RISK BOUNDS WHEN LEARNING INFINITELY MANY RESPONSE FUNCTIONS BY ORDINARY LINEAR REGRESSION

Plassier, V., Portier, F. And J. Segers

DISCUSSION PAPER | 2020 / 19

UCLouvain

ISBA | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

RISK BOUNDS WHEN LEARNING INFINITELY MANY RESPONSE FUNCTIONS BY ORDINARY LINEAR REGRESSION

Vincent Plassier
LIDAM/ISBA, UCLouvain
and ENS Paris-Saclay
1348 Louvain-la-Neuve, Belgium
vincent.plassier@ens-paris-saclay.fr

Francois Portier
LTCI, Télécom Paris
Institut polytechnique de Paris
91120 Palaiseau, France
francois.portier@gmail.com

Johan Segers
LIDAM/ISBA, UCLouvain
1348 Louvain-la-Neuve, Belgium
johan.segers@uclouvain.be

June 17, 2020

ABSTRACT

Consider the problem of learning a large number of response functions simultaneously based on the same input variables. The training data consist of a single independent random sample of the input variables drawn from a common distribution together with the associated responses. The input variables are mapped into a high-dimensional linear space, called the feature space, and the response functions are modelled as linear functionals of the mapped features, with coefficients calibrated via ordinary least squares. We provide convergence guarantees on the worst-case excess prediction risk by controlling the convergence rate of the excess risk uniformly in the response function. The dimension of the feature map is allowed to tend to infinity with the sample size. The collection of response functions, although potentially infinite, is supposed to have a finite Vapnik–Chervonenkis dimension. The bound derived can be applied when building multiple surrogate models in a reasonable computing time.

1 Introduction

Context. When the outcome of interest is generated by a black box model which cannot be easily evaluated, a well-spread technique is to build a surrogate model allowing to reproduce the behavior of the true model while being computationally cheaper. This approach, known as *response surface*, is popular in many fields of engineering such as *reliability analysis* (Bucher and Bourgund, 1990), *aerospace science* (Forrester and Keane, 2009), *energy science* (Nguyen et al., 2014), or electromagnetic dosimetry (Azzi et al., 2019), to name a few. In addition, the response surface methodology is useful in applied mathematics, for instance, in *optimization* when the objective function is difficult to evaluate (Jones, 2001) and in *Monte Carlo integration*, where a surrogate function can be used to reduce the variance of the Monte Carlo estimate with the help of control variates (Portier and Segers, 2019). For a general presentation of the response surface methodology, we refer to Myers et al. (2016).

Framework. The statistical framework of a response surface is as follows. Consider a real-valued function f defined on a state space $\mathscr{X}$, that is, $f : \mathscr{X} \to \mathbb{R}$. In many applications, the function f is a black box and difficult to evaluate. For instance, a request to f might be obtained by running a heavy computer program. In such a situation, one can afford only a few requests to f, that is, one can obtain $f(X_1), \dots, f(X_n)$, where $n \in \mathbb{N}$ and $X_1, \dots, X_n$ is an independent sample of $\mathscr{X}$-valued random variables with distribution P, called the inputs or the covariates. Based on those evaluations, the goal is to build a *surrogate model*, that is, an approximation of f which is easier to calculate than f itself.

Many different methods can be used to build a surrogate model. The simplest one consists in learning f as a linear combination of the covariates by minimizing the sum of squared errors. A well spread extension is to fit a polynomial function instead of a linear one as presented in Myers et al. (2016) or Konakli and Sudret (2016). Since building a response surface consists in the same task as regression, any regression method might be used in principle. Popular methods include moving least-squares (Breitkopf et al., 2005), Gaussian processes (Frean and Boyle, 2008) or neural nets (Bauer et al., 2019). For surrogate models, the approximation method needs to be sufficiently flexible to fit the black box function f well and simple enough to require only a small amount of computations. It is perhaps due to its connection to the regression framework that the problem of building surrogate models has received little specific attention in the statistical learning literature. Our purpose is to address the question of learning many surrogate models simultaneously from a single, random design.

Learning several models simultaneously. The fundamental question raised in this paper deals with the ability of building several, possibly infinitely many, surrogate models such that (a) they share the same quality and (b) they are constructed with the help of a single input sample. From a theoretical standpoint, it relates to the question of *uniformity* over the tasks: can a broad family of models be learnt with a uniform level of accuracy? From a more practical point of view, by working with the same inputs to solve multiple tasks simultaneously, one benefits from a certain computational advantage, as explained below.

Consider a broad class $\mathscr{F}$ of black box models f. For each such model f, a least-squares estimate is obtained out of the linear span of the components of the d-dimensional feature map $h = (h_1, \dots, h_d)^\intercal : \mathscr{X} \to \mathbb{R}^d$, where $^\intercal$ denotes matrix transposition. The feature map h should be known and easy to evaluate. Define the ordinary least squares estimate

$$\hat{\beta}_f \in \operatorname*{arg\,min}_{b \in \mathbb{R}^d} \sum_{i=1}^{n} \{f(X_i) - h(X_i)^\intercal b\}^2.$$

The surrogate model for f is then defined as $x \mapsto \hat{f}(x) = h(x)^\intercal \hat{\beta}_f$.

This approach is known as *series estimators* or simply as *least squares estimators* (Härdle, 1990; Györfi et al., 2006). It is quite general as several basis functions might be considered such as polynomials, indicators, spline functions or the Fourier basis. In the regression framework, series estimators have been studied for instance in Newey (1997) and Belloni et al. (2015). Series estimators are a convenient way to include shape constraints on f, as for instance when f is partially linear. They also facilitate the computation of derivatives (Zhou and Wolfe, 2000). In this respect, series estimators can help to build *easy-to-interpret* models, a desirable feature when one is in need of some knowledge about the effects of certain inputs on the black box model f.

One motivation for the use of series estimators is the computational advantage they provide when several models are to be learnt at the same time. When a large number m of such surrogate models f are built with different covariates, running m least-squares algorithms, for instance using the Cholesky decomposition, requires $\mathrm{O}(mnd^2)$ operations (Friedman et al., 2001, Section 3.5), assuming that $d = \mathrm{O}(n)$. In our framework of a single training sample, however, the Cholesky decomposition needs to be done only once, and the time needed to compute m least squares estimates is rather $\mathrm{O}(nd^2 + mnd)$.

Uniform convergence rate in random design with increasing dimension. For a given model $f \in \mathscr{F}$, the error made by a surrogate model $x \mapsto h(x)^\intercal b$ with coefficient vector $b \in \mathbb{R}^d$ is measured

through the $L^2(P)$-risk

$$L_f(b) = \int_{\mathscr{X}} \{f(x) - h(x)^\intercal b\}^2 \, \mathrm{d}P(x) = \mathbb{E}[\{f(X) - h(X)^\intercal b\}^2],$$

where the $\mathscr{X}$-valued random variable X has distribution P. We place ourselves in the random design setting and study the risk $L_f(\hat{\beta}_f)$ of the least squares estimator $\hat{\beta}_f$. This risk is a random variable whose randomness stems from the one of the training sample $X_1, \ldots, X_n$.

The main result of the paper concerns the *excess risk* $L_f(\hat{\beta}_f) - \min_{b \in \mathbb{R}^d} L_f(b)$ when $n \to \infty$ and $d \to \infty$, uniformly over $f \in \mathscr{F}$. That is, we study the convergence rate to zero of the random variable $\sup_{f \in \mathscr{F}} \{L_f(\hat{\beta}_f) - \min_{b \in \mathbb{R}^d} L_f(b)\}$. A key quantity is the *leverage function* $q : \mathscr{X} \to [0, \infty)$ defined by

$$\forall x \in \mathscr{X}, \qquad q(x) = h(x)^\intercal G^{-1} h(x),$$

where $G = \mathbb{E}[h(X)h(X)^\intercal]$ is the $d \times d$ Gram matrix of the feature map. The leverage function is the population version of the *statistical leverage* of a feature vector $h(X_i)$ in the linear regression model. It plays an important role when analyzing regression with random design (Hsu et al., 2014). Note that q does not change if the feature map is composed with an invertible linear transformation. Let $\varepsilon_f = f - h^\intercal \beta_f$ be the error function, with $\beta_f = \arg\min_{b \in \mathbb{R}^d} L_f(b)$ the risk-minimizing coefficient vector. Our main result, expressed in Corollary 1, is that

$$\sup_{f \in \mathscr{F}} \left\{ L_f(\hat{\beta}_f) - \min_{b \in \mathbb{R}^d} L_f(b) \right\} = \mathrm{O}_{\mathbb{P}} \left(\frac{\log n}{n} \sup_{f \in \mathscr{F}} \mathbb{E}[q(X)\varepsilon_f^2(X)] \right), \qquad n \to \infty.$$

Apart from the fact that the obtained bound is invariant under invertible linear transformations of the feature map, this result is remarkable for the three following reasons. First, the dimension d of the feature space is allowed to tend to infinity with the input sample size n at a speed which depends on the leverage function q via Condition 1. Second, in case the class $\mathcal{F}$ contains only a single response function f, a simple analysis leads to a bound for the excess risk that scales as $\mathbb{E}[q(X)\varepsilon_f^2(X)]/n$, see Eq. (6), a bound that matches the one of our main result up to a logarithmic term. Third, the quantity $\mathbb{E}[q(X)\varepsilon_f^2(X)]$ takes over the role of the quantity $\sigma_f^2 d$ in the fixed-design setting, where σ_f^2 is the variance of the error variable in the linear model.

The uniformity in $f \in \mathscr{F}$ is achieved by a decomposition of a quadratic form in terms of a sample mean and a U-statistic in combination with a concentration inequality for the suprema of such statistics. The analysis is focused on the ordinary least squares estimator, whereas the extension to ridge regression as in Hsu et al. (2014) is left for further research.

Paper outline. The mathematical background is presented in Section 2. The main result and a sketch of its proof are presented in Sections 3 and 4, respectively. Detailed proofs are deferred to the appendices.

2 Learning multiple response functions simultaneously

Linear model and ordinary least squares estimator. Consider a collection $\mathscr{F}$ of functions $f : \mathscr{X} \to \mathbb{R}$ on a probability space $(\mathscr{X}, \mathscr{A}, P)$. We think of $f(x)$ as the real-valued response given an input $x \in \mathscr{X}$. Given an independent random sample $X_1, \ldots, X_n$ from P together with the associated responses $f(X_1), \ldots, f(X_n)$ for every $f \in \mathscr{F}$, we wish to learn the values $f(x)$ of the response functions $f \in \mathscr{F}$ for new but yet unobserved inputs $x \in \mathscr{X}$. To this end, we map the input space $\mathscr{X}$ into a feature space $\mathbb{R}^d$ via a feature map $h : \mathscr{X} \to \mathbb{R}^d : x \mapsto h(x) = (h_1(x), \ldots, h_d(x))^\intercal$. One of the feature functions h_j could be the constant function 1, corresponding to an intercept. The response functions are modelled as linear functionals of the mapped features with coefficients estimated by ordinary least squares. The approximation to the response function $f \in \mathscr{F}$ is thus

$$\forall x \in \mathscr{X}, \qquad \hat{f}(x) = h(x)^\intercal \hat{\beta}_f \quad \text{where} \quad \hat{\beta}_f = \operatorname*{arg\,min}_{b \in \mathbb{R}^d} \sum_{i=1}^n \{f(X_i) - h(X_i)^\intercal b\}^2.$$

We wish to control the learning error $\hat{f} - f$ uniformly in $f \in \mathscr{F}$.

Sharing the same inputs and the same feature map across multiple response functions brings computational gains. Classical least-squares theory yields

$$\forall f \in \mathscr{F}, \qquad \hat{\beta}_f = G_n^{-1} \frac{1}{n} \sum_{i=1}^{n} h(X_i) f(X_i) \quad \text{where} \quad G_n = \frac{1}{n} \sum_{i=1}^{n} h(X_i) h(X_i)^\intercal.$$

In case the empirical Gram matrix G_n is not invertible, a pseudo-inverse is used instead. It follows that the predicted responses are linear in the observed responses:

$$\forall f \in \mathscr{F}, x \in \mathscr{X}, \qquad \hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} w(x, X_i) f(X_i) \quad \text{where} \quad w(x, X_i) = h(x)^\intercal G_n^{-1} h(X_i).$$

The weights $w(x, X_i)$ do not depend on the response function f. This invariance is a computational advantage if multiple response functions $f \in \mathscr{F}$ are to be learned simultaneously.

Worst-case excess prediction risk. The response functions are modelled as linear functionals of the mapped features via

$$\forall f \in \mathscr{F}, \forall x \in \mathscr{X}, \qquad f(x) = h(x)^\intercal \beta_f + \varepsilon_f(x). \tag{1}$$

The coefficient vector β_f is defined as the minimizer over $b \in \mathbb{R}^d$ of the prediction risk

$$\forall f \in \mathscr{F}, \forall b \in \mathbb{R}^d, \qquad L_f(b) = \mathbb{E}[\{f(X) - h(X)^\intercal b\}^2].$$

Here, the expectation is taken with respect to a random feature X with distribution P and it is assumed that f and h have finite second moments. Let $G = \mathbb{E}[h(X)h(X)^\intercal]$ denote the $d \times d$ Gram matrix of the feature map, assumed to be positive definite. Classical least squares theory yields

$$\forall f \in \mathscr{F}, \qquad \beta_f = \operatorname*{arg\,min}_{b \in \mathbb{R}^d} L_f(b) = G^{-1} \mathbb{E}[h(X) f(X)].$$

Since $\varepsilon_f(X) = f(X) - h(X)^\intercal \beta_f$ is orthogonal to $h(X)$, i.e., $\mathbb{E}[h(X)\varepsilon_f(X)] = 0$, the *excess risk* associated to any other coefficient vector $b \in \mathbb{R}^d$ is

$$L_f(b) - L_f(\beta_f) = \mathbb{E}[\{h(X)^\intercal (b - \beta_f)\}^2] = (b - \beta_f)^\intercal G (b - \beta_f).$$

The optimal coefficient vector β_f is unknown, so we estimate it by the least-squares estimator $\hat{\beta}_f$. The expected squared error made by the approximated response function for a new input distributed according to P is

$$\int_{\mathscr{X}} \{\hat{f}(x) - f(x)\}^2 \, \mathrm{d}P(x) = L_f(\hat{\beta}_f) = L_f(\beta_f) + (\hat{\beta}_f - \beta_f)^\intercal \, G \, (\hat{\beta}_f - \beta_f). \tag{2}$$

The right-hand side of (2) decomposes the expected squared error into two parts.

- The first term, $L_f(\beta_f)$, is deterministic. It represents the *modelling error* stemming from the linear model in (1). Given the model, this term is incompressible and does not depend on the learning algorithm nor on the training data.
- The second term on the right-hand side in (2) is random. It represents the *learning error* due to the estimation step and the randomness of the training data.

It is on the learning error that we focus our analysis, with the particularity that we consider the error uniformly in the response function. Our object of interest is thus the *worst-case excess prediction risk*

$$\sup_{f \in \mathscr{F}} \{L_f(\hat{\beta}_f) - L_f(\beta_f)\} = \sup_{f \in \mathscr{F}} (\hat{\beta}_f - \beta_f)^\intercal \, G \, (\hat{\beta}_f - \beta_f). \tag{3}$$

We are interested in the rate at which this supremum tends to zero as the sample size n and the dimension d of the feature map tend to infinity.

It is to be emphasized that our setting is that of a random design. The excess prediction risk $L_f(\hat{\beta}_f) - L_f(\beta_f)$ is a nonnegative random variable that is constructed out of the random training sample $X_1, \ldots, X_n$. As is clear from (2), it incorporates the risk associated to a new and yet unobserved input $x \in \mathscr{X}$, averaged over P. The expression in (3) is thus a supremum over potentially infinitely many random variables, each variable being built on the same inputs.

Excess prediction risk of a single response. Fix a response function $f \in \mathscr{F}$. Let P_n denote the empirical distribution of the training sample $X_1, \dots, X_n$, assigning probability $1/n$ to each observed input. For a real-valued, vector-valued or matrix-valued function g on $\mathscr{X}$, expectations with respect to the unknown sampling distribution P and the empirical distribution P_n are denoted respectively by

$$P(g) = \mathbb{E}[g(X)] = \int_{\mathscr{X}} g(x)\,\mathrm{d}P(x), \qquad\qquad P_n(g) = \frac{1}{n}\sum_{i=1}^{n} g(X_i).$$

Both operators are linear: for instance, $P(Ag) = AP(g)$ and $P_n(Ag) = AP_n(g)$ if A is a linear map between Euclidean spaces of suitable dimension. Using this operator notation, we get

$$G = P(hh^\intercal), \qquad\qquad \beta_f = G^{-1}P(hf),$$
$$G_n = P_n(hh^\intercal), \qquad\qquad \hat{\beta}_f = G_n^{-1}P_n(hf).$$

Since $f = h^\intercal\beta_f + \varepsilon_f$, the estimated coefficient vector is

$$\hat{\beta}_f = G_n^{-1}P_n[h(h^\intercal\beta_f + \varepsilon_f)] = \beta_f + G_n^{-1}P_n(h\varepsilon_f).$$

The excess prediction risk is thus

$$L_f(\hat{\beta}_f) - L_f(\beta_f) = P_n(h\varepsilon_f)^\intercal G_n^{-1} G G_n^{-1} P_n(h^\intercal\varepsilon_f). \tag{4}$$

The predicted response functions $\hat{f}$ are linear combinations of the components $h_1, \dots, h_d$ of the feature map h. They only depend on the feature map h through the linear span of the functions $h_1, \dots, h_d$. Therefore, the predicted responses remain unchanged if we compose the feature map with an invertible linear transformation A of $\mathbb{R}^d$. Let $G^{1/2}$ be the unique symmetric square root matrix of G and let $G^{-1/2}$ be its inverse. The whitened feature map is $\hbar = G^{-1/2}h : \mathscr{X} \to \mathbb{R}^d$. If the random input X has distribution P, then $\mathbb{E}[\hbar(X)\hbar(X)^\intercal] = P(\hbar\hbar^\intercal) = I_d$, the $d \times d$ identity matrix. The empirical Gram matrix of the whitened feature map is

$$P_n(\hbar\hbar^\intercal) = G^{-1/2}P_n(hh^\intercal)G^{-1/2} = G^{-1/2}G_nG^{-1/2}.$$

Since $h = G^{1/2}\hbar$ and $P_n(\hbar\hbar^\intercal)^{-1} = G^{1/2}G_n^{-1}G^{1/2}$, the excess prediction risk in (4) becomes

$$L_f(\hat{\beta}_f) - L_f(\beta_f) = |P_n(\hbar\hbar^\intercal)^{-1}P_n(\hbar\varepsilon_f)|_2^2, \tag{5}$$

where $|y|_2 = (y^\intercal y)^{1/2}$ denotes the Euclidean norm of a vector $y \in \mathbb{R}^d$.

Since $P(\hbar\hbar^\intercal) = I_d$, it is reasonable to expect that $P_n(\hbar\hbar^\intercal)^{-1}$ is approximately equal to I_d, at least if n is large and d is not too large compared to n; see Lemma 1 below for a precise statement. In that case, the excess prediction risk in (5) is approximately equal to $|P_n(\hbar\varepsilon_f)|_2^2$. The expectation of the latter random variable can be easily calculated: since $P(\hbar\varepsilon_f) = 0$, the terms with $i \neq j$ in the double sum below vanish and we find

$$\mathbb{E}\left[|P_n(\hbar\varepsilon_f)|_2^2\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}[\varepsilon_f(X_i)\hbar(X_i)^\intercal\hbar(X_j)\varepsilon_f(X_j)] = \frac{1}{n}P(\hbar^\intercal\hbar\varepsilon_f^2). \tag{6}$$

The excess prediction risk for a single response function f can thus be expected to have an order of magnitude equal to $n^{-1}P(\hbar^\intercal\hbar\varepsilon_f)$. In comparison, in the fixed-design case, where the training sample is considered as non-random, the expected excess risk is equal to $\sigma_f^2 d/n$, where σ_f^2 is the error variance (Hsu et al., 2014). Eq. (6) motivates why in Theorem 1, the convergence rate for the worst-case prediction risk over the whole response family $\mathscr{F}$ involves the quantity $\sup_{f\in\mathscr{F}} P(\hbar^\intercal\hbar\varepsilon_f^2)$.

3 Convergence rate of the worst-case excess prediction risk

Notation. Consider an asymptotic setting where the size n of the training sample tends to infinity. The feature map may change with n: with a slight change of notation, we write henceforth

$$h_n = (h_{n,1}, \dots, h_{n,d_n})^\intercal : \mathscr{X} \to \mathbb{R}^{d_n}.$$

The feature dimension $d_n \geq 1$ depends on n and may tend to infinity. The whitened feature map is

$$\hbar_n = P(h_n h_n^\intercal)^{-1/2} h_n : \mathscr{X} \to \mathbb{R}^{d_n},$$

where the $d_n \times d_n$ Gram matrix $P(h_n h_n^\intercal)$ is supposed to be invertible; otherwise, we can omit some components $h_{n,j}$ without affecting the linear span of the component functions. The least squares coefficient vectors and modelling errors depend on n as well: we write

$$\forall f \in \mathscr{F}, \forall x \in \mathscr{X}, \qquad f(x) = h_n(x)^\intercal \beta_{n,f} + \varepsilon_{n,f}(x) \quad \text{where} \quad \beta_{n,f} = P(h_n h_n^\intercal)^{-1} P(h_n f).$$

The least squares estimator of $\beta_{n,f}$ is $\hat{\beta}_{n,f} = P_n(h_n h_n^\intercal)^{-1} P_n(h_n f)$.

In this setting, the *leverage function* $q_n : \mathscr{X} \to [0, \infty)$ is defined by

$$\forall x \in \mathscr{X}, \qquad q_n(x) = h_n(x)^\intercal P(h_n h_n^\intercal)^{-1} h_n(x) = \hbar_n(x)^\intercal \hbar_n(x) = |\hbar_n(x)|_2^2.$$

The name of q_n is derived from the notion of leverage of a design point in multiple linear regression. Note that q_n does not change if the feature map is composed with an invertible linear transformation. We always have $P(q_n) = \operatorname{tr}[P(\hbar_n \hbar_n^\intercal)] = d_n$, where $\operatorname{tr}(A)$ denotes the trace of a square matrix A.

Let $\mathbb{P}$ denote the probability measure on the probability space on which the random inputs $X_1, \ldots, X_n$, taking values in $\mathscr{X}$, are defined. For any sequence $(Y_n)_n$ of real-valued random variables on that space and for any positive sequence $(a_n)_n$, the expression $Y_n = \mathrm{O}_{\mathbb{P}}(a_n)$ as $n \to \infty$ signifies that Y_n / a_n is bounded in probability, that is, for every $\epsilon > 0$ there exists $K > 0$ such that $\limsup_{n\to\infty} \mathbb{P}(|Y_n| > a_n K) \leq \epsilon$. Similarly, the expression $Y_n = \mathrm{o}_{\mathbb{P}}(a_n)$ as $n \to \infty$ signifies that Y_n / a_n converges to zero in probability, that is, $\lim_{n\to\infty} \mathbb{P}(|Y_n| > a_n \epsilon) = 0$ for every $\epsilon > 0$. Our aim is to determine a positive sequence a_n such that $a_n \to 0$ and, under reasonable assumptions,

$$\sup_{f \in \mathscr{F}} \{L_f(\hat{\beta}_{n,f}) - L_f(\beta_{n,f})\} = \mathrm{O}_{\mathbb{P}}(a_n), \qquad n \to \infty.$$

Lastly, the supremum norm of a function $g : \mathscr{X} \to \mathbb{R}$ is denoted by $\|g\|_\infty = \sup_{x \in \mathscr{X}} |g(x)|$.

Conditions. The conditions under which the main result holds concern the leverage function q_n and the response functions $\mathscr{F}$.

Condition 1. *One of the two alternative conditions hold:*

(a) $P(q_n^2) = \mathrm{o}(n)$ *and* $\log(\|q_n\|_\infty) = \mathrm{O}(\log n)$ *as* $n \to \infty$*;*

(b) $\|q_n\|_\infty \log(2 d_n) = \mathrm{o}(n)$ *as* $n \to \infty$.

Since $P(q_n) = d_n$ and $P(q_n^2) \geq [P(q_n)]^2$, condition (a) implies that $d_n = \mathrm{o}(n^{1/2})$ as $n \to \infty$. As $P(q_n^2) \leq \|q_n\|_\infty P(q_n) = \|q_n\|_\infty d_n$, a sufficient condition for (a) moreover is that $\|q_n\|_\infty = \mathrm{o}(n/d_n)$, which is the leverage condition in Portier and Segers (2019). Here, we have just assumed that the speed at which $\|q_n\|_\infty$ tends to infinity is at most polynomial in n. Condition (b) implies that $d_n \log(2 d_n) = \mathrm{o}(n)$ as $n \to \infty$ but, compared to (a), imposes a stronger condition on $\|q_n\|_\infty$.

Condition 2. *The collection $\mathscr{F}$ of response functions admits a uniformly bounded envelope $F : \mathscr{X} \to [0, \infty)$, i.e., $|f(x)| \leq F(x)$ for any $f \in \mathscr{F}$ and any $x \in \mathscr{X}$, and $\|F\|_\infty$ is finite. In addition, $\mathscr{F}$ is suppose to be at most countably infinite.*

As $\|q_n\|_\infty$ and $\|F\|_\infty$ are finite, the collection of error functions $\{\varepsilon_{n,f} : f \in \mathscr{F}\}$ is uniformly bounded, see Lemma 2 in the appendices.

The assumption in Condition 2 that $\mathscr{F}$ is countable assures that suprema over random variables indexed by $f \in \mathscr{F}$ are measurable. Otherwise, probabilities involving such suprema would need to be replaced by outer probabilities (van der Vaart and Wellner, 1996, Part 1). In practice, the countability assumption is harmless insofar as $\mathscr{F}$ can usually be approximated by a countable dense subfamily anyway without affecting the value of the supremum (van der Vaart and Wellner, 1996, Section 2.3.3).

Covering numbers capture the complexity of a subset of a metric space and play a central role in a number of areas in information theory and statistics, including nonparametric function estimation, density estimation, empirical processes, and machine learning.

Definition 1 (Covering number)**.** *For a subset $\mathscr{F}$ of a metric space $(\mathscr{Y}, \rho)$, the η-covering number $\mathscr{N}(\mathscr{F}, \rho, \eta)$ is the smallest number of open ρ-balls of radius $\eta > 0$ required to cover $\mathscr{F}$, i.e.,*

$$\mathscr{N}(\mathscr{F}, \rho, \eta) = \min\left\{p \geq 1 : \exists f_1, \ldots, f_p \in \mathscr{Y},\ \mathscr{F} \subset \bigcup_{i=1}^{p} B_\rho(f_i, \eta)\right\},$$

where $B_\rho(f, \eta) = \{g \in \mathscr{Y} : \rho(g, f) < \eta\}$ for $f \in \mathscr{Y}$ and $\eta > 0$.

Definition 2 (VC-class)**.** *A class $\mathscr{F}$ of real functions on a measurable space $(\mathscr{X}, \mathscr{A})$ is called a VC-class (Vapnik–Chervonenkis) of parameters $(v, A) \in (0, \infty)^2$ with respect to the envelope F if for any $0 < \eta < 1$ and any probability measure Q on $(\mathscr{X}, \mathscr{A})$, we have*

$$\mathscr{N}\left(\mathscr{F}, L^2(Q), \eta\|F\|_{L^2(Q)}\right) \leq (A/\eta)^v.$$

In Definition 2, we view $\mathscr{F}$ as a subset of the metric space $L^2(Q) \equiv L^2(\mathscr{X}, \mathscr{A}, Q)$ of Q-square-integrable functions $f : \mathscr{X} \to \mathbb{R}$ equipped with the metric $\rho(f, g) = \|f - g\|_{L^2(Q)}$, where $\|h\|_{L^2(Q)} = [Q(h^2)]^{1/2}$ for measurable $h : \mathscr{X} \to \mathbb{R}$.

Condition 3. *With respect to the envelope F, the collection $\mathscr{F}$ is VC with parameters (v, A).*

Main result. The maximal error standard deviation is

$$\sigma_n = \sup_{f \in \mathscr{F}} [P(\varepsilon_{n,f}^2)]^{1/2},$$

which under Condition 2 is bounded by $[P(F^2)]^{1/2}$ since $P(\varepsilon_{n,f}^2) \leq P(f^2) \leq P(F^2)$ for every $f \in \mathscr{F}$ by orthogonality of $\varepsilon_{n,f}$ with h_n. The growth rate of the worst-case excess prediction risk will be expressed in terms of

$$\gamma_n = \sup_{f \in \mathscr{F}} [P(q_n \varepsilon_{n,f}^2)]^{1/2}, \qquad L_n = \|q_n\|_\infty^{1/2} \sup_{f \in \mathscr{F}} \|\varepsilon_{n,f}\|_\infty. \tag{7}$$

Clearly, $\gamma_n \leq L_n$. Write $a \vee b = \max(a, b)$ for real a and b.

Theorem 1 (Convergence rate of worst-case excess prediction risk)**.** *If Conditions 1, 2, and 3 hold and if $\log(1/\sigma_n) = \mathrm{O}(\log n)$ as $n \to \infty$, then*

$$\sup_{f \in \mathscr{F}} \{L_f(\hat{\beta}_f) - L_f(\beta_f)\} = \mathrm{O}_{\mathbb{P}}\left(\left(\gamma_n^2 \vee \frac{L_n^2 \log n}{\sqrt{n}}\right) \frac{\log n}{n}\right), \qquad n \to \infty.$$

The quantity γ_n^2 is related to the expectation of the functions $q_n \varepsilon_{n,f}^2$ while L_n^2 is related to their supremum norm. It is reasonable to hope that the latter will not be too large in comparison to the former. This motivates the growth rate (8) assumed in the next corollary.

Corollary 1. *In Theorem 1, if, in addition, also*

$$\|q_n\|_\infty \sup_{f \in \mathscr{F}} \|\varepsilon_{n,f}^2\|_\infty = \mathrm{O}\left(\frac{\sqrt{n}}{\log n} \sup_{f \in \mathscr{F}} P(q_n \varepsilon_{n,f}^2)\right), \qquad n \to \infty, \tag{8}$$

then

$$\sup_{f \in \mathscr{F}} \{L_f(\hat{\beta}_f) - L_f(\beta_f)\} = \mathrm{O}_{\mathbb{P}}\left(\frac{\log n}{n} \sup_{f \in \mathscr{F}} P(q_n \varepsilon_{n,f}^2)\right), \qquad n \to \infty. \tag{9}$$

Apart from the factor $\log n$, the convergence rate in (9) corresponds to the one for a single response function f as discussed in the paragraph around Eq. (6). The additional factor $\log n$ stems from a concentration inequality for U-statistics in combination with bounds on the covering numbers of function classes derived from $\mathscr{F}$.

4 Sketch of proof of Theorem 1

Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalue, respectively, of the symmetric matrix A. Consider the matrix norm $|A|_2 = \sup\{|Ay|_2 : y \in \mathbb{R}^d, |y|_2 \leq 1\}$ for $A \in \mathbb{R}^{d\times d}$. If A is symmetric and positive semi-definite, then $|A|_2 = \lambda_{\max}(A)$. If, moreover, A is positive definite, then $\lambda_{\max}(A^{-1}) = \{\lambda_{\min}(A)\}^{-1}$. In view of (5), the worst-case excess prediction risk is bounded by

$$\sup_{f\in\mathscr{F}}\{L_f(\hat{\beta}_f) - L_f(\beta_f)\} \leq [\lambda_{\min}\{P_n(\hbar_n\hbar_n^\intercal)\}]^{-2} \cdot \sup_{f\in\mathscr{F}} |P_n(\hbar_n\varepsilon_{n,f})|_2^2. \tag{10}$$

Under reasonable conditions permitting $d_n \to \infty$, the smallest eigenvalue of $P_n(\hbar_n\hbar_n^\intercal)$ remains bounded away from zero with high probability.

Lemma 1. *Suppose one of the following two conditions holds:*

(a) $P(q_n^2) = \mathrm{o}(n)$ *as* $n \to \infty$*;*
(b) $\|q_n\|_\infty \log(2d_n) = \mathrm{o}(n)$ *as* $n \to \infty$*.*

Then $P_n(\hbar_n\hbar_n^\intercal)$ *is invertible with probability tending to one and* $\lambda_{\min}\{P_n(\hbar_n\hbar_n^\intercal)\} \geq 1 + \mathrm{o}_{\mathbb{P}}(1)$ *as* $n \to \infty$*.*

The proof of Lemma 1 in case (a) builds upon Portier and Segers (2019, Lemmas 2 and 3), while the one in case (b) is based upon Leluc et al. (2019, Lemma A.2), relying on a matrix Chernoff inequality due to Tropp (2015, Theorem 5.1.1). The proof is given in Section A in the appendices.

In view of the bound (10) in combination with Lemma 1, it is sufficient to show the claimed convergence rate with the excess risk $L_f(\hat{\beta}_n) - L_f(\beta_f)$ replaced by $|P_n(\hbar_n\varepsilon_{n,f})|_2^2$. For $f \in \mathscr{F}$, define $g_{n,f} : \mathscr{X}^2 \to \mathbb{R}$ by

$$\forall (x,y) \in \mathscr{X}^2, \qquad g_{n,f}(x,y) = \varepsilon_{n,f}(x)\hbar_n(x)^\intercal\hbar_n(y)\varepsilon_{n,f}(y).$$

Note that $g_{n,f}(x,x) = q_n(x)\varepsilon_{n,f}^2(x)$ for $x \in \mathscr{X}$. The quantity of interest is bounded by

$$n^2|P_n(\hbar_n\varepsilon_{n,f})|_2^2 \leq nP(q_n\varepsilon_{n,f}^2) + \left|\sum_{i=1}^n\{(q_n\varepsilon_{n,f}^2)(X_i) - P(q_n\varepsilon_{n,f}^2)\}\right| + \left|\sum_{1\leq i\neq j\leq n} g_{n,f}(X_i,X_j)\right|. \tag{11}$$

We will bound the supremum over $f \in \mathscr{F}$ of the sum on the right-hand side of (11) by the sum of the suprema of the three terms.

- The first supremum is just $n\gamma_n^2$, which is dominated by the stated convergence rate.
- The second supremum is a supremum over a sum of independent and identically distributed random variables with mean zero and finite variance. We will find its rate of convergence to zero via Proposition 2 below.
- The third supremum involves a U-statistic of order two with a degenerate kernel: by orthogonality of h_n and $\varepsilon_{n,f}$, we have $\mathbb{E}[g_{n,f}(X,x)] = \mathbb{E}[\varepsilon_{n,f}(X)\hbar_n(X)^\intercal]\hbar_n(x)\varepsilon_{n,f}(x) = 0$ and similarly $\mathbb{E}[g_{n,f}(x,X)] = 0$ for every $x \in \mathscr{X}$. We determine its convergence rate via a concentration inequality due to Major (2006) quoted as Theorem 2 in the appendices.

For the suprema over $f \in \mathscr{F}$ of the second and third terms on the right-hand side of (11), we will need to control the covering numbers of the classes $\mathscr{G}_n$ and $\mathscr{G}_n^{(d)}$ ("*d*" for "diagonal") given by

$$\mathscr{G}_n = \{g_{n,f} : f \in \mathscr{F}\}, \qquad \mathscr{G}_n^{(d)} = \{q_n\varepsilon_{n,f}^2 : f \in \mathscr{F}\}. \tag{12}$$

We will find bounds on their covering numbers in terms of those of the ones of the collection of response functions $\mathscr{F}$. The bounds are of potentially independent interest. The proof of Proposition 1 is given in Section B in the appendices.

Proposition 1 (Preservation VC-class)**.** *Let $\mathscr{F}$ be a VC-class with parameters (v, A) with respect to the envelope function F. Assume that the associated residuals $\varepsilon_{n,f}$ are uniformly bounded, i.e., there exists $M_n > 0$ such that $\sup_{f\in\mathscr{F}}\|\varepsilon_{n,f}\|_\infty \leq M_n$. Then $\mathscr{G}_n$ and $\mathscr{G}_n^{(d)}$ defined in* (12) *are VC-classes with respect to the envelope $M_n^2\|q_n\|_\infty$ with parameters $(4v, 4A_n)$ and $(2v, 2A_n)$ respectively, where*

$$A_n := 8A\|F\|_\infty\|q_n\|_\infty/M_n. \tag{13}$$

As $\mathscr{G}_n^{(d)}$ is a VC-class of functions by Proposition 1, we will be able to bound the supremum over $f \in \mathscr{F}$ of the second term on the right-hand side in (11) by an application of the following result.

Proposition 2. *On the probability space $(\mathscr{X}, \mathscr{A}, P)$, let $\mathscr{H}_n$ be a sequence of VC-classes of parameters (w_n, B_n) with respect to envelopes $U_n \geq \sup_{h\in\mathscr{H}_n}\|h\|_\infty$. Suppose the following conditions hold:*

(i) $\tau_n^2 \geq \sup_{h\in\mathscr{H}_n} \operatorname{var}_P(h)$ *and* $2\tau_n \leq U_n$*;*
(ii) $w_n \geq 1$ *and* $B_n \geq 3\sqrt{e}$*;*
(iii) $w_n U_n^2 \log(B_n U_n/\tau_n) = \mathrm{O}(n\tau_n^2), \qquad n \to \infty.$

Then, for P_n the empirical distribution of an independent random sample $X_1, \ldots, X_n$ from P, we have

$$\sup_{h\in\mathscr{H}_n} |P_n\{h - P(h)\}| = \mathrm{O}_{\mathbb{P}}\left(\frac{\tau_n}{\sqrt{n}}\sqrt{w_n \log\left(\frac{B_n U_n}{\tau_n}\right)}\right), \qquad n \to \infty.$$

The proof of Proposition 2 is given in Section C in the appendices. In the proof, we bound the expectation of the supremum by combining a well-known symmetrization inequality (Devroye et al., 2013, Theorem 12.4) with Proposition 2.1 in Giné and Guillou (2001), and we will find a rate on the deviation of the supremum around its expectation by Corollary 3.4 in Talagrand (1994).

Following the above plan, the proof of Theorem 1 is given in detail in Section D in the appendices.

A Auxiliary lemmas

Proof of Lemma 1. The lemma states an asymptotic lower bound for the smallest eigenvalue of the empirical Gram matrix $P_n(\hbar_n\hbar_n^\intercal)$ under two alternative conditions, (a) or (b). The two cases are treated in the next two paragraphs.

Under assumption (a). Lemma 3 in Portier and Segers (2019) states that $P_n(h_n h_n^\intercal)$ and thus $P_n(\hbar_n\hbar_n^\intercal)$ fails to be invertible with probability at most $n^{-1}P(q_n^2)$. This probability tends to zero by assumption.

Recall the spectral norm $|\cdot|_2$ and let $|A|_F = (\sum_{i,j} A_{ij}^2)^{1/2}$ denote the Frobenius norm of a matrix A. Lemma 2 in Portier and Segers (2019) states that $\mathbb{E}[|P_n(\hbar_n\hbar_n^\intercal) - I_{d_n}|_F^2]$ is bounded by $n^{-1}P(q_n^2)$ and thus converges to zero as $n \to \infty$. But then the same is true for $\mathbb{E}[|P_n(\hbar_n\hbar_n^\intercal) - I_{d_n}|_2^2]$, since $|A|_2 \le |A|_F$ for any square matrix A. It follows that $|P_n(\hbar_n\hbar_n^\intercal) - I_{d_n}|_2 = \mathrm{o}_{\mathbb{P}}(1)$ as $n \to \infty$.

On the event that $P_n(\hbar_n\hbar_n^\intercal)$ is invertible, we have

$$\begin{aligned} |P_n(\hbar_n\hbar_n^\intercal)^{-1}|_2 &= |I_{d_n} + P_n(\hbar_n\hbar_n^\intercal)^{-1}\{I_{d_n} - P_n(\hbar_n\hbar_n^\intercal)\}|_2 \\ &\le 1 + |P_n(\hbar_n\hbar_n^\intercal)^{-1}|_2 \cdot |P_n(\hbar_n\hbar_n^\intercal) - I_{d_n}|_2 \end{aligned}$$

from which

$$\frac{1}{\lambda_{\min}\{P_n(\hbar_n\hbar_n^\intercal)\}} = |P_n(\hbar_n\hbar_n^\intercal)^{-1}|_2 \le \frac{1}{1 - |P_n(\hbar_n\hbar_n^\intercal) - I_{d_n}|_2} = 1 + \mathrm{o}_{\mathbb{P}}(1), \qquad n \to \infty.$$

Under assumption (b). Lemma A.2 in Leluc et al. (2019), which is based on Theorem 5.1.1 in Tropp (2015), states that for $0 < \delta < 1$ and for n sufficiently large such that $n > 2\|q_n\|_\infty \log(d_n/\delta)$, we have

$$\mathbb{P}\left[\lambda_{\min}\{P_n(\hbar_n\hbar_n^\intercal)\} \le 1 - \sqrt{(2/n)\|q_n\|_\infty \log(d_n/\delta)}\right] \le \delta.$$

By assumption, $\|q_n\|_\infty \log(d_n/\delta) = \mathrm{o}(n)$ as $n \to \infty$, for any $0 < \delta < 1$. It follows that $\lambda_{\min}\{P_n(\hbar_n\hbar_n^\intercal)\} \ge 1 - \mathrm{o}_{\mathbb{P}}(1)$ as $n \to \infty$. □

Lemma 2. *If Conditions 1 and 2 hold, then*

$$\sup_{f \in \mathscr{F}} \|\varepsilon_{n,f}\|_\infty \le \|F\|_\infty + [\|q_n\|_\infty P(F^2)]^{1/2} \le (1 + \|q_n\|_\infty^{1/2})\|F\|_\infty.$$

Proof of Lemma 2. Let $f \in \mathscr{F}$. We have $f = h_n^\intercal \beta_{n,f} + \varepsilon_{n,f}$ with $\beta_{n,f} = P(h_n h_n^\intercal)^{-1} P(h_n f)$ and $P(h_n \varepsilon_{n,f}) = 0$. Since $\hbar_n = P(h_n h_n^\intercal)^{-1/2} h_n$, we get $f = \hbar_n^\intercal P(\hbar_n f) + \varepsilon_{n,f}$. Now $\hbar_n$ and $\varepsilon_{n,f}$ are orthogonal while $P(\hbar_n\hbar_n^\intercal) = I_{d_n}$, so that

$$P(f^2) = |P(\hbar_n f)|_2^2 + P(\varepsilon_{n,f}^2) \ge |P(\hbar_n f)|_2^2.$$

It follows that

$$[\hbar_n^\intercal P(\hbar_n f)]^2 = q_n |P(\hbar_n f)|_2^2 \le q_n P(f^2) \le q_n P(F^2).$$

But then

$$|\varepsilon_{n,f}| \le |f| + |\hbar_n^\intercal P(\hbar_n f)| \le |F| + [q_n P(F^2)]^{1/2}.$$

Since $P(F^2) \le \|F\|_\infty^2$, the result follows. □

B Proof of Proposition 1

The idea of the proof is to create a grid of functions on $\mathscr{X}$ based on a covering of $\mathscr{F}$ to cover $\mathscr{E}_n = \{\varepsilon_{n,f} : f \in \mathscr{F}\}$. From this grid, we will deduce coverings of $\mathscr{G}_n$ and $\mathscr{G}_n^{(d)}$.

Step 1: covering of $\mathscr{E}_n$. By assumption, the class $\mathscr{F}$ is VC of parameters (v, A) with respect to an envelope F. That means that for any $0 < \eta < 1$ and for any probability measure Q on $\mathscr{X}$, we have

$$\mathscr{N}\left(\mathscr{F}, L^2(Q), \eta \|F\|_{L^2(Q)}\right) \leq \left(\frac{A}{\eta}\right)^v.$$

Moreover, when $\eta \geq 1$, the ball centered at the constant function equal to zero and with radius $\|F\|_{L^2(Q)}$ is enough to cover $\mathscr{F}$. Thus, for any positive choice of η, the covering number is bounded from above by

$$\forall \eta \in (0, \infty), \quad \mathscr{N}\left(\mathscr{F}, L^2(Q), \eta\right) \leq \left(\frac{A\|F\|_{L^2(Q)}}{\eta}\right)^v \vee 1.$$

Fix $\eta > 0$, write $\eta_P = \eta/(4\|q_n\|_\infty)$ and $\eta_Q = \eta/4$, and define the covering number

$$N_P = \mathscr{N}\left(\mathscr{F}, L^2(P), \eta_P/2\right), \qquad N_Q = \mathscr{N}\left(\mathscr{F}, L^2(Q), \eta_Q/2\right) \tag{14}$$

associated to the open balls

$$B_P(f, \delta) = \left\{g \in L^2(P) : \|g - f\|_{L^2(P)} < \delta\right\}, \quad B_Q(f, \delta) = \left\{g \in L^2(Q) : \|g - f\|_{L^2(Q)} < \delta\right\},$$

for $\delta > 0$. The balls in the definition of the covering numbers in (14) have their centers in $L^2(P)$ and $L^2(Q)$ but not necessarily in $\mathscr{F}$. At the price of doubling the radii, the triangle inequality permits us to find functions $f_1^{(P)}, \ldots, f_{N_P}^{(P)}$ and $f_1^{(Q)}, \ldots, f_{N_Q}^{(Q)}$ in $\mathscr{F}$ such that

$$\mathscr{F} \subset \bigcup_{i=1}^{N_P} B_P(f_i^{(P)}, \eta_P), \qquad \mathscr{F} \subset \bigcup_{j=1}^{N_Q} B_Q(f_j^{(Q)}, \eta_Q). \tag{15}$$

Therefore, the class $\mathscr{F}$ is covered by the union of the intersections between the balls, that is to say

$$\mathscr{F} \subset \bigcup_{\substack{1 \leq i \leq N_P \\ 1 \leq j \leq N_Q}} \left(B_P(f_i^{(P)}, \eta_P) \cap B_Q(f_j^{(Q)}, \eta_Q)\right). \tag{16}$$

Define the support of this covering as

$$\mathscr{S} = \left\{(i, j) \in \{1, \ldots, N_P\} \times \{1, \ldots, N_Q\} : B_P(f_i^{(P)}, \eta_P) \cap B_Q(f_j^{(Q)}, \eta_Q) \neq \varnothing\right\}.$$

For every $(i, j) \in \mathscr{S}$, we fix an arbitrary function $f_{i,j} \in B_P(f_i^{(P)}, \eta_P) \cap B_Q(f_j^{(Q)}, \eta_Q)$.

Let $f \in \mathscr{F}$ and let $(i, j) \in \mathscr{S}$ be such that $f \in B_P(f_i^{(P)}, \eta_P) \cap B_Q(f_j^{(Q)}, \eta_Q)$. We will show that $\|\varepsilon_{n,f} - \varepsilon_{n,f_{i,j}}\|_{L^2(Q)} \leq \eta$. Since f and $f_{i,j}$ belong to the same intersection in (16), we have

$$\|f - f_{i,j}\|_{L^2(P)} \leq 2\eta_P, \qquad \|f - f_{i,j}\|_{L^2(Q)} \leq 2\eta_Q. \tag{17}$$

The residual functions can be expressed in terms of the whitened feature map $\hbar_n$ via

$$\varepsilon_{n,f} = f - \hbar_n^\intercal P(\hbar_n f), \qquad \varepsilon_{n,f_{i,j}} = f_{i,j} - \hbar_n^\intercal P(\hbar_n f_{i,j}).$$

By the triangle inequality, we find

$$\|\varepsilon_{n,f} - \varepsilon_{n,f_{i,j}}\|_{L^2(Q)} \leq \|f - f_{i,j}\|_{L^2(Q)} + \|\hbar_n^\intercal P[\hbar_n(f - f_{i,j})]\|_{L^2(Q)}. \tag{18}$$

Recall $M_n \geq \sup_{f \in \mathscr{F}} \|\varepsilon_{n,f}\|_\infty$, a constant envelope for the class $\mathscr{E}_n$, and recall the leverage function $q_n = |\hbar_n|_2^2$. By the Cauchy–Schwarz inequality,

$$\begin{aligned}
\|\hbar_n^\intercal P[\hbar_n(f - f_{i,j})]\|_{L^2(Q)}^2 &= \int_{y \in \mathscr{X}} \left\{\hbar_n(y)^\intercal \int_{x \in \mathscr{X}} \hbar_n(x)(f - f_{i,j})(x) \, \mathrm{d}P(x)\right\}^2 \mathrm{d}Q(y) \\
&\leq \int_{y \in \mathscr{X}} |\hbar_n(y)|_2^2 \left|\int_{x \in \mathscr{X}} \hbar_n(x)(f - f_{i,j})(x) \, \mathrm{d}P(x)\right|_2^2 \mathrm{d}Q(y) \\
&\leq \int_{y \in \mathscr{X}} |\hbar_n(y)|_2^2 \int_{x \in \mathscr{X}} |\hbar_n(x)|_2^2 (f - f_{i,j})^2(x) \, \mathrm{d}P(x) \, \mathrm{d}Q(y) \\
&\leq \|q_n\|_\infty^2 \|f - f_{i,j}\|_{L^2(P)}^2.
\end{aligned} \tag{19}$$

The combination of (17), (18) and (19) yields

$$\|\varepsilon_{n,f} - \varepsilon_{n,f_{i,j}}\|_{L^2(Q)} \leq 2\eta_Q + 2\|q_n\|_\infty \eta_P = \eta/2 + \eta/2 = \eta.$$

We have thus constructed the covering

$$\mathscr{E}_n \subset \bigcup_{(i,j)\in\mathscr{S}} B_Q(\varepsilon_{n,f_{i,j}}, \eta)$$

of $\mathscr{E}_n$ with $L^2(Q)$ balls of radius at most η. The covering number of $\mathscr{E}_n$ is thus bounded by

$$\mathscr{N}\left(\mathscr{E}_n, L^2(Q), \eta\right) \leq \#\mathscr{S} \leq \mathscr{N}\left(\mathscr{F}, L^2(P), \eta_P/2\right) \cdot \mathscr{N}\left(\mathscr{F}, L^2(Q), \eta_Q/2\right).$$

Using the definition of a VC-class, we get that

$$\begin{aligned}
&\mathscr{N}\left(\mathscr{E}_n, L^2(Q), \eta\right) \\
&\leq \left(\frac{2A\|F\|_{L^2(P)}}{\eta_P} \vee 1\right)^v \cdot \left(\frac{2A\|F\|_{L^2(Q)}}{\eta_Q} \vee 1\right)^v \\
&\leq \left(\frac{64A^2\|F\|_\infty^2\|q_n\|_\infty}{\eta^2}\right)^v \vee \left(\frac{8A\|F\|_\infty\|q_n\|_\infty}{\eta}\right)^v \vee 1 \\
&= \left(\frac{64A^2\|F\|_\infty^2\|q_n\|_\infty}{\eta^2}\right)^v \mathbb{1}_{\eta\leq t_-} + \left(\frac{8A\|F\|_\infty\|q_n\|_\infty}{\eta}\right)^v \mathbb{1}_{t_-<\eta<t_+} + \mathbb{1}_{t_+\leq\eta} \\
&\leq \left(\frac{A_n M_n}{\eta}\right)^{2v} \vee 1,
\end{aligned}$$

with $t_- = 8A\|F\|_\infty, t_+ = 8A\|F\|_\infty\|q_n\|_\infty$ and

$$A_n = 8A\|F\|_\infty\|q_n\|_\infty / M_n. \tag{20}$$

We deduce that the residual class $\mathscr{E}_n$ is VC of parameters $(2v, A_n)$ with respect to the envelope M_n.

Step 2: covering of $\mathscr{G}_n$. Consider two functions $f, \tilde{f} \in \mathscr{F}$ and a probability measure Q on $\mathscr{X}^2$ with marginals Q_1, Q_2 on $\mathscr{X}$, that is, $Q_1(B) = Q(B \times \mathscr{X})$ and $Q_2(B) = Q(\mathscr{X} \times B)$ for measurable $B \subset \mathscr{X}$. By definition, every function in $\mathscr{G}_n$ is written as $(x, y) \mapsto \hbar_n(x)^\intercal \hbar_n(y) \varepsilon_{n,f}(x)\varepsilon_{n,f}(y)$. The Cauchy–Schwarz inequality gives

$$\forall (x,y) \in \mathscr{X}^2, \qquad |\hbar_n(x)^\intercal \hbar_n(y)|^2 \leq |\hbar_n(x)|_2^2 \, |\hbar_n(y)|_2^2 = q_n(x)q_n(y) \leq \|q_n\|_\infty^2. \tag{21}$$

Furthermore, since $(a+b)^2 \leq 2a^2 + 2b^2$ for $a, b \in \mathbb{R}$, we have

$$\begin{aligned}
&\{\varepsilon_{n,f}(x)\varepsilon_{n,f}(y) - \varepsilon_{n,\tilde{f}}(x)\varepsilon_{n,\tilde{f}}(y)\}^2 \\
&\leq 2\varepsilon_{n,f}^2(x)\{\varepsilon_{n,f}(y) - \varepsilon_{n,\tilde{f}}(y)\}^2 + 2\{\varepsilon_{n,f}(x) - \varepsilon_{n,\tilde{f}}(x)\}^2 \varepsilon_{n,\tilde{f}}^2(y) \\
&\leq 2M_n^2[\{\varepsilon_{n,f}(y) - \varepsilon_{n,\tilde{f}}(y)\}^2 + \{\varepsilon_{n,f}(x) - \varepsilon_{n,\tilde{f}}(x)\}^2]
\end{aligned}$$

for all $(x, y) \in \mathscr{X}^2$. We obtain

$$\begin{aligned}
&\|g_{n,f} - g_{n,\tilde{f}}\|_{L^2(Q)}^2 \\
&= \int_{(x,y)\in\mathscr{X}^2} |\hbar_n(x)^\intercal \hbar_n(y)|^2 \left\{\varepsilon_{n,f}(x)\varepsilon_{n,f}(y) - \varepsilon_{n,\tilde{f}}(x)\varepsilon_{n,\tilde{f}}(y)\right\}^2 \mathrm{d}Q(x,y) \\
&\leq 2\|q_n\|_\infty^2 M_n^2 \left(\|\varepsilon_{n,f} - \varepsilon_{n,\tilde{f}}\|_{L^2(Q_1)}^2 + \|\varepsilon_{n,f} - \varepsilon_{n,\tilde{f}}\|_{L^2(Q_2)}^2\right).
\end{aligned} \tag{22}$$

Fix $\eta > 0$. Following the approach in Step 1, we can for $\ell = 1, 2$ construct a covering of $\mathscr{E}_n$ by $L^2(Q_\ell)$-balls of at most radius η. The centers of the balls are of the form

$$\forall \ell = 1, 2, \ \forall k = 1, \ldots, m_\ell, \qquad \varepsilon_{n,f_k^{(\ell)}} \in \mathscr{E}_n,$$

where $f_k^{(\ell)}$ belongs to $\mathscr{F}$ and where m_ℓ is the number of such balls needed, a number which is bounded by $(A_n M_n/\eta)^{2v} \vee 1$ for A_n defined in (20). Consider the intersections

$$\forall i = 1, \ldots, m_1,\ \forall j = 1, \ldots, m_2, \qquad B(i,j) = B_{Q_1}(\varepsilon_{n,f_i^{(1)}}, \eta) \cap B_{Q_2}(\varepsilon_{n,f_j^{(2)}}, \eta).$$

The set $\mathscr{E}_n$ is covered by the union of all those intersections $B(i,j)$. For each $(i,j) \in \{1, \ldots, m_1\} \times \{1, \ldots, m_2\}$ such that $\mathscr{E}_n$ intersects $B(i,j)$, pick an arbitrary $f_{i,j} \in \mathscr{F}$ such that $\varepsilon_{n,f_{i,j}} \in \mathscr{E}_n \cap B(i,j)$. Note that the functions $f_{i,j}$ are different from the ones denoted in the same way in Step 1.

Let $f \in \mathscr{F}$ and let (i,j) be such that $\varepsilon_{n,f}$ and $\varepsilon_{n,f_{i,j}}$ belong to the same intersection $B(i,j)$. Since the diameters of the two balls in the definition of $B(i,j)$ are bounded by 2η in view of the triangle inequality, we find

$$\forall \ell = 1, 2, \qquad \|\varepsilon_{n,f} - \varepsilon_{n,f_{i,j}}\|_{L^2(Q_\ell)} \le 2\eta.$$

By (22), it follows that

$$\|g_{n,f} - g_{n,f_{i,j}}\|^2_{L^2(Q)} \le 2\|q_n\|^2_\infty M_n^2[(2\eta)^2 + (2\eta)^2] = 16\|q_n\|^2_\infty M_n^2 \eta^2.$$

We find that $\mathscr{G}_n$ is covered by the union of the balls $B_Q(g_{n,f_{i,j}}, 4\|q_n\|_\infty M_n \eta)$. Its covering number is thus bounded by

$$\mathscr{N}\left(\mathscr{G}_n, L^2(Q), 4\|q_n\|_\infty M_n \eta\right) \le m_1 m_2 \le \left(\frac{A_n M_n}{\eta}\right)^{4v} \vee 1.$$

Rescaling $M_n \eta' = 4\eta$, we get

$$\mathscr{N}\left(\mathscr{G}_n, L^2(Q), \|q_n\|_\infty M_n^2 \eta'\right) \le \left(\frac{4A_n}{\eta'}\right)^{4v} \mathbb{1}_{\eta' < 4} + \mathbb{1}_{\eta' \ge 4} \le \left(\frac{4A_n}{\eta'}\right)^{4v} \vee 1.$$

In view of (21), the functions in $\mathscr{G}_n$ are uniformly bounded by $\|q_n\|_\infty M_n^2$. Since Q was an arbitrary probability measure on $\mathscr{X}^2$, we conclude that $\mathscr{G}_n$ is a VC-class with parameters $(4v, 4A_n)$ with respect to the constant envelope $\|q_n\|_\infty M_n^2$.

Step 3: covering of $\mathscr{G}_n^{(d)}$. Let Q be a probability measure on $\mathscr{X}$ and let $\eta > 0$. In Step 1, we found functions $f_{i,j} \in \mathscr{F}$ such that $\mathscr{E}_n$ is covered by the balls $B_Q(\varepsilon_{n,f_{i,j}}, \eta)$, and we needed at most $(A_n M_n/\eta)^{2v}$ of such functions. For $f \in \mathscr{F}$ we can thus find (i,j) such that

$$\|\varepsilon_{n,f} - \varepsilon_{n,f_{i,j}}\|_{L^2(Q)} \le \eta$$

and thus

$$\begin{aligned}
\|q_n \varepsilon^2_{n,f} - q_n \varepsilon^2_{n,f_{i,j}}\|_{L^2(Q)} &= \int_{x \in \mathscr{X}} \left\{q_n(x)\varepsilon^2_{n,f}(x) - q_n(x)\varepsilon^2_{n,f_{i,j}}(x)\right\}^2 \mathrm{d}Q(x) \\
&\le 4\|q_n\|^2_\infty M_n^2 \|\varepsilon_{n,f} - \varepsilon_{n,f_{i,j}}\|^2_{L^2(Q)} \\
&\le 4\|q_n\|^2_\infty M_n^2 \eta^2.
\end{aligned}$$

It follows that the number of $L^2(Q)$ balls of radius $\sqrt{4\|q_n\|^2_\infty M_n^2 \eta^2}$ needed to cover $\mathscr{G}_n^{(d)}$ is bounded by the number of functions $f_{i,j}$ in the construction in Step 1, and so

$$\begin{aligned}
\mathscr{N}\left(\mathscr{G}_n^{(d)}, L^2(Q), 2\|q_n\|_\infty M_n \eta\right) &\le \mathscr{N}\left(\mathscr{E}_n, L^2(Q), \eta\right)^2 \\
&\le \left(\frac{A_n M_n}{\eta}\right)^{2v} \mathbb{1}_{\eta < M_n} + \mathbb{1}_{\eta \ge M_n}.
\end{aligned}$$

Upon rescaling $M_n \eta' = 2\eta$, we find

$$\begin{aligned}
\mathscr{N}\left(\mathscr{G}_n^{(d)}, L^2(Q), \|q_n\|_\infty M_n^2 \eta'\right) &\le \left(\frac{2A_n}{\eta'}\right)^{2v} \mathbb{1}_{\eta' < 2} + \mathbb{1}_{\eta' \ge 2} \\
&\le \left(\frac{2A_n}{\eta'}\right)^{2v} \vee 1.
\end{aligned}$$

We conclude that $\mathscr{G}_n^{(d)}$ is a VC-class with parameters $(2v, 2A_n)$ with respect to the constant envelope $\|q_n\|_\infty M_n^2$. The proof of Proposition 1 is complete.

C Proof of Proposition 2

In this proof, we consider $Z_n := \sup_{h \in \mathscr{H}_n} |P_n(h) - P(h)|$. Thanks to the triangle inequality, we get

$$Z_n \leq \mathbb{E}(Z_n) + |Z_n - \mathbb{E}(Z_n)|. \tag{23}$$

We treat the two terms on the right-hand side of (23) in Steps 1 and 2, respectively. The convergence rate for Z_n then follows by adding both bounds in Step 3.

Step 1. Let $(\eta_i)_i$ denote a sequence of independent Rademacher variables, that is, $\mathbb{P}(\eta_i = +1) = \mathbb{P}(\eta_i = -1) = 1/2$ for all i, and such that $(\eta_i)_i$ and $(X_i)_i$ are independent. The symmetrization inequality (Devroye et al., 2013, Theorem 12.4) gives

$$n\,\mathbb{E}(Z_n) \leq 2\,\mathbb{E}\left[\sup_{h \in \mathscr{H}_n} \left| \sum_{i=1}^{n} \eta_i \{h(X_i) - P(h)\} \right| \right].$$

Further, Proposition 2.1 in Giné and Guillou (2001) shows the existence of a universal constant $C > 0$ such that

$$\mathbb{E}\left[\sup_{\mathscr{H}_n} \left| \sum_{i=1}^{n} \eta_i \{h(X_i) - P(h)\} \right| \right] \leq C \left(2 w_n U_n \log(2\theta_n) + \tau_n \sqrt{w_n n \log(2\theta_n)} \right),$$

where we wrote

$$\theta_n = B_n U_n / \tau_n.$$

By the asymptotic relation in (iii), the latter inequality is bounded from above by $\tilde{C} r_n$ where

$$\begin{aligned} r_n &= \tau_n \sqrt{w_n n \log(\theta_n)}, \\ \tilde{C} &= 2C \left[1 + \sup_n \sqrt{w_n U_n^2 \log(2\theta_n)/(n\tau_n^2)} \right]. \end{aligned} \tag{24}$$

We find that $n\,\mathbb{E}(Z_n) \leq \tilde{C} r_n$.

Step 2. Corollary 3.4 in Talagrand (1994) states the existence of universal constants $K, L > 0$ such that

$$\forall t > 0, \qquad \mathbb{P}\left(n|Z_n - \mathbb{E}(Z_n)| \geq r_n t\right) \leq K \exp\left(-\frac{r_n t}{2KU_n} \log\left(1 + 2t \frac{r_n U_n}{R_n} \right) \right) \tag{25}$$

where r_n is defined in (24) and where

$$R_n = \left(\sqrt{n}\tau_n + 2LU_n \sqrt{w_n \log(2\theta_n)} \right)^2.$$

Since $\log(1+x) \geq x/(1+x/2)$ for all $x \geq 0$, we get

$$\frac{r_n t}{2KU_n} \log\left(1 + 2t \frac{r_n U_n}{R_n} \right) \geq \frac{r_n^2 t^2}{K R_n \left(1 + \frac{r_n U_n t}{R_n} \right)}. \tag{26}$$

Eventually, we deduce from (25) and (26) that

$$\forall t > 0, \qquad \mathbb{P}\left(n|Z_n - \mathbb{E}(Z_n)| \geq r_n t\right) \leq K \exp\left(-\frac{r_n^2 t^2}{K R_n \left(1 + \frac{r_n U_n t}{R_n} \right)} \right).$$

For any $\delta \in (0, 1]$, solving in t the equation $K \exp\left(-r_n^2 t^2 / \{K R_n (1 + r_n U_n t / R_n)\}\right) = \delta$ gives

$$t_n^{(\delta)} = K(U_n/2r_n) \log(K/\delta) + \sqrt{K^2 (U_n/2r_n)^2 \log(K/\delta)^2 + K(R_n/r_n^2) \log(K/\delta)}. \tag{27}$$

The first term U_n/r_n in (27) can be rewritten as

$$\begin{aligned}\frac{U_n}{r_n} &= \frac{U_n}{\tau_n\sqrt{w_n n\log(\theta_n)}} \\ &= \frac{U_n\sqrt{w_n\log(\theta_n)}}{\sqrt{n}\tau_n}\{w_n\log(\theta_n)\}^{-1}.\end{aligned}$$

while the second argument R_n/r_n^2 in (27) is bounded from above by

$$\begin{aligned}\frac{R_n}{r_n^2} &\leq \frac{4\max\left\{n\tau_n^2, 4L^2U_n^2 w_n\log(2\theta_n)\right\}}{n\tau_n^2 w_n\log(\theta_n)} \\ &\leq 4\max\left\{[w_n\log(\theta_n)]^{-1}, \frac{4L^2U_n^2}{n\tau_n^2}\frac{\log(2\theta_n)}{\log(\theta_n)}\right\}.\end{aligned}$$

By the assumption (iii), we obtain $\sup_n U_n/r_n < \infty$ and $\sup_n R_n/r_n^2 < \infty$. For any $\delta \in (0,1]$, it follows that $\sup_n t_n^{(\delta)} < \infty$.

Step 3. Combining the bound (23) with the rates found in Steps 1 and 2 for the two terms in the bound, we obtain

$$\forall\delta\in(0,1], \quad \mathbb{P}\left(nZ_n > (\sup_n t_n^{(\delta)} + \tilde{C})r_n\right) \leq \delta.$$

Divide by n and plug in the definition (24) of r_n to conclude the proof of Proposition 2.

D Proof of Theorem 1

We recall and introduce the following quantities:

$$\begin{aligned}\sigma_n &= \sup_{f\in\mathscr{F}}[P(\varepsilon_{n,f}^2)]^{1/2}, \qquad M_n = \sup_{f\in\mathscr{F}}\|\varepsilon_{n,f}\|_\infty, \\ \gamma_n &= \sup_{f\in\mathscr{F}}[P(q_n\varepsilon_{n,f}^2)]^{1/2}, \quad L_n = \|q_n\|_\infty^{1/2}M_n.\end{aligned} \tag{28}$$

For all $f\in\mathscr{F}$, we have $P(\varepsilon_{n,f}^2) \leq P(f^2) \leq P(F^2)$. Since $P(q_n) = d_n$, we find

$$\begin{aligned}&\sigma_n \leq M_n \wedge [P(F^2)]^{1/2} \leq M_n \wedge \|F\|_\infty, \\ &\gamma_n \leq d_n^{1/2}M_n \wedge \|q_n\|_\infty^{1/2}\sigma_n \leq L_n.\end{aligned} \tag{29}$$

Step 1. We follow the plan laid out in Section 4 in the paper. In view of (10) and Lemma 1, we have

$$\sup_{f\in\mathscr{F}}\{L_f(\hat{\beta}_{n,f}) - L_f(\beta_f)\} \leq \{1+\mathrm{O}_{\mathbb{P}}(1)\}\sup_{f\in\mathscr{F}}|P_n(\hbar\varepsilon_{n,f})|_2^2.$$

The inequality (11) provides a bound on $n^2|P_n(\hbar_n\varepsilon_{n,f})|_2^2$ consisting of a sum of three terms. The first term is just $nP(q_n\varepsilon_{n,f}^2) \leq n\gamma_n^2$. The second and third terms require a separate analysis (Steps 2 and 3). Finally, we collect the bounds to arrive at the stated rate (Step 4).

Step 2. We apply Proposition 2 to the class $\mathscr{G}_n^{(d)} = \{q_n\varepsilon_{n,f}^2 : f\in\mathscr{F}\}$ introduced in (12). We get

$$\sup_{g\in\mathscr{G}_n^{(d)}}\left|\sum_{i=1}^n\{g(X_i) - P(g)\}\right| = \mathrm{O}\left(\sqrt{n\tau^2 w\log(BU/\tau)}\right), \qquad n\to\infty, \tag{30}$$

where the positive quantities τ, U, w, B depend on n and must satisfy the following conditions:

(i) $\tau^2 \geq \sup_{f\in\mathscr{F}}\mathrm{var}[(q_n\varepsilon_{n,f}^2)(X_1)]$;

(ii) $U \geq \sup_{f \in \mathscr{F}} \|q_n \varepsilon_{n,f}^2\|_\infty$ and $U \geq 2\tau$;

(iii) (w, B) are VC parameters of $\mathscr{G}_n^{(d)}$ with respect to the envelope U, and $w \geq 1$ and $B \geq 3\sqrt{e}$;

(iv) $w\,(U/\tau)^2 \log(BU/\tau) = \mathrm{O}(n)$.

Recall $M_n = \sup_{f \in \mathscr{F}} \|\varepsilon_{n,f}\|_\infty$ and $L_n = \|q_n\|_\infty^{1/2} M_n$ in (28). We have $\sup_{f \in \mathscr{F}} \|q_n \varepsilon_{n,f}^2\|_\infty \leq L_n^2$ and therefore we put

$$U = 2L_n^2. \tag{31}$$

Further, we have

$$\mathrm{var}[(q_n \varepsilon_{n,f}^2)(X_1)] \leq \mathbb{E}[q_n^2(X_1)\varepsilon_{n,f}^4(X_1)] \leq \|q_n\|_\infty M_n^2 P(q_n \varepsilon_{n,f}^2) \leq L_n^2 \gamma_n^2.$$

To meet (i) and at the same time have an upper bound on U/τ, we set

$$\tau = L_n \gamma_n \vee \frac{L_n^2 \sqrt{\log n}}{\sqrt{n}}. \tag{32}$$

Since $\gamma_n \leq L_n$, conditions (i) and (ii) are met.

Next, Proposition 1 says that, for any probability measure Q on $\mathscr{X}$ and any $0 < \eta < 1$,

$$\mathscr{N}\left(\mathscr{G}_n^{(d)}, L^2(Q), \eta\right) \leq \left(\frac{2A_n M_n^2 \|q_n\|_\infty}{\eta}\right)^{2v} = \left(\frac{A_n U}{\eta}\right)^{2v}$$

with A_n defined in (13). We can therefore set $w = \max(2v, 1)$, while for B it is sufficient to ensure that $B \geq \max(A_n, 3\sqrt{e})$. But

$$A_n \leq C_{\mathscr{F}} \|q_n\|_\infty / M_n \tag{33}$$

for some constant $C_{\mathscr{F}}$ that depends only on $\mathscr{F}$. Therefore, condition (iii) is met for

$$B = (C_{\mathscr{F}} \|q_n\|_\infty / M_n) \vee (3\sqrt{e}).$$

As $M_n \geq \sigma_n$, the conditions in the main theorem imply that $\log(B) = \mathrm{O}(\log n)$ as $n \to \infty$. By definition of U and τ in (31) and (32) respectively, we have $2 \leq U/\tau \leq 2(n/\log n)^{1/2}$. It follows that $\log(U/\tau) = \mathrm{O}(\log n)$ as $n \to \infty$ and thus that (iv) is met since

$$(U/\tau)^2 \log(BU/\tau) \leq 4(n/\log n)\,\mathrm{O}(\log n) = \mathrm{O}(n), \qquad n \to \infty.$$

We can now develop the rate obtained in (30). As $\log(BU/\tau) = \mathrm{O}(\log n)$ as $n \to \infty$, we get

$$\sup_{g \in \mathscr{G}_n^{(d)}} \left| \sum_{i=1}^n \{g(X_i) - P(g)\} \right| = \mathrm{O}\left(\left(L_n \gamma_n \vee \frac{L_n^2 \sqrt{\log n}}{\sqrt{n}} \right) \sqrt{n \log n} \right), \qquad n \to \infty. \tag{34}$$

In the computation of the rate of the excess prediction risk, we must divide the rate above by n^2.

Step 3. The term $\sum_{1 \leq i \neq j \leq n} g_{n,f}(X_i, X_j)$ is a degenerate U-statistic of order two. Note indeed that $\mathbb{E}[g_{n,f}(X, x)] = \mathbb{E}[g_{n,f}(x, X)] = 0$ for any $x \in \mathscr{X}$, where the random variable X has distribution P. We apply a special case of Theorem 2 in Major (2006), cited for convenience as Theorem 2 below. The functions $g_{n,f}$ are uniformly bounded by

$$\sup_{x,y \in \mathscr{X}} |g_{n,f}(x, y)| \leq \|q_n\|_\infty M_n^2 = L_n^2.$$

Consider the class of rescaled functions

$$\widetilde{\mathscr{G}}_n = \left\{ g_{n,f}/L_n^2 : f \in \mathscr{F} \right\}.$$

In view of Proposition 1, its covering numbers are bounded by

$$\mathscr{N}\left(\widetilde{\mathscr{G}}_n, L^2(Q), \eta\right) \leq (4A_n)^{4v} \eta^{-4v}$$

for any probability measure Q on $\mathscr{X}^2$ and for any $\eta \in (0,1]$. In the terminology of Major (2006, p. 490), this means that $\widetilde{\mathscr{G}}_n$ is an L^2-dense class of functions with parameter D and exponent w defined by

$$D = (4A_n)^{4v}, \qquad\qquad w = 4v \vee 1.$$

For u, we take the maximum with 1 in order to apply Theorem 2 later on.

By the Cauchy–Schwarz inequality, we have $g_{n,f}(x,y)^2 \leq q_n(x)\varepsilon_{n,f}^2(x)q_n(y)\varepsilon_{n,f}^2(y)$ for all $(x,y) \in \mathscr{X}^2$ and $f \in \mathscr{F}$. It follows that, for independent random variables X_1 and X_2 with common distribution P, we have

$$\forall f \in \mathscr{F}, \qquad \left(\mathbb{E}[g_{n,f}^2(X_1,X_2)]\right)^{1/2} \leq P(q_n\varepsilon_{n,f}^2) \leq \gamma_n^2.$$

Put

$$\upsilon = \frac{\gamma_n^2}{L_n^2} \vee \frac{\log n}{\sqrt{n}}.$$

Then $\upsilon^2 \in (0,1]$ is an upper bound on the second moment of the functions in $\widetilde{\mathscr{G}}_n$.

Theorem 2 yields that

$$\mathbb{P}\left(\sup_{f\in\mathscr{F}}\left|\sum_{1\leq i\neq j\leq n} g_{n,f}(X_i,X_j)\right| \geq 2n\upsilon L_n^2 y\right) \leq CD\exp(-\alpha y), \tag{35}$$

for all $y \in [y_{n,-}, y_{n,+}]$, where

$$\begin{aligned} y_{n,-} &= K\left(w + \left(\tfrac{\log D}{\log n}\right)_+\right)^{3/2} \log\left(\tfrac{2}{\upsilon}\right), \\ y_{n,+} &= n\upsilon^2, \end{aligned} \tag{36}$$

and where α, C and K are universal constants.

By definition, we have $1/\upsilon \leq \sqrt{n}/\log n$ and thus $\log(1/\upsilon) = \mathrm{O}(\log n)$ as $n \to \infty$. Furthermore,

$$\log(\|q_n\|_\infty) + |\log M_n| = \mathrm{O}(\log n), \qquad n \to \infty, \tag{37}$$

by assumption and by Lemma 2, whence $\log(D) = \mathrm{O}(\log n)$ as $n \to \infty$ in view of (33). It follows that

$$y_{n,-} = \mathrm{O}(\log n) = \mathrm{o}(y_{n,+}), \qquad n \to \infty,$$

and thus $y_{n,-} < y_{n,+}$ for all sufficiently large n.

Let $0 < \delta < 1$ and define

$$y(\delta) = \frac{1}{\alpha}\log\left(\frac{CD}{\delta}\right) \vee y_{n,-}.$$

Then $y(\delta) \in [y_{n,-}, y_{n,+}]$ for all sufficiently large n and moreover

$$CD\exp\{-\alpha y(\delta)\} \leq \delta.$$

Since moreover $y(\delta) = \mathrm{O}(\log n)$ as $n \to \infty$, we get from (35) and $\upsilon L_n^2 = \gamma_n^2 \vee L_n^2 n^{-1/2}\log n$ that

$$\sup_{f\in\mathscr{F}}\left|\sum_{1\leq i\neq j\leq n} g_{n,f}(X_i,X_j)\right| = \mathrm{O}_{\mathbb{P}}\left(\left(\gamma_n^2 \vee \frac{L_n^2\log n}{\sqrt{n}}\right) n\log n\right), \qquad n \to \infty. \tag{38}$$

Step 4. Assemble the bounds for the suprema over $f \in \mathscr{F}$ of the three terms on the right-hand side of (11). From (34) and 38, dividing by n^2, we get

$$\sup_{f\in\mathscr{F}}\{L_f(\hat{\beta}_{n,f}) - L_f(\beta_f)\} = \mathrm{O}\left(\frac{\gamma_n^2}{n}\right) + \mathrm{O}_{\mathbb{P}}\left(\left(L_n\gamma_n \vee \frac{L_n^2\sqrt{\log n}}{\sqrt{n}}\right)\frac{\sqrt{\log n}}{n\sqrt{n}}\right) + \mathrm{O}_{\mathbb{P}}\left(\left(\gamma_n^2 \vee \frac{L_n^2\log n}{\sqrt{n}}\right)\frac{\log n}{n}\right), \qquad n\to\infty.$$

As $\gamma_n \leq L_n$, the first and second terms are dominated by the third one. The convergence rate of the excess prediction risk is thus

$$\sup_{f\in\mathscr{F}}\{L_f(\hat{\beta}_{n,f}) - L_f(\beta_f)\} = \mathrm{O}_{\mathbb{P}}\left(\left(\gamma_n^2 \vee \frac{L_n^2\log n}{\sqrt{n}}\right)\frac{\log n}{n}\right), \qquad n\to\infty. \tag{39}$$

This completes the proof of Theorem 1.

E A concentration inequality for degenerate U-statistics

The main term in the proof of Theorem 1 concerned a degenerate U-statistic of order two, see Step 3 in Section D. We dealt with it via a special case of the concentration inequality in Theorem 2 in Major (2006), stated next.

Theorem 2 (Special case of Theorem 2 in Major (2006))**.** *Let $(\mathscr{X}, \mathscr{A}, P)$ be a probability space and let $\mathscr{G}$ be an at most countably infinite collection of measurable functions $g : \mathscr{X}^2 \to [-1, 1]$ such that $\int_{\mathscr{X}} g(x, z)\,\mathrm{d}P(z) = \int_{\mathscr{X}} g(z, x)\,\mathrm{d}P(z) = 0$ for every $x \in \mathscr{X}$. Assume that $\mathscr{G}$ is a countable VC-class of parameters (w, B), with $w \geq 1$ and $B > 0$. Let $\upsilon \in (0, 1]$ be such that $\sup_{g\in\mathscr{G}} \mathbb{E}[g^2(X_1, X_2)] \leq \upsilon^2$. Let $X_1, \ldots, X_n$ be an independent random sample from P. There exist universal positive constants α, C and K such that*

$$\forall y \in [y_-, y_+], \qquad \mathbb{P}\left(\sup_{g\in\mathscr{G}}\left|\sum_{1\leq i\neq j\leq n} g(X_i, X_j)\right| \geq 2n\upsilon\right) \leq CB^w e^{-\alpha y}$$

where

$$y_- = K[w + (w\log B/\log n)_+]^{3/2}\log(2/\upsilon), \qquad y_+ = n\upsilon^2.$$

Acknowledgments and Disclosure of Funding

The authors thank very particularly Rémi Leluc for his valuable feedback and gratefully acknowledge funding from FNRS-F.R.S. grant CDR J.0146.19.

References

Azzi, S., Y. Huang, B. Sudret, and J. Wiart (2019). Surrogate modeling of stochastic functions – application to computational electromagnetic dosimetry. *International Journal for Uncertainty Quantification 9*(4).

Bauer, B., F. Heimrich, M. Kohler, and A. Krzyżak (2019). On estimation of surrogate models for multivariate computer experiments. *Annals of the Institute of Statistical Mathematics 71*(1), 107–136.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics 186*(2), 345–366.

Breitkopf, P., H. Naceur, A. Rassineux, and P. Villon (2005). Moving least squares response surface approximation: formulation and metal forming applications. *Computers & Structures 83*(17-18), 1411–1428.

Bucher, C. G. and U. Bourgund (1990). A fast and efficient response surface approach for structural reliability problems. *Structural Safety 7*(1), 57–66.

Devroye, L., L. Györfi, and G. Lugosi (2013). *A Probabilistic Theory of Pattern Recognition*, Volume 31. New York: Springer Science+Business Media.

Forrester, A. I. and A. J. Keane (2009). Recent advances in surrogate-based optimization. *Progress in Aerospace Sciences 45*(1–3), 50–79.

Frean, M. and P. Boyle (2008). Using Gaussian processes to optimize expensive functions. In *Australasian Joint Conference on Artificial Intelligence*, pp. 258–267. Springer.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The Elements of Statistical Learning*. New York: Springer Science+Business Media.

Giné, E. and A. Guillou (2001). On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. *Annales de l'IHP Probabilités et Statistiques 37*(4), 503–522.

Györfi, L., M. Kohler, A. Krzyzak, and H. Walk (2006). *A Distribution-Free Theory of Nonparametric Regression*. New York: Springer Science & Business Media.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambride: Cambridge University Press.

Hsu, D., S. M. Kakade, and T. Zhang (2014). Random design analysis of ridge regression. *Foundations of Computational Mathematics 14*(3), 569–600.

Jones, D. R. (2001). A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization 21*(4), 345–383.

Konakli, K. and B. Sudret (2016). Polynomial meta-models with canonical low-rank approximations: Numerical insights and comparison to sparse polynomial chaos expansions. *Journal of Computational Physics 321*, 1144–1169.

Leluc, R., F. Portier, and J. Segers (2019). Control variate selection for monte carlo integration. *arXiv preprint arXiv:1906.10920*.

Major, P. (2006). An estimate on the supremum of a nice class of stochastic integrals and U-statistics. *Probability Theory and Related Fields 134*(3), 489–537.

Myers, R. H., D. C. Montgomery, and C. M. Anderson-Cook (2016). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments* (Fourth ed.). Hoboken, NJ: John Wiley & Sons.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics 79*(1), 147–168.

Nguyen, A.-T., S. Reiter, and P. Rigo (2014). A review on simulation-based optimization methods applied to building performance analysis. *Applied Energy 113*, 1043–1058.

Portier, F. and J. Segers (2019). Monte Carlo integration with a growing number of control variates. *Journal of Applied Probability 56*(4), 1168–1186.

Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. *The Annals of Probability 22*(1), 28–76.

Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning 8*(1-2), 1–230. arXiv:1501.01571.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Process. With Applications to Statistics*. New York: Springer-Verlag.

Zhou, S. and D. A. Wolfe (2000). On derivative estimation in spline regression. *Statistica Sinica 10*, 93–108.