Ann Inst Stat Math (2010) 62:515–546
DOI 10.1007/s10463-008-0183-3

Bootstrap model selection for possibly dependent and heterogeneous data

Alessio Sancetta

Received: 28 August 2006 / Revised: 20 February 2008 / Published online: 16 July 2008

Abstract This paper proposes the use of the bootstrap in penalized model selection for possibly dependent heterogeneous data. The results show that we can establish (at least asymptotically) a direct relationship between estimation error and a data based complexity penalization. This requires redefinition of the target function as the sum of the individual expected predicted risks. In this framework, the wild bootstrap and related approaches can be used to estimate the penalty with no need to account for heterogeneous dependent data. The methodology is highlighted by a simulation study whose results are particularly encouraging.

Keywords Complexity regularization · Random penalty · Wild bootstrap

1 Introduction

This paper derives a bound for the penalized model selection problem. This bound is then used to derive and study bootstrap penalties uniform over each class of competing models. Improvements with penalties over subsets of competing models are also studied and their performance is highlighted via simulation.

The model selection problem using penalties that approximate the estimation error uniformly over each class of models has been pioneered by Vapnik and Chervonenkis and it is usually referred to as the structural minimization approach (e.g., Vapnik

I thank the associate editor and the referee for comments that improved the quality and presentation of the paper.

A. Sancetta (✉)
Faculty of Economics, University of Cambridge, Austin Robinson Building,
Sidgwick Avenue, Cambridge CB3 9DD, UK
e-mail: asancetta@gmail.com

1998). A problem with the original approach is that the penalty does not depend on the sample sequence; consequently it overestimates the estimation error. Subsequent literature has focused on more data dependent penalties in order to obtain better uniform estimates of the estimation error (e.g., Koltchinskii 2001; Bartlett et al. 2002; Lugosi and Wegkamp 2004; Bartlett et al. 2005; Fromont 2007, and references therein). In particular, Fromont (2007) suggests to use the bootstrap (Efron 1983) to obtain tighter penalties and provides oracle inequalities. The literature in this area has looked for improvements in penalty estimation and the selection of subclasses of functions over which to estimate the estimation error uniformly, but for technical reasons independent identically distributed (iid) random variables have been assumed. In the iid framework, powerful inequalities (e.g., McDiarmid inequality and extensions based on the martingale method) can be used to obtain uniform bounds of the estimation error and related quantities. As soon as we allow for dependence, these inequalities cannot be used and the model selection problem becomes harder both to define and to study. The goal of this paper is to provide a framework for structural risk minimization for dependent heterogeneous data sets using bootstrap penalties. An asymptotic inequality is derived to show that we can expect the bootstrap to work in this case as in the case of iid random variables. Because of the use of the bootstrap, the results of this paper can be related to the ones in Fromont (2007). In order to allow for dependence, we need to restrict attention to smooth classes of loss functions as defined in terms of an entropy integral under the uniform distance and we cannot derive so powerful results as the ones in the literature based on iid observations. Hence, this rules out the classification problem. Essentially, the class of functions allowed is the same as in Cesa-Bianchi and Lugosi (2001), where a different problem is considered. Further remarks on this can be found at the end of Sect. 2.

Some background material can be found below. Section 2 states an inequality with uniform asymptotic rates for the structural risk minimization problem using some suitable penalties. Then, it is shown that the bootstrap can be used to define these penalties. In Sect. 3, a simulation study shows that the proposed methodology works well in practice. In a variety of situations it seems to outperform other methods like Akaike information criterion and V-fold cross validation. Section 4 contains proofs of results.

1.1 Background

1.1.1 IID case

Suppose $(Z_i)_{i\in\mathbb{N}}$ is a sequence of iid random variables with values in some set $\mathcal{Z}$. Define $Z_a^b := (Z_a, \dots, Z_b)$ $(a < b, a, b \in \mathbb{N})$. Suppose that using the sample Z_1^n we want to minimize the expected risk $\mathcal{R}(Z_1^n, f) := \mathbb{E} f(Z_1)$, where $f \in \mathfrak{F}$, and $\mathfrak{F}$ is some class of loss functions. Suppose also $\mathfrak{F} := \bigcup_{k=1}^{K} \mathfrak{F}_k$. Our goal is to minimize $\mathcal{R}(Z_1^n, f)$ with respect to $f \in \mathfrak{F}_k$ and k, i.e., to identify the "right" model (i.e., $\mathfrak{F}_k$).

Example 1 Suppose $Z_i := (Y_i, X_i) \in \mathbb{R} \times \mathbb{R}^K$, and $f(Z_i) = f_\theta(Z_i) := \left(Y_i - \theta' X_i\right)^2$, where $\theta \in \mathbb{R}^K$. Minimizing $\mathcal{R}\left(Z_1^n, f\right)$, with respect to f, implies minimization with

respect to θ. From a technical point of view, what matters is the structure of f, hence, it is more convenient to see the minimization with respect to f and not θ. Allowing for some entries in θ to be zero leads to a model selection problem for regression under the square loss.

1.1.2 Non-IID case

If $(Z_i)_{i\in\mathbb{N}}$ are not iid random variables, the definition of risk as unconditional expectation of $f(Z_i)$ may not be suitable. In particular, each Z_i may take values in $\mathcal{Z}_i$, where $\mathcal{Z}_i \neq \mathcal{Z}_j$ for $i \neq j$.

Example 2 In Example 1, suppose $X_i := (Y_{i-1}, \ldots, Y_1)$ (so that θ also depends on i). Then, $\mathcal{Z}_{i-1} \subset \mathcal{Z}_i$. While f depends on i, for simplicity, this dependence will not be made explicit in the sequel.

When dealing with possibly dependent observations, the goal is to use the X_i variable as a predictor for Y_i.

Example 3 In Example 1 suppose $X_i := (Y_{i-1}, \ldots, Y_{i-K})$. If $Y_i = \theta_0' X_i + \varepsilon_i$, where $(\varepsilon_i)_{i\in\mathbb{Z}}$ are iid, then, $(f(Z_i))_{i\in\mathbb{N}}$ is not iid, unless θ is evaluated at θ_0. In this case, it is less sensible to consider full expectation, as Y_i depends on X_i which is known at time $i-1$. If X_i is a valid predictor, it must be an exogenous variable and as such the estimation problem to choose f should be formulated as minimization of the sum of the prediction errors (Seillier-Moiseiwitsch and Dawid 1993), i.e., minimize $\mathcal{R}(Z_1^n, f) := n^{-1}\sum_{i=1}^n \mathbb{E}_{i-1} f(Z_i)$, where $\mathbb{E}_{i-1}$ is expectation conditioning on the sigma algebra generated by Z_0^{i-1}. If $(Z_i)_{i\in\mathbb{N}}$ is iid, risk minimization using unconditional and conditional expectations is identical.

1.1.3 Prequential definition of risk minimization

Example 3 shows that conditional expectation rather than full expectation might be required in a time series context. Hence, the risk should be

$$\mathcal{R}\left(Z_1^n, f\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{i-1} f\left(Z_i\right). \tag{1}$$

Clearly, under suitable conditions, $\mathcal{R}(Z_1^n, f) \to \mathbb{E} f(Z_1)$ in some mode of convergence. This does not need to be always the case, especially for dependent heterogeneous data in misspecified models (e.g., Skouras and Dawid 2000, for examples). This definition of risk is in line with the prequential principle of Dawid (e.g., Dawid 1984, 1985, 1986).

Unless we know the true conditional distribution, $\mathcal{R}(Z_1^n, f)$ is unknown and, in practice, we would replace $\mathcal{R}(Z_1^n, f)$ with

$$\mathcal{R}_n\left(Z_1^n, f\right) := \frac{1}{n}\sum_{i=1}^{n} f\left(Z_i\right),$$

which is its empirical counterpart. To see that this makes sense, notice that $\mathcal{R}_n(Z_1^n, f) - \mathcal{R}(Z_1^n, f)$ is the average of martingale differences and converges to zero under regularity conditions. Clearly, $\mathcal{R}(Z_1^n, f)$ may have a limit under regularity conditions and this limit may correspond to the expectation of $\mathcal{R}(Z_1^n, f)$ with respect to the asymptotically stationary measure, when it exists (Gray and Kieffer 1980, for details). While we will not directly refer to this, we tacitly assume that the sigma algebra generated by $Z_{-\infty}^0$ is trivial with no further mention.

2 Risk minimization problem for possibly dependent heterogeneous data

Suppose $\mathfrak{F}$ can be represented as the union of the models $(\mathfrak{F}_k)_{k\in\{1,\dots,K\}}$, where K may tend to infinity with the sample size. This covers the case of estimation by the method of sieves (e.g., Bühlmann 1997, in the autoregressive case).

Suppose $\hat{f}_{n,k} \in \mathfrak{F}_k$ is a data based estimator. To provide the usual intuition for the structural minimization problem, consider the following identity:

$$\min_{k\in\{1,\dots,K\}} \mathcal{R}\left(Z_1^n, \hat{f}_{n,k}\right) - \inf_f \mathcal{R}\left(Z_1^n, f\right) = \left[\min_{k\in\{1,\dots,K\}} \mathcal{R}\left(Z_1^n, \hat{f}_{n,k}\right) - \inf_{f\in\mathfrak{F}} \mathcal{R}\left(Z_1^n, f\right)\right] + \left[\inf_{f\in\mathfrak{F}} \mathcal{R}\left(Z_1^n, f\right) - \inf_f \mathcal{R}\left(Z_1^n, f\right)\right].$$

The first term on the right is usually called the estimation error, while the second is the approximation error. The approximation error summarizes the loss incurred in restricting attention to the class $\mathfrak{F}$, where the $\inf_f$ is taken within a larger class that includes the "true model". Clearly, the larger is $\mathfrak{F}$, the smaller is the approximation error. However, a large $\mathfrak{F}$ makes the estimation problem more difficult. This resembles the usual trade off between bias and variance in the L_2 nonparametric problem.

Since $\mathcal{R}(Z_1^n, f)$ is unknown, we may use the following identity and its upperbound:

$$\begin{aligned}\min_{k\in\{1,\dots,K\}} \mathcal{R}\left(Z_1^n, \hat{f}_{n,k}\right) - \inf_f \mathcal{R}\left(Z_1^n, f\right) &= \min_{k\in\{1,\dots,K\}} \left[\mathcal{R}_n\left(Z_1^n, \hat{f}_{n,k}\right) - \inf_f \mathcal{R}\left(Z_1^n, f\right) \right.\\ &\quad \left. + \mathcal{R}\left(Z_1^n, \hat{f}_{n,k}\right) - \mathcal{R}_n\left(Z_1^n, \hat{f}_{n,k}\right)\right]\\ &\le \min_{k\in\{1,\dots,K\}} \left[\left(\mathcal{R}_n\left(Z_1^n, \hat{f}_{n,k}\right) - \inf_f \mathcal{R}\left(Z_1^n, f\right)\right)\right.\\ &\quad \left. + \sup_{f\in\mathfrak{F}_k} \left(\mathcal{R}\left(Z_1^n, f\right) - \mathcal{R}_n\left(Z_1^n, f\right)\right)\right],\end{aligned}$$

so that the second term in the last inequality is a uniform loss for the estimation error. Since we do not know $\mathcal{R}$, the above upperbound can be used if we can find a good estimate of the second term (the one in the supremum). Then, the strategy is to choose $\hat{f}_{n,k}$ such that the upperbound is minimized (note that $\inf_f \mathcal{R}(Z_1^n, f)$ is independent

of $\hat{f}_{n,k}$ and k). For the sake of clarity, but at the cost of some repetition, we introduce the following notation.

Notation 1 *Let $\mathfrak{F} = \bigcup_{k=1}^{K} \mathfrak{F}_k$. The symbol f_k defines a fixed but arbitrary element of $\mathfrak{F}_k$, and*

$$\hat{f}_{n,k} := \arg \inf_{f \in \mathfrak{F}_k} \mathcal{R}_n \left(Z_1^n, f \right)$$

is the empirical risk minimizer for model k. For some function $pen_n(\mathfrak{F}_k)$ (to be characterized in Condition 3) define

$$\hat{\mathcal{R}}_n \left(Z_1^n, f_k \right) := \mathcal{R}_n \left(Z_1^n, f_k \right) + \frac{pen_n(\mathfrak{F}_k)}{\sqrt{n}}$$

and

$$\hat{f}_{n,\hat{k}} := \arg \min_{k \in \{1,\ldots,K\}} \hat{\mathcal{R}}_n \left(Z_1^n, \hat{f}_{n,k} \right).$$

We shall also use the symbol $\lesssim$ to denote inequality up to a multiplicative finite absolute constant.

Remark 1 Note that $\mathcal{R}(Z_1^n, \hat{f}_{n,k}) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{i-1} f(Z_i)|_{f=\hat{f}_{n,k}}$, i.e., the conditional expectation is taken before evaluating f at $\hat{f}_{n,k}$.

Usually, $\hat{f}_{n,\hat{k}}$ is asymptotically consistent if $|\mathcal{R}_n(Z_1^n, f) - \mathcal{R}(Z_1^n, f)| \to 0$ in some appropriate mode of convergence uniformly over some subset of $\mathfrak{F}_k$, and $pen_n(\mathfrak{F}_k) \to 0$ as $n \to \infty$ (see van der Vaart and Wellner 2000; Skouras and Dawid 2000, for details on the nonpenalized case). To ease notation, set $\hat{f}_k := \hat{f}_{n,k}$. For a random variable X, $M(X)$ stands for the median of X, i.e., $\Pr(X < M(X)) = \Pr(X > M(X))$.

Introduce the following condition.

Condition 1 *The following holds for any $k \in \{1, \ldots, K\}$:*
(i) *If $f \in \mathfrak{F}_k$, then $\|f\|_\infty := \sup_z |f(z)| < \infty$ a.s.;*

$$\text{(ii)} \quad n^{-1} \sum_{i=1}^{n} \left[(1 - \mathbb{E}_{i-1}) f(Z_i) \right] \left[(1 - \mathbb{E}_{i-1}) g(Z_i) \right] \xrightarrow{p} \sigma(f, g), \qquad (f, g \in \mathfrak{F}_k) \tag{2}$$

where $\sigma : \mathfrak{F} \times \mathfrak{F} \to \mathbb{R}$ is some limiting function such that $\sigma(f, f) > 0$ $(\forall f \in \mathfrak{F}_k)$.

Remark 2 The dependence conditions on the data series are the ones implicit in Condition 1(ii).

The following definition is needed for the next condition.

Definition 1 The entropy number $N(s, \mathfrak{G}, d)$ is the minimal number of balls $\{g : d(f, g) \leq s\}$ of radius s required to cover the set $\mathfrak{G}$, under the distance d. The entropy integral of $\mathfrak{G}$ is defined as

$$H(\mathfrak{G}, d) := \int_0^{\text{diam}(\mathfrak{G})} \sqrt{\ln N(s, \mathfrak{G}, d)} \mathrm{d}s,$$

where $\text{diam}(\mathfrak{G}) = \sup_{f,g \in \mathfrak{G}} d(f, g)$.

Condition 2 *Define*

$$d_n(f, g) := \sqrt{n^{-1} \sum_{i=1}^{n} \sup_{z \in \mathcal{Z}_i} |f(z) - g(z)|^2}.$$

Then, the following hold:
(i)

$$d_n(f, g) \lesssim d_\infty(f, g) := \lim_{n \to \infty} d_n(f, g), \tag{3}$$

where $\lesssim$ *is inequality up to a finite absolute multiplictive constant on the right hand side;*
(ii)

$$H_{\mathfrak{F}} := \max_{k \in \{1, \ldots, K\}} \left[\|f_{0k}\|_\infty + H(\mathfrak{F}_k, d_\infty) \right] < \infty,$$

for any $f_{0k} \in \mathfrak{F}_k$ *such that* $\|f_{0k}\|_\infty \geq \inf_{f \in \mathfrak{F}_k} \|f\|_\infty$.

Remark 3 As mentioned in the previous section, we may allow $Z_i \in \mathcal{Z}_i$, $Z_j \in \mathcal{Z}_j$, with $\mathcal{Z}_i \neq \mathcal{Z}_j$ $(i \neq j)$ (Example 2). To ease notation, this is not made explicit in the notation for $f \in \mathfrak{F}$ and we also use (3) to simplify some arguments and avoid trivialities in the notation. Clearly, if $\mathcal{Z}_i = \mathcal{Z}$ for any i, then $d_\infty(f, g) = \sup_{z \in \mathcal{Z}} |f(z) - g(z)|$. Note that $\max_{k \in \{1, \ldots, K\}} H(\mathfrak{F}_k, d_\infty) \leq H(\mathfrak{F}, d_\infty)$. The odd looking condition $\|f_{0k}\|_\infty \geq \inf_{f \in \mathfrak{F}_k} \|f\|_\infty$ is used just in case the inf is not in $\mathfrak{F}_k$.

Remark 4 It is necessary to impose extra conditions to assure that quantities that are supposed to be random variables are measurable (otherwise, they fail to be random variables). Since these issues are well understood (e.g., van der Vaart and Wellner 2000) measurability conditions are overlooked and everything is assumed to be measurable with no further mention. The simplest option is to take $\mathfrak{F}$ countable.

Finally, the penalty needs to satisfy the following.

Condition 3 *Let* $(\mathbb{G}(f))_{f \in \mathfrak{F}}$ *be a mean zero Gaussian process with covariance function* $\sigma(f, g)$*, where* $\sigma(f, g)$ *is as in* (2)*. For any* $k \in \{1, \ldots, K\}$*, define*

$$pen_n(\mathfrak{F}_k) := pen_\infty(\mathfrak{F}_k) + p_n(k) \tag{4}$$

such that either
(i)

$$\mathbb{E} \sup_{f \in \mathfrak{F}_k} \mathbb{G}(f) \leq pen_\infty(\mathfrak{F}_k)$$

or
(ii)

$$M\left(\sup_{f \in \mathfrak{F}_k} \mathbb{G}(f) \right) \leq pen_\infty(\mathfrak{F}_k) \tag{5}$$

and in both cases, there exists a sequence $r_n \to 0$ *such that for any* $\tau > 0$ *with probability at least* $1 - e^{-\tau}$

$$|p_n(k)| \lesssim \sqrt{H_{\mathfrak{F}}^2 \ln(1 + r_n e^{\tau})}. \tag{6}$$

Remark 5 It is worth providing some intuition about (6). We shall show that we can find a version of $(\mathbb{G}(f))_{f \in \mathfrak{F}}$ in Condition 3 such that with probability at least $1 - e^{-\tau}$,

$$\left| \sup_{f \in \mathfrak{F}_k} \sqrt{n} \left[\mathcal{R}\left(Z_1^n, f\right) - \mathcal{R}_n\left(Z_1^n, f\right) \right] - \sup_{f \in \mathfrak{F}_k} \mathbb{G}(f) \right| \lesssim \sqrt{H_{\mathfrak{F}}^2 \ln(1 + r_n e^{\tau})}$$

(Lemma 8) so that we can replace control over

$$\sup_{f \in \mathfrak{F}_k} \sqrt{n} \left[\mathcal{R}\left(Z_1^n, f\right) - \mathcal{R}_n\left(Z_1^n, f\right) \right]$$

with control over $\sup_{f \in \mathfrak{F}_k} \mathbb{G}(f)$. Hence, any additional error incurred in the procedure shall not be larger than the error due to this approximation. This is the requirement in (6).

The estimator $\hat{f}_{n,\hat{k}}$ satisfies the following asymptotic bound.

Theorem 1 *Suppose Conditions 1, 2 and 3 are satisfied. Then, for any* $k \in \{1, \ldots, K\}$, $f_k \in \mathfrak{F}_k$, *and* $\tau > 0$, *with probability at least* $1 - e^{-\tau}$,

$$\mathcal{R}\left(Z_1^n, \hat{f}_{\hat{k}}\right) \leq \mathcal{R}\left(Z_1^n, f_k\right) + \frac{pen_\infty(\mathfrak{F}_k)}{\sqrt{n}} + 8\sqrt{\frac{2(\ln(K) + \tau)\sigma_{\mathfrak{F}}^2 + C H_{\mathfrak{F}}^2 \ln(1 + r_n K e^{\tau})}{n}},$$

for some finite absolute constant C *and some sequence* $r_n \to 0$ *both independent of* τ *and* K *and we have defined*

$$\sigma_{\mathfrak{F}}^2 := \sup_{f \in \mathfrak{F}} \sigma(f, f).$$

The above result provides a rough bound for penalized risk minimization in terms of the penalty, the maximum asymptotic variance $\sigma^2_{\mathfrak{F}}$, the number K of competing models and a term $H^2_{\mathfrak{F}} \ln(1 + r_n K e^{\tau})$ which goes to zero for any $\tau > 0$ and $K < \infty$. The complication in the proof of this result is to show that the constant C and the sequence $r_n \to 0$ can be chosen independently of $\tau > 0$ and K, hence providing a uniform rate of convergence. However, the sequence r_n does depend on $\max_k H(\mathfrak{F}_k, d_n)$, i.e., on the size of the maximal entropy integral. This dependence could be made explicit by the use of a more refined argument based on an estimate of the Prohorov distance between $\sqrt{n}\left[\mathcal{R}\left(Z_1^n, f\right) - \mathcal{R}_n\left(Z_1^n, f\right)\right]$ and the limiting Gaussian process $\mathbb{G}(f)$ (e.g., Doukhan et al. 1987). However, this would also require to impose explicit conditions on the rate of convergence in Condition 1. For the sake of simplicity as well of generality, we avoid more refined statements. The purpose of the bound is to identify the main terms that contribute to the error. If we naively choose k without penalization (i.e., $pen_n := 0$), the bound is still of root-order, but should be replaced by a bound of the following form

$$\mathcal{R}\left(Z_1^n, \hat{f}_{\hat{k}}\right) \leq \mathcal{R}\left(Z_1^n, f_k\right) + C' \sqrt{\frac{H^2_{\mathfrak{F}} \ln(2K + \tau)}{n}}$$

for some finite absolute constant C' (see Lemma 13). Since $pen_\infty(\mathfrak{F}_k)/\sqrt{n} = O\left(n^{-1/2}\right)$ uniformly in τ, and $\sigma^2_{\mathfrak{F}}$ is smaller than $H^2_{\mathfrak{F}}$ [note that $(\ln(K) + \tau) \simeq \ln(1 + 2Ke^{\tau})$ for large τ] there is an improvement as soon as we require high confidence (i.e., large τ). When the penalty needs to be estimated, the median might be preferred because of its robustness. It seems plausible that we may avoid the $\ln K$ term in the bound at the expense of a larger penalty (e.g., Bartlett et al. 2002). The derived bound reveals a fundamental weakness of penalties that try to control the fluctuations of the estimation error over the whole set $\mathfrak{F}_k$. The term $pen_n(\mathfrak{F}_k)/\sqrt{n}$ can be quite large relatively to the actual fluctuations of $\mathcal{R}(Z_1^n, \hat{f}_k) - \mathcal{R}_n(Z_1^n, \hat{f}_k)$. As noted by several authors (e.g., Bartlett et al. 2002), penalties that provide control uniformly over the whole set $\mathfrak{F}_k$ tend to perform quite poorly when the noise level is not high. For this reason, it is worthwhile to derive a uniform bound for the estimation error only over regions where the minimum is likely to be positioned. If we can obtain information on where the data driven estimator is more likely to be positioned, then we can improve on the estimation error.

Condition 4 *For any $\tau > 0$ there is a sequences of random functions $\hat{u}_{n,k} = \hat{u}_{n,k}(\tau)$, such that $\Pr(|\hat{f}_{n,k}| < \hat{u}_{n,k}) \geq 1 - e^{-\tau}$, for $k \in \{1, \ldots, K\}$.*

For convenience introduce the following notation.

Notation 2 *For any $\tau > 0$, define $\mathfrak{U}_{n,k}(\tau) := \{f \in \mathfrak{F}_k : |f_{n,k}| < \hat{u}_{n,k}\}$ and $\mathfrak{U}_n(\tau) := \bigcup_{k=1}^K \mathfrak{U}_{n,k}(\tau)$, where the argument τ stresses the fact that $\hat{u}_{n,k}$ depends on τ. Moreover,*

$$\sigma^2_{\mathfrak{U}_n(\tau)} := \sup_{f \in \mathfrak{U}_n(\tau)} \sigma(f, f), \quad H_{\mathfrak{U}_n(\tau)} := \max_{k \in \{1,\ldots,K\}} \left[\|f_{0k}\|_\infty + H\left(\mathfrak{U}_{n,k}(\tau), d_\infty\right)\right].$$

Note that using Condition 4, we still find that the bound of Theorem 1 is valid but with $\mathfrak{F}_k$ replaced by a smaller set.

Corollary 1 *Under Conditions 1, 2, 3 and 4, for any* $k \in \{1, \ldots, K\}$, $f_k \in \mathfrak{F}_k$, $\tau > \ln 2$, *with probability at least* $1 - 2e^{-\tau}$,

$$\mathcal{R}\left(Z_1^n, \hat{f}_{\hat{k}}\right) \leq \mathcal{R}\left(Z_1^n, f_k\right) + \frac{pen_\infty\left(\mathfrak{U}_{n,k}(\tau)\right)}{\sqrt{n}} + 8\sqrt{\frac{2\left(\ln(K) + \tau\right)\sigma^2_{\mathfrak{U}_n(\tau)} + C H^2_{\mathfrak{U}_n(\tau)} \ln\left(1 + r_n K e^{\tau}\right)}{n}},$$

for some finite absolute constant C *and some sequence* $r_n \to 0$ *both independent of* τ *and* K.

The improvement of the above result is that if we can identify $\hat{u}_{n,k}$, then, we can reduce considerably the size of the error both in terms of pen_∞, the maximal asymptotic variance and the entropy integral. Note that the confidence probability has decreased from $1 - e^{-\tau}$ to $1 - 2e^{-\tau}$ due to Condition 4. Equipped with these results, the goal of this paper is to obtain a data based algorithm that would allow us to satisfy Conditions 3 and 4.

2.1 Bootstrap penalty estimators

In this section, we consider a simple bootstrap empirical process. It can be used to construct a penalty that satisfies Condition 3. Suppose $\{(M_{i,b})_{i\in\mathbb{Z}}, b = 1, \ldots, B\}$ are sequences of iid bounded random variables independent of each other and of $(Z_i)_{i\in\mathbb{N}}$ with mean and variance equal to one. The variables $\{(M_{i,b})_{i\in\mathbb{Z}}, b = 1, \ldots, B\}$ might be continuous. We shall define the following wild bootstrap empirical process

$$\mathcal{R}_n^*\left(Z_1^n, f, M_{i,b}\right) := \frac{1}{n}\sum_{i=1}^n M_{i,b} f\left(Z_i\right). \tag{7}$$

This is a generalization of the wild bootstrap (as named in Mammen 1992) to the empirical risk. Then, conditioning on the sample values,

$$\mathcal{R}_n\left(Z_1^n, f\right) - \mathcal{R}_n^*\left(Z_1^n, f, M_{i,b}\right) = \frac{1}{n}\sum_{i=1}^n \left(1 - M_{i,b}\right) f\left(Z_i\right) \tag{8}$$

is the average of martingale differences like

$$\mathcal{R}\left(Z_1^n, f\right) - \mathcal{R}_n\left(Z_1^n, f\right) = \frac{1}{n}\sum_{i=1}^n \left(\mathbb{E}_{i-1} - 1\right) f\left(Z_i\right). \tag{9}$$

We define

$$pen_{n,B}\left(\mathfrak{F}_k\right) := \frac{1}{B}\sum_{b=1}^{B}\sup_{f^b\in\mathfrak{F}_k}\left(\mathcal{R}_n\left(Z_1^n, f^b\right) - \mathcal{R}_n^*\left(Z_1^n, f^b, M_{i,b}\right)\right), \tag{10}$$

and show that it can be used to satisfy Condition 3.

Theorem 2 *Define* $pen_n\left(\mathfrak{F}_k\right) := \lim_{B\to\infty} pen_{n,B}\left(\mathfrak{F}_k\right)$ *with* $pen_{n,B}\left(\mathfrak{F}_k\right)$ *in* (10). *Then, under Conditions 1 and 2,* $pen_n\left(\mathfrak{F}_k\right)$ *satisfies Condition 3.*

A more general penalty can be derived in place of (10). Suppose $\{(\pi_{i,b})_{i\in\mathbb{Z}}, b = 1, \ldots, B\}$ are sequences of iid bounded random variables independent of each other with mean zero and variance one. Define

$$pen_{n,B}\left(\mathfrak{F}_k\right) := \frac{1}{B}\sum_{b=1}^{B}\sup_{f^b\in\mathfrak{F}_k}\left(\frac{1}{n}\sum_{i=1}^{n}\pi_{i,b}f^b\left(Z_i\right)\right). \tag{11}$$

Then, we have the following generalization of Theorem 2.

Corollary 2 *Define* $pen_n\left(\mathfrak{F}_k\right) := \lim_{B\to\infty} pen_{n,B}\left(\mathfrak{F}_k\right)$ *with* $pen_{n,B}\left(\mathfrak{F}_k\right)$ *in* (11). *Then, under Conditions 1 and 2,* $pen_n\left(\mathfrak{F}_k\right)$ *satisfies Condition 3.*

Note that the penalty defined in (11) reminds quite closely Rademacker penalties which are an effective mean to upperbound an empirical process via symmetrization in the iid case. A similar idea is applied here, but asymptotically.

As mentioned previously, a penalty uniform over $\mathfrak{F}_k$ may perform poorly (cf. the bound of Theorem 1) and it is desirable to apply Corollary 1. To this end, we need an estimate of the set $\mathfrak{U}_{n,k}\left(\tau\right) := \{f \in \mathfrak{F}_k : |f_{n,k}| < \hat{u}_{n,k}\left(\tau\right)\}$ for any $\tau > 0$. Again, the bootstrap empirical process can be used. Define

$$\hat{f}_{n,k}^b := \arg\inf_{f\in\mathfrak{F}_k}\mathcal{R}_n^*\left(Z_1^n, f, M_{i,b}\right).$$

Then, set $\hat{u}_{n,k}^B = \max_{b\in\{1,\ldots,B\}}\left|\hat{f}_{n,k}^b\right| + \delta$, for some $\delta = \delta_n^B \to 0$ as either n or B go to infinity. We have the following.

Theorem 3 *For* $k \in \{1, \ldots, K\}$, *suppose* $\hat{f}_{n,k}$ *is such that, a.s.,*

$$\mathcal{R}_n\left(Z_1^n, \hat{f}_k\right) < \inf_{f\notin G_k}\mathcal{R}_n\left(Z_1^n, f\right) \tag{12}$$

for any open set $G_k \subset \mathfrak{F}_k$ *that contains* $\hat{f}_k$. *Then, for any* $\tau, \delta > 0$ *and* $n \in \mathbb{N}$, *there exists a* $B_0 = B_0\left(\tau, \delta, n\right)$ *such that for* $B \geq B_0$, *Condition 4 is satisfied with* $\hat{u}_{n,k} = \hat{u}_{n,k}^B$. *In particular,* $B_0 \to \infty$ *as either n and/or* $\tau \to \infty$.

Theorem 3 allows us to find an estimator for the size of the set over which to perform optimization of $\mathcal{R}_n$ with respect to f so that Corollary 1 applies. This leads to considerable improvement in many applications. While B_0 is unknown, Theorem 3 says that we can choose δ independently of τ and n as long as B is chosen large. Hence, for sufficiently large B, the performance based on the constrained penalty will be superior to the one based on a penalty over $\mathfrak{F}_k$ (Corollary 1 vs. Theorem 1). The condition in (12) is required for identifiably of the minimizer.

2.2 Bootstrap model selection in practice

From the previous results, the following approach for bootstrap model selection should be a good choice:
For $k = 1, \ldots, K$:
(1) Estimate $\hat{f}_{n,k}$ from the empirical risk $\mathcal{R}_n$;
(2) Use weights $\{(M_{i,b})_{i\in\mathbb{Z}}, b = 1, \ldots, B_1\}$ with mean and variance equal to one, and estimate $(\hat{f}^b_{n,k})_{b\in\{1,\ldots,B_1\}}$ from the bootstrapped process $\mathcal{R}^*_n$ in (7) and define the set

$$\mathfrak{B}^{B_1}_{n,k} := \left\{\hat{f}^b_{n,k}; b = 1, \ldots, B_1\right\};$$

(3) Use mean zero, variance one weights $\{(\pi_{i,b})_{i\in\mathbb{Z}}, b = 1, \ldots, B_2\}$ independent of the weights in (2) and estimate

$$\frac{pen_{n,B_2}\left(\mathfrak{B}^{B_1}_{n,k}\right)}{\sqrt{n}} := \frac{1}{B_2}\sum_{b=1}^{B_2}\max_{f^b\in\mathfrak{B}^{B_1}_{n,k}}\left(\frac{1}{n}\sum_{i=1}^{n}\pi_{i,b}f^b(Z_i)\right);$$

(4) Use (1) and (3) to find the penalized risk $\mathcal{R}_n(\hat{f}_{n,k}) + pen_{n,B_2}(\mathfrak{B}^{B_1}_{n,k})/\sqrt{n}$;
(5) Choose k to minimize $\mathcal{R}_n(\hat{f}_{n,k}) + pen_{n,B_2}(\mathfrak{B}^{B_1}_{n,k})/\sqrt{n}$.

The only step that requires some further comment is (3). For practical reasons, instead of using the set $\{|f| < \hat{u}^B_{n,k}\}$, the countable set $\mathfrak{B}^{B_1}_{n,k}$ is used where $\delta := 0$ for simplicity. Since $(\hat{f}^b_{n,k})_{b\in\{1,\ldots,B_1\}}$ is random we may expect $\mathfrak{B}^{B_1}_{n,k}$ to be a good approximation for $\{|f| < \hat{u}^B_{n,k}\}$ when B_1 is large.

If we use the median instead of the mean, in step (3) we should set $pen_{n,B_2}(\mathfrak{B}^{B_1}_{n,k})/\sqrt{n}$ equal to the $n/2$ order statistic of

$$\left\{\max_{f^b\in\mathfrak{B}^{B_1}_{n,k}}\left(\frac{1}{n}\sum_{i=1}^{n}\pi_{i,b}f^b(Z_i)\right), b = 1, \ldots, B_2\right\}$$

($n/2 \in \mathbb{N}$ to avoid trivialities in the notation).

2.3 A few remarks on condition 2

In their paper on regret minimization for the logarithmic loss, Cesa-Bianchi and Lugosi (2001) use the same entropy condition used here. To draw a more direct relation, suppose $f = -\ln p$, where $p \in \mathfrak{P}$, and $\mathfrak{P}$ is a class of probability density functions (with respect to some suitable dominating measure). As remarked by these authors, for most "smooth parametric" classes $\mathfrak{P}$, $\ln N\left(s, \mathfrak{P}, d_n'\right) \leq a \ln\left(bn^{1/2}/s\right)$ $(a, b > 0)$, where

$$d_n'(p, q) := \sqrt{\sum_{i=1}^{n} \sup_{z \in \mathcal{Z}_i} |\ln p(z) - \ln q(z)|^2},$$

implying that for $\mathfrak{F}$ being the class of functions $f = -\ln p$ with $p \in \mathfrak{P}$, $\ln N(s, \mathfrak{F}, d_n) \leq a \ln(b/s)$. Hence all their comments about the entropy numbers under the metric d_n' apply in this paper for the metric d_n, replacing s balls with $sn^{-1/2}$ balls, so that the reader is referred to them for a discussion and examples (their Sect. 4).

The use of the semimetric d_n may still lead to large entropy numbers ruling out some "less smooth" classes. This semimetric is used because our results are based on a combination of Azuma inequality for bounded martingales and bounds for the Orlicz norm of the empirical process. An interesting question is if using a Bernstein inequality for martingales (e.g., De la Peña 1999) and a combination of Orlicz norms (e.g., van der Vaart and Wellner 2000, Lemma 2.2.10), together with a conditioning argument, the semimetric d_n could be replaced by a weaker one which would allow us to consider less "smooth" classes of functions. This should be possible in some circumstances. If we are not interested in uniform control w.r.t. τ in the second term in the square root of Theorem 1, Condition 2 could be considerably weakened, by use of weak convergence results for families of martingales (Levental 1989).

3 Simulation study

The performance of the bootstrap model selection strategy relative to other methods may depend on the specific problem to which it is applied. For this reason, following Friedman (2001), the performance will be tested on a series of randomly generated models which can describe a large variety of continuous functions. Consider a function $F: \mathbb{R} \to \mathbb{R}$ that admits the following representation

$$\begin{aligned} F(x) &= \sum_{s=1}^{S} a_s g_s(x) \\ g_s(x) &= \exp\left\{-\frac{(x - b_s)^2}{2c_s^2}\right\}, \end{aligned} \tag{13}$$

where $a_s \in [-1, 1]$, $b_s, c_s \in \mathbb{R}$, $s = 1, \ldots, S$. For $S \to \infty$ the class of functions F (parametrized in terms of a_s, b_s, c_s, $s = 1, \ldots, S$) is dense in the class of continuous

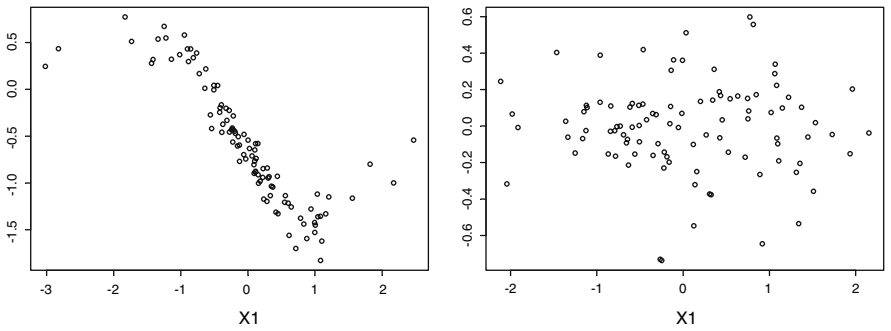

Fig. 1 Cross plot for two different data samples

bounded functions on $\mathbb{R}$ (e.g., Ripley 1996). For the simulation study, we shall consider $S = 20$, $(a_s)_{s\in\{1,\ldots,S\}}$ iid uniform in $[-1, 1]$, and $(b_s)_{s\in\{1,\ldots,S\}}$ iid normal with mean zero and variance one ($N(0, 1)$). For simplicity, $c_s = c$ $(\forall s)$ is also $N(0, 1)$. The scaling parameters c_s are set all equal in order to avoid particularly irregular functions that might be very uncommon in any practical application. One hundred functions are simulated using this approach and data $(Z_i^{(r)})_{i\in\mathbb{N}} = (Y_i^{(r)}, X_i^{(r)})_{i\in\mathbb{N}}$ $(r = 1, \ldots, 100)$ are simulated adding correlated noise:

$$Y_i^{(r)} = F^{(r)}\left(X_i^{(r)}\right) + U_i^{(r)}$$
$$U_i^{(r)} = 0.8U_{i-1}^{(r)} + \varepsilon_i^{(r)},$$

where, for each r, $(X_i^{(r)})_{i\in\mathbb{N}}$ and $(\varepsilon_i^{(r)})_{i\in\mathbb{N}}$ are, respectively, sequences of iid $N(0, 1)$ and $N(0, \sigma^2)$. Hence, each r corresponds to a simulated function $F^{(r)}$ which is identified by the parameters $a_s, b_s, c_s = c$ in (13). For each of these functions, results are tested on $\sigma = 0.05, 0.1, 0.2$ and sample sizes $n = 50, 100, 200, 400, 800$. In this case, the signal to noise ratio is quite high for all σ's and r's. However, this is necessary because of the highly nonlinear structure of the target functions $F^{(r)}$. Figure 1 gives the cross plot of $(Y_i^{(r)}, X_i^{(r)})_{1\le i\le 100}$ for two different $F^{(r)}$'s when $\sigma = 0.1$, representing two opposite extreme cases. Despite the high signal to noise ratio, the second panel displays a high degree of noise/randomness. Moreover, as mentioned previously, penalties that provide uniform control of the estimation error tend to perform better (relative to other methods) when the noise level is high. Hence, comparison with other methods is of more interest in a framework with a lower level of noise.

Each $F^{(r)}$ is approximated by a $k = 0, \ldots, K$ order polynomial $P_k(x) = \sum_{l=0}^{k} x^l \beta_l$, where the β_l's coefficients are estimated by least square, so that the empirical risk is $\mathcal{R}_n(f_k) := n^{-1}\sum_{i=1}^{n} |Y_i - P_k(X_i)|^2$ for the square loss $f_k(y, x) = |y - P_k(x)|^2$. For each r, the estimated loss $\hat{f}_k(y, x) = |y - \hat{P}_k(x)|^2$ is then used to compute the prediction error on a validation sample $(Y_i^{(r)}, X_i^{(r)})_{n<i\le n+10{,}000}$ of 10,000 observations. Then, the average and median prediction error over $r = 1, \ldots, 100$ is computed. For some r's, the prediction error tends to be quite large and also reporting the median prediction error might be informative. Results for the bootstrap penalties

Table 1 Prediction error

n	A/C Median	Mean	V-CV Median	Mean	BMean Median	Mean	BMedian Median	Mean
Sigma=0.05								
50	0.18	1.58	0.11	1.08	0.23	0.37	0.19	0.32
100	0.08	0.46	0.09	0.31	0.13	0.27	0.12	0.24
200	0.08	0.23	0.08	0.22	0.09	0.19	0.08	0.24
400	0.04	0.14	0.04	0.15	0.06	0.16	0.05	0.15
800	0.04	0.11	0.04	0.13	0.06	0.13	0.04	0.11
Sigma=0.1								
50	0.24	1.75	0.14	1.20	0.25	0.40	0.21	0.35
100	0.12	0.52	0.11	0.32	0.16	0.29	0.14	0.26
200	0.10	0.26	0.09	0.24	0.12	0.24	0.10	0.26
400	0.06	0.16	0.07	0.17	0.08	0.18	0.08	0.17
800	0.06	0.13	0.06	0.15	0.08	0.15	0.06	0.13
Sigma=0.2								
50	0.37	2.30	0.26	1.56	0.35	0.50	0.31	0.45
100	0.23	0.72	0.21	0.53	0.25	0.45	0.24	0.36
200	0.19	0.35	0.19	0.33	0.21	0.30	0.20	0.33
400	0.15	0.25	0.15	0.26	0.17	0.27	0.16	0.26
800	0.14	0.22	0.15	0.24	0.16	0.24	0.15	0.22

based on mean and median (BMean and BMedian) are in Table 1, and are compared to competitors based on Akaike's information criterion (AIC) and V-fold cross validation (V-CV). The bootstrap penalties are computed according to the procedure described in Sect. 2.2 with $\left(M_{i,b}\right)_{i\leq n}$ Poisson with mean 1 and $\left(\pi_{i,b}\right)_{i\leq n}$ standard Gaussian, and $B_1 = B_2 = 100$. The penalized risk for AIC is given by $\mathcal{R}_n\left(f_k\right)(1+2k/n)$. For V-CV, the sample is randomly partitioned into $V = 10$ validation samples. For each validation sample, estimation is carried out using the remaining observations and the prediction error is estimated using the validation sample (e.g., van der Laan and Dudoit 2003).

Note that for this problem, the conditional risk converges to the unconditional risk (when we divide by n) because of the weak dependence of the simulated data. Given that model selection procedures are usually studied in terms of unconditional risk, it makes sense to compare methods within this more restrictive framework.

It is clear that the conditions used in the theoretical analysis are not satisfied by these simulations: (1) the loss function is not bounded, (2) the bootstrap weights are not bounded. Despite being unbounded, these quantities have thin tails and we could easily truncate in the theoretical derivations (but did not do so for the sake of simplicity). Hence, it is of additional practical interest to test the procedure allowing for unbounded quantities.

The results show that penalized bootstrap model selection and in particular BMedian should be favored when we compare in terms of mean prediction error over each simulated sample r. This is particularly so for small sample size n. When the sample

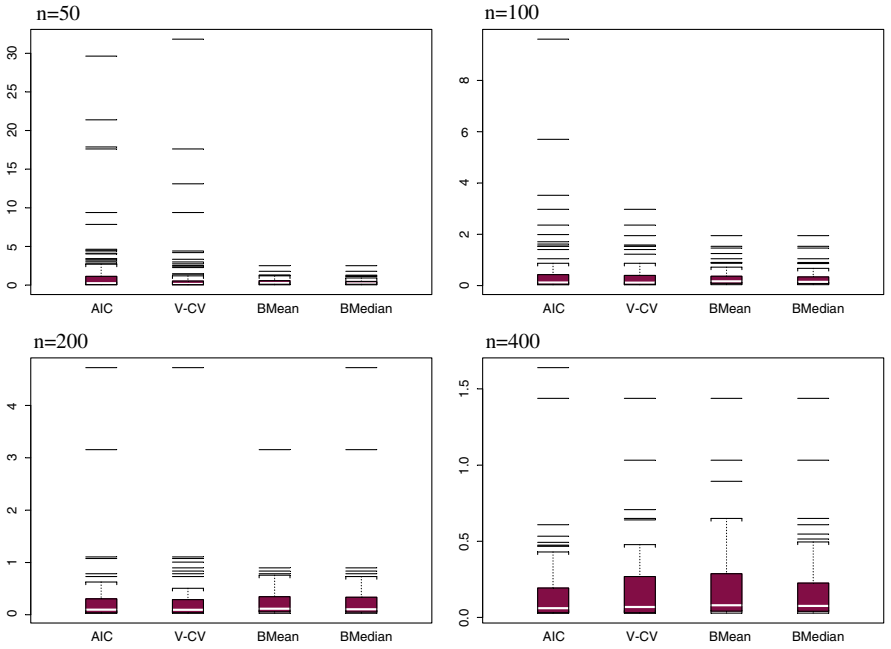

Fig. 2 Boxplot of prediction error for $\sigma = 0.1$

size increases, the performance remains comparable to the one of the other methods. The differences between the mean and median prediction error over r is due to the difficulty of selecting a good model for some of the r's. Some target functions $F^{(r)}$'s are very challenging to estimate and approximate. While AIC and V-CV often perform quite well, sometimes they select a model that leads to huge prediction error when n is small. Figure 2 shows the boxplot for the prediction error when $\sigma = 0.1$ and $n = 50, 100, 200, 400$ ($n = 800$ is not reported for economy of space). Recall that these boxplots are constructed using the prediction error of $R = 100$ simulated samples each based on a different target function $F^{(r)}$. The boxplot clearly shows that the bootstrap penalties do a relatively good job for small n. For small n, AIC and V-CV tend to be less stable producing the many outliers shown in the boxplot. When n increases, BMean and BMedian still perform comparably well with respect to AIC and V-CV. These results confirm the ones in Table 1. For higher noise levels (e.g., $\sigma = 0.2$, not reported in Fig. 2) the relative performance of the bootstrap penalties improves.

4 Technical details and proofs

Notation 3 *The following notation will be used:* $Y_n(f) := \sqrt{n}[\mathcal{R}(Z_1^n, f) - \mathcal{R}_n(Z_1^n, f)]$ *and* $X_{n,b}(f) := n^{-1/2} \sum_{i=1}^n \pi_{i,b} f(Z_i)$ *where* $\{(\pi_i)_{i\in\mathbb{Z}}, b = 1, \ldots, B\}$ *is as in the previous section. The symbol* $\xrightarrow{w}$ *stands for weak convergence and* $\stackrel{d}{=}$ *for equality*

in distribution. Recall that for any two sequences a_n and b_n, $a_n \lesssim b_n$ means that there is a finite absolute constant C such that $a_n \leq Cb_n$. Reference to van der Vaart and Wellner (2000) will be abbreviated to VW00.

The proof of Theorem 1 is based on the following steps: replace control over $Y_n(f)$ with control over a Gaussian process plus a term shown to be small (recall Remark 5), control a centered version of the supremum of the Gaussian process by standard inequalities. In particular, all the terms involved in these approximations are given in Lemma 1 below, and their control proves Theorem 1. Proof of the other results follow at the end. For the sake of clarity, from time to time, reference will be made to four simple supplementary lemmata stated at the very end of this section. Since, we are considering two kinds of penalties (based on mean and median), proofs will deal with the penalty based on the mean first, without necessarily mentioning it.

4.1 Upperbound for conditional risk

The following upperbound is the starting point for the proof of Theorem 1.

Lemma 1 *Suppose Condition 3*(i) *holds. Then,*

$$\mathcal{R}\left(Z_1^n, \hat{f}_{\hat{k}}\right) \leq \mathcal{R}\left(Z_1^n, f_k\right) + \frac{pen_\infty(\mathfrak{F}_k)}{\sqrt{n}} + \max_{k\in\{1,\ldots,K\}}\left[\frac{(1-\mathbb{E})\sup_{f\in\mathfrak{F}_k}\mathbb{G}(f)}{\sqrt{n}}\right] + \max_{k\in\{1,\ldots,K\}}\left[\frac{\left(\sup_{f\in\mathfrak{F}_k}Y_n(f) - \sup_{f\in\mathfrak{F}_k}\mathbb{G}(f)\right) - p_n(k)}{\sqrt{n}}\right] + \frac{p_n(k) - Y_n(f_k)}{\sqrt{n}}.$$

If Condition 3(ii) *holds, the third term in the above display holds with*

$$(1-\mathbb{E})\sup_{f\in\mathfrak{F}_k}\mathbb{G}(f)$$

replaced by

$$\sup_{f\in\mathfrak{F}_k}\mathbb{G}(f) - M\left(\sup_{f\in\mathfrak{F}_k}\mathbb{G}(f)\right).$$

Proof Start with the following identity

$$\begin{aligned}\mathcal{R}\left(Z_1^n, \hat{f}_{\hat{k}}\right) &= \mathcal{R}\left(Z_1^n, f_k\right) + \left[\mathcal{R}\left(Z_1^n, \hat{f}_{\hat{k}}\right) - \hat{\mathcal{R}}_n\left(Z_1^n, \hat{f}_{\hat{k}}\right)\right] \\ &\quad + \left[\hat{\mathcal{R}}_n\left(Z_1^n, \hat{f}_{\hat{k}}\right) - \mathcal{R}\left(Z_1^n, f_k\right)\right] \\ &= \mathrm{I} + \mathrm{II} + \mathrm{III}.\end{aligned}$$

We shall deal with II and III separately.

Control over II.

$$\begin{aligned}
\text{II} &= \left[\mathcal{R}\left(Z_1^n, \hat{f}_{\hat{k}}\right) - \mathcal{R}_n\left(Z_1^n, \hat{f}_{\hat{k}}\right) - \frac{pen_n\left(\mathfrak{F}_{\hat{k}}\right)}{\sqrt{n}}\right] \\
&\quad \text{[by definition of } \hat{\mathcal{R}}_n\text{]} \\
&\leq \left\{\mathcal{R}\left(Z_1^n, \hat{f}_{\hat{k}}\right) - \mathcal{R}_n\left(Z_1^n, \hat{f}_{\hat{k}}\right) - \frac{\mathbb{E}\sup_{f\in\mathfrak{F}_{\hat{k}}}\mathbb{G}(f)}{\sqrt{n}} - \frac{p_n(k)}{\sqrt{n}}\right\} \\
&\quad \text{[by Condition 3(i)]} \\
&\leq \max_{k\in\{1,\ldots,K\}}\left[\sup_{f\in\mathfrak{F}_k}\left(\mathcal{R}\left(Z_1^n, f\right) - \mathcal{R}_n\left(Z_1^n, f\right)\right) - \frac{\mathbb{E}\sup_{f\in\mathfrak{F}_k}\mathbb{G}(f)}{\sqrt{n}} - \frac{p_n(k)}{\sqrt{n}}\right]
\end{aligned}$$

by a uniform bound over k and then over f.
Control over III.

$$\begin{aligned}
\text{III} &= \left[\mathcal{R}_n\left(Z_1^n, \hat{f}_{\hat{k}}\right) + pen_n\left(\mathfrak{F}_{\hat{k}}\right) - \mathcal{R}\left(Z_1^n, f_k\right)\right] \\
&\quad \text{[by definition of } \hat{\mathcal{R}}_n\text{]} \\
&\leq \left[\mathcal{R}_n\left(Z_1^n, f_k\right) - \mathcal{R}\left(Z_1^n, f_k\right)\right] + \frac{pen_n\left(\mathfrak{F}_k\right)}{\sqrt{n}},
\end{aligned}$$

because $\hat{f}_{\hat{k}}$ is the minimizer of $\hat{\mathcal{R}}_n$. Using the definition of $Y_n(f)$ and (4), the result follows once we add and subtract $\sup_{f\in\mathfrak{F}_k}\mathbb{G}(f)/\sqrt{n}$ and use the fact that the maximum of a sum is bounded above by the sum of the maxima. The proof when (ii) in Condition 3 holds is identical. □

4.2 Uniform bound for $Y_n(f)$

A uniform bound for $Y_n(f)$ is found by Gaussian approximation.

4.2.1 Gaussian approximation

To show a Gaussian approximation, finite dimensional (fidi) convergence and stochastic equicontinuity are shown. Together they imply weak convergence to a Gaussian process.

Lemma 2 (Fidi convergence) *Suppose* $\bar{\mathfrak{F}}(\subset\mathfrak{F})$ *is a finite set. Under Condition 1,* $(Y_n(f))_{f\in\bar{\mathfrak{F}}} \xrightarrow{w} (\mathbb{G}(f))_{f\in\bar{\mathfrak{F}}}$, *where* $(\mathbb{G}(f))_{f\in\bar{\mathfrak{F}}}$ *is a vector of mean zero Gaussian random variables with covariance matrix* $(\sigma(f,g))_{f,g\in\bar{\mathfrak{F}}}$.

Proof Condition 1 satisfies the conditions of Theorem 2.3 in McLeish (1974) which implies that $Y_n(f) \to \mathbb{G}(f)$, weakly for any fixed f, where $\mathbb{G}(f)$ is a $(0, \sigma(f,f))$ Gaussian random variable. By the Cramér Wold device fidi convergence follows. □

To show stochastic equicontinuity, we shall control the oscillations of $Y_n(f)$ in terms of the Orlicz norm defined next.

Definition 2 For a random variable R, its $\psi(x) := e^{|x|^2} - 1$ Orlicz norm is defined as

$$\|R\|_\psi := \inf\left\{C > 0 : \mathbb{E}\psi\left(\frac{R}{C}\right) \leq 1\right\}.$$

For the sake of clarity, we recall the statement of a set of inequalities that shall be used momentarily. At first, we recall Azuma's inequality (e.g., Devroye et al. 1996).

Lemma 3 (Azuma inequality) *Suppose $(R_n)_{n\in\mathbb{N}}$ is a martingale sequence such that $|R_n - R_{n-1}| \leq c_n$ a.s. for any $n > 0$ and $R_0 = 0$. Then,*

$$\Pr(|R_n| > x) \leq 2\exp\left\{-\frac{x^2}{2\sum_{i=1}^n c_i^2}\right\}.$$

Azuma inequality can be used to verify the condition of the following lemma that relates the tails of a random variable to its Orlicz norm (Lemma 2.2.1 in VW00).

Lemma 4 *Suppose R is a random variable such that, for some finite absolute constants a and C,*

$$\Pr(|R| > x) \leq a\exp\left\{-Cx^2\right\}.$$

Then, $\|R\|_\psi \leq [(1+a)/C]^{1/2}$.

Finally, one uses a bound for the Orlicz norm of the oscillations of a stochastic process to derive an entropy condition (Corollary 2.2.5 in VW00).

Lemma 5 *Suppose that $(R(f))_{f\in\mathfrak{G}}$ is a stochastic process and $(\mathfrak{G}, d)$ an arbitrary semimetric space. If*

$$\|R(f) - R(g)\|_\psi \lesssim d(f,g)$$

then,

$$\left\|\sup_{f,g\in\mathfrak{G}} |R(f) - R(g)|\right\|_\psi \lesssim H(\mathfrak{G}, d)$$

where $H(\mathfrak{G}, d)$ is the entropy integral in Definition 1.

Putting the above ingredients together, we can prove the following result.

Lemma 6 (Orlicz Norm) *For d_n as in Condition 2,*

$$\|Y_n(f) - Y_n(g)\|_\psi \lesssim d_n(f,g),$$

which, for any $\mathfrak{F}_k$, implies

$$\left\| \sup_{f,g\in\mathfrak{F}_k} |Y_n(f) - Y_n(g)| \right\|_\psi \lesssim H(\mathfrak{F}_k, d_n). \tag{14}$$

Proof For any fixed f, $\sqrt{n}Y_n(f)$ is the sum of martingale differences. Hence, $\sqrt{n}(Y_n(f) - Y_n(g))$ is also a sum of martingale differences

$$\sqrt{n}[Y_n(f) - Y_n(g)] = \sum_{i=1}^n (1 - \mathbb{E}_{i-1})(f(Z_i) - g(Z_i))$$

where

$$|(1 - \mathbb{E}_{i-1})(f(Z_i) - g(Z_i))| \le 2 \sup_{z\in\mathcal{Z}_i} |f(z) - g(z)|.$$

Then, Lemma 3 gives

$$\begin{aligned}\Pr(|Y_n(f) - Y_n(g)| \ge x) &= \Pr\left(\sqrt{n}|Y_n(f) - Y_n(g)| \ge x\sqrt{n}\right)\\ &\le 2\exp\left\{-\frac{nx^2}{8\sum_{i=1}^n \sup_{z\in\mathcal{Z}_i}|f(z) - g(z)|^2}\right\}\\ &= 2\exp\left\{-\frac{x^2}{8d_n(f,g)^2}\right\}.\end{aligned}$$

Lemma 4 and the last display imply $\|Y_n(f) - Y_n(g)\|_\psi \lesssim d_n(f,g)$. This inequality and Lemma 5 give the result. □

Remark 6 Lemma 6 implies that $(Y_n(f))_{f\in\mathfrak{F}}$ is stochastically equicontinuous. In fact, define the set $\mathfrak{F}_{\delta,n} := \{f, g \in \mathfrak{F} : d_n(f,g) \le \delta\}$, for any $\delta > 0$. Then,

$$\left\| \sup_{f,g\in\mathfrak{F}_{\delta,n}} |Y_n(f) - Y_n(g)| \right\|_\psi \lesssim H\left(\mathfrak{F}_{\delta,n}, d\right) \le \int_0^\delta \sqrt{\ln N(s,\mathfrak{F},d_n)}\,ds.$$

Another implication is that

$$\left\| \sup_{f\in\mathfrak{F}_k} |Y_n(f)| \right\|_\psi \le \|Y_n(f_{0k})\|_\psi + CH(\mathfrak{F}_k, d_n) \tag{15}$$

for any $f_{0k} \in \mathfrak{F}_k$ and some finite absolute constant C independent of $\mathfrak{F}_k$ (VW00, p.100). Then, using Lemma 3 (with $c_n = \|f\|_\infty$) an application of Lemma 4 gives $\|Y_n(f_{0k})\|_\psi \lesssim \|f_{0k}\|_\infty$. Inserting this last relation in (15) together with Condition 2(i) gives

$$\left\| \sup_{f \in \mathfrak{F}_k} |Y_n(f)| \right\|_\psi \lesssim \left[\|f_{0k}\|_\infty + H(\mathfrak{F}_k, d_\infty) \right]. \tag{16}$$

Hence, Condition 2 is used to control this Orlicz norm once we take max over k.

Weak convergence easily follows.

Lemma 7 (weak convergence) *Under Conditions 1 and 2,*

$$(Y_n(f))_{f \in \mathfrak{F}} \xrightarrow{w} (\mathbb{G}(f))_{f \in \mathfrak{F}},$$

where $(\mathbb{G}(f))_{f \in \mathfrak{F}}$ *is a mean zero Gaussian process with covariance function* $\sigma(f, g)$.

Proof Fidi convergence to a Gaussian random process, stochastic equicontinuity and total boundedness imply weak convergence of the process (e.g., Example 1.5.10 in VW00). Hence Lemmas 2 and 6 together with $\operatorname{diam}(\mathfrak{F}_k) < \infty$ [by Condition 1(i)] and $K < \infty$ give the result. □

Lemma 7 is used to prove a uniform bound by Gaussian approximation using Borell inequality.

4.2.2 Approximation by Borell inequality

The following approximation is crucial.

Lemma 8 *For any* $k \in \{1, \ldots, K\}$, *under Conditions 1 and 2, there exist Gaussian processes* $(\mathbb{G}'_n(f))_{f \in \mathfrak{F}_k} \overset{d}{=} (\mathbb{G}(f))_{f \in \mathfrak{F}_k}$ *and a sequence* $r_n \to 0$ *such that for any* $\epsilon \in (0, 1)$, *with probability at least* $1 - \epsilon$,

$$\left| \sup_{f \in \mathfrak{F}_k} Y_n(f) - \sup_{f \in \mathfrak{F}_k} \mathbb{G}'(f) \right| \lesssim \sqrt{H^2_{\mathfrak{F}} \ln\left(1 + \frac{r_n}{\epsilon}\right)}$$

where $H^2_{\mathfrak{F}}$ *is the maximum entropy integral in Condition 2.*

Proof Set $W_n = W_{n,k} := \sup_{f \in \mathfrak{F}_k} Y_n(f)$ and $W = W_k := \sup_{f \in \mathfrak{F}_k} \mathbb{G}(f)$ and write F_n and F for their distribution functions. Lemma 7 and the continuous mapping theorem (VW00, Theorem 1.3.6) imply $F_n(x) \to F(x)$ as $n \to \infty$ [weak convergence where $F(x)$ is continuous]. Using weak convergence, we construct a sequence of random variables $(W_n)_{n \in \mathbb{N}}$ distributed as W and such that $|W_n - W'_n| \to 0$ in probability. Redefine $(W_n)_{n \in \mathbb{N}}$ on a common probability space by enlarging the original

probability space so that there exists a sequence $(V_n)_{n\in\mathbb{N}}$ of iid uniform [0, 1] random variables independent of $(W_n)_{n\in\mathbb{N}}$ and W. Define

$$\tilde{F}(x, v) := \Pr(W_n < x) + v \Pr(W_n = x),$$

so that $U_n := \tilde{F}(W_n, V_n)$ is a [0, 1] uniform random variable and $F_n^{-1}(U_n) \overset{a.s.}{=} W_n$ (where $F_n^{-1}(u) := \inf(x : \Pr(W_n \le x) \ge u)$) (Rüschendorf and de Valk 1993, Proposition 1). We shall show that $|W_n' - W_n| \xrightarrow{p} 0$ where $W_n' := F^{-1}(U_n)$, and it is obvious that W_n' is distributed as W for any n. To this end,

$$\begin{aligned}\mathbb{E}\left|W_n - W_n'\right| &= \int |F_n(x) - F(x)|\,\mathrm{d}x \\ &\quad \text{[Dudley 2002, Problem 2, p. 425]} \\ &\to 0, \end{aligned} \tag{17}$$

because if W_n has an $r > 1$ absolute moment, then $F_n(x) \to F(x)$ implies the convergence of the above integral (Petrov 1995, Theorem 1.12). Note that W_n has moments of all orders by Lemma 6 so that the above convergence does indeed hold. The first display in the statement of the lemma is proved if we show that with probability at least $1 - \epsilon$,

$$\left|W_n - W_n'\right| \lesssim \sqrt{\frac{1}{t}\ln\left(1 + \frac{r_n}{\epsilon}\right)} \tag{18}$$

for some $t \lesssim H_{\mathfrak{F}}^{-2}$. By Markov inequality for some $t \lesssim H_{\mathfrak{F}}^{-2}$

$$\Pr(W_n > x) \le \frac{\mathbb{E}\exp\left\{tW_n^2\right\}}{\exp\left\{tx^2\right\}} \lesssim \exp\left\{-tx^2\right\} \tag{19}$$

using Lemma 15 with $\|W_n\|_\psi \lesssim H_{\mathfrak{F}}$ by (16). Moreover, for some $t \lesssim \left[\mathbb{E}\sup_{f\in\mathfrak{F}_k} |\mathbb{G}(f)|\right]^{-2}$,

$$\Pr\left(W_n' > x\right) \le \frac{\mathbb{E}\exp\left\{tW_n^2\right\}}{\exp\left\{tx^2\right\}} \lesssim \exp\{-tx^2\}$$

by Lemma 15 with $\|W_n'\|_\psi \lesssim \mathbb{E}\sup_{f\in\mathfrak{F}_k} |\mathbb{G}(f)|$ by (23), in Lemma 9 below, and Lemma 4. Hence, by the exponential bounds in the last two displays and (17), we can apply Lemma 14 implying (18) with $t \lesssim H_{\mathfrak{F}}^{-2} \lesssim \left(H_{\mathfrak{F}}^{-2} \wedge \left[\mathbb{E}\sup_{f\in\mathfrak{F}_k} |\mathbb{G}(f)|\right]^{-2}\right)$. Hence, we only need to show that $H_{\mathfrak{F}}^{-2} \lesssim \left(H_{\mathfrak{F}}^{-2} \wedge \left[\mathbb{E}\sup_{f\in\mathfrak{F}_k} |\mathbb{G}(f)|\right]^{-2}\right)$, which requires a bound for $\mathbb{E}\sup_{f\in\mathfrak{F}_k} |\mathbb{G}(f)|$. To this end, note that we can apply the same argument used to bound $Y_n(f)$ also to bound $\mathbb{G}(f)$. We just need to apply Lemma 5 to $\mathbb{G}(f)$. Continuity of $\mathbb{G}(f) - \mathbb{G}(g)$ under the ψ Orlicz norm is found by an application

of the sub-Gaussian inequality of Lemma 6. Note that

$$\begin{aligned}\rho(f,g)^2 :&= \lim_n \frac{1}{n}\sum_{i=1}^n \left[(1-\mathbb{E}_{i-1})\, f(Z_i) - (1-\mathbb{E}_{i-1})\, g(Z_i)\right]^2 \\ &= [\sigma(f,f) + \sigma(g,g) - 2\sigma(f,g)] \end{aligned} \tag{20}$$

by Condition 1(ii). Hence, by Gaussianity,

$$\Pr\left(|\mathbb{G}(f) - \mathbb{G}(g)| > x\right) < \exp\left\{-\frac{x^2}{2\rho(f,g)^2}\right\}. \tag{21}$$

By (20) and Condition 1(i), we also have convergence of the expectation:

$$\begin{aligned}&\lim_n \frac{1}{n}\sum_{i=1}^n \left[(1-\mathbb{E}_{i-1})\, f(Z_i) - (1-\mathbb{E}_{i-1})\, g(Z_i)\right]^2 \\ &= \mathbb{E}\liminf \frac{1}{n}\sum_{i=1}^n \left[(1-\mathbb{E}_{i-1})\, f(Z_i) - (1-\mathbb{E}_{i-1})\, g(Z_i)\right]^2 \\ &\le \liminf \mathbb{E}\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{i-1}\left[(1-\mathbb{E}_{i-1})\, f(Z_i) - (1-\mathbb{E}_{i-1})\, g(Z_i)\right]^2 \end{aligned} \tag{22}$$

by Fatou lemma and the tower law of conditional expectations. It is then easy to see that the above display implies $\rho(f,g) \le \lim_n d_n(f,g)$ and that this last relation together with (21) and Lemma 5 gives $\mathbb{E}\sup_{f\in\mathfrak{F}_k} |\mathbb{G}(f)| \le H_{\mathfrak{F}}$ [see (16)]. □

Recall Borell inequality for Gaussian processes (VW00, Proposition A.2.1).

Lemma 9 *Suppose* $(\mathbb{G}(f))_{f\in\mathfrak{F}}$ *is a separable mean zero Gaussian process with* $\mathbb{E}\sup_{f\in\mathfrak{F}} \mathbb{G}(f) < \infty$. *Define* $\sigma^2_{\mathfrak{F}} := \sup_{f\in\mathfrak{F}} Var(\mathbb{G}(f))$. *For any* $x > 0$,

$$\begin{aligned}\Pr\left((1-\mathbb{E})\sup_{f\in\mathfrak{F}} \mathbb{G}(f) > x\right) &\le \exp\left\{-\frac{x^2}{2\sigma^2_{\mathfrak{F}}}\right\} \\ \Pr\left(\sup_{f\in\mathfrak{F}} \mathbb{G}(f) - M\left(\sup_{f\in\mathfrak{F}_k} \mathbb{G}(f)\right) > x\right) &\le \frac{1}{2}\exp\left\{-\frac{x^2}{2\sigma^2_{\mathfrak{F}}}\right\} \\ \Pr\left(\sup_{f\in\mathfrak{F}} \mathbb{G}(f) > x\right) &\le \exp\left\{-\frac{x^2}{8\left[\mathbb{E}\sup_{f\in\mathfrak{F}} |\mathbb{G}(f)|\right]^2}\right\}. \end{aligned} \tag{23}$$

4.3 Proof of Theorem 1

Proof of Theorem 1 By Lemma 1 it is sufficient to bound

$$\mathrm{I} := \max_{k\in\{1,\ldots,K\}}\left(\sup_{f\in\mathfrak{F}_k} Y_n(f) - \sup_{f\in\mathfrak{F}_k}\mathbb{G}(f)\right), \mathrm{II} := p_n(k), \mathrm{III} := \max_{k\in\{1,\ldots,K\}} -p_n(k),$$
$$\mathrm{IV} := \max_{k\in\{1,\ldots,K\}}(1-\mathbb{E})\sup_{f\in\mathfrak{F}_k}\mathbb{G}(f), \mathrm{V} := -Y_n(f_k),$$

where in the case of the median, IV is changed accordingly. We shall deal with each term separately. To avoid trivialities in the notation, $r_n \to 0$ is a sequence that may change in the control of each term. Similarly, C, C' are finite absolute constants that my change from line to line. By Lemmas 8 and 17,

$$\mathrm{I} \lesssim \sqrt{H_{\mathfrak{F}}^2 \ln\left(1 + r_n\frac{K}{\epsilon}\right)},$$

with probability at least $1-\epsilon$. By (6) in Condition 3, with probability at least $1-\epsilon$, $\mathrm{II} \lesssim \sqrt{H_{\mathfrak{F}}^2 \ln(1+r_n/\epsilon)}$; hence by Lemma 17,

$$\mathrm{III} \lesssim \sqrt{\ln\left(1 + r_n\frac{K}{\epsilon}\right)}$$

with probability at least $1-\epsilon$. By Lemmas 9 and 17, with probability at least $1-\epsilon$,

$$\mathrm{IV} \leq \sqrt{2\sigma_{\mathfrak{F}}^2 \ln\left(\frac{K}{\epsilon}\right)}.$$

Finally, rewrite

$$\mathrm{V} = [\mathbb{G}(f_k) - Y_n(f_k)] - \mathbb{G}(f_k)$$

so that it is not difficult to deduce that we can bound the first term in the above display with the upperbound for I and the second term with the upperbound for IV (note that the results for I and IV hold for -I and -IV as well). Hence deduce the crude upperbound

$$\mathrm{V} \leq \sqrt{2\sigma_{\mathfrak{F}}^2 \ln\left(\frac{K}{\epsilon}\right)} + \sqrt{C H_{\mathfrak{F}}^2 \ln\left(1 + r_n\frac{K}{\epsilon}\right)}$$

with probability at least $1-2\epsilon$. By Lemma 16, the bounds for I–V imply, with probability at least $(1-6\epsilon)$,

$$\begin{aligned}
\mathcal{R}\left(Z_1^n, \hat{f}_{\hat{k}}\right) &\leq \mathcal{R}\left(Z_1^n, f_k\right) + \frac{pen_\infty(\mathfrak{F}_k)}{\sqrt{n}} \\
&\quad + 2\sqrt{\frac{4\sigma_{\mathfrak{F}}^2 \ln(K/\epsilon) + CH_{\mathfrak{F}}^2 \ln(1+r_n K/\epsilon)}{n}} + C'\sqrt{\frac{\ln(1+r_n K/\epsilon)}{n}} \\
&\quad \text{[absorbing I and IV (and V) together]} \\
&= \mathcal{R}\left(Z_1^n, f_k\right) + \frac{pen_\infty(\mathfrak{F}_k)}{\sqrt{n}} \\
&\quad + 2\sqrt{\frac{8\sigma_{\mathfrak{F}}^2 \ln(K/\epsilon) + CH_{\mathfrak{F}}^2 \ln(1+r_n K/\epsilon)}{n}} \\
&\quad \text{[absorbing the fourth term into the third]} \\
&\leq \mathcal{R}\left(Z_1^n, f_k\right) + \frac{pen_\infty(\mathfrak{F}_k)}{\sqrt{n}} \\
&\quad + 2\sqrt{\frac{8\sigma_{\mathfrak{F}}^2 (\ln(6K) + \tau) + CH_{\mathfrak{F}}^2 \ln(1+r_n 6K e^{\tau})}{n}} \\
&\quad \text{[equating } (1-6\epsilon) \text{ to } 1-e^{-\tau}\text{, solving for } \epsilon, \\
&\quad \text{and substituting in } \ln(1/\epsilon)\text{]} \\
&\leq \mathcal{R}\left(Z_1^n, f_k\right) + \frac{pen_\infty(\mathfrak{F}_k)}{\sqrt{n}} \\
&\quad + 2\sqrt{\frac{32\sigma_{\mathfrak{F}}^2 (\ln(K) + \tau) + CH_{\mathfrak{F}}^2 \ln(1+r_n K e^{\tau})}{n}}
\end{aligned}$$

where the last step follows by some further bounding because $K \geq 2$ implying that $\ln(6K) \leq 4\ln(K)$. Moreover, we absorbed the constant 6 into the sequence r_n. In the case of the median, mutatis mutandis, IV is controlled using the bound for the median in Lemma 9 and we get the same result (actually with a slightly smaller constant). □

4.4 Proof of other results

Corollary 1 is proved next.

Proof of Corollary 1 Recall $\mathfrak{U}_{n,k} := \left\{f \in \mathfrak{F}_k : |f| < \hat{u}_{n,k}\right\}$. Then,

$$\begin{aligned}
Y_n\left(\hat{f}_k\right) &= \inf_{g_k \in \mathfrak{U}_{n,k}} \left[Y_n(g_k) + Y_n\left(\hat{f}_k\right) - Y_n(g_k)\right] \\
&\leq \sup_{g_k \in \mathfrak{U}_{n,k}} Y_n(g_k) + \inf_{g_k \in \mathfrak{U}_{n,k}} \left[Y_n\left(\hat{f}_k\right) - Y_n(g_k)\right], \qquad (24)
\end{aligned}$$

and note that by Condition 4,

$$\Pr\left(\inf_{g_k\in\mathfrak{U}_{n,k}}\left[Y_n\left(\hat{f}_k\right)-Y_n\left(g_k\right)\right]>0\right)\leq\Pr\left(\hat{f}_k\notin\mathfrak{U}_{n,k}\right)\leq e^{-\tau}.$$

Then, Lemma 1 applies with $\mathfrak{F}_k$ replaced by $\mathfrak{U}_{n,k}$ with probability $1-e^{-\tau}$. To see this just use (24) in the control of II in the proof of Lemma 1. Then, the proof is identical to the one of Theorem 1 but using $\mathfrak{U}_{n,k}$ rather than $\mathfrak{F}_k$. However, by Lemma 16, the stated bound now holds with probability at least $1-2e^{-\tau}$, which is a well defined probability value for $\tau>\ln 2$. □

Results related to the bootstrap are proved next. To this end, the following bootstrap approximation is required.

Lemma 10 (Bootstrap approximation) *Let* $(\mathbb{G}(f))_{f\in\mathfrak{F}_k}$ *be a Gaussian process with covariance function* $\sigma(f,g)$ *as in Condition 1. For any k, under Conditions 1 and 2, there exist mean zero Gaussian processes* $(\mathbb{G}'_{b,n}(f))_{f\in\mathfrak{F}_k}\overset{d}{=}(\mathbb{G}''_{b,n}(f))_{f\in\mathfrak{F}_k}\overset{d}{=}(\mathbb{G}_b(f))_{f\in\mathfrak{F}_k}$ *and a sequence* $r_n\to 0$ *such that*

$$\mathbb{E}\left|\mathbb{G}(f)-\mathbb{G}(g)\right|^2\leq\mathbb{E}\left|\mathbb{G}_b(f)-\mathbb{G}_b(g)\right|^2$$

and for any $\epsilon\in(0,1)$*, with probability at least* $1-\epsilon$*,*

$$\left|\mathbb{E}\left[\sup_{f\in\mathfrak{F}_k}X_{n,b}(f)\,|Z_1^n\right]-\mathbb{E}\sup_{f\in\mathfrak{F}_k}\mathbb{G}'_{b,n}(f)\right|\lesssim\sqrt{H_{\mathfrak{F}}^2\ln\left(1+\frac{r_n}{\epsilon}\right)}$$

and

$$\left|M\left(\sup_{f\in\mathfrak{F}_k}X_{n,b}(f)\,|Z_1^n\right)-M\left(\sup_{f\in\mathfrak{F}_k}\mathbb{G}''_{b,n}(f)\right)\right|\lesssim\sqrt{H_{\mathfrak{F}}^2\ln\left(1+\frac{r_n}{\epsilon}\right)}.$$

Proof Condition 1 and linearity of lim imply

$$n^{-1}\sum_{i=1}^{n}f(Z_i)\,g(Z_i)\xrightarrow{p}\eta(f,g),$$

for some finite function $\eta(f,g)$: $\mathfrak{F}\times\mathfrak{F}\to\mathbb{R}$. Hence, conditioning on the sample values Z_1^n, the bootstrap process $(X_{n,b}(f))_{f\in\mathfrak{F}}$ converges weakly in probability to a mean zero Gaussian process $(\mathbb{G}_b(f))_{f\in\mathfrak{F}}$ with covariance function $\eta(f,g)$. Fidi convergence follows from the Lindeberg Central Limit Theorem. To show stochastic equicontinuity note that $X_{n,b}(f)$ is a martingale with bounded increments $|X_{n,b}(f)-X_{n-1,b}(f)|\leq 2\|\pi_{i,b}\|_\infty\|f\|_\infty$ so that we can apply Lemma 3 and just follow the proof of Lemma 6 step by step to show that (14) holds for $(X_{n,b}(f))_{f\in\mathfrak{F}_k}$ as well with the same semi-metric d_n. This holds both unconditionally and conditioning on the sample sequence Z_1^n. Therefore, Lemma 6 (uniform integrability) implies convergence of moments for the supremum, e.g.,

$$\mathbb{E}\left[\sup_{f\in\mathfrak{F}_k} X_{n,b}(f)\,|Z_1^n\right] \xrightarrow{p} \mathbb{E}\sup_{f\in\mathfrak{F}_k}\mathbb{G}'(f). \tag{25}$$

Then, we just replicate the proof of Lemma 8 with $W_n = \mathbb{E}\left[\sup_{f\in\mathfrak{F}_k} X_{n,b}(f)\,|Z_1^n\right]$ and $W_n' = \mathbb{E}\sup_{f\in\mathfrak{F}_k}\mathbb{G}'_{b,n}(f)$. Note that now W_n' is a constant, but for ease of reference we keep the same notation used in the proof of Lemma 8. We want to show a result analogous to (18) for some suitable choice of t. We can use (25) in place of (17), and, mutatis mutandis, we only need to show that (19) holds for W_n as defined here. We note that

$$\begin{aligned}
\mathrm{I} &= \mathbb{E}\exp\left\{tW_n^2\right\} = \mathbb{E}\exp\left\{t\left[\mathbb{E}\left(\sup_{f\in\mathfrak{F}_k} X_{n,b}(f)\,|Z_1^n\right)\right]^2\right\} \text{ [by definition of } W_n\text{]}\\
&\le \mathbb{E}\exp\left\{t\mathbb{E}\left[\sup_{f\in\mathfrak{F}_k} X_{n,b}(f)\,|Z_1^n\right]^2\right\}\\
&\quad \text{[by convexity]}\\
&\le \mathbb{E}\exp\left\{t\left[\sup_{f\in\mathfrak{F}_k} X_{n,b}(f)\right]^2\right\}
\end{aligned} \tag{26}$$

again by convexity and the tower law for conditional expectations. Since $X_{n,b}(f)$ is a martingale with bounded increments as $Y_n(f)$, by the same arguments used for $Y_n(f)$ we deduce $\|\sup_{f\in\mathfrak{F}_k} X_{n,b}(f)\|_\psi \lesssim H_{\mathfrak{F}}$ implying, mutatis mutandis, (19) for some $t \lesssim H_{\mathfrak{F}}^{-2}$ by Lemma 15. Hence, we just apply Lemma 14 implying the first result. For convergence of the median, note that by the continuous mapping theorem, weak convergence of $(X_{n,b}(f))_{f\in\mathfrak{F}_k}$ implies weak convergence of the supremum and that the median is just the 50% quantile which converges to $M(\sup_{f\in\mathfrak{F}_k}\mathbb{G}_b(f))$ (convergence of distributions implies convergence of all quantiles for smooth distributions assuming the quantiles to be finite). Carrying out a coupling argument for the conditional median rather than the conditional mean, we need to show (26) in the case of the median: $\mathrm{II} = \mathbb{E}\exp\{tM(\sup_{f\in\mathfrak{F}_k} X_{n,b}(f)\,|Z_1^n)\}$ is bounded for some suitably chosen t. Here, $M[\sup_{f\in\mathfrak{F}_k} X_{n,b}(f)\,|Z_1^n]$ is the median of $\sup_{f\in\mathfrak{F}_k} X_{n,b}(f)$ conditioning on Z_1^n. Note that

$$\begin{aligned}
\mathrm{II} &= \mathbb{E}\exp\left\{t\left[M\left(\sup_{f\in\mathfrak{F}_k} X_{n,b}(f)\,|Z_1^n\right)\right]^2\right\}\\
&= \mathbb{E}M\left(\exp\left\{t\left(\sup_{f\in\mathfrak{F}_k} X_{n,b}(f)\right)^2\right\}|Z_1^n\right)\\
&= M\left(\exp\left\{t\left(\sup_{f\in\mathfrak{F}_k} X_{n,b}(f)\right)^2\right\}\right)
\end{aligned}$$

where the second equality follows because the median of a strictly increasing function is the strictly increasing function of the median, and e^{x^2} is strictly increasing for $x > 0$. The third equality follows by taking expectation. We need to show that the above display is bounded. To ease notation, write $\varphi_t = \exp\{t[\sup_{f\in\mathfrak{F}_k} X_{n,b}(f)]^2\}$. Since $\varphi_t \geq 0$, by Markov inequality, $\Pr(\varphi_t \geq 4) \leq \mathbb{E}\varphi_t/4 \leq 1/2$ for some $t \lesssim H_{\mathfrak{F}}^{-2}$, using (26) and Lemma 15. By this remark,

$$M\left(\varphi_t\left(\sup_{f\in\mathfrak{F}_k} X_{n,b}(f)\right)\right) \leq 4, \tag{27}$$

implying $\mathrm{II} \leq 4$ and the proof is completed along the lines of the proof for the conditional mean by an application of Lemma 14.

We finish the proof showing that

$$\mathbb{E}\left|\mathbb{G}(f) - \mathbb{G}(g)\right|^2 \leq \mathbb{E}\left|\mathbb{G}_b(f) - \mathbb{G}_b(g)\right|^2.$$

By (20) and (22),

$$\mathbb{E}\left|\mathbb{G}(f) - \mathbb{G}(g)\right|^2 = \lim_n \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[(1-\mathbb{E}_{i-1})f(Z_i) - (1-\mathbb{E}_{i-1})g(Z_i)\right]^2$$

and mutatis mutandis,

$$\begin{aligned}\mathbb{E}\left|\mathbb{G}_b(f) - \mathbb{G}_b(g)\right|^2 &= \lim_n \frac{1}{n}\mathbb{E}\left[\left(\sum_{i=1}^n \pi_{i,b} f(Z_i)\right) - \left(\sum_{i=1}^n \pi_{i,b} g(Z_i)\right)\right]^2 \\ &= \lim_n \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[f(Z_i) - g(Z_i)\right]^2,\end{aligned}$$

by independence of $\left(\pi_{i,b}\right)_{i\in\mathbb{N}}$. Noting that

$$\mathbb{E}_{i-1}\left[(1-\mathbb{E}_{i-1})f(Z_i) - (1-\mathbb{E}_{i-1})g(Z_i)\right]^2 \leq \mathbb{E}_{i-1}\left[f(Z_i) - g(Z_i)\right]^2$$

the result is deduced using the tower law of conditional expectations. □

Recall the Sudakov–Fernique Inequality (Proposition A.2.6 in VW00).

Lemma 11 *Suppose* $(\mathbb{G}(f))_{f\in\mathfrak{F}}$ *and* $(\mathbb{G}'(f))_{f\in\mathfrak{F}}$ *are separable mean zero Gaussian processes such that*

$$\mathbb{E}\left|\mathbb{G}(f) - \mathbb{G}(g)\right|^2 \leq \mathbb{E}\left|\mathbb{G}'(f) - \mathbb{G}'(g)\right|^2$$

for any $f, g \in \mathfrak{F}$. *Then, for any* $x > 0$,

$$\Pr\left(\sup_{f\in\mathfrak{F}} \mathbb{G}(f) \geq x\right) \leq \Pr\left(\sup_{f\in\mathfrak{F}} \mathbb{G}'(f) \geq x\right)$$

Then, Theorem 2 and Corollary 2 are a direct consequence of the following.

Lemma 12 (Bootstrap inequality) *For* $k \in \{1, \ldots, K\}$, *under Conditions 1 and 2, there exists a sequence* $r_n \to 0$ *and a finite absolute constant* C *such that, for any* $\epsilon \in (0, 1)$, *with probability at least* $1 - \epsilon$,

$$\mathbb{E}\sup_{f\in\mathfrak{F}_k} \mathbb{G}(f) - \mathbb{E}\left[\sup_{f\in\mathfrak{F}_k} X_{n,b}(f) \,|Z_1^n\right] \lesssim \sqrt{H_{\mathfrak{F}}^2 \ln\left(1 + \frac{r_n}{\epsilon}\right)},$$
$$M\left(\sup_{f\in\mathfrak{F}_k} Y_n(f)\right) - M\left(\sup_{f\in\mathfrak{F}_k} X_{n,b}(f) \,|Z_1^n\right) \lesssim \sqrt{H_{\mathfrak{F}}^2 \ln\left(1 + \frac{r_n}{\epsilon}\right)}.$$

Proof By Lemma 10,

$$\mathbb{E}\,|\mathbb{G}(f) - \mathbb{G}(g)|^2 \leq \mathbb{E}\,|\mathbb{G}_b(f) - \mathbb{G}_b(g)|^2, \tag{28}$$

where $\mathbb{G}_b(f)$ and $\mathbb{G}(f)$ are the Gaussian processes of Lemma 10, so that, by Lemma 11,

$$\Pr\left(\sup_{f\in\mathfrak{F}} \mathbb{G}(f) \geq x\right) \leq \Pr\left(\sup_{f\in\mathfrak{F}} \mathbb{G}_b(f) \geq x\right) \tag{29}$$

for all x, implying

$$\mathbb{E}\sup_{f\in\mathfrak{F}} \mathbb{G}(f) \leq \mathbb{E}\sup_{f\in\mathfrak{F}} \mathbb{G}_b(f). \tag{30}$$

Now, consider the following identity,

$$\begin{aligned}
&\mathbb{E}\sup_{f\in\mathfrak{F}} \mathbb{G}(f) - \mathbb{E}\left(\sup_{f\in\mathfrak{F}} X_{n,b}(f) \,|Z_1^n\right) \\
&= \left[\mathbb{E}\sup_{f\in\mathfrak{F}} \mathbb{G}(f) - \mathbb{E}\sup_{f\in\mathfrak{F}} \mathbb{G}_b'(f)\right] + \left[\mathbb{E}\sup_{f\in\mathfrak{F}} \mathbb{G}_b'(f) - \mathbb{E}\left(\sup_{f\in\mathfrak{F}} X_{n,b}(f) \,|Z_1^n\right)\right] \\
&= \mathrm{I} + \mathrm{II}.
\end{aligned}$$

The result of the Lemma follows bounding I by (30) (i.e., $\mathrm{I} \leq 0$) and II by Lemma 10. The inequality for the median also follows using (29) and Lemma 10. □

Finally, this is the proof of Theorem 3.

Proof of Theorem 3 We need to show that for any $\tau > 0$ and $\delta = \delta_n^B > 0$ we can find a B_0 such that for $B \geq B_0$,

$$\begin{aligned}\Pr\left(\max_{b\in\{1,\ldots,B\}}\left|\hat{f}_{n,k}^b\right| + \delta \geq \left|\hat{f}_{n,k}\right|\right) &= \Pr\left(\left(\min_{b\in\{1,\ldots,B\}}\hat{f}_{n,k}^b \wedge 0\right)\right. \\ &\quad \left. - \delta \leq \hat{f}_{n,k} \leq \left(\max_{b\in\{1,\ldots,B\}}\hat{f}_{n,k}^b \vee 0\right) + \delta\right) \\ &\geq 1 - e^{-\tau}.\end{aligned}$$

For simplicity, we assume $\mathfrak{F}_k$ only contains positive functions, so that we only need to show that

$$\Pr\left(\hat{f}_{n,k} \leq \max_{b\in\{1,\ldots,B\}}\hat{f}_{n,k}^b + \delta\right) \geq 1 - e^{-\tau}.$$

Conditioning on the sample sequence Z_1^n,

$$\sup_{f\in\mathfrak{F}_k}\left|\mathcal{R}_n^*\left(Z_1^n, f, M_{i,b}\right) - \mathcal{R}_n\left(Z_1^n, f\right)\right| = \sup_{f\in\mathfrak{F}_k}\left|\frac{1}{n}\sum_{i=1}^n\left(M_{i,b} - 1\right) f\left(Z_i\right)\right| \xrightarrow{p} 0$$

by Markov inequality and (25). This together with (12) implies that, for any b and for any $\delta > 0$ there exists a $\gamma_{n,\delta} \in (0, 1)$ such that

$$\Pr\left(\hat{f}_{n,k} \leq \hat{f}_{n,k}^b + \delta | Z_1^n\right) \geq \left(1 - \gamma_{n,\delta}\right) \uparrow 1, \ a.s.$$

i.e., conditioning on Z_1^n, $\hat{f}_{n,k}^b \xrightarrow{p} \hat{f}_{n,k}$ for any b as $n \to \infty$ (VW00, Corollary 3.2.3). Hence, for any n,

$$\begin{aligned}\Pr\left(\max_{b\in\{1,\ldots,B\}}\hat{f}_{n,k}^b - \hat{f}_{n,k} < -\delta\right) &= \mathbb{E}\Pr\left(\max_{b\in\{1,\ldots,B\}}\hat{f}_{n,k}^b - \hat{f}_{n,k} < -\delta | Z_1^n\right) \\ &= \mathbb{E}\left[\Pr\left(\hat{f}_{n,k}^b - \hat{f}_{n,k} < -\delta | Z_1^n\right)\right]^B \\ &\quad \text{[by independence conditioning on } Z_1^n\text{]} \\ &= \mathbb{E}\left[1 - \Pr\left(\hat{f}_{n,k} \leq \hat{f}_{n,k}^b + \delta | Z_1^n\right)\right]^B \\ &\leq \mathbb{E}\left(\gamma_{n,\delta}\right)^B.\end{aligned}$$

Since $\gamma_{n,\delta}$ is bounded and coverges to zero a.s., there is a non random sequence $\gamma_{n,\delta}$ such that $\mathbb{E}\left(\gamma_{n,\delta}\right)^B \leq \left(\gamma'_{n,\delta}\right)^B \to 0$. This means that for any $\tau > 0, \delta > 0$ and $n > 0$, we can choose a B_0 such that, for $B \geq B_0$, $\left(\gamma'_{n,\delta}\right)^B \leq e^{-\tau}$. □

4.5 Supplementary Lemmata

The following is cited in the text after Theorem 1.

Lemma 13 *Set* $pen_n(\mathfrak{F}_k) = 0$ *so that*

$$\hat{\mathcal{R}}_n\left(Z_1^n, f_k\right) := \mathcal{R}_n\left(Z_1^n, f_k\right)$$

and

$$\hat{f}_{n,\hat{k}} := \arg \min_{k\in\{1,\ldots,K\}} \hat{\mathcal{R}}_n\left(Z_1^n, \hat{f}_k\right) = \arg \min_{k\in\{1,\ldots,K\}} \mathcal{R}_n\left(Z_1^n, \hat{f}_k\right).$$

Then, there is a finite absolute constant C such that, for all $\tau > 0$*, with probability at least* $1 - e^{-\tau}$

$$\mathcal{R}\left(Z_1^n, \hat{f}_{\hat{k}}\right) = \mathcal{R}\left(Z_1^n, f_k\right) + C\sqrt{\frac{H_{\mathfrak{F}}^2 \ln(2K+\tau)}{n}}.$$

Proof Note that

$$\begin{aligned}\mathcal{R}\left(Z_1^n, \hat{f}_{\hat{k}}\right) = \mathcal{R}\left(Z_1^n, f_k\right) &+ \left[\mathcal{R}\left(Z_1^n, \hat{f}_{\hat{k}}\right) - \mathcal{R}_n\left(Z_1^n, \hat{f}_{\hat{k}}\right)\right] \\ &+ \left[\mathcal{R}_n\left(Z_1^n, \hat{f}_{\hat{k}}\right) - \mathcal{R}\left(Z_1^n, f_k\right)\right]\end{aligned}$$

and

$$\left[\mathcal{R}\left(Z_1^n, \hat{f}_{\hat{k}}\right) - \mathcal{R}_n\left(Z_1^n, \hat{f}_{\hat{k}}\right)\right] \le \max_{k\in\{1,\ldots,K\}} \sup_{f\in\mathfrak{F}_k} \left[\mathcal{R}\left(Z_1^n, f\right) - \mathcal{R}_n\left(Z_1^n, f\right)\right].$$

By Markov inequality and the union bound,

$$\begin{aligned}&\Pr\left(\max_{k\in\{1,\ldots,K\}} \sup_{f\in\mathfrak{F}_k} \sqrt{n}\left[\mathcal{R}\left(Z_1^n, f\right) - \mathcal{R}_n\left(Z_1^n, f\right)\right] > x\right) \\ &\le K \frac{\mathbb{E}\exp\left\{t \sup_{f\in\mathfrak{F}_k} \sqrt{n}\left[\mathcal{R}\left(Z_1^n, f\right) - \mathcal{R}_n\left(Z_1^n, f\right)\right]\right\}}{\exp\left\{tx^2\right\}}\end{aligned}$$

for some suitable $t > 0$. The expectation can be bound by the ψ Orlicz norm so that by Remark 6 this expectation is finite if $t \lesssim H_{,\mathfrak{F}}^{-2}$. This implies that with probability at least $1 - \epsilon$

$$\max_{k\in\{1,\ldots,K\}} \sup_{f\in\mathfrak{F}_k} \left[\mathcal{R}\left(Z_1^n, f\right) - \mathcal{R}_n\left(Z_1^n, f\right)\right] \lesssim \sqrt{\frac{H_{\mathfrak{F}}^2 \ln(K/\epsilon)}{n}}.$$

The result follows by crudely bounding $\sqrt{n}[\mathcal{R}_n(Z_1^n, \hat{f}_{\hat{k}}) - \mathcal{R}(Z_1^n, f_k)]$ with the above display along the same lines of the proof of Theorem 1. □

The following lemma is simple, but convenient.

Lemma 14 *Suppose that* $(X_n)_{n\in\mathbb{N}}$ *is a sequence of random variables converging in probability to a random variable* X. *Suppose that for any* $x > 0$, *and* $n > 0$, $\Pr(|X_n| > x) \lesssim \exp\{-tx^2\}$ *and* $\Pr(|X| > x) \lesssim \exp\{-tx^2\}$ *for some* $t > 0$. *Then, there exists a sequence* $r_n \to 0$ *as* $n \to \infty$ *such that for any* $\epsilon \in (0, 1)$, *with probability at least* $1 - \epsilon$,

$$|X_n - X| \leq \sqrt{\frac{4}{t} \ln\left(1 + \frac{r_n}{\epsilon}\right)}.$$

Proof We claim that by the conditions of the lemma, for $\psi(x) = e^{x^2} - 1$ and some $z > 0$, $r_n := \mathbb{E}\psi(z|X_n - X|) \to 0$. Since $\psi(0) = 0$, and $|X_n - X_n'| \xrightarrow{p} 0$, to show convergence of this expectation, it is sufficient to show uniform integrability of $\psi(z|X_n - X|)$ which is implied by integrability of $\psi(2z|X_n - X|)$. Clearly

$$\mathbb{E}\psi(2z|X_n - X|) \leq \mathbb{E}\psi(4zX_n) + \mathbb{E}\psi(4zX) \lesssim z^{-1/2}$$

by Lemma 4 for $4z \leq t$. Hence, for $z \leq t/4$,

$$\begin{aligned}\Pr(|X_n - X| > x) &\leq \frac{\mathbb{E}\psi\left(z\left|X_n - X_n'\right|\right)}{\psi(zx)} = \frac{r_n}{\psi(zx)} \\ &= \epsilon \in (0, 1) \quad \text{for } x = \sqrt{\frac{4}{t} \ln\left(1 + \frac{r_n}{\epsilon}\right)}.\end{aligned}$$

The last equality is found by solving $r_n/\psi(zx) = \epsilon$ and replacing the constraint on z. □

The next three results are elementary and stated for convenience of repeated reference. The first follows by definition of the Orlicz norm (Definition 2), the second by a simple application of Bonferroni inequality while the third by the union bound.

Lemma 15 *Suppose that* X *is a random variable with* ψ *Orlicz norm satisfying* $\|X\|_\psi \leq C$ *for some finite absolute constant* C. *Then* $\mathbb{E}\exp\{tX^2\} \leq 2$ *for* $t \leq C^{-2}$.

Lemma 16 *Suppose* $X_1, \ldots, X_I$ *are real valued random variables. Then, for any* $x_i \in \mathbb{R}$ $(i = 1, \ldots, I)$,

$$\Pr\left(\sum_{i=1}^{I} X_i \leq \sum_{i=1}^{I} x_i\right) \geq 1 - \sum_{i=1}^{I} \Pr(X_i > x_i).$$

Lemma 17 *Suppose* $X_1, \ldots, X_K$ *are random variables and there is a function* Q: $(0, 1) \to \mathbb{R}$ *such that, for any* $k \in \{1, \ldots, K\}$ *and* $\epsilon \in (0, 1)$, $\Pr(X_k > Q(\epsilon)) \leq \epsilon$. *Then,*

$$\Pr\left(\max_{k\in\{1,\ldots,K\}} X_k > Q(\epsilon/K)\right) \leq \epsilon.$$

References

Bartlett, P., Boucheron, G., Lugosi, G. (2002). Model selection and error estimation. *Machine Learning*, *48*, 85–113.

Bartlett, P., Bousquet, O., Mendelson, S. (2005). Local rademacher complexities. *Annals of Statistics*, *33*, 1497–1537.

Bühlmann, P. (1997). Sieve Bootstrap for time series. *Bernoulli*, *3*, 123–148.

Cesa-Bianchi, N., Lugosi, G. (2001). Worst-case bounds for the logarithmic loss of predictors. *Machine Learning*, *43*, 247–264.

Dawid, A. P. (1984). Present position and potential developments: some personal views: statistical theory: the prequential approach. *Journal of the Royal Statistical Society Series A*, *147*, 278–292.

Dawid, A. P. (1985). Calibration-based empirical probability. *The Annals of Statistics*, *13*, 1251–1274.

Dawid, A. P. (1986). Probability forecasting. In S. Kotz, N. L. Johnson, C. B. Read (Eds.), *Encyclopedia of statistical sciences* (Vol. 7, pp. 210–218). New York: Wiley.

De la Peña, V. H. (1999). A general class of exponential inequalities for Martingales and ratios. *Annals of Probability*, *27*, 537–564.

Devroye, L., Györfi, L., Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. New York: Springer.

Doukhan, P., Leon, J. R., Portal, F. (1987). Principes d'Invariance Faible pour la Mesure Empirique d'un Suite de Variables Aléatoires Mélangeante. *Probability Theory and Related Fields*, *76*, 51–70.

Dudley, R. M. (2002). *Real analysis and probability*. Cambridge: Cambridge University Press.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal American Statistical Association*, *78*, 316–331.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, *29*, 1189–1232.

Fromont, M. (2007). Model selection by bootstrap penalization for classification. *Machine Learning*, *66*, 165–207.

Gray, R. M., Kieffer, J. C. (1980) Asymptotically mean stationary measures. *Annals of Probability*, *8*, 962–973.

Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, *47*, 1902–1914.

Levental, S. (1989). A uniform CLT for uniformly bounded families of Martingale differences. *Journal of Theoretical Probability*, *2*, 271–287.

Lugosi, G., Wegkamp, M. (2004). Complexity regularization via localized random penalties. *Annals of Statistics*, *32*, 1679–1697.

Mammen, E. (1992). Bootstrap, wild bootstrap, and asymptotic normality. *Probability Theory Related Fields*, *93*, 439–455.

McLeish, D. L. (1974). Dependent central limit theorems and invariance principles. *Annals of Probability*, *2*, 620–628.

Petrov, V. (1995). *Limit Theorems of probability theory*. Oxford: Oxford University Press.

Ripley, B. (1996). *Pattern recognition and neural networks*. Cambridge: Cambridge University Press.

Rüschendorf, L., de Valk, V. (1993). On regression representation of stochastic processes. *Stochastic Processes and their Applications*, *46*, 183–198.

Seillier-Moiseiwitsch, P., Dawid, A. P. (1993). On testing the validity of sequential probability forecasts. *Journal of the American Statistical Association*, *88*, 355–359.

Skouras, K., Dawid, P. (2000). *Consistency in misspecified models.* Research report 218. Department of Statistical Science, University College London.

Van der Laan, M. J., Dudoit, S. (2003). *Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples*. U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper 130.

Van der Vaart, A., Wellner, J. A. (2000). *Weak convergence of empirical processes.* Springer series in statistics. New York: Springer.

Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley.