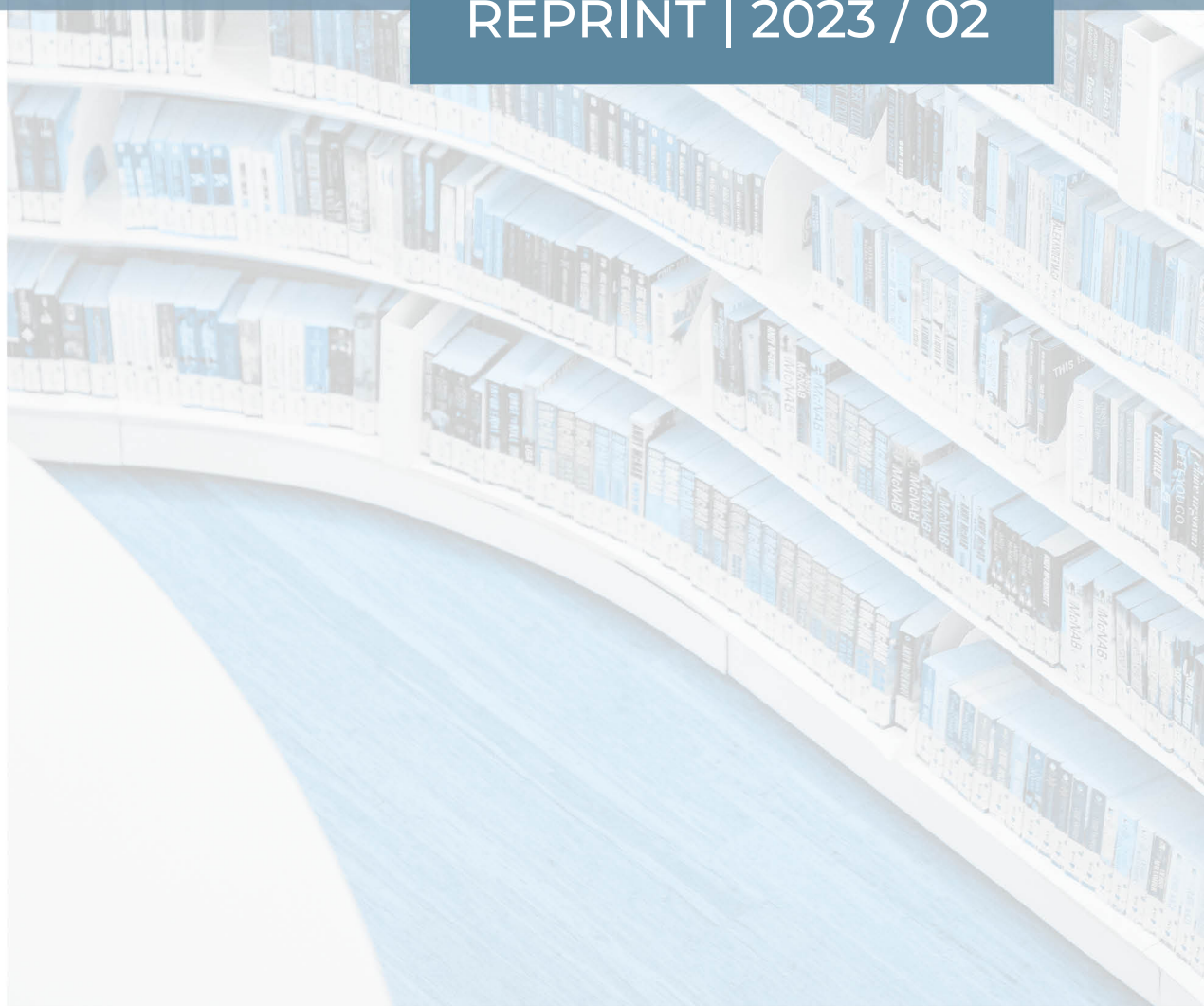
RISK BOUNDS WHEN LEARNING INFINITELY MANY RESPONSE FUNCTIONS BY ORDINARY LINEAR REGRESSION

Vincent Plassier, François Portier, Johan Segers

REPRINT | 2023 / 02



UCLouvain

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

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/isba/publication.html

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

Vincent Plassier
CMAP, École Polytechnique
Institut Polytechnique de Paris
Lagrange Mathematics and Computing Research Center
75007 Paris, France
vincent.plassier@ens-paris-saclay.fr

Francois Portier
LTCI, Télécom Paris and CREST, ENSAI
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

November 30, 2021

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.

Application to Monte Carlo integration with control variates. The past few years, Monte Carlo integration has received increasing interest because of its simplicity and its success facing complex high-dimensional approximation problems. The standard Monte Carlo error has a convergence rate of $1/\sqrt{n}$, independently of the dimension—see Novak (2016) for a review of deeper results around this point. Although a dimension-free convergence rate is comfortable, $1/\sqrt{n}$ is still relatively slow and some difficulties might arise in situations where we can only make a limited number of requests to the integrand. The use of control variates (Owen, 2013; Glasserman, 2013) has then represented an interesting avenue as it allows to reduce the variance of the standard Monte Carlo estimate without requiring additional evaluations of the integrand. Recently, it has been shown (Oates et al., 2017; Portier and Segers, 2019) that control variates allow to accelerate the $1/\sqrt{n}$ convergence rate substantially. However, whether this acceleration occurs when the error is measured uniformly over a class of integrand functions is, to the best of our knowledge, still unknown. Motivated by several applications in which uniform results are needed (see Section 5 for details), we obtain, as a by-product of the bound in Corollary 1, a uniform convergence rate for control variate Monte Carlo estimates. The fact that the rate established is faster than the standard Monte Carlo rate furnishes an additional argument for the use of control variates in Monte Carlo methods.

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. The application to control variate Monte Carlo methods is considered in Section 5. 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 $X_1, \ldots, X_n$, 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

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

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

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

The predicted response functions $\hat{f}$ are linear combinations of the components $h_1, \ldots, h_d$ of the feature map h. They only depend on the feature map h through the linear span of the functions $h_1, \ldots, h_d$. 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

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

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 $n^{-1}P(\hbar^\intercal\hbar\varepsilon_f^2)$. 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

3.1 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

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

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 $P(q_n) = \mathrm{tr}[P(\hbar_n \hbar_n^\intercal)] = d_n$, where $\mathrm{tr}(A)$ 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,

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

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

3.2 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) $\|q_n\|_\infty \log(2d_n) = \mathrm{o}(n)$ *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 $d_n \log(2d_n) = \mathrm{o}(n)$ 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 supposed 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 Appendix A.

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* $\eta > 0$.

For the definition of Vapnik–Chervonenkis (VC) classes, we follow Giné and Guillou (1999).

Definition 2 (VC-class)**.** *A class* $\mathscr{F}$ *of real functions on a measurable space* $(\mathscr{X}, \mathscr{A})$ *is called a VC-class of parameters* $(v, A) \in (0, \infty) \times [1, \infty)$ *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).

3.3 Main result

The maximal error is

$$M_n = \sup_{f \in \mathscr{F}} \|\varepsilon_{n,f}\|_\infty \tag{7}$$

and the growth rate of the worst-case excess prediction risk will be expressed in terms of

$$\gamma_n^2 = \sup_{f \in \mathscr{F}} P(q_n \varepsilon_{n,f}^2) \qquad \text{and} \qquad L_n^2 = M_n^2 \left\| q_n \right\|_\infty . \tag{8}$$

Clearly, $\gamma_n^2 \leq L_n^2$. For $a, b \in \mathbb{R}$, write $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. The positive part of $a \in \mathbb{R}$ is $(a)_+ = a \vee 0$.

Theorem 1 (Convergence rate of worst-case excess prediction risk)**.** *If Conditions 1, 2, and 3 hold, then, as* $n \to \infty$*,*

$$\sup_{f \in \mathscr{F}} \left\{ L_f(\hat{\beta}_f) - L_f(\beta_f) \right\} = \mathrm{O}_{\mathbb{P}} \left(\left\{ \gamma_n^2 \vee \left(\tau_n r_n^{1/2} \right) \right\} r_n \right), \qquad n \to \infty,$$

where

$$\tau_n = L_n^2 \left(1 \vee (a_n / M_n) \right) \qquad \textit{and}$$
$$r_n = 1 \wedge \left[\frac{\log n}{n} \left\{ 1 + \left(\frac{(\log M_n^{-1})_+}{\log n} \right)^{3/2} \right\} \right]$$

and where a_n *is defined in* (38) *and is of the order* $\mathrm{o}(\exp\{-n^{2/3}\})$ *as* $n \to \infty$*.*

When all the functions $f \in \mathscr{F}$ are in the vector space $\{\beta^\intercal h_n : \beta \in \mathbb{R}^{d_n}\}$, the maximal error becomes $M_n = 0$ and therefore we find that the worst-case excess prediction risk is equal to zero. The quantity γ_n^2 is related to the L^2-norm 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 assumptions in the next corollary.

Corollary 1 (Simplified rates)**.** *If* $(\log M_n^{-1})_+ = \mathrm{O}(\log n)$ *as* $n \to \infty$*, i.e., if there is some* $\alpha > 0$ *such that* $\liminf_{n \to \infty} n^\alpha M_n > 0$*, then*

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

If, moreover, $L_n^2 = \mathrm{O}\left(\gamma_n^2 \sqrt{n / \log n} \right)$*, then*

$$\sup_{f \in \mathscr{F}} \left\{ L_f(\hat{\beta}_f) - L_f(\beta_f) \right\} = \mathrm{O}\left(\gamma_n^2 \frac{\log n}{n} \right), \qquad n \to \infty. \tag{10}$$

If, on the other hand, the sequence with general term $(\log M_n^{-1})_+$ is of larger order than $\log n$, then for every function $f \in \mathscr{F}$, the sequence $(\left\| \varepsilon_{n,f} \right\|_\infty)_{n \in \mathbb{N}}$ converges very quickly to zero. This corresponds to the "ideal case" where the family $\mathscr{F}$ is well approximated by the chosen regressors $(h_{n,1}, \ldots, h_{n,d_n})$. In this case, (M_n) converges to zero faster than $(n^{-\alpha})$ for any $\alpha > 0$ and the general form of the rate in Theorem 1 implies that the worst-case excess prediction risk converges to zero very fast.

Apart from the factor $\log n$, the convergence rate in (10) 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

$$\sup_{f \in \mathscr{F}} \left\{ L_f(\hat{\beta}_f) - L_f(\beta_f) \right\} \leq \left\{ \lambda_{\min}\big(P_n(\hbar_n \hbar_n^\intercal) \big) \right\}^{-2} \cdot \sup_{f \in \mathscr{F}} |P_n(\hbar_n \varepsilon_{n,f})|_2^2 . \tag{11}$$

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. (2021, 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 (11) 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

$$n^2 |P_n(\hbar_n \varepsilon_{n,f})|_2^2 \leq n P_n(q_n \varepsilon_{n,f}^2) + \left| \sum_{1 \leq i \neq j \leq n} g_{n,f}(X_i, X_j) \right|. \tag{12}$$

We will bound the supremum over $f \in \mathscr{F}$ of the sum on the right-hand side of (12) by the sum of the suprema of the two terms.

- Using concentration inequalities established in Talagrand (1994) and Giné and Guillou (2001), we show that the first supremum is dominated by the stated convergence rate.
- The second supremum involves a U-statistic of order two with a degenerate kernel: by orthogonality of h_n and $\varepsilon_{n,f}$, we have $\mathbb{E}\left[g_{n,f}(X, x)\right] = \mathbb{E}\left[\varepsilon_{n,f}(X) \hbar_n(X)^\intercal\right] \hbar_n(x) \varepsilon_{n,f}(x) = 0$ and similarly $\mathbb{E}\left[g_{n,f}(x, X)\right] = 0$ for every $x \in \mathscr{X}$. We determine its convergence rate via a concentration inequality due to Major (2006) quoted as Theorem 4 in Appendix G.

To deal with suprema over $f \in \mathscr{F}$, we will need to control the covering numbers of the classes $\mathscr{G}_n$ and $\mathscr{G}_n^{(d)}$ ("d" for "diagonal") given by

$$\mathscr{G}_n = \{g_{n,f} : f \in \mathscr{F}\}, \qquad \mathscr{G}_n^{(d)} = \{q_n^{1/2} \varepsilon_{n,f} : f \in \mathscr{F}\}. \tag{13}$$

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 Appendix B.

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* (13) *are VC-classes with respect to the envelopes* $M_n^2 \|q_n\|_\infty$ *and* $M_n \|q_n\|_\infty^{1/2}$ *with parameters* $(4v, 4A_n)$ *and* $(2v, A_n)$ *respectively, where*

$$A_n = 8A \|F\|_\infty \|q_n\|_\infty^{1/2} / M_n. \tag{14}$$

Following the above plan, the proof of Theorem 1 is given in detail in Appendix C.

5 Uniform bound for Monte Carlo integration with control variates

This section investigates the application of the previous results to Monte Carlo estimates constructed with the help of control variates in order to reduce the variance. We start by presenting several applications in which uniform bounds for Monte Carlo methods are of interest. Then we give the mathematical background and finally provide sharp uniform error bounds for Monte Carlo estimates that use control variates.

5.1 Uniformity in Monte Carlo procedures

Proving uniform bounds on the error of Monte Carlo methods is motivated by the following three applications.

Latent variable models. Suppose we are interested in estimating the distribution of the variable X whose density is assumed to lie in the model

$$p_\theta(y) = \int p_\theta(y|z)\, p(z)\, \mathrm{d}z$$

where $\theta \in \Theta$ is the parameter to estimate, $p(z)$ is the known density of the so-called latent variable and $\{(y,z) \mapsto p_\theta(y|z)\}_{\theta\in\Theta}$ is a model of conditional densities (Y conditionally on the latent variable). This is actually a frequent situation in economics (McFadden, 2001) and medicine (McCulloch and Neuhaus, 2005, example 4, 6 and 9). Given independent and identically distributed random variables $Y_1, \ldots, Y_n$ observed from the previous model with parameter θ_0, the log-likelihood function takes the form

$$\theta \mapsto \sum_{i=1}^{n} \log\left(\int p_\theta(Y_i|z)\, p(z)\, \mathrm{d}z\right).$$

In most cases, each term in the previous sum is intractable and the approach proposed in McFadden and Ruud (1994) consists in replacing the unknown integrals by Monte Carlo estimates. In such a procedure, there is an additional estimation error compared to the statistical error of the standard maximum likelihood estimator. This additional error can be controlled in terms of the approximation error of $\int p_\theta(y|z)\, p(z)\, \mathrm{d}z$ by the Monte Carlo estimate uniformly in (y,θ).

Stochastic programming. Consider the stochastic optimization problem

$$\min_{\theta\in\Theta} F(\theta) \qquad \text{with} \qquad F(\theta) = \mathbb{E}[f(\theta, X)],$$

where X is a random variable in some space $\mathscr{X}$ with distribution P and where Θ is a Euclidean set. The response functions are thus the maps $x \mapsto f(\theta, x)$ as θ ranges over Θ. This problem is different from standard optimization because it takes into account some uncertainty in the output of the function f. One might think of the following toy example: f is the output of a laboratory experiment, e.g., the amount of salt in a solution, θ is the input of the experiment, e.g., the temperature of the solution and X gathers unobserved random factors that influence the output. Optimizing such kind of functions is of interest in many different fields and we refer the reader to Shapiro et al. (2014) for more concrete examples (see also the example below that deals with quantile estimation). This problem might be solved using two competitive approaches: the sample average approximation (SAA) and stochastic approximation techniques such as gradient descent—see Nemirovski et al. (2009) for a comparison between both approaches. The SAA approach follows from approximating the function F by a functional Monte Carlo estimate defined as

$$F_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} f(\theta, X_i),$$

where $X_1, \ldots, X_n$ is a random sample from P. Then the minimizer of the first stochastic optimization problem is approximated by the minimizer θ_n of the functional estimate F_n. Following the reference textbook Shapiro et al. (2014), the analysis of θ_n is carried out through error bounds for the Monte Carlo estimate $F_n(\theta)$ uniformly in $\theta \in \Theta$. When sampling from P is expensive, one may want to reduce the variance of the Monte Carlo estimate $F_n(\theta)$ by the method of control variates or some other method. In that case, a control on the error uniformly in θ is required.

Quantile estimation in simulation modeling. Quantiles are of prime importance when it comes to measure the uncertainty of random models (Law and Kelton, 2000). When the stochastic experiments are costly, variance reduction techniques such as the use of control variates are helpful (Hesterberg and Nelson, 1998; Cannamela et al., 2008). Let $F(y) = \mathbb{P}\{g(X) \leq y\}$, for $y \in \mathbb{R}$, be

the cumulative distribution function of a transformation $g : \mathscr{X} \to \mathbb{R}$ of a random element X in some space $\mathscr{X}$. Suppose the interest is in the quantile $F^-(u) = \inf\{y \in \mathbb{R} \,:\, F(y) \geq u\}$ for $u \in (0,1)$. The functions $f_y : \mathscr{X} \to \mathbb{R}$ to be integrated with respect to the distribution P of X are thus the indicators $x \mapsto f_y(x) = I\{g(x) \leq y\}$, indexed by $y \in \mathbb{R}$. It is necessary to control the accuracy of an estimate of the probability $F(y) = \mathbb{E}[f_y(X)]$ uniformly in $y \in \mathbb{R}$ in order to have a control on the accuracy of an estimate of the quantile $F^-(u)$, even for a single $u \in (0,1)$; see for instance Lemma 12 in Portier and Segers (2018). If drawing samples from P or evaluating g is expensive, it may be of interest to limit the number of Monte Carlo draws X_i and function evaluations $g(X_i)$. Finally, note that due to the formulation of a quantile as the minimiser of the expectation of the check function, see e.g., Hjort and Pollard (2011), this example is an instance of the stochastic programming framework described before.

Related results. The previous examples underline the need of error bounds for Monte Carlo methods that are uniform over a family of response functions. The uniform consistency of standard Monte Carlo estimates over certain collections of functions can easily be shown by relying on Glivenko-Cantelli classes (van der Vaart and Wellner, 1996); see for instance Shapiro et al. (2014) for applications to stochastic programming problems. Similarly, uniform convergence rates for standard Monte Carlo estimates can be derived from classical empirical process theory. We refer to Giné and Guillou (2002) and the references therein for suprema over VC-type classes and to Kloeckner (2020) and the references therein for suprema over Hölder-type classes. For variance reduction methods based on adaptive importance sampling, uniform consistency has been proven recently in Delyon and Portier (2018) and Feng et al. (2018). For control variates, however, we are not aware of any uniform error bounds and we believe the next results to be the first of their kind.

5.2 Mathematical background for control variates

Let $\mathscr{F} \subset L^2(P)$ be a collection of square-integrable, real-valued functions f on a probability space $(\mathcal{X}, \mathcal{A}, P)$ of which we would like to calculate the integral $P(f) = \int_{\mathcal{X}} f(x)\,\mathrm{d}P(x)$. Let $X_1, \ldots, X_n$ be independent random variables taking values in $\mathcal{X}$ and with common distribution P. The standard Monte Carlo estimate of $P(f)$ simply takes the form $P_n(f) = \frac{1}{n}\sum_{i=1}^n f(X_i)$. However, this estimator may converge slowly to $P(f)$ due to a high variance. To tackle this issue, it is common practice to use control variates, which are functions in $L^2(P)$ with known integrals. Without loss of generality, we can center the control variates $g_{n,1} \ldots, g_{n,d_n}$ and assume they have zero expectation, that is, $P(g_{n,k}) = 0$ for all $k \in \{1, \ldots, d_n\}$. Let $g_n = (g_{n,1}, \ldots, g_{n,d_n})$ denote the $\mathbb{R}^{d_n}$-valued function with the d_n control variates as elements and put $h_n = (1, g_n^\intercal)^\intercal$. Similarly as before, we assume that the Gram matrix $P(g_n g_n^\intercal)$ is invertible. The control variate Monte Carlo estimate of $P(f)$ is given by $\hat{\alpha}_{n,f}$ defined as (see for instance Portier and Segers, 2019, Section 1),

$$(\hat{\alpha}_{n,f}, \hat{\beta}_{n,f}) \in \operatorname*{arg\,min}_{\alpha \in \mathbb{R},\, \beta \in \mathbb{R}^{d_n}} P_n(f - \alpha - g_n^\intercal \beta)^2. \tag{15}$$

The vector $\hat{\beta}_{n,f}$ contains the regression coefficients for the prediction of f based on the covariates h_n. Remark that the control variate integral estimate $\hat{\alpha}_{n,f}$ coincides with the integral of the least square estimate of f, i.e., $\hat{\alpha}_{n,f} = P(\hat{\alpha}_{n,f} + g_n^\intercal \hat{\beta}_{n,f})$. In addition, since $\hat{\alpha}_{n,f}$ can be expressed as a weighted estimate $\sum_{i=1}^n w_i f(X_i)$ where the weights $(w_i)_{i=1,\ldots,n}$ do not depend on the integrand f, there is a computational benefit to integrating multiple functions (Leluc et al., 2021, Remark 4). It is useful to define

$$(\alpha_{n,f}, \beta_{n,f}) \in \operatorname*{arg\,min}_{\alpha \in \mathbb{R},\, \beta \in \mathbb{R}^{d}} P(f - \alpha - g_n^\intercal \beta)^2,$$

as well as the residual function

$$\varepsilon_{n,f} = f - \alpha_{n,f} - g_n^\intercal \beta_{n,f}.$$

Note that $\alpha_{n,f} = P(f)$. If $\beta_{n,f}$ would be known, the resulting oracle estimator would be

$$\hat{\alpha}_{n,f}^{\mathrm{or}} = P_n[f - g_n^\intercal \beta_{n,f}]. \tag{16}$$

The question raised in the next section is whether the control variate estimate $\hat{\alpha}_{n,f}$ can achieve a similar accuracy uniformly in $f \in \mathscr{F}$ as the oracle estimator $\hat{\alpha}_{n,f}^{\mathrm{or}}$.

5.3 Uniform error bounds

Motivated by the examples above, we provide an error bound for the control variate Monte Carlo estimate $\hat{\alpha}_{n,f}$ in (15) uniformly in $f \in \mathscr{F}$. Before doing so, we give a uniform error bound for the oracle estimate $\hat{\alpha}^{\mathrm{or}}_{n,f}$ in (16). This serves two purposes: first, it will be useful in the analysis of $\hat{\alpha}_{n,f}$ and second, it will provide sufficient conditions for the two estimates to achieve the same level of performance. Recall M_n, γ_n^2 and L_n^2 in (7) and (8) and put

$$\sigma_n^2 = \sup_{f \in \mathscr{F}} P(\varepsilon_{n,f}^2).$$

A new assumption, $M_n^2 = \mathrm{O}\left(\sigma_n^2 n / \log(n)\right)$ as $n \to \infty$, in the same vein but weaker[1] than $L_n^2 = \mathrm{O}\left(\gamma_n^2 \sqrt{n/\log(n)}\right)$, turns out to be useful to obtain the result.

Theorem 2 (Uniform bound on error of oracle estimator)**.** *Assume the framework of Section 5.2 and suppose that Conditions 1, 2 and 3 hold. If* $\liminf_{n\to\infty} n^{\alpha} M_n > 0$ *for some* $\alpha > 0$ *and if* $M_n^2 = \mathrm{O}\left(\sigma_n^2 n / \log(n)\right)$*, then*

$$\sup_{f \in \mathcal{F}} \left|\hat{\alpha}^{\mathrm{or}}_{n,f} - P(f)\right| = \mathrm{O}_{\mathbb{P}}\left(\sigma_n \sqrt{n^{-1}\log(n)}\right), \qquad n \to \infty.$$

The proof of Theorem 2 is provided in Appendix E. The derivation of the stated rate relies on the property that the residual class $\mathscr{E}_n = \{\varepsilon_{n,f} : f \in \mathscr{F}\}$ is a VC-class of functions (as detailed in the proof of Proposition 1). Indeed, noticing that $\hat{\alpha}^{\mathrm{or}}_{n,f} - P(f) = P_n(\varepsilon_{n,f})$ allows to rely on the next proposition, dedicated to suprema of empirical process.

Proposition 2 (Bound of supremum of empirical process)**.** *On the probability space* $(\mathscr{X}, \mathscr{A}, P)$*, let* $\mathscr{S}$ *be a VC-class of parameters* (w, B) *with respect to the constant envelope* $U \geq \sup_{s\in\mathscr{S}} \|s\|_{\infty}$*. Suppose the following two conditions hold:*

(i) $\tau^2 \geq \sup_{s\in\mathscr{S}} \mathrm{var}_P(s)$ *and* $\tau \leq 2U$*;*
(ii) $w \geq 1$ *and* $B \geq 1$*.*

Then, for P_n *the empirical distribution of an independent random sample* $X_1, \ldots, X_n$ *from* P*, we have with probability* $1 - \delta$*:*

$$\sup_{s\in\mathscr{S}} |P_n(s) - P(s)| \leq L\left(\tau\sqrt{wn^{-1}\log(L\theta/\delta)} + Uwn^{-1}\log(L\theta/\delta)\right),$$

with $\theta = BU/\tau$ *and* $L > 0$ *a universal constant.*

The proof of Proposition 2 is given in Appendix D. In the proof, we bound the expectation of the supremum by combining a well-known symmetrization inequality (van der Vaart and Wellner, 1996, Lemma 2.3.1) 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 Theorem 1.4 in Talagrand (1996). Compared to existing results such as Proposition 2.2 stated in Giné and Guillou (2001), our version is more precise due to the explicit role played by the VC constants in the bound.

The next result follows from an application of Corollary 1 and Theorem 2 combined with some other bounds that are standard when analyzing control variates estimates.

Theorem 3 (Uniform error bound on control variate Monte Carlo estimator)**.** *Assume the framework of Section 5.2 and suppose that Conditions 1, 2 and 3 hold. If* $\liminf_{n\to\infty} n^{\alpha} M_n > 0$ *for some* $\alpha > 0$ *and if* $L_n^2 = \mathrm{O}\left(\gamma_n^2 \sqrt{n/\log(n)}\right)$*, then*

$$\sup_{f \in \mathscr{F}} |\hat{\alpha}_f - P(f)| = \mathrm{O}_{\mathbb{P}}\left(\sigma_n \sqrt{n^{-1}\log(n)}\left(1 + \sqrt{d_n n^{-1} \|q_n\|_{\infty}}\right)\right), \qquad n \to \infty.$$

[1] Since $L_n^2 = M_n^2 \|q_n\|_{\infty}$ and $\gamma_n^2 \leq \sigma_n^2 \|q_n\|_{\infty}$.

The proof of Theorem 3 is provided in Appendix F. Compared to the error bound given in Theorem 2 for the oracle estimator, the error bound in Theorem 3 for the control variate estimator has an additional term. This term, which is due to the additional learning step that is needed to estimate the optimal control variate, vanishes as soon as $d_n \|q_n\|_\infty = \mathrm{o}(n)$ as $n \to \infty$. This condition, which was used in Newey (1997) as well as in Portier and Segers (2019), is meaningful as it relates the model complexity to the sample size, i.e., the computing time of the experiment.

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).

First, suppose first condition (a) holds. 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 $|A|_2 \leq |A|_F$ 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

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

from which

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

Second, suppose condition (b) holds. Lemma A.2 in Leluc et al. (2021), 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

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

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

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

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

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

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

It follows that

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

But then

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

Since $P(F^2) \leq \|F\|_\infty^2$, 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, a single 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 $\eta \in (0, \infty)$, the covering number is bounded from above by

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

Fix $\eta > 0$, write $\eta_P = \eta/(4\left\|q_n\right\|_\infty^{1/2})$ and $\eta_Q = \eta/4$, and define the covering numbers

$$N_P = \mathscr{N}\left(\mathscr{F}, L^2(P), \eta_P/2\right), \qquad N_Q = \mathscr{N}\left(\mathscr{F}, L^2(Q), \eta_Q/2\right) \tag{17}$$

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 (17) 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

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

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

$$\mathscr{F} \subset \bigcup_{\substack{1 \leq i \leq N_P \\ 1 \leq j \leq N_Q}} \left(B_P(f_i^{(P)}, \eta_P) \cap B_Q(f_j^{(Q)}, \eta_Q)\right). \tag{18}$$

Define the support of this covering as

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

For every $(i, j) \in \mathsf{S}$, 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 $(i, j) \in \mathsf{S}$ 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 (18), we have

$$\|f - f_{i,j}\|_{L^2(P)} < 2\eta_P, \qquad \|f - f_{i,j}\|_{L^2(Q)} < 2\eta_Q. \tag{19}$$

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

$$\|\varepsilon_{n,f} - \varepsilon_{n,f_{i,j}}\|_{L^2(Q)} \leq \|f - f_{i,j}\|_{L^2(Q)} + \|\hbar_n^\intercal P[\hbar_n(f - f_{i,j})]\|_{L^2(Q)}. \tag{20}$$

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 $q_n = |\hbar_n|_2^2$. The Cauchy–Schwarz inequality and the orthonormality of $(\hbar_{n,1},\dots,\hbar_{n,d_n})$ give

$$\begin{aligned}\|\hbar_n^\intercal P[\hbar_n(f-f_{i,j})]\|_{L^2(Q)}^2 &= \int_{y\in\mathscr{X}} \left\{\hbar_n(y)^\intercal \int_{x\in\mathscr{X}} \hbar_n(x)(f-f_{i,j})(x)\,\mathrm{d}P(x)\right\}^2 \mathrm{d}Q(y)\\ &\leq \int_{y\in\mathscr{X}} |\hbar_n(y)|_2^2 \left|\int_{x\in\mathscr{X}} \hbar_n(x)(f-f_{i,j})(x)\,\mathrm{d}P(x)\right|_2^2 \mathrm{d}Q(y)\\ &\leq \|q_n\|_\infty \|f-f_{i,j}\|_{L^2(P)}^2. \end{aligned} \tag{21}$$

The combination of (19), (20) and (21) yields

$$\|\varepsilon_{n,f}-\varepsilon_{n,f_{i,j}}\|_{L^2(Q)} < 2\eta_Q + 2\|q_n\|_\infty^{1/2}\eta_P = \eta/2+\eta/2 = \eta.$$

We have thus constructed the covering

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

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

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

Using the definition of a VC-class and since $\|q_n\|_\infty \geq P(q_n) = P(|\hbar_n|_2^2) = d_n \geq 1$, a case-by-case analysis reveals that

$$\begin{aligned}\mathscr{N}\left(\mathscr{E}_n, L^2(Q),\eta\right) &\leq \left(\frac{2A\|F\|_{L^2(P)}}{\eta_P}\vee 1\right)^v \cdot \left(\frac{2A\|F\|_{L^2(Q)}}{\eta_Q}\vee 1\right)^v\\ &\leq \left(\frac{64A^2\|F\|_\infty^2\|q_n\|_\infty^{1/2}}{\eta^2}\right)^v \vee \left(\frac{8A\|F\|_\infty\|q_n\|_\infty^{1/2}}{\eta}\right)^v \vee 1\\ &\leq \left(\frac{8A\|F\|_\infty\|q_n\|_\infty^{1/2}}{\eta}\right)^{2v} \vee 1. \end{aligned} \tag{22}$$

Lemma 2 implies that the sequence A_n defined in (14) satisfies $A_n \geq 1$. From (22) we deduce

$$\forall \eta\in(0,1], \qquad \mathscr{N}\left(\mathscr{E}_n, L^2(Q), \eta M_n\right) \leq (A_n/\eta)^{2v}.$$

Therefore, 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

$$\forall (x,y)\in\mathscr{X}^2, \qquad |\hbar_n(x)^\intercal\hbar_n(y)|^2 \leq |\hbar_n(x)|_2^2\,|\hbar_n(y)|_2^2 = q_n(x)q_n(y) \leq \|q_n\|_\infty^2. \tag{23}$$

For $f,\tilde f\in\mathscr{F}$ and $(x,y)\in\mathscr{X}^2$, define

$$g_{n,f,\tilde f}(x,y) = \varepsilon_{n,f}(x)\hbar_n(x)^\intercal\hbar_n(y)\varepsilon_{n,\tilde f}(y).$$

Note that $g_{n,f} = g_{n,f,\tilde f}$. By the Minkowski inequality,

$$\|g_{n,f}-g_{n,\tilde f}\|_{L^2(Q)} \leq \|g_{n,f}-g_{n,f,\tilde f}\|_{L^2(Q)} + \|g_{n,f,\tilde f}-g_{n,\tilde f}\|_{L^2(Q)}.$$

Let us look at square of the first term on the right-hand side: by (23),

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

The term $\|g_{n,f,\tilde{f}} - g_{n,\tilde{f}}\|_{L^2(Q)}$ can be treated similarly, yielding

$$\begin{aligned}&\|g_{n,f} - g_{n,\tilde{f}}\|_{L^2(Q)} \\ &\leq \left\|q_n \varepsilon^2_{n,f}\right\|^{1/2}_\infty \|q_n\|^{1/2}_\infty \|\varepsilon_{n,f} - \varepsilon_{n,\tilde{f}}\|_{L^2(Q_2)} + \left\|q_n \varepsilon^2_{n,\tilde{f}}\right\|^{1/2}_\infty \|q_n\|^{1/2}_\infty \|\varepsilon_{n,f} - \varepsilon_{n,\tilde{f}}\|_{L^2(Q_1)} \\ &\leq M_n \|q_n\|_\infty \left(\|\varepsilon_{n,f} - \varepsilon_{n,\tilde{f}}\|_{L^2(Q_1)} + \|\varepsilon_{n,f} - \varepsilon_{n,\tilde{f}}\|_{L^2(Q_2)}\right). \qquad (24)\end{aligned}$$

Fix $\eta > 0$. Following the approach in Step 1, we can for $\ell \in \{1,2\}$ 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

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

where $f^{(\ell)}_k$ 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 (14) Consider the intersections

$$\forall i = 1,\ldots,m_1,\ \forall j = 1,\ldots,m_2, \qquad B(i,j) = B_{Q_1}\left(\varepsilon_{n,f^{(1)}_i}, \eta\right) \cap B_{Q_2}\left(\varepsilon_{n,f^{(2)}_j}, \eta\right).$$

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

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

By (24), it follows that

$$\|g_{n,f} - g_{n,f_{i,j}}\|_{L^2(Q)} < M_n \|q_n\|_\infty [(2\eta) + (2\eta)] = 4M_n \|q_n\|_\infty \eta.$$

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

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

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

$$\mathscr{N}\left(\mathscr{G}_n, L^2(Q), M_n^2 \|q_n\|_\infty \eta'\right) \leq \left(\frac{4A_n}{\eta'}\right)^{4v} \vee 1. \qquad (25)$$

In view of (23), the functions in $\mathscr{G}_n$ are uniformly bounded by $M_n^2 \|q_n\|_\infty$. 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 $M_n^2 \|q_n\|_\infty$.

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 $(A_n M_n/\eta)^{2v} \vee 1$ of such functions. For $f \in \mathscr{F}$ we can thus find (i,j) such that

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

and thus

$$\begin{aligned}\|\sqrt{q_n}\varepsilon_{n,f} - \sqrt{q_n}\varepsilon_{n,f_{i,j}}\|_{L^2(Q)}^2 &\leq \|q_n\|_\infty \left\|\varepsilon_{n,f} - \varepsilon_{n,f_{i,j}}\right\|_{L^2(Q)}^2 \\ &< \|q_n\|_\infty \eta^2.\end{aligned}$$

It follows that the number of $L^2(Q)$ balls of radius $\|q_n\|_\infty^{1/2}\eta$ 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

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

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

$$\mathscr{N}\left(\mathscr{G}_n^{(d)}, L^2(Q), M_n\|q_n\|_\infty^{1/2}\eta'\right) \leq \left(\frac{A_n}{\eta'}\right)^{2v} \vee 1. \tag{26}$$

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

C Proof of Theorem 1 and Corollary 1

We recall the following quantities:

$$M_n = \sup_{f\in\mathscr{F}} \|\varepsilon_{n,f}\|_\infty, \qquad \gamma_n^2 = \sup_{f\in\mathscr{F}} P(q_n\varepsilon_{n,f}^2), \qquad L_n^2 = M_n^2\|q_n\|_\infty.$$

For all $f \in \mathscr{F}$, we have $P(\varepsilon_{n,f}^2) \leq P(f^2) \leq P(F^2)$ and $\gamma_n \leq L_n$.

Step 1: Overview. We follow the plan laid out in Section 4 in the paper. In view of (11) and Lemma 1, we have

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

The inequality (12) provides a bound on $n^2|P_n(\hbar_n\varepsilon_{n,f})|_2^2$ consisting of a sum of two terms which require a separate analysis (Steps 2 and 3). Finally, we collect the bounds to arrive at the stated rate (Step 4).

Step 2: First term in (12)**.** Recall the definition $\mathscr{G}_n^{(d)} = \{q_n^{1/2}\varepsilon_{n,f} : f \in \mathscr{F}\}$ introduced in (13). We first apply Corollary 3.4 given in Talagrand (1994) to the class $\mathscr{G}_n^{(d)}$ normalized by its envelope L_n so that the resulting class is valued in $[-1,1]$. We obtain that

$$\mathbb{E}\left[\sup_{g\in\mathscr{G}_n^{(d)}}\left|\sum_{i=1}^n g^2(X_i)\right|\right] \leq n\gamma_n^2 + 8L_n\,\mathbb{E}\left[\sup_{g\in\mathscr{G}_n^{(d)}}\left|\sum_{i=1}^n \eta_i g(X_i)\right|\right]. \tag{27}$$

Next, we apply Proposition 2.1 stated in Giné and Guillou (2001) to the class $\mathscr{G}_n^{(d)}$ to get

$$\mathbb{E}\left[\sup_{g\in\mathscr{G}_n^{(d)}}\left|\sum_{i=1}^n \eta_i g(X_i)\right|\right] \leq C\left(\tau_n\sqrt{w_n n\log(\theta_n)} + U_n w_n\log(\theta_n)\right), \tag{28}$$

where $C > 0$ is a universal constant and $\theta_n = B_n U_n/\tau_n$ and where the positive quantities τ_n, U_n, w_n, B_n need to be chosen to satisfy the following two conditions:

(i) $\tau_n^2 \geq \sup_{f\in\mathscr{F}} P(q_n\varepsilon_{n,f}^2)$, $U_n \geq \sup_{f\in\mathscr{F}} \|\sqrt{q_n}\varepsilon_{n,f}\|_\infty$ and $U_n \geq \tau_n$;

(ii) (w_n, B_n) are VC parameters of $\mathscr{G}_n^{(d)}$ with respect to the envelope U_n, and $w_n \geq 1$ and $B_n \geq 3\sqrt{\mathrm{e}}$.

We set

$$\tau_n = \|q_n\|_\infty^{1/2} (a_n \vee M_n) \qquad \text{and} \qquad U_n = \tau_n,$$

where a_n is a sequence tending to zero that is introduced for some technical reason in Step 3, Eq. (38). Because $\tau_n \geq \|q_n\|_\infty^{1/2} M_n$, Condition (i) is satisfied. To meet (ii), we set $w_n = (2v) \vee 1$, independently of n, and

$$B_n = \frac{10A \|F\|_\infty \|q_n\|_\infty^{1/2}}{M_n \vee a_n}. \tag{29}$$

Since $A \geq 1$, since $M_n \leq 2\|F\|_\infty \|q_n\|_\infty^{1/2}$ (Lemma 2), since $\|q_n\|_\infty = d_n \geq 1$ and since $a_n = \mathrm{o}(1)$, it follows that $B_n \geq 5 \geq 3\sqrt{\mathrm{e}}$ for all sufficiently large n. We get, using the bound on the covering numbers of $\mathscr{G}_n^{(d)}$ in (26) and the definition of A_n in (14), after some calculations,

$$\mathscr{N}\left(\mathscr{G}_n^{(d)}, L^2(Q), U_n\eta\right) \leq \left(\frac{A_n}{(1 \vee (a_n/M_n))\,\eta}\right)^{2v} \vee 1 \leq (B_n/\eta)^{2v} \vee 1, \qquad 0 < \eta \leq 1,$$

from which Condition (ii) follows. As $\theta_n = B_n U_n/\tau_n = B_n$, (28) implies

$$\begin{aligned}\mathbb{E}\left[\sup_{g\in\mathscr{G}_n^{(d)}} \left|\sum_{i=1}^n \eta_i g(X_i)\right|\right] &\leq C\tau_n w_n \left(\sqrt{n\log(B_n)} + \log(B_n)\right), \\ &= \mathrm{O}\left(\tau_n\sqrt{n}\log(B_n)\right), \qquad n\to\infty. \end{aligned} \tag{30}$$

The bounds (27) and (30) combined with the Markov inequality give

$$\sup_{f\in\mathscr{F}} nP_n(q_n\varepsilon_{n,f}^2) = \sup_{g\in\mathscr{G}_n^{(d)}} \left|\sum_{i=1}^n g^2(X_i)\right| = \mathrm{O}_{\mathbb{P}}\left(n\gamma_n^2 + L_n\tau_n\sqrt{n}\log(B_n)\right), \qquad n\to\infty.$$

As we will see in Step 4, the latter rate is of smaller order than the one for the third term in (12), which is derived in Step 3.

Step 3: Third term in (12)**.** 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: for any $x \in \mathscr{X}$, we have we have $\mathbb{E}\left[g_{n,f}(X, x)\right] = \mathbb{E}\left[g_{n,f}(x, X)\right] = 0$, where the random variable X has distribution P. We apply a special case of Theorem 2 in Major (2006), cited for convenience as Theorem 4 below. The functions $g_{n,f}$ are uniformly bounded by

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

Let

$$\tilde{\tau}_n = L_n\tau_n = M_n\|q_n\|_\infty (M_n \vee a_n) = L_n^2 (1 \vee (a_n/M_n)).$$

Scale the functions $g_{n,f}$ by $\tilde{\tau}_n$, yielding the class

$$\widetilde{\mathscr{G}}_n = \{g_{n,f}/\tilde{\tau}_n : f \in \mathscr{F}\}$$

of functions on $\mathscr{X}^2$ taking values in $[-1, 1]$. For any $\eta \in (0, 1]$, by applying (25) with $\eta' = \tilde{\tau}_n\eta/\left(M_n^2\|q_n\|_\infty\right)$, we get, recalling the definition of A_n in (14),

$$\begin{aligned}\mathscr{N}\left(\widetilde{\mathscr{G}}_n, L^2(Q), \eta\right) = \mathscr{N}\left(\mathscr{G}_n, L^2(Q), \tilde{\tau}_n\eta\right) &\leq \left(\frac{4A_nM_n^2\|q_n\|_\infty}{\tilde{\tau}_n\eta}\right)^{4v} \vee 1 \\ &\leq \left(\frac{32A\|F\|_\infty\|q_n\|_\infty^{1/2}}{(M_n \vee a_n)\,\eta}\right)^{4v} \vee 1. \end{aligned} \tag{31}$$

Using Lemma 2, we get $M_n \le 2\|F\|_\infty \|q_n\|_\infty^{1/2}$ and since $A \ge 1$ we obtain $32A\|F\|_\infty \|q_n\|_\infty^{1/2} \ge M_n$. In addition, the sequence $(a_n)_{n\in\mathbb{N}}$ defined in (38) is taken such that, for n sufficiently large, $32A\|F\|_\infty \|q_n\|_\infty^{1/2} \ge a_n$. Then, for n sufficiently large (31) we have, for B_n defined in (29),

$$\mathscr{N}\left(\widetilde{\mathscr{G}_n}, L^2(Q), \eta\right) \le \left(\tfrac{32}{10} B_n/\eta\right)^{4v}, \qquad \eta \in (0,1].$$

In the terminology of Major (2006, p. 490), $\widetilde{\mathscr{G}_n}$ is an L^2-dense class of functions with parameter D_n and exponent w defined by

$$D_n = \left(\tfrac{32}{10} B_n\right)^{4v} \qquad \text{and} \qquad w = (4v) \vee 1. \tag{32}$$

For w, we take the maximum with 1 in order to apply Theorem 4 later on.

By the Cauchy–Schwarz inequality, we have, for any $(x,y) \in \mathscr{X}^2$ and any $f \in \mathscr{F}$,

$$\begin{aligned} g_{n,f}(x,y)^2 &= \varepsilon_{n,f}(x)^2 \varepsilon_{n,f}(y)^2 \left|\hbar_n(x)^\intercal \hbar_n(y)\right|^2 \\ &\le \varepsilon_{n,f}(x)^2 \varepsilon_{n,f}(y)^2 \, q_n(x)\, q_n(y). \end{aligned}$$

It follows that, for independent random variables X_1 and X_2 with common distribution P, we have

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

Upon rescaling, we get

$$\sup_{g \in \widetilde{\mathscr{G}_n}} \left(\mathbb{E}[g(X_1,X_2)^2]\right)^{1/2} \le \gamma_n^2/\tilde{\tau}_n.$$

For a sequence $b_n \in (0,1]$ to be determined shortly, put

$$\nu_n = (\gamma_n^2/\tilde{\tau}_n) \vee b_n. \tag{33}$$

We have $\nu_n \le 1$; recall that $\tilde{\tau}_n \ge L_n^2 \ge \gamma_n^2$. Moreover, ν_n^2 is an upper bound of the second moments of the functions in $\widetilde{\mathscr{G}_n}$. Theorem 4 yields

$$\mathbb{P}\left(\sup_{f\in\mathscr{F}} \left|\sum_{1\le i\ne j\le n} g_{n,f}(X_i,X_j)\right| \ge 2n\nu_n\tilde{\tau}_n y\right) \le C D_n \, \mathrm{e}^{-\alpha y} \tag{34}$$

for all $y \in [y_{n,-}, y_{n,+}]$, where, for some universal constants α, C, and K, the endpoints of the interval are (note from (32) that $D_n \ge B_n^{4v} \ge 1$ so $\log D_n \ge 0$)

$$y_{n,-} = K\left(w + \frac{\log D_n}{\log n}\right)^{3/2} \log(2/\nu_n) \qquad \text{and} \qquad y_{n,+} = n\nu_n^2.$$

We still need to determine a_n and b_n. We would like to choose y in (34) in such a way that the right-hand side is bounded by a pre-specified $\delta \in (0,1]$. Therefore, we need to ensure two things:

$$y_{n,-} \le y_{n,+}, \qquad \text{and} \tag{35}$$

$$C D_n \mathrm{e}^{-\alpha y_{n,+}} \to 0, \qquad n \to \infty. \tag{36}$$

We need (35) in order to ensure that the interval $[y_{n,-}, y_{n,+}]$ on which (34) holds is non-empty; we need (36) to ensure that for any $\delta \in (0,1]$, we can find $y \in [y_{n,-}, y_{n,+}]$ such that the right-hand side of (34) is bounded by δ. To this end, define

$$D_n^* = \exp\left[\left\{\left(\frac{2n}{K\log n}\right)^{2/3} - w\right\}\log n\right], \tag{37}$$

$$a_n = \frac{32A\|F\|_\infty \|q_n\|_\infty^{1/2}}{(D_n^*)^{1/(4v)}}, \tag{38}$$

$$b_n = \left(\frac{K\log n}{2n}\right)^{1/2}\left(w + \frac{\log D_n}{\log n}\right)^{3/4}. \tag{39}$$

To explain these definitions, note that D_n^* is the solution to

$$\left(w + \frac{\log D_n^*}{\log n}\right)^{3/2} = \frac{2n}{K \log n},$$

whereas a_n is chosen in such a way that

$$D_n = D_n^* \wedge \left(\frac{32A \left\|F\right\|_\infty \left\|q_n\right\|_\infty^{1/2}}{M_n}\right)^{4v}. \tag{40}$$

Furthermore, since $D_n \leq D_n^*$, we have, as required earlier,

$$b_n \leq \left(\frac{K \log n}{2n}\right)^{1/2} \left(w + \frac{\log D_n^*}{\log n}\right)^{3/4} = 1.$$

We can now verify that both (35) and (36) hold:

- The definition of ν_n in (33) implies $\nu_n \geq b_n$, from which we get the chain of inequalities

$$y_{n,+} = n\nu_n^2 \geq nb_n^2 = \frac{K \log n}{2} \left(w + \frac{\log D_n}{\log n}\right)^{3/2} = \frac{\log n}{2 \log(2/\nu_n)} y_{n,-} \geq y_{n,-}, \tag{41}$$

  which is (35). To see the last inequality in (41), note that $1/\nu_n \leq 1/b_n = \mathrm{o}(n^{1/2})$, from which $(2/\nu_n)^2 = \mathrm{o}(n)$ as $n \to \infty$ and thus $2\log(2/\nu_n) \leq \log n$ for sufficiently large n.
- Enlarging K if necessary to ensure that $\alpha K \geq 2$, we have, in view of $y_n \geq nb_n^2$ and (39),

$$\begin{aligned}
\log D_n - \alpha y_{n,+} &\leq \log D_n - \alpha n b_n^2 \\
&= \log D_n - \frac{\alpha K}{2} (\log n) \left(w + \frac{\log D_n}{\log n}\right)^{3/2} \\
&\leq \log D_n - (\log n) \left(1 + \frac{\log D_n}{\log n}\right)^{3/2} \\
&= (-\log n) \left[\left(1 + \frac{\log D_n}{\log n}\right)^{3/2} - \frac{\log D_n}{\log n}\right] \\
&\leq -\log n \to -\infty,
\end{aligned}$$

  from which (36) follows.

Let $\delta \in (0, 1]$ and define

$$y_n(\delta) = y_{n,-} \vee \left(\alpha^{-1} \log(CD_n/\delta)\right).$$

We already know from (41) that $y_{n,-} \leq y_{n,+}$ for large n. Moreover, since $y = \alpha^{-1} \log(CD_n/\delta)$ is the solution to $CD_n \mathrm{e}^{-\alpha y} = \delta$, the asymptotic relation in (36) implies $\alpha^{-1} \log(CD_n/\delta) \leq y_{n,+}$ for large n. It follows that $y_n(\delta) \in [y_{n,-}, y_{n,+}]$ and $CD_n \mathrm{e}^{-\alpha y_n(\delta)} \leq \delta$ for all (sufficiently large) n. Defining

$$u_n = 1 \vee \frac{\log D_n}{\log n}$$

we have, as $n \to \infty$, the asymptotic relations [recall from the lines following (41) that $2\log(2/\nu_n) \leq \log n$ for large n]

$$y_{n,-} = \mathrm{O}(u_n^{3/2} \log n) \qquad \text{and} \qquad \log D_n = \mathrm{O}(u_n \log n).$$

Since $\nu_n \tilde{\tau}_n = \gamma_n^2 \vee (\tilde{\tau}_n b_n)$, we find by (34) that

$$\sup_{f \in \mathscr{F}} \left| \sum_{1 \leq i \neq j \leq n} g_{n,f}(X_i, X_j) \right| = \mathrm{O}_{\mathbb{P}} \left(n\nu_n \tilde{\tau}_n \left(y_{n,-} \vee \log D_n\right)\right) \tag{42}$$

$$= \mathrm{O}_{\mathbb{P}} \left(n \left(\gamma_n^2 \vee (\tilde{\tau}_n b_n)\right) u_n^{3/2} \log n\right), \qquad n \to \infty. \tag{43}$$

Write

$$\rho_n = u_n^{3/2} \frac{\log n}{n} \tag{44}$$

and note that $b_n = \mathrm{O}(\rho_n^{1/2})$ as $n \to \infty$. It follows that

$$\sup_{f \in \mathscr{F}} \left| \frac{1}{n^2} \sum_{1 \le i \ne j \le n} g_{n,f}(X_i, X_j) \right| = \mathrm{O}_{\mathbb{P}} \left(\left\{ \gamma_n^2 \vee \left(\tilde{\tau}_n \rho_n^{1/2} \right) \right\} \rho_n \right), \qquad n \to \infty. \tag{45}$$

Step 4: Comparison and simplification of rates. The first term in the expansion (12) was shown in Step 2 to have rate $n\gamma_n^2 + L_n \tau_n \sqrt{n} \log(B_n)$. By definition, $n\gamma_n^2$ is of smaller order than the rate in (43). The other term, $L_n \tau_n \sqrt{n} \log(B_n)$, is of smaller order too: since $\nu_n \ge b_n$ and since b_n in (39) is of larger order than $1/\sqrt{n}$, the rate $L_n \tau_n \sqrt{n} \log B_n$ is of smaller order than $\mathrm{O}\left(n \nu_n L_n \tau_n \log B_n\right)$, which is bounded by the rate in (42) in view of $L_n \tau_n = \tilde{\tau}_n$ and the connection between B_n and D_n in (32).

We can therefore conclude that the rate in (45) for the third term is actually the dominating one. It remains to work out the sequence $(\rho_n)_{n \in \mathbb{N}}$ in (44), that is, to analyse $(u_n^{3/2})_{n \in \mathbb{N}}$ further. By (37),

$$\frac{\log D_n^*}{\log n} = \mathrm{O}\left(\left(\frac{n}{\log n} \right)^{2/3} \right), \qquad n \to \infty.$$

Since M_n is bounded by a constant multiple of $\|q_n\|_\infty^{1/2}$ (Lemma 2), which grows at most at a polynomial rate in n by Condition 1, we have

$$\frac{\log \left(\|q_n\|_\infty^{1/2} / M_n \right)}{\log n} = \mathrm{O}\left(1 + \frac{(\log M_n^{-1})_+}{\log n} \right), \qquad n \to \infty.$$

In view of the connection between D_n and D_n^* in (40), it follows from the combination of the two estimates above that

$$\begin{aligned} \left(\frac{\log D_n}{\log n} \right)^{3/2} &= \mathrm{O}\left(\frac{n}{\log n} \wedge \left\{ 1 + \frac{(\log M_n^{-1})_+}{\log n} \right\}^{3/2} \right) \\ &= \mathrm{O}\left(\frac{n}{\log n} \wedge \left\{ 1 + \left(\frac{(\log M_n^{-1})_+}{\log n} \right)^{3/2} \right\} \right), \qquad n \to \infty. \end{aligned}$$

Since both members of the minimum on the right-hand side are larger than one, it follows that

$$\rho_n = u_n^{3/2} \frac{\log n}{n} = \mathrm{O}\left(1 \wedge \left[\frac{\log n}{n} \left\{ 1 + \left(\frac{(\log M_n^{-1})_+}{\log n} \right)^{3/2} \right\} \right] \right), \qquad n \to \infty.$$

It is the latter form that is stated in the theorem. □

Proof of Corollary 1. If $(\log M_n^{-1})_+ = \mathrm{O}(\log n)$ as $n \to \infty$, then $M_n > a_n$ and thus $\tau_n = L_n^2$ for sufficiently large n. Moreover, r_n can then be replaced by $(\log n)/n$. The simpler rate (9) follows. Under the additional condition on L_n^2, the latter rate implies the one in (10). □

D Proof of Proposition 2

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

$$Z_n \le \mathbb{E}(Z_n) + |Z_n - \mathbb{E}(Z_n)|. \tag{46}$$

We treat the two terms on the right-hand side of (46) in Steps 1 and 2, respectively. The bound for Z_n then follows by adding both bounds in Step 3.

Step 1: Expectation of the supremum. 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 detailed in van der Vaart and Wellner (1996, Lemma 2.3.1) gives

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

Further, by applying Proposition 2.1 in Giné and Guillou (2001) to the class $\{s - P(s) \,:\, s \in \mathscr{S}\}$, (with VC parameters $(w, 3\sqrt{\mathrm{e}}B)$, envelope $2U$ and variance bound τ^2), we obtain the existence of a universal constant $C_1 > 0$ such that

$$\mathbb{E}\left[\sup_{s\in\mathscr{S}}\left|\sum_{i=1}^{n}\eta_i\{s(X_i)-P(s)\}\right|\right] \leq C_1\left(2wU\log\left(C_2\theta\right) + \tau\sqrt{wn\log\left(C_2\theta\right)}\right) =: \xi_n \tag{47}$$

where $C_2 = 6\sqrt{\mathrm{e}}$ and where we wrote $\theta = BU/\tau$ as in the statement of the proposition.

Step 2: Concentration of the supremum around its expectation. Since $\sup_{s\in\mathscr{S}}\|s - P(s)\|_\infty \leq 2U$, Theorem 1.4 in Talagrand (1996) states the existence of a universal constant $K > 0$ such that

$$\forall t > 0, \qquad \mathbb{P}\left(n|Z_n - \mathbb{E}(Z_n)| \geq t\right) \leq K\exp\left\{-\frac{t}{2KU}\log\left(1+\frac{2tU}{V_n}\right)\right\} \tag{48}$$

where

$$V_n = \mathbb{E}\left[\sup_{s\in\mathscr{S}}\sum_{i=1}^{n}\{s(X_i)-P(s)\}^2\right].$$

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

$$\frac{t}{2KU}\log\left(1+\frac{2tU}{V_n}\right) \geq \frac{t}{2KU}\frac{\frac{2tU}{V_n}}{1+\frac{1}{2}\frac{2tU}{V_n}} = \frac{t^2}{K\left(V_n+tU\right)}$$

and thus,

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

Assuming without loss of generality that $K \geq 1$ and inverting the previous bound (see for instance Peel et al., 2010, Lemma 1) gives that with probability at least $1-\delta$,

$$n|Z_n - \mathbb{E}(Z_n)| \leq \sqrt{V_nK\log(K/\delta)} + UK\log(K/\delta).$$

Corollary 3.4 in Talagrand (1994) applied to the family $\{s - P(s) : s \in \mathscr{S}\}$ yields

$$\begin{aligned} V_n &\leq n\tau^2 + 16U\,\mathbb{E}\left[\sup_{s\in\mathscr{S}}\left|\sum_{i=1}^{n}\eta_i\{s(X_i)-P(s)\}\right|\right] \\ &\leq n\tau^2 + 16U\xi_n \end{aligned}$$

in view of (47). Note that in the cited corollary, the functions are bounded in absolute value by 1, whereas here, the functions $s - P(s)$ are bounded uniformly by $2U$, and this is reflected in the bound above. Since $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for nonnegative a and b,

$$\sqrt{V_n} \leq \tau\sqrt{n} + 4\sqrt{U\xi_n}.$$

As a consequence, with probability at least $1-\delta$, it holds that

$$n|Z_n - \mathbb{E}(Z_n)| \leq \left(\tau\sqrt{n} + 4\sqrt{U\xi_n}\right)\sqrt{K\log(K/\delta)} + UK\log(K/\delta).$$

Step 3: Bound on the supremum. Combining the bound (46) with the inequalities obtained in Steps 1 and 2, we obtain that with probability at least $1-\delta$,

$$\begin{aligned} nZ_n &\leq 2\xi_n + 2\times 2\sqrt{\xi_n}\sqrt{UK\log(K/\delta)} + UK\log(K/\delta) + \tau\sqrt{nK\log(K/\delta)} \\ &\leq 4\xi_n + 3UK\log(K/\delta) + \tau\sqrt{nK\log(K/\delta)} \end{aligned}$$

where we have just used that $2ab \leq a^2+b^2$ with $a=\sqrt{\xi_n}$ et $b=\sqrt{UK\log(K/\delta)}$. Injecting the value of ξ_n from (47) and factorizing, we get

$$\begin{aligned} nZ_n &\leq 4C_1\left(2wU\log(C_2\theta) + \tau\sqrt{wn\log(C_2\theta)}\right) + 3UK\log(K/\delta) + \tau\sqrt{nK\log(K/\delta)} \\ &\leq U\left((8C_1w)\vee(3K)\right)\log(C_2K\theta/\delta) + \tau\sqrt{n}\left(\left(4C_1\sqrt{w}\right)\vee\sqrt{K}\right)\left(\sqrt{\log(C_2\theta)}+\sqrt{\log(K/\delta)}\right) \\ &\leq U\left((8C_1w)\vee(3K)\right)\log(C_2K\theta/\delta) + \tau\sqrt{n}\left(\left(4C_1\sqrt{w}\right)\vee\sqrt{K}\right)\sqrt{2\log(C_2K\theta/\delta)}, \end{aligned}$$

where, in the last step, we have used that $\sqrt{a}+\sqrt{b}\leq\sqrt{2(a+b)}$ with $a=\log(2\theta)$ and $b=\log(K/\delta)$. Conclude the proof by applying the inequality $(aw)\vee b \leq (a\vee b)w$ for $w\geq 1$ and $a,b\geq 0$; similarly for w replaced by $\sqrt{w}$.

E Proof of Theorem 2

Start by noting that

$$\hat{\alpha}^{\mathrm{or}}_{n,f} - P(f) = P_n(\varepsilon_{n,f}).$$

As shown in Step 1 in Section B, the residual class $\mathscr{E}_n$ is VC of parameters $(2v, A_n)$ with respect to the envelope M_n with $A_n = 8A\left\|F\right\|_\infty\left\|q_n\right\|_\infty^{1/2}/M_n$. Hence we can apply Proposition 2 with $w_n=(2v)\vee 1$, $B_n = 8\left(A\vee(3/4)\sqrt{\mathrm{e}}\right)\left\|F\right\|_\infty\left\|q_n\right\|_\infty^{1/2}/M_n \geq A_n$, $\tau_n=\sigma_n$ and $U_n = M_n\vee(2\sigma_n)$. Condition (i) of Proposition 2 is easily met. Because $M_n \leq 2\left\|F\right\|_\infty\left\|q_n\right\|_\infty^{1/2}$ we have

$$B_n \geq 4\left(A\vee\left((3/4)\sqrt{\mathrm{e}}\right)\right) \geq 3\sqrt{\mathrm{e}}.$$

Therefore Condition (ii) of Proposition 2 is also met. Since w_n is constant, we obtain

$$\sup_{f\in\mathscr{F}}|P_n(\varepsilon_{n,f})| = \mathrm{O}_{\mathbb{P}}\left(\sigma_n\sqrt{n^{-1}\log(\theta_n)} + U_n n^{-1}\log(\theta_n)\right), \qquad n\to\infty,$$

with $\theta_n = B_n\left(M_n\vee(2\sigma_n)\right)/\sigma_n$. Note that $\liminf_{n\to\infty}\theta_n>0$ and, since $M_n\vee(2\sigma_n)\leq 2M_n$,

$$\theta_n = \left(8\left(A\vee\left((3/4)\sqrt{\mathrm{e}}\right)\right)\left\|F\right\|_\infty\left\|q_n\right\|_\infty^{1/2}\right)\left(M_n\vee(2\sigma_n)\right)/\left(\sigma_n M_n\right) \leq A'\left\|q_n\right\|_\infty^{1/2}/\sigma_n,$$

with $A' = 16\left(A\vee((3/4)\sqrt{\mathrm{e}})\right)\left\|F\right\|_\infty$. From Condition 1, it holds that $\log(\left\|q_n\right\|_\infty) = \mathrm{O}(\log(n))$. Using the fact that $M_n^2 = \mathrm{O}(\sigma_n^2 n/\log(n))$, which is also $\mathrm{O}(\sigma_n^2 n)$, and since by assumption $M_n^{-1} = \mathrm{O}(n^\alpha)$, we get $\sigma_n^{-2} = \mathrm{O}(n^{1+2\alpha})$ as $n\to\infty$. Consequently, $\log(\theta_n)=\mathrm{O}(\log(n))$ and we find that

$$\sup_{f\in\mathscr{F}}|P_n(\varepsilon_{n,f})| = \mathrm{O}_{\mathbb{P}}\left(\sigma_n\sqrt{n^{-1}\log(n)} + M_n n^{-1}\log(n)\right), \qquad n\to\infty.$$

Moreover, using again that $M_n^2 = \mathrm{O}(\sigma_n^2 n/\log(n))$ as $n\to\infty$, we find the stated rate. □

F Proof of Theorem 3

Let $H_n = P(h_n h_n^\intercal) \in \mathbb{R}^{(d_n+1)\times(d_n+1)}$ and $G_n = P(g_n g_n^\intercal) \in \mathbb{R}^{d_n\times d_n}$. By assumption, both matrices are invertible. Using Equation (22) in Leluc et al. (2021), we obtain

$$|\hat{\alpha}_{n,f} - P(f)| \leq |P_n(\varepsilon_{n,f})| + \left|G_n^{1/2}\left(\hat{\beta}_{n,f}-\beta_{n,f}\right)\right|_2\left|G_n^{-1/2}P_n(g_n)\right|_2.$$

Since $h_n = (1, g_n^\intercal)^\intercal$, we have

$$\left| H_n^{1/2} \begin{pmatrix} \hat{\alpha}_{n,f} - \alpha_{n,f} \\ \hat{\beta}_{n,f} - \beta_{n,f} \end{pmatrix} \right|_2^2 = (\hat{\alpha}_{n,f} - \alpha_{n,f})^2 + \left| G_n^{1/2} \left(\hat{\beta}_{n,f} - \beta_{n,f} \right) \right|_2^2,$$

which, combined with the identity (2), gives

$$\begin{aligned} \left| G_n^{1/2} \left(\hat{\beta}_{n,f} - \beta_{n,f} \right) \right|_2^2 &\leq \left| H_n^{1/2} \begin{pmatrix} \hat{\alpha}_{n,f} - \alpha_{n,f} \\ \hat{\beta}_{n,f} - \beta_{n,f} \end{pmatrix} \right|_2^2 \\ &= L_f(\hat{\alpha}_{n,f}, \hat{\beta}_{n,f}) - L_f(\alpha_{n,f}, \beta_{n,f}) \end{aligned}$$

with $L_f(\alpha, \beta) = P[(f - \alpha - \beta^\intercal h_n)^2]$; note the slight change in notation of the excess risk due to the presence of an intercept. Consequently, we have shown that

$$|\hat{\alpha}_{n,f} - P(f)| \leq |P_n(\varepsilon_{n,f})| + \sqrt{L_f(\hat{\alpha}_{n,f}, \hat{\beta}_{n,f}) - L_f(\alpha_{n,f}, \beta_{n,f})} \left| G_n^{-1/2} P_n(g_n) \right|_2,$$

and the rest of the proof consists in bounding the three terms on the right-hand side uniformly in $f \in \mathscr{F}$. First, using Corollary 1 and the fact that $\gamma_n^2 \leq \sigma_n^2 \|q_n\|_\infty$, we get the uniform bound

$$\sup_{f \in \mathscr{F}} \{L_f(\hat{\alpha}_{n,f}, \hat{\beta}_{n,f}) - L_f(\alpha_{n,f}, \beta_{n,f})\} = \mathrm{O}_{\mathbb{P}} \left(\sigma_n^2 \|q_n\|_\infty \frac{\log n}{n} \right), \qquad n \to \infty.$$

Second, using $\mathbb{E}(|G_n^{-1/2} g_n|_2^2) = \mathrm{tr}(G_n^{-1} G_n) = d_n$, we obtain $\mathbb{E}(|G_n^{-1/2} P_n(g_n)|_2^2) = \mathrm{O}(d_n/n)$ and it follows by Markov's inequality that

$$\left| G_n^{-1/2} P_n(g_n) \right|_2 = \mathrm{O}_{\mathbb{P}} \left(\sqrt{d_n/n} \right), \qquad n \to \infty.$$

Third, because $L_n^2 = \mathrm{O}(\gamma_n^2 \sqrt{n/\log(n)})$ implies $M_n^2 = \mathrm{O}(\sigma_n^2 \sqrt{n/\log(n)})$, which is also $\mathrm{O}(\sigma_n^2 n/\log(n))$, we can apply Theorem 2 to get

$$\sup_{f \in \mathscr{F}} |P_n(\varepsilon_{n,f})| = \mathrm{O}_{\mathbb{P}} \left(\sigma_n \sqrt{n^{-1} \log(n)} \right), \qquad n \to \infty.$$

□

G 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 C. We dealt with it via a special case of the concentration inequality in Theorem 2 in Major (2006), stated next.

Theorem 4 (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 $\nu \in (0, 1]$ be such that $\sup_{g \in \mathscr{G}} \mathbb{E}[g^2(X_1, X_2)] \leq \nu^2$. Let $X_1, \ldots, X_n$ be an independent random sample from P. There exist universal positive constants α, C and K such that*

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

where

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

Acknowledgments and Disclosure of Funding

The authors gratefully acknowledge comments and suggestions by an anonymous Reviewer that stimulated us to sharpen the main theorem and work out the application to Monte Carlo methods. The authors also thank Rémi Leluc for his valuable feedback and Aigerim Zhuman for her insightful remarks. This project was supported financially by 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.

Cannamela, C., J. Garnier, B. Iooss, et al. (2008). Controlled stratification for quantile estimation. *Annals of Applied Statistics 2*(4), 1554–1580.

Delyon, B. and F. Portier (2018). Asymptotic optimality of adaptive importance sampling. *Thirty-second Conference on Neural Information Processing Systems*.

Feng, M. B., A. Maggiar, J. Staum, and A. Wächter (2018). Uniform convergence of sample average approximation with adaptive multiple importance sampling. In *2018 Winter Simulation Conference (WSC)*, pp. 1646–1657. IEEE.

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 (1999). Laws of the iterated logarithm for censored data. *The Annals of Probability 27*(4), 2042–2067.

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.

Giné, E. and A. Guillou (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. H. Poincaré Probab. Statist. 38*(6), 907–921.

Glasserman, P. (2013). *Monte Carlo Methods in Financial Engineering*, Volume 53. Springer Science & Business Media.

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.

Hesterberg, T. C. and B. L. Nelson (1998). Control variates for probability and quantile estimation. *Management Science 44*(9), 1295–1312.

Hjort, N. L. and D. Pollard (2011). Asymptotics for minimisers of convex processes. *arXiv preprint arXiv:1107.3806*.

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.

Kloeckner, B. R. (2020). Empirical measures: regularity is a counter-curse to dimensionality. *ESAIM: Probability and Statistics 24*, 408–434.

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.

Law, A. M. and W. D. Kelton (2000). *Simulation Modeling and Analysis* (third ed.). McGraw-Hill New York.

Leluc, R., F. Portier, and J. Segers (2021). Control variate selection for Monte Carlo integration. *Statistics and Computing 31*, 50.

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.

McCulloch, C. E. and J. M. Neuhaus (2005). Generalized linear mixed models. In P. Armitage and T. Colton (Eds.), *Encyclopedia of Biostatistics*. Wiley Online Library.

McFadden, D. (2001). Economic choices. *American Economic Review 91*(3), 351–378.

McFadden, D. and P. A. Ruud (1994). Estimation by simulation. *The Review of Economics and Statistics 76*(4), 591–608.

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.

Nemirovski, A., A. Juditsky, G. Lan, and A. Shapiro (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization 19*(4), 1574–1609.

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.

Novak, E. (2016). Some results on the complexity of numerical integration. In *Monte Carlo and Quasi-Monte Carlo Methods*, pp. 161–183. Springer.

Oates, C. J., M. Girolami, and N. Chopin (2017). Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(3), 695–718.

Owen, A. B. (2013). *Monte Carlo Theory, Methods and Examples*.

Peel, T., S. Anthoine, and L. Ralaivola (2010). Empirical Bernstein inequalities for U-statistics. In *Neural Information Processing Systems (NIPS)*, Number 23, pp. 1903–1911.

Portier, F. and J. Segers (2018). On the weak convergence of the empirical conditional copula under a simplifying assumption. *Journal of Multivariate Analysis 166*, 160–181.

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

Shapiro, A., D. Dentcheva, and A. Ruszczyński (2014). *Lectures on Stochastic Programming: Modeling and Theory*. SIAM.

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

Talagrand, M. (1996). New concentration inequalities in product spaces. *Inventiones mathematicae 126*(3), 505–563.

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.