CONDITIONAL PREDICTIVE INFERENCE FOR STABLE ALGORITHMS

By Lukas Steinberger* and Hannes Leeb†

University of Vienna

We investigate generically applicable and intuitively appealing prediction intervals based on \( k \)-fold cross validation. We focus on the conditional coverage probability of the proposed intervals, given the observations in the training sample (hence, training conditional validity), and show that it is close to the nominal level, in an appropriate sense, provided that the underlying algorithm used for computing point predictions is sufficiently stable when feature-response pairs are omitted. Our results are based on a finite sample analysis of the empirical distribution function of \( k \)-fold cross validation residuals and hold in non-parametric settings with only minimal assumptions on the error distribution. To illustrate our results, we also apply them to high-dimensional linear predictors, where we obtain uniform asymptotic training conditional validity as both sample size and dimension tend to infinity at the same rate and consistent parameter estimation typically fails. These results show that despite the serious problems of resampling procedures for inference on the unknown parameters (cf. Bickel and Freedman, 1983; El Karoui and Purdom, 2018; Mammen, 1996), cross validation methods can be successfully applied to obtain reliable predictive inference even in high dimensions and conditionally on the training data.

1. Introduction. It is the fundamental task of (supervised) statistical learning, when given an i.i.d. training sample of feature-response pairs \((x_i, y_i)\) and an additional feature vector \(x_0\), to provide a point prediction for the corresponding unobserved response variable \(y_0\). In such a situation, a prediction interval that contains the unobserved response variable with a

---

* Lukas Steinberger was supported by the Austrian Science Fund (FWF): P 28233-N32 and I 5484-N, the latter project is part of the Research Unit 5381 of the German Research Foundation. Part of this research was conducted while Lukas Steinberger was funded by the German Research Foundation (DFG): RO 3766/401.

† Hannes Leeb was supported by the Austrian Science Fund (FWF): P 28233-N32 and P 26354-N26.

MSC 2010 subject classifications: Primary 62G15, 62J02; secondary 62G20, 62J07

Keywords and phrases: prediction intervals, high-dimensional regression, algorithmic stability, cross-validation
prescribed probability provides valuable additional information to the practitioner. In many applications a training sample is obtained only once and is subsequently used to repeatedly construct point and interval predictions as new measurements of feature vectors become available. In such a situation, it is desirable to control the conditional coverage probability of the prediction interval given the observations in the training sample, rather than the unconditional probability (the latter is controlled, for example, by procedures discussed in Barber et al., 2021b; Vovk et al., 2005). We call this notion of validity of a prediction interval, where the conditional coverage probability given the training data is (approximately) controlled, training conditional validity (see also Vovk, 2013). Notice how this is markedly different from the notion of object conditional validity where the conditioning is on the new feature vector $x_0$ (cf. Barber et al., 2021a).

We study a very simple method based on cross validation (CV) residuals, which is generic, in the sense that it can be constructed on top of any algorithm that produces a point predictor, and that yields asymptotic training conditionally valid prediction intervals provided the underlying point prediction algorithm is sufficiently stable. To illustrate the idea and keep notation simple, we begin with the case of $n$-fold cross validation, that is, leave-one-out. For an i.i.d. training sample $T_n = (x_i, y_i)_{i=1}^n$ of size $n$, consisting of $\mathbb{R}^p \times \mathbb{R}$-valued feature-response pairs, and an additional feature vector $x_0$ in $\mathbb{R}^p$, suppose that we have decided to use a prediction algorithm $M_{n,p} : (\mathbb{R}^p \times \mathbb{R})^n \times \mathbb{R}^p \to \mathbb{R}$ to produce a point prediction $\hat{y}_0 = M_n(T_n, x_0)$ for the real unobserved response $y_0$. If $T_n[i] = (x_j, y_j)_{j \neq i}$ is the sample without the $i$-th observation pair, compute leave-one-out residuals $\hat{u}_i = y_i - M_{n-1,p}(T_n[i], x_i)$, $1 \leq i \leq n$. Finally, to obtain a prediction interval for $y_0$, compute appropriate empirical quantiles $\hat{q}_{\alpha_1}$ and $\hat{q}_{\alpha_2}$ from the collection $\hat{u}_1, \ldots, \hat{u}_n$ and report the leave-one-out prediction interval

$$PI_{(LO)}^{(LO)}(T_n, x_0) = [\hat{y}_0 + \hat{q}_{\alpha_1}, \hat{y}_0 + \hat{q}_{\alpha_2}].$$

In this paper we investigate the following conditional coverage probability given the training data, where the randomness comes only from the feature-response pair $(x_0, y_0)$ in the prediction period,

$$\mathbb{P}(y_0 \in PI_{(LO)}^{(LO)}(T_n, x_0) | T_n).$$

We investigate this probability in finite samples and in more specific asymptotic settings, including those where the dimension $p$ of the feature vectors $x_i$ increases at the same rate as sample size $n$. We find that even in these
challenging scenarios where both \( n \) and \( p \) are large, the conditional coverage probability of \( \text{Pl}_{\alpha_1, \alpha_2}(T_n, x_0) \) is typically close to the nominal level \( \alpha_2 - \alpha_1 \), that is, \( \text{Pl}_{\alpha_1, \alpha_2}(T_n, x_0) \) guarantees (approximate) training conditional validity, provided some regularity conditions on the data generating model and the predictor are satisfied. Note that the analogous procedure based on ordinary residuals \( y_i - M_{n,p}(T_n, x_i) \) instead of leave-one-out residuals would, in general, not be valid in such a large-\( p \) scenario (cf. Bickel and Freedman, 1983). Extending these results to the general \( k \)-fold cross-validation case relies on a generalization of a result by Bousquet and Elisseeff (2002) on the estimation of the test error of a learning algorithm by its empirical leave-one-out error, which might be of independent interest (see Lemma C.1 in the supplement).

Despite the remarkable simplicity of this method, and its apparent similarity to the Jackknife, we are not aware of any rigorous analysis of its statistical properties. Notable exceptions are Steinberger and Leeb (2016, 2021), which are precursors of this paper, and Barber et al. (2021b); the latter will be discussed in Subsection 1.1.3. Our approach is similar, in spirit, to the methods proposed in Butler and Rothman (1980), Stine (1985), Schmoyer (1992), Olive (2007) and Politis (2013), in the sense that it relies on resampling and leave-one-out ideas for predictive inference. But the methods from these references, like most resampling procedures in the literature, are investigated only in the classical large sample asymptotic regime where the number of available explanatory variables is fixed. Prominent exceptions are Bickel and Freedman (1983), Mammen (1996) and, more recently, El Karoui and Purdom (2018). These articles draw mainly negative conclusions about resampling methods in high dimensions, arguing, for instance, that the famous residual bootstrap in linear regression, which relies on the consistent estimation of the true unknown error distribution, is unreliable when the number of variables in the model is not small compared to sample size. In contrast, we show that under mild conditions the leave-one-out prediction interval \( \text{Pl}_{\alpha_1, \alpha_2}(T_n, x_0) \) does not suffer from these problems because it relies on estimation of the conditional distribution of the prediction error \( P(y_0 - \hat{y}_0 \leq t || T_n) \) instead of an estimator for the unconditional distribution of the error term \( y_0 - E[y_0 || x_0] \). That the use of leave-one-out residuals leads to more reliable methods in high dimensions was also observed by El Karoui and Purdom (2018).

Our contribution is threefold. First, we show that the leave-one-out prediction interval (and its extension to \( k \)-fold CV) is approximately training
conditionally valid given the training sample $T_n$, that is

$$P\left(y_0 \in PI_{\alpha_1,\alpha_2}(T_n, x_0) \mid T_n\right) \approx \alpha_2 - \alpha_1.$$  

The error term of the above approximation can be controlled in finite samples and asymptotically, provided that the employed prediction algorithm $M_{n,p}$ is sufficiently stable under the omission of feature-response pairs and that it has a bounded (in probability) estimation error as an estimator for the true unknown regression function. It is of paramount importance, however, to point out that we do not need to assume consistent estimability of the regression function and our leading examples are such that consistency fails.

Second, we show that the required stability and approximation properties are satisfied in many cases, including many linear predictors in high dimensional regression problems and even if the true model is not exactly linear. In particular, the proposed method is always valid if the employed predictor is consistent for the unknown regression function (or for an appropriate surrogate target), and is therefore applicable to complex data structures and methods such as non-parametric regression, LASSO prediction, random forests or deep learning (see Section 4).

Third, we discuss issues of interval length and find that in typical situations predictors with smaller mean squared prediction error lead to shorter prediction intervals. For ordinary least squares prediction, we also investigate the impact of the dimensionality of the regression problem on the interval length and discuss the relationship between the leave-one-out method and an obvious sample splitting technique. All our results hold uniformly over large classes of data generating processes and under weak assumptions on the unknown error distribution (e.g., the errors may be heavy tailed and non-symmetric, and the standardized design vectors $\text{Cov}[x_i]^{-1/2}x_i$ may have dependent components and a non-spherical distribution).

Our work is greatly inspired by El Karoui et al. (2013) and Bean et al. (2013) (see also El Karoui, 2013, 2018), who investigate efficiency of general $M$-estimators in linear regression when the number of regressors $p$ is of the same order of magnitude as sample size $n$. In particular, the $M$-estimators studied in these references provide one leading example of a class of linear predictors for which our construction of prediction intervals leads to training conditionally valid predictive inference even in high dimensions.

The remainder of the paper is organized as follows. In the following Subsection 1.1 we give a brief overview of alternative methods from the large
body of literature on predictive inference in regression. Subsection 1.2 introduces the notation that is used throughout the paper. Sections 2 and 3 proceed along a general-to-specific scheme. We begin, in Subsection 2.1, by introducing the general cross validation method and the notion of training conditional validity and we relate this notion to estimation of the conditional prediction error distribution in Kolmogorov distance. In Subsection 2.2, we draw the connection between training conditional validity and algorithmic stability and present our main results which provide sufficient conditions for training conditional validity. In Section 3 we then show that these conditions can be verified in challenging statistical scenarios where regression function estimators and the bootstrap usually fail to be consistent. In particular, we consider linear predictors based on regularized $M$-estimators and based on James-Stein-type estimators in a situation where the number of regressors $p$ is not small relative to sample size $n$. We also take a closer look at the ordinary least squares estimator, because its simplicity allows for a rigorous analysis of the resulting interval length. In Section 4, we then also discuss the important case where the employed predictor is consistent (possibly for some pseudo target rather than the true regression function) and we provide examples on non-parametric regression, high-dimensional LASSO, random forests and deep neural networks. The case of consistency is an important test case for our method. Finally, we present the results of an extensive simulation study in Section 5. Further discussions and possible extensions of our results as well as most of the technical proofs are deferred to the supplementary material Steinberger and Leeb (2022).

1.1. Related work. In a fully parametric setting, predictive inference is essentially a special case of parametric inference (see, e.g., Cox and Hinkley, 1974, Subsection 7.5). Constructing valid prediction sets becomes much more challenging, however, if one is interested in a non-parametric setting. By non-parametric, we do not only mean that the regression function can not be indexed by a finite dimensional Euclidean space, but also that the random fluctuations $y_i - E[y_i\|x_i]$ about the conditional mean function can not be described by a parametric family of distributions.

1.1.1. Tolerance regions. A rather well researched and classical topic in the statistics literature is the construction of so called tolerance regions or tolerance limits, which are closely related to prediction regions. A tolerance region is a set valued estimate $TR(T_n) \subseteq \mathbb{R}^m$ based on i.i.d. $m$-variate data $z_1, \ldots, z_n$, $T_n = (z_1, \ldots, z_n)$, such that the probability of covering an independent copy $z_0$ is close to a prescribed confidence level. More pre-
cisely, a \((1 - \alpha, \rho)\) tolerance region \(TR\) is such that \(P(P(z_0 \in TR|T_n) \geq 1 - \alpha) = \rho\), and \(TR\) is called a \((1 - \alpha)\)-expectation tolerance region, if \(E[P(z_0 \in TR|T_n)] = P(z_0 \in TR) = 1 - \alpha\) (cf. Krishnamoorthy and Mathew, 2009). The study of non-parametric tolerance regions goes back at least to Wilks (1941, 1942), Wald (1943) and Tukey (1947) (see Krishnamoorthy and Mathew, 2009, for an overview and further references) and is traditionally based on the theory of order statistics of i.i.d. data. These researchers already obtained multivariate distribution-free methods, that is, tolerance regions that achieve a certain type of validity in finite samples without imposing parametric assumptions. The connection to prediction regions is apparent: If \(z_i = (x_i, y_i)\), then a tolerance region \(TR(T_n)\) for \(z_0 = (x_0, y_0)\) can be immediately used to obtain a prediction region for \(y_0\) by setting \(PR(T_n, x_0) = \{y : (x_0, y) \in TR(T_n)\}\). However, this is arguably not the most economical way of constructing a prediction region. In fact, the construction of a multivariate and possibly high-dimensional tolerance region appears to be a more ambitious goal than the construction of a prediction region for a univariate response variable. In particular, since estimation of the full density of \(z_0\) – which could be used to compute an optimal highest density region – is usually not feasible if the dimension \(m\) is non-negligible compared to sample size \(n\), one has to specify a shape for the tolerance region \(TR\) and it is not obvious which shapes are preferable in a non-parametric setting. For example, Bucchianico et al. (2001) provide results for smallest possible hyperrectangles and ellipsoids, but obtain only the classical large sample asymptotic results with fixed dimension. Chatterjee and Patra (1980) estimate the density non-parametrically, which fails in high dimensions. Li and Liu (2008) use a notion of data depth to avoid the specification of the shape, but the fully data driven method, again, is only shown to be valid asymptotically, with the dimension fixed. Finally, numerically computing the \(x_0\)-cut of \(TR\) to obtain \(PR\) is computationally demanding and the result is sensitive to the shape of \(TR\).

1.1.2. Conformal prediction. A strand of literature which has emerged from the early ideas of non-parametric tolerance regions, but which is more prominent within the machine learning community than the statistics community, is called conformal prediction (Vovk et al., 1999, 2005, 2009). Conformal prediction is a very flexible general framework for construction of prediction regions that can be used in conjunction with any learning algorithm. The general idea is to construct a pivotal \(p\)-value \(\pi(y_*)\) to test \(H_* : y_0 = y_*\) based on the sample \(T_n\) and \(x_0\), for each possible value \(y_*\) of \(y_0\), and to invert the test to obtain a prediction region for \(y_0\), i.e., \(PR = \{y : \pi(y) \geq \alpha\}\). The
method was primarily designed for an on-line learning setup (cf. Vovk et al., 2009), but has recently been popularized in the statistics community by Lei et al. (2018, 2013) and Lei and Wasserman (2014), who study it as a batch method. Aside from their flexibility, conformal prediction methods have the advantage that they are valid in finite samples, in the sense that the unconditional coverage probability \( P(y_0 \in \mathcal{P} R) \) is no less than the nominal level \( 1 - \alpha \), provided only that the feature-response pairs \((x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)\) are exchangeable. On the other hand, their practical implementation is not so straightforward, because, for the test inversion, the \( p \)-value \( \pi \) has to be evaluated on a grid of possible \( y \) values, which is especially tricky if the conformal prediction region is not an interval (see Chen et al., 2018; Lei, 2019, for further discussion of these issues). Moreover, it is not clear if the classical conformal methods can also provide training conditional validity. In Vovk (2012), a version of conformal prediction was presented that achieves also a certain type of training conditional validity. More recently, Bates et al. (2021) have constructed conformal predictive \( p \)-values that also achieve a form of training conditional validity. However, these methods rely on a sample splitting idea, which usually makes the prediction region unnecessarily wide (see Subsection 3.4 as well as Subsection A.2 in the supplement for further discussion of sample splitting techniques). A different version of conditional validity (conditioning on \( x_0 \)), is discussed in Barber et al. (2021a) (see also Remark A.3 in the supplement).

1.1.3. The jackknife+. Barber et al. (2021b) recently proposed a modification of the leave-one-out method considered here. For the modified method, which they call jackknife+, they derived a finite-sample lower bound for the unconditional coverage probability, under the assumption that the feature-response pairs are exchangeable and without requiring that the prediction algorithm is stable. For a jackknife+ interval with nominal coverage probability \( 1 - \alpha \), the lower bound is \( 1 - 2\alpha \); if the prediction algorithm is stable (under omission of a single feature-response pair), the lower bound moves closer towards \( 1 - \alpha \) (provided that the interval is slightly modified further). In simulations and data-examples, Barber et al. (2021b) found that the jackknife+ performs essentially like the jackknife, i.e., like the method considered in this paper (see also our simulation section 5), unless the prediction algorithm is highly unstable; in the unstable case, jackknife+ outperforms jackknife. Whether or not a certain algorithm is stable depends on the unknown data generating process and also cannot be tested in a fully assumption-free setting (cf. Kim and Barber, 2021). However, many classical methods are provably stable under common and even high-dimensional
modeling assumptions. The training conditional performance of the jackknife+ interval, that is, its coverage probability conditional on the training data, is yet to be analyzed.

1.2. Preliminaries and notation. For \( p \in \mathbb{N} \), let \( \mathcal{Y} \subseteq \mathbb{R} \) and \( \mathcal{X} \subseteq \mathbb{R}^p \) be Borel measurable sets and let \( \mathcal{Z} = \mathcal{X} \times \mathcal{Y} \). For \( n \in \mathbb{N} \), \( n \geq 2 \), a realization of a training sample is denoted by \( T_n := (z_i)_{i=1}^n \in \mathcal{Z}^n \), where \( z_i = (x_i, y_i) \in \mathcal{Z} \). A feature-response pair. The realizations of feature-response pairs in the prediction period are denoted by \( z_0 = (x_0, y_0) \in \mathcal{Z} \). All the randomness we consider comes from drawing training samples \( T_n \) and \( z_0 \), and we assume that all feature-response pairs are drawn independently and identically distributed from some probability distribution \( P \) belonging to a class \( \mathcal{P} \) of Borel probability measures on \( \mathcal{Z} \). By \( \mathbb{P} = \mathbb{P}(P) \) we denote the joint product measure of all occurring feature-response pairs with marginal distribution \( P \in \mathcal{P} \) and we write \( \mathbb{E} \) for the corresponding expectation operator. Thus, strictly speaking, \( \mathbb{P} \) and \( \mathbb{E} \) can change their meaning from line to line if the number of feature-response pairs changes. We use expressions like \( \mathbb{P}_{T_n}, \mathbb{P}_{z_1}, \mathbb{P}_{x_0}, \mathbb{E}_{z_0}, \) etc., to denote integration only with respect to the indicated variables (or marginal distributions). By \( P_{y_0|x_0} \) we denote a regular conditional distribution of \( y_0 \) given \( x_0 \) under \( P \). By \( \mu_P(x) := \mathbb{E}_{P|y_0=x} := \int_{\mathcal{Y}} y P_{y_0|x_0}(dy|x) \), \( \mu_P : \mathcal{X} \to \mathbb{R} \), we denote (a version of) the true unknown regression function, if it exists. We sometimes express the training data \( T_n \) as \( (X, Y) \), where \( X = [x_1, \ldots, x_n]' \in \mathbb{R}^{n \times p} \) and \( Y = (y_1, \ldots, y_n)' \in \mathbb{R}^n \). Moreover, \( X' \) denotes the transpose of \( X \), and we write \( (X'X)' \) for the Moore-Penrose inverse of \( X'X \). Similarly, we write \( X_{[i]} = [x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n]' \) and \( Y_{[i]} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)' \) for the corresponding leave-one-out quantities.

Next, we formally define the notion of a (learning) algorithm, of a predictor (or estimator) \( \hat{\mu}_n \), and of the cross validation residuals. For \( n, p \in \mathbb{N} \), write \( \lfloor n \rfloor = \{1, \ldots, n\} \) and consider a collection of measurable functions \( (M_{n', p})_{n' \in \lfloor n \rfloor}, M_{n', p} : \mathcal{Z}^{n'} \times \mathcal{X} \to \mathbb{R} \). The collection \( (M_{n', p})_{n' \in \lfloor n \rfloor} \) is called a learning algorithm. For \( k \in \lfloor n \rfloor \), let \( K_1, \ldots, K_k \subseteq \lfloor n \rfloor \) be a partition of \( \lfloor n \rfloor \) and abbreviate \( m_l := |K_l|, l \in [k] \). For a vector \( x \in \mathcal{X} \), we set \( \hat{\mu}_n(x) := M_{n, p}(T_n, x) \) and \( \mu_n^{(l)}(x) := M_{n-m_l, p}(T_n \setminus K_l, x) \), where \( T_n \setminus K_l := (z_j)_{j \notin K_l} \) denotes the reduced training sample where the observations with indices in \( K_l \) have been excluded. We call \( \hat{\mu}_n : \mathcal{X} \to \mathbb{R} \) a predictor or estimator and say that it, or the underlying learning algorithm, is symmetric if for every \( n' \in \lfloor n \rfloor \), for every choice of \( z_1, \ldots, z_{n'} \in \mathcal{Z} \), every \( x \in \mathcal{X} \) and every per-
mutation \( \pi \) of \( n' \) elements, \( M_{n', \pi}(z_1^{n'}, x) = M_{n', \pi}(z_{\pi(i)}^{n'}, x) \).\(^1\) Finally, we define \( k \)-fold cross validation (\( k \)-CV) residuals for \( l \in [k] \) and \( i \in K_l \), by 
\[
\hat{u}_i := y_i - \hat{\mu}_n(x_i).
\]

If \( f : D \to \mathbb{R} \) is a real valued function on some domain \( D \), then \( \|f\|_{\infty} = \sup_{s \in D} |f(s)| \). For \( a, b \in \mathbb{R} \), we also write \( a \vee b = \max(a, b) \) and \( a \wedge b = \min(a, b) \) and \( a_+ = a \vee 0 \), and let \( \lceil \delta \rceil \) denote the smallest integer no less than \( \delta \in \mathbb{R} \). If \( F : \mathbb{R} \to [0, 1] \) is a cumulative distribution function, we write \( \alpha \mapsto F^\dagger(\alpha) := \inf\{s \in \mathbb{R} : \alpha \leq F(s)\} \) for the corresponding quantile function. We write \( U \overset{\mathcal{L}}{\equiv} V \), if the random quantities \( U \) and \( V \) are equal in distribution and the underlying probability spaces are clear from the context. By a slight abuse of notation, we also write \( U \overset{\mathcal{L}}{\equiv} \mathcal{L}_0 \) if the random variable \( U \) is distributed according to the probability law \( \mathcal{L}_0 \) and, again, the underlying probability space is clear from the context.

For our asymptotic statements, we will also adopt the following array setup: Unless stated otherwise, all quantities introduced above are allowed to also depend on sample size \( n \). We consider sequences \((p_n)_{n \in \mathbb{N}}\) and \((k_n)_{n \in \mathbb{N}}\) of positive integers and a collection of data generating distributions \((P_n)_{n \in \mathbb{N}}\), where \( P_n \) is a probability measure on \( Z_n = X_n \times \mathcal{Y}_n \subseteq \mathbb{R}^{p_n+1} \). We use symbols like \( \mathbb{P}_n \), \( \mathbb{E}_n \), \( x_{0,n} \), \( y_{0,n} \), etc., to emphasize dependence on \( n \). For a collection of measurable functions \( \phi_n : Z_n^{n+1} \to \mathbb{R} \), we say that \( \phi_n \) is \( P_n \)-bounded in probability if \( \limsup_{n \to \infty} \mathbb{P}_n(|\phi_n| > M) \to 0 \), as \( M \to \infty \), and write \( \phi_n = O_{P_n}(1) \). If \( \mathbb{P}_n(|\phi_n| > \varepsilon) \to 0 \), as \( n \to \infty \), for every \( \varepsilon > 0 \), then we say that \( \phi_n \) converges in \( P_n \)-probability to zero and write \( \phi_n = o_{P_n}(1) \). Similarly, we say that \( \phi_n \) converges in \( P_n \)-probability to \( \psi_n : Z_n^{n+1} \to \mathbb{R} \), which is also assumed to be measurable, if \( |\phi_n - \psi_n| = o_{P_n}(1) \). Since the probability measure \( P_n \) will be chosen arbitrarily from some class \( \mathcal{P}_n \) of probability measures on \( Z_n \), our asymptotic results will be uniform over the respective classes.

2. Main results.

2.1. \( k \)-CV prediction intervals and training conditional validity. For \( \alpha \in (0, 1) \), we want to construct a prediction interval \( PI_\alpha(T_n, x_0) = (\hat{\mu}_n(x_0) + L_\alpha(T_n), \hat{\mu}_n(x_0) + U_\alpha(T_n)) \) for \( y_0 \), where \( L_\alpha \) and \( U_\alpha \) are measurable functions

\(^1\)Some of the predictors that appear in this paper happen to be symmetric, but symmetry is actually never imposed as a condition in our general theory. We mention this concept only because it can sometimes simplify things and also appears in the literature.
on $\mathcal{Z}$, such that

\[(2.1) \quad \mathbb{E} \left[ \mathbb{P} \left( y_0 \in PI_\alpha(T_n, x_0) \bigg\lvert T_n \right) - (1 - \alpha) \right] \]

is small uniformly over a class $\mathcal{P}$ of data generating distributions $P$ for the feature-response pairs $(x_i, y_i)$ and $(x_0, y_0)$. We can not expect the expression in (2.1) to be equal to zero for some fixed $n$ and a reasonably large class $\mathcal{P}$ (see Remark A.1 in the supplement). Therefore, we are content with (2.1) being close to zero as $n$, and possibly also $p$, is large. A similar but slightly different notion of training conditional validity is studied by Vovk (2013), and is closely related to the conventional notion of a $(1 - \alpha, \rho)$ tolerance region for $\rho$ close to 1 (cf. Krishnamoorthy and Mathew, 2009).

However, these conventional definitions require only that the conditional coverage probability

\[
\mathbb{P} \left( y_0 \in PI_\alpha(T_n, x_0) \bigg\lvert T_n \right) = \mathbb{P} \left( y_0 \in PI_\alpha(T_n, x_0) \right)
\]

is no less than the prescribed confidence level $1 - \alpha$, with high probability, whereas the requirement that (2.1) is small also excludes overly conservative procedures. Nevertheless, we also refer to our requirement in (2.1) as training conditional validity. Note that if (2.1) is small, then also

\[
|\mathbb{P} (y_0 \in PI_\alpha(T_n, x_0)) - (1 - \alpha)| \leq \mathbb{E} \left[ \mathbb{P} \left( y_0 \in PI_\alpha(T_n, x_0) \bigg\lvert T_n \right) - (1 - \alpha) \right]
\]

will be small. Hence, the prediction interval is then also approximately unconditionally valid, uniformly over $P \in \mathcal{P}$.

If the conditional distribution function $s \mapsto \tilde{F}_n(s) := \mathbb{P}(y_0 - \hat{\mu}_n(x_0) \leq s \bigg\lvert T_n)$ is continuous, then, for $0 \leq \alpha_1 < \alpha_2 \leq 1$ fixed, there is an optimal shortest but infeasible interval

\[(2.2) \quad PI_{\alpha_1, \alpha_2}^{(OPT)} = [\hat{\mu}_n(x_0) + \tilde{q}_{\alpha_1}, \hat{\mu}_n(x_0) + \tilde{q}_{\alpha_2}]\]

in the set of all prediction intervals $PI$ of the form $PI = PI(T_n, x_0) = [\hat{\mu}_n(x_0) + L(T_n), \hat{\mu}_n(x_0) + U(T_n)]$ which also satisfy

\[(2.3) \quad \mathbb{P} \left( y_0 \leq \inf PI \bigg\lvert T_n \right) = \alpha_1, \quad \text{and}\]

\[(2.4) \quad \mathbb{P} \left( y_0 \geq \sup PI \bigg\lvert T_n \right) = 1 - \alpha_2 :\]

Simply choose $\tilde{q}_{\alpha_1}$ to be the largest $\alpha_1$-quantile of $\tilde{F}$ and $\tilde{q}_{\alpha_2}$ to be the smallest $\alpha_2$-quantile of $\tilde{F}_n$. This gives the user the flexibility to choose precisely
what error probability of under- and over-prediction she is willing to accept. This also shows that what we are really interested in is the unknown conditional distribution of the prediction error \( y_0 - \hat{\mu}_n(x_0) \). Thus, for \( P \left( I^{(OPT)}_{\alpha_1, \alpha_2} \right) \), (2.1) is actually equal to zero (for \( \alpha_1 + 1 - \alpha_2 = \alpha \)), at least if \( P \) contains only probability distributions on \( \mathcal{Z} \) for which \( \tilde{F}_n : \mathbb{R} \rightarrow [0,1] \) is almost surely continuous.

We propose the following simple cross validation idea to approximate the optimal infeasible procedure: Consider the weighted empirical distribution function

\[
\hat{F}_n(s) := \hat{F}_n(s; T_n) := \sum_{l=1}^{k} \sum_{i \in K_l} \frac{1}{km_l} \mathbb{1}(\hat{u}_i) \tag{2.5}
\]

of the \( k \)-CV residuals \( \hat{u}_i = y_i - \hat{\mu}^{(l)}_n(x_i) \), \( i \in K_l \), where \( \hat{\mu}^{(l)}_n(x_i) = M_{n-m_l,p}(T_n \setminus K_l, x_i) \) is the prediction of the learning algorithm at \( x_i \) when all observations from the fold \( K_l \) are removed from the training data \( T_n \). For \( \alpha \in [0,1] \), let \( \hat{q}_\alpha := \hat{F}_n^\dagger(\alpha) \) denote the empirical \( \alpha \)-quantile. Then the \( k \)-CV prediction interval is given by

\[
P \left( I^{(kCV)}_{\alpha_1, \alpha_2}(T_n, x_0) \right) = \hat{\mu}_n(x_0) + \left[ \hat{q}_{\alpha_1}, \hat{q}_{\alpha_2} \right]. \tag{2.6}
\]

The idea behind the \( k \)-CV procedure is remarkably simple. To estimate the conditional distribution \( \tilde{F}_n \) of the prediction error \( y_0 - \hat{\mu}_n(x_0) \), i.e.,

\[
\tilde{F}_n(s) := \tilde{F}_n(s; T_n) := \mathbb{P}(y_0 - \hat{\mu}_n(x_0) \leq s | T_n), \quad s \in \mathbb{R}, \tag{2.7}
\]

we simply use the (weighted) empirical distribution \( \hat{F}_n \) of the \( k \)-CV residuals \( \hat{u}_i = y_i - \hat{\mu}^{(l)}_n(x_i) \), \( i \in K_l \). Notice that \( \hat{\mu}_n \) is independent of \( (x_0, y_0) \), and \( \hat{\mu}^{(l)}_n \) is independent of \( (x_i, y_i) \), since \( i \in K_l \), and thus \( \hat{u}_i \) has almost the same distribution as the prediction error, except that \( \hat{\mu}^{(l)}_n \) is calculated from \( m_l = |K_l| \) observations less than \( \hat{\mu}_n \). In many cases this difference turns out to be negligible if \( n \) is large, even if \( p \) is relatively large too, provided that \( |K_l| \) is chosen to be sufficiently small. Note, however, that the \( k \)-CV residuals \( \hat{u}_i \) are not independent, which leads to substantial technical challenges.

The following result shows that, indeed, a sufficient condition for asymptotic training conditional validity (2.1) of the \( k \)-CV prediction interval in (2.6) is consistent estimation of \( \tilde{F}_n \) in Kolmogorov distance.
Proposition 2.1. Fix $0 \leq \alpha_1 < \alpha_2 \leq 1$, a training sample $T_n \in \mathbb{Z}^n$, an arbitrary cumulative distribution function $\hat{F}$ (possibly depending on $T_n$) and write $\text{s}_{\text{max}}(\hat{F}) := \sup_{s \in \mathbb{R}}[\hat{F}(s) - \hat{F}(s^-)]$ for the size of the largest jump of $\hat{F}$.

Define the interval

$$\text{(2.8)} \quad \text{PI}_{\alpha_1, \alpha_2}(T_n, x_0) := \mu_n(x_0) + [\hat{q}_{\alpha_1}, \hat{q}_{\alpha_2}],$$

where $\hat{q}_\alpha := \hat{F}^{-1}(\alpha)$. Let $\tilde{F}_n$ be as in (2.7).

i) We have

$$|P_{(x_0, y_0)}(y_0 \in \text{PI}_{\alpha_1, \alpha_2}(T_n, x_0)) - (\alpha_2 - \alpha_1)| \leq 2\|\tilde{F}_n - \hat{F}\|_\infty + 2\text{s}_{\text{max}}(\hat{F}).$$

ii) If the cdf $s \mapsto \tilde{F}_n(s; T_n)$ is continuous, then the prediction interval defined in (2.8) satisfies

$$|P_{(x_0, y_0)}(y_0 \in \text{PI}_{\alpha_1, \alpha_2}(T_n, x_0)) - (\alpha_2 - \alpha_1)| \leq 4\|\tilde{F}_n - \hat{F}\|_\infty.$$

Remark 2.2. Note that the inequalities of Proposition 2.1 are purely algebraic statements for a fixed training set $T_n$. Also note that the coverage probability $P_{(x_0, y_0)}(y_0 \in \text{PI}_{\alpha_1, \alpha_2}(T_n, x_0))$ is a version of the conditional probability $\mathbb{P}(y_0 \in \text{PI}_{\alpha_1, \alpha_2}(T_n, x_0)|T_n)$.

Proof of Proposition 2.1. For arbitrary $\alpha \in [0, 1]$, we have $\alpha \leq \hat{F} \circ \hat{F}^{-1}(\alpha) \leq \alpha + \text{s}_{\text{max}}(\hat{F})$ and therefore

$$\text{(2.9)} \quad |\hat{F} \circ \hat{F}^{-1}(\alpha_2) - \hat{F} \circ \hat{F}^{-1}(\alpha_1) - (\alpha_2 - \alpha_1)| \leq \text{s}_{\text{max}}(\hat{F}).$$

Now Part i) follows upon noticing that

$$P_{(x_0, y_0)}(y_0 \in \text{PI}_{\alpha_1, \alpha_2}(x_0)) = \tilde{F}_n(\hat{q}_{\alpha_2}) - \tilde{F}_n(\hat{q}_{\alpha_1})
= \tilde{F}_n(\hat{q}_{\alpha_2}) - \hat{F}(\hat{q}_{\alpha_2}) + \hat{F}(\hat{q}_{\alpha_1}) - \tilde{F}_n(\hat{q}_{\alpha_1}) + \hat{F}(\hat{q}_{\alpha_2}) - \hat{F}(\hat{q}_{\alpha_1}) + \hat{F}(\hat{q}_{\alpha_1}) - \hat{F}(\hat{q}_{\alpha_1}),$$

and $|\hat{F}(\hat{q}_{\alpha_1}) - \hat{F}(\hat{q}_{\alpha_1})| \leq \text{s}_{\text{max}}(\hat{F})$.

For Part ii) let $s_* \in \mathbb{R}$ denote a point at which the largest jump of $\hat{F}$ occurs, that is, $\text{s}_{\text{max}}(\hat{F}) = \hat{F}(s_*) - \hat{F}(s_-^*)$. Then either $\tilde{F}_n(s_*) \leq \hat{F}(s_*) - \text{s}_{\text{max}}/2$ or $\tilde{F}_n(s_*) > \hat{F}(s_*) - \text{s}_{\text{max}}/2 = \hat{F}(s_-^*) + \text{s}_{\text{max}}/2$. The former case implies $\tilde{F}_n(s_*) - \hat{F}(s_*) \leq -\text{s}_{\text{max}}/2$ whereas, by continuity of $\tilde{F}_n$, the latter implies $\tilde{F}_n(s_-^*) - \hat{F}(s_-^*) > \text{s}_{\text{max}}/2$. Either way, we have $\|\hat{F} - \tilde{F}_n\|_\infty \geq \text{s}_{\text{max}}/2$. Using (2.9) again, we obtain

$$|\hat{F}(\hat{q}_{\alpha_2}) - \hat{F}(\hat{q}_{\alpha_1}) - (\alpha_2 - \alpha_1)| \leq \text{s}_{\text{max}}(\hat{F}) \leq 2\|\hat{F} - \tilde{F}_n\|_\infty.$$
which concludes the proof, since for continuous \( \hat{F} \), we have
\[
P_{(x_0, y_0)}(y_0 \in P_{I_{\alpha_1, \alpha_2}}(x_0)) = \hat{F}_n(\hat{q}_{\alpha_2}) - \hat{F}_n(\hat{q}_{\alpha_1}) = \hat{F}_n(q_{\alpha_2}) - \hat{F}_n(q_{\alpha_1})
\]
\[
= \hat{F}_n(q_{\alpha_2}) - \hat{F}(q_{\alpha_2}) + \hat{F}(q_{\alpha_1}) - \hat{F}(q_{\alpha_1}) - \hat{F}(q_{\alpha_2}) + \hat{F}(q_{\alpha_1}) = \hat{F}_n(q_{\alpha_2}) - \hat{F}_n(q_{\alpha_1}) - \hat{F}(q_{\alpha_2}) + \hat{F}(q_{\alpha_1}) = \hat{F}_n(q_{\alpha_2}) - \hat{F}_n(q_{\alpha_1}) - \hat{F}_n(q_{\alpha_2}) + \hat{F}_n(q_{\alpha_1})
\]

\[\Box\]

By virtue of Proposition 2.1, most of what follows will be concerned with the analysis of \( \|\hat{F}_n - \hat{F}_n\|_\infty \). We are particularly interested in situations where, for a fixed \( x \in X \), \( \hat{\mu}_n(x) \) does not concentrate around \( \mu_P(x) = \mathbb{E}[y_0 | x_0 = x] \) with high probability but remains random (cf. Remark A.2 in the supplement). In such cases, the unconditional distribution function of the prediction error \( P(y_0 - \hat{\mu}_n(x_0) \leq s) = \mathbb{E}[\hat{F}_n(s)] \), the empirical distribution function of the ordinary residuals \( s \mapsto \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty,s]}(y_i - \hat{\mu}_n(x_i)) \) and the true error distribution function \( P(y_0 - \mu_P(x_0) \leq s) \) need not be close to one another, because \( \hat{\mu}_n \) may not contain enough information about the true regression function \( \mu_P \) (see, for instance, Bean et al., 2013; Bickel and Freedman, 1983, for a linear regression example where \( \mu_P(x) = x'\beta_P \)). Nevertheless, we will see that even in such a challenging scenario, it is often possible to consistently estimate the conditional distribution \( \hat{F}_n \) of \( y_0 - \hat{\mu}_n(x_0) \), given the training sample \( T_n \), by the (weighted) empirical distribution \( \hat{F}_n \) of the \( k \)-CV residuals.

2.2. The role of algorithmic stability. In this section we present general results that relate the uniform estimation error \( \|\hat{F}_n - \hat{F}_n\|_\infty \) to a measure of stability of the underlying predictor \( \hat{\mu}_n \). For our first result, sample size \( n \geq 2 \) and dimension \( p \geq 1 \) are fixed. We only need the following condition on the data generating distribution.

\( (C1) \) The conditional distribution \( P_{y_0 | x_0} \) has a Lebesgue density \( f_{y_0 | x_0} \).

Notice that, although the training data \( (x_i, y_i), i \in [n], \) are assumed to be realizations of i.i.d. random vectors, Condition (C1) allows to model heteroskedasticity, because the conditional variance of the response \( y_i \) given \( x_i \) is allowed to depend on the value of the feature vector \( x_i \). Building on terminology from Bousquet and Elisseeff (2002) (see also Devroye and Wagner, 1979), we use the following probabilistic notion to quantify algorithmic

\[\text{It turns out, however, at least in the linear model } \mu_P(x) = x'\beta_P \text{ and for appropriate estimators of } \beta_P, \text{ that the conditional distribution of the prediction error } \hat{F}_n \text{ does concentrate at its mean, i.e., the unconditional distribution, even if } n \text{ and } p \text{ are of the same order of magnitude (cf. Subsection 3.3 and Lemma C.9 in the proof of Theorem 3.4).} \]
stability, which is adjusted to the general case of \( k \)-fold cross validation and which depends on the true data generating distribution \( P \) on \( \mathcal{Z} \).

**Definition 2.3.** The \( k \)-stability coefficient of the predictor \( \hat{\mu}_n \) is defined as

\[
\eta_{n,k} := \eta_{n,k}(P) := \frac{1}{k} \sum_{l=1}^{k} \mathbb{E}\left[ \left( \| f_{y_0 \mid x_0} \|_\infty \left| \hat{\mu}_n(x_0) - \hat{\mu}_n^{(l)}(x_0) \right| \right) \wedge 1 \right].
\]

The stability coefficient \( \eta_{n,k} \) measures the average change of the prediction of \( \hat{\mu}_n \) as the observations of each of the \( k \)-folds are removed in turn for training. A highly stable algorithm has small stability coefficient. In the extreme case of an algorithm that does not even make use of the training data at all, we have \( \eta_{n,k} = 0 \). Since the feature-response pairs are assumed to be i.i.d. under \( P \), it is easy to see that in case of equal fold sizes \( m_l = \frac{n}{k} \), \( l \in [k] \), a symmetric predictor has stability coefficient \( \eta_{n,k} = \mathbb{E}\left[ \| f_{y_0 \mid x_0} \|_\infty | \hat{\mu}_n(x_0) - \hat{\mu}_n^{(1)}(x_0) | \right] \wedge 1 \). Also note that a predictor with \( \eta_{n,k} = 0 \) can not depend on the training data in a non-trivial way (cf. Lemma D.6 in the supplement).

We are now in the position to state our main result on the estimation of \( \tilde{F}_n(s) = P(y_0 - \hat{\mu}_n(x_0) \leq s \| T_n \) by \( \hat{F}_n(s) = \sum_{l=1}^{k} \sum_{i \in K_l} \frac{1}{k} \mathbf{1}_{(-\infty, s]}(\hat{u}_i) \). To that end, we adopt the triangular array setup described in Subsection 1.2.

**Theorem 2.4.** For \( n \in \mathbb{N} \), let \( p = p_n \) and \( k = k_n \) be positive integers, with \( k_n \to \infty \) as \( n \to \infty \), and let \( P_n \) be a data generating distribution on \( \mathcal{Z}_n = \mathcal{X}_n \times \mathcal{Y}_n \), with \( \mathcal{X}_n \subseteq \mathbb{R}^{p_n} \), \( \mathcal{Y}_n \subseteq \mathbb{R} \), that satisfies (C1). Furthermore, suppose that there exists a scaling constant \( \sigma_n^2 \in (0, \infty) \) and a measurable function \( g_n : \mathcal{X}_n \to \mathbb{R} \), such that the following hold true:

\[
\begin{align*}
\sigma_n \| f_{y_0 \mid x_0, n} \|_\infty &= O_{P_n}(1), \\
\frac{|y_0 - n - g_n(x_0, n)|}{\sigma_n} &= O_{P_n}(1), \\
\frac{|g_n(x_0, n) - \hat{\mu}_n(x_0, n)|}{\sigma_n} &= O_{P_n}(1), \\
\eta_{n,k_n} &= \frac{1}{k_n} \sum_{l=1}^{k_n} \mathbb{E}_n \left[ \left( \| f_{y_0 \mid x_0, n} \|_\infty \left| \hat{\mu}_n(x_0, n) - \hat{\mu}_n^{(l)}(x_0, n) \right| \right) \wedge 1 \right] \xrightarrow{n \to \infty} 0.
\end{align*}
\]

Then, the (weighted) empirical cdf \( \hat{F}_n \) of the \( k_n \)-fold cross validation residuals satisfies

\[
\mathbb{E}_n \left[ \| \hat{F}_n - \tilde{F}_n \|_\infty \right] \xrightarrow{n \to \infty} 0.
\]
Moreover, for $0 \leq \alpha_1 < \alpha_2 \leq 1$, the $k_n$-fold cross validation interval is asymptotically training conditionally valid, i.e.,

$$
\mathbb{E}_n \left[ \mathbb{P}_n \left( y_{0,n} \in P_{(k_n,\alpha_2)}(T_n, x_{0,n}) \mid T_n \right) - (\alpha_2 - \alpha_1) \right] \xrightarrow{n \to \infty} 0.
$$

**Remark 2.5 (Uniform training conditional validity).** For each $n$, let $p_n$ and $k_n$ be as in Thorem 2.4, and let $\mathcal{P}_n$ be a class of probability measures on $\mathbb{Z}^n$. If Theorem 2.4 applies for any sequence $P_n \in \mathcal{P}_n$, $n \geq 1$, then the $k_n$-fold cross validation interval is uniformly asymptotically training conditionally valid, i.e., the left-hand-side of (2.14) converges to zero even when taking the supremum over all $P_n \in \mathcal{P}_n$.

In Theorem 2.4, one should think of the scaling constant $\sigma_n^2$ and the function $g_n$ as the error variance and the regression function in a non-linear regression model $y_0 = g_n(x_0) + u_0$, with additive error $u_0$ that is independent of $x_0$ and has variance $\sigma_n^2$. However, they can be taken as appropriate pseudo parameters if needed. Also notice that Condition (C1) allows for a heteroskedastic error term. Theorem 2.4 is actually a simple consequence of the following finite sample bound.

**Theorem 2.6.** Fix positive finite constants $\varepsilon, c_1, c_2$ and a measurable function $g : \mathcal{X} \to \mathbb{R}$. If Condition (C1) holds, then

$$
\mathbb{E} \left[ \| \hat{F}_n - \tilde{F}_n \|_\infty \right] \leq P(|y_0 - g(x_0)| \geq c_1) + \mathbb{P}(|\hat{\mu}_n(x_0) - g(x_0)| > c_2) + \mathbb{E}_{x_0} \left[ (\varepsilon \| f_{y_0 \mid x_0} \|_\infty) \wedge 1 \right] + \sqrt{\left( \frac{2(c_1 + c_2)}{\varepsilon} + 2 \right) \left[ \frac{1}{4(k - 1)} + 5\eta_{n,k} \right]},
$$

where $\eta_{n,k}$ is the stability coefficient of the predictor $\hat{\mu}_n$.

**Proof of Theorem 2.4.** Set $\nu_n := \frac{1}{4(k_n - 1)} + 5\eta_{n,k_n} = o(1)$ and apply Theorem 2.6 with $c_1 = c_2 = \sigma_n \nu_n^{-1/3}$, $\varepsilon = \sigma_n \nu_n^{-1/3}$. For the second claim, note that under (C1) we have $\mathbb{P}_n(y_0, n - \hat{\mu}_n(x_{0,n}) = c \mid T_n) = \mathbb{E}_{x_0,n} \left[ \mathbb{P}_n(y_0, n - \hat{\mu}_n(x_{0,n}) = c \mid T_n, x_{0,n}) \right] = 0$, for all $c \in \mathbb{R}$ and all $T_n \in \mathbb{Z}_n^n$, which means $s \mapsto \hat{F}_n(s; T_n)$ is continuous. Now apply Proposition 2.1.(ii).

Theorem 2.6 provides an upper bound on the risk of estimating the conditional prediction error distribution $\tilde{F}_n$ by the (weighted) empirical distribution of the cross validation residuals $\hat{F}_n$. The upper bound crucially relies
on the properties of the chosen estimator \( \hat{\mu}_n \) for the (pseudo) regression function \( g \). If the sample size is sufficiently large and if the estimator is sufficiently stable and has a moderate estimation error, then the parameters \( \varepsilon, c_1, c_2 \) can be chosen such that the upper bound is small. This is what we do in Theorem 2.4. It is important to note that Theorem 2.4 and Theorem 2.6 are informative also in case the estimator \( \hat{\mu}_n \) is not consistent for \( g \), as is often the case when \( p/n \not\to 0 \). The bound of Theorem 2.6 also exhibits an interesting trade-off between the \( \eta \)-stability of \( \hat{\mu}_n \) and the magnitude of its estimation error. More stable estimators are allowed to be less accurate whereas less stable estimators need to achieve higher accuracy in order to be as reliable for predictive inference purposes as a more stable algorithm.

Theorem 2.4 and Theorem 2.6 show that the \( k \)-CV prediction interval in (2.6) is approximately uniformly training conditionally valid, i.e., has the property that (2.1) is uniformly small at least for large \( n \), provided that the underlying estimator \( \hat{\mu}_n \) has two essential properties: First, the \( k \)-stability coefficient of the estimator must be (uniformly) small if \( n \) is large, as in (2.13). This is an intuitively appealing assumption since otherwise the \( k \)-CV residuals \( \hat{u}_i = y_i - \hat{\mu}_n^{(i)}(x_i) \) may not be well suited to estimate the distribution of the prediction error \( y_0 - \hat{\mu}_n(x_0) \). Second, the scaled estimation error \( (\mu_P(x_0) - \hat{\mu}_n(x_0))/\sigma_P \) at the new observation \( x_0 \) must be \( P_n \)-bounded (cf. (2.12) with \( g_n = \mu_P \)). This is used to guarantee that the conditional distribution \( \hat{\mu}_n(x_0) \) given the training data \( T_n \) is tight in an appropriate sense (cf. Lemma C.5(ii) in the supplement), so that a pointwise bound on \( |\hat{F}_n(t) - \hat{F}_n(t)| \) can be turned into a uniform bound. In the following sections we demonstrate that these two conditions on the estimator \( \hat{\mu}_n \) are satisfied in several different contexts. From now on, as in Theorem 2.4, we will take on an asymptotic point of view.

3. Linear prediction with many variables. In this section we investigate a scenario in which both consistent parameter estimation as well as bootstrap consistency fail (cf. Bickel and Freedman, 1983; El Karoui and Purdom, 2018), but the \( k \)-CV prediction interval is still asymptotically uniformly training conditionally valid. See Section 4 for a discussion of scenarios where consistent parameter estimation is possible. For simplicity, in this section we consider only the leave-one-out case, that is, \( k_n = n \). For \( \kappa \in [0, 1) \), we fix a sequence of positive integers \( (p_n) \), such that \( p_n/n \to \kappa \) as \( n \to \infty \) and \( n > p_n + 1 \) for all \( n \in \mathbb{N} \). In case \( \kappa > 0 \), this type of ‘large \( p \), large \( n \)’ asymptotics has the advantage that certain finite sample features of the problem are preserved in the limit, while offering a workable simplification.
It turns out that conclusions drawn from this type of asymptotic analyses often provide remarkably accurate descriptions of finite sample phenomena.

We work with linear predictors of the form \(\hat{\mu}_n(x_0) = x_0'\hat{\beta}_n\) for various estimators \(\hat{\beta}_n\). Suppose the second moment matrix \(\Sigma_P = \mathbb{E}_{x_0} [x_0 x_0']\) exists. Then the conditions (2.13) and (2.12) can be verified as follows: For \(\varepsilon > 0,

\[
\mathbb{E} \left[ \left( \|f_{y_0|x_0}\|=\infty |\hat{\mu}_n(x_0) - \hat{\mu}_n^{(1)}(x_0) \right) \right] \leq \mathbb{P} \left( \frac{|x_0'\hat{\beta}_n - x_0'\hat{\beta}_{n,[1]}|}{\sigma_n} > \varepsilon \right) + \mathbb{E}_{x_0} \left[ (\varepsilon \|\sigma_n f_{y_0|x_0}\|=\infty ) \right. \leq \mathbb{E} \left[ \left( \frac{1}{\varepsilon^2} \left\|\Sigma_P^{1/2} \right( \hat{\beta}_n - \hat{\beta}_{n,[1]} \right) \|_2^2 / \sigma_n^2 \right) \right] + \mathbb{E}_{x_0} \left[ (\varepsilon \|\sigma_n f_{y_0|x_0}\|=\infty ) \right]
\]

where, for the second inequality, we have used the conditional Markov inequality along with independence of \(x_0\) and \(T_n\) under \(P\), and \(\hat{\beta}_{n,[1]}\) denotes the estimator based on the reduced training sample \((X_{[1]}, Y_{[1]})\). Thus (2.13) follows if

\[
(3.1) \quad \|\sigma_n f_{y_0,n|x_0,n}\|=\infty = O_{P_n}(1) \quad \text{and}

(3.2) \quad \left\|\Sigma_P^{1/2} \left( \hat{\beta}_n - \hat{\beta}_{n,[1]} \right) \right\|_2 / \sigma_n^2 = o_{P_n}(1).
\]

Notice that (3.1) coincides with (2.10). By a similar argument, we find that (2.12) with \(g_n(x_0) = x_0'\beta_n\) follows if

\[
(3.3) \quad \left\|\Sigma_P^{1/2} \left( \hat{\beta}_n - \beta_n \right) \right\|_2 / \sigma_n = O_{P_n}(1),
\]

for some vectors \(\beta_n \in \mathbb{R}^p\), and (2.11) can be expressed as

\[
(3.4) \quad |y_{0,n} - x_{0,n}'\beta_n| / \sigma_n = O_{P_n}(1).
\]

3.1. Regularized M-estimators. An important class of linear predictors for which our theory on the leave-one-out prediction interval applies are those based on regularized M-estimators investigated by El Karoui (2018) in the challenging scenario where \(p/n\) is not close to zero (see also Bean et al., 2013; El Karoui, 2013; El Karoui et al., 2013). For a given convex loss function \(\rho : \mathbb{R} \rightarrow \mathbb{R}\) and a fixed tuning parameter \(\gamma \in (0, \infty)\) (both not depending on \(n\)), consider the estimator

\[
\hat{\beta}_n^{(\rho)} := \arg\min_{b \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \rho(y_i - x_i'b) + \frac{\gamma}{2} \|b\|_2^2.
\]
In a remarkable tour de force, El Karoui (2018) studied the estimation error \(\|\hat{\beta}_n^{(\rho)} - \beta\|_2\) as \(p/n \to \kappa \in (0, \infty)\), in a linear model \(y_i = x_i'\beta + u_i\), allowing for heavy tailed errors (including the Cauchy distribution) and non-spherical design (see Subsection 2.1 in El Karoui, 2018, for details on the technical assumptions). In particular, the author shows that \(\|\hat{\beta}_n^{(\rho)} - \beta\|_2\) converges in probability to a deterministic positive and finite quantity \(r_\rho(\kappa)\) and characterizes the limit through a system of non-linear equations. On the way to this result, El Karoui (2018, Theorem 3.9 together with Lemma 3.5 and the ensuing discussion) also establishes the stability property \(\|\hat{\beta}_n^{(\rho)} - \hat{\beta}_n^{(\rho)}\|_2 \to 0\) in probability. Thus, under the assumptions maintained in that reference, (3.2), (3.3) and (3.4) hold, and the leave-one-out prediction interval based on the linear predictor \(\hat{\mu}_n(x_0) = x_0'\hat{\beta}_n^{(JS)}\) is asymptotically training conditionally valid, provided that also the boundedness condition (3.1) is satisfied. Finally, we note that a detailed assessment of the predictive performance of \(\hat{\beta}_n^{(\rho)}\) in dependence on \(\rho\) requires a highly non-trivial analysis of \(r_\rho(\kappa)\). For the asymptotic validity of the leave-one-out prediction interval, however, all the information needed on \(r_\rho(\kappa)\) is that it is finite.

3.2. James-Stein type estimators. Another important example is the class of linear predictors \(\hat{\mu}_n(x_0) = x_0'\hat{\beta}_n^{(JS)}\) based on James-Stein type estimators \(\hat{\beta}_n^{(JS)}\) defined below. Here, we can allow for the class of data generating processes defined in Condition (C2) below. The feature-response pairs are realizations of a non-Gaussian random design non-linear homoskedastic regression model with regression function \(\mu_P\) and error variance \(\sigma_P\). Moreover, the feature vectors \(x_i\) are allowed to have a complex geometric structure, in the sense that the standardized design vector \(\Sigma_2^{-1/2}P x_1\) is not necessarily concentrated on a sphere of radius \(\sqrt{p/n}\), as would be the case if \(L_t\) in Condition (C2) was supported on \{\(-1, 1\) (see, e.g., Subsection 3.2 in El Karoui, 2010 and Subsection 2.3.1 in El Karoui, 2018 for further discussion of this point).

**(C2)** Fix finite constants \(C_0 > 0\) and \(c_0 > 0\) and probability measures \(L_t\) and \(L_w\) on \((\mathbb{R}, \mathcal{B}(\mathbb{R}))\), such that \(L_w\) has mean zero, unit variance and finite fourth moment, \(\int s^2 L_t(ds) = 1\) and \(L_t((-c_0, c_0)) = 0\). For every \(n \in \mathbb{N}\), the class \(\mathcal{P}_n = \mathcal{P}_n(L_t, L_w, C_0)\) consists of all probability measures \(P\) on \(\mathcal{Z} \subseteq \mathbb{R}^{p_n+1}\), such that the following hold:

(a) The \(x_0\)-marginal distribution of \(P\) is given by \(P_{x_0} \equiv l_0 \Sigma_2^{-1/2}(w_1, \ldots, w_{p_n})'\).
where \( w_1, \ldots, w_{p_n} \) are i.i.d. according to \( L_w \), \( l_0 \overset{E}{=} L_l \) is independent of the \( w_j \) and \( \Sigma_{n}^{1/2} \) is the unique symmetric positive definite square root of a positive definite \( p_n \times p_n \) covariance matrix \( \Sigma_n \).

(b) The conditional distribution of the response given the regressors is
\[
P_{y_0|x_0} \overset{E}{=} \mu_P(x_0) + \sigma_P v_P,
\]
where \( v_P \) is independent of \( x_0 \) and has mean zero, unit variance and fourth moment bounded by \( C_0 \), where \( \mu_P : \mathbb{R}^{p_n} \to \mathbb{R} \) is some measurable regression function with \( \mathbb{E}_{x_0}[\mu_P(x_0)] = 0 \) and \( \sigma_P \in (0, \infty) \).

The model \( P_n \) in (C2) is non-parametric, because the regression function \( \mu_P \) is unrestricted, up to being centered, and the independent error distribution is arbitrary, up to the requirements \( \mathbb{E}_P[v_P] = 0, \mathbb{E}_P[v_P^2] = 1 \) and \( \mathbb{E}_P[v_P^4] \leq C_0 \).

To predict the value of \( y_0 \) from \( x_0 \) and a training sample \( T_n = (x_i, y_i)_{i=1}^n \) with \( n \geq p_n + 2 \), generated from \( P \), we consider linear predictors \( \hat{\beta}_n(x_0) = x_0' \hat{\beta}_n(c) \), where \( \hat{\beta}_n(c) \) is a James-Stein-type estimator given by
\[
\hat{\beta}_n(c) = \begin{cases} 
\left(1 - \frac{cp_n \sigma^2_n}{\hat{\beta}_n' X' \hat{\beta}_n}\right) + \hat{\beta}_n, & \text{if } \hat{\beta}_n' X' \hat{\beta}_n > 0, \\
0, & \text{if } \hat{\beta}_n' X' \hat{\beta}_n = 0,
\end{cases}
\]
for a tuning parameter \( c \in [0, 1] \). Here \( \hat{\beta}_n = (X'X)^{-1}X'Y \), \( \hat{\sigma}^2_n = ||Y - X\hat{\beta}_n||^2_2/(n - p_n) \). The corresponding leave-one-out estimator \( \hat{\beta}^*_n(c) \) is defined equivalently, but with \( X \) and \( Y \) replaced by \( X_{[i]} \) and \( Y_{[i]} \). Note that the leave-one-out equivalent of \( \hat{\sigma}^2_n = \hat{\sigma}^2_{n-1}(X, Y) \) is given by
\[
\hat{\sigma}^2_{n,[i]}(X_{[i]}, Y_{[i]}) = \hat{\sigma}^2_{n-1}(X_{[i]}, Y_{[i]}) = ||Y_{[i]} - X_{[i]} \hat{\beta}^{[i]}_n||^2_2/(n - 1 - p_n).
\]

The ordinary least squares estimator \( \hat{\beta}_n \) belongs to the class of James-Stein estimators. In particular, \( \hat{\beta}_n(0) = \hat{\beta}_n \), because, with \( P_X := X(X'X)^{-1}X' \), we have \( ||P_X Y||^2_2 = \hat{\beta}_n' X' X \hat{\beta}_n = 0 \) if, and only if, \( Y \in \text{span}(P_X)^\perp = \text{span}(X)^\perp \), and the latter clearly implies \( \hat{\beta}_n = 0 \).

Using James-Stein type estimators for prediction is motivated, e.g., by the optimality results of Dicker (2013) and the discussion in Huber and Leeb (2013). The next result shows that in the model (C2) with \( p_n/n \to \kappa \in (0, 1) \) and if the deviation from a linear model is not too severe, the James-Stein-type estimators are sufficiently stable and their estimation errors are uniformly bounded in probability, just as required in (3.2) and (3.3).
Theorem 3.1. For every \( n \in \mathbb{N} \), let \( P_n \in \mathcal{P}_n = \mathcal{P}_n(\mathcal{L}_l, \mathcal{L}_w, C_0) \) as in Condition (C2) and suppose that under \( P_n \), the error term \( v_{P_n} \) in (C2) has a Lebesgue density. Define \( \beta_n = \beta_{P_n} \) to be the minimizer of \( \beta \rightarrow \mathbb{E}_{(x_0, y_0)}[(y_0 - \beta'x_0)^2] \) over \( \mathbb{R}^p \). If \( p_n/n \to \kappa \in [0, 1) \), \( c_n \in [0, 1] \) for all \( n \in \mathbb{N} \), and

\[
\limsup_{n \to \infty} \mathbb{E}_{x_{0,n}} \left[ \left( \frac{\mu_{P_n}(x_{0,n}) - x'_{0,n}\beta_n}{\sigma_{P_n}} \right)^2 \right] < \infty,
\]

then the positive part James-Stein estimator \( \hat{\beta}_n(c_n) \) satisfies (3.3), i.e.,

\[
\limsup_{n \to \infty} \mathbb{P}_n \left( \left\| \Sigma_{P_n}^{1/2}(\hat{\beta}_n(c_n) - \beta_n) / \sigma_{P_n} \right\|_2 > M \right) \xrightarrow{M \to \infty} 0.
\]

If, in addition, \( \kappa > 0 \), then (3.2) is also satisfied, i.e., for every \( \varepsilon > 0 \),

\[
\mathbb{P}_n \left( \left\| \Sigma_{P_n}^{1/2}(\hat{\beta}_n(c_n) - \hat{\beta}_n^{[1]}(c_n)) / \sigma_{P_n} \right\|_2 > \varepsilon \right) \xrightarrow{n \to \infty} 0.
\]

Remark 3.2. Under the assumptions of Theorem 3.1, uniform asymptotic training conditional validity of the leave-one-out prediction interval follows, provided that, in addition, the errors \( v_P \) in Condition (C2) have uniformly bounded densities so that (3.1) holds. To see this, note that (3.2) and (3.3) are conclusions of the theorem, that uniformly bounded fourth moment of the error implies \( P_n \)-uniform boundedness, such that (3.4) is a consequence of assumption (3.6).

Remark 3.3. The last statement of Theorem 3.1 can also be established for the case \( \kappa = 0 \) but would require a different proof strategy. Since this case is statistically less interesting we omit it for the sake of brevity.

3.3. Ordinary least squares and interval length. We investigate the special case of the ordinary least squares predictor \( \hat{\mu}_n(x) = x'\hat{\beta}_n = x'(X'X)^{\dagger}X'Y \) in some more detail, because here also the length

\[
\left| PI_{\alpha_1, \alpha_2}^{(LO)} \right| = \hat{q}_{\alpha_2} - \hat{q}_{\alpha_1},
\]

of the leave-one-out prediction interval (cf. (2.6) with \( k = n \)) admits a reasonably simple asymptotic characterization. We consider a class \( \mathcal{P}_n^{(lin)} \) which is a subset of the one of Condition (C2), with the additional assumption that the regression function \( \mu_P \) is linear and that the error distribution is fixed (up to arbitrary scaling).
(C3) Fix a finite constant $c_0 > 0$ and probability measures $\mathcal{L}_t$, $\mathcal{L}_w$ and $\mathcal{L}_v$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, such that $\mathcal{L}_w$ and $\mathcal{L}_v$ have mean zero, unit variance and finite fourth moment, $\int s^2 \mathcal{L}_t(ds) = 1$ and $\mathcal{L}_t((-c_0, c_0)) = 0$. For every $n \in \mathbb{N}$, the class $\mathcal{P}_n^{(\text{lin})} = \mathcal{P}_n^{(\text{lin})}(\mathcal{L}_t, \mathcal{L}_w, \mathcal{L}_v)$ consists of all probability measures $P$ on $Z \subseteq \mathbb{R}^{p_n+1}$, such that the following hold:

(a) The $x_0$-marginal distribution of $P$ is given by

$$P_{x_0} \overset{\mathcal{L}}{=} l_0 \Sigma_P^{1/2}(w_1, \ldots, w_{p_n}),$$

where $w_1, \ldots, w_{p_n}$ are i.i.d. according to $\mathcal{L}_w$, $l_0 \overset{\mathcal{L}}{=} \mathcal{L}_t$ is independent of the $w_j$ and $\Sigma_P$ is the unique symmetric positive definite square root of a positive definite $p_n \times p_n$ covariance matrix $\Sigma_P$.

(b) The conditional distribution of the response given the regressors is

$$P_{y_0|x_0} \overset{\mathcal{L}}{=} x_0' \beta + \sigma_P v_0,$$

where $v_0 \overset{\mathcal{L}}{=} \mathcal{L}_v$ is independent of $x_0$, and where $\beta_P \in \mathbb{R}^{p_n}$ and $\sigma_P \in (0, \infty)$.

Note that under (C3), the distributions $\mathcal{L}_t$, $\mathcal{L}_w$ and $\mathcal{L}_v$ are fixed, so that $\mathcal{P}_n^{(\text{lin})}$ is a parametric model indexed by $\beta_P$, $\Sigma_P$ and $\sigma_P$. However, these parameters may depend on sample size $n$, and the dimension $p_n$ of $\beta_P \in \mathbb{R}^{p_n}$ and of $\Sigma_P \in \mathbb{R}^{p_n \times p_n}$ may increase with $n$. Subsequently, we aim at uniformity in these parameters.

**Theorem 3.4.** Fix $\alpha \in [0, 1]$. For every $n \in \mathbb{N}$, let $P_n \in \mathcal{P}_n$ with $\mathcal{P}_n = \mathcal{P}_n^{(\text{lin})}(\mathcal{L}_t, \mathcal{L}_w, \mathcal{L}_v)$ as in (C3). If $p_n/n \to \kappa \in (0, 1)$ then the scaled empirical $\alpha$-quantile $\hat{q}_\alpha/\sqrt{p_n}$ of the leave-one-out residuals $\hat{u}_i = y_i - x_i' \hat{\beta}_n^{[i]}$ based on the OLS estimator $\hat{\beta}_n = (X'X)^{-1}X'Y$ converges in $P_n$-probability to the corresponding $\alpha$-quantile $q_\alpha$ of the asymptotic distribution of the scaled prediction error $(y_0 - x_0' \hat{\beta}_n)/\sqrt{p_n}$. The latter is equal in distribution to

$$\norm{\tau} + v,$$

where $l, N, \tau$ and $v$ are defined as follows: $l \overset{\mathcal{L}}{=} \mathcal{L}_t$, $N \overset{\mathcal{L}}{=} N(0, 1)$, and $v \overset{\mathcal{L}}{=} \mathcal{L}_v$ are independent, and $\tau = \tau(\mathcal{L}_t, \kappa)$ is non-random.

The same statement holds also for $\kappa = 0$, provided that, in addition, $\mathcal{L}_v$ has a continuous and strictly increasing cdf and $p_n \to \infty$ as $n \to \infty$.

Here, the function $\kappa \mapsto \tau(\mathcal{L}_t, \kappa) \in [0, \infty)$ defined on $[0, 1)$ has the following properties: For any $\mathcal{L}_t$ as in (C3), $\tau(\mathcal{L}_t, \kappa) = 0$ if, and only if, $\kappa = 0$. If $\mathcal{L}_t([-1, 1]) = 1$, then $\tau(\mathcal{L}_t, \kappa) = \sqrt{\kappa}/(1 - \kappa)$. 

Theorem 3.4 shows how the length $\hat{q}_{\alpha_2} - \hat{q}_{\alpha_1}$ of the leave-one-out prediction interval for the OLS predictor depends (asymptotically) on $L_i$, $L_v$ and $\kappa = \lim_{n \to \infty} p_n/n$. For simplicity, let $L_i(\{-1, 1\}) = 1$ and consider an equal tailed interval, i.e., $\alpha_1 = \alpha/2 = 1 - \alpha_2$. Figure 1 shows asymptotic interval lengths as functions of $\kappa \in [0, 1]$ for different values of error level $\alpha$ in the cases $L_v = \text{Unif}\{\{-1, 1\}\}$ and $L_v = \mathcal{N}(0, 1)$. For a wide range of $\kappa$ values ($\kappa > 0.8$), the interval length increases dramatically, as expected, because here the asymptotic estimation error $\tau = \sqrt{\kappa/(1 - \kappa)}$ explodes. We also get an idea about the impact of the error distribution, on which the practitioner has no handle. In particular, for large error levels ($\alpha = 0.6$) we even observe a non-monotonic dependence of the interval length on $\kappa$, which seems rather counterintuitive. This results from the non-monotonicity of $\tau^2 \mapsto \text{IQR}_\alpha(\mathcal{N}(0, \tau^2) \ast L_v) = q_{1-\alpha/2} - q_{\alpha/2}$, where $\ast$ denotes convolution. This non-monotonicity may only occur if the error distribution $L_v$ is not log-concave (e.g., the blue curve for $\alpha = 0.6$ in Figure 1; cf. the discussion in Subsection A.1 of the supplement). Finally, for large values of $\kappa$, and thus, for large values of $\tau$, the error distribution has little effect on the interval length, because in that case the term $N\tau$ dominates the distribution of $N\tau + v$.

The result of Theorem 3.4 can be intuitively understood as follows. If the true model $P_n^{(lin)}$ is linear and satisfies (C3) then the scaled prediction error under $P = P^{n+1}$, $P \in P_n^{(lin)}$, is distributed as

$$\frac{y_0 - x'_0\hat{\beta}_n}{\sigma_P} \leq l_0(w_1, \ldots, w_{p_n})\Sigma_{P}^{1/2}(\beta_P - \hat{\beta}_n)/\sigma_P + v_0,$$

and for $n$ large, $\|\Sigma_{P}^{1/2}(\beta_P - \hat{\beta}_n)/\sigma_P\|_2 \approx \tau$ is approximately non-random, so that $(w_1, \ldots, w_{p_n})\Sigma_{P}^{1/2}(\beta_P - \hat{\beta}_n)/\sigma_P \approx w'_0Z\tau$, where $Z := \Sigma_{P}^{1/2}(\beta_P - \hat{\beta}_n)/\|\Sigma_{P}^{1/2}(\beta_P - \hat{\beta}_n)\|_2$ is a random unit vector that is independent of $w_0 := (w_1, \ldots, w_{p_n})'$. Thus, if $p_n$ is large and $Z$ satisfies the Lyapounov condition $\|Z\|_{2+\delta} \to 0$, then $w'_0Z \approx \mathcal{N}(0, 1)$ (see Lemma C.9(ii) in the supplement). This effect of additional Gaussian noise in the prediction error was also observed by El Karoui (2013, 2018); El Karoui et al. (2013); El Karoui and Purdom (2018). Note that the conditions $\|\Sigma_{P}^{1/2}(\beta_P - \hat{\beta}_n)/\sigma_P\|_2 \approx \tau$ and $\|Z\|_{2+\delta} \to 0$ need not be satisfied by a given estimator $\hat{\beta}_n$. However, the former condition is always satisfied by the robust $M$-estimators

$$\hat{\beta}_n^{(\rho)} = \arg\min_{b \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(y_i - x'_ib),$$
considered in El Karoui (2013, 2018) and under the model assumptions in that reference (cf. Subsection 3.1). Here, \( \rho : \mathbb{R} \to \mathbb{R} \) is an appropriate convex loss function. The Lyapounov condition \( \|Z\|_{2+\delta} \to 0 \) is also satisfied by \( \hat{\beta}_n^{(\rho)} \), provided that the standardized design vectors \( \Sigma P^{-1/2} x_i \) follow an orthogonally invariant distribution, because then one easily sees that under (C3)

\[
\hat{\beta}_n^{(\rho)} = \beta_p + \Sigma P^{-1/2} \hat{\beta}_n^{(\rho)} \equiv \beta_p + \|\hat{\beta}_n^{(\rho)}\|_2 \Sigma P^{-1/2} U,
\]

where \( \hat{\beta}_n^{(\rho)} = \arg\min_{b \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \rho(\sigma P v_i - x_i' \Sigma P^{-1/2} b) \) and \( U \) is uniformly distributed on the unit sphere and independent of \( \|\hat{\beta}_n^{(\rho)}\|_2 = \|\Sigma P^{-1/2} (\beta_p - \hat{\beta}_n^{(\rho)})\|_2 \).

Therefore, \( \|Z\|_{2+\delta} \leq \|U\|_{2+\delta} \leq \left( \|V\|_{2+\delta}^2 / \|V\|_2^{2(1+\delta/2)} \right)^{\frac{1}{2+\delta}} \to 0 \), in probability, for \( V \sim \mathcal{N}(0, I_p) \), \( p_n \to \infty \) as \( n \to \infty \). However, the Lyapounov property is not satisfied, e.g., by the James-Stein estimators (cf. Lemma D.5 in the supplement). If the mentioned conditions are not satisfied, much more complicated limiting distributions of the prediction error than the one of Theorem 3.4 may arise.

### 3.4. Sample splitting

An obvious alternative to the \( k \)-fold cross validation prediction interval (2.6) is to use a sample splitting method as follows. Decide on a fraction \( \nu \in (0,1) \) and use only a number \( n_1 = \lfloor \nu n \rfloor \) of observation pairs \((x_i, y_i), i \in S_\nu \subseteq \{1, \ldots, n\}, |S_\nu| = n_1 \), to compute an estimate \( \hat{\mu}_{n_1} \). Now use the remaining \( n - n_1 \) observations to compute residuals \( \hat{u}_i^{(\nu)} = y_i - \hat{\mu}_{n_1}(x_i), i \in \{1, \ldots, n\} \setminus S_\nu \). Since, conditionally on the observations corresponding to \( S_\nu \), these residuals are i.i.d. and distributed as \( y_0 - \hat{\mu}_{n_1}(x_0) \), constructing a prediction interval of the form \( [\hat{\mu}_{n_1}(x_0) + L, \hat{\mu}_{n_1}(x_0) + U] \) for \( y_0 \) is now equivalent to constructing a tolerance interval for \( y_0 - \hat{\mu}_{n_1}(x_0) \) based on i.i.d. observations with the same distribution. One can now simply use appropriate empirical quantiles \( L = \hat{q}_\alpha^{(\nu)} \) and \( U = \hat{q}_{\alpha^2}^{(\nu)} \) from the sample splitting residuals \( \hat{u}_i^{(\nu)} \) (see also Subsection A.2 in the supplement). Such a procedure is suggested and studied, e.g., by Vovk (2012) and Lei et al. (2018). Notice, however, that this prediction interval is not centered at the more accurate point predictor \( \hat{\mu}_{n} \) that is computed from the full sample of size \( n \).

In order to formally study the length of this sample splitting interval we restrict to the case of OLS estimation, i.e., \( \hat{\mu}_{n_1}(x) = x' \hat{\beta}_{n_1} \), where \( \hat{\beta}_{n_1} \) is the OLS estimator based on the sample corresponding to \( S_\nu \). Note that in this case, the estimator will not be unique if \( n_1 < p_n \), so one usually requires \( n_1 \geq p_n \). Now, in view of Theorem 3.4, the empirical quantiles of
Fig 1. Lengths of leave-one-out prediction intervals as a function of $\kappa = \lim_{n \to \infty} p_n/n$ for confidence level $1 - \alpha$ and with $\text{Unif}(-1, 1)$ (binary) and $N(0, 1)$ (Gauss) errors.
the residuals \( \hat{u}_i^{(\nu)} \), \( i \in S^{(\nu)}_c \), converge (unconditionally) to the quantiles of \( lN \tau' + u \), where now \( \tau' \) is the non-random limit of \( \| \Sigma P^{1/2} (\beta_P - \hat{\beta}_{n_1})/\sigma_P \|_2 \). In particular, if \( \mathcal{L}_i \) degenerates to \( \{-1, 1\} \), then \( \tau' = \sqrt{\kappa'/(1 - \kappa')} \), where \( \kappa' = \lim_{n \to \infty} p_n/n_1 = \kappa/\nu \). Thus, we can read off the asymptotic interval length of the sample splitting procedure from Figure 1 by simply adjusting the value of \( \kappa \) to \( \kappa/\nu \). For instance, in the binary error case with \( \alpha = 0.05 \), if \( \kappa = 0.4 \) and we use sample splitting with \( \nu = 1/2 \), then \( \kappa' = 0.8 \) and the asymptotic length of the leave-one-out prediction interval is about 4.7, while the asymptotic length of the sample splitting interval is about 9, so almost twice as wide.

### 4. Asymptotically degenerate (non-random) estimators.

Another important class of problems, where the conditions (2.12) and (2.13) of Subsection 2.2 are satisfied, are those where the estimator \( \hat{\mu}_n \) (and those computed without the observations in the \( l \)-th fold \( \hat{\mu}_{n}^{(l)} \)) asymptotically degenerates to some non-random function which need not be the true regression function \( \mu_P(x_0) = \mathbb{E}_P[y_0|x_0] \). We point out that in the scenario considered in this section, the naive approach that tries to estimate the true unknown distribution of the errors \( u_i = y_i - \mu_P(x_i) \) in the additive error model (C2) based on the ordinary residuals \( y_i - \hat{\mu}_n(x_i) \) is often successful (asymptotically) for constructing training conditionally valid prediction intervals, provided that consistent estimation of \( \mu_P \) is possible. This less challenging but more classical setting of asymptotically non-random predictors is an important test case for the CV method. We still consider asymptotic results where the number of explanatory variables \( p = p_n \) and the number of folds \( k = k_n \) can grow with sample size \( n \). Thus, we consider sequences \( (p_n)_{n \in \mathbb{N}} \) and \( (k_n)_{n \in \mathbb{N}} \), and a sequence \( (P_n)_{n \in \mathbb{N}} \) of probability measures, where \( P_n \) is supported on \( Z_n \subseteq \mathbb{R}^{p_n+1} \). Moreover, we have to slightly extend the usual definition of uniform consistency of an estimator sequence to cover also the estimates \( \hat{\mu}_{n}^{(l)} \) omitting the \( l \)-th fold and to allow the possibility of an asymptotically non-vanishing bias.

**Definition 4.1 (Uniform Asymptotic Degeneracy (UAD)).** For every \( n \in \mathbb{N} \), let \( k_n, p_n \in \mathbb{N} \), let \( P_n \) be a probability measure on \( Z_n \) and let \( \sigma_n^2 \in (0, \infty) \). We say that a sequence of predictors \( \hat{\mu}_n \) is uniformly asymptotically degenerate (UAD) with respect to \( (P_n)_{n \in \mathbb{N}} \) and relative to \( (\sigma_n^2)_{n \in \mathbb{N}} \), if there
exists measurable functions $g_n : \mathbb{R}^{p_n} \to \mathbb{R}$, such that for every $\varepsilon > 0$,

\begin{equation}
|g_n(x_{0,n}) - \hat{\mu}_n(x_{0,n})| = o_{P_n}(\sigma_n) \quad \text{and}
\end{equation}

\begin{equation}
\frac{1}{k_n} \sum_{l=1}^{k_n} \mathbb{E}_n \left[ \left( \frac{|g_n(x_{0,n}) - \hat{\mu}_n^{(l)}(x_{0,n})|}{\sigma_n} \right) \wedge 1 \right] = o(1).
\end{equation}

In the classical case of consistent estimation, the function $g_n$ would be the true unknown regression function $g_n(x_{0,n}) = \mu_{P_n}(x_{0,n})$ and the constant $\sigma^2_n$ can be thought of as the error variance $\sigma^2_n = \text{Var}_n[y_{0,n} - \mu_{P_n}(x_{0,n})]$, if it exists, or can be chosen equal to 1.

It is convenient and common practice to choose the $k$-folds $K_1, \ldots, K_k$ of approximately the same size. We consider the following convention: In case $n/k \in \mathbb{N}$, choose all folds of equal size $\frac{n}{k}$ and otherwise choose $\ell := n - \left\lfloor \frac{n}{k} \right\rfloor k \in [k - 1]$ and $m_1 = m_2 = \cdots = m_\ell = \left\lceil \frac{n}{k} \right\rceil + 1$ and $m_{\ell+1} = \cdots = m_k = \frac{n}{k}$.

With this convention, the left hand side of (4.2) can be written as

\begin{equation}
\frac{\ell}{k_n} \mathbb{E}_n \left[ \left( \frac{|g_n(x_{0,n}) - \hat{\mu}_n^{(1)}(x_{0,n})|}{\sigma_n} \right) \wedge 1 \right] + \frac{k_n - \ell}{k_n} \mathbb{E}_n \left[ \left( \frac{|g_n(x_{0,n}) - \hat{\mu}_n^{(k_n)}(x_{0,n})|}{\sigma_n} \right) \wedge 1 \right],
\end{equation}

irrespective of whether $\hat{\mu}_n$ is symmetric or not. Hence, (4.2) follows if

\begin{equation}
|g_n(x_{0,n}) - \hat{\mu}_n^{(1)}(x_{0,n})| + |g_n(x_{0,n}) - \hat{\mu}_n^{(k_n)}(x_{0,n})| = o_{P_n}(\sigma_n).
\end{equation}

Thus, under this construction it actually suffices to verify (4.1) and (4.3) to establish UAD. Furthermore, in many cases (4.3) will even be equivalent to (4.1), provided that $n - m_1$ is growing sufficiently fast with $n$.

It is easy to see that if $\hat{\mu}_n$ is UAD with respect to $(P_n)_{n \in \mathbb{N}}$ and relative to $(\sigma^2_n)_{n \in \mathbb{N}}$ and if (C1) holds, then the sequence of stability constants $\eta_{n,k_n}$ satisfies (2.13), i.e.,

\begin{equation}
\eta_{n,k_n} = \frac{1}{k_n} \sum_{l=1}^{k_n} \mathbb{E}_n \left[ \left( \|f_{y_{0,n},||x_{0,n}||}\|_\infty |\hat{\mu}_n(x_{0,n}) - \hat{\mu}_n^{(l)}(x_{0,n})| \right) \wedge 1 \right] \to 0, \quad n \to \infty,
\end{equation}

provided that $\sigma_n f_{y_{0,n},||x_{0,n}||}_\infty = O_{P_n}(1)$. Simply use the inequality $(a+b) \wedge 1 \leq (a \wedge 1) + (b \wedge 1)$, for $a, b \geq 0$. Also note that $\|\sigma_n f_{y_{0,n},||x_{0,n}||}_\infty \|$ is the sup-norm of the conditional density of the scaled error term $(y_{0,n} - g_n(x_{0,n}))/\sigma_n$ given $x_{0,n}$. 
In the remainder of this subsection we list a number of examples where
the UAD property of $\hat{\mu}_n$ holds. Therefore (assuming (C1) and the bounded-
ness conditions of Theorem 2.4) also asymptotic training conditional validity
of the $k$-CV prediction interval follows. We emphasize that the conditions
on the statistical model $\mathcal{P}$, that are imposed in the subsequent examples,
are taken from the respective reference and we do not claim that they are
minimal.

Example 4.2 (Non-parametric regression estimation). Consider a con-
stant sequence of dimension parameters $p_n = p \in \mathbb{N}$. For positive finite
constants $L$ and $C$, let $\mathcal{P}(L, C)$ denote the class of probability distributions
$P$ on $Z = \mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^{p+1}$ such that $P_{y_0}(|y_0| \leq L) = 1 = P_{x_0}(\|x_0\|_2 \leq L)$
and whose corresponding regression function $\mu_P : \mathbb{R}^p \rightarrow \mathbb{R}$ is C-Lipschitz,

\[ \|\mu_P(x_1) - \mu_P(x_2)\| \leq C\|x_1 - x_2\|_2 \] for all $x_1, x_2 \in \mathcal{X}$. Györfi et al. (2002,
Chapter 7) show that if $\hat{\mu}_n$ is either an appropriate kernel estimate, a par-
titioning estimate or a nearest-neighbor estimate, all with fully data driven
choice of tuning parameter, then, if $P_n \in \mathcal{P}(L, C)$,

\[ P_n(\|\hat{\mu}_n(x_{0,n}) - \mu_{P_n}(x_{0,n})\| > \varepsilon) \rightarrow 0 \]

for every $\varepsilon > 0$. Thus (4.1) holds with $g_n = \mu_{P_n}$ and $\sigma_n = 1$, and under the
(approximate) equal fold size construction mentioned above, (4.3) also holds
provided $n - m_1 \rightarrow \infty$, because here $p_n = p$ is constant, and UAD follows.

Example 4.3 (Deep neural networks). Schmidt-Hieber (2020) considers
a nonparametric regression model similar to (C2), but with $\sigma_P = 1$ and $v_P \sim \mathcal{N}(0, 1)$. The distribution of the regressors $x_i$ is supported on $\mathcal{X} = [0, 1]^p$ but
is otherwise unrestricted. He assumes a smooth compositional structure for
the true unknown regression function, that is, $\mu_P \in \mathcal{G}(q, d, t, \beta, K)$, where

\[ \mathcal{G}(q, d, t, \beta, K) = \left\{ f = g_q \circ \cdots \circ g_0 | g_i : [a_i, b_i]^{d_i} \rightarrow [a_{i+1}, b_{i+1}]^{d_{i+1}}, \right. \]

\[ g_i \in C^\beta_{t_i}([a_i, b_i]^{t_i}, K), \text{ for some } |a_i|, |b_i| \leq K \}, \]

$q \in \mathbb{N}$, $d = (d_0, \ldots, d_q) \in \mathbb{N}^{q+2}$, $d_0 = p, d_{q+1} = 1$, $t = (t_0, \ldots, t_q) \in \mathbb{N}^{q+1}$,
$t_i \leq d_i$, for $i \leq q + 1$, $\beta = (\beta_0, \ldots, \beta_q) \in (0, \infty)^{q+1}$, $C^\beta_{t_i}([a_i, b_i]^{t_i}, K)$ is the set of Hölder functions on $[a_i, b_i]^{t_i}$ with Hölder-exponent $\beta_i$ and Hölder
norm bounded by $K$, and it is implicitly assumed that $g_i$ depends only on
$t_i$ of its $d_i$ possible input variables. Consider a predictor $\hat{\mu}_n$ that is obtained
by empirical risk minimization over a certain class $F(L, (p_i)_{i=0}^{L+1}, s, F)$ of $F$-
bounded (i.e., $\|\hat{\mu}_n\|_\infty \leq F$) ReLU networks with $L + 1$ layers where the $i$-th
layer has output dimension $p_i$, $p_0 = p$ and $p_{L+1} = 1$, and which have bounded and $s$-sparse network parameters. If the tuning parameters $L, (p_i)_{i=0}^{L+1}, s$ and $F$ are chosen appropriately, and if, in particular, the network depth $L = L_n$ is logarithmically growing with $n$, Schmidt-Hieber (2020, Corollary 1) shows that

$$
\mathbb{E} \left[ (\hat{\mu}_n(x_0) - \mu_P(x_0))^2 \right] \leq C' L_n \phi_n \log^2 n,
$$

where $C' = C_p$ does not depend on $n$ (but on $p$),

$$
\phi_n = \max_{i=0, \ldots, q} n^{-\frac{2 \beta_i^*}{2 t_i^* + 1}},
$$

and $\beta_i^* = \beta_i \prod_{\ell=i+1}^q (\beta_\ell \land 1)$. Thus, (4.1) and (4.2) are satisfied with $g_n = \mu_{P_n}$ and $\sigma_n = 1$, provided that $C'_p L_n \phi_n \log^2 n^* = o(1)$, where $n^* = n - \max_l m_l$ is the number of observed feature-response pairs excluding the ones in the largest fold.

**Example 4.4 (Random forests).** Scornet et al. (2015) consider an additive regression model similar to (C2), but with $v_P \sim N(0,1)$, $P_{x_0}$ the uniform distribution on $[0,1]^p$ and additive regression function

$$
\mu(x) = \mu_P(x) = \sum_{j=1}^p \mu_j(x^{(j)}), \quad x = (x^{(1)}, \ldots, x^{(p)})' \in \mathbb{R}^p,
$$

where the $\mu_j : [0,1] \rightarrow \mathbb{R}$ are continuous. A random forest predictor $\hat{\mu}_n$ is trained in the following way: Each tree is grown by first randomly subsampling a number of $a_n \in [n]$ feature-response pairs from the training set and then growing a tree according to the CART-split criterion (see Scornet et al., 2015, Section 2, for details). The tree growing process is terminated when a number of $t_n \in [n]$ leaves is reached. In this way, $M$ random regression trees are grown, each of which predicts $y_0$ by averaging all the $y_i$ whose features $x_i$ belong to the leaf containing $x_0$. Scornet et al. (2015) study the idealized predictor $\hat{\mu}_n(x_0)$ obtained from averaging these $M$ predictions and letting $M \rightarrow \infty$. They show that if $a_n, t_n \rightarrow \infty$ in such a way that $t_n (\log a_n)^{\alpha} / a_n \rightarrow 0$, we have for every fixed data generating distribution $P$ as above,

$$
\mathbb{E}[(\hat{\mu}_n(x_0) - \mu(x_0))^2] \rightarrow 0, \quad \text{as } n \rightarrow \infty.
$$

Thus (4.1) holds, and under the (approximate) equal fold size construction mentioned above, (4.3) also holds provided $n - m_1 \rightarrow \infty$, because here $p_n = p$ is constant.
Example 4.5 (High-dimensional linear regression with the LASSO). Consider a non-decreasing sequence \( (K_n)_{n \in \mathbb{N}} \) of positive real numbers and a sequence of dimension parameters \( (p_n)_{n \in \mathbb{N}} \) such that \( K_n^4 \log(p_n)/n \to 0 \) as \( n \to \infty \). For a positive finite constant \( M \), let \( \mathcal{P}_n(M) \) denote the class of probability distributions on \( \mathbb{R}^{p_n+1} \), such that under \( P \in \mathcal{P}_n(M) \), the pair \( (x_0, y_0) \) has the following properties:

- \( \|x_0\|_\infty \leq M \), almost surely.
- Conditional on \( x_0, y_0 \) is distributed as \( \mathcal{N}(x_0' \beta_P, \sigma_P^2) \), for some \( \beta_P \in \mathbb{R}^{p_n} \) and \( \sigma_P^2 \in (0, \infty) \).
- The parameters \( \beta_P \) and \( \sigma_P^2 \) satisfy \( \max(\|\beta_P\|_1, \sigma_P^2) \leq K_n \).

In particular, we have \( \mu_P(x_0) = x_0' \beta_P \). Chatterjee (2013, Theorem 1) shows that any estimate \( \hat{\beta}_{n,K_n} \) which minimizes

\[
\beta \mapsto \sum_{i=1}^{n} (y_i - \beta' x_i)^2 \quad \text{subject to} \quad \|\beta\|_1 \leq K_n
\]

satisfies

\[
\mathbb{P}_n \left( \left| x_0' \hat{\beta}_{n,K_n} - \mu_P(x_0) \right| > \varepsilon \right) \overset{n \to \infty}{\longrightarrow} 0
\]

for every \( \varepsilon > 0 \) and if \( P_n \in \mathcal{P}_n(M) \). Clearly, here the estimate \( \hat{\beta}_{n,K_n}^{(1)} \) excluding the data from the first fold has the same asymptotic property, provided that \( K_n^4 d_1 \log(p_n)/(n - m_1) \to 0 \), with \( m_1 = \lfloor n/2 \rfloor \). Thus, under the (approximately) equal fold size construction, UAD holds. Note that in this example, consistent estimation of the parameters \( \beta_P \) and \( \sigma_P^2 \) would require additional assumptions on the distribution of the feature vector \( x_0 \) (so called ‘compatibility conditions’, see Bühlmann and van de Geer, 2011), and therefore, it is not immediately clear whether the usual Gaussian prediction interval, based on estimates \( \hat{\beta}_n \) and \( \hat{\sigma}^2_n \) and a Gaussian quantile, is asymptotically valid in the present setting. Furthermore, the result of Chatterjee (2013) can be extended also to the non-Gaussian case, where the usual Gaussian prediction interval certainly fails.

Example 4.6 (Ridge regression with many variables). A qualitatively different parameter space is considered in Lopes (2015), who shows uniform consistency of ridge regularized estimators in a linear model under a boundedness assumption on the regression parameter \( \beta_P \) and a specific decay rate of eigenvalues of \( \mathbb{E}_x_0[x_0x_0'] \).
Example 4.7 (Misspecified regression estimation). A classical strand of literature on the asymptotics of Maximum-Likelihood under misspecification has established various conditions under which the MLE is not consistent for the true unknown parameter, but for a pseudo parameter that corresponds to the projection of the true data generating distribution onto the maintained working model. See, for example, Huber (1967), White (1980a,b) or Fahrmeir (1990). A common pseudo target in random design regression is the minimizer of $\beta \mapsto \mathbb{E}_{z_0}[(y_0 - \beta'x_0)^2]$.

5. Numerical results. We conduct an extensive simulation study to assess the quality of our theoretical approximations in a small sample situation. For our numerical experiments we closely follow the setup of Barber et al. (2021b). We compute prediction intervals with nominal coverage probability of $1 - \alpha = 0.9$ using i.i.d. training samples of different sizes $n$ and $p = n$ explanatory variables. Each data point $(x_i, y_i)$ is generated as $x_i \sim \mathcal{N}(0, I_p)$ and $y_i \sim \mathcal{N}(\beta' x_i, 1), \quad i = 1, \ldots, n$.

We also generate 100 test data points from the same distribution. The true coefficient vector $\beta$ is drawn once for each parameter configuration as $\beta = \sqrt{10} \cdot u$ for a uniform random unit vector $u$ in $\mathbb{R}^p$ and is then fixed throughout the Monte Carlo iterations. Instead of the least squares estimator (or the minimum norm interpolator in case $p > n$) studied in the simulation section 7.1 of Barber et al. (2021b), which was exemplified to be unstable in the present regime of $n \approx p$ and results in anti-conservative performance, we here investigate the LASSO as our prediction algorithm. For simplicity and reproducibility we take the default implementation of the LASSO in the R-package glmnet (Friedman et al., 2010) with data driven choice of tuning parameter determined by 5-fold cross validation.

First, we try to experimentally confirm our theoretical prediction that the cross-validation prediction intervals are asymptotically training conditionally valid even in scenarios where no consistent parameter estimation is possible. For simplicity, we consider only leave-one-out cross validation ($k = n$). Recall that in the challenging scenario where $n = p$, stability of the OLS predictor fails. Notice also that the condition $\|x_0\|_{\infty} \leq M$ for consistency of the LASSO predictor of Example 4.5 is not satisfied in the current setup. For each $n$, using the 100 test data points, we compute the empirical probability of coverage. For one fixed training sample $T_n$, this gives us a Monte Carlo

---

3 cv.glmnet(x=X, y=Y, family="gaussian", intercept=FALSE, nfolds=5, alpha=1) in glmnet version 4.1-2.
estimate of the conditional coverage probability \( \mathbb{P}(y_0 \in PI_{\alpha}^{(L1O)}(T_n, x_0) | T_n) \).

In Figure 2 we show histograms of these 100 Monte Carlo estimates, their sample means (black solid lines) and the nominal coverage level \( 1 - \alpha = 0.9 \) (red dashed lines). For small samples we observe an anti-conservative bias as well as a considerable variability in the true conditional coverage probabilities, both of which are reduced as sample size \( n \) increases, even though \( p = n \) increases as well.

In Figure 3 we show average coverage (i.e., Monte Carlo estimates of the unconditional coverage probabilities \( \mathbb{P}(y_0 \in PI_{\alpha}(T_n, x_0)) \)), average length (i.e., Monte Carlo estimates of the expected length) and a Monte Carlo estimate \( \hat{\eta}_{n,k} \) of the stability coefficient \( \eta_{n,k} \) (cf. Definition 2.3) as functions of \( n \) (here, still, \( n = k = p \)). We estimate the stability of the predictor \( \hat{\mu}_n \) on a given training sample \( T_n \), by

\[
\hat{\eta}_{n,k}(T_n) = \frac{1}{k} \sum_{l=1}^{k} \frac{1}{100} \sum_{j=1}^{100} \left( \frac{1}{\sqrt{2\pi}} \left| \hat{\mu}_n(x_{0,j}) - \hat{\mu}_n^{(l)}(x_{0,j}) \right| \right) \wedge 1,
\]

where \( x_{0,j}, j = 1, \ldots, 100 \) are the regressors in the test set, and then average over the 100 Monte Carlo samples \( T_n \) to approximate \( \eta_{n,k} \). As we already saw in Figure 2 (black bold vertical lines), Figure 3 also shows that the unconditional coverage probabilities increase towards the nominal level with increasing \( n \) and \( p \). Simultaneously, the expected interval lengths and the stability coefficients decrease with increasing \( n \). Notice that a feasible but trivial \( 1 - \alpha \) prediction (or tolerance) interval for \( y_0 \) in the present scenario is

\[
TI_{\alpha} = \pm q_{1 - \alpha/2} \cdot \sigma,
\]

with \( \sigma^2 = \text{Var}[y_0] = \|\beta\|_2^2 + 1 = 11 \), which has length \( 2 \cdot q_{0.95}^N \cdot \sqrt{11} = 10.91 \). This is well outperformed by the leave-one-out prediction interval even in small samples. For reference, the oracle prediction interval for \( y_0 \) which uses knowledge of the true coefficient vector \( \beta \), has length \( 2 \cdot q_{0.95}^N = 3.29 \).

Next, we investigate the impact of the number of folds \( k \) on the \( k \)-CV prediction interval. We compute (2.6) for a range of \( k \) from 2 up to sample size \( n \). In Figure 4 we plot histograms of conditional coverage probabilities for different values of \( k \), in the case \( n = p = 100 \). We find that for small values of \( k \), most training samples produce conservative prediction intervals, whereas for larger \( k \) the conditional coverages are more symmetrically spread around the nominal level with a slight anti-conservative bias (as was already observed in Figure 2). The number of folds \( k \) does not seem to have a strong impact on the spread of these conditional probabilities. The conservative behavior for small \( k \) can be explained in a similar way as for the case of
Fig 2. Histograms of conditional coverage probabilities $P(y_0 \in PI^{(L1O)}(T_n, x_0)\|T_n)$ from 100 Monte Carlo training samples $T_n$ for different values of $n$ at confidence level of $1 - \alpha = 0.9$ (red dashed line). Here $k = p = n$. 
sample splitting (cf. Subsection 3.4 and Subsections A.1 and A.2 in the supplement). In a nutshell, if $k$ is small, then $\hat{\mu}_n^{(l)}$, $l = 1, \ldots, k$, will be based on much fewer observations than $\hat{\mu}_n$ and will thus be less accurate for prediction, leading to residuals that will have much larger variance than the true prediction error $y_0 - \hat{\mu}_n(x_0)$. Hence, the resulting prediction interval will be too wide. This reasoning is also supported by Figure 5.

From Figure 5 we learn that coverage probabilities, expected interval lengths and stability all decrease with increasing $k$ in a quite similar fashion. This was to be expected for the stability, because for nearly equal fold sizes, the number of observations in each fold decreases as $k$ increases, which means that, on average, $|\hat{\mu}_n(x_0) - \hat{\mu}_n^{(l)}(x_0)|$ will decrease for each $l$, as the two predictors are computed on nearly the same observations. We also see in Figure 5 that apparently for smaller values of $k$ below $n/2$, the $k$-CV prediction interval is too wide, as we observe both, large interval lengths as well as over-coverage (cf. the discussion in the previous paragraph). On the other hand, there does not seem to be any practical reason to choose $k$ substantially larger than $n/2$, as there is no benefit in terms of validity or interval length for increasing the number of folds above that threshold. However, keep in mind that the runtime of the $k$-CV prediction interval increases (roughly) linearly in $k$ (cf. Remark A.5 in the supplement).

We also investigated the CV+ procedure of Barber et al. (2021b), which is called Jackknife+ in case of leave-one-out (i.e., the number of folds is $k = n$). Again, in Figure 6 we see, similarly to Figure 2, that the conditional coverage probabilities of the Jackknife+ method concentrate around the nominal
Fig 4. Histograms of conditional coverage probabilities $\mathbb{P}(y_0 \in P_1^{(kCV)}(T_n, x_0)||T_n)$ from 100 Monte Carlo training samples $T_n$ for different values of $k$ at confidence level of $1 - \alpha = 0.9$ (red dashed line). Here $n = p = 100$. 
level as $n \to \infty$, even though $p/n = 1$. However, here we observe a slight conservative bias in small samples as opposed to the slight anti-conservative bias of the Jackknife in Figure 2. Figure 7 is analogous to Figure 4 and exhibits the effect of the choice of $k$ on the conditional coverage probabilities of the CV+ procedure. Somewhat surprisingly, we find that the number of folds does not seem to have a strong impact on the coverage probability. The picture is generally a more conservative one with slightly wider intervals compared to the $k$-CV procedure, as seen in Figure 8. Generally speaking, the two procedures seem to behave quite similarly in this stable scenario while it is known that CV+ is more robust against instability. If CV+ also proves to be (asymptotically) training conditionally valid under reasonable stability assumptions, then it could be a practically preferable alternative to the more classical $k$-CV method investigated here.

Finally, we repeated the simulations for heteroskedastic $y_i$ generated as

$$y_i \parallel x_i \sim N(\beta' x_i, \sigma^2(x_i)), \quad \sigma^2(x_i) = 1 + \frac{|\beta' x_i|}{\sqrt{10}} - \frac{\sqrt{2}}{\Gamma\left(\frac{1}{2}\right)}.$$  

For comparability we have chosen the conditional variance of $y_i$ such that $E[\sigma^2(x_i)] = 1$ in order to achieve $\text{Var}(y_i) = 11$, as in the homoskedastic case. The resulting plots are very similar to those of the homoskedastic case and are deferred to Section B in the supplement. For heteroskedastic data we observe slightly larger spread of conditional probabilities around the nominal level, slightly wider intervals and slightly slower asymptotic convergence of average (unconditional) coverage to the nominal level.
Fig 6. Histograms of conditional coverage probabilities $P(y_0 \in P_{\alpha}^{\text{Jackknife}}(T_n, x_0) | T_n)$ from 100 Monte Carlo training samples $T_n$ for different values of $n$ at confidence level of $1 - \alpha = 0.9$ (red dashed line). Here $k = p = n$. 
Fig 7. Histograms of conditional coverage probabilities \( P(y_0 \in PI^{CV+}_n(T_n, x_0) \| T_n) \) from 100 Monte Carlo training samples \( T_n \) for different values of \( k \) at confidence level of \( 1 - \alpha = 0.9 \) (red dashed line). Here \( n = p = 100 \).
Fig 8. Unconditional coverage probability, expected interval length and estimated stability of \( CV^+ \) for different numbers of folds \( k \) from 100 Monte Carlo simulations with \( n = p = 100 \).

**Acknowledgements.** The authors thank the participants of the “ISOR Research Seminar in Statistics and Econometrics” at the University of Vienna for discussion of an early version of the paper. In particular, we want to thank Benedikt M. Pötscher and David Preinerstorfer for valuable comments. We are also grateful to three anonymous referees and an associate editor for their constructive feedback to improve the paper.

**SUPPLEMENTARY MATERIAL**

Supplement to “Conditional predictive inference for stable algorithms” (supplement.pdf). The supplementary material contains further discussions, simulation results and most of the proofs.

**References.**

Barber, R. F., E. J. Candes, A. Ramdas, and R. J. Tibshirani (2021a). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA* 10(2), 455–482.
Barber, R. F., E. J. Candes, A. Ramdas, and R. J. Tibshirani (2021b). Predictive inference with the jackknife+. *Ann. Statist.* 49(1), 486–507.
Bates, S., E. Candès, L. Lei, Y. Romano, and M. Sesia (2021). Testing for outliers with conformal p-values. *arXiv preprint arXiv:2104.08279*.
Bean, D., P. J. Bickel, N. El Karoui, and B. Yu (2013). Optimal m-estimation in high-dimensional regression. *Proc. Natl. Acad. Sci. USA* 110(36), 14563–14568.
Bickel, P. J. and D. A. Freedman (1983). Bootstrapping regression models with many parameters. In P. Bickel, K. Doksum, and J. Hodges (Eds.), *A Festschrift for Erich L. Lehmann*, pp. 28–48. Wadsworth Inc.

Bousquet, O. and A. Elisseeff (2002). Stability and generalization. *J. Mach. Learn. Res.* 2, 499–526.

Bucchianico, A. D., J. H. J. Einmahl, and N. A. Mushkudiani (2001). Smallest nonparametric tolerance regions. *Ann. Statist.* 29(5), 1320–1343.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-dimensional Data*. Berlin: Springer.

Butler, R. and E. D. Rothman (1980). Predictive intervals based on reuse of the sample. *J. Amer. Statist. Assoc.* 75(372), 881–889.

Chatterjee, S. (2013). Assumptionless consistency of the lasso. *arXiv preprint arXiv:1303.5817*.

Chatterjee, S. K. and N. K. Patra (1980). Asymptotically minimal multivariate tolerance sets. *Calcutta Statist. Assoc. Bull.* 29(1-2), 73–94.

Chen, W., K.-J. Chun, and R. F. Barber (2018). Discretized conformal prediction for efficient distribution-free inference. *Stat* 7(1), e173.

Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics*. London: Chapman and Hall.

Devroye, L. and T. J. Wagner (1979). Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Trans. Inform. Theory* 25(2), 202–207.

Dicker, L. H. (2013). Optimal equivariant prediction for high-dimensional linear models with arbitrary predictor covariance. *Electron. J. Stat.* 7, 1806–1834.

El Karoui, N. (2010). The spectrum of kernel random matrices. *Ann. Statist.* 38(1), 1–50.

El Karoui, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*.

El Karoui, N. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probab. Theory Relat. Fields* 170(1-2), 95–175.

El Karoui, N., D. Bean, P. J. Bickel, C. Lim, and B. Yu (2013). On robust regression with high-dimensional predictors. *Proc. Natl. Acad. Sci. USA* 110(36), 14557–14562.

El Karoui, N. and E. Purdom (2018). Can we trust the bootstrap in high-dimensions? the case of linear models. *J. Mach. Learn. Res.* 19(1), 170–235.

Fahrmeir, L. (1990). Maximum likelihood estimation in misspecified generalized linear models. *Statistics* 21(4), 487–502.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1), 1–22.

Györfi, L., M. Kohler, A. Krzyżak, and H. Walk (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. New York: Springer.

Huber, N. and H. Leeb (2013). Shrinkage estimators for prediction out-of-sample: Conditional performance. *Commun. Statist. - Theory Methods* 42(7), 1246–1264.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1, 221–233.

Kim, B. and R. F. Barber (2021). Black box tests for algorithmic stability. *arXiv preprint arXiv:2111.15546*.

Krishnamoorthy, K. and T. Mathew (2009). *Statistical Tolerance Regions: Theory, Applications, and Computation*. Wiley Series in Probability and Statistics. Hoboken, New Jersey: Wiley.

Lei, J. (2019). Fast exact conformalization of the lasso using piecewise linear homotopy.
Biometrika 106(4), 749–764.
Lei, J., M. G’Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman (2018). Distribution-free predictive inference for regression. J. Amer. Statist. Assoc. 113(523), 1094–1111.
Lei, J., J. Robins, and L. Wasserman (2013). Distribution-free prediction sets. J. Amer. Statist. Assoc. 108(501), 278–287.
Lei, J. and L. Wasserman (2014). Distribution-free prediction bands for non-parametric regression. J. Roy. Statist. Soc. Ser. B 76, 71–96.
Li, J. and R. Y. Liu (2008). Multivariate spacings based on data depth: I. construction of nonparametric multivariate tolerance regions. Ann. Statist. 36(3), 1299–1323.
Lopes, M. E. (2015). Some Inference Problems in High-Dimensional Linear Models. Ph. D. thesis, UC Berkeley.
Mammen, E. (1996). Empirical process of residuals for high-dimensional linear models. Ann. Statist. 24(1), 307–335.
Olive, D. J. (2007). Prediction intervals for regression models. Comput. Statist. Data Anal. 51(6), 3115–3122.
Politis, D. N. (2013). Model-free model-fitting and predictive distributions. Test 22(2), 183–221.
Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with relu activation function. Ann. Statist. 48(4), 1875–1897.
Schmoyer, R. L. (1992). Asymptotically valid prediction intervals for linear models. Technometrics 34(4), 399–408.
Scornet, E., G. Biau, and J.-P. Vert (2015). Consistency of random forests. Ann. Statist. 43(4), 1716–1741.
Steinberger, L. and H. Leeb (2016). Leave-one-out prediction intervals in linear regression models with many variables. arXiv preprint arXiv:1602.05801.
Steinberger, L. and H. Leeb (2021). Conditional predictive inference for high-dimensional stable algorithms. arXiv preprint arXiv:1809.01412.
Steinberger, L. and H. Leeb (2022). Supplement to “Conditional predictive inference for stable algorithms”.
Stine, R. A. (1985). Bootstrap prediction intervals for regression. J. Amer. Statist. Assoc. 80(392), 1026–1031.
Tukey, J. W. (1947). Non-parametric estimation ii. statistically equivalent blocks and tolerance regions – the continuous case. Ann. Math. Statist. 18(4), 529–539.
Vovk, V. (2012). Conditional validity of inductive conformal predictors. In Asian conference on machine learning, pp. 475–490.
Vovk, V. (2013). Conditional validity of inductive conformal predictors. Machine Learning 92(2), 349–376.
Vovk, V., A. Gammerman, and C. Saunders (1999). Machine-learning applications of algorithmic randomness. In I. Bratko and S. Dzeroski (Eds.), Proceedings of the Sixteenth International Conference on Machine Learning, ICML ’99, pp. 444–453. Morgan Kaufmann Publishers Inc.
Vovk, V., A. Gammerman, and G. Shafer (2005). Algorithmic Learning in a Random World. New York: Springer.
Vovk, V., I. Nouretdinov, and A. Gammerman (2009). On-line predictive linear regression. Ann. Statist. 37(3), 1566–1590.
Wald, A. (1943). An extension of Wilks’ method for setting tolerance limits. Ann. Math. Statist. 14(1), 45–55.
White, H. (1980a). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48(4), 817–838.
White, H. (1980b). Using least squares to approximate unknown regression functions.
Wilks, S. S. (1941). Determination of sample sizes for setting tolerance limits. *Ann. Math. Statist.* 12(1), 91–96.

Wilks, S. S. (1942). Statistical prediction with special reference to the problem of tolerance limits. *Ann. Math. Statist.* 13(4), 400–409.