A high-dimensional Wilks phenomenon

Stéphane Boucheron · Pascal Massart

Abstract   A theorem by Wilks asserts that in smooth parametric density estimation the difference between the maximum likelihood and the likelihood of the sampling distribution converges toward a Chi-square distribution where the number of degrees of freedom coincides with the model dimension. This observation is at the core of some goodness-of-fit testing procedures and of some classical model selection methods. This paper describes a non-asymptotic version of the Wilks phenomenon in bounded contrast optimization procedures. Using concentration inequalities for general functions of independent random variables, it proves that in bounded contrast minimization (as for example in Statistical Learning Theory), the difference between the empirical risk of the minimizer of the true risk in the model and the minimum of the empirical risk (the excess empirical risk) satisfies a Bernstein-like inequality where the variance term reflects the dimension of the model and the scale term reflects the noise conditions. From a mathematical statistics viewpoint, the significance of this result comes from the recent observation that when using model selection via penalization, the excess empirical risk represents a minimum penalty if non-asymptotic guarantees concerning prediction error are to be provided. From the perspective of empirical process theory, this paper describes a concentration inequality for the supremum of a bounded non-centered (actually non-positive) empirical process. Combining the now classical analysis of M-estimation (building on Talagrand's inequality for suprema of

S. Boucheron
Laboratoire Probabilités et Modèles Aléatoires, Université Paris-Diderot,
175 rue du Chevaleret, 75013 Paris, France
e-mail: stephane.boucheron@math.jussieu.fr

P. Massart (✉)
Département de Mathématiques, Université Paris-Sud, 91405 Orsay, France
e-mail: pascal.massart@math.u-psud.fr

empirical processes) and versatile moment inequalities for functions of independent random variables, this paper develops a genuine Bernstein-like inequality that seems beyond the reach of traditional tools.

Keywords Wilks phenomenon · Risk estimates · Suprema of empirical processes · Concentration inequalities · Statistical learning

Mathematics Subject Classification (2000) 60E15 · 62G08 · 62H30

1 Introduction

Soon after Fisher popularized maximum likelihood methods, Wilks [49] described the so-called Wilks phenomenon. Assume $(P_\theta, \theta \in \Theta \subseteq \mathbb{R}^m)$ defines a family of probability distributions over some space $\mathcal{X}$ where for each θ, p_θ is a version of the density of P_θ with respect to some dominating measure μ. Given a sample $x_1, \ldots, x_n$, of elements of $\mathcal{X}$, the log-likelihood at θ is defined as $\ell_n(\theta) = \sum_{i=1}^n \log p_\theta(x_i)$. Assume that $\mu^{\otimes n}$-almost surely, there exists a maximizer $\widehat{\theta} \in \Theta$ of the log-likelihood, that is

$$\ell_n(\widehat{\theta}) = \sup_{\theta \in \Theta} \sum_{i=1}^n \log p_\theta(x_i).$$

Then, if the model is smooth enough (see [8,46], for precise definitions and demonstrations), assuming that $X_1, \ldots, X_n, \ldots$ are sampled independently according to P_θ, where $\theta \in \Theta$, the sequences of random variables $2\left(\ell_n(\widehat{\theta}) - \ell_n(\theta)\right)$ and $2nD(P_\theta, P_{\widehat{\theta}})$ both converge in distribution toward χ_m^2 (Chi-square distribution with m degrees of freedom) where $D(P, Q)$ denotes the relative entropy or Kullback information between distributions P and Q. This remarkable phenomenon can be used to build goodness-of-fit tests. In Gaussian linear models with known variance, the Wilks phenomenon is used to perform variable selection (although in disguised form) [1,28].

Considering maximum likelihood estimation as a special case of M-estimation, one may wonder whether extensions of the Wilks phenomenon exist. Does it hold in non-parametric density estimation problems where maximum likelihood estimation or generalizations of it can still be used? Does it still hold when dealing with other contrast optimization problems? A short glimpse at the proof of the Wilks phenomenon reveals that it depends on the asymptotic normality of $\sqrt{n}(\widehat{\theta} - \theta)$. When this last property does not hold, for example in non-parametric inference, we should look for reasonable generalizations. Such an endeavor is the main topic of the paper entitled "Generalized likelihood ratio statistics and Wilks phenomenon" by Fan et al. [17]. Though our aims are quite different, the findings of Fan et al. help to motivate our approach. They start by exploring a simple non-parametric Gaussian regression model where the parameter space is a Sobolev ball. The authors aim at testing whether the regression function is affine or not. They analyze a statistics that compares the maximum log-likelihood in the Sobolev ball and the maximum log-likelihood in the submodel constituted by affine functions, and generalizations of this approach. The traditional Wilks phenomenon does not hold in such an infinite-dimensional

setting. But, Fan et al. point out that the number of non-zero coefficients of the maximum likelihood estimator (the effective dimension of the maximum-likelihood estimator) inside the Sobolev ball tends to increase with sample size. Meanwhile, as m tends to infinity, if $X_m \sim \chi_m^2$ $(X_m - \mathbb{E}[X_m])/\sqrt{\mathbb{E}[X_m]}$ converges in distribution towards a centered Gaussian random variable with variance 2. This observation allows Fan et al. to propose a convenient generalization of the Wilks phenomenon: when centered, the difference between the maximum log-likelihood and the likelihood in the submodel should converge toward a Gamma distribution whose shape parameter depends mostly on the effective dimension of the maximum likelihood estimator and whose scale parameter is slowly varying with respect to sample size and effective dimension. This asymptotic pivotality property paves the way to non-trivial statistical applications. A generalized Wilks phenomenon has been reported in several settings [16,18,35]. But Fan et al. observe:

> While we have observed the Wilks phenomenon and demonstrated it for a few useful cases, it is impossible to verify the phenomenon for all non-parametric hypothesis problems. The Wilks phenomenon needs to be checked for other problems not covered in [17].

Rather than attempting to prove asymptotic distributional results, we are looking for non-asymptotic moment or tail bounds concerning analogues of $\ell_n(\widehat{\theta}) - \ell_n(\theta)$ that exhibit a Gamma-like behavior where the shape parameter reflects the model dimension rather than the sample size and the scale parameter depends only moderately on model dimension and sample size. To get a more concrete understanding, remember that if Z is a $\Gamma(p, 1/c)$-distributed random variable with expectation $\mathbb{E}[Z] = cp$ and variance $\sigma^2 = c^2 p$, then for $q \geq 2$, the qth norm of $(Z - \mathbb{E}[Z])_+$ can be upper-bounded by $\kappa\left(\sqrt{q\sigma^2} + cq\right)$ where κ is a universal constant that depends neither on the expectation nor on the variance of the Γ distribution.

The settings considered in [16,17,35] and the references we are aware of share a common feature: they are connected with density estimation in a Gaussian framework. They disregard robustness considerations. This contrasts with the setting we are interested in: statistical learning [39,48]. In statistical learning, a product set $\mathcal{X} \times \mathcal{Y}$ is endowed with a probability distribution P, the components of $\omega = (x, y) \in \mathcal{X} \times \mathcal{Y}$ are denoted by X $(X(\omega) = x)$ and Y $(Y(\omega) = y)$. In binary classification, $\mathcal{Y} = \{-1, 1\}$. A loss function ℓ maps $\mathcal{Y} \times \mathcal{Y}$ to $\mathbb{R}_+$. In the sequel, all loss functions are assumed to be bounded. For example, the classification loss is defined by $\ell(y, y') = \mathbf{1}_{y \neq y'}$ [48].

In the sequel, if U is distributed according to P, then for any (measurable) function g, the expected value of $g(U)$ is denoted either by Pg, $Pg(U)$, $\mathbb{E}_P[g(U)]$ or even $\mathbb{E}[g(U)]$ when P is clear from context. If $U_1, \ldots, U_n$ is a collection of independent random variables, P_n denotes the associated empirical distribution: $P_n g = \frac{1}{n} \sum_{i=1}^{n} g(U_i)$.

The learning problem consists in finding a function f on $\mathcal{X}$ such that the risk $R(f) = P\ell(f(X), Y)$ is as small as possible. Henceforth f^* is assumed to minimize $R(f)$ among all functions such that $(x, y) \mapsto \ell(f(x), y)$ is measurable. For example, in binary classification, the best possible classifier f^* which is called the Bayes classifier is defined from the regression function $\eta(x) = \mathbb{E}[Y \mid X = x]$ by $f^*(x) = \mathrm{sign}(\eta(x))$.

A learning algorithm typically looks for a good approximation $\widehat{f}$ of f^* in some class $\mathcal{F}$ of classifiers by minimizing an empirical risk like $R_n(f) = P_n\ell(f(X), Y)$. We do not assume that $f^* \in \mathcal{F}$ but we assume that some $\bar{f} \in \mathcal{F}$ minimizes the risk $R(\cdot)$ over $\mathcal{F}$. Henceforth, the model bias is $R(\bar{f}) - R(f^*)$. The connections between the richness of $\mathcal{F}$, the model bias, the distribution of $|\eta(\cdot) - 1/2|$ (the so-called noise conditions from [43]), and the distribution of the risk $R(\widehat{f}) = P\ell(\widehat{f}(X), Y)$ and more recently the excess risk $R(\widehat{f}) - R(\bar{f})$ have been the subject of intense research during the last 35 years [6,24,29,31–33,48]. The present paper is concerned with a rather atypical object, the empirical analogue of the excess risk, that from now on, will be called the unnormalized *empirical excess risk*:

$$n\left(R_n\left(\bar{f}\right) - R_n\left(\widehat{f}\right)\right) = n \sup_{f \in \mathcal{F}} \left(R_n\left(\bar{f}\right) - R_n(f)\right).$$

The purpose of the paper is to show that the right-tail bounds for the excess empirical risk described in [24,32] can be refined into concentration inequalities.

The main results of the paper are stated in Sects. 3 and 4. Upper-bounds on the variance of the empirical excess risk are first derived by building on the Efron–Stein inequalities [Inequality (2)] and classical tail bounds for excess risk and excess empirical risk (Theorem 1). Those variance estimates (Theorem 2) already show the merits of an ad hoc analysis of this atypical supremum of an empirical process. Variance estimates are illustrated in a simple setting: bounded regression using piecewise constant functions (histograms) with quadratic loss. It has been possible to check that the expected empirical excess risk is almost a linear function of the model dimension (that is the number D of pieces of the partition that defines the model) [4]: under mild conditions on the relationship between the number of classes and the sample size, $\mathbb{E}Z/\sigma^2 \approx D$ where $Z = n\left(R_n(\bar{f}) - R_n(\widehat{f}_n)\right)$. And in this simple context, Theorem 2 implies that the variance of the unnormalized empirical excess risk is upper-bounded by $\kappa\sigma^2 D$ where κ is some universal constant.

These variance estimates are amplified in Sect. 4 in order to estimate the higher moments of the centered empirical excess risk. Taking advantage on moment estimates for general functions of independent random variables derived in [10], it is possible to show that the unnormalized empirical excess risk satisfies a Bernstein-like moment inequality where the variance term is of the order of the model dimension, while the scale term does not depend too much on either sample size or model dimension.

The end of Sect. 4 turns back to the regression problem. It follows that for a fixed noise level σ, there exists a universal constant κ such that the unnormalized empirical excess risk satisfies:

$$\|Z - \mathbb{E}Z\|_q \leq \kappa \left(\sqrt{Dq} + q\right)$$

for all sample sizes n and all partition cardinalities D where κ is some universal constant. In this simplified setting, the empirical excess risk exemplifies the high dimensional Wilks phenomenon. The significance of this observation is best understood by having a second look at the already mentioned paper by Arlot and Massart [4] on model selection. The concentration of the empirical excess risk is the keystone

in the derivation of an adaptive model selection method: it justifies a data-dependent penalty calibration technique.

Section 5 describes what can be obtained from theorems from Sect. 4 in classification problems. When using Vapnik–Chervonenkis classes of classifiers, the VC-dimension is a well-known surrogate for the traditional metric dimension that appears in the analysis of smooth parametric models. Under the so-called random classification noise [3,7,23], noise conditions turn out to be very simple and the sharp estimates on the excess risk derived by Massart and Nedelec [33] can be injected in Theorems 2 and 5. In the absence of model bias, variance and higher moment upper-bounds for excess empirical risk provide another version of the Wilks phenomenon: variance is upper-bounded by the VC-dimension multiplied by a slowly varying function of the ratio between sample size and VC-dimension. And, as far as the higher moment estimates are concerned, the scale parameter only depends on the noise conditions.

2 Notation and background

The expression $n\left(R_n(\bar{f}) - R_n(\widehat{f})\right) = n\sup_{f\in\mathcal{F}}\left(R_n(\bar{f}) - R_n(f)\right)$ stresses the fact that the empirical excess risk is the supremum of an empirical process. The latter is atypical since the expected value of the supremum is always non-negative while for each $f \in \mathcal{F}$, $\mathbb{E}\left[R_n(\bar{f} - R_n(f)\right] \leq 0$. We are not aware of any direct method that could be used to analyze the excess empirical risk.

In the sequel, tools from empirical process theory are not directly applied to $\mathcal{F}$ but rather to the associated (relative) loss class $\mathcal{H}$. The loss class $\mathcal{H}$ generated by $\mathcal{F}$ is defined as

$$\mathcal{H} = \left\{(x, y) \mapsto \ell\left(\bar{f}(x), y\right) - \ell(f(x), y) \; : \; f \in \mathcal{F}\right\}.$$

Henceforth h^* denotes the loss associated with the global risk minimizer f^*

$$h^*(x, y) = \ell\left(\bar{f}(x), y\right) - \ell(f^*(x), y),$$

and $\widehat{h}_n$ denotes the loss function associated with the empirical risk minimizer $\widehat{f}$:

$$\widehat{h}_n(x, y) = \ell\left(\bar{f}(x), y\right) - \ell\left(\widehat{f}(x), y\right).$$

When n is clear from context, we may abbreviate $\widehat{h}_n$ by $\widehat{h}$. The excess empirical risk is thus rewritten as a supremum of an empirical process: $n\left(R_n(\bar{f}) - R_n(\widehat{f})\right) = nP_n\widehat{h}_n = n\sup_{h\in\mathcal{H}} P_n h$. Note that the model bias $R\left(\bar{f}\right) - R(f^*)$ is equal to Ph^* and that the excess risk $R\left(\widehat{f}\right) - R(\bar{f})$ can be rewritten as $-P\widehat{h}$.

For the sake of simplicity, in the sequel, we assume that $\mathcal{H}$ is separable, that is $\mathcal{H}$ contains a dense countable subset $\mathcal{G}$ (with respect to the topology of pointwise convergence). Since $\mathcal{H}$ is also assumed to be uniformly bounded, dominated convergence arguments ensure for example that:

$$\sup_{h\in\mathcal{H}}(P_n - P)h = \sup_{h\in\mathcal{G}}(P_n - P)h.$$

This ensures that the supremum over $\mathcal{H}$ is measurable. Suprema over intersection of $\mathcal{H}$ with various balls are checked to be measurable by similar arguments. We assume furthermore that all suprema over $\mathcal{H}$ are achieved. Dealing with approximate minimizers of the empirical risk makes proof slightly lengthier without providing more insights.

We need to recall a few important notions that allow us to derive tail-bounds for excess risk and empirical excess risk. Those tail bounds are not meant to precisely describe the fluctuations of the empirical excess risk around its expectations. This is the purpose of the concentration inequalities stated in this paper.

Thanks to ideas borrowed from robust statistics [21], empirical process theory [40, 45,47] and the theory of concentration inequalities [26,32,41,42], we now have a rather precise idea of the aforementioned connexions ([9,32], for surveys).

In the sequel, various properties of the following class of positive sub-linear functions are repeatedly used. The class $\mathcal{C}_1$ comprises non-decreasing and continuous functions ϕ from $\mathbb{R}_+$ to $\mathbb{R}_+$ such that $\phi(x)/x$ is non-increasing and $\phi(1) \geq 1$. Note that if ψ and ρ belong to $\mathcal{C}_1$, so do $\psi \circ \rho$ and $\psi + \rho$. Moreover, for any $\alpha \geq \psi(1)$, the equation $\alpha r^2 = \psi(r)$ has a unique solution in $(0, 1]$. One can check that any function $\psi \in \mathcal{C}_1$, is sub-additive : $\psi(u + v) \leq \psi(u) + \psi(v)$. Note also that if U is distributed according to P, $\mathbb{E}_P[\psi(U)] \leq 2\psi(\mathbb{E}_P[U])$.

In the sequel, we assume that there exists a pseudo-distance d on the set of bounded real functions on $\mathcal{X}$, such that for $f \in \mathcal{F}$

$$\mathbb{E}_P\left[\left(\ell(f(X), Y) - \ell(f^*(X), Y)\right)^2\right] \leq d^2(f, f^*), \tag{1}$$

or equivalently, using relative loss functions: $P(h^* - h)^2 \leq d^2(f, f^*)$. In some situations, as when analyzing classification problems, it is possible to define d as $d^2(f, f^*) = P(h^* - h)^2$, but in other circumstances, for example when investigating bounded regression, it is convenient to define d in other ways.

The complexity assumption below aims at describing the richness of the L_2 neighborhood of 0 in the loss class $\mathcal{H}$.

Assumption 1 (*Complexity assumption*) There exists some function $\psi \in \mathcal{C}_1$ such that for all r and all integers n satisfying $\psi(r) \leq \sqrt{n}r^2$:

$$\sqrt{n}\mathbb{E}\left[\sup_{f \in \mathcal{F}:\, d^2(f,\bar{f}) \leq r^2} \left|(P_n - P)\left(\ell(f(X), Y) - \ell\left(\bar{f}(X), Y\right)\right)\right|\right] \leq \psi(r).$$

The possibility to derive interesting bounds (fast rates) for the excess risk critically relies on the next assumption [6,9,24,29,31,32]. The name "noise conditions" will be clarified through examples in Sects. 3, 4 and 5.

Assumption 2 (*noise conditions*) There exists some function $\omega \in \mathcal{C}_1$ such that for all f in $\mathcal{F}$ (with associated relative loss function $h \in \mathcal{H}$):

$$d^2(f, f^*) \leq \omega^2\left(\sqrt{P(h^* - h)}\right) = \omega^2\left(\sqrt{R(f) - R(f^*)}\right).$$

Throughout the paper, a *learning task* is defined by a sampling probability P over $\mathcal{X} \times \mathcal{Y}$, a loss function ℓ, and a collection of functions $\mathcal{F}$ (a relative loss class $\mathcal{H}$ is automatically associated with the learning task). If the complexity and noise conditions assumptions are satisfied, the positive root r_* of the equation

$$\sqrt{n} r^2 = \psi(\omega(r))$$

is a key quantity [45–47]. It appears in upper-bounds for the expected excess risk $\mathbb{E}[-P\widehat{h}_n]$ and the expected empirical risk $\mathbb{E}[P_n\widehat{h}_n]$. Moreover, the analysis reveals that a function of r_* upper-bounds the expected value of $P\widehat{h}_n^2$ and $P_n\widehat{h}_n^2$ (Theorems 1 and 4).

The positive root r_* of equation $\sqrt{n} r^2 = \psi(\omega(r))$ also shows up in the detailed analysis of our main object of concern, the unnormalized empirical excess risk $n P_n\widehat{h}_n$.

3 The variance of the empirical excess risk

3.1 A generic upper-bound

The first step in understanding the fluctuations of the empirical excess risk around its expectation consists of upper-bounding the variance and checking that it can be related with the expectation. In order to carry out this program we use the Efron–Stein inequalities.

Let Z denote a (square-integrable) function of a sequence of independent random variables $(U_1, U_2, \ldots, U_n)$, that is $Z = F(U_1, \ldots, U_n)$ for some function F. Let $U_1', \ldots, U_n'$ be distributed as $U_1, \ldots, U_n$ and be independent of $U_1, \ldots, U_n$. For each $i \leq n$, let $Z_i' = F(U_1, \ldots, U_{i-1}, U_i', U_{i+1}, \ldots, U_n)$, let $U^{(i)} = (U_1, \ldots, U_{i-1}, U_{i+1}, \ldots, U_n)$. Let F_i denote a function of $n-1$ arguments and let $Z_i = F_i(U_1, \ldots, U_{i-1}, U_{i+1}, U_n) = F_i(U^{(i)})$. The jackknife estimates of variance V_+ and V [36,44] are defined by

$$V_+ = \sum_{i=1}^{n} \mathbb{E}\left[(Z - Z_i')_+^2 \mid U_1, \ldots, U_n \right] \quad \text{and} \quad V = \sum_{i=1}^{n} (Z - Z_i)^2.$$

The Efron–Stein inequalities [15] assert that the expectation of the jackknife estimates of variance are upper-bounds:

$$\text{Var}[Z] \leq \mathbb{E}[V_+] \leq \mathbb{E}[V]. \tag{2}$$

Before handling the variance of the empirical excess risk, let us recall the following upper-bounds on the variance of suprema of general empirical processes. In the statement of the next proposition $\mathcal{H}$ denotes a generic class of bounded functions over some space $\mathcal{X}$. The random variables U_i are assumed to be independent and $\mathcal{X}$-valued. The following inequality follows from $\text{Var}[Z] \leq \mathbb{E}[V_+]$: let $Z = n \sup_{h \in \mathcal{H}} P_n h$, then

$$\text{Var}[Z] \leq 2\mathbb{E}\left[n \sup_{h \in \mathcal{H}} P_n h^2\right] + 2 \sup_{h \in \mathcal{H}} \mathbb{E}\left[\sum_{i=1}^{n} h^2(U_i)\right]. \tag{3}$$

The upper-bound can be further simplified and replaced by $4\mathbb{E}\left[n \sup_{h \in \mathcal{H}} P_n h^2\right]$.

When the random variables $h(U_i)$ are centered and identically distributed, the following handy bound can be derived as in [38]:

$$\text{Var}[Z] \leq 2b\mathbb{E}[Z] + n\sigma^2 \tag{4}$$

where $b = \sup_{h \in \mathcal{H}} \sup_x |h(x)|$ and $\sigma^2 = \sup_{h \in \mathcal{H}} \text{Var}[h(U)]$.

As the empirical excess risk is also the supremum of an empirical process, it is not too surprising that the Efron–Stein inequalities allow us to relate the variance of the empirical excess risk with its expectation and with various functions of the expected excess risk.

Proposition 1 *Let $\mathcal{H}$ denote a class of measurable functions over $\mathcal{U}$.*

Let P denote a probability distribution over $\mathcal{U}$. Assume that functions in $\mathcal{H}$ are square-integrable. Let $\widehat{h}_n$ maximize $P_n h$ over $\mathcal{H}$. Then

$$\frac{Var\left[n P_n \widehat{h}_n\right]}{2n} \leq \mathbb{E}\left[P_n \widehat{h}_n^2\right] + \mathbb{E}\left[P \widehat{h}_n^2\right]. \tag{5}$$

Assume now that for all $h \in \mathcal{H}$, $Ph \leq 0$, that the null function belongs to $\mathcal{H}$ and that $\|h\|_\infty \leq 1$, then

$$\frac{Var\left[n P_n \widehat{h}_n\right]}{n} \leq 2\mathbb{E}\left[(P_{n-1} - P)\widehat{h}_{n-1}\right] + \mathbb{E}\left[P \widehat{h}_{n-1}^2\right]. \tag{6}$$

Proof Let $(U_1, \ldots, U_n)$ be independently, identically distributed according to P. Let $(U_1', \ldots, U_n')$ denote an independent copy of $(U_1, \ldots, U_n)$. The first inequality in the proposition is obtained by using the first Efron–Stein inequality ($\text{Var}[Z] \leq \mathbb{E}[V_+]$), and taking advantage of the fact that for each $i \in 1, \ldots, n$

$$\left(Z - Z_i'\right) \leq \widehat{h}_n(U_i) - \widehat{h}_n\left(U_i'\right).$$

Thus, as $(a + b)^2 \leq 2(a^2 + b^2)$, taking expectations over primed variables:

$$V_+ \leq 2n\left(P_n \widehat{h}_n^2 + P \widehat{h}_n^2\right).$$

Taking expectations over $U_1, \ldots, U_n$:

$$\text{Var}\left[n P_n \widehat{h}_n\right] \leq \mathbb{E}[V_+] \leq 2n\mathbb{E}\left[P_n \widehat{h}_n^2\right] + 2n\mathbb{E}\left[P \widehat{h}_n^2\right].$$

Let us now turn to the proof of the second inequality. Let $\widehat{h}_n^{(i)}$ denote the maximizer of the empirical process when the ith element has been removed from the sample, that is $\widehat{h}_n^{(i)}$ maximizes $\sum_{j \leq n, j \neq i} h(U_j)$. Thus

$$Z = \sum_{i \le n} \widehat{h}_n(U_i) \quad \text{and let} \quad Z_i = \sum_{j \le n, j \ne i} \widehat{h}_n^{(i)}(U_j).$$

Note first that for each $i \in 1, \dots, n$

$$-1 \le \widehat{h}_n^{(i)}(U_i) \le Z - Z_i \le \widehat{h}_n(U_i) \le 1,$$

hence, the so-called self-boundedness property holds: $\sum_{i=1}^{n}(Z - Z_i) \le Z$. For each sample $(u_1, \dots, u_n)$, for each i,

$$(Z - Z_i)^2 - \left(\widehat{h}_n^{(i)}(u_i)\right)^2 = \left(Z - Z_i + \widehat{h}_n^{(i)}(u_i)\right)\left(Z - Z_i - \widehat{h}_n^{(i)}(u_i)\right)$$

and since the second factor is non-negative

$$\le 2\left(Z - Z_i - \widehat{h}_n^{(i)}(u_i)\right).$$

Summing over all i, using the self-boundedness property:

$$\sum_{i=1}^{n}(Z - Z_i)^2 \le 2Z + \sum_{i=1}^{n}\left(\widehat{h}_n^{(i)}(u_i)\right)^2 - 2\sum_{i=1}^{n}\widehat{h}_n^{(i)}(u_i).$$

Now, taking expectations and invoking the Efron–Stein inequality:

$$\begin{aligned}
\text{Var}[Z] &\le 2\mathbb{E}[Z] + \sum_{i=1}^{n}\mathbb{E}\left[\left(\widehat{h}_n^{(i)}(U_i)\right)^2\right] - 2\sum_{i=1}^{n}\mathbb{E}\left[\widehat{h}_n^{(i)}(U_i)\right] \\
&= 2\mathbb{E}[Z] + n\mathbb{E}\left[\left(\widehat{h}_n^{(n)}(U_n)\right)^2\right] - 2n\mathbb{E}\left[P\widehat{h}_n^{(n)}\right] \quad \text{by exchangeability} \\
&= 2n\mathbb{E}[P_n\widehat{h}_n] + n\mathbb{E}\left[P\widehat{h}_{n-1}^2\right] - 2n\mathbb{E}\left[P\widehat{h}_{n-1}\right].
\end{aligned}$$

It remains to check that $\mathbb{E}\left[P_n\widehat{h}_n\right] \le \mathbb{E}[P_{n-1}\widehat{h}_{n-1}]$. This follows from a classical argument that works for suprema of empirical processes provided some measurability conditions are met. Let $\mathcal{G}_n$ denote the sub-σ-algebra of $\sigma(U_i, i \in \mathbb{N})$ consisting of events that are invariant under permutation of the first n coordinates. Then P_n is a measure-valued backward martingale: for any $f \in \mathcal{H}, \mathbb{E}\left[P_{n-1}h \mid \mathcal{G}_n\right] = P_n h$ a.s. As a supremum of linear functionals, $P_n\widehat{h}_n$ is a backward submartingale with respect to the decreasing filtration $(\mathcal{G}_n)$ [34]. Almost surely, we have

$$\begin{aligned}
\mathbb{E}\left[P_{n-1}\widehat{h}_{n-1} \mid \mathcal{G}_n\right] &= \mathbb{E}\left[\sup_{h \in \mathcal{H}} P_{n-1}h \mid \mathcal{G}_n\right] \\
&\ge \sup_{h \in \mathcal{H}} \mathbb{E}\left[P_{n-1}h \mid \mathcal{G}_n\right] \\
&= \sup_{h \in \mathcal{H}} P_n h \\
&= P_n\widehat{h}_n.
\end{aligned}$$

So that by deconditioning $\mathbb{E}\left[P_{n-1}\widehat{h}_{n-1}\right] \geq \mathbb{E}\left[P_n\widehat{h}_n\right]$. □

Remark 1 Note that we refrained from using the immediate upper-bounds

$$P_n\widehat{h}_n^2 \leq \sup_{h \in \mathcal{H}} P_n h^2 \quad \text{or} \quad P\widehat{h}_n^2 \leq \sup_{h \in \mathcal{H}} P h^2$$

as is usually done in the proof of the above-stated upper-bounds on the variance of suprema of general empirical processes [11,38]. This would lead to bounds like (3) and would not allow us to take advantage of the fact that, under favorable noise conditions, with high probability, the supremum of the empirical process is achieved at some $\widehat{f}_n$ in the L_2 neighborhood of $\bar{f}$.

3.2 Learning rates and variance upper bounds

Up to now, it is not clear whether cautiousness is to be rewarded. In order to check that this is the case, we turn back to statistical learning problems, and invoke Proposition 1 when the function class denoted by $\mathcal{H}$ in Proposition 1 turns out to be the loss class associated with a learning task as explained in Assumption 2. In order to take full advantage of Proposition 1, we need non-trivial bounds on the expected values of what becomes the excess risk $-P\widehat{h}_n$ and the excess empirical risk $P_n\widehat{h}_n$.

Remark 2 When $\mathcal{H}$ is the loss class associated with a learning task, inequality (5) stresses the fact that the variance of the empirical excess risk is upper-bounded using the second moments of the $L_2(P)$ and the $L_2(P_n)$ distances between the risk minimizer and the empirical risk minimizer. A similar connection between the so-called L_1-stability $\mathbb{E}\left[P|\widehat{h}_n - \widehat{h}_{n-1}|\right]$ and the bias of the leave-one-out estimate of the risk had been pointed out by Devroye and Wagner [14] and further investigated by Kearns and Ron [22] and Rakhlin et al. [37] (see Theorem 4.1 there). Note that $\mathbb{E}\left[P|\widehat{h}_n - \widehat{h}_{n-1}|\right] \leq \mathbb{E}P\widehat{h}_n^2 + \mathbb{E}P\widehat{h}_{n-1}^2$.

The next theorem gathers a slight refinement of classical deviation inequalities for the excess risk and the excess empirical risk. Complete derivations can be found in [32,33]. A different style of presentation and derivation can be found in [24]. A roadmap of the proof is given in the appendix following the blend introduced in [9].

Theorem 1 *Let $\mathcal{H}$ denote the loss class associated with a learning task that satisfies the complexity Assumption (1) with function ψ and the noise conditions Assumption (2) with function ω. Let r_* denote the positive solution of*

$$\sqrt{n}r^2 = \psi(\omega(r)).$$

Let h^* denote the relative loss function associated with the minimizer of the risk over all measurable functions. Then, with probability larger than $1 - 2\delta$, the excess risk $-P\widehat{h}_n$ and the excess empirical risk $P_n\widehat{h}_n$ satisfy:

$$\max\left(-P\widehat{h}_n, P_n\widehat{h}_n\right) \leq \kappa\left(Ph^* + r_*^2 + \frac{\omega^2(r_*)}{nr_*^2}\log\frac{1}{\delta}\right).$$

and

$$\max \left(\mathbb{E}\left[-P\widehat{h}_n \right], \mathbb{E}\left[P_n\widehat{h}_n \right] \right) \leq \kappa \left(Ph^* + r_*^2 \right).$$

In both inequalities, κ denotes a universal constant larger than 1 that does not depend on the learning task.

Remark 3 In the stochastic upper-bound, $\omega^2(r_*)/(nr_*^2)$ may be further upper-bounded by r_*^2. This upper-bound may be used to simplify the statement of the Theorem, but it is important to keep the result in this form in order to derive meaningful results in Sects. 4 and 5.

We may now combine the second inequality from Proposition 1 and the preceding Theorem to get meaningful upper-bounds for the variance of the empirical excess risk.

Theorem 2 *There exists a universal constant κ such that, using notations and assumptions of Theorem 1, the variance of the empirical excess risk is upper-bounded by*

$$Var\left[nP_n\widehat{h}_n \right] \leq n\kappa \left(\omega^2(r_*) + \omega^2\left(\sqrt{Ph^*} \right) \right).$$

Proof Thanks to Proposition 1, we have:

$$Var\left[nP_n\widehat{h}_n \right] \leq 2n\mathbb{E}\left[(P_{n-1} - P)\widehat{h}_{n-1} \right] + 2n\mathbb{E}\left[P\widehat{h}_n^2 \right].$$

Thanks to the fact that $\omega^2(\sqrt{.}) \in \mathcal{C}_1$, the last summand can be upper-bounded:

$$\begin{aligned} 2n\mathbb{E}\left[P\widehat{h}_n^2 \right] &\leq 4n \left(\mathbb{E}\left[P(\widehat{h}_n - h^*)^2 \right] + Ph^{*2} \right) \quad \text{as } (x+y)^2 \leq 2(x^2 + y^2) \\ &\leq 4n \left(\mathbb{E}\left[\omega^2\left(\sqrt{P(h^* - \widehat{h}_n)} \right) \right] + \omega^2\left(\sqrt{Ph^*} \right) \right) \\ &\leq 4n \left(2\omega^2\left(\sqrt{\mathbb{E}\left[P(h^* - \widehat{h}_n) \right]} \right) + \omega^2\left(\sqrt{Ph^*} \right) \right). \end{aligned}$$

Let us now relate $\mathbb{E}\left[(P_{n-1} - P)\widehat{h}_{n-1} \right]$ with bounds on the excess empirical risk and the excess risk over samples of size n. Let s_* denote the positive solution of $\sqrt{n-1}r^2 = \psi(2\omega(r))$. It follows from Theorem 1 that:

$$\mathbb{E}\left[(P_{n-1} - P)\widehat{h}_{n-1} \right] \leq 2\kappa \left(Ph^* + s_*^2 \right).$$

Now as $\sqrt{n-1}s_*^2 = \psi(\omega(s_*))$ and $\sqrt{n}r_*^2 = \psi(\omega(r_*))$, we have $1 \leq s_*^2/r_*^2 \leq \frac{n}{n-1}$.
So, plugging upper-bounds on expected and empirical excess risk from Theorem 1:

$$
\begin{aligned}
\mathrm{Var} &\left[n \, P_n \widehat{h}_n \right] \\
&\leq 4\kappa n \left(\frac{nr_*^2}{n-1} + Ph^* \right) + 8n\omega^2 \left(\sqrt{\kappa(Ph^* + r_*^2)} \right) + 12n\omega^2 \left(\sqrt{Ph^*} \right) \\
&\leq \kappa' n \left(r_*^2 + Ph^* + \omega^2 \left(\sqrt{Ph^*} \right) + \omega^2 (r_*) \right) \\
&\leq \kappa'' n \left(\omega^2 \left(\sqrt{Ph^*} \right) + \omega^2 (r_*) \right).
\end{aligned}
$$

$\square$

Remark 4 The second bound in Proposition 1 provides an easy way to relate the variance of the empirical excess risk with r_* and the model bias Ph^*. With a little more work, the first bound would lead to the same kind of variance upper-bounds and it will prove very convenient when deriving higher moment estimates in the next section.

3.3 Application to bounded regression

Let us illustrate the variance bounds on a problem that provides a bridge between statistical learning and traditional statistics: bounded regression using least square estimation. This simple setting has been used to investigate model selection using (careful) data-dependent penalization by Arlot and Massart [4]. The latter paper takes advantage of the generalized Wilks phenomenon to build and calibrate nearly minimal penalties and to derive adaptive estimators.

Recall that the quadratic loss is defined by $\ell(y, y') = (y - y')^2$. Labels Y are assumed to be $[-1/2, 1/2]$-valued. In the setting of the paper, the pseudo-distance d satisfying (1) and used in the statement of the complexity assumption and of the noise conditions assumption in Sect. 2 is defined in the following way. Let $f, f' \in \mathcal{F}$ then

$$
d^2(f, f') = 2P\ell(f(X), f'(X)).
$$

If $\bar{f} = f^*$ (we will see that this is a sensible assumption), the pseudo-distance can be related with the $L_2(P)$ norm of the (relative) loss functions:

$$
\begin{aligned}
Ph^2 &= P \left((f(X) - f^*(X))^2 \left((f(X) - f^*(X)) - 2(Y - f^*(X)) \right)^2 \right) \\
&= P \left((f(X) - f^*(X))^2 \left((f(X) - f^*(X))^2 + 4(Y - f^*(X))^2 \right) \right) \\
&\leq P \left(f(X) - f^*(X) \right)^4 + 4P \left((f(X) - f^*(X))^2 (Y - f^*(X))^2 \right) \\
&\leq 2P(f - f^*)^2 \\
&= d^2(f, f^*)
\end{aligned}
$$

where the second equation is obtained by reasoning conditionally on X, and the inequalities follow from the boundedness assumptions on Y, f and f'.

As $R(f) - R(f^*) = P(f - f^*)^2$, we may choose $\omega(r) = \sqrt{2}r$ in order to comply with Assumption 2.

We consider a very simple estimator, the *regressogram*. The input space is partitioned into a finite (say D) number of pairwise disjoint subsets. Let Π denote the partition. The class $\mathcal{F}$ is defined as the set of functions that are constant over the subsets defining Π. Any function f in $\mathcal{F}$ is defined by a set of coefficients $(a_I)_{I \in \Pi}$:

$$f(x) = \sum_{I \in \Pi} a_I \mathbb{1}_I(x).$$

Given a sample of labelled points $(X_i, Y_i)_{i \le n}$ where Y_i are $[-1/2, 1/2]$-valued, the regressogram is the minimizer $\widehat{f}$ of the empirical quadratic risk over $\mathcal{F}$. It is defined by the set of coefficients $(\widehat{a}_I)_{I \in \Pi}$

$$\widehat{a}_I = \begin{cases} \frac{P_n(Y\mathbb{1}_I(X))}{P_n\mathbb{1}_I(X)} = \frac{P_n(Y\mathbb{1}_I(X))}{P_n(I)} & \text{if } P_n(I) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Note that the pseudo-distance is generated by a weighted pseudo-norm, if we let $f, f' \in \mathcal{F}$ be defined by vectors $\mathbf{a} = (a_I)_{I \in \Pi}$ and $\mathbf{b} = (b_I)_{I \in \Pi}$,

$$\|\mathbf{a}\|_2^2 = \sum_{I \in \Pi} P(I)a_I^2.$$

then $d^2(f, f') = 2\|\mathbf{a} - \mathbf{b}\|_2^2$.

Thanks to Pythagorean relationships, as $\mathcal{F}$ is a closed subspace of $L_2(P)$, we may replace f^* by its projection onto $\mathcal{F}$. And thus, without loss of generality, we may assume that f^* belongs to $\mathcal{F}$. The function f^* is defined by the vector of coefficients $\mathbf{a}^*$ defined by

$$a_I^* = \begin{cases} \frac{P(Y\mathbb{1}_I(X))}{P\mathbb{1}_I(X)} = \frac{P(Y\mathbb{1}_I(X))}{P(I)} & \text{if } P(I) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

In order to simplify notations and computations, we further assume that f^* is the null function and that Y may be considered as additive homoschedastic centered noise. Let σ^2 denote the noise variance. Details concerning the general setting can be found in [4].

In that setting, we may choose

$$\psi(r) = 2r\sqrt{D}. \tag{7}$$

Indeed

$$(P_n - P)h = 2P_n(Yf(X)) - P_n f^2 + P f^2.$$

$$\mathbb{E}\left[\sup_{f:\, d^2(f,f^*)\leq r^2} |(P_n - P)h|\right] \leq \mathbb{E}\left[\sup_{f:\, d^2(f,f^*)\leq r^2} |2P_n(Yf(X))|\right]$$

$$+ \mathbb{E}\left[\sup_{f:\, d^2(f,f^*)\leq r^2} \left|(P - P_n)f^2\right|\right].$$

Recall that thanks to the assumption $f^* = 0$, $d^2(f, f^*) = 2\|\mathbf{a}\|^2$. Hence

$$\sup_{f:\, d^2(f,f^*)\leq r^2} |2P_n(Yf(X))| = \frac{2}{n} \sup_{\mathbf{a}:2\|\mathbf{a}\|^2\leq r^2} \sum_{I\in\Pi}\sum_{i\leq n} (Y_i \mathbb{1}_I(X_i)) a_I$$

$$\leq \frac{2}{n} \sup_{\mathbf{a}:\sqrt{2}\|\mathbf{a}\|\leq r} \|\mathbf{a}\| \sqrt{\sum_{I\in\Pi} \frac{\left(\sum_i Y_i \mathbb{1}_I(X_i)\right)^2}{P(I)}}$$

by Cauchy–Schwarz inequality.

$$\leq \frac{\sqrt{2}r}{n} \sqrt{\sum_{I\in\Pi} \frac{\left(\sum_i Y_i \mathbb{1}_I(X_i)\right)^2}{P(I)}}.$$

Taking expected values

$$\mathbb{E}\left[\sup_{f:\, d^2(f,f^*)\leq r^2} |2P_n(Yf(X))|\right] \leq \frac{\sqrt{2}r}{\sqrt{n}}\sqrt{D}.$$

On the other hand, using Cauchy–Schwarz inequality once again,

$$\mathbb{E}\left[\sup_{f:\, d^2(f,f^*)\leq r^2} \left|(P - P_n)f^2\right|\right] \leq \frac{r}{\sqrt{8}}\mathbb{E}\left[\left(\sum_{I\in\Pi} \frac{((P_n - P)(I))^2}{P(I)}\right)^{1/2}\right].$$

The right-hand side is the expected value of the square root of a Pearson statistics. Its expected value can be upper-bounded by $\sqrt{D/n}$. Then (7) follows from $\sqrt{2} + 1/\sqrt{8} \leq 2$.

Hence, r_* may be chosen as

$$r_* = \sqrt{\frac{8D}{n}}.$$

Hence, from Theorem 2, we obtain the following upper-bound on the variance of the unnormalized empirical excess risk:

$$8\kappa D.$$

In the simple setting considered here, it is possible to lower bound the expected value of the empirical excess risk. In order to alleviate computations, we further assume that $P(I) = 1/D$ for all $I \in \Pi$.

$$R_n(f^*) - R_n\left(\widehat{f}\right) = P_n\left(\sum_{I \in \Pi} \left(Y^2 - (Y - \widehat{a}_I)^2\right) \mathbb{1}_I(X)\right)$$

$$= \sum_{I \in \Pi} P_n(I)\widehat{a}_I^2.$$

Now, taking expectation over samples,

$$\mathbb{E}\left[R_n(f^*) - R_n\left(\widehat{f}\right)\right] = \frac{\sigma^2}{n} \sum_{I \in \Pi} \mathbb{E}[\mathbb{1}_{P_n(I)>0}]$$

$$= \frac{\sigma^2}{n} \sum_{I \in \Pi} \left((1 - (1 - P(I))^n\right)$$

$$\geq \frac{\sigma^2}{n} \sum_{I \in \Pi} \left(1 - \exp(-n/D)\right).$$

If $n/D \geq \log(2)$, the last sum is larger than $D/2$. Note that $\mathbb{E}\left[R_n(f^*) - R_n(\widehat{f})\right] \leq D\sigma^2/n$.

For a fixed value of σ^2, the expected value and the variance of the unnormalized empirical excess risk are both proportional to D. As far as the first two moments are concerned, the bounded regression problem exemplifies the high-dimensional Wilks phenomenon.

4 Higher moments

The purpose of this section is to check that the excess empirical risk satisfies a Bernstein-like inequality. The classical Bernstein inequality for sums of independent random variables asserts the following. Let $X_1, \ldots, X_n$ be independent real-valued random variables. Assume that there exist positive numbers v and c such that $\sum_{i=1}^n \mathbb{E}\left[X_i^2\right] \leq v$ and

$$\sum_{i=1}^n \mathbb{E}\left[(X_i)_+^k\right] \leq \frac{k!}{2} v c^{k-2} \quad \text{for all integers } k \geq 3.$$

If $S = \sum_{i=1}^n (X_i - \mathbb{E}X_i)$, then for all $\lambda \in (0, 1/c)$ and $t > 0$,

$$\log \mathbb{E}e^{\lambda S} \leq \frac{v\lambda^2}{2(1 - c\lambda)},$$

so that for all $t > 0$,

$$P\left\{S \geq \sqrt{2vt} + ct\right\} \leq e^{-t}.$$

Tail bounds, bounds on the logarithmic moment generating functions are known to be equivalent to moment bounds (see [27], for the "equivalence of moments principle"). Let Y be a centered random variable. It can be checked that if for some positive v and c

$$P\left\{Y > \sqrt{2vt} + ct\right\} \wedge P\left\{-Y > \sqrt{2vt} + ct\right\} \leq e^{-t} \quad \text{for } t > 0,$$

then for every integer $q \geq 1$

$$\mathbb{E}[Y^{2q}] \leq q!(8v)^q + (2q)!(4c)^{2q}.$$

And conversely, if for some positive constants A, B, for every integer $q \geq 1$

$$\mathbb{E}[Y^{2q}] \leq q!A^q + (2q)!B^{2q}$$

then

$$P\left\{Y > \sqrt{8(A + B^2)t} + 2Bt\right\} \wedge P\left\{-Y > \sqrt{8(A + B^2)t} + 2Bt\right\} \leq e^{-t} \quad \text{for } t > 0.$$

Recall that if the X_i are recentered independent $\Gamma(v/c^2, 1/c)$-distributed, then the preceding conditions are satisfied. This is true if the X_i are recentered χ_1^2 random variables, then $c = 2$.

Asserting that an M-estimation problem (for example a learning problem) exhibits a high dimensional Wilks phenomenon means that the q-norm of the centered empirical excess risk is upper-bounded by $\kappa\left(\sqrt{qv} + cq\right)$ where κ is a(another) universal constant, v is a sharp variance upper-bound and c depends slowly on sample size and model complexity/dimension, and is related to the ratio between the expected value and the variance.

Theorem 1 from the preceding section, provides deviation inequalities for empirical excess risk. Moreover, the statement of the theorem reveals that the upper-tail is not heavier than the upper-tail of a gamma-shaped distribution. Those deviation inequalities can be completed by concentration inequalities that show that the excess empirical risk $P_n\widehat{h}_n$ is sharply concentrated around its mean. Deviation inequalities provide information on the probability that a random variable deviates from its median or expectation by a large amount, they may be loose concerning the fluctuations of the random variable around its expectation. Concentration inequalities attempt to provide meaningful information on the probability of small as well as large fluctuations. The results of this section readily follow from the following moment inequalities [10]. Henceforth for $q > 0$, for any random variable X, $\|X\|_q^q = \mathbb{E}[|X|^q]$.

Theorem 3 *Let Z be a measurable function F of n independent random variables $X_1, \ldots, X_n$, let $X'_1, \ldots, X'_n$, be independent copies of $X_1, \ldots, X_n$.*

Let $Z'_i = F(X_1, \ldots, X_{i-1}, X'_i, X_{i+1}, \ldots, X_n)$. Let $V_+ = \sum_{i=1}^n \mathbb{E}' \left[(Z - Z'_i)_+^2 \right]$ where $\mathbb{E}'$ means conditional expectation conditioned on $X_1, \ldots, X_n$. Then for any $q \geq 2$:

$$\| (Z - \mathbb{E}[Z])_+ \|_q \leq \sqrt{3q \, \|V_+\|_{q/2}} = \sqrt{3q} \, \left\| \sqrt{V_+} \right\|_q.$$

Let M be a random variable satisfying $(Z - Z'_i)_+ \leq M$ for all $i \leq n$, then for all $q \geq 2$

$$\| (Z - \mathbb{E}[Z])_- \|_q \leq \sqrt{5q} \left(\left\| \sqrt{V_+} \right\|_q \vee \|M\|_q \right).$$

In the sequel, just as in statement of Theorems 1 and 2, let $\mathcal{H}$ denote the relative loss class associated with a learning task. From the preceding section, we know that if Z is the excess empirical risk, then the corresponding V_+ is upper-bounded by $2n(P_n \widehat{h}_n^2 + P \widehat{h}_n^2)$. Combining this observation with the theorem above immediately leads to the following proposition.

Proposition 2 *Let $\mathcal{H}$ denote a separable class of functions. Let $Z = n \sup_{h \in \mathcal{H}} P_n h$ denote the unnormalized supremum of the empirical risk indexed by $\mathcal{H}$. Assume that the supremum is achieved at some $\widehat{h}_n \in \mathcal{H}$, that is $Z = n P_n \widehat{h}_n$. Then for $q \geq 2$*

$$\| (Z - \mathbb{E}[Z])_+ \|_q \leq \sqrt{6nq} \left(\sqrt{\left\| P_n \widehat{h}_n^2 \right\|_{q/2}} + \sqrt{\left\| P \widehat{h}_n^2 \right\|_{q/2}} \right).$$

Proof Starting from

$$V_+ \leq 2n \left(P_n \widehat{h}_n^2 + P \widehat{h}_n^2 \right),$$

by Theorem 3, for $q \geq 2$:

$$
\begin{aligned}
&\| (Z - \mathbb{E}[Z])_+ \|_q \\
&\leq \sqrt{3q} \left\| \sqrt{2n \left(P_n \widehat{h}_n^2 + P \widehat{h}_n^2 \right)} \right\|_q \\
&\leq \sqrt{6nq} \left(\sqrt{\left\| P_n \widehat{h}_n^2 \right\|_{q/2}} + \sqrt{\left\| P \widehat{h}_n^2 \right\|_{q/2}} \right).
\end{aligned}
$$

$\square$

Note that the Talagrand Inequality for suprema of empirical processes [42] does provide us with a Bernstein-like inequality for the empirical excess risk $Z = n P_n \widehat{h}_n$: for some universal constant κ,

$$\| Z - \mathbb{E}[Z] \|_q \leq \kappa \left(\sqrt{q \mathbb{E} \left[\sup_{h \in \mathcal{H}} n P_n h^2 \right]} + q \right).$$

But, this inequality does not meet our requirements. In many situations, the variance upper-bounds it relies on, far exceed the upper-bounds described in the preceding section. Those bounds grow at least as fast as $n \sup_{h \in \mathcal{H}} Ph^2$, that is linearly with sample size and cannot be easily related with model complexity that is with a sensible notion of degrees of freedom.

In the sequel, we assume that the learning task satisfies the complexity Assumption (1) with function ψ and the noise conditions Assumption (2) with function ω. Let r_* denote the positive solution of

$$\sqrt{n} r^2 = \psi(\omega(r)).$$

Recall that h^* denotes the relative loss function associated with the minimizer of the risk over all measurable functions, so Ph^* denotes the model bias. In the sequel, we will use the fact that with high probability, $P_n \widehat{h}_n^2$ is not much larger than $P\widehat{h}_n^2$. This is embodied in the following theorem whose proof can be found in the appendix.

Theorem 4 *Let $\mathcal{H}$ denote the loss class associated with a learning task that satisfies Assumption (1) about complexity with function ψ and Assumption (2) about noise conditions with function ω. Let r_* denote the positive solution of $\sqrt{n} r^2 = \psi(\omega(r))$. Let $\widehat{h}_n$ denote the relative loss function associated with the empirical risk minimizer. For $q \geq 2$*

$$\left\| P\widehat{h}_n^2 \right\|_q \vee \left\| P_n \widehat{h}_n^2 \right\|_q$$
$$\leq \kappa \left(\omega^2 \left(\sqrt{Ph^*} \right) + \omega^2 (r_*) + \omega^2 \left(\frac{\omega(r_*)}{\sqrt{n} r_*} \right) q \right),$$

where h^* denotes the relative loss function associated with the minimizer of the risk over all measurable functions.

The universal constant κ does not depend on the learning task.

Combining this theorem, the moment inequalities stated above and Proposition 2, leads to the main result:

Theorem 5 *Let $\mathcal{H}$ denote the loss class associated with a learning task that satisfies the complexity Assumption (1) with function ψ and the noise conditions Assumption (2) with function ω. Let r_* denote the positive solution of $\sqrt{n} r^2 = \psi(\omega(r))$. Let $\widehat{h}_n$ (resp. h^*) denote the relative loss function associated with the empirical risk minimizer (resp. the global minimizer of the risk). Let $Z = n P_n \widehat{h}_n$ denote the unnormalized empirical excess risk. There exists a universal constant κ such that for $q \geq 2$,*

$$\| Z - \mathbb{E}[Z] \|_q$$
$$\leq \kappa \left(\sqrt{n} \left(\omega \left(\sqrt{Ph^*} \right) + \omega(r_*) \right) q^{1/2} + \sqrt{n} \, \omega \left(\frac{\omega(r_*)}{\sqrt{n} r_*} \right) q \right).$$

Remark 5 Comparing this inequality with the moment bounds that follow from the naive application of Talagrand's inequality shows that we are likely to obtain a significant improvement on the variance term which describes the sub-Gaussian part of the

tail. But we may feel less confident with respect to the second term on the right-hand side. The latter describes the sub-exponential part of the tail. We will see in the next section that in some favorable situations $\sqrt{n}\,\omega\left(\omega(r_*)/\left(\sqrt{n}r_*\right)\right)$ can be upper-bounded by an expression that depends neither on sample size nor on model complexity but on noise conditions.

Proof The qth norm of $P_n\widehat{h}_n^2$ is upper-bounded due to Theorem 4. Now plugging the bounds from Theorem 4 into the last inequality leads to:

$$
\begin{aligned}
&\|(Z-\mathbb{E}[Z])_+\|_q \\
&\quad \leq 2\sqrt{6n\kappa}\left(\omega\left(\sqrt{Ph^*}\right)+\omega(r_*)\right)q^{1/2}+2\sqrt{3n\kappa}\,\omega\left(\frac{\omega(r_*)}{\sqrt{n}r_*}\right)q.
\end{aligned}
$$

The proof of the theorem is completed by observing that

$$
\|(Z-\mathbb{E}[Z])\|_q \leq \|(Z-\mathbb{E}[Z])_+\|_q + \|(Z-\mathbb{E}[Z])_-\|_q
$$

and that the random variable denoted by M in the statement of Theorem 3 is upper-bounded by 1.   □

We may turn again to the bounded regression problem which served as a benchmark for variance bounds at the end of Sect. 3. From Theorem 5, it follows that for a fixed noise level σ, there exists a universal constant κ such that the empirical excess risk satisfies:

$$
\|Z-\mathbb{E}Z\|_q \leq \kappa\left(\sqrt{Dq}+q\right)
$$

for all sample sizes n and all partition cardinalities D. On the other hand, under mild conditions on the relationship between the number of classes and the sample size, $\mathbb{E}Z/\sigma^2 \approx D$, so in this setting again, the empirical excess risk exemplifies the high dimensional Wilks phenomenon.

5 Application: classification with Vapnik–Chervonenkis classes

In this section, we plug the above-described moment estimates into the various bounds described by Massart and Nedelec [33]. Details concerning Vapnik–Chervonenkis classes can be found in [5,47,48].

Proposition 3 *There exists a universal constant κ such that the following holds. Let $\mathcal{F}$ denote a* VC-*class of sets with* VC *dimension V. Assume the following noise conditions: the Bayes classifier f^* belongs to $\mathcal{F}$, and conditionally on X, with probability*

$(1 + \beta)/2$, *the label Y equals* $f^*(X)$. *The excess empirical risk satisfies*

$$\mathbb{E}\left[n(R_n(\bar{f}) - R_n(\widehat{f}))\right] \leq \kappa\left[\left(\frac{V\left(1 + \log((n\beta^2/V) \vee 1)\right)}{\beta}\right) \wedge \sqrt{nV}\right],$$

$$Var\left[n\left(R_n(\bar{f}) - R_n\left(\widehat{f}\right)\right)\right] \leq \kappa\left[\left(\frac{V(1 + \log\left((n\beta^2/V) \vee 1\right))}{\beta^2}\right) \wedge \sqrt{\frac{nV}{\beta^2}}\right]$$

and for $q \geq 2$

$$\left\|\left(n\left(R_n(\bar{f}) - R_n\left(\widehat{f}\right)\right) - \mathbb{E}\left[n(R_n(\bar{f}) - R_n(\widehat{f}))\right]\right)_+\right\|_q$$

$$\leq \sqrt{\kappa\left(\frac{V(1 + \log((n\beta^2/V) \vee 1))}{\beta^2}\right) \wedge \sqrt{\frac{nV}{\beta^2}}}q^{1/2} + \frac{\kappa}{\beta}q.$$

Remark 6 These noise conditions have sometimes been called random classification noise in early Machine Learning papers (see [3,7,23]).

Remark 7 For $n \geq V/\beta^2$, the upper-bounds stated in the proposition coincides with the one obtained when deriving tail bounds for a Gamma distribution with shape parameter $V(1 + \log((n\beta^2/V))$ and scale parameter $1/\beta$. Up to the slowly varying function $\log((n\beta^2/V))$, this is in-line with what we meant using the expression "Wilks phenomenon".

The proof consists in putting together the generic moment bounds described in this paper and the detailed computations carried out in [33].

Proof In the random classification noise setting, for any classifier f

$$R(f) - R(f^*) = \beta\mathbb{E}\left[\mathbb{1}_{f \neq f^*}\right] = \beta\mathbb{E}\left[(f(X) - f^*(X))^2\right].$$

Hence the function ω may be chosen as $\omega(r) = 1 \vee r/\sqrt{\beta}$. On the other hand, computations carried out in [33] show that ψ may be chosen as

$$\psi(r) = Kr\sqrt{V(1 + \log\left(r^{-1} \vee 1\right))},$$

where K is some universal constant that depends neither on the VC-class under investigation nor on the sampling probability distribution. The quantity denoted by r_* is the solution of

$$\sqrt{n}r^2 = K\frac{r}{\sqrt{\beta}}\sqrt{V\left((1 + \log\left(\frac{\sqrt{\beta}}{r} \vee 1\right)\right)}.$$

A few lines of algebra (see [33]) lead to

$$r_*^2 \leq K^2 \left[\left(\frac{V(1 + \log((n\beta^2/V) \vee 1))}{n\beta} \right) \wedge \sqrt{\frac{V}{n}} \right]$$

so that Theorem 1 and Proposition 2 lead to the first two inequalities.

Note that $\frac{\omega(r_*)}{\sqrt{nr_*}} = \frac{1}{\sqrt{n\beta}}$. Hence the bounds on the qth moment of the excess empirical risk translate into the last inequality. □

Appendix: Tools from empirical process theory

This appendix provides a sketch of proof of Theorems 1 and 4. The proof of these two results depends on the following concentration inequality for suprema of empirical processes. It is due to [42, Theorem 1.4]. The sharp constants in the last inequality are due to [12] building on previous work by several authors [25,30,38].

Theorem 6 *There exists a number κ with the following property. Consider n independent random variables $U_1, \ldots, U_n$ valued in some measurable space Ω. Consider a countable class $\mathcal{H}$ of measurable functions on Ω. Consider the random variable $Z = \sup_{h \in \mathcal{H}} \sum_{i \leq n} h(U_i)$. Consider $u = \sup_{h \in \mathcal{H}, x \in \Omega} |h(x)|$ and $v = \mathbb{E} \sup_{h \in \mathcal{H}} \sum_{i=1}^n h(X_i)^2$. Then for each $t > 0$*

$$\Pr\{|Z - \mathbb{E}Z| > t\} \leq K \exp\left(-\frac{1}{K} \frac{t}{u} \log\left(1 + \frac{t}{uv}\right) \right).$$

Ih $Ph = 0$ for all $h \in \mathcal{H}$ and $u = 1$, then letting $\sigma^2 = \sup_{h \in \mathcal{H}} Ph^2$ and $v = 2\mathbb{E}[Z] + n\sigma^2$, for all $\delta > 0$,

$$\mathbb{P}\left\{ Z \geq \mathbb{E}[Z] + \sqrt{2v \log \frac{1}{\delta}} + \frac{1}{3} \log \frac{1}{\delta} \right\} \leq \delta.$$

The development of tight bounds for excess risk and empirical excess risk relies on a good understanding of the stochastic behavior of the increments of the empirical process indexed by $\mathcal{F}$. Indeed, the sum of the excess risk and the empirical excess risk (which are both positive) coincides with the increment of the centered empirical process defined by the loss functions between $\bar{f}$ and $\widehat{f}$:

$$\begin{aligned} R(\widehat{f}) - R(\bar{f}) + R_n(\bar{f}) - R_n(\widehat{f}) &= (P - P_n)\left(\ell\left(\widehat{f}(X), Y\right) - \ell\left(\bar{f}(X), Y\right) \right) \\ &= (P_n - P)\widehat{h}. \end{aligned}$$

If we are given a (random) function ϕ_n such that for all classifiers f in $\mathcal{F}$,

$$\left|(P - P_n)\left(\ell(f(X), Y) - \ell(\bar{f}(X), Y)\right)\right| \leq \phi_n\left(R(f)\right),$$

a sharp upper-bound on $R(\widehat{f}) - R(\bar{f})$ is obtained by solving the inequality $R(f) - R(\bar{f}) \leq \phi_n(R(f))$. The analysis of the modulus of oscillation of empirical processes indexed by general classes of sets goes back to the work of [2] (see also [45] for applications to M-estimation), and the impact of recently obtained concentration inequalities on this topic is thoroughly investigated by Giné et al. [19] and Giné and Koltchinskii [20]. However, the analysis of the excess risk does not require an exhaustive understanding of the modulus of continuity but a tight control of this modulus over a special range of values. Hence, rather than resorting to omnibus results like those presented in the aforementioned references, we rely on ad hoc analysis which has already proved convenient and successful [31–33].

The following lemma is borrowed from [33].

Lemma 1 *Let $\mathcal{H}$ be some countable index set. Let L denote a non-negative function on $\mathcal{H}$, and assume there exists some $\bar{h} \in \mathcal{H}$ such that $L(\bar{h}) = \inf_{h \in \mathcal{H}} L(h)$.*

Let Z denote a stochastic process indexed by $\mathcal{H}$. Assume that there exists a function ψ from the class $\mathcal{C}_1$ of positive sub-linear functions such that for all r satisfying $r^2 \geq \psi(r)$:

$$\mathbb{E}\left[\sup_{h:h\in\mathcal{H}, L(h)\leq r^2} \left|Z(h) - Z(\bar{h})\right|\right] \leq \psi(r).$$

Then, for any r satisfying $r^2 \geq \psi(r)$:

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{r^2}{r^2 + L(h)} \left|Z(h) - Z(\bar{h})\right|\right] \leq 4\psi(r).$$

The next theorem is the key technical ingredient in the derivation of tail bounds for excess risk and excess empirical risk (Theorem 1) as well as for the $L_2(P)$ and $L_2(P_n)$ distances of relative loss $\widehat{h}_n$ to 0 (Theorem 4).

Theorem 7 *Let $\mathcal{H}$ denote a countable class of measurable functions over $\mathcal{X}$. Let L denote a non-negative function on $\mathcal{H}$, which achieves its minimum at $0 \in \mathcal{H}$. Assume furthermore that $\sup_{x\in\mathcal{X}} |h(x)| \leq 1$.*

Let σ be some non-negative mapping on $\mathcal{H}$ such that for every $h \in \mathcal{H}$, $Ph^2 \leq \sigma^2(h)$. Assume furthermore that there exists some function ρ from $\mathcal{C}_1$ such that for any $h \in \mathcal{H} : \sigma(h) \leq \rho\left(\sqrt{L(h)}\right)$. Assume that there exists some function ψ from $\mathcal{C}_1$ such that for r satisfying $\psi(r) \leq \sqrt{n}r^2$

$$\sqrt{n}\mathbb{E}\left[\sup_{h\in\mathcal{H}, \sigma(h)\leq r} |(P_n - P)h|\right] \leq \psi(r).$$

Let K denote a number larger than 1. *Let* δ *and* ϵ *be positive numbers. Assume that the positive solution* $r(\delta)$ *of equation*

$$r^2 = K(1 + \epsilon)\left(4\frac{\psi(\rho(r))}{\sqrt{n}} + \rho(r)\sqrt{\frac{\log\frac{1}{\delta}}{n}} + \frac{1}{\epsilon}\frac{\log\frac{1}{\delta}}{n}\right),$$

satisfies $\psi(r) \leq \sqrt{n}r^2$. *Then, with probability larger than* $1 - 2\delta$, *for all* $h \in \mathcal{H}$

$$|(P - P_n)h| \leq \frac{L(h) + r^2(\delta)}{K}.$$

Proof We prove that with probability larger than $1 - \delta$, for all $h \in \mathcal{H}$

$$(P_n - P)h \leq \frac{L(h) + r^2(\delta)}{K}.$$

Let r satisfy $\psi(r) \leq \sqrt{n}r^2$, and let the random variable V_r be defined as

$$V_r = \sup_{h \in \mathcal{H}} r^2 \frac{(P_n - P)h}{L(h) + r^2}.$$

Then V_r is the supremum of a centered empirical process indexed by $\mathcal{H}$. Moreover

$$\mathbb{E}\left[\sup_{h, L(h) \leq r^2} |(P_n - P)h|\right] \leq \mathbb{E}\left[\sup_{\sigma(h) \leq \rho(r)} |(P_n - P)h|\right] \leq \frac{\psi(\rho(r))}{\sqrt{n}}.$$

As $\psi \circ \rho \in \mathcal{C}_1$, by Lemma 1, we have $\mathbb{E}[V_r] \leq 4\frac{\psi(\rho(r))}{\sqrt{n}}$. Now, note that if $L(h) \leq r^2$,

$$\text{Var}\left[\frac{r^2 h}{L(h) + r^2}\right] \leq Ph^2 \leq \rho^2(r).$$

Moreover, for each $h \in \mathcal{F}$,

$$\text{Var}\left[\frac{r^2 h}{L(h) + r^2}\right] \leq \left(\frac{r^2\rho\left(\sqrt{L(h)}\right)}{L(h) + r^2}\right)^2 \leq \left(\frac{r^2\rho\left(\sqrt{L(h)}\right)}{2\sqrt{L(h)}r}\right)^2 \leq \left(\frac{r\rho\left(\sqrt{L(h)}\right)}{2\sqrt{L(h)}}\right)^2.$$

Thus if $L(h) \geq r^2$, as ρ is assumed to belong to the class $\mathcal{C}_1$ (of positive sub-linear functions),

$$\text{Var}\left[\frac{r^2 h}{L(h) + r^2}\right] \leq \frac{\rho^2(r)}{4}.$$

On the other hand, for all $h \in \mathcal{H}$

$$\sup_{x \in \mathcal{X}} \left| \frac{r^2 h(x)}{L(h) + r^2} \right| \leq 1.$$

We may now invoke Talagrand's inequality. With probability larger than $1 - \delta$ for all $h \in \mathcal{H}$

$$V_r \leq \mathbb{E}[V_r] + \sqrt{\frac{2}{n} \left(2\mathbb{E}[V_r] + \rho^2(r) \right) \log \frac{1}{\delta}} + \frac{1}{3n} \log \frac{1}{\delta}$$

$$\leq (1 + \epsilon) \mathbb{E}[V_r] + \rho(r) \sqrt{\frac{\log \frac{1}{\delta}}{n}} + \left(\frac{1}{3} + \frac{1}{\epsilon} \right) \frac{\log \frac{1}{\delta}}{n}$$

$$\leq 4(1 + \epsilon) \frac{\psi(\rho(r))}{\sqrt{n}} + \rho(r) \sqrt{\frac{\log \frac{1}{\delta}}{n}} + \left(\frac{1}{3} + \frac{1}{\epsilon} \right) \frac{\log \frac{1}{\delta}}{n}.$$

Let $r(\delta)$ be defined as in the statement of the Theorem. With probability larger than $1 - \delta$, for all $h \in \mathcal{H}$:

$$(P_n - P)h \leq \frac{1}{K} \left(L(h) + r^2(\delta) \right).$$

The same line of reasoning can be applied to the process $((P - P_n)h)_{h \in \mathcal{H}}$. This leads to the theorem. $\square$

The following corollary reveals that if the neighborhood of 0 in $\mathcal{H}$ is not too rich, the empirical $L_2(P_n)$ structure of $\mathcal{H}$ around 0 faithfully approximates the $L_2(P)$ structure of $\mathcal{H}$.

Corollary 1 *Let $\mathcal{H}$ denote a countable class of measurable functions over $\mathcal{X}$. And assume that $\sup_{x \in \mathcal{X}} |h(x)| \leq 1$. Assume that there exists some function $\phi \in \mathcal{C}_1$ such that for all r such that $\sqrt{n} r^2 \geq \phi(r)$:*

$$\sqrt{n} \mathbb{E} \left[\sup_{h \in \mathcal{H}, P h^2 \leq r^2} |(P_n - P)h| \right] \leq \phi(r).$$

Let K be larger than 1, let ϵ be a positive number. Assume that be the positive solution $r(\delta)$ of equation

$$r^2 = K(1 + \epsilon) \left(\frac{32\phi(r)}{\sqrt{n}} + r \sqrt{\frac{\log \frac{1}{\delta}}{n}} + \frac{1}{\epsilon} \frac{\log \frac{1}{\delta}}{n} \right)$$

satisfies $\sqrt{n} r^2(\delta) \geq \phi(r(\delta))$. Then, with probability larger than $1 - 2\delta$, for all $h \in \mathcal{H}$

$$-\frac{1}{K} \left(P h^2 + r^2(\delta) \right) \leq (P_n - P)h^2 \leq \frac{1}{K} \left(P h^2 + r^2(\delta) \right).$$

Proof The proof of Corollary 1 consists in applying Theorem 7 to the class of functions $\left\{h^2 : h \in \mathcal{H}\right\}$. Note first that

$$\sup_{h \in \mathcal{H}, x \in \mathcal{X}} \left| h^2(x) - Ph^2 \right| \leq 1.$$

Second, choosing $L(h) = \sigma^2(h) = Ph^2$, L is minimized by 0, and letting $\rho = \mathrm{Id}$,

$$P\left(h^2\right)^2 \leq \sigma^2(h) \leq \rho^2\left(\sqrt{L(h)}\right).$$

The only point that needs to be checked is that

$$\sqrt{n}\mathbb{E}\left[\sup_{h:\,\sigma(h)\leq r} \left|(P_n - P)h^2\right|\right] \leq 8\phi(r).$$

Let now $\epsilon_1, \ldots, \epsilon_n$ be a sequence of independent symmetric, $\{-1, 1\}$-valued random variables, the so-called Rademacher variables. By a standard symmetrization argument:

$$\mathbb{E}\left[\sup_{h:\,\sigma(h)\leq r} \left|(P_n - P)h^2\right|\right] \leq \frac{2}{n}\mathbb{E}\left[\mathbb{E}\left[\sup_{h:\,\sigma(h)\leq r} \left|\sum_{i=1}^n \epsilon_i h^2(X_i)\right| \mid X_1, \ldots, X_n\right]\right]$$

where the internal expectation is taken over the Rademacher variables.

Now as $x^\iota \rightarrow x^2$ is 2-Lipschitz over $[-1, 1]$, the contraction principle [13,27] implies that

$$\frac{1}{n}\mathbb{E}\left[\sup_{h:\,\sigma(h)\leq r} \left|\sum_{i=1}^n \epsilon_i h^2(X_i)\right| \mid X_1, \ldots, X_n\right]$$

$$\leq \frac{2}{n}\mathbb{E}\left[\sup_{h:\,\sigma(h)\leq r} \left|\sum_{i=1}^n \epsilon_i h(X_i)\right| \mid X_1, \ldots, X_n\right].$$

Another standard symmetrization argument entails

$$\mathbb{E}\left[\sup_{h:\,\sigma(h)\leq r} \left|(P_n - P)h^2\right|\right] \leq 8\mathbb{E}\left[\sup_{h:\,\sigma(h)\leq r} |(P - P_n)h|\right] \leq 8\phi(r).$$

We may now invoke Theorem 7 on the class $\left\{h^2 : h \in \mathcal{H}\right\}$, with $L(h) = \sigma^2(h) = Ph^2$, $\rho = \mathrm{Id}$ and $\psi = 8\phi$. □

The following elementary proposition allows us to translate stochastic bounds in the style of [24] into bounds stated in the style of [33]. It provides the ultimate ingredient in the proof of Theorems 1 and 4.

Proposition 4 *Let ψ and ρ denote non-constant, functions from $\mathcal{C}_1$. Let r_* denote the unique positive solution of equation $\sqrt{n}r^2 = \psi(\rho(r))$. For some a, b, c in $\mathbb{R}_+$, with $a \geq 1$, let s denote the unique solution of equation $r^2 = \frac{a}{\sqrt{n}}\psi(\rho(r)) + \frac{b}{\sqrt{n}}\rho(r) + c$. Then $s^2 \leq 2a^2 r_*^2 + 2b^2 \frac{\rho^2(r_*)}{nr_*^2} + 2c$.*

Proof (Theorem 1) The proof consists of invoking Theorem 7 for the relative loss class $\mathcal{H}$ as defined earlier. From the definitions of $\mathcal{H}$ and $L(h) = P(h^* - h)$, L is minimized by 0. Moreover, the function σ and ρ can be chosen as $\sigma(h) = d(f, f^*)$ and $\rho = 2\omega$. Indeed,

$$\begin{aligned}
Ph^2 &\leq 2\left(P(h^* - h)^2 + P(h^*)^2\right) \\
&\leq 2\left(d^2(f, f^*) + d^2(\bar{f}, f^*)\right) \\
&\leq 2\left(\omega^2\left(\sqrt{P(h^* - h)}\right) + \omega^2\left(\sqrt{Ph^*}\right)\right) \\
&\leq 4\omega^2\left(\sqrt{P(h^* - h)}\right) = 4\omega^2\left(\sqrt{L(h)}\right).
\end{aligned}$$

We may now invoke Theorem 7 to check that, with probability larger than $1 - 2\delta$, for all $h \in \mathcal{H}$:

$$(P - P_n)h \leq \frac{1}{K}\left(L(h) + r^2(\delta)\right).$$

Going back to classifiers and risks, this translates into

$$\frac{K-1}{K}(-Ph) + P_n h \leq \frac{1}{K}\left(P(h^* - h) + r^2(\delta)\right) = \frac{1}{K}\left(Ph^* + r^2(\delta)\right).$$

Specializing the preceding bound to $\widehat{h}_n$, we can take advantage of the fact that both $-P\widehat{h}_n$ and $P_n\widehat{h}_n$ are both non-negative. With probability larger $1 - 2\delta$, the excess risk is upper-bounded by

$$-P\widehat{h}_n \leq \frac{1}{K-1}\left(Ph^* + r^2(\delta)\right),$$

while the empirical excess risk is upper-bounded by

$$P_n\widehat{h}_n \leq \frac{1}{K}\left(Ph^* + r^2(\delta)\right).$$

The proof can be completed by invoking Proposition 4 and integration by parts. □

Proof (Theorem 4) From the proof of Theorem 1, it follows that with probability larger than $1 - 2\delta$

$$L\left(\widehat{h}_n\right) \leq \frac{KL(0)}{K-1} + \frac{r^2(\delta)}{K-1}.$$

Then, from the relation

$$P\widehat{h}_n^2 \le 4\omega^2\left(\sqrt{L\left(\widehat{h}_n\right)}\right).$$

and thanks to the fact that $\omega^2\left(\sqrt{\cdot}\right) \in \mathcal{C}_1$, with probability $1 - 2\delta$

$$\frac{1}{4}P\widehat{h}_n^2 \le \frac{K}{K-1}\left(\omega^2\left(\sqrt{Ph^*}\right) + \omega^2\left(\sqrt{\frac{r^2(\delta)}{K}}\right)\right).$$

Now, invoking the relation between $r^2(\delta)$ and r_*^2 (Proposition 4), this last inequality translates into

$$\frac{1}{4}P\widehat{h}_n^2 \le \frac{K}{K-1}\left(\omega^2\left(\sqrt{Ph^*}\right) + 2Ka^2\omega^2(r_*) + 2K\omega^2\left(\frac{\omega(r_*)}{\sqrt{n}r_*}\right)\log\frac{1}{\delta}\right). \quad (8)$$

The inequality concerning $P\widehat{h}_n^2$ follows by integration by parts.

The inequality concerning $P_n\widehat{h}^2$ follows by combining the inequality concerning $P\widehat{h}_n^2$ and the stochastic relation between the $L_2(P)$ and the $L_2(P_n)$ structure of the loss class $\mathcal{H}$ (Corollary 1). Note that if we keep using K and δ, and call $s(\delta)$ the solution of

$$r^2 = K\left(8a\frac{\phi(r)}{\sqrt{n}} + r\sqrt{\frac{\log\frac{1}{\delta}}{n}} + \frac{c\log\frac{1}{\delta}}{n}\right),$$

then $s(\delta) \le r(\delta)$ since $2\omega(r) \ge r$ on $[0, 1]$. This entails that with probability $1 - 4\delta$, we have both (8) and

$$P_n\widehat{h}_n^2 \le \frac{K+1}{K-1}P\widehat{h}_n^2 + \frac{r^2(\delta)}{K}.$$

The Theorem follows by observing that $r(\delta) \ge r_*$ and $\omega(r) \ge r$ on $[0, 1]$, and using Proposition 4. $\square$

References

1. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Automat. Contr. **19**(6), 716–723 (1974)
2. Alexander, K.: Rates of growth and sample moduli for weighted empirical processes indexed by sets. Probab. Theory Relat. Fields **75**, 379–423 (1987)
3. Angluin, D., Laird, P.: Learning from noisy examples. Mach. Learn. **2**(4), 343–370 (1987)
4. Arlot, S., Massart, P.: Data-driven calibration of penalties for least-squares regression. J. Mach. Learn. Res. **10**, 245–279 (2009)
5. Assouad, P.: Densité et dimension. Ann. Inst. Fourier (Grenoble) **33**(3), 233–282 (1983)
6. Bartlett, P., Mendelson, S.: Empirical minimization. Probab. Theory Relat. Fields **135**(3), 311–334 (2006)

7. Bartlett, P., Boucheron, S., Lugosi, G.: Model selection and error estimation. Mach. Learn. **48**, 85–113 (2002)
8. Bickel, P., Doksum, K.: Mathematical Statistics. Holden-Day Inc., San Francisco (1976)
9. Boucheron, S., Bousquet, O., Lugosi, G.: Theory of classification: some recent advances. ESAIM Probab. Stat. **9**, 329–375 (2005)
10. Boucheron, S., Bousquet, O., Lugosi, G., Massart, P.: Moment inequalities for functions of independent random variables. Ann. Probab. **33**(2), 514–560 (2005)
11. Bousquet, O.: Concentration inequalities for sub-additive functions using the entropy method. In: Stochastic Inequalities and Applications. Progress in Probability, vol. 56, pp. 213–247. Birkhäuser, Basel (2003)
12. Bousquet, O.: A Bennett concentration inequality and its application to suprema of empirical processes. C. R. Math. Acad. Sci. Paris **334**(6), 495–500 (2002)
13. de la Pena, V., Giné, E.: Decoupling. Springer, Berlin (1999)
14. Devroye, L., Wagner, T.: Distribution-free inequalities for the deleted and holdout error estimates. IEEE Trans. Inform. Theory **25**, 202–207 (1977)
15. Efron, B., Stein, C.: The jackknife estimate of variance. Ann. Stat. **9**(3), 586–596 (1981)
16. Fan, J.: Local linear regression smoothers and their minimax efficiency. Ann. Stat. **21**, 196–216 (1993)
17. Fan, J., Zhang, C., Zhang, J.: Generalized likelihood ratio statistics and Wilks phenomenon. Ann. Stat. **29**(1), 153–193 (2001)
18. Gayraud, G., Pouet, C.: Minimax testing composite null hypotheses in the discrete regression scheme. Math. Methods Stat. **10**(4), 375–394 (2001)
19. Giné, E., Koltchinskii, V.: Concentration inequalities and asymptotic results for ratio type empirical processes. Ann. Probab. **34**(3), 1143–1216 (2006)
20. Giné, E., Koltchinskii, V., Wellner, J.: Stochastic inequalities and applications. In: Ratio Limit Theorems for Empirical Processes, pp. 249–278. Birkhaüser, Basel (2003)
21. Huber, P.: The behavior of the maximum likelihood estimates under non-standard conditions. In: Proceedings of Fifth Berkeley Symposium on Probability and Mathematical Statistics, pp. 221–233. University of California Press, Berkeley (1967)
22. Kearns, M., Ron, D.: Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. Neural Comput. **11**(6), 1427–1453 (1999)
23. Kearns, M., Mansour, Y., Ng, A., Ron, D.: An experimental and theoretical comparison of model selection methods. Mach. Learn. **27**, 7–50 (1997)
24. Koltchinskii, V.: Localized rademacher complexities and oracle inequalities in risk minimization. Ann. Stat. **34**, 2593–2656 (2006)
25. Ledoux, M.: On Talagrand's deviation inequalities for product measures. ESAIM Probab. Stat. **1**, 63–87 (1995/1997)
26. Ledoux, M.: The concentration of measure phenomenon. American Mathematical Society, Providence (2001)
27. Ledoux, M., Talagrand, M.: Probability in Banach spaces. Springer, Berlin (1991)
28. Mallows, C.: Some comments on C_p. Technometrics **15**(4), 661–675 (1973)
29. Mammen, E., Tsybakov, A.: Smooth discrimination analysis. Ann. Stat. **27**(6), 1808–1829 (1999)
30. Massart, P.: About the constants in Talagrand's concentration inequality. Ann. Probab. **28**, 863–885 (2000)
31. Massart, P.: Some applications of concentration inequalities to statistics. Ann. Fac. Sci. Toulouse **IX**(2), 245–303 (2000)
32. Massart, P.: Concentration inequalities and model selection. Ecole d'Eté de Probabilité de Saint-Flour xxxiv. In: Lecture Notes in Mathematics, vol. 1896. Springer, Berlin (2007)
33. Massart, P., Nedelec, E.: Risk bounds for classification. Ann. Stat. **34**(5), 2326–2366 (2006)
34. Pollard, D.: Convergence of Stochastic Processes. Springer, Berlin (1984)
35. Portnoy, S.: Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. Ann. Stat. **16**, 356–366 (1988)
36. Quenouille, M.: Approximate test of correlation in time series. J. R. Stat. Soc. Ser. B **11**, 68–84 (1949)
37. Rakhlin, A., Mukherjee, S., Poggio, T.: Stability results in learning theory. Anal. Appl. (Singapore) **3**, 397–417 (2005)
38. Rio, E.: Inégalités de concentration pour les processus empiriques de classes de parties. Probab. Theory Relat. Fields **119**, 163–175 (2001)
39. Schoelkopf, B., Smola, A.: Learning with Kernels. MIT Press, Cambridge (2002)

40. Shorack, G.R., Wellner, J.A.: Empirical Processes with Applications to Statistics. Wiley, New York (1986)
41. Talagrand, M.: A new look at independence. Ann. Probab. **24**, 1–34 (1996)
42. Talagrand, M.: New concentration inequalities in product spaces. Invent. Math. **126**, 505–563 (1996)
43. Tsybakov, A.B.: Optimal aggregation of classifiers in statistical learning. Ann. Stat. **32**, 135–166 (2004)
44. Tukey, J.: Bias and confidence in not quite large samples. Ann. Math. Stat. **29**, 614 (1958)
45. van de Geer, S.: Applications of Empirical Process Theory. Cambridge University Press, London (2000)
46. van der Vaart, A.: Asymptotic Statistics. Cambridge University Press, London (1998)
47. van der Vaart, A., Wellner, J.: Weak Convergence and Empirical Processes. Springer, Berlin (1996)
48. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, Berlin (1995)
49. Wilks, S.: The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Stat. **9**, 60–62 (1938)