Uniform Risk Bounds for Learning with Dependent Data Sequences

Fabien Lauer

Uniform Risk Bounds for Learning with Dependent Data Sequences

Fabien Lauer

LORIA, Université de Lorraine, CNRS, France

Abstract

This paper extends standard results from learning theory with independent data to sequences of dependent data. Contrary to most of the literature, we do not rely on mixing arguments or sequential measures of complexity and derive uniform risk bounds with classical proof patterns and capacity measures. In particular, we show that the standard classification risk bounds based on the VC-dimension hold in the exact same form for dependent data, and further provide Rademacher complexity-based bounds, that remain unchanged compared to the standard results for the identically and independently distributed case. Finally, we show how to apply these results in the context of scenario-based optimization in order to compute the sample complexity of random programs with dependent constraints.

1 Introduction

Statistical learning theory offers probabilistic guarantees on the accuracy of models learned from data. Most of these results assume that the data come from a realization of a sample of independent and identically distributed (i.i.d.) random variables, which allows one to build the theory upon standard concentration arguments. However, this assumption is often unrealistic as dependent data are ubiquitous in real-world applications, such as signal processing, speech recognition, biological sequence annotation (Baldi and Brunak, 2001), dynamical system identification (Ljung, 1987), or even handwritten character recognition where the images collected for training come from a string of letters forming a meaningful text.

This paper extends several classical results to sequences of dependent data, such as risk bounds based on the Vapnik-Chervonenkis (VC) dimension or the Rademacher complexity. In particular, we focus on *uniform* risk bounds that are more suitable for nonconvex loss functions difficult to minimize in practice.

As a motivating application, we also consider the consequences of these results in the framework of scenario-based optimization for solving uncertain optimization problems. Here, robust solutions are those that typically satisfy an infinite number of constraints: one for each value of the uncertain parameter of the problem. Scenario-based optimization computes instead probably approximately correct solutions by sampling the set of uncertainties and

solving the problem with a finite number of constraints. In this context, the computing complexity is directly related to the number of sampled constraints and it is thus of primary importance to compute the sample complexity of the corresponding random program, i.e., the smallest sample size for which we can guarantee with a high confidence that the probability of violation of the constraint is low.

Related work Dependence between training instances typically occur for machine learning in the context of ranking problems, where the training algorithms deal with overlapping pairs of input data. In this setting, prior work on generalization bounds (Usunier et al., 2006; Ralaivola and Amini, 2015) relied on a decomposition of the training sample into independent subsamples for which concentration could be applied by following the graph-coloring scheme of Janson (2004).

The literature also contains numerous results that do not assume that such a decomposition is possible, e.g., when a single training instance depends on all the others as is common when handling time-series or sequential data. In that case, the general approach is to measure the degree of dependence between the data points with a mixing coefficient (Bradley, 2005) and assume that this coefficient tends to zero sufficiently quickly to allow for the derivation of meaningful bounds. Works in this line include that of Meir (2000); Steinwart and Christmann (2009); Mohri and Rostamizadeh (2009, 2010) and rely on technical arguments inherited from Yu (1994). Though the obtained risk bounds share most of their structure with their counterpart for i.i.d. data, they also involve the mixing coefficient, which slightly degrades the convergence and remains difficult to determine or estimate in practice. Note that a connection between the mixing and graph-coloring arguments is discussed in Ralaivola et al. (2010).

Other approaches based on Rademacher complexities include that of Rakhlin et al. (2015), which can also deal with non-stationary sequences (Kuznetsov and Mohri, 2015), but involves complex computations with tree processes and decoupling techniques from de la Peña and Giné (1999). Also, these works consider a different form of the risk (a conditional forecasting risk) and rely on sequential counterparts of standard capacity measures.

A more recent line of research developped by Simchowitz et al. (2018); Faradonbeh et al. (2018) relies on techniques from Mendelson (2014, 2018) to bypass the need for mixing arguments. However, these results apply only to the empirical risk minimizer (the orthogonal least-squares estimator is considered in Simchowitz et al. (2018); Faradonbeh et al. (2018)). Thus, they do not provide uniform risk bounds that apply to any model in a given class, which is critical for applications with nonconvex loss functions where the empirical risk minimizer remains elusive (such as unsupervised learning or hybrid dynamical system identification (Lauer and Bloch, 2019)).

Regarding our motivating application, i.e., scenario-based optimization, we can distinguish two main lines of research. The first, developed for instance in Alamo et al. (2009); Lauer (2023), builds upon learning theory to derive bounds on the probability of violation and sample complexities. The second, pioneered by Campi and Garatti (2008); Calafiore (2010), relies instead on convex analysis arguments that lead to tighter bounds for the specific case

of convex optimization. A number of works also tried to extend the latter to various forms of nonconvex programs (Esfahani et al., 2014; Campi et al., 2018). However, all these works, either based on learning theory or convex analysis, assume that the scenarios are sampled independently, which could be problematic for certain applications, as, e.g., the one of Wang et al. (2021).

Contribution We show that many standard results from learning theory, such as classification VC bounds and Rademacher complexity-based bounds, apply *without any modification* to dependent data. In addition, we derive these results with proofs that follow the standard patterns and rely on the classical capacity measures, which stands in contrast to the approach of Rakhlin et al. (2015) which requires additional arguments and more complex computations with sequential complexities. In comparison with other works from the literature, we obtain *uniform* risk bounds that do not introduce additional terms or mixing coefficients and that are thus more widely applicable than those of Simchowitz et al. (2018); Faradonbeh et al. (2018) and tighter than those of, e.g., Mohri and Rostamizadeh (2009).

Technically, our results rely on a simple construction of the ghost sample that enjoys the necessary properties, and, for Rademacher complexity-based bounds, a concentration inequality adapted to dependent variables by following van De Geer (2007). Specifically, these allow us to prove the following claims along the different sections of the paper.

- The standard classification risk bounds of Vapnik (1998) that are based on the VC-dimension hold in the exact same form for non-i.i.d. training sequences. Interestingly, no new concentration result is required for the proof (Sect. 3).
- With an additional stationarity assumption, similar conclusions hold for VC relative deviation bounds and classification risk bounds with fast rates, thus generalizing the results of Vapnik (1998); Cortes et al. (2019) (Sect. 3.1).
- Extending the results above to deal with regression problems poses no difficulty in the non-i.i.d. context (Sect. 3.2).
- When the Rademacher complexity can be bounded in terms of the marginal distributions of the data, Rademacher complexity-based bounds identical to the ones for the standard i.i.d. case hold for dependent data. This is true in particular for linear and kernel machines, thus generalizing the results of Bartlett and Mendelson (2002) (Sect. 4).
- Using the chaining method, standard risk bounds based on uniform covering numbers or the fat-shattering dimension are shown to hold also for non-i.i.d. data (Sect. 4.1).
- In the framework of robust optimization, the sample complexities of random programs derived in Alamo et al. (2009); Lauer (2023) for independent scenarios also hold with dependent scenarios (Sect. 5).

Finally, the main drawback of our approach is that it does not allow the derivation of data-dependent bounds, i.e., risk bounds in which the complexity term is evaluated with respect to the available training sample. Data-dependent bounds may be tighter than worst-case or average estimates. How-

ever, this improvement remains limited in many standard cases. In addition, keeping in mind our motivation stemming from the field of scenario-based optimization, we note that data-dependent bounds are irrelevant for computing sample complexity estimates.

2 Preliminaries

This section presents the general learning framework with dependent data sequences and the basic tools needed to derive our main results.

2.1 Learning framework

Let $(\mathcal{A}_i)_{0\leq i\leq n}$ denote a filtration and $\boldsymbol{Z}_n = (Z_i)_{1\leq i\leq n}$ a sequence of random variables $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y} = \mathcal{Z}$ adapted to $(\mathcal{A}_i)_{1\leq i\leq n}$,[1] without any other assumption. Given a model class $\mathcal{F}$ of functions from $\mathcal{X} \to \mathcal{Y}$ and a bounded loss function $\ell : \mathcal{Y}^2 \to [0, B]$, we consider learning a model $f \in \mathcal{F}$ from such a training sequence $\boldsymbol{Z}_n$ and aim at the estimation of its risk,

$$L_n(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\ell(f(X_i), Y_i), \tag{1}$$

where $\mathbb{E}$ denotes the expectation, by its empirical risk

$$\hat{L}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i). \tag{2}$$

Note that for stationary sequences, the Z_i variables are identically distributed and the risk (1) merely boils down to the standard learning risk $\mathbb{E}\ell(f(X_1), Y_1)$, also considered for sequential data, e.g., in Mohri and Rostamizadeh (2009).

For non-stationary sequences, the risk in (1) does not really assess the ability of the model to predict future values of Y_i, since these need not share the same distribution with the Y_i's in (1). However, it remains a valuable quantity for other applications, such as vector quantization or clustering problems, where, after the obvious reformulation of the loss as $\ell : \mathcal{X}^2 \to [0, B]$, the risk in (1) stands for the distortion or the clustering risk and evaluates how well the model approximates the distribution of the training sample.

Regarding prediction problems, note that the risk (1) differs from the conditional risk

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\ell(f(X_i), Y_i) \mid \mathcal{A}_{i-1}\right] \tag{3}$$

studied, e.g., in Kuznetsov and Mohri (2015); Rakhlin et al. (2015). The conditional risk (3) only measures the ability of the model to predict the (immediate) future of the training sequence, which is particularly well-suited for time-series forecasting problems. However, it is not suitable for applications such as dynamical system identification where the model is learned from

[1] A filtration is a sequence of increasing σ-algebras $\emptyset = \mathcal{A}_0 \subset \mathcal{A}_1 \subset \mathcal{A}_2 \subset \cdots \subset \mathcal{A}_n$ and a sequence of random variables is adapted to it if each Z_i is $\mathcal{A}_i$-measurable.

training data offline, and then applied to predict the output of the system that might have been reset in the meantime. In other words, the conditional risk (3) refers only to the sample path of the random process used for training, whereas the risk in (1) averages over all sample paths.

2.2 Basic tools

Most results in statistical learning theory rely on concentration arguments, which are easily available for samples of independent variables. While we will see in Sect. 3 that these are sufficient to extend some standard risk bounds to dependent sequences, others will require a concentration inequality for samples of dependent variables. It is stated here as a specification of the results in van De Geer (2007). The detailed proof can be found in Appendix A for completeness.

Theorem 1 (Bounded difference inequality for sequences of dependent variables)**.** *Let $(Z_i)_{1\leq i\leq n}$ denote a sequence of random variables (not necessarily stationary) taking values in $\mathcal{Z}$ and g be a real-valued function of Z_1, ..., Z_n such that it is $\mathcal{A}_n$-measurable and*

$$\sup_{\substack{(z_j)_{1\leq j\leq n}\in\mathcal{Z}^n \\ z'\in\mathcal{Z}}} |g(z_1,\ldots,z_i,\ldots,z_n) - g(z_1,\ldots,z',\ldots,z_n)| \leq c_i, \quad i=1,\ldots,n.$$

Then, for any $\epsilon > 0$,

$$P\left\{g(Z_1,\ldots,Z_n) - \mathbb{E}g(Z_1,\ldots,Z_n) > \epsilon\right\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

Note that Theorem 1 provides the exact same result as McDiarmid's inequality for i.i.d. variables (McDiarmid, 1989), with the same exponential rate.

We also recall a famous result from Hoeffding (1963), in its original form for *independent* variables, which is all we will require in the sequel.

Theorem 2 (Hoeffding's inequality)**.** *Let $(W_i)_{1\leq i\leq n}$ denote a sequence of independent random variables satisfying $W_i \in [a_i, b_i]$. Then, for any $\epsilon > 0$,*

$$P\left\{\frac{1}{n}\sum_{i=1}^n W_i - \frac{1}{n}\sum_{i=1}^n \mathbb{E}W_i > \epsilon\right\} \leq \exp\left(\frac{-2n^2\epsilon^2}{\sum_{i=1}^n (b_i-a_i)^2}\right).$$

3 Classification risk bounds based on the VC-dimension

We first discuss how to derive the standard VC classification risk bounds for dependent data. The first ingredient of such bounds is a symmetrization lemma, which in turn requires two things: a concentration inequality and a ghost sample. In the classical scenario where the training sample $\boldsymbol{Z}_n$ is made of independent copies of some random variable Z, the ghost sample $\boldsymbol{Z}'_n$ can merely be taken as an independent copy of $\boldsymbol{Z}_n$. Then, standard concentration inequalities apply to this sample of i.i.d. variables Z'_i.

In the case of a sequence of dependent variables, we instead build the ghost sample with variables Z'_i taken as independent copies of the variables Z_i. Thus, we obtain a sample $\boldsymbol{Z}'_n$ of independent variables that is also independent of $\boldsymbol{Z}_n$ and with each $Z'_i \sim Z_i$ (the notation $A \sim B$ indicates that the two random variables A and B share the same distribution). Note however that the resulting ghost sample $\boldsymbol{Z}'_n$ need not share the same joint distribution with $\boldsymbol{Z}_n$ as in the classical i.i.d. scenario. But, a careful look at the proof of the standard symmetrization lemma of Vapnik (1998) shows that this is not needed for symmetrization to hold. In addition, in this proof, the concentration inequality is only applied to the ghost sample, not the training sample. Since our ghost sample is made of independent variables by construction, we have the following (detailed proof in App. B).

Lemma 1 (Symmetrization)**.** *Let $\boldsymbol{Z}_n$ denote a sequence of possibly dependent variables and consider the ghost sample $\boldsymbol{Z}'_n = (Z'_i)_{1\leq i\leq n}$ with each $Z'_i = (X'_i, Y'_i)$ built as an independent copy of $Z_i = (X_i, Y_i)$. Then, for any ϵ such that $n\epsilon^2 \geq 2B^2$,*

$$P\left\{\sup_{f\in\mathcal{F}} L_n(f) - \hat{L}_n(f) \geq \epsilon\right\} \leq 2P\left\{\sup_{f\in\mathcal{F}} \hat{L}'_n(f) - \hat{L}_n(f) \geq \frac{\epsilon}{2}\right\},$$

where $\hat{L}'_n(f) = \frac{1}{n}\sum_{i=1}^n \ell(f(X'_i), Y'_i)$.

Equipped with this tool, classification risk bounds based on the zero-one loss[2]

$$\ell(\hat{y}, y) = \mathbf{1}_{\hat{y}\neq y} \tag{4}$$

can be obtained in terms of the growth function $\Pi_{\mathcal{F}}(n)$ of the class $\mathcal{F}$ that can be bounded by the VC-dimension d_{VC}.

Definition 1 (Growth function)**.** *For a set $\mathcal{F}$ of classifiers $f : \mathcal{X} \to \mathcal{Y}$ with a discrete set $\mathcal{Y}$ of finite cardinality, its* growth function *is the largest cardinality of the set of classifications produced by its classifiers over all sets $\boldsymbol{x}_n = (x_i)_{1\leq i\leq n}$ of n points:*

$$\Pi_{\mathcal{F}}(n) = \sup_{\boldsymbol{x}_n\in\mathcal{X}^n} |\{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\}|.$$

Definition 2 (VC-dimension)**.** *The* Vapnik-Chervonenkis (VC) dimension *d_{VC} of a set $\mathcal{F}$ of binary classifiers $f : \mathcal{X} \to \mathcal{Y}$, $|\mathcal{Y}| = 2$, is the largest number of points n such that $\Pi_{\mathcal{F}}(n) = 2^n$.*

More precisely, it is a remarkable fact that no new concentration result is needed to derive a risk bound for dependent data from Lemma 1. The trick is that, by introducing independent Rademacher variables and conditioning, we only require concentration with respect to these variables rather than for the samples themselves. Therefore, the classical Hoeffding inequality of Theorem 2 applies and the proof of the next result (given in Appendix C) only requires that the quantity $\mathbf{1}_{f(X_i)\neq Y_i} - \mathbf{1}_{f(X'_i)\neq Y'_i}$ is a symmetric random variable, which is ensured by the fact that (X'_i, Y'_i) is an independent copy of (X_i, Y_i).

[2]With the loss (4), the risk (1) is just the probability of misclassification $P(f(X) \neq Y)$ for stationary sequences.

Theorem 3 (Basic VC risk bound for dependent data)**.** *Let $\mathcal{F}$ be a set of classifiers $f : \mathcal{X} \to \{-1, +1\}$ with VC-dimension d_{VC} and ℓ denote the classification loss (4). Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\forall f \in \mathcal{F}, \quad L_n(f) \leq \hat{L}_n(f) + 2\sqrt{2\frac{\log \Pi_{\mathcal{F}}(2n) + \log \frac{2}{\delta}}{n}}$$
$$\leq \hat{L}_n(f) + 2\sqrt{2\frac{d_{VC} \log \frac{2en}{d_{VC}} + \log \frac{2}{\delta}}{n}}.$$

Note that Theorem 3 applies to dependent training sequences $\boldsymbol{Z}_n$ and provides the exact same result as Vapnik (1998) did for the i.i.d. case.

Example 1 (Linear classification)**.** *Consider the set of linear classifiers, $\mathcal{F} = \{f : f(x) = sign(\langle w, x\rangle + b),\ w \in \mathbb{R}^d,\ b \in \mathbb{R}\}$, of VC-dimension $d_{VC} = d + 1$ (Vapnik, 1998). For all $d \geq 3$, Theorem 3 then yields for $n = 100\,000$ that, with 95% probability, the risk of any classifier $f \in \mathcal{F}$ can be estimated from its empirical risk on the training sample of dependent instances with accuracy no less than $0.031\sqrt{d+1} + 0.015$.*

3.1 Relative deviation bounds

Risk bounds with a faster convergence rate close to $O(1/n)$ instead of $O(1/\sqrt{n})$ can be derived from relative deviation bounds of the form

$$P\left\{\sup_{f\in\mathcal{F}} \frac{L_n(f) - \hat{L}_n(f)}{\sqrt{L_n(f)}} \geq \epsilon\right\} \leq A \exp(-nC),$$

for some constants A and C. In turn, obtaining such results requires a different form of symmetrization, associated to a concentration argument.

The detailed proof of the corresponding symmetrization given in Cortes et al. (2019) shows that independence is used at only two different places. It is first required between data points to apply a binomial tail bound, but only on the ghost sample. Then, the fact that the training sample is independent of the ghost sample is used.

Considering now a construction of the ghost sample as in Theorem 1, we directly obtain the required indepedence: $\boldsymbol{Z}_n$ is independent of $\boldsymbol{Z}'_n$ and all Z'_i in the ghost sample are independent of each other. This leads to the following generalization of the result of Vapnik (1998); Cortes et al. (2019) to stationary sequences of training data (the detailed proof can be found in Appendix D for completeness). Note that stationarity, which is required for the binomial tail bound, is a reasonable assumption for a prediction task as considered here.

Lemma 2 (Symmetrization for relative deviations)**.** *Let $\boldsymbol{Z}_n = (Z_i)_{1\leq i\leq n}$ denote a stationary sequence of possibly dependent variables and consider the ghost sample $\boldsymbol{Z}'_n = (Z'_i)_{1\leq i\leq n}$ with each $Z'_i = (X'_i, Y'_i)$ built as an independent copy of $Z_i = (X_i, Y_i)$. Let ℓ denote the classification loss (4). Then, for any ϵ such that $n\epsilon^2 > 1$,*

$$P\left\{\sup_{f\in\mathcal{F}} \frac{L_n(f) - \hat{L}_n(f)}{\sqrt{L_n(f)}} \geq \epsilon\right\} \leq 4P\left\{\sup_{f\in\mathcal{F}} \frac{\hat{L}'_n(f) - \hat{L}_n(f)}{\sqrt{(\hat{L}_n(f) + \hat{L}'_n(f) + \frac{1}{n})/2}} \geq \frac{\epsilon}{2}\right\},$$

where $\hat{L}'_n(f) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{f(X'_i)\neq Y'_i}$.

As for Theorem 3, we can derive a fast-rate risk bound from Lemma 2 using only standard concentration arguments, i.e., by following the proof of Cortes et al. (2019) and introducing independent Rademacher variables, and the symmetry of $\mathbf{1}_{f(X_i)\neq Y_i} - \mathbf{1}_{f(X'_i)\neq Y'_i}$ ensured by our ghost sample construction (detailed proof given in App. E).

Theorem 4 (General VC risk bound for dependent data)**.** *Let $\mathcal{F}$ be a class of classifiers $f : \mathcal{X} \to \{-1, +1\}$ with VC-dimension d_{VC} and ℓ denote the classification loss (4). Let $\boldsymbol{Z}_n = (Z_i)_{1\leq i\leq n}$ denote a stationary sequence of possibly dependent variables. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\begin{aligned}\forall f \in \mathcal{F}, \quad L_n(f) &\leq \hat{L}_n(f) + 2\sqrt{\hat{L}_n(f)\frac{\log \Pi_{\mathcal{F}}(2n) + \log\frac{4}{\delta}}{n}} + 4\frac{\log \Pi_{\mathcal{F}}(2n) + \log\frac{4}{\delta}}{n} \\ &\leq \hat{L}_n(f) + 2\sqrt{\hat{L}_n(f)\frac{d_{VC}\log\frac{2en}{d_{VC}} + \log\frac{4}{\delta}}{n}} + 4\frac{d_{VC}\log\frac{2en}{d_{VC}} + \log\frac{4}{\delta}}{n}.\end{aligned}$$

Here again, the bounds of Theorem 4 provide the same guarantees as the classical results of Vapnik (1998); Cortes et al. (2019), while holding more generally for dependent data. In particular, for classifiers achieving a small empirical error $\hat{L}_n(f)$, these are tighter than the ones of Theorem 3, with a convergence rate in $O(1/n)$ in the optimistic scenario where a perfect fit of the data can be ensured.

Finally, we note that the fast-rate and margin-based risk bounds provided in Cortes et al. (2021) and that involve covering numbers instead of the growth function can also be proved to hold for dependent data using similar arguments as above. However, we refrain from giving the details and will instead consider margin classifiers when discussing Rademacher complexity-based bounds in Sect. 4.

3.2 Regression bounds

When $\ell : \mathcal{Y}^2 \to [0, B]$ is a regression loss bounded by B, the VC theory for classification applies thanks to Inequality (5.11) in Vapnik (1998):

$$P\left\{\sup_{f\in\mathcal{F}} L_n(f) - \hat{L}_n(f) > \epsilon\right\} \leq P\left\{\sup_{f\in\mathcal{F},\beta\in(0,B)} \tilde{L}(f) - \hat{\tilde{L}}_n(f) > \frac{\epsilon}{B}\right\}, \quad (5)$$

where $\tilde{L}(f)$ and $\hat{\tilde{L}}_n(f)$ are defined as in (1)–(2) with the classification loss $\tilde{\ell}(f(X_i), Y_i) = \mathbf{1}_{\ell(f(X_i),Y_i)-\beta\geq 0}$. Interestingly, it can be checked that Inequality (5) holds irrespective of the independence of the Z_i. Thus, the classification bounds above can be applied to its right-hand side with the VC-dimension of a function class that has one additional parameter $\beta \in (0, B)$, and this yields bounds on the regression risk of models learned from dependent data sequences.

Example 2 (VC bound for linear system identification)**.** *Consider a regression problem with data generated by a linear dynamical system as $y_i =$*

$\langle \theta, x_i \rangle + \nu_i$ *with a random noise term* ν_i *and* $x_i = (y_{i-1}, \ldots, y_{i-d})$. *In this case, the data collected from a single trajectory of the system clearly cannot be assumed to be i.i.d., but the results above still yield performance guarantees. Specifically, for the squared loss* $\ell(\hat{y}, y) = (\hat{y} - y)^2$ *and a class of linear models* $\mathcal{F} = \{f : f(x) = \langle w, x \rangle, \ w \in \mathbb{R}^d\}$ *over* $\mathcal{X} \subset \mathbb{R}^d$, *the VC-dimension of the set of functions* $\mathbf{1}_{\ell(f(x),y)-\beta \geq 0}$ *induced by* $\mathcal{F} \times (0, B)$ *can be computed as follows. Let* ϕ *be the function that maps* $x \in \mathcal{X}$ *into a vector of all the* d^2 *monomials of degree* 2 *over the components of* x: $\phi_{i+(j-1)d}(x) = x_i x_j$, $i = 1, \ldots, d$, $j = 1, \ldots, d$. *Let* $z = (x, y)$ *and* $\psi(z) = (\phi(x), yx, y^2)$. *Then, for any* $f \in \mathcal{F}$ *and* $\beta \in (0, B)$,

$$\ell(f(x), y) - \beta = (\langle w, x \rangle - y)^2 - \beta = (\langle w, x \rangle)^2 - 2 \langle w, yx \rangle + y^2 - \beta,$$

which is a quadratic function of z *but a linear function of* $\psi(z)$. *Therefore, the VC-dimension of the set of functions* $\mathbf{1}_{\ell(f(x),y)-\beta \geq 0}$ *cannot be larger than the VC-dimension of the set of linear classifiers of* $\mathbb{R}^{d^2+d+1}$ *to which all the* $\psi(z)$ *belong, i.e.,* $d_{VC} \leq d^2 + d + 2$. *Thus, Inequality* (5) *ensures with Theorem 3 that, with probability at least* $1 - \delta$,

$$\forall f \in \mathcal{F}, \quad L_n(f) \leq \hat{L}_n(f) + 2B\sqrt{2\frac{(d^2 + d + 2)\log \frac{2en}{d^2+d+2} + \log \frac{2}{\delta}}{n}}.$$

4 Rademacher complexity-based bounds

We now consider the derivation of risk bounds based on Rademacher complexities instead of VC-dimensions. This also relies on a symmetrization step, which we conduct using a ghost sample defined as in Section 3. In addition, we will also require a bounded difference inequality for dependent variables, which we established in Theorem 1.

Let us first define the relevant capacity measures.

Definition 3 (Rademacher complexities)**.** *Let* $\boldsymbol{T}_n = (T_i)_{1,\leq i \leq n}$ *be a sequence of (not necessarily independent nor identically distributed) random variables* $T_i \in \mathcal{T}$ *and* $\boldsymbol{\sigma}_n = (\sigma_i)_{1 \leq i \leq n}$ *an i.i.d. sequence of uniformly distributed* $\sigma_i \in \{-1, +1\}$. *Let* $\mathcal{F}$ *be a class of real-valued functions over* $\mathcal{T}$. *The* empirical Rademacher complexity *of* $\mathcal{F}$ *given* $\boldsymbol{T}_n$ *is*

$$\hat{\mathcal{R}}_{\boldsymbol{T}_n}(\mathcal{F}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(T_i) \,\middle|\, \boldsymbol{T}_n\right]$$

and the Rademacher complexity *of* $\mathcal{F}$ *is*

$$\mathcal{R}_{\boldsymbol{T}_n}(\mathcal{F}) = \mathbb{E}\hat{\mathcal{R}}_{\boldsymbol{T}_n}(\mathcal{F}) = \mathbb{E}\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(T_i).$$

Note that we keep the subscript $\boldsymbol{T}_n$ in the notation of the Rademacher complexity to emphasize that its definition depends on the distribution of $\boldsymbol{T}_n$. For a sequence of values $\boldsymbol{t}_n = (t_i)_{1 \leq i \leq n} \in \mathcal{T}^n$, we will also occasionally write $\hat{\mathcal{R}}_{\boldsymbol{t}_n}(\mathcal{F})$ with a lowercase $\boldsymbol{t}_n$ in the subscript to refer to the value of the empirical Rademacher complexity computed given $\boldsymbol{T}_n = \boldsymbol{t}_n$.

We can now state the first result of this section (proved in App. F).

Theorem 5 (Rademacher complexity-based bound)**.** *Let $\boldsymbol{Z}_n$ denote a sequence of possibly dependent variables and consider the ghost sample $\boldsymbol{Z}'_n = (Z'_i)_{1\leq i\leq n}$ with each $Z'_i = (X'_i, Y'_i)$ built as an independent copy of $Z_i = (X_i, Y_i)$. Let $\ell : \mathcal{Y}\times\mathcal{Y} \to [0, B]$ be a loss function. Then, for any $\delta \in (0, 1)$, with probability at least $1-\delta$,*

$$\forall f \in \mathcal{F}, \quad L_n(f) \leq \hat{L}_n(f) + \mathcal{R}_{\boldsymbol{Z}_n}(\mathcal{L}) + \mathcal{R}_{\boldsymbol{Z}'_n}(\mathcal{L}) + B\sqrt{\frac{\log\frac{1}{\delta}}{2n}},$$

where $\mathcal{L} = \{\ell_f : \ell_f(z) = \ell(f(x), y),\ f \in \mathcal{F}\}$, and $\mathcal{R}_{\boldsymbol{Z}'_n}(\mathcal{L})$ is the Rademacher complexity of $\mathcal{L}$ defined over the ghost sample $\boldsymbol{Z}'_n$ instead of $\boldsymbol{Z}_n$.

The main difference with classical Rademacher complexity-based bounds of, e.g., Bartlett and Mendelson (2002), is due to the possibility of having different joint distributions for $\boldsymbol{Z}_n$ and $\boldsymbol{Z}'_n$, which prevents us from collapsing $\mathcal{R}_{\boldsymbol{Z}_n}(\mathcal{L}) + \mathcal{R}_{\boldsymbol{Z}'_n}(\mathcal{L})$ into $2\mathcal{R}_{\boldsymbol{Z}_n}(\mathcal{L})$. In particular, this compromises the possibility to obtain data-dependent bounds based on the empirical Rademacher complexities, since the ghost sample is not really available to compute $\hat{\mathcal{R}}_{\boldsymbol{Z}'_n}(\mathcal{L})$.

However, even in the standard i.i.d. case where we can perform such a simplification, it is common practice to upper bound the Rademacher complexity by its empirical version in a worst-case manner, which we reproduce here:

$$\mathcal{R}_{\boldsymbol{Z}_n}(\mathcal{L}) \leq \sup_{\boldsymbol{z}_n\in\mathcal{Z}^n} \hat{\mathcal{R}}_{\boldsymbol{z}_n}(\mathcal{L}) \qquad \text{and} \quad \mathcal{R}_{\boldsymbol{Z}'_n}(\mathcal{L}) \leq \sup_{\boldsymbol{z}_n\in\mathcal{X}^n} \hat{\mathcal{R}}_{\boldsymbol{z}_n}(\mathcal{L}).$$

Therefore, we can obtain a risk bound that does not depend on the ghost sample anymore, and which coincides with common bounds for the i.i.d. case:

Corollary 1. *For any $\delta \in (0, 1)$, with probability at least $1-\delta$,*

$$\forall f \in \mathcal{F}, \quad L_n(f) \leq \hat{L}_n(f) + 2\sup_{\boldsymbol{z}_n\in\mathcal{Z}^n} \hat{\mathcal{R}}_{\boldsymbol{z}_n}(\mathcal{L}) + B\sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

Example 3 (Linear and kernel regression)**.** *Consider the squared loss, $\ell(\hat{y}, y) = (y - \hat{y})^2$, computed from a bounded output $Y \in [-M, M]$ and a clipped prediction $\hat{y} = \min(M, \max(-M, \hat{y}))$ given by a linear model $\hat{y} = f(x)$ from $\mathcal{F}_{lin} = \{f : f(x) = \langle w, x\rangle,\ \|w\| \leq \Lambda\}$. Then, standard computations (detailed in App. G) yield*

$$\sup_{\boldsymbol{z}_n\in\mathcal{Z}^n} \hat{\mathcal{R}}_{\boldsymbol{z}_n}(\mathcal{L}) \leq \frac{4M\Lambda \sup_{x\in\mathcal{X}} \|x\|}{\sqrt{n}}, \tag{6}$$

which can only be improved in data-dependent bounds through the substitution of $\sqrt{\sum_{i=1}^n \|X_i\|^2/n}$ for $\sup_{x\in\mathcal{X}} \|x\|$. If we now consider $\mathcal{F}$ as the ball of radius Λ in the Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ induced by the Gaussian kernel, $K(x, x') = \exp(-\|x - x'\|/2\sigma^2)$, we have instead

$$\sup_{\boldsymbol{z}_n\in\mathcal{Z}^n} \hat{\mathcal{R}}_{\boldsymbol{z}_n}(\mathcal{L}) \leq \frac{4M\Lambda}{\sqrt{n}},$$

which is exactly what the corresponding data-dependent bound would give in an i.i.d. setting.

Instead of relying on worst-case estimates, another possibility is to use an upper bound on the Rademacher complexity expressed in terms of the marginal distributions of the variables Z_i, which coincide with those of the Z_i'.

Theorem 6. *Let $\boldsymbol{Z}_n$ denote a sequence of possibly dependent variables $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$. For a bounded loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, B]$ and a class $\mathcal{F}$ of functions from $\mathcal{X}$ to $\mathcal{Y}$, let $\mathcal{L} = \{\ell_f : \ell_f(z) = \ell(f(x), y),\ f \in \mathcal{F}\}$. If*

$$\mathcal{R}_{\boldsymbol{Z}_n}(\mathcal{L}) \leq \overline{\mathcal{R}}_n(\mathcal{L}) \tag{7}$$

with an upper bound $\overline{\mathcal{R}}_n(\mathcal{L})$ on the Rademacher complexity of $\mathcal{L}$ expressed in terms of the marginal distributions of the Z_i's, then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\forall f \in \mathcal{F}, \quad L_n(f) \leq \hat{L}_n(f) + 2\overline{\mathcal{R}}_n(\mathcal{L}) + B\sqrt{\frac{\log \frac{1}{\delta}}{2n}}.$$

Proof. Since $\overline{\mathcal{R}}_n(\mathcal{L})$ only involves the marginal distributions of the Z_i, which are identical to the ones of the Z_i' by construction of the ghost sample in Theorem 5, the bound in (7) also implies $\mathcal{R}_{\boldsymbol{Z}_n'}(\mathcal{L}) \leq \overline{\mathcal{R}}_n(\mathcal{L})$. Then, both Rademacher complexities in the bound of Theorem 5 can be upper bounded by $\overline{\mathcal{R}}_n(\mathcal{L})$, which yields the desired result. □

Example 4 (Linear and kernel regression—continued)**.** *Consider again the squared loss with the class $\mathcal{F}_{lin}$ for linear regression. Then, the classically used upper bound on the Rademacher complexity is precisely of a form suitable for Theorem 6:*

$$\mathcal{R}_{\boldsymbol{Z}_n}(\mathcal{L}) \leq \frac{4M\Lambda\sqrt{\sum_{i=1}^n \mathbb{E}\|X_i\|^2}}{n} = \overline{\mathcal{R}}_n(\mathcal{L}),$$

and, since $X_i' \sim X_i$,

$$\mathcal{R}_{\boldsymbol{Z}_n'}(\mathcal{L}) \leq \frac{4M\Lambda\sqrt{\sum_{i=1}^n \mathbb{E}\|X_i'\|^2}}{n} = \frac{4M\Lambda\sqrt{\sum_{i=1}^n \mathbb{E}\|X_i\|^2}}{n} = \overline{\mathcal{R}}_n(\mathcal{L}).$$

Thus, we obtain that, with probability at least $1 - \delta$,

$$\forall f \in \mathcal{F}_{lin}, \quad L_n(f) \leq \hat{L}_n(f) + \frac{8M\Lambda\sqrt{\sum_{i=1}^n \mathbb{E}\|X_i\|^2}}{n} + 4M^2\sqrt{\frac{\log \frac{1}{\delta}}{2n}},$$

which coincides with the standard result for linear regression with i.i.d. data. In addition, similar results can also be derived for classes of kernel models with $\mathbb{E}\|X_i\|^2$ replaced by $\mathbb{E}K(X_i, X_i)$, which only depends on the marginal distribution of X_i.

Example 5 (Margin classifiers)**.** *Binary margin classifiers are classifiers that implement $f(x) = sign(g(x))$ with a real-valued function $g \in \mathcal{G}$. For those, the piecewise linear margin loss function is usually considered: for a margin $\gamma > 0$,*

$$\ell_\gamma(g(x), y) = \min\{\gamma, \max\{0,\ (1 - yg(x))/\gamma\}.$$

Then, the contraction principle (Ledoux and Talagrand, 1991) yields $\mathcal{R}_{\boldsymbol{Z}_n}(\mathcal{L}) \leq \mathcal{R}_{\boldsymbol{Z}_n}(\mathcal{G})/\gamma$ and an upper bound on $\mathcal{R}_{\boldsymbol{Z}_n}(\mathcal{G})$ in terms of the marginal

distributions also leads to a bound on $\mathcal{R}_{\boldsymbol{Z}_n}(\mathcal{L})$ *suitable for Theorem 6. For linear classifiers,* $\mathcal{G} = \{g : g(x) = \langle w, x\rangle\,,\ \|w\| \leq \Lambda\}$*, we can use the bound of Example 4 to obtain*

$$\overline{\mathcal{R}}_n(\mathcal{L}) = \frac{\Lambda\sqrt{\sum_{i=1}^n \mathbb{E}\|X_i\|^2}}{\gamma n},$$

and a bound on the probability of misclassification via $P(f(X) \neq Y) = \mathbb{E}\mathbf{1}_{f(X)\neq Y} \leq \mathbb{E}\ell_\gamma(g(X), Y) = L_n(f)$*, where* $L_n(f)$ *is bounded by Theorem 6. Of course, similar results hold for kernel machines by exchanging* $\mathbb{E}\|X_i\|^2$ *with* $\mathbb{E}K(X_i, X_i)$*.*

Theorem 6 applies also more broadly to other nonconvex loss functions for which the contraction principle does not apply as directly as in Example 5. Note that, for such losses, the empirical risk minimizer remains most often elusive and the obtention of *uniform* risk bounds is critical to allow for their application to the model returned by a training algorithm.

Example 6 (Vector quantization)**.** *Consider the problem of vector quantization in* $\mathcal{X} \subset \mathcal{H}$ *for some Hilbert space* $\mathcal{H}$ *as discussed in Bartlett et al. (1998); Biau et al. (2008), where one aims at finding a collection* $f = (f_k)_{1\leq k\leq C} \in \mathcal{H}^C$ *of* C *codepoints* f_k *that can well approximate the observations of* $X_i \in \mathcal{X}$*. The framework of Sect. 2.1 can be modified in the obvious manner to account for this setting by letting* $Z_i = X_i$ *and computing the loss from* $\mathcal{X}^2 \to [0, B]$ *instead of* $\mathcal{Y}^2 \to [0, B]$*. For nearest neighbors quantizers, the error of the model* f *is computed by the loss function*

$$\ell(f, x) = \min_{k\in\{1,\dots,C\}} \|x - f_k\|^2,$$

and the risk (1) is the so-called distortion *of* f*. Risk bounds in this setting were derived for the i.i.d. scenario via the Rademacher complexity in Biau et al. (2008); Lauer (2020b), and their computations led to bounds in terms of the marginals. For instance, for* $\mathcal{F} = \mathcal{F}_0^C$*,* $\mathcal{F}_0 = \{f_0 \in \mathcal{H} : \|f_0\|_{\mathcal{H}} \leq \Lambda\}$*, we have*

$$\mathcal{R}_{\boldsymbol{X}_n}(\mathcal{L}) \leq \frac{2C\Lambda\sqrt{\sum_{i=1}^n \mathbb{E}\|X_i\|^2}}{n} + \frac{C\Lambda^2}{\sqrt{n}} = \overline{\mathcal{R}}_n(\mathcal{L})$$

and therefore a risk bound that applies in the non-i.i.d. setting via Theorem 6.

Other examples, such as in switching regression or subspace clustering (Lauer, 2020a,b), can be found and lead to the same conclusion: in all these settings, common bounds on the Rademacher complexity is expressed in terms of the marginal distributions, and can thus be used verbatim to produce risk bounds with dependent data. For switching regression, this leads to guarantees on the accuracy of models in switched system identification (Lauer and Bloch, 2019) that are tighter than those derived in Massucci et al. (2022) with mixing arguments.

More generally, Theorem 6 can be compared with the results of Mohri and Rostamizadeh (2009) that provide generic risk bounds based on Rademacher complexities and mixing arguments, used for instance by Massucci et al. (2022). For a training sequence $\boldsymbol{Z}_n$ generated by a β-mixing process of mixing coefficient $\beta(a)$ that measures the degree of dependence between data

points separated by a time steps (see Bradley (2005) for proper definitions of these terms), and in a setting otherwise similar to Theorem 6, Mohri and Rostamizadeh (2009) provides the bound:

$$\forall f \in \mathcal{F}, \quad L_n(f) \leq \hat{L}_n(f) + 2\overline{\mathcal{R}}_\mu(\mathcal{L}) + B\sqrt{\frac{\log \frac{1}{\delta - 4(\mu-1)\beta(a)}}{2\mu}}, \tag{8}$$

where a and μ are positive integers such that $n = 2a\mu$, and δ must satisfy $\delta > 4(\mu-1)\beta(a)$. Theorem 6 offers several advantages over the mixing bound (8). First, the mixing bound involves the so-called "effective sample size" $\mu = n/2a < n$ instead of n itself, which makes it less tight, even when focusing only on the complexity term of the bound and $\overline{\mathcal{R}}_\mu(\mathcal{L})$ that is computed on a subset of μ points. Second, it is only applicable with values of the confidence index δ larger than a quantity that grows with μ, which prevents us from reaching the "practical certainty" of $\delta \approx 10^{-9}$ that is often needed in applications such as scenario-based optimization. Finally, the last term of (8) involves the mixing coefficient $\beta(a)$ that is often unknown and difficult to estimate from the data itself (see McDonald et al. (2015) for a discussion of this topic).

4.1 Chaining and covering numbers

Chaining (Talagrand, 2014) is a generic method for estimating the empirical Rademacher complexity of a function class from its covering numbers, defined in terms of an empirical pseudo-metric.

Definition 4 (Pseudo-metric)**.** *Given a sequence $\boldsymbol{t}_n \in \mathcal{T}^n$, $d_{2,\boldsymbol{t}_n}$ is the empirical pseudo-metric over the set of functions from $\mathcal{T}$ to $\mathbb{R}$ defined by*

$$d_{2,\boldsymbol{t}_n}(f, f') = \left(\frac{1}{n}\sum_{i=1}^{n} |f(t_i) - f'(t_i)|^2\right)^{\frac{1}{2}}.$$

Definition 5 (Covering numbers)**.** *Given a class $\mathcal{F}$ of functions of $\mathcal{T} \to \mathbb{R}$ and a pseudo-metric ρ, the* covering number *$\mathcal{N}(\epsilon, \mathcal{F}, \rho)$ at scale ϵ of $\mathcal{F}$ for the distance ρ is the smallest cardinality of the proper ϵ-net $\mathcal{H} \subseteq \mathcal{F}$ of $\mathcal{F}$ such that $\forall f \in \mathcal{F}$, $\rho(f, \mathcal{H}) < \epsilon$.* Uniform covering numbers *are defined for the pseudo-metric of Definition 4 by*

$$\mathcal{N}_2(\epsilon, \mathcal{F}, n) = \sup_{\boldsymbol{t}_n \in \mathcal{T}^n} \mathcal{N}(\epsilon, \mathcal{F}, d_{2,\boldsymbol{t}_n}).$$

Theorem 7 (Chaining)**.** *Let $\mathcal{F}$ be a real-valued function class over $\mathcal{T}$ and, for any $\boldsymbol{t}_n \in \mathcal{T}^n$, let $D_{\mathcal{F}} = \sup_{(f,f') \in \mathcal{F}^2} d_{2,\boldsymbol{t}_n}(f, f')$ denote its diameter. Then, for any integer $N > 0$,*

$$\hat{\mathcal{R}}_{\boldsymbol{t}_n}(\mathcal{F}) \leq \frac{D_{\mathcal{F}}}{2^N} + 6D_{\mathcal{F}} \sum_{j=1}^{N} 2^{-j} \sqrt{\frac{\log \mathcal{N}(D_{\mathcal{F}} 2^{-j}, \mathcal{F}, d_{2,\boldsymbol{t}_n})}{n}}.$$

As we have seen in the examples above, the Rademacher complexity of the loss class $\mathcal{L}$ can usually be bounded in terms of the one of the function class

$\mathcal{F}$ via contraction arguments that merely introduce a factor L_φ.[3] Thus, if we let $C(\boldsymbol{X}_n, \mathcal{F})$ denote the bound on the empirical Rademacher complexity of $\mathcal{F}$ given by Theorem 7, then after taking expectation and using Jensen's inequality, we obtain

$$\begin{aligned}\mathcal{R}_{\boldsymbol{Z}_n}(\mathcal{L}) &\leq L_\varphi \mathbb{E} C(\boldsymbol{X}_n, \mathcal{F}) \\ &= \frac{L_\varphi D_\mathcal{F}}{2^N} + 6 L_\varphi D_\mathcal{F} \sum_{j=1}^{N} 2^{-j} \sqrt{\frac{\log \mathbb{E}\mathcal{N}(D_\mathcal{F} 2^{-j}, \mathcal{F}, d_{2,\boldsymbol{Z}_n})}{n}}, \end{aligned} \tag{9}$$

and a bound on the Rademacher complexity given in terms of expected covering numbers. This bound thus depends on the distribution of $\boldsymbol{Z}_n$ and would differ on the ghost sample $\boldsymbol{Z}'_n$. However, in many cases, the covering numbers are merely bounded by their uniform (worst-case) counterpart, i.e.,

$$\mathbb{E}\mathcal{N}(D_\mathcal{F} 2^{-j}, \mathcal{F}, d_{2,\boldsymbol{X}_n}) \leq \sup_{\boldsymbol{x}_n \in \mathcal{X}^n} \mathcal{N}(D_\mathcal{F} 2^{-j}, \mathcal{F}, d_{2,\boldsymbol{x}_n}) = \mathcal{N}_2(D_\mathcal{F} 2^{-j}, \mathcal{F}, n),$$

which are themselves bounded in terms of the fat-shattering dimension (see, e.g., Alon et al. (1997); Mendelson (2002); Mendelson and Vershynin (2003)). Then, the resulting bound,

$$\overline{\mathcal{R}}_n(\mathcal{F}) = \frac{L_\varphi D_\mathcal{F}}{2^N} + 6 L_\varphi D_\mathcal{F} \sum_{j=1}^{N} 2^{-j} \sqrt{\frac{\log \mathcal{N}_2(D_\mathcal{F} 2^{-j}, \mathcal{F}, n)}{n}},$$

on the Rademacher complexity on the training sample applies similarly to the one on the ghost sample, $\mathcal{R}_{\boldsymbol{Z}'_n}(\mathcal{F})$, and this leads, with Theorem 6, to the same result one would have obtained under an i.i.d. assumption.

In addition, if one is reluctant to use the uniform (worst-case) covering numbers, many bounds in the literature on the expected covering numbers are given in terms of the marginal distributions.

For instance, in the multi-category classification setting, Bartlett et al. (2017) derived a covering numbers bound for a class $\mathcal{L}$ of margin loss functions induced by a class of spectrally-regularized deep neural networks, which is of the form

$$\log \mathcal{N}(\epsilon, \mathcal{L}, d_{2,\boldsymbol{Z}_n}) \leq \frac{A \sum_{i=1}^{n} \|X_i\|^2}{\epsilon^2}.$$

This directly yields a bound on the expected log covering numbers that could be used in (9) and that is expressed in terms of the marginal distributions, and thus is valid for both the training and the ghost samples.[4]

[3] The factor L_φ is the Lipschitz constant of the univariate function φ used to compute the loss $\ell(f(x), y) = \varphi(u)$, where u is for instance $yg(x)$ for margin classification or $f(x) - y$ for regression.

[4] Note that when formulating (9), we could have applied Jensen's inequality only to the square root and express the bound in terms of expected log coverging numbers.

5 Application to scenario-based optimization

A robust optimization program can typically be written as

$$\begin{aligned} &\min_{\theta\in\Theta} \; J(\theta) \\ &\text{s.t. } f(x,\theta) \leq 0, \quad \forall x \in \mathcal{X}, \end{aligned} \tag{10}$$

with an infinite number of constraints induced by the set $\mathcal{X}$ that accounts for the uncertainties in the parameters of the problem. Scenario-based optimization computes solutions to (10) with probabilistic guarantees of feasibility by solving random programs of the form

$$\begin{aligned} &\hat{\theta} \in \operatorname*{argmin}_{\theta\in\Theta} \; J(\theta) \\ &\text{s.t. } f(X_i,\theta) \leq 0, \quad i = 1,\dots,n, \end{aligned} \tag{11}$$

where the uncertainties are sampled in order to yield a finite number of constraints. Note that in order to avoid confusion, we kept the notations of the rest of the paper and denote the random quantity as X and the optimization variable as θ (instead of x as is usually done in the robust optimization literature). Here, $\boldsymbol{X}_n = (X_i)_{1\leq i\leq n}$ is a random sample of "scenarios" $X_i \in \mathcal{X}$ and Θ is a general and deterministic feasible set (think of $\Theta = \mathbb{R}^d$ for instance).

The two critical issues in scenario-based optimization are to estimate the probability of violation of the computed solution,

$$L_n(\hat{\theta}) = P\{f(X_1,\hat{\theta}) > 0\},$$

where we assume stationarity of $\boldsymbol{X}_n$, and the number of scenarios one should sample to guarantee a certain reliability, i.e., the sample complexity of (11), defined for given ϵ, $\delta \in (0,1)$ as the smallest n such that

$$P\{L_n(\hat{\theta}) > \epsilon\} \leq \delta \tag{12}$$

holds.

For independent scenarios, Alamo et al. (2009) proposed upper bounds on the probability of violation and sample complexity estimates based on the VC-dimension of the set of indicator functions

$$\mathcal{I} = \{I_\theta \in \{0,1\}^{\mathcal{X}} : I_\theta(x) = \mathbf{1}_{f(x,\theta)>0},\ \theta \in \Theta\}. \tag{13}$$

Random programs for which the VC-dimension cannot be bounded can be handled with other results from Lauer (2023) that are instead based on the Rademacher complexity of the class of margin loss functions,

$$\mathcal{L}_{\gamma,\Theta} = \{\ell_\theta \in [0,1]^{\mathcal{X}} : \ell_\theta(x) = \min\{1, \max\{0, 1 + f(x,\theta)/\gamma\}\},\ \theta \in \Theta\}, \tag{14}$$

for a margin $\gamma > 0$.

We generalize these results to the case of dependent scenarios below.

5.1 Sample complexity of random programs with dependent scenarios and finite VC-dimension

We now consider the case of dependent scenarios, i.e., when $\boldsymbol{X}_n$ is a sample of dependent variables. Building on the results of Section 3, we obtain the following sample complexity, which we only state for the zero-error case that is rather standard in the framework of scenario optimization.

Theorem 8. *Let $\boldsymbol{X}_n$ denote a stationary sequence of random variables of values in $\mathcal{X}^n$. Let d_{VC} denote the VC-dimension of the set (13). Assume that, for any $\boldsymbol{x}_n \in \mathcal{X}^n$, the random program (11) is feasible and that an algorithm computes a feasible point $\hat{\theta}$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$P\{f(X_1, \hat{\theta}) > 0\} \leq \frac{4 \log \Pi_{\mathcal{I}}(2n) + \log \frac{4}{\delta}}{n} \leq \frac{4 d_{VC} \log(\frac{2en}{d_{VC}}) + \log \frac{4}{\delta}}{n},$$

and, for any $\epsilon \in (0,1)$, $\delta \in (0,1)$, the sample complexity of (11) is no more than

$$n(\epsilon, \delta) = \frac{5}{\epsilon}\left(d_{VC} \log \frac{40}{\epsilon} + \log \frac{4}{\delta}\right).$$

Proof. Theorem 4 can be reformulated for the loss $\ell(\theta, x) = \mathbf{1}_{f(x,\theta)>0}$ where $f \in \mathcal{F}$ is identified with $\theta \in \Theta$ and the growth function and VC-dimension of $\mathcal{F}$ are replaced by those of $\mathcal{I}$ (13). This yields the bound on the probability of violation above under the assumption that $\hat{\theta}$ is a feasible point for (11), i.e., $\hat{L}_n(\hat{\theta}) = 0$.

Then, (12) holds as soon as

$$\frac{4 \log \Pi_{\mathcal{I}}(2n) + \log \frac{4}{\delta}}{n} \leq \epsilon,$$

which is implied by

$$4\left(\frac{2en}{d_{VC}}\right)^{d_{VC}} \exp(-n\epsilon/4) \leq \delta.$$

Using Theorem 6 in Alamo et al. (2009), this can be ensured by setting

$$n \geq \inf_{\mu>1} \frac{4}{\epsilon} \frac{\mu}{\mu - 1}\left(\log \frac{4}{\delta} + d_{VC} \log \frac{8\mu}{\epsilon}\right),$$

in which the suboptimal choice $\mu = 5$ suffices to conclude. □

Here, Theorem 8 yields the same guarantees as Alamo et al. (2009) for scenario optimization, but applies more broadly to dependent scenarios.

5.2 Margin-based sample complexity of pseudo-linear random programs

For random programs with infinite (or merely too large) VC-dimension, we instead follow the margin-based approach of Lauer (2023) which, given a margin

parameter $\gamma > 0$, implements the following random program instead of (11):

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \Theta} \; J(\theta) \qquad (15)$$
$$\text{s.t. } f(X_i, \theta) \leq -\gamma, \quad i = 1, \ldots, n.$$

More specifically, we focus here on problems in which the constraint function is of the form

$$f(x, \theta) = \max_{k \in \{1,\ldots,C\}} f_k(x, \theta) \qquad (16)$$

for a collection of C functions

$$f_k(x, \theta) = \psi_k(x)^\top \phi_k(\theta) + \eta_k(x), \quad k = 1, \ldots, C, \qquad (17)$$

based on functions $\psi_k : \mathcal{X} \to \mathbb{R}^{n_k}$, $\phi_k : \Theta \to \mathbb{R}^{n_k}$ and $\eta_k : \mathcal{X} \to \mathbb{R}$.[5] Then, combining the contraction principle (with Lipschitz constant $1/\gamma$) and Theorem 2 in Lauer (2023) yields

$$\mathcal{R}_{\boldsymbol{X}_n}(\mathcal{L}_{\gamma,\Theta}) \leq \overline{\mathcal{R}}_n(\mathcal{L}_{\gamma,\Theta}) = \frac{1}{\gamma} \sum_{k=1}^{C} \frac{\tau_k \Lambda_k}{\sqrt{n}}, \qquad (18)$$

where $\mathcal{L}_{\gamma,\Theta}$ is as in (14), $\tau_k = \sup_{x \in \mathcal{X}} \|\psi_k(x)\|$, and $\Lambda_k = \sup_{\theta \in \Theta} \|\phi_k(\theta)\|$. Note that this bound holds similarly for any ghost sample $\boldsymbol{X}'_n \in \mathcal{X}^n$, which leads to the following.

Theorem 9. *Let $\boldsymbol{X}_n$ denote a stationary sequence of random variables of values in $\mathcal{X}^n$. Let $\tau_k = \sup_{x \in \mathcal{X}} \|\psi_k(x)\|$ and $\Lambda_k = \sup_{\theta \in \Theta} \|\phi_k(\theta)\|$. Fix the margin parameter $\gamma > 0$ and assume that, for any $\boldsymbol{x}_n \in \mathcal{X}^n$, the random program (15) is feasible and that an algorithm computes a feasible point $\hat{\theta}$ satisfying the margin condition $f(x_i, \hat{\theta}_i) < -\gamma$, $i = 1, \ldots, n$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$P\{f(X, \hat{\theta}) > 0\} \leq \frac{2}{\gamma} \sum_{k=1}^{C} \frac{\tau_k \Lambda_k}{\sqrt{n}} + \sqrt{\frac{\log \frac{1}{\delta}}{2n}},$$

and, for any $\epsilon \in (0, 1)$, $\delta \in (0, 1)$, the sample complexity of (11) is no more than

$$n(\epsilon, \delta) = \frac{1}{\epsilon^2} \left(\frac{2}{\gamma} \sum_{k=1}^{C} \tau_k \Lambda_k + \sqrt{\log \frac{1}{\delta}} \right)^2.$$

Proof. Apply Theorem 5 to the loss class (14) and use (18) to bound the two Rademacher complexities. Then, the bound on the probability of violation follows from the feasibility of $\hat{\theta}$, that ensures that $\hat{L}_n(\hat{\theta})$ computed with the margin loss $\ell_\gamma(\theta, x) = \min\{1, \max\{0, 1 + f(x, \theta)/\gamma\}\}$ is zero, and the fact that $P\{f(X, \hat{\theta}) > 0\} \leq P\{f(X, \hat{\theta}) > -\gamma\} = L_n(\hat{\theta})$.

Then, (12) holds as soon as

$$\frac{2}{\gamma} \sum_{k=1}^{C} \frac{\tau_k \Lambda_k}{\sqrt{n}} + \sqrt{\frac{\log \frac{1}{\delta}}{2n}} \leq \epsilon,$$

which is ensured whenever $n \geq n(\epsilon, \delta)$. □

[5] The results actually hold for a more general form of composition in (16), as detailed in Lauer (2023).

Again, the proposed method yields with Theorem 9 the same guarantees as Lauer (2023) for scenario optimization, but in a more general setting that allows the sampling of dependent scenarios.

6 Conclusions

This paper explored the possible extensions of standard risk bounds to dependent data. This resulted in a number of uniform risk bounds applicable to a wide range of problems. Compared with other approaches from the literature for deriving uniform bounds without assuming independence, the proposed method provides tighter bounds and does not rely on a mixing condition difficult to verify in practice.

Yet, one drawback of the proposed approach is the lack of data-dependent counterparts of the bounds. However, as seen on a number of examples, data-dependent bounds do not offer improved guarantees in several classical cases. In addition, regarding the motivating application of scenario-based optimization, where the focus is on the sample complexity, data-dependent bounds are irrelevant. In this context, we showed that standard results so far limited to the i.i.d. case also hold with dependent scenarios. This provides the basis for a refined analysis of several problems in control theory such as that of Wang et al. (2021), and the basic tool for investigating new sampling strategies in robust optimization.

Another open issue arises in scenario-based optimization, where the standard bounds for *convex* random programs (Campi and Garatti, 2008; Calafiore, 2010) heavily rely on a binomial tail bound and independence. As these follow a proof scheme rather different than that of learning theory bounds, the current work does not allow their extension to dependent scenarios, which would allow for a straightforward strengthening of the results of Wang et al. (2021).

A Proof of Theorem 1

Let $(W_i)_{1\leq i\leq n}$ be a random sequence adapted to $(\mathcal{A}_i)$, i.e., a sequence of real-valued $\mathcal{A}_i$-measurable random variables. Recall that an $\mathcal{A}_{i-1}$-measurable variable is said to be *predictable*. Define

$$S_n = \sum_{i=1}^{n} W_i.$$

We will first state a few intermediate results on generic variables W_i that are necessary for the proof of Theorem 1.

Lemma 3 (Hoeffding's lemma, slightly extended)**.** *For predictable random variables $L_i \leq W_i \leq U_i$ with $\mathbb{E}[W_i \mid \mathcal{A}_{i-1}] = 0$,*

$$\mathbb{E}[e^{\beta W_i} \mid \mathcal{A}_{i-1}] \leq e^{\beta^2 (U_i - L_i)^2/8}.$$

Proof. Replace the constant bounds by L_i, U_i and the expectation by the conditional expectation in the standard proof of Hoeffding (1963): by convexity

of the exponential,

$$e^{\beta W_i} \leq \frac{U_i - W_i}{U_i - L_i} e^{\beta L_i} + \frac{W_i - L_i}{U_i - L_i} e^{\beta U_i}$$

and

$$\mathbb{E}[e^{\beta W_i} \mid \mathcal{A}_{i-1}] \leq \mathbb{E}\left[\frac{U_i - W_i}{U_i - L_i} e^{\beta L_i} \middle| \mathcal{A}_{i-1} \right] + \mathbb{E}\left[\frac{W_i - L_i}{U_i - L_i} e^{\beta U_i} \middle| \mathcal{A}_{i-1} \right].$$

Then, use the fact that the L_i and U_i are predictable and $\mathbb{E}[W_i \mid \mathcal{A}_{i-1}] = 0$:

$$\begin{aligned}
\mathbb{E}[e^{\beta W_i} \mid \mathcal{A}_{i-1}] &\leq \frac{U_i - \mathbb{E}[W_i \mid \mathcal{A}_{i-1}]}{U_i - L_i} e^{\beta L_i} + \frac{\mathbb{E}[W_i \mid \mathcal{A}_{i-1}] - L_i}{U_i - L_i} e^{\beta U_i} \\
&= \frac{U_i}{U_i - L_i} e^{\beta L_i} + \frac{-L_i}{U_i - L_i} e^{\beta U_i} \\
&= e^{h(\beta(U_i - L_i))}
\end{aligned}$$

with $h(u) = uL_i/(U_i - L_i) + \log(1 + L_i[(1 - e^u)/(U_i - L_i)])$. By computing derivatives and the Taylor expansion, we know that $h(u) \leq u^2/8$, which completes the proof. □

Lemma 4 (Lemma 2.4 in van De Geer (2007))**.** *Assume $\mathbb{E}[W_i \mid \mathcal{A}_{i-1}] = 0$ and $L_i \leq W_i \leq U_i$ with L_i and U_i that are predictable (i.e., $\mathcal{A}_{i-1}$-measurable). For any $\beta > 0$, the sequence of random variables*

$$\zeta_n(\beta) = \exp\left(\beta S_n - \beta^2 \sum_{i=1}^{n} (U_i - L_i)^2/8 \right), \quad n = 1, 2, \ldots,$$

is a supermartingale, i.e., $\mathbb{E}[\zeta_n(\beta) \mid \mathcal{A}_{n-1}] \leq \zeta_{n-1}(\beta)$.

Proof.

$$\begin{aligned}
&\mathbb{E}[\zeta_n(\beta) \mid \mathcal{A}_{n-1}] \\
&= \mathbb{E}\left[\exp\left(\beta \sum_{i=1}^{n} (W_i - \beta(U_i - L_i)^2/8) \right) \mid \mathcal{A}_{n-1} \right] \\
&= \mathbb{E}\left[\exp\left(\beta(W_n - \beta(U_n - L_n)^2/8) \right) \prod_{i=1}^{n-1} \exp\left(\beta(W_i - \beta(U_i - L_i)^2/8) \right) \mid \mathcal{A}_{n-1} \right]
\end{aligned}$$

Here, the assumed predictability of L_i, U_i implies that, for any function f, $\forall i \leq n-1$, $\mathbb{E}[f(W_i, L_i, U_i) \mid \mathcal{A}_{n-1}] = f(W_i, L_i, U_i)$. Thus,

$$\begin{aligned}
\mathbb{E}[\zeta_n(\beta) \mid \mathcal{A}_{n-1}] &= \left(\prod_{i=1}^{n-1} \exp\left(\beta(W_i - \beta(U_i - L_i)^2/8) \right) \right) \\
&\quad \times \mathbb{E}\left[\exp\left(\beta(W_n - \beta(U_n - L_n)^2/8) \right) \mid \mathcal{A}_{n-1} \right] \\
&\leq \prod_{i=1}^{n-1} \exp\left(\beta(W_i - \beta(U_i - L_i)^2/8) \right) \\
&= \exp\left(\beta \sum_{i=1}^{n-1} (W_i - \beta(U_i - L_i)^2/8) \right) \\
&= \zeta_{n-1}(\beta),
\end{aligned}$$

where we used Lemma 3 for the inequality. □

Theorem 10 (Hoeffding inequality for sums—simplified version of Theorem 2.5 in van De Geer (2007) with fixed n)**.** *Assume that $L_i \leq W_i \leq U_i$ for predictable random variables L_i and U_i, i.e., they are $\mathcal{A}_{i-1}$-measurable. Also assume that $\mathbb{E}[W_i \mid \mathcal{A}_{i-1}] = 0$. Then, for any $t > 0$ and $c > 0$,*

$$P\left\{S_n \geq t \text{ and } \sum_{i=1}^{n}(U_i - L_i)^2 \leq c^2\right\} \leq \exp(-2t^2/c^2).$$

Proof. By Lemma 4, $\zeta_n(\beta)$ is a supermartingale. Thus, for any n,

$$\mathbb{E}\zeta_n(\beta) = \mathbb{E}\mathbb{E}[\zeta_n(\beta) \mid \mathcal{A}_{n-1}] \leq \mathbb{E}\zeta_{n-1}(\beta)$$

and, by induction,

$$\begin{aligned}\mathbb{E}\zeta_n(\beta) \leq \mathbb{E}\zeta_1(\beta) &= \mathbb{E}\exp(\beta S_1 - \beta^2(U_1 - L_1)^2/8) = \mathbb{E}\exp(\beta W_1 - \beta^2(U_1 - L_1)^2/8) \\ &= \mathbb{E}[\exp(\beta W_1 - \beta^2(U_1 - L_1)^2/8) \mid \mathcal{A}_0] \qquad (\textit{since } \mathcal{A}_0 = \emptyset) \\ &\leq 1,\end{aligned}$$

where we used Lemma 3 for the last inequality.

Now consider the event

$$A = \left\{S_n > t \text{ and } \sum_{i=1}^{n}(U_i - L_i)^2 \leq c^2\right\}.$$

Then, since $\zeta_n(\beta)\mathbf{1}_A \leq \zeta_n(\beta)$,

$$\mathbb{E}[\zeta_n(\beta)\mathbf{1}_A] \leq \mathbb{E}[\zeta_n(\beta)] \leq 1.$$

But on A, we have

$$\zeta_n(\beta) = \exp\left(\beta S_n - \beta^2 \sum_{i=1}^{n}(U_i - L_i)^2/8\right) \geq \exp(\beta t - \beta^2 c^2/8),$$

so that

$$\begin{aligned}\mathbb{E}[\zeta_n(\beta)\mathbf{1}_A] \geq \mathbb{E}[\exp(\beta t - \beta^2 c^2/8)\mathbf{1}_A] &= \exp(\beta t - \beta^2 c^2/8)\mathbb{E}[\mathbf{1}_A] \\ &= \exp(\beta t - \beta^2 c^2/8)P(A).\end{aligned}$$

Thus,

$$P(A) \leq \frac{1}{\exp(\beta t - \beta^2 c^2/8)} = \exp(-\beta t + \beta^2 c^2/8).$$

Now set $\beta = 4t/c^2$, this yields

$$P(A) \leq \exp(-4t^2/c^2 + 16t^2/8c^2) = \exp(-2t^2/c^2).$$

□

Now, we are ready to state the proof of Theorem 1.

Let us define the random variables $G = g(Z_1, \ldots, Z_n)$ and $W_i = \mathbb{E}[G \mid \mathcal{A}_i] - \mathbb{E}[G \mid \mathcal{A}_{i-1}]$. Then,

$$\begin{aligned} g(Z_1, \ldots, Z_n) - \mathbb{E}g(Z_1, \ldots, Z_n) &= G - \mathbb{E}G \\ &= \mathbb{E}[G \mid \mathcal{A}_n] - \mathbb{E}[G \mid \mathcal{A}_0] \\ &= \sum_{i=1}^{n} \mathbb{E}[G \mid \mathcal{A}_i] - \mathbb{E}[G \mid \mathcal{A}_{i-1}] \\ &= \sum_{i=1}^{n} W_i. \end{aligned}$$

Therefore, the theorem is just an application of Theorem 10 to $S_n = \sum_{i=1}^{n} W_i$, after checking that

$$\begin{aligned} \mathbb{E}[W_i \mid \mathcal{A}_{i-1}] &= \mathbb{E}\left[\mathbb{E}[G \mid \mathcal{A}_i] \mid \mathcal{A}_{i-1}\right] - \mathbb{E}\left[\mathbb{E}[G \mid \mathcal{A}_{i-1}] \mid \mathcal{A}_{i-1}\right] \\ &= \mathbb{E}[G \mid \mathcal{A}_{i-1}] - \mathbb{E}[G \mid \mathcal{A}_{i-1}] = 0 \end{aligned}$$

(due to the tower property, since $\mathcal{A}_{i-1} \subset \mathcal{A}_i$). Actually, to apply Theorem 10, we also need predictable ($\mathcal{A}_{i-1}$-measurable) bounds on W_i. First note that

$$L_i = \inf_z \mathbb{E}\left[g(Z_1, \ldots, Z_{i-1}, z, Z_{i+1}, \ldots, Z_n) \mid \mathcal{A}_i\right]$$

and

$$U_i = \sup_z \mathbb{E}\left[g(Z_1, \ldots, Z_{i-1}, z, Z_{i+1}, \ldots, Z_n) \mid \mathcal{A}_i\right]$$

are $\mathcal{A}_{i-1}$-measurable (since Z_i is replaced by an arbitrary z in their definition) and satisfy

$$L_i \leq \mathbb{E}[G \mid \mathcal{A}_i] \leq U_i,$$

while the bounded difference assumption ensures that $U_i - L_i \leq c_i$. So, since $W_i = \mathbb{E}[G \mid \mathcal{A}_i] - \mathbb{E}[G \mid \mathcal{A}_{i-1}]$, we have the following predictable bounds:

$$\tilde{L}_i = L_i - \mathbb{E}[G \mid \mathcal{A}_{i-1}] \leq W_i \leq U_i - \mathbb{E}[G \mid \mathcal{A}_i] = \tilde{U}_i$$

with $\tilde{U}_i - \tilde{L}_i \leq c_i$. Thus, with $c^2 = \sum_{i=1}^{n} c_i^2$, $\sum_{i=1}^{n} (\tilde{U}_i - \tilde{L}_i)^2 \leq c^2$ always holds and, by Theorem 10: for any $t > 0$,

$$P(S_n > t) = P\left\{S_n > t \text{ and } \sum_{i=1}^{n} (\tilde{U}_i - \tilde{L}_i)^2 \leq c^2\right\} \leq \exp(-2t^2/c^2).$$

Recalling that $S_n = g(Z_1, \ldots, Z_n) - \mathbb{E}g(Z_1, \ldots, Z_n)$ and choosing $t = \epsilon$ completes the proof.

B Proof of Lemma 1

Define the events

$$A(f) = \left\{L_n(f) - \hat{L}_n(f) \geq \epsilon\right\}, \qquad A = \left\{\sup_{f \in \mathcal{F}} L_n(f) - \hat{L}_n(f) \geq \epsilon\right\}$$

$$B(f) = \left\{ \hat{L}'_n(f) - \hat{L}_n(f) \geq \frac{\epsilon}{2} \right\}, \qquad B = \left\{ \sup_{f \in \mathcal{F}} \hat{L}'_n(f) - \hat{L}_n(f) \geq \frac{\epsilon}{2} \right\}.$$

Let $f^* \in \mathcal{F}$ denote the function that depends solely on $\boldsymbol{Z}_n$ and such that[6]

$$L_n(f^*) - \hat{L}_n(f^*) = \sup_{f \in \mathcal{F}} L_n(f) - \hat{L}_n(f).$$

Then, since $f^* \in \mathcal{F}$, we have

$$P(B) \geq P\{B(f^*)\}.$$

For any real numbers a, b, c,

$$(c-a) \geq \epsilon \ \wedge \ (b-c) \geq \frac{-\epsilon}{2} \quad \Rightarrow \quad b - a \geq \frac{\epsilon}{2},$$

and thus

$$P(b - a \geq \epsilon/2) \geq P(c - a \geq \epsilon, \ b - c \geq -\epsilon/2).$$

Applied with $a = \hat{L}_n(f^*)$, $b = \hat{L}'_n(f^*)$ and $c = L_n(f^*)$, this gives:

$$\begin{aligned}
P\{B(f^*)\} &\geq P\left\{ L_n(f^*) - \hat{L}_n(f^*) \geq \epsilon, \ \hat{L}'_n(f^*) - L_n(f^*) \geq \frac{-\epsilon}{2} \right\} \\
&= P\left\{ A(f^*), \ \hat{L}'_n(f^*) - L_n(f^*) \geq \frac{-\epsilon}{2} \right\} \\
&= \mathbb{E}\mathbf{1}_{A(f^*)} \mathbf{1}_{\hat{L}'_n(f^*) - L_n(f^*) \geq \frac{-\epsilon}{2}} \\
&= \mathbb{E}\mathbb{E}\left[\mathbf{1}_{A(f^*)} \mathbf{1}_{\hat{L}'_n(f^*) - L_n(f^*) \geq \frac{-\epsilon}{2}} \middle| \boldsymbol{Z}_n \right] \\
&= \mathbb{E}\mathbf{1}_{A(f^*)} \mathbb{E}\left[\mathbf{1}_{\hat{L}'_n(f^*) - L_n(f^*) \geq \frac{-\epsilon}{2}} \middle| \boldsymbol{Z}_n \right] \\
&= \mathbb{E}\mathbf{1}_{A(f^*)} P\left\{ \hat{L}'_n(f^*) - L_n(f^*) \geq \frac{-\epsilon}{2} \middle| \boldsymbol{Z}_n \right\}.
\end{aligned}$$

The probability in the latter can be bounded as

$$\begin{aligned}
P\left\{ \hat{L}'_n(f^*) - L_n(f^*) \geq \frac{-\epsilon}{2} \middle| \boldsymbol{Z}_n \right\} &= P\left\{ \left| \hat{L}'_n(f^*) - L_n(f^*) \right| \leq \frac{\epsilon}{2} \middle| \boldsymbol{Z}_n \right\} \\
&\quad + P\left\{ \hat{L}'_n(f^*) - L_n(f^*) > \frac{\epsilon}{2} \middle| \boldsymbol{Z}_n \right\} \\
&\geq P\left\{ \left| \hat{L}'_n(f^*) - L_n(f^*) \right| \leq \frac{\epsilon}{2} \middle| \boldsymbol{Z}_n \right\} \\
&\geq 1 - P\left\{ \left| \hat{L}'_n(f^*) - L_n(f^*) \right| > \frac{\epsilon}{2} \middle| \boldsymbol{Z}_n \right\} \\
&\geq \frac{1}{2},
\end{aligned}$$

where the last inequality is obtained from (a conditional form of) Bienaymé–Chebyshev's inequality applied to the random variable $\hat{L}'_n(f^*)$ whose expectation is $L_n(f^*)$ by construction of the ghost sample ($Z'_i \sim Z_i$):

$$\mathbb{E}\hat{L}'_n(f^*) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\ell(f^*(X'_i), Y'_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\ell(f^*(X_i), Y_i) = L_n(f^*).$$

[6] In case this is not possible, choose f^* as the function for which the difference is η-close to the supremum and take the limit $\eta \to 0$ to conclude as in Cortes et al. (2019).

Note that we can apply this inequality here since given $\boldsymbol{Z}_n$, f^* is fixed and independent of $\boldsymbol{Z}'_n$. Moreover, by the independence of $\boldsymbol{Z}_n$ and $\boldsymbol{Z}'_n$:

$$\begin{aligned}\mathbb{E}\left[(\hat{L}'_n(f^*) - L_n(f^*))^2 \mid \boldsymbol{Z}_n\right] &\leq \sup_{f\in\mathcal{F}} \mathbb{E}\left[(\hat{L}'_n(f) - L_n(f))^2\right] \\ &= \sup_{f\in\mathcal{F}} Var\left[\hat{L}'_n(f)\right] \\ &= \frac{1}{n^2}\sup_{f\in\mathcal{F}}\sum_{i=1}^{n} Var\left[\ell(f(X'_i), Y'_i)\right] \\ &\leq \frac{B^2}{4n},\end{aligned}$$

where the two last lines are due to the independence of the Z'_i and the fact that, for any f, $\ell(f(X'_i), Y'_i) \in [0, B]$. Thus,

$$\begin{aligned}P\left\{\left|\hat{L}'_n(f^*) - L_n(f^*)\right| \geq \frac{\epsilon}{2}\middle| \boldsymbol{Z}_n\right\} &\leq \frac{4\mathbb{E}\left[(\hat{L}'_n(f^*) - L_n(f^*))^2 \mid \boldsymbol{Z}_n\right]}{\epsilon^2} \\ &\leq \frac{B^2}{n\epsilon^2} \leq \frac{1}{2},\end{aligned}$$

since, by assumption, $n\epsilon^2 > 2B^2$.

Therefore, we have shown that

$$P(B) \geq P\{B(f^*)\} \geq \frac{1}{2}\mathbb{E}\mathbf{1}_{A(f^*)} = \frac{1}{2}P\{A(f^*)\} = \frac{1}{2}P(A)$$

and this concludes the proof.

C Proof of Theorem 3

We will prove that

$$P\left\{\sup_{f\in\mathcal{F}} L_n(f) - \hat{L}_n(f) \geq \epsilon\right\} \leq 2\Pi_{\mathcal{F}}(2n)\exp(-n\epsilon^2/8), \tag{19}$$

which implies the first inequality of Theorem 3 if we let $\delta = 2\Pi_{\mathcal{F}}(2n)\exp(-n\epsilon^2/8)$ and solve this equation for ϵ. The second inequality is then obtained by bounding the growth function by the VC-dimension with Sauer's lemma (Sauer, 1972; Vapnik, 1998), as usual.

By Lemma 1, we only have to bound the deviation between the empirical risk on the training sample and the one on the ghost sample. Let $\boldsymbol{\sigma}_n = (\sigma_i)_{1\leq i\leq n}$ denote a sequence of independent Rademacher variables uniformly distributed in $\{+1, -1\}$. Then, this random quantity,

$$D = \hat{L}'_n(f) - \hat{L}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{f(X'_i)\neq Y'_i} - \mathbf{1}_{f(X_i)\neq Y_i},$$

is identically distributed to

$$D = \frac{1}{n}\sum_{i=1}^{n} \sigma_i(\mathbf{1}_{f(X'_i)\neq Y'_i} - \mathbf{1}_{f(X_i)\neq Y_i}),$$

since (X_i, Y_i) and (X_i', Y_i') are independent and identically distributed, which makes $(\mathbf{1}_{f(X_i')\neq Y_i'} - \mathbf{1}_{f(X_i)\neq Y_i})$ a symmetric random variable whose distribution remains unchanged after multiplication by a random sign σ_i. Thus,

$$
\begin{aligned}
&P\left\{\sup_{f\in\mathcal{F}} \hat{L}_n'(f) - \hat{L}_n(f) > \frac{\epsilon}{2}\right\} \\
&= \mathbb{E}\mathbf{1}\left\{\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i(\mathbf{1}_{f(X_i')\neq Y_i'} - \mathbf{1}_{f(X_i)\neq Y_i}) > \frac{\epsilon}{2}\right\} \\
&= \mathbb{E}\ \mathbb{E}\left[\mathbf{1}\left\{\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i(\mathbf{1}_{f(X_i')\neq Y_i'} - \mathbf{1}_{f(X_i)\neq Y_i}) > \frac{\epsilon}{2}\right\}\middle| \boldsymbol{Z}_n, \boldsymbol{Z}_n'\right] \\
&= \mathbb{E}\ P\left\{\boldsymbol{\sigma}_n : \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i(\mathbf{1}_{f(X_i')\neq Y_i'} - \mathbf{1}_{f(X_i)\neq Y_i}) > \frac{\epsilon}{2}\middle| \boldsymbol{Z}_n, \boldsymbol{Z}_n'\right\} \\
&= \mathbb{E}\ P\left\{\boldsymbol{D}_n : \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n D_i > \frac{\epsilon}{2}\middle| \boldsymbol{Z}_n, \boldsymbol{Z}_n'\right\},
\end{aligned}
$$

where we let $\boldsymbol{D}_n = (D_i)_{1\leq i\leq n}$ and $D_i = \sigma_i(\mathbf{1}_{f(X_i')\neq Y_i'} - \mathbf{1}_{f(X_i)\neq Y_i})$. For any fixed f and $\boldsymbol{Z}_n$, $\boldsymbol{Z}_n'$, the conditional expectation of these variables is zero: $\mathbb{E}[D_i \mid \boldsymbol{Z}_n, \boldsymbol{Z}_n'] = (\mathbf{1}_{f(X_i')\neq Y_i'} - \mathbf{1}_{f(X_i)\neq Y_i})\mathbb{E}\sigma_i = 0$. Thus, by applying Hoeffding inequality (Theorem 2) on the sequence of independent variables $\boldsymbol{D}_n$, we obtain

$$
P\left\{\frac{1}{n}\sum_{i=1}^n D_i > \frac{\epsilon}{2}\middle| \boldsymbol{Z}_n, \boldsymbol{Z}_n'\right\} \leq \exp\left(\frac{-n^2\epsilon^2}{2\sum_{i=1}^n c_i^2}\right),
$$

where $c_i = b_i - a_i$ with $a_i \leq D_i \leq b_i$ and $a_i = -1$, $b_i = 1$. This gives $c_i^2 = 4$ and

$$
\forall f \in \mathcal{F}, \quad P\left\{\frac{1}{n}\sum_{i=1}^n D_i > \frac{\epsilon}{2}\middle| \boldsymbol{Z}_n, \boldsymbol{Z}_n'\right\} \leq \exp\left(\frac{-n\epsilon^2}{8}\right).
$$

Therefore, with $\mathcal{F}_{\boldsymbol{X}_n\boldsymbol{X}_n'} = \{(f(X_1), \ldots, f(X_n), f(X_1'), \ldots, f(X_n')) : f \in \mathcal{F}\}$, by the union bound,

$$
P\left\{\boldsymbol{D}_n : \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n D_i > \frac{\epsilon}{2}\middle| \boldsymbol{Z}_n, \boldsymbol{Z}_n'\right\} \leq |\mathcal{F}_{\boldsymbol{X}_n\boldsymbol{X}_n'}| \exp\left(\frac{-n\epsilon^2}{2}\right)
$$

and the conclusion stems from Lemma 1 and Definition 1.

D Proof of Lemma 2

The proof follows the lines of that of Lemma 2 in Cortes et al. (2019) and makes use of the following.

Lemma 5 (Theorem 1 in Greenberg and Mohri (2014))**.** *For any positive integer m and any probability p such that $p > 1/m$, let X be a random variable distributed according to the binomial distribution with m trials and probability of success of each trial p. Then, $\mathbb{E}X = mp$ and*

$$
P(X \geq \mathbb{E}X) > \frac{1}{4}.
$$

For any $f \in \mathcal{F}$, consider the events

$$A(f) = \left\{ \frac{L_n(f) - \hat{L}_n(f)}{\sqrt{L_n(f)}} > \epsilon \right\} \quad \text{and} \quad B(f) = \{\hat{L}'_n(f) > L_n(f)\}.$$

Then

$$A(f) \cap B(f) \quad \subset \quad C(f) = \left\{ \frac{\hat{L}'_n(f) - \hat{L}_n(f)}{\sqrt{(\hat{L}_n(f) + \hat{L}'_n(f) + \frac{1}{n})/2}} > \epsilon \right\}.$$

To see this, first note that

$$A(f) \quad \Rightarrow \quad \hat{L}_n(f) < L_n(f) - \epsilon\sqrt{L_n(f)},$$

which also implies $\epsilon < \sqrt{L_n(f)}$. Then, also note that, for all $a, b > 0$, the function $\frac{a-b}{\sqrt{(a+b+\frac{1}{n})/2}}$ is increasing in a and decreasing in b (the derivatives are positive and negative). Thus,

$$A(f) \Rightarrow \quad \frac{\hat{L}'_n(f) - \hat{L}_n(f)}{\sqrt{(\hat{L}_n(f) + \hat{L}'_n(f) + \frac{1}{n})/2}} \geq \frac{\hat{L}'_n(f) - L_n(f) + \epsilon\sqrt{L_n(f)}}{\sqrt{(\hat{L}'_n(f) + L_n(f) - \epsilon\sqrt{L_n(f)} + \frac{1}{n})/2}},$$

where, on $B(f)$,

$$\begin{aligned}
\frac{\hat{L}'_n(f) - L_n(f) + \epsilon\sqrt{L_n(f)}}{\sqrt{(\hat{L}'_n(f) + L_n(f) - \epsilon\sqrt{L_n(f)} + \frac{1}{n})/2}} &\geq \frac{\epsilon\sqrt{L_n(f)}}{\sqrt{(2L_n(f) - \epsilon\sqrt{L_n(f)} + \frac{1}{n})/2}} \\
&\geq \frac{\epsilon\sqrt{L_n(f)}}{\sqrt{(2L_n(f) - \epsilon^2 + \frac{1}{n})/2}}
\end{aligned}$$

(because $\epsilon < \sqrt{L_n(f)}$). Since we assume $n\epsilon^2 > 1$ and thus $\epsilon^2 > 1/n$, we have $-\epsilon^2 + \frac{1}{n} < 0$ and

$$\frac{\hat{L}'_n(f) - L_n(f) + \epsilon\sqrt{L_n(f)}}{\sqrt{(\hat{L}'_n(f) + L_n(f) - \epsilon\sqrt{L_n(f)} + \frac{1}{n})/2}} \geq \frac{\epsilon\sqrt{L_n(f)}}{\sqrt{L_n(f)}} = \epsilon.$$

Now, let $f^* \in \mathcal{F}$ denote the function that depends solely on $\boldsymbol{Z}_n$ and such that[7]

$$\frac{L_n(f^*) - \hat{L}_n(f^*)}{\sqrt{L_n(f)}} = \sup_{f \in \mathcal{F}} \frac{L_n(f) - \hat{L_{nn}}(f)}{\sqrt{L_n(f)}}.$$

Then, by the discussion above,

$$\begin{aligned}
P\{C(f^*)\} &\geq P\{A(f^*) \cap B(f^*)\} \\
&= \mathbb{E}\mathbf{1}_{A(f^*)}\mathbf{1}_{B(f^*)} \\
&= \mathbb{E}\mathbb{E}\left[\mathbf{1}_{A(f^*)}\mathbf{1}_{B(f^*)} \middle| \boldsymbol{Z}_n\right] \\
&= \mathbb{E}\mathbf{1}_{A(f^*)}\mathbb{E}\left[\mathbf{1}_{B(f^*)} \middle| \boldsymbol{Z}_n\right] \\
&= \mathbb{E}\mathbf{1}_{A(f^*)}P\{B(f^*) \mid \boldsymbol{Z}_n\}.
\end{aligned}$$

[7] In case this is not possible, choose f^* as the function for which the ratio is η-close to the supremum and take the limit $\eta \to 0$ to conclude as in Cortes et al. (2019).

Moreover, given $\boldsymbol{Z}_n$, f^* is fixed and if $L_n(f^*) > \epsilon^2$, then under the assumption $n\epsilon > 1$, $L_n(f) > 1/n$. Then, by stationarity and the construction of the ghost sample, the Z'_i are i.i.d., which makes the quantity

$$n\hat{L}'_n(f^*) = \sum_{i=1}^{n} \mathbf{1}_{f^*(X'_i)\neq Y'_i}$$

distributed according to a binomial distribution with $m = n$ trials and probability of success $p = P(f^*(X'_i) \neq Y'_i) = P(f^*(X_i) \neq Y_i) = L_n(f^*) > 1/m$. Thus, by Lemma 5, $P\{B(f^*) \mid \boldsymbol{Z}_n\} > 1/4$ and, in the case $L_n(f^*) > \epsilon^2$, $P\{C(f^*)\} \geq \mathbb{E}\mathbf{1}_{A(f^*)}/4$ and $P\{A(f^*)\} \leq 4P\{C(f^*)\}$. In the other case, if $L_n(f^*) \leq \epsilon^2$, then $A(f^*)$ cannot hold and $P\{A(f^*)\} = 0 \leq 4P\{C(f^*)\}$. So, in any case:

$$\begin{aligned} P\left\{\sup_{f\in\mathcal{F}} \frac{L_n(f) - \hat{L}_n(f)}{\sqrt{L_n(f)}} \geq \epsilon\right\} &= P\{A(f^*)\} \\ &\leq 4P\{C(f^*)\} \\ &\leq 4P\left\{\sup_{f\in\mathcal{F}} \frac{\hat{L}'_n(f) - \hat{L}_n(f)}{\sqrt{(\hat{L}_n(f) + \hat{L}'_n(f) + \frac{1}{n})/2}} > \epsilon\right\}, \end{aligned}$$

where the first equality is due to the choice of f^* and the last inequality is due to $f^* \in \mathcal{F}$.

E Proof of Theorem 4

We proceed as in the proof of Theorem 3 (see App. C) with a sequence $\boldsymbol{\sigma}_n = (\sigma_i)_{1\leq i\leq n}$ of independent Rademacher variables uniformly distributed in $\{+1, -1\}$ and by taking advantage of the symmetry of the $(\mathbf{1}_{f(X'_i)\neq Y'_i} - \mathbf{1}_{f(X_i)\neq Y_i})$:

$$\begin{aligned} &P\left\{\sup_{f\in\mathcal{F}} \frac{\hat{L}'_n(f) - \hat{L}_n(f)}{\sqrt{(\hat{L}_n(f) + \hat{L}'_n(f) + \frac{1}{n})/2}} > \epsilon\right\} \\ &= \mathbb{E}\mathbf{1}\left\{\sup_{f\in\mathcal{F}} \frac{\frac{1}{n}\sum_{i=1}^{n} \sigma_i(\mathbf{1}_{f(X'_i)\neq Y'_i} - \mathbf{1}_{f(X_i)\neq Y_i})}{\sqrt{(\hat{L}_n(f) + \hat{L}'_n(f) + \frac{1}{n})/2}} > \epsilon\right\} \\ &= \mathbb{E}\ \mathbb{E}\left[\mathbf{1}\left\{\sup_{f\in\mathcal{F}} \frac{\frac{1}{n}\sum_{i=1}^{n} \sigma_i(\mathbf{1}_{f(X'_i)\neq Y'_i} - \mathbf{1}_{f(X_i)\neq Y_i})}{\sqrt{(\hat{L}_n(f) + \hat{L}'_n(f) + \frac{1}{n})/2}} > \epsilon\right\} \middle| \boldsymbol{Z}_n, \boldsymbol{Z}'_n\right] \\ &= \mathbb{E}\ P\left\{\boldsymbol{\sigma}_n : \sup_{f\in\mathcal{F}} \frac{\frac{1}{n}\sum_{i=1}^{n} \sigma_i(\mathbf{1}_{f(X'_i)\neq Y'_i} - \mathbf{1}_{f(X_i)\neq Y_i})}{\sqrt{(\hat{L}_n(f) + \hat{L}'_n(f) + \frac{1}{n})/2}} > \epsilon \middle| \boldsymbol{Z}_n, \boldsymbol{Z}'_n\right\} \\ &= \mathbb{E}\ P\left\{\boldsymbol{D}_n : \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} D_i > \epsilon \middle| \boldsymbol{Z}_n, \boldsymbol{Z}'_n\right\}, \end{aligned}$$

where we let $\boldsymbol{D}_n = (D_i)_{1\leq i\leq n}$ and $D_i = \frac{\sigma_i(\mathbf{1}_{f(X'_i)\neq Y'_i} - \mathbf{1}_{f(X_i)\neq Y_i})}{\sqrt{(\hat{L}_n(f) + \hat{L}'_n(f) + \frac{1}{n})/2}}$. Since $\mathbb{E}\sigma_i = 0$, for any fixed f and $\boldsymbol{Z}_n$, $\boldsymbol{Z}'_n$, we have $\mathbb{E}[D_i \mid \boldsymbol{Z}_n, \boldsymbol{Z}'_n] = 0$. Thus, by applying

Hoeffding inequality (Theorem 2) on the sequence of independent variables $\boldsymbol{D}_n$, we obtain

$$P\left\{\frac{1}{n}\sum_{i=1}^{n} D_i > \epsilon \,\middle|\, \boldsymbol{Z}_n, \boldsymbol{Z}'_n\right\} \leq \exp\left(\frac{-2n^2\epsilon^2}{\sum_{i=1}^n c_i^2}\right),$$

where $c_i = b_i - a_i$ with $a_i \leq D_i \leq b_i$. In particular, we have

$$\begin{aligned} a_i = \frac{-|\mathbf{1}_{f(X'_i)\neq Y'_i} - \mathbf{1}_{f(X_i)\neq Y_i}|}{\sqrt{(\hat{L}_n(f) + \hat{L}'_n(f) + \frac{1}{n})/2}} &= \frac{-|\mathbf{1}_{f(X'_i)\neq Y'_i} - \mathbf{1}_{f(X_i)\neq Y_i}|}{\sqrt{\frac{1}{n}(1 + \sum_{i=1}^n \mathbf{1}_{f(X_i)\neq Y_i} + \mathbf{1}_{f(X'_i)\neq Y'_i})/2}} \\ &= \frac{-|\mathbf{1}_{f(X'_i)\neq Y'_i} - \mathbf{1}_{f(X_i)\neq Y_i}|\sqrt{2n}}{\sqrt{1 + \sum_{i=1}^n \mathbf{1}_{f(X_i)\neq Y_i} + \mathbf{1}_{f(X'_i)\neq Y'_i}}} \end{aligned}$$

and, similarly:

$$b_i = \frac{|\mathbf{1}_{f(X'_i)\neq Y'_i} - \mathbf{1}_{f(X_i)\neq Y_i}|\sqrt{2n}}{\sqrt{1 + \sum_{i=1}^n \mathbf{1}_{f(X_i)\neq Y_i} + \mathbf{1}_{f(X'_i)\neq Y'_i}}},$$

which gives

$$\begin{aligned} c_i^2 &= \left(\frac{2|\mathbf{1}_{f(X'_i)\neq Y'_i} - \mathbf{1}_{f(X_i)\neq Y_i}|\sqrt{2n}}{\sqrt{1 + \sum_{i=1}^n \mathbf{1}_{f(X_i)\neq Y_i} + \mathbf{1}_{f(X'_i)\neq Y'_i}}}\right)^2 = \frac{8n(\mathbf{1}_{f(X'_i)\neq Y'_i} - \mathbf{1}_{f(X_i)\neq Y_i})^2}{1 + \sum_{i=1}^n \mathbf{1}_{f(X_i)\neq Y_i} + \mathbf{1}_{f(X'_i)\neq Y'_i}} \\ &\leq \frac{8n(\mathbf{1}_{f(X'_i)\neq Y'_i} - \mathbf{1}_{f(X_i)\neq Y_i})^2}{\sum_{i=1}^n \mathbf{1}_{f(X_i)\neq Y_i} + \mathbf{1}_{f(X'_i)\neq Y'_i}} \end{aligned}$$

and

$$\sum_{i=1}^n c_i^2 \leq \frac{8n\sum_{i=1}^n (\mathbf{1}_{f(X'_i)\neq Y'_i} - \mathbf{1}_{f(X_i)\neq Y_i})^2}{\sum_{i=1}^n \mathbf{1}_{f(X_i)\neq Y_i} + \mathbf{1}_{f(X'_i)\neq Y'_i}} \leq 8n,$$

where the last inequality is due to

$$\begin{aligned} (\mathbf{1}_{f(X'_i)\neq Y'_i} - \mathbf{1}_{f(X_i)\neq Y_i})^2 &= \mathbf{1}^2_{f(X'_i)\neq Y'_i} + \mathbf{1}^2_{f(X_i)\neq Y_i} - 2\mathbf{1}_{f(X'_i)\neq Y'_i}\mathbf{1}_{f(X_i)\neq Y_i} \\ &\leq \mathbf{1}_{f(X'_i)\neq Y'_i} + \mathbf{1}_{f(X_i)\neq Y_i}. \end{aligned}$$

Thus, for any $f \in \mathcal{F}$,

$$P\left\{\boldsymbol{\sigma}_n : \frac{\frac{1}{n}\sum_{i=1}^n \sigma_i(\mathbf{1}_{f(X'_i)\neq Y'_i} - \mathbf{1}_{f(X_i)\neq Y_i})}{\sqrt{(\hat{L}_n(f) + \hat{L}'_n(f) + \frac{1}{n})/2}} > \epsilon \,\middle|\, \boldsymbol{Z}_n, \boldsymbol{Z}'_n\right\} \leq \exp\left(\frac{-n\epsilon^2}{4}\right)$$

and, with $\mathcal{F}_{\boldsymbol{X}_n\boldsymbol{X}'_n} = \{(f(X_1), \ldots, f(X_n), f(X'_1), \ldots, f(X'_n)) : f \in \mathcal{F}\}$, by the union bound,

$$\begin{aligned} P\left\{\boldsymbol{\sigma}_n : \sup_{f\in\mathcal{F}} \frac{\frac{1}{n}\sum_{i=1}^n \sigma_i(\mathbf{1}_{f(X'_i)\neq Y'_i} - \mathbf{1}_{f(X_i)\neq Y_i})}{\sqrt{(\hat{L}_n(f) + \hat{L}'_n(f) + \frac{1}{n})/2}} > \epsilon \,\middle|\, \boldsymbol{Z}_n, \boldsymbol{Z}'_n\right\} \\ \leq |\mathcal{F}_{\boldsymbol{X}_n\boldsymbol{X}'_n}| \exp\left(\frac{-n\epsilon^2}{4}\right). \end{aligned}$$

Gathering it all and using Lemma 2:

$$P\left\{\sup_{f\in\mathcal{F}}\frac{L(f)-\hat{L}_n(f)}{\sqrt{L(f)}}>\epsilon\right\}\leq 4P\left\{\sup_{f\in\mathcal{F}}\frac{\hat{L}'_n(f)-\hat{L}_n(f)}{\sqrt{(\hat{L}_n(f)+\hat{L}'_n(f)+\frac{1}{n})/2}}>\epsilon\right\}$$
$$\leq 4\mathbb{E}|\mathcal{F}_{\boldsymbol{X}_n\boldsymbol{X}'_n}|\exp\left(\frac{-n\epsilon^2}{4}\right)$$
$$\leq 4\Pi_{\mathcal{F}}(2n)\exp\left(\frac{-n\epsilon^2}{4}\right).$$

By setting $\delta = 4\Pi_{\mathcal{F}}(2n)\exp(-n\epsilon^2/4)$, we obtain that, with probability at least $1-\delta$,

$$\forall f\in\mathcal{F},\quad \frac{L(f)-\hat{L}_n(f)}{\sqrt{L(f)}}\leq 2\sqrt{\frac{\log\Pi_{\mathcal{F}}(2n)+\log\frac{4}{\delta}}{n}}$$

and, using $A\leq B\sqrt{A}+C\Rightarrow A\leq B^2+B\sqrt{C}+C$ yields

$$L(f)\leq\hat{L}_n(f)+\sqrt{L(f)}2\sqrt{\frac{\log\Pi_{\mathcal{F}}(2n)+\log\frac{4}{\delta}}{n}}$$
$$\Rightarrow\quad L(f)\leq\hat{L}_n(f)+2\sqrt{\hat{L}_n(f)\frac{\log\Pi_{\mathcal{F}}(2n)+\log\frac{4}{\delta}}{n}}+4\frac{\log\Pi_{\mathcal{F}}(2n)+\log\frac{4}{\delta}}{n},$$

in which the growth function $\Pi_{\mathcal{F}}(2n)$ can be bounded in terms of the VC-dimension d_{VC} by Sauer's lemma (Sauer, 1972; Vapnik, 1998).

F Proof of Theorem 5

We will show that, with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}}\left(L_n(f)-\hat{L}_n(f)\right)\leq\mathcal{R}_{\boldsymbol{Z}_n}(\mathcal{L})+\mathcal{R}_{\boldsymbol{Z}'_n}(\mathcal{L})+B\sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

Let us first rewrite, with a slight abuse of notation, the loss as $\ell(Z_i)=\ell(f(X_i),Y_i)$. Then, we apply a bounded difference inequality to

$$g(Z_1,\ldots,Z_n)=\sup_{f\in\mathcal{F}}\left(L_n(f)-\hat{L}_n(f)\right).$$

If $\ell(Z_i)\in[0,B]$, then the bounded difference condition with $c_i=B/n$ is satisfied and we can apply Theorem 1. By setting $\delta=\exp(-2t^2/\sum_{i=1}^n c_i^2)=\exp(-2nt^2/B^2)$, and thus $t=B\sqrt{\log(1/\delta)/2n}$, we get that, with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}}\left(L_n(f)-\hat{L}_n(f)\right)\leq\mathbb{E}\sup_{f\in\mathcal{F}}\left(L_n(f)-\hat{L}_n(f)\right)+B\sqrt{\frac{\log\frac{1}{\delta}}{2n}}.\tag{20}$$

Then, introduce the ghost sample $\boldsymbol{Z}'_n$ and use its independence with $\boldsymbol{Z}_n$ and $Z'_i\sim Z_i$ to write

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\ell(Z'_i)\,\middle|\,\boldsymbol{Z}_n\right]=\mathbb{E}\frac{1}{n}\sum_{i=1}^n\ell(Z'_i)=\frac{1}{n}\sum_{i=1}^n\mathbb{E}\ell(Z_i)=L_n(f),$$

and thus

$$\begin{aligned}
\mathbb{E}\sup_{f\in\mathcal{F}}\left(L_n(f)-\hat{L}_n(f)\right) &= \mathbb{E}\sup_{f\in\mathcal{F}}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(Z_i')\middle|\boldsymbol{Z}_n\right]-\frac{1}{n}\sum_{i=1}^{n}\ell(Z_i)\right)\\
&= \mathbb{E}\sup_{f\in\mathcal{F}}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(Z_i')-\frac{1}{n}\sum_{i=1}^{n}\ell(Z_i)\middle|\boldsymbol{Z}_n\right]\right)\\
&\leq \mathbb{E}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\left(\ell(Z_i')-\ell(Z_i)\right)\middle|\boldsymbol{Z}_n\right]\\
&= \mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\left(\ell(Z_i')-\ell(Z_i)\right),
\end{aligned}$$

where we used Jensen's inequality in the third line.

Then, we can introduce Rademacher variables as follows. As in the proof of Theorem 3, the variables $(\ell(Z_i')-\ell(Z_i))$ are symmetric by construction of the ghost sample that ensures that $Z_i \sim Z_i'$ and Z_i is independent of Z_i'. Thus, for any $\boldsymbol{\sigma}_n = (\sigma_i)_{1\leq i\leq n} \in \{-1,+1\}^n$ chosen independently of Z_i and Z_i', we can replace them by $\sigma_i\left(\ell(Z_i')-\ell(Z_i)\right)$ without changing the resulting expectation. Furthermore, averaging over all the 2^n sequences of σ_i does not change the result either. This leads to

$$\begin{aligned}
\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\left(\ell(Z_i')-\ell(Z_i)\right) &= \mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(\ell(Z_i')-\ell(Z_i)\right)\\
&\leq \mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell(Z_i')+\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}-\sigma_i\ell(Z_i)\\
&= \mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell(Z_i')+\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell(Z_i)\\
&= \mathcal{R}_n(\mathcal{L})+\mathcal{R}_n'(\mathcal{L}),
\end{aligned}$$

where we used $\sigma_i \sim -\sigma_i$ in the third line. Together with (20), this completes the proof.

G Proof of (6)

For any $\boldsymbol{z}_n \in \mathcal{Z}^n$, by the contraction principle (Ledoux and Talagrand, 1991), we have $\hat{\mathcal{R}}_{\boldsymbol{z}_n}(\mathcal{L}) \leq 4M\hat{\mathcal{R}}_{\boldsymbol{x}_n}(\mathcal{F})$. Then, the computations of Bartlett and Mendelson (2002) give

$$\hat{\mathcal{R}}_{\boldsymbol{x}_n}(\mathcal{F}) \leq \frac{\Lambda\sqrt{\sum_{i=1}^{n}\|X_i\|^2}}{n} \leq \frac{\Lambda\sup_{x\in\mathcal{X}}\|x\|}{\sqrt{n}}$$

for linear models. For kernel models, the norms $\|X_i\|$ are replaced by $\|K(X_i,\cdot)\| = \sqrt{K(X_i,X_i)}$, which is 1 for the Gaussian kernel.

References

Alamo, T., Tempo, R., and Camacho, E. (2009). Randomized strategies for probabilistic solutions of uncertain feasibility and optimization problems. *IEEE Transactions on Automatic Control*, 54(11):2545–2559.

Alon, N., Ben-David, S., Cesa-Bianchi, N., and Haussler, D. (1997). Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631.

Baldi, P. and Brunak, S. (2001). *Bioinformatics: the Machine Learning Approach*. MIT press.

Bartlett, P., Foster, D., and Telgarsky, M. (2017). Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 30.

Bartlett, P., Linder, T., and Lugosi, G. (1998). The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information Theory*, 44(5):1802–1813.

Bartlett, P. and Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482.

Biau, G., Devroye, L., and Lugosi, G. (2008). On the performance of clustering in Hilbert spaces. *IEEE Transactions on Information Theory*, 54(2):781–790.

Bradley, R. (2005). Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144.

Calafiore, G. (2010). Random convex programs. *SIAM Journal on Optimization*, 20(6):3427–3464.

Campi, M. and Garatti, S. (2008). The exact feasibility of randomized solutions of uncertain convex programs. *SIAM Journal on Optimization*, 19(3):1211–1230.

Campi, M., Garatti, S., and Ramponi, F. (2018). A general scenario theory for nonconvex optimization and decision making. *IEEE Transactions on Automatic Control*, 63(12):4067–4078.

Cortes, C., Greenberg, S., and Mohri, M. (2019). Relative deviation learning bounds and generalization with unbounded loss functions. *Annals of Mathematics and Artificial Intelligence*, 85:45–70.

Cortes, C., Mohri, M., and Suresh, A. (2021). Relative deviation margin bounds. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 2122–2131.

de la Peña, V. and Giné, E. (1999). *Decoupling: From Dependence to Independence*. Springer.

Esfahani, P., Sutter, T., and Lygeros, J. (2014). Performance bounds for the scenario approach and an extension to a class of non-convex programs. *IEEE Transactions on Automatic Control*, 60(1):46–58.

Faradonbeh, M., Tewari, A., and Michailidis, G. (2018). Finite time identification in unstable linear systems. *Automatica*, 96:342–353.

Greenberg, S. and Mohri, M. (2014). Tight lower bound on the probability of a binomial exceeding its expectation. *Statistics & Probability Letters*, 86:91–98.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Janson, S. (2004). Large deviations for sums of partly dependent random variables. *Random Structures and Algorithms*, 24(3):234–248.

Kuznetsov, V. and Mohri, M. (2015). Learning theory and algorithms for forecasting non-stationary time series. In *Advances in Neural Information Processing Systems*, volume 28.

Lauer, F. (2020a). Error bounds for piecewise smooth and switching regression. *IEEE Transactions on Neural Networks and Learning Systems*, 31(4):1183–1195.

Lauer, F. (2020b). Risk bounds for learning multiple components with permutation-invariant losses. In *Proc. of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1178–1187.

Lauer, F. (2023). Margin theory for the scenario-based approach to robust optimization in high dimension. *arXiv preprint*, arXiv:2303.03891.

Lauer, F. and Bloch, G. (2019). *Hybrid System Identification: Theory and Algorithms for Learning Switching Models*. Springer.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, Berlin.

Ljung, L. (1987). *System Identification: Theory for the User*. Prentice-Hall, Englewood Cliffs, NJ.

Massucci, L., Lauer, F., and Gilson, M. (2022). A statistical learning perspective on switched linear system identification. *Automatica*, 145:110532.

McDiarmid, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics*, pages 148–188. Cambridge University Press.

McDonald, D., Shalizi, C., and Schervish, M. (2015). Estimating beta-mixing coefficients via histograms. *Electronic Journal of Statistics*, 9:2855–2883.

Meir, R. (2000). Nonparametric time series prediction through adaptive model selection. *Machine Learning*, 39:5–34.

Mendelson, S. (2002). Rademacher averages and phase transitions in Glivenko-Cantelli classes. *IEEE Transactions on Information Theory*, 48(1):251–263.

Mendelson, S. (2014). Learning without concentration. In *Proc. of the Conference on Learning Theory (COLT)*, pages 25–39.

Mendelson, S. (2018). Learning without concentration for general loss functions. *Probability Theory and Related Fields*, 171(1-2):459–502.

Mendelson, S. and Vershynin, R. (2003). Entropy and the combinatorial dimension. *Inventiones mathematicae*, 152:37–55.

Mohri, M. and Rostamizadeh, A. (2009). Rademacher complexity bounds for non-iid processes. In *Advances in Neural Information Processing Systems*, pages 1097–1104.

Mohri, M. and Rostamizadeh, A. (2010). Stability bounds for stationary φ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11:789–814.

Rakhlin, A., Sridharan, K., and Tewari, A. (2015). Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161:111–153.

Ralaivola, L. and Amini, M.-R. (2015). Entropy-based concentration inequalities for dependent variables. In *Proc. of the 32nd International Conference on Machine Learning, (ICML), Lille, France*, pages 2436–2444.

Ralaivola, L., Szafranski, M., and Stempfel, G. (2010). Chromatic PAC-Bayes bounds for non-iid data: Applications to ranking and stationary β-mixing processes. *Journal of Machine Learning Research*, 11:1927–1956.

Sauer, N. (1972). On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147.

Simchowitz, M., Mania, H., Tu, S., Jordan, M., and Recht, B. (2018). Learning without mixing: Towards a sharp analysis of linear system identification. In *Proc. of the 31st Conf. On Learning Theory (COLT)*, pages 439–473.

Steinwart, I. and Christmann, A. (2009). Fast learning from non-i.i.d. observations. In *Advances in Neural Information Processing Systems 22*, pages 1768–1776.

Talagrand, M. (2014). *Upper and Lower Bounds for Stochastic Processes*. Springer.

Usunier, N., Amini, M.-R., and Gallinari, P. (2006). Generalization error bounds for classifiers trained with interdependent data. In *Advances in Neural Information Processing Systems*, volume 18, pages 1369–1376.

van De Geer, S. (2007). On Hoeffding's inequality for dependent random variables. In Dehling, H., Mikosch, T., and Sørensen, M., editors, *Empirical Process Techniques for Dependent Data*, pages 161–169. Springer.

Vapnik, V. (1998). *Statistical Learning Theory*. John Wiley & Sons.

Wang, Z., Berger, G., and Jungers, R. (2021). Data-driven feedback stabilization of switched linear systems with probabilistic stability guarantees. In *Proceedings of the 60th IEEE Conference on Decision and Control (CDC)*, pages 4400–4405.

Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116.