta} < \sqrt{n}$: As $\min(1,x)^2 \leq \min(1,x)$ for every $x \geq 0$, we have

$$b_2^2 \leq \min\left(1, \tilde{\alpha}|\hat{\delta} - \tilde{\delta}|\right) \|\Sigma^{1/2}\widehat{\beta}_{LS}\|_2^2 \leq \left|\frac{\tilde{\delta}}{\hat{\delta}} - 1\right| \underbrace{\tilde{\alpha}\frac{Y'(I-P_X)Y}{n} \frac{n\|\Sigma^{1/2}\widehat{\beta}_{LS}\|_2^2}{\|X\widehat{\beta}_{LS}\|_2^2}}_{\mathcal{O}_p(1)}.$$

Now, in the case $0 < \hat{\delta} < \sqrt{n}, 0 < \tilde{\delta}$ Lemma A.19 implies that the term $|(\tilde{\delta}/\hat{\delta}) - 1|$ converges to 0 in probability.

To sum it up, we showed that the term

$$b_0 = \|\Sigma^{1/2}\widehat{\beta}_{LS}\|_2 \, |\min(1,\hat{\gamma}) - \min(1,\tilde{\gamma})|$$

converges to 0 regardless of the values of $\hat{\delta}$ and $\tilde{\delta}$. To formalize this, we can proceed as follows:

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}(b_0 \geq \varepsilon) \leq \lim_{n\to\infty} \big[&\mathbb{P}(b_0 \geq \varepsilon, \hat{\delta} = 0) + \mathbb{P}(b_0 \geq \varepsilon, \hat{\delta} > 0, \tilde{\delta} = 0) \\ &+ \mathbb{P}(b_0 \geq \varepsilon, 0 < \hat{\delta} < \sqrt{n}, \tilde{\delta} > 0) + \mathbb{P}(b_0 \geq \varepsilon, \sqrt{n} \leq \hat{\delta}, \tilde{\delta} > 0)\big] = 0. \end{aligned}$$

□

A.4.4. Proofs for binary classification

Proof of Proposition 5.14: As y_0 and $\hat{y}_0$ can only take the values 1 or -1, the prediction error is clearly bounded.

Thus, it remains to show the stability of $\hat{y}_0$, i.e.,

$$\lim_{n\to\infty} \mathbb{P}\left(|\hat{y}_0 - \tilde{y}_0| \geq \varepsilon\right) = 0$$

for every $\varepsilon > 0$. However, $|\hat{y}_0 - \tilde{y}_0|$ can only be non-zero if the signs of $x_0'\widehat{\beta}$ and $x_0'\widetilde{\beta}$ differ, which implies that $\widehat{\beta}' x_0 x_0' \widetilde{\beta}$ is smaller or equal to 0. Hence, it suffices to show that

$$\lim_{n\to\infty} \mathbb{P}\left(\widehat{\beta}' x_0 x_0' \widetilde{\beta} \leq 0\right) = 0.$$

We can rewrite the probability above as

$$\mathbb{P}\left(\widehat{\beta}' x_0 x_0' \widehat{\beta} \leq \widehat{\beta}' x_0 x_0' (\widehat{\beta} - \widetilde{\beta})\right),$$

which can be bounded from above by

$$\mathbb{P}\left(|x_0'\widehat{\beta}| \leq |x_0'(\widehat{\beta} - \widetilde{\beta})|\right) \leq \mathbb{P}\left(|x_0'\widehat{\beta}| \leq \delta_n\right) + \mathbb{P}\left(|x_0'(\widehat{\beta} - \widetilde{\beta})| > \delta_n\right),$$

where $\delta_n > 0$ can be chosen arbitrarily. Recalling that $x_0'(\widehat{\beta} - \widetilde{\beta})$ converges to 0 in probability by assumption, we can apply Lemma A.5 to find a null-sequence $(v_n)_{n\in\mathbb{N}}$ of positive numbers, such that $x_0'(\widehat{\beta} - \widetilde{\beta})/v_n$ still converges to 0 in probability. Defining $\delta_n = v_n$, this yields

$$\lim_{n\to\infty} \mathbb{P}\left(|x_0'(\widehat{\beta} - \widetilde{\beta})| > \delta_n\right) = \lim_{n\to\infty} \mathbb{P}\left(|x_0'(\widehat{\beta} - \widetilde{\beta})|/\delta_n > 1\right) = 0.$$

To finish the proof, it remains to show that

$$\limsup_{n\to\infty} \mathbb{P}\left(|x_0'\widehat{\beta}| \leq \delta_n\right) = 0.$$

For this, we fix an $\varepsilon > 0$. Since $\varepsilon > \lim_{n\to\infty} \delta_n$, this yields

$$\limsup_{n\to\infty} \mathbb{P}\left(|x_0'\widehat{\beta}| \leq \delta_n\right) \leq \limsup_{n\to\infty} \mathbb{P}\left(|x_0'\widehat{\beta}| \leq \varepsilon\right).$$

Since $\varepsilon > 0$ was arbitrary, we conclude

$$\limsup_{n\to\infty} \mathbb{P}\left(|x_0'\widehat{\beta}| \leq \delta_n\right) \leq \lim_{\varepsilon \searrow 0} \limsup_{n\to\infty} \mathbb{P}\left(|x_0'\widehat{\beta}| \leq \varepsilon\right) = 0$$

by assumption.

□

Before proving Proposition 5.17 we have to ensure the existence of a support vector classifier if we are given a linearly separable model as in Definition 5.16, which will be guaranteed by Lemma A.21. Before stating it, we need another definition:

Definition A.20. We call a vector $x_i \in \mathbb{R}^p$ a support vector of the support vector classifier $(\widehat{\alpha}, \widehat{\gamma}')' \in \mathbb{R}^{p+1}$ if it fulfills $y_i = x_i'\widehat{\gamma} + \widehat{\alpha}$, where y_i is the response variable corresponding to x_i.

Lemma A.21. *Let $n \in \mathbb{N}$, $p \in \mathbb{N}$ and $(y_i, x_i')' \in \mathbb{R}^{p+1}$ with $|y_i| = 1$ for $1 \leq i \leq n$. Assume there exists a vector $(a_*, b_*')' \in \mathbb{R}^{p+1}$, such that $y_i = \widetilde{\operatorname{sgn}}(x_i'b_* + a_*)$ holds true for all $1 \leq i \leq n$. Then, the optimization problem in (5.2) has (at least) one solution.*

- *(One-class) If all the y_i belong to the same group, that is, $y_i = y_1$ for all $1 \leq i \leq n$, then every solution $(\widehat{\alpha}, \widehat{\gamma}')'$ to the optimization problem in (5.2) is given by $(\alpha y_1, \mathbf{0}')' \in \mathbb{R}^{p+1}$ where $\alpha \geq 1$ and $\mathbf{0} = (0, \ldots, 0)' \in \mathbb{R}^p$.*

- *(Two-classes) If, otherwise, there are indices $i, j \in \{1, \dots, n\}$ such that $y_i = -y_j$, then the solution $(\widehat{\alpha}, \widehat{\gamma}')'$ to the optimization problem in (5.2) is unique. Furthermore, there is at least one support vector in each group, that is, there exist $k, l \in \{1, \dots, n\}$ with $x_k'\widehat{\gamma} + \widehat{\alpha} = 1$ and $x_l'\widehat{\gamma} + \widehat{\alpha} = -1$.*

In particular, a solution $(\widehat{\alpha}, \widehat{\gamma}')'$ to the optimization problem in (5.2) fulfills $\widehat{\gamma} = \mathbf{0}$ if and only if all y_i belong to the same group.

Proof. We start with the case where all y_i belong to the same group: Clearly, every vector $(\alpha y_1, \mathbf{0}')' \in \mathbb{R}^{p+1}$ with $\alpha \geq 1$ is a solution to the optimization problem. Furthermore, every other solution $(\widetilde{\alpha}, \widetilde{\gamma}')' \in \mathbb{R}^{p+1}$ of the optimization problem fulfills $\widetilde{\gamma} = \mathbf{0}$ due to its optimality. As it also fulfills the constraints we have

$$y_1(x_1'\mathbf{0} + \widetilde{\alpha}) \geq 1 \iff y_1\widetilde{\alpha} \geq 1 \iff \operatorname{sign}(\widetilde{\alpha}) = y_1 \text{ and } |\widetilde{\alpha}| \geq 1.$$

We now consider the case of two classes. Thus, the generating vector $(a_*, b_*')' \in \mathbb{R}^{p+1}$ fulfills $b_* \neq \mathbf{0}$, as otherwise the equation $y_i = \widetilde{\operatorname{sgn}}(x_i'\mathbf{0} + a_*) = \widetilde{\operatorname{sgn}}(a_*)$ would imply all y_i to belong to the same class. With the same argument we can show that there exists (at least) one $i \in \{1, \dots, n\}$ such that $x_i \neq \mathbf{0}$. We start showing that the optimization problem given in (5.2) indeed has (at least) one solution. For this, we will find a feasible point $(a, b')' \in \mathbb{R}^{p+1}$. We would like to emphasize that by definition we have $y_i(x_i'b_* + a_*) \geq 0$, but not necessarily $y_i(x_i'b_* + a_*) \geq 1$ for all $1 \leq i \leq n$. Define $c := \min_{1 \leq i \leq n} y_i(x_i'b_* + a_*)$. If $c > 0$, then the vector $(a_*, b_*')'/c$ fulfills the constraints. If $c = 0$, we define $\delta := \min_{i:y_i=-1} |x_i'b_* + a_*|$. Since $y_i = -1$ implies $x_i'b_* + a_* < 0$, we conclude $\delta > 0$. We then have

$$y_i \left(x_i'b_* + a_* + \frac{\delta}{2} \right) > x_i'b_* + a_* \geq 0$$

whenever $y_i = 1$ and

$$\begin{aligned} y_i \left(x_i'b_* + a_* + \frac{\delta}{2} \right) &= y_i(x_i'b_* + a_*) - \frac{\delta}{2} \\ &= |x_i'b_* + a_*| - \frac{\delta}{2} \geq \frac{\delta}{2} > 0. \end{aligned}$$

whenever $y_i = -1$. Defining $\tilde{c} := \min_{1 \leq i \leq p} |x_i'b_* + a_* + \delta/2|$ we conclude that $\tilde{c} > 0$. Thus, $(a_* + \delta/2, b_*')'/\tilde{c}$ is a feasible solution to (5.2). To sum it up, we showed that there exists (at least) one point fulfilling the constraints in (5.2).[5]

Let $(a_f, b_f')'$ denote the feasible point we found. Then, every other feasible point $(\tilde{a}, \tilde{b}')'$ with $\|\tilde{b}\|_2 \leq \|b_f\|_2$ satisfies

$$y_i\tilde{a} \geq 1 - y_i x_i'\tilde{b} \geq 1 - \max_{1 \leq i \leq n} \|x_i\|_2 \|\tilde{b}\|_2 \geq 1 - \max_{1 \leq i \leq n} \|x_i\|_2 \|b_f\|_2$$

[5] Alternatively, we could also use the Hahn-Banach separation theorem to show the existence of a supporting hyperplane.

for every $1 \leq i \leq n$. Since there are indices k, l with $y_l = -y_k$, we conclude that every feasible point $(\tilde{a}, \tilde{b}')'$ with $\|\tilde{b}\|_2 \leq \|b_f\|_2$ fulfills $|\tilde{a}| \leq |\max_{1\leq i\leq n} \|x_i\|_2 \|b_f\|_2 - 1|$. Now, the restricted optimization problem

$$\begin{aligned} &\min_{(\alpha,\gamma')' \in \mathbb{R}^{p+1}} \|\gamma\|_2 \\ &\text{s.t. } y_i(x_i'\gamma + \alpha) \geq 1 \text{ for all } 1 \leq i \leq n, \\ &|\alpha| \leq |\max_{1\leq i\leq n} \|x_i\|_2 \|b_f\|_2 - 1| \text{ and } \|\gamma\|_2 \leq \|b_f\|_2 \end{aligned} \tag{A.29}$$

has a solution as the objective function is convex and the feasible region is a convex, compact set and nonempty (because it contains the point $(a_f, b_f')'$). Since any feasible point $(a, b')'$ of the original optimization problem given in (5.2) is either also a feasible point of the restricted optimization problem given in (A.29) or fulfills $\|b\|_2 > \|b_f\|_2$, every solution of the restricted optimization problem is also a solution of the original optimization problem. Thus, the original optimization problem has a solution.

Before showing the uniqueness of the solution we prove that for every solution $(\widehat{\alpha}, \widehat{\gamma}')'$ of the optimization problem given in (5.2) there is at least one support vector in each group. For this, define $\mathcal{A} := \{1 \leq i \leq n : |x_i'\widehat{\gamma} + \widehat{\alpha}| = 1\}$ the indices of the support vectors. The set $\mathcal{A}$ is not empty as otherwise scaling $(\widehat{\alpha}, \widehat{\gamma}')'$ by $1/\min_{1\leq i\leq n} |x_i'\widehat{\gamma} + \widehat{\alpha}|$ would yield a better solution to the optimization problem, which is a contradiction to the optimality of $(\widehat{\alpha}, \widehat{\gamma}')'$. Furthermore, we define $\mathcal{P} := \{1 \leq i \leq n : x_i'\widehat{\gamma} + \widehat{\alpha} \geq 1\} \neq \emptyset$ and $\mathcal{N} := \{1 \leq i \leq n : x_i'\widehat{\gamma} + \widehat{\alpha} \leq -1\} \neq \emptyset$ as the indices of the two groups. We would like to point out that we have $\mathcal{P} \cup \mathcal{N} = \{1, \ldots, n\}$, $i \in \mathcal{P} \Leftrightarrow y_i = 1$ and $i \in \mathcal{N} \Leftrightarrow y_i = -1$. We proceed by contradiction: Suppose all support vectors were in group $\mathcal{N}$, that is $\mathcal{A} \subseteq \mathcal{N}$. We will show that this leads to a contradiction. If $\mathcal{A} \subseteq \mathcal{N}$, this would imply $x_i'\widehat{\gamma} + \widehat{\alpha} > 1$ for all $i \in \mathcal{P}$. Let c denote $\min_{i\in\mathcal{P}} x_i'\widehat{\gamma} + \widehat{\alpha} - 1 > 0$ and define $M := \max_{1\leq i\leq n} \|x_i\|_2$. As explained before, we have $M > 0$ and $\|\widehat{\gamma}\|_2 > 0$ in the case of two classes. We now claim that the point $(\widehat{\alpha} - c/2, (1 - \frac{c}{2\|\widehat{\gamma}\|_2 M})\widehat{\gamma}')'$ also fulfills the constraints: For every $i \in \mathcal{P}$ we have

$$\begin{aligned} y_i\left(x_i'\widehat{\gamma}\left(1 - \frac{c}{2\|\widehat{\gamma}\|_2 M}\right) + \widehat{\alpha} - \frac{c}{2}\right) &= x_i'\widehat{\gamma} + \widehat{\alpha} - \frac{c}{2} - x_i'\widehat{\gamma}\frac{c}{2\|\widehat{\gamma}\|_2 M} \\ &\geq 1 + c - \frac{c}{2} - \frac{c}{2}\frac{|x_i'\widehat{\gamma}|}{\|\widehat{\gamma}\|_2 M} \geq 1. \end{aligned}$$

Furthermore, for every $i \in \mathcal{N}$ we have

$$\begin{aligned} y_i\left(x_i'\widehat{\gamma}\left(1 - \frac{c}{2\|\widehat{\gamma}\|_2 M}\right) + \widehat{\alpha} - \frac{c}{2}\right) &= y_i\left(x_i'\widehat{\gamma} + \widehat{\alpha}\right) + \frac{c}{2} + x_i'\widehat{\gamma}\frac{c}{2\|\widehat{\gamma}\|_2 M} \\ &\geq 1 + \frac{c}{2} - \frac{c}{2}\frac{|x_i'\widehat{\gamma}|}{\|\widehat{\gamma}\|_2 M} \geq 1. \end{aligned}$$

We will now show that $c \leq 2\|\widehat{\gamma}\|_2 M$ holds true. Since the set of support vectors $\mathcal{A}$ is not empty, we can find an index $j \in \mathcal{A}$, such that $y_j(x_j\widehat{\gamma} + \widehat{\alpha}) = 1$ holds true, which implies

$|\widehat{\alpha}| \leq 1 + M\|\widehat{\gamma}\|_2$. Thus, we have

$$c = \min_{i \in \mathcal{P}} x_i'\widehat{\gamma} + \widehat{\alpha} - 1 \leq M\|\widehat{\gamma}\|_2 + \widehat{\alpha} - 1 \leq 2M\|\widehat{\gamma}\|_2.$$

Since this implies $\frac{c}{2\|\widehat{\gamma}\|_2 M} \leq 1$ and c is positive, we have $\|(1 - \frac{c}{2\|\widehat{\gamma}\|_2 M})\widehat{\gamma}\|_2 < \|\widehat{\gamma}\|_2$, which is a contradiction to the optimality of $(\widehat{\alpha}, \widehat{\gamma}')'$. With the same argument we can find a contradiction to $\mathcal{A} \subseteq \mathcal{P}$. Thus, every group has (at least) one support vector.

We now prove the uniqueness of the solution of the optimization problem given in (5.2). Assume $(\alpha_1, \gamma_1')'$ and $(\alpha_2, \gamma_2')'$ are two solutions of the optimization problem. Since $f(x) = \|x\|_2$ is strictly convex, we conclude $\gamma_1 = \gamma_2$ as otherwise a convex combination of the two points would give a better solution. In order to show that $\alpha_1 = \alpha_2$ we consider the support vectors of the solution $(\alpha_1, \gamma_1')'$. As shown before, there are indices $i, j \in \{1, \ldots, n\}$ such that

$$\begin{aligned} x_i'\gamma_1 + \alpha_1 &= 1 \text{ and} \\ x_j'\gamma_1 + \alpha_1 &= -1. \end{aligned}$$

As $(\alpha_2, \gamma_2')'$ is also a solution, we have

$$\begin{aligned} 1 \leq x_i'\gamma_2 + \alpha_2 = x_i'\gamma_1 + \alpha_2 = 1 + \alpha_2 - \alpha_1 \text{ and} \\ -1 \geq x_j'\gamma_2 + \alpha_2 = x_j'\gamma_1 + \alpha_2 = -1 + \alpha_2 - \alpha_1. \end{aligned}$$

Rearranging the equations in the preceding display yields $\alpha_2 - \alpha_1 \geq 0$ and $\alpha_2 - \alpha_1 \leq 0$, which proves $\alpha_1 = \alpha_2$. Hence, the solution is unique.

It remains to show the last statement. If all y_i belong to the same group, then every solution $(\widehat{\alpha}, \widehat{\gamma}')'$ fulfills $\|\widehat{\gamma}\|_2 = 0$ as shown before. Furthermore, if the solution fulfills $\widehat{\gamma} = \mathbf{0}$, we then have $1 \leq y_i(x_i'\widehat{\gamma} + \widehat{\alpha}) = y_i\widehat{\alpha}$ for $1 \leq i \leq n$, which implies $y_i = \text{sign}(\widehat{\alpha})$. Hence, all y_i belong to the same group. □

For the proof of Proposition 5.17 we need the following result:

Lemma A.22. *Assume a linearly separable binary model as in Definition 5.16. Let $(\widehat{\alpha}, \widehat{\gamma}')' \in \mathbb{R}^{p+1}$ denote the support vector classifier with respect to the training data $T_n = (y_i, x_i')_{i=1}^n$ and $(\widetilde{\alpha}, \widetilde{\gamma}')' \in \mathbb{R}^{p+1}$ denote the support vector classifier with respect to the training data $T_{n-1} = (y_i, x_i')_{i=1}^{n-1}$. If $\|\widehat{\gamma}\|_2 > \|\widetilde{\gamma}\|_2$, then $y_n(x_n'\widehat{\gamma} + \widehat{\alpha}) = 1$, that is, x_n is a support vector for $(\widehat{\alpha}, \widehat{\gamma}')'$.*

Proof. We start with the observation that $y_n(x_n'\widehat{\gamma} + \widehat{\alpha}) \geq 1$ as $(\widehat{\alpha}, \widehat{\gamma}')'$ satisfies the constraints. Moreover, we have $y_n(x_n'\widetilde{\gamma} + \widetilde{\alpha}) < 1$ as otherwise $(\widetilde{\alpha}, \widetilde{\gamma}')$ would also satisfy the constraints contradicting the optimality of $(\widehat{\alpha}, \widehat{\gamma}')'$. We now assume that $y_n(x_n'\widehat{\gamma} + \widehat{\alpha}) > 1$ and find a contradiction. Define $c_1 := y_n(x_n'\widehat{\gamma} + \widehat{\alpha}) - 1 > 0$ and $c_2 := 1 - y_n(x_n'\widetilde{\gamma} + \widetilde{\alpha}) > 0$ and $\varepsilon := c_1/(c_1 + c_2) \in (0, 1)$. We now claim that the point $(\bar{\alpha}, \bar{\gamma}')' := (1 - \varepsilon)(\widehat{\alpha}, \widehat{\gamma}')' + \varepsilon(\widetilde{\alpha}, \widetilde{\gamma}')'$ also satisfies the constraints. As $(\bar{\alpha}, \bar{\gamma}')'$ is a convex combination, it fulfills the

constraints $y_i(x_i'\bar{\gamma} + \bar{\alpha}) \geq 1$ for all $1 \leq i \leq n-1$. Furthermore, we have

$$\begin{aligned} y_n(x_n'\bar{\gamma} + \bar{\alpha}) &= (1-\varepsilon) y_n(x_n'\widehat{\gamma} + \widehat{\alpha}) + \varepsilon y_n(x_n'\widetilde{\gamma} + \widetilde{\alpha}) \\ &= \frac{c_2}{c_1 + c_2}(c_1 + 1) + \frac{c_1}{c_1 + c_2}(1 - c_2) = 1. \end{aligned}$$

To put it in other words, $(\bar{\alpha}, \bar{\gamma}')'$ also satisfies the constraints $y_i(x_i'\bar{\gamma} + \bar{\alpha}) \geq 1$ for all $1 \leq i \leq n$. As $\|\widetilde{\gamma}\|_2 < \|\widehat{\gamma}\|_2$ we have $\|\bar{\gamma}\|_2 < \|\widehat{\gamma}\|_2$, which is a contradiction to the optimality of $(\widehat{\alpha}, \widehat{\gamma}')' \in \mathbb{R}^{p+1}$. □

Proof of Proposition 5.17: By Lemma A.21 there is indeed at least one support vector classifier in the linearly separable case of Definition 5.16. Furthermore, Lemma A.21 shows that in the case of two classes the support vector classifier is unique and hence the corresponding predictor is unique. In the case where all y_i belong to the same group the predictor with respect to these training data assigns every new observation to y_1. To see this, we point out that every support vector classifier is given by $(\alpha y_1, \mathbf{0}')'$ with $\alpha \geq 1$, which yields

$$\hat{y}_0 = \widetilde{\operatorname{sgn}}(x_0'\mathbf{0} + \alpha y_1) = \widetilde{\operatorname{sgn}}(\alpha y_1) = y_1$$

regardless of the choice of $\alpha \geq 1$. Thus, in any of the two cases the predictor is unique.

We now show that the number of support vectors $|\mathcal{A}|$ is bounded by $p+1$ almost surely unless all of the $(y_i)_{i=1}^n$ are in the same group. In principle, our argument will be the following: all support vectors are contained in two hyperplanes in $\mathbb{R}^p$, which can be defined by the first $p+1$ support vectors. Then, the probability of another vector x_i lying on that hyperplane is 0 as their distribution is absolutely continuous (with respect to the p-dimensional Lebesgue-measure).

To make the argument precise, we recall that a support vector fulfills $y_i = x_i'\widehat{\gamma} + \widehat{\alpha}$ by definition. Thus, whenever there are more than $p+1$ support vectors, there exist distinct indices $i_1, \ldots, i_{p+2}$ and a vector $(\alpha, \gamma')' \in \mathbb{R}^{p+1}$, such that $y_{i_j} = x_{i_j}'\gamma + \alpha$ for all $1 \leq j \leq p+2$. As the pairs (y_i, x_i') are i.i.d and there are exactly $\binom{n}{p+2}$ possible combinations of $p+2$ vectors of a group of n data, we have

$$\begin{aligned} &\mathbb{P}\left(|\mathcal{A}| > p+1, |\mathbf{1}_n'Y| < n\right) \\ &\leq \binom{n}{p+2} \mathbb{P}\left(\exists (\alpha, \gamma')' \in \mathbb{R}^{p+1} : y_i = x_i'\gamma + \alpha \text{ for all } 1 \leq i \leq p+2, |\mathbf{1}_n'Y| < n\right), \end{aligned}$$

where $Y = (y_1, \ldots, y_n)' \in \mathbb{R}^n$ and $\mathbf{1}_n = (1, \ldots, 1)' \in \mathbb{R}^n$. We would like to emphasize that $|\mathbf{1}_n'Y|$ equals n if and only if all y_i belong to the same group and thus $|\mathbf{1}_n'Y| < n$ is just the mathematical formulation of the two-classes case. Hence, we conclude that γ cannot be the zero vector as long as $|\mathbf{1}_n'Y| < n$ holds true. Thus, we can bound the last line in the preceding display from above by

$$\binom{n}{p+2} \mathbb{P}\left(\exists (\alpha, \gamma')' \in \mathbb{R}^{p+1}, \gamma \neq \mathbf{0} : y_i = x_i'\gamma + \alpha \text{ for all } 1 \leq i \leq p+2\right).$$

Define $X = (x_1, \dots, x_{p+1})' \in \mathbb{R}^{(p+1)\times p}$. Since the x_i are independent and absolutely continuous random vectors, the matrix $\left[\mathbf{1}_{p+1}, X\right]$ has full rank $p+1$ almost surely and hence is invertible. Thus, the vector $(\alpha, \gamma')' \in \mathbb{R}^{p+1}$ fulfilling the first $p+1$ equations $y_i = x_i'\gamma + \alpha$ for $1 \leq i \leq p+1$ is unique and given by

$$\begin{pmatrix} \alpha \\ \gamma \end{pmatrix} = \left[\mathbf{1}_{p+1}, \quad X\right]^{-1} \bar{Y},$$

where $\bar{Y} = (y_1, \dots, y_{p+1})' \in \mathbb{R}^{p+1}$. Hence, we have

$$\begin{aligned} &\mathbb{P}\left(\exists (\alpha, \gamma')' \in \mathbb{R}^{p+1}, \gamma \neq \mathbf{0} : y_i = x_i'\gamma + \alpha \text{ for all } 1 \leq i \leq p+2\right) \\ &\leq \mathbb{P}\left(y_{p+2} = \bar{\alpha} + \bar{\gamma}' x_{p+2}, (\bar{\alpha}, \bar{\gamma}')' = \left[\mathbf{1}_{p+1}, \quad X\right]^{-1} \bar{Y}, \bar{\gamma} \neq \mathbf{0}\right) \\ &= \mathbb{E}\left(\mathbb{1}\{\bar{\gamma} \neq \mathbf{0}\} \mathbb{P}\left(y_{p+2} = \bar{\alpha} + \bar{\gamma}' x_{p+2} \| \bar{\alpha}, \bar{\gamma}\right)\right). \end{aligned}$$

Now, as $\bar{\alpha}$ and $\bar{\gamma}$ are measurable functions of the first $p+1$ observations and hence independent of (y_{p+2}, x_{p+2}'), we have for every $\bar{\gamma} \neq \mathbf{0}$

$$\mathbb{P}\left(y_{p+2} = \bar{\alpha} + \bar{\gamma}' x_{p+2} \| \bar{\alpha}, \bar{\gamma}\right) \leq \mathbb{P}\left(|\bar{\gamma}' x_{p+2} + \bar{\alpha}| = 1 \| \bar{\alpha}, \bar{\gamma}\right).$$

Since for every fixed $(\bar{\alpha}, \bar{\gamma})$ with $\bar{\gamma} \neq \mathbf{0}$ the set $\mathcal{S} := \{x \in \mathbb{R}^p : |\bar{\gamma}' x + \bar{\alpha}| = 1\}$ is the union of two hyperplanes in $\mathbb{R}^p$, we have $\lambda_p(\mathcal{S}) = 0$, where λ_p denotes the p-dimensional Lebesgue-measure. By the absolute continuity of x_{p+2} we have $\mathbb{P}(x_{p+2} \in \mathcal{S}) = 0$. Thus, the number of support vectors is bounded by $p+1$ almost surely unless all y_i are in the same group.

To show the stability of the predictor based on the support vector classifier, we distinguish two cases: If all y_i are in the same group, any support vector classifier $(\widehat{\alpha}, \widehat{\gamma}')$ of the training data fulfills $\widehat{\gamma} = \mathbf{0}$. Hence, the predictor $\hat{y}_0$ automatically assigns the new observation to the same class as y_1. With the same argumentation one can show that $\tilde{y}_0 = y_1$ as well. Thus, in that case we have $\hat{y}_0 - \tilde{y}_0 = 0$.

We now consider the case where not all y_i are in the same group. Let $(\widehat{\alpha}, \widehat{\gamma}')' \in \mathbb{R}^{p+1}$ denote the (unique) support vector classifier based on T_n, $(\widetilde{\alpha}, \widetilde{\gamma}')' \in \mathbb{R}^{p+1}$ denote a support vector classifier based on T_{n-1} and $\mathcal{A} := \{1 \leq i \leq n : |x_i'\widehat{\gamma} + \widehat{\alpha}| = 1\}$ denote the indices of the support vectors of $(\widehat{\alpha}, \widehat{\gamma}')'$. We would like to emphasize that $\|\widehat{\gamma}\|_2 > 0$ as we are in the case of two classes. We now claim that $(\widehat{\alpha}, \widehat{\gamma}')' \neq (\widetilde{\alpha}, \widetilde{\gamma}')'$ implies x_n to be a support vector of $(\widehat{\alpha}, \widehat{\gamma}')'$. For this, we distinguish two cases:

If $\|\widehat{\gamma}\|_2 = \|\widetilde{\gamma}\|_2$, then $(\widehat{\alpha}, \widehat{\gamma}')'$ is also a support vector classifier with respect to the data T_{n-1}. As $\|\widetilde{\gamma}\|_2 = \|\widehat{\gamma}\|_2 > 0$, we conclude that even in the set $\{1, \dots, n-1\}$ there are indices i, j with $y_i = -y_j$. Thus, by Lemma A.21 the support vector classifier with respect to the data T_{n-1} is unique, which yields $(\widehat{\alpha}, \widehat{\gamma}')' = (\widetilde{\alpha}, \widetilde{\gamma}')'$. Hence, the case $(\widehat{\alpha}, \widehat{\gamma}')' \neq (\widetilde{\alpha}, \widetilde{\gamma}')'$ can only occur if $\|\widehat{\gamma}\|_2 > \|\widetilde{\gamma}\|_2$.

However, if $\|\widehat{\gamma}\|_2 > \|\widetilde{\gamma}\|_2$ we are able to apply Lemma A.22, which implies x_n to be a support vector of $(\widehat{\alpha}, \widehat{\gamma}')'$. To sum it up, in the case where not all y_i belong to the same class we have $\hat{y}_0 \neq \tilde{y}_0$ only if x_n is a support vector of $(\widehat{\alpha}, \widehat{\gamma}')'$, that is, $n \in \mathcal{A}$. Now, in the case of two classes the probability of being a support vector can be bounded from

above by $(p+1)/n$: To see this, we recall that in this case the number of support vectors is bounded by $p+1$ almost surely, which yields

$$p+1 \geq \mathbb{E}\left(|\mathcal{A}|\mathbb{1}\{|Y'\mathbf{1}_n| < n\}\right) = \mathbb{E}\left(\sum_{i=1}^{n} \mathbb{1}\{i \in \mathcal{A}\}\mathbb{1}\{|Y'\mathbf{1}_n| < n\}\right)$$
$$= \sum_{i=1}^{n} \mathbb{E}\left(\mathbb{1}\{i \in \mathcal{A}, |Y'\mathbf{1}_n| < n\}\right) = \sum_{i=1}^{n} \mathbb{P}\left(i \in \mathcal{A}, |Y'\mathbf{1}_n| < n\right).$$

As the data are i.i.d. and $Y'\mathbf{1}_n$ does not change by a permutation of the data we conclude $p+1 \geq n\mathbb{P}\left(n \in \mathcal{A}, |Y'\mathbf{1}_n| < n\right)$. Putting the pieces together, we have

$$\mathbb{P}(\hat{y}_0 \neq \tilde{y}_0) \leq \mathbb{P}(\hat{y}_0 \neq \tilde{y}_0, |Y'\mathbf{1}_n| < n) + \mathbb{P}(\hat{y}_0 \neq \tilde{y}_0, |Y'\mathbf{1}_n| = n)$$
$$\leq \mathbb{P}\left(n \in \mathcal{A}, |Y'\mathbf{1}_n| < n\right) + 0 \leq \frac{p+1}{n}.$$

In the case where $\lim_{n\to\infty} p/n = 0$, this also proves the asymptotic stability. □

A.5. Proofs of Chapter 6

Proof of Lemma 6.1: For the first part we can find for every $\delta > 0$ a t_δ, such that

$$\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n) = \sup_{t\in\mathbb{R}} \inf_{x,y\in K_{\varepsilon/2}(t)} |\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(y)| \leq \inf_{x,y\in K_{\varepsilon/2}(t_\delta)} |\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(y)| + \delta.$$

Now, for any $p > 0$ we have

$$\inf_{x,y\in K_{\varepsilon/2}(t_\delta)} |\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(y)| \leq \inf_{x\in K_{\varepsilon/2}(t_\delta)} |\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(x)|$$
$$= \left[\inf_{x\in K_{\varepsilon/2}(t_\delta)} |\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(x)|^p\right]^{\frac{1}{p}}.$$

Since we have $(b-a)\inf_{x\in[a,b]} f(x) \leq \int_{[a,b]} f(x)d\lambda(x)$, the last line in the preceding display can be bounded from above by

$$\left[\frac{1}{\lambda(K_{\varepsilon/2}(t_\delta))}\int_{K_{\varepsilon/2}(t_\delta)} |\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(x)|^p d\lambda(x)\right]^{\frac{1}{p}} \leq \left[\frac{\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_p}^p}{\varepsilon}\right]^{\frac{1}{p}},$$

where λ denotes the Lebesgue-measure on $\mathbb{R}$. Moreover, it can also easily be seen that

$$\inf_{x\in K_{\varepsilon/2}(t_\delta)} |\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(x)| \leq \|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_\infty}.$$

As $\delta > 0$ can be made arbitrarily small, the statement follows.

For the continuous case we start with $c \in [0, \varepsilon)$, which will be defined later on. For

every $a, b \in \mathbb{R}$ and $s \in [a, b)$ we have

$$\begin{aligned}|\widehat{\mathrm{F}}_n(s) - \mathrm{F}_n(s)| &\leq \max(\widehat{\mathrm{F}}_n(b) - \mathrm{F}_n(a), \mathrm{F}_n(b) - \widehat{\mathrm{F}}_n(a)) \\ &\leq \max(|\widehat{\mathrm{F}}_n(a) - \mathrm{F}_n(a)|, |\widehat{\mathrm{F}}_n(b) - \mathrm{F}_n(b)|) + \mathrm{F}_n(b) - \mathrm{F}_n(a) \\ &\leq \max(|\widehat{\mathrm{F}}_n(a) - \mathrm{F}_n(a)|, |\widehat{\mathrm{F}}_n(b) - \mathrm{F}_n(b)|) + \mathbb{E}(\min(1, (b-a)\|f_{y_0\|x_0}\|_\infty)),\end{aligned}$$

where we used equation (A.17) for the last inequality. As the upper bound does not depend on s (but on the length of the interval $[a, b)$), we can use it as a uniform bound on the interval and get

$$\begin{aligned}\sup_{s\in\mathbb{R}} |\widehat{\mathrm{F}}_n(s) - \mathrm{F}_n(s)| &= \sup_{j\in\mathbb{Z}} \sup_{s\in[j\varepsilon+c,(j+1)\varepsilon+c)} |\widehat{\mathrm{F}}_n(s) - \mathrm{F}_n(s)| \\ &\leq \sup_{j\in\mathbb{Z}} \max(|\widehat{\mathrm{F}}_n(j\varepsilon+c) - \mathrm{F}_n(j\varepsilon+c)|, |\widehat{\mathrm{F}}_n((j+1)\varepsilon+c) - \mathrm{F}_n((j+1)\varepsilon+c)|) \\ &\quad + \mathbb{E}(\min(1, \varepsilon\|f_{y_0\|x_0}\|_\infty)),\end{aligned}$$

which again can be bounded from above by

$$\begin{aligned}&\mathbb{E}(\min(1, \varepsilon\|f_{y_0\|x_0}\|_\infty)) + \sup_{j\in\mathbb{Z}} |\widehat{\mathrm{F}}_n(j\varepsilon+c) - \mathrm{F}_n(j\varepsilon+c)| \\ &\leq \mathbb{E}(\min(1, \varepsilon\|f_{y_0\|x_0}\|_\infty)) + \left[\sum_{j\in\mathbb{Z}} |\widehat{\mathrm{F}}_n(j\varepsilon+c) - \mathrm{F}_n(j\varepsilon+c)|^p\right]^{\frac{1}{p}}.\end{aligned}$$

We would like to emphasize that the inequality above holds true for every $c \in [0, \varepsilon)$. By choosing our c arbitrarily close to the infimum provided in Lemma A.8, we have

$$\sup_{s\in\mathbb{R}} |\widehat{\mathrm{F}}_n(s) - \mathrm{F}_n(s)| \leq \mathbb{E}(\min(1, \varepsilon\|f_{y_0\|x_0}\|_\infty)) + \left[\frac{1}{\varepsilon}\int_{\mathbb{R}} |\widehat{\mathrm{F}}_n(s) - \mathrm{F}_n(s)|^p d\lambda(s)\right]^{\frac{1}{p}}.$$

□

Proof of Proposition 6.2: We fix an $M \geq 0$. Applying Tonelli's theorem and Jensen's inequality yields

$$\begin{aligned}\mathbb{E}\left(\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_1}\right) &= \int_{\mathbb{R}} \mathbb{E}\left(\left|\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(x)\right|\right) d\lambda(x) \\ &\leq \int_{[-M,M]} \left(\mathbb{E}\left((\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(x))^2\right)\right)^{\frac{1}{2}} d\lambda(x) \\ &\quad + \int_{(M,\infty)} \mathbb{E}((1 - \widehat{\mathrm{F}}_n(x)) + (1 - \mathrm{F}_n(x)))d\lambda(x) \\ &\quad + \int_{(-\infty,-M)} \mathbb{E}(\widehat{\mathrm{F}}_n(x) + \mathrm{F}_n(x))d\lambda(x).\end{aligned}$$

By Lemma A.13 we have $\mathbb{E}((\widehat{\mathrm{F}}_n(x)-\mathrm{F}_n(x))^2) \leq \frac{1}{4(n-1)}+\frac{5}{n}\sum_{i=1}^n \mathbb{E}(\min(1, \|f_{y_0\|x_0}\|_\infty|\hat{y}_0-\tilde{y}_0^{[-i]}|)$. By the symmetry of the predictor we have $\mathbb{E}(\min(1, \|f_{y_0\|x_0}\|_\infty|\hat{y}_0 - \tilde{y}_0^{[-i]}|)) = \mathbb{E}(\min(1, \|f_{y_0\|x_0}\|_\infty|\hat{y}_0 - \tilde{y}_0|))$ for all $1 \leq i \leq n$.

As $\hat{u}_i$ are identically distributed like $y_0 - \tilde{y}_0$, we have $\mathbb{E}(\widehat{\mathrm{F}}_n(x)) = \mathbb{P}(y_0 - \tilde{y}_0 \leq x)$. Furthermore, we have $\mathbb{E}(\mathrm{F}_n(x)) = \mathbb{P}(y_0 - \hat{y}_0 \leq x)$, which yields

$$\int_{(M,\infty)} \mathbb{E}((1 - \mathrm{F}_n(x))d\lambda(x) = \int_{(M,\infty)} \mathbb{P}(y_0 - \hat{y}_0 > x)d\lambda(x).$$

As the set of all discontinuity points of F_n is countable, the term in the preceding display coincides with

$$\int_{(M,\infty)} \mathbb{P}(y_0 - \hat{y}_0 \geq x)d\lambda(x).$$

We can bound the expression from above by

$$\int_{[0,\infty)} \mathbb{P}((y_0 - \hat{y}_0)\mathbb{1}_{[M,\infty)}(y_0 - \hat{y}_0) \geq x)d\lambda(x) = \mathbb{E}((y_0 - \hat{y}_0)\mathbb{1}_{[M,\infty)}(y_0 - \hat{y}_0)),$$

which holds true because $M \geq 0$. With a similar argument one can show that

$$\int_{(-\infty,-M)} \mathbb{E}(\mathrm{F}_n(x))d\lambda(x) \leq -\mathbb{E}((y_0 - \hat{y}_0)\mathbb{1}_{(-\infty,-M]}(y_0 - \hat{y}_0)),$$

which yields

$$\begin{aligned}&\int_{(M,\infty)} \mathbb{E}((1 - \mathrm{F}_n(x)))d\lambda(x) + \int_{(-\infty,-M)} \mathbb{E}(\mathrm{F}_n(x))d\lambda(x)\\ &\leq \mathbb{E}\left(|y_0 - \hat{y}_0|\mathbb{1}_{[M,\infty)}(|y_0 - \hat{y}_0|)\right).\end{aligned}$$

This procedure can be repeated for $\widehat{\mathrm{F}}_n$ with a minor modification: We have

$$\mathbb{E}(1 - \widehat{\mathrm{F}}_n(x)) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(1 - \mathbb{1}_{[\hat{u}_i,\infty)}(x)) = \frac{1}{n}\sum_{i=1}^n \mathbb{P}(\hat{u}_i > x).$$

Now, as for every $1 \leq i \leq n$ the leave-one-out errors $\hat{u}_i = y_i - \tilde{y}_i^{[-i]}$ have the same distribution as $y_0 - \tilde{y}_0$, the expression in the preceding display equals $\mathbb{P}(y_0 - \tilde{y}_0 > x)$. Now, we can proceed as before to get

$$\begin{aligned}&\int_{(M,\infty)} \mathbb{E}((1 - \widehat{\mathrm{F}}_n(x)))d\lambda(x) + \int_{(-\infty,-M)} \mathbb{E}(\widehat{\mathrm{F}}_n(x))d\lambda(x)\\ &\leq \mathbb{E}\left(|y_0 - \tilde{y}_0|\mathbb{1}_{[M,\infty)}(|y_0 - \tilde{y}_0|)\right),\end{aligned}$$

which finishes the proof. □

Proof of Proposition 6.3: By the boundedness of $\|f_{y_0\|x_0}\|_\infty$ and the asymptotic stability we have

$$c_n := \frac{1}{4(n-1)} + 5\,\mathbb{E}(\min(1, \|f_{y_0\|x_0}\|_\infty |\hat{y}_0 - \tilde{y}_0|))) \underset{n\to\infty}{\longrightarrow} 0.$$

Now we can define a sequence $(M_n)_{n\in\mathbb{N}}$, such that $M_n \to \infty$ while at the same time $M_n c_n^{1/2} \to 0$ holds true. The uniform integrability of $y_0 - \hat{y}_0$ and $y_0 - \tilde{y}_0$ implies the existence of their first moments. Thus, we can apply for every $n \in \mathbb{N}$ the inequality of Proposition 6.2. Taking the limit $n \to \infty$, the uniform integrability yields

$$\lim_{n\to\infty} \mathbb{E}(\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_1}) = 0.$$

To finish the proof note that $|\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(x)| \leq 1$ for all x, which yields the bound

$$\mathbb{E}(\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_p}) \leq \mathbb{E}\left[\left(\int_{\mathbb{R}} |\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(x)| d\lambda(x)\right)^{\frac{1}{p}}\right] \leq \left[\mathbb{E}(\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_1})\right]^{\frac{1}{p}}$$

for any $p \geq 1$. □

Proof of Lemma 6.4: The first part is basically a small modification of Lemma C.3 in Steinberger and Leeb (2023), which itself is a conclusion from Lemma C.1 therein. While Lemma C.1 of Steinberger and Leeb (2023) can be applied to the more general case of k-fold cross-validation, we here are dealing with the case of leave-one-out residuals. Thus, we apply Lemma C.1 to the case where $k = n$. As in Lemma C.3. of Steinberger and Leeb (2023), our loss function is $L(f, z) = \mathbb{1}_{(-\infty,s]}(y - f(x))$, which implies $C = 1$. Our main modification will be to find an alternative bound for the expression

$$\frac{1}{n}\frac{1}{n(n-1)} \sum_{i\neq j} \mathbb{E}\left(\mathbb{1}_{(-\infty,s]}(\hat{u}_i)(1 - \mathbb{1}_{(-\infty,s]}(\hat{u}_j))\right),$$

which is the last line of the upper bound of $\mathbb{E}(R_{CV}^2)$ in Lemma C.1. In Steinberger and Leeb (2023) the expression of the preceding display is bounded from above by $1/(4(n-1))$. However, we need an upper bound which is integrable over the whole real line. Since $\|f\|_{\mathcal{L}_2} = \infty$ for any constant nonzero function f, we need to find a better bound. In order to do so, we will use the following equality instead:

$$\begin{aligned}
&\frac{1}{n}\frac{1}{n(n-1)} \sum_{i\neq j} \mathbb{E}\left(\mathbb{1}_{(-\infty,s]}(\hat{u}_i)(1 - \mathbb{1}_{(-\infty,s]}(\hat{u}_j))\right) = \frac{1}{n}\frac{1}{n(n-1)} \sum_{i\neq j} \mathbb{P}\left(\hat{u}_i \leq s < \hat{u}_j\right) \\
&= \frac{1}{2n}\left(\frac{2}{n(n-1)} \sum_{i<j} \mathbb{P}\left(\hat{u}_i \leq s < \hat{u}_j\right) + \frac{2}{n(n-1)} \sum_{i>j} \mathbb{P}\left(\hat{u}_i \leq s < \hat{u}_j\right)\right) \\
&= \frac{1}{2n}\left(\mathbb{P}\left(\hat{u}_1 \leq s < \hat{u}_2\right) + \mathbb{P}\left(\hat{u}_2 \leq s < \hat{u}_1\right)\right),
\end{aligned}$$

where the last equality is given by the symmetry of the predictor. If we replace the factor

$1/(4(n-1))$ by $\frac{1}{2n}(\mathbb{P}(\hat{u}_1 \leq s < \hat{u}_2) + \mathbb{P}(\hat{u}_2 \leq s < \hat{u}_1))$ in the remaining part of the proof of Lemma C.1 in Steinberger and Leeb (2023), we end up with

$$\begin{aligned}\mathbb{E}\left((\widehat{\mathrm{F}}_n(s) - \mathrm{F}_n(s))^2\right) \leq \frac{1}{2n}\left(\mathbb{P}(\hat{u}_1 \leq s < \hat{u}_2) + \mathbb{P}(\hat{u}_2 \leq s < \hat{u}_1)\right) \\ + 5\,\mathbb{E}\left|\mathbb{1}_{(-\infty,s]}(y_0 - \hat{y}_0) - \mathbb{1}_{(-\infty,s]}(y_0 - \tilde{y}_0)\right|,\end{aligned}$$

where the right-hand side of the preceding display can be rewritten as

$$\begin{aligned}&\frac{1}{2n}\left(\mathbb{P}(\hat{u}_1 \leq s < \hat{u}_2) + \mathbb{P}(\hat{u}_2 \leq s < \hat{u}_1)\right) \\ &+ 5\left(\mathbb{P}(y_0 - \hat{y}_0 \leq s < y_0 - \tilde{y}_0) + \mathbb{P}(y_0 - \tilde{y}_0 \leq s < y_0 - \hat{y}_0)\right).\end{aligned}$$

Turning back to our original task, we use Tonelli's theorem to bound $\mathbb{E}\left(\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_2}^2\right)$ by

$$\begin{aligned}&\int_{\mathbb{R}} \mathbb{E}\left((\widehat{\mathrm{F}}_n(s) - \mathrm{F}_n(s))^2\right) d\lambda(s) \\ &\leq \frac{1}{2n}\int_{\mathbb{R}} \left(\mathbb{P}(\hat{u}_1 \leq s < \hat{u}_2) + \mathbb{P}(\hat{u}_2 \leq s < \hat{u}_1)\right) d\lambda(s) \\ &+ 5\int_{\mathbb{R}} \left(\mathbb{P}(y_0 - \hat{y}_0 \leq s < y_0 - \tilde{y}_0) + \mathbb{P}(y_0 - \tilde{y}_0 \leq s < y_0 - \hat{y}_0)\right) d\lambda(s).\end{aligned}$$

Now the claim follows by applying Lemma A.7. □

Proof of Corollary 6.5: The case $p = 2$ can be trivially concluded from Lemma 6.4 and Jensen's inequality. For the case $p \in (2, \infty)$ we use the fact that $|\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(x)| \leq 1$ for all $x \in \mathbb{R}$ together with Jensen's inequality to get

$$\begin{aligned}\mathbb{E}\left(\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_p}\right) &\leq \mathbb{E}\left(\left[\int_{\mathbb{R}} |\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(x)|^2 d\lambda(x)\right]^{\frac{1}{p}}\right) \\ &\leq \left[\mathbb{E}\left(\left(\int_{\mathbb{R}} |\widehat{\mathrm{F}}_n(x) - \mathrm{F}_n(x)|^2 d\lambda(x)\right)^{\frac{1}{2}}\right)\right]^{\frac{2}{p}} = \left[\mathbb{E}\left(\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_2}\right)\right]^{\frac{2}{p}}.\end{aligned}$$

Finally, the statements for $\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n)$ and $\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_\infty$ follow from Lemma 6.1 and an appropriate choice of ε_n, for example $\varepsilon_n = \mathbb{E}(\|\widehat{\mathrm{F}}_n - \mathrm{F}_n\|_{\mathcal{L}_2})$. □

Proof of Lemma 6.6: Let x_0 take the value 1 with probability q and -1 otherwise, y_0 be independent from x_0 and uniformly distributed on the interval $[-1/2, 1/2]$. Let $K > \varepsilon + 1$ and define $\hat{y}_0 := Kx_0\Pi_{i=1}^n x_i$ and its leave-one-out analogue $\tilde{y}_0^{[-i]} := Kx_0\Pi_{j\neq i} x_j$. We then

have for any $0 < \delta < K$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(|\hat{y}_0 - \tilde{y}_0^{[-i]}| \geq \delta) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(|x_0(x_i - 1)\Pi_{j\neq i}x_j| \geq \frac{\delta}{K}) \\ = \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(x_i = -1) = 1 - q.$$

Hence, the predictor is not stable whenever $q < 1$. For the leave-one-out residuals we have

$$\hat{u}_i = y_i - \tilde{y}_0^{[-i]} = y_i - Kx_i\Pi_{j\neq i}x_j = y_i - K\Pi_{i=1}^{n}x_i.$$

We distinguish two cases: if $\Pi_{i=1}^{n}x_i = 1$, then all the leave-one-out residuals are smaller than (or equal to) $1/2 - K$. Thus, any ε-inflated prediction interval $PI^{+}_{\alpha_1,\alpha_2}(\varepsilon)$ contains no values larger than $x_0K + 1/2 + \varepsilon - K < x_0K - 1/2$. Hence, the prediction interval has no intersection with $[-1/2, 1/2]$ whenever $x_0 = -1$, which yields

$$\mathbb{P}\left(y_0 \in PI^{+}_{\alpha_1,\alpha_2}(\varepsilon) \| \Pi_{i=1}^{n}x_i = 1\right) \leq \mathbb{P}\left(x_0 = 1 \| \Pi_{i=1}^{n}x_i = 1\right) = q.$$

We can show with the same argument that in the case $\Pi_{i=1}^{n}x_i = -1$ the prediction interval contains only values larger than $-x_0K + 1/2$, which yields the same result for the conditional coverage probability of $PI^{+}_{\alpha_1,\alpha_2}(\varepsilon)$. To sum it up, we have for any training data T_n

$$\mathbb{P}\left(y_0 \in PI^{+}_{\alpha_1,\alpha_2}(\varepsilon) \| T_n\right) \leq q$$

regardless of the choice of α_1, α_2 and ε as long as we choose $K > 1 + \varepsilon$. The absolute value of the prediction error $|y_0 - \hat{y}_0|$ is bounded by $K + 1/2$ and by the independence of y_0 and x_0 we have $\|f_{y_0\|x_0}\|_\infty = 1$, fulfilling Assumption **CC1**. Hence, the only reason for the failing of the Jackknife-approach is the predictor's instability. We would like to point out that - although Assumption **CC1** is fulfilled - even an enlargement of the prediction interval by a finite length ε will not solve the problem. □

Proof of Lemma 6.7. We point out that F_1 is strictly increasing if and only if F_2 is. W.l.o.g. we assume $\lambda > 1$ as otherwise we can change the roles of F_1 and F_2, yielding $F_2(t) = F_1(t\frac{1}{\lambda})$ with $\frac{1}{\lambda} > 1$ (and the case $\lambda = 1$ would imply $F_1 \equiv F_2$).

Next, we show that every prediction interval $\hat{y}_0 + (\hat{L}, \hat{U}]$ for y_0 satisfying a conditional coverage probability of $1-\alpha$ fulfills $-\infty \leq \hat{L} < 0 < \hat{U} \leq \infty$: We start with the observation that

$$1 - \alpha = F_1(\hat{U}) - F_1(\hat{L}) \leq \min\left(1 - F_1(\hat{L}), F_1(\hat{U})\right),$$

which implies

$$\alpha \geq 1 - \min\left(1 - F_1(\hat{L}), F_1(\hat{U})\right) = \max\left(F_1(\hat{L}), 1 - F_1(\hat{U})\right).$$

Combining the inequality from the preceding display with the assumption $\min(F_1(0), 1 - F_1(0)) > \alpha$, we get

$$\min(F_1(0), 1 - F_1(0)) > \max\left(F_1(\hat{L}), 1 - F_1(\hat{U})\right),$$

which yields

$$F_1(0) > F_1(\hat{L}) \text{ as well as } 1 - F_1(0) > 1 - F_1(\hat{U}),$$

where the latter coincides with $F_1(0) < F_1(\hat{U})$. As F_1 is strictly increasing, this implies

$$\hat{L} < 0 < \hat{U}.$$

Next, we notice that either $\hat{U} < \infty$ or $\hat{L} > -\infty$ as otherwise the prediction interval $(-\infty, \infty]$ has conditional coverage probability of $1 > 1 - \alpha$.

We now deal with the case $0 < \hat{U} < \infty$. As F_2 is strictly increasing and $\lambda > 1$, we have

$$F_1(\hat{U}) = F_2(\lambda\hat{U}) > F_2(\hat{U}). \tag{A.30}$$

Moreover, as $-\infty \leq \hat{L} < 0$, we get

$$F_1(\hat{L}) = F_2(\lambda\hat{L}) \leq F_2(\hat{L}). \tag{A.31}$$

Combining equation (A.30) with equation (A.31) yields

$$\begin{aligned} 1 - \alpha = F_1(\hat{U}) - F_1(\hat{L}) &\geq F_1(\hat{U}) - F_2(\hat{L}) \\ &> F_2(\hat{U}) - F_2(\hat{L}), \end{aligned}$$

which is a contradiction to the exact conditional coverage probability of $1 - \alpha$ under the distribution F_2.

In the second case $-\infty < \hat{L} < 0$ we can use the same arguments as before resulting in

$$\begin{aligned} F_1(\hat{U}) = F_2(\lambda\hat{U}) &\geq F_2(\hat{U}) \text{ and} \\ F_1(\hat{L}) = F_2(\lambda\hat{L}) &< F_2(\hat{L}), \end{aligned}$$

which also yields the same contradiction

$$\begin{aligned} 1 - \alpha = F_1(\hat{U}) - F_1(\hat{L}) &\geq F_2(\hat{U}) - F_1(\hat{L}) \\ &> F_2(\hat{U}) - F_2(\hat{L}). \end{aligned}$$

□

Proof of Lemma 6.8: As the probability of PI^r coinciding with the empty set is α, the actual (conditional) coverage probability cannot exceed $1-\alpha$. Thus, it suffices to show that the following holds true:

$$\mathbb{P}\left(y_0 \in PI^r \| T_n\right) \geq 1 - \alpha - 2\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n) \text{ a.s.}$$

Recalling equation (3.4) of Lemma 3.2 the following holds true almost surely:

$$\begin{aligned}
\mathbb{P}\left(y_0 \in PI^r \| T_n\right) &= (1-\alpha)\left(\mathrm{F}_n\left(\max_{1\leq i\leq n}(\hat{u}_i) + \varepsilon\right) - \lim_{\delta \searrow 0} \mathrm{F}_n\left(\min_{1\leq i\leq n}(\hat{u}_i) - \varepsilon - \delta\right)\right) \\
&\geq (1-\alpha)\left(\widehat{\mathrm{F}}_n\left(\max_{1\leq i\leq n}(\hat{u}_i)\right) - \lim_{\delta \searrow 0} \widehat{\mathrm{F}}_n\left(\min_{1\leq i\leq n}(\hat{u}_i) - \delta\right) - 2\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n)\right) \\
&= (1-\alpha)(1 - 2\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n)) \geq 1 - \alpha - 2\ell_\varepsilon(\widehat{\mathrm{F}}_n, \mathrm{F}_n)),
\end{aligned}$$

which finishes the proof. □

Abstract

The aim of the present work is to construct prediction intervals via a Jackknife-approach whose coverage probability conditional on the training data is close to its nominal level in finite samples and can be asymptotically valid in high-dimensions. The main innovation is to generalize the results of Steinberger and Leeb (2023) to a non-continuous response distribution and to the case of non-linear models.

More specifically, this work is split into four parts: in the first part we link the prediction interval's coverage probability to the accuracy of estimating the distribution of the prediction error in different metrics. While in the case of a continuous distribution the Kolmogorov distance is a suitable choice, we introduce the ε-variational divergence to deal with the non-continuous case and discuss advantages to the Kolmogorov distance, the $\mathcal{L}_p$-norm and the Lévy metric. Moreover, the usability (i.e. the informativeness) of the ε-variational divergence extends to the estimation of other functions of the prediction error, such as the mean-squared prediction error or the mean-absolute prediction error.

In the second part of the work, we define an approach based on the Jackknife for the estimation of the prediction error's distribution conditional on the training data. Thirdly, we present upper bounds for the distance between the conditional prediction error's distribution and its estimate measured in terms of different measurements of distance. We state our results both in finite sample and asymptotically. Our results include both the low-dimensional and the high-dimensional case. Moreover, we show that the prediction error's distribution can be estimated consistently if two conditions are fulfilled: the prediction error should be bounded in probability and the prediction algorithm should satisfy a stability condition. In the last part we show that under mild assumptions these two properties are fulfilled for the OLS estimator and the James-Stein estimator in a low-dimensional setting, for the minimum-norm interpolator in high-dimensions and for the ridge regression regardless of the number of regressors. Furthermore, we also present an example in the case of binary classification where the corresponding predictor fulfills these properties.

Zusammenfassung

Das Ziel der vorliegenden Arbeit ist die Konstruktion von Prognoseintervallen mithilfe eines Jackknife Ansatzes, deren tatsächliche Überdeckungswahrscheinlichkeit bedingt auf die Trainingsdaten in endlicher Stichprobe nahe an dem nominalen Wert liegt und asymptotisch valide sein kann im hochdimensionalen Fall. Die Hauptinnovation besteht in der Verallgemeinerung der Resultate von Steinberger und Leeb (2023) auf unstetige Verteilungen der abhängigen Variable und den Fall von nicht-linearen Modellen.

Genauer gesagt teilt sich diese Arbeit in vier Teile auf: Im ersten Teil stellen wir einen Zusammenhang zwischen der Überdeckungswahrscheinlichkeit eines Prognoseintervalls und der in verschiedenen Metriken gemessenen Genauigkeit der Schätzung der Verteilungsfunktion des Prognosefehlers auf. Während im Falle einer stetigen Verteilung die Kolmogorov Distanz eine geeignete Wahl ist, führen wir die *ε-variational divergence* ein um den nicht-stetigen Fall zu behandeln und diskutieren Vorteile gegenüber der Kolmogorov Distanz, der $\mathcal{L}_p$-norm und der Lévy Metrik. Des Weiteren erstreckt sich die Verwendung der *ε-variational divergence* auch auf die Schätzung von anderen Funktionen des Prognosefehlers wie beispielsweise den mittleren quadratischen Prognosefehler oder den mittleren absoluten Prognosefehler.

Im zweiten Teil der Arbeit definieren wir einen Jackknife Ansatz zur Schätzung der Verteilung des Prognosefehlers bedingt auf die Trainingsdaten. Drittens präsentieren wir obere Schranken für die auf verschiedene Arten gemessene Distanz zwischen der bedingten Verteilung des Prognosefehlers und deren Schätzung. Unsere Resultate werden sowohl in endlicher Stichprobe als auch asymptotisch angegeben und umfassen sowohl den niedrig-dimensionalen als auch den hoch-dimensionalen Fall. Des Weiteren zeigen wir, dass die Verteilung des Prognosefehlers konsistent geschätzt werden kann, wenn die folgenden zwei Bedingungen erfüllt sind: Der Prognosefehler sollte beschränkt in Wahrscheinlichkeit sein und der Prognosealgorithmus sollte eine Stabilitätsbedingung erfüllen. Im letzten Teil zeigen wir, dass unter schwachen Annahmen diese zwei Eigenschaften für den Kleinste-Quadrate-Schätzer und den James-Stein Schätzer im niedrig-dimensionalen Setting, für den Minimum-norm Interpolator im hoch-dimensionalen Fall und für den Ridge Schätzer unabhängig von der Anzahl der Regressoren erfüllt sind. Außerdem präsentieren wir ein Beispiel im Fall von binärer Klassifizierung, in dem der dazugehörige Prädiktor ebenfalls diese Eigenschaften erfüllt.