Risk bounds in linear regression through PAC-Bayesian truncation

Jean-Yves Audibert, Olivier Catoni

Risk bounds in linear regression through PAC-Bayesian truncation

JEAN-YVES AUDIBERT[1,2], OLIVIER CATONI [3]

July 3, 2010

ABSTRACT : We consider the problem of predicting as well as the best linear combination of d given functions in least squares regression, and variants of this problem including constraints on the parameters of the linear combination. When the input distribution is known, there already exists an algorithm having an expected excess risk of order d/n, where n is the size of the training data. Without this strong assumption, standard results often contain a multiplicative $\log n$ factor, and require some additional assumptions like uniform boundedness of the d-dimensional input representation and exponential moments of the output.

This work provides new risk bounds for the ridge estimator and the ordinary least squares estimator, and their variants. It also provides shrinkage procedures with convergence rate d/n (i.e., without the logarithmic factor) in expectation and in deviations, under various assumptions. The key common surprising factor of these results is the absence of exponential moment condition on the output distribution while achieving exponential deviations. All risk bounds are obtained through a PAC-Bayesian analysis on truncated differences of losses. Finally, we show that some of these results are not particular to the least squares loss, but can be generalized to similar strongly convex loss functions.

2000 MATHEMATICS SUBJECT CLASSIFICATION: 62J05, 62J07.

KEYWORDS: Linear regression, Generalization error, Shrinkage, PAC-Bayesian theorems, Risk bounds, Robust statistics, Resistant estimators, Gibbs posterior distributions, Randomized estimators, Statistical learning theory

CONTENTS

[1]Université Paris-Est, Ecole des Ponts ParisTech, Imagine, 6 avenue Blaise Pascal, 77455 Marne-la-Vallée, France, audibert@imagine.enpc.fr

[2]Willow, CNRS/ENS/INRIA — UMR 8548, 45 rue d'Ulm, F75230 Paris cedex 05, France

[3]Département de Mathématiques et Applications, CNRS – UMR 8553, École Normale Supérieure, 45 rue d'Ulm, F75230 Paris cedex 05, olivier.catoni@ens.fr

INTRODUCTION

OUR STATISTICAL TASK. Let $Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n)$ be $n \geq 2$ pairs of input-output and assume that each pair has been independently drawn from the same unknown distribution P. Let $\mathcal{X}$ denote the input space and let the output space be the set of real numbers $\mathbb{R}$, so that P is a probability distribution on the product space $\mathcal{Z} \triangleq \mathcal{X} \times \mathbb{R}$. The target of learning algorithms is to predict the output Y associated with an input X for pairs $Z = (X, Y)$ drawn from the distribution P. The quality of a (prediction) function $f : \mathcal{X} \to \mathbb{R}$ is measured by the least squares *risk*:

$$R(f) \triangleq \mathbb{E}_{Z\sim P}\left\{[Y - f(X)]^2\right\}.$$

Through the paper, we assume that the output and all the prediction functions we consider are square integrable. Let Θ be a closed convex set of $\mathbb{R}^d$, and $\varphi_1, \ldots, \varphi_d$ be d prediction functions. Consider the regression model

$$\mathcal{F} = \left\{ f_\theta = \sum_{j=1}^{d} \theta_j \varphi_j ; (\theta_1, \ldots, \theta_d) \in \Theta \right\}.$$

The best function f^* in $\mathcal{F}$ is defined by

$$f^* = \sum_{j=1}^{d} \theta_j^* \varphi_j \in \operatorname*{argmin}_{f \in \mathcal{F}} R(f).$$

Such a function always exists but is not necessarily unique. Besides it is unknown since the probability generating the data is unknown.

We will study the problem of predicting (at least) as well as function f^*. In other words, we want to deduce from the observations $Z_1, \ldots, Z_n$ a function $\hat{f}$ having with high probability a risk bounded by the minimal risk $R(f^*)$ on $\mathcal{F}$ plus a small remainder term, which is typically of order d/n up to a possible logarithmic factor. Except in particular settings (e.g., Θ is a simplex and $d \geq \sqrt{n}$), it is known that the convergence rate d/n cannot be improved in a minimax sense (see [20], and [21] for related results).

More formally, the target of the paper is to develop estimators $\hat{f}$ for which the excess risk is controlled *in deviations*, i.e., such that for an appropriate constant $\kappa > 0$, for any $\varepsilon > 0$, with probability at least $1 - \varepsilon$,

$$R(\hat{f}) - R(f^*) \leq \kappa \frac{d + \log(\varepsilon^{-1})}{n}. \tag{0.1}$$

Note that by integrating the deviations (using the identity $\mathbb{E}W = \int_0^{+\infty} \mathbb{P}(W > t)dt$ which holds true for any nonnegative random variable W), Inequality (0.1) implies

$$\mathbb{E}R(\hat{f}) - R(f^*) \leq \kappa \frac{d+1}{n}. \tag{0.2}$$

In this work, we do not assume that the function

$$f^{(\text{reg})} : x \mapsto \mathbb{E}[Y|X = x],$$

which minimizes the risk R among all possible measurable functions, belongs to the model $\mathcal{F}$. So we might have $f^* \neq f^{(\text{reg})}$ and in this case, bounds of the form

$$\mathbb{E}R(\hat{f}) - R(f^{(\text{reg})}) \leq C[R(f^*) - R(f^{(\text{reg})})] + \kappa \frac{d}{n}, \tag{0.3}$$

with a constant C larger than 1 do not even ensure that $\mathbb{E}R(\hat{f})$ tends to $R(f^*)$ when n goes to infinity. This kind of bounds with $C > 1$ have been developed to analyze nonparametric estimators using linear approximation spaces, in which case the dimension d is a function of n chosen so that the bias term $R(f^*) - R(f^{(\text{reg})})$ has the order d/n of the estimation term (see [11] and references within). Here we intend to assess the generalization ability of the estimator even when the

model is misspecified (namely when $R(f^*) > R(f^{(\text{reg})})$). Moreover we do not assume either that $Y - f^{(\text{reg})}(X)$ and X are independent.

Notation. When $\Theta = \mathbb{R}^d$, the function f^* and the space $\mathcal{F}$ will be written f^*_{lin} and $\mathcal{F}_{\text{lin}}$ to emphasize that $\mathcal{F}$ is the whole linear space spanned by $\varphi_1, \ldots, \varphi_d$:

$$\mathcal{F}_{\text{lin}} = \operatorname{span}\{\varphi_1, \ldots, \varphi_d\} \qquad \text{and} \qquad f^*_{\text{lin}} \in \operatorname*{argmin}_{f \in \mathcal{F}_{\text{lin}}} R(f).$$

The Euclidean norm will simply be written as $\|\cdot\|$, and $\langle \cdot, \cdot \rangle$ will be its associated inner product. We will consider the vector valued function $\varphi : \mathcal{X} \to \mathbb{R}^d$ defined by $\varphi(X) = \big[\varphi_k(X)\big]_{k=1}^d$, so that for any $\theta \in \Theta$, we have

$$f_\theta(X) = \langle \theta, \varphi(X) \rangle.$$

The Gram matrix is the $d \times d$-matrix $Q = \mathbb{E}\big[\varphi(X)\varphi(X)^T\big]$, and its smallest and largest eigenvalues will respectively be written as $q_{\min}$ and $q_{\max}$. The empirical risk of a function f is

$$r(f) = \frac{1}{n} \sum_{i=1}^n \big[f(X_i) - Y_i\big]^2$$

and for $\lambda \geq 0$, the ridge regression estimator on $\mathcal{F}$ is defined by $\hat{f}^{(\text{ridge})} = f_{\hat{\theta}^{(\text{ridge})}}$ with

$$\hat{\theta}^{(\text{ridge})} \in \arg\min_{\theta \in \Theta} r(f_\theta) + \lambda \|\theta\|^2,$$

where λ is some nonnegative real parameter. In the case when $\lambda = 0$, the ridge regression $\hat{f}^{(\text{ridge})}$ is nothing but the empirical risk minimizer $\hat{f}^{(\text{erm})}$. In the same way, we introduce the optimal ridge function optimizing the expected ridge risk: $\tilde{f} = f_{\tilde{\theta}}$ with

$$\tilde{\theta} \in \arg\min_{\theta \in \Theta} \big\{R(f_\theta) + \lambda \|\theta\|^2\big\}. \tag{0.4}$$

Finally, let $Q_\lambda = Q + \lambda I$ be the ridge regularization of Q, where I is the identity matrix.

WHY SHOULD WE BE INTERESTED IN THIS TASK. There are three main reasons. First we aim at a better understanding of the parametric linear least squares method (classical textbooks can be misleading on this subject as we will point out later), and intend to provide a non-asymptotic analysis of it.

Secondly, the task is central in nonparametric estimation for linear approximation spaces (piecewise polynomials based on a regular partition, wavelet expansions, trigonometric polynomials…)

Thirdly, it naturally arises in two-stage model selection. Precisely, when facing the data, the statistician has often to choose several models which are likely to

be relevant for the task. These models can be of similar structures (like embedded balls of functional spaces) or on the contrary of very different nature (e.g., based on kernels, splines, wavelets or on parametric approaches). For each of these models, we assume that we have a learning scheme which produces a 'good' prediction function in the sense that it predicts as well as the best function of the model up to some small additive term. Then the question is to decide on how we use or combine/aggregate these schemes. One possible answer is to split the data into two groups, use the first group to train the prediction function associated with each model, and finally use the second group to build a prediction function which is as good as (i) the best of the previously learnt prediction functions, (ii) the best convex combination of these functions or (iii) the best linear combination of these functions. This point of view has been introduced by Nemirovski in [17] and optimal rates of aggregation are given in [20] and references within. This paper focuses more on the linear aggregation task (even if (ii) enters in our setting), assuming implicitly here that the models are given in advance and are beyond our control and that the goal is to combine them appropriately.

OUTLINE AND CONTRIBUTIONS. The paper is organized as follows. Section 1 is a survey on risk bounds in linear least squares. Theorems 1.3 and 1.5 are the results which come closer to our target. Section 2 provides a new analysis of the ridge estimator and the ordinary least squares estimator, and their variants. Theorem 2.1 provides an asymptotic result for the ridge estimator while Theorem 2.2 gives a non asymptotic risk bound of the empirical risk minimizer, which is complementary to the theorems put in the survey section. In particular, the result has the benefit to hold for the ordinary least squares estimator and for heavy-tailed outputs. We show quantitatively that the ridge penalty leads to an implicit reduction of the input space dimension. Section 3 shows a non asymptotic d/n exponential deviation risk bound under weak moment conditions on the output Y and on the d-dimensional input representation $\varphi(X)$. Section 4 presents stronger results under boundedness assumption of $\varphi(X)$. However the latter results are concerned with a not easily computable estimator. Section 5 gives risk bounds for general loss functions from which the results of Section 4 are derived.

The main contribution of this paper is to show through a PAC-Bayesian analysis on truncated differences of losses that the output distribution does not need to have bounded conditional exponential moments in order for the excess risk of appropriate estimators to concentrate exponentially. Our results tend to say that truncation leads to more robust algorithms. Local robustness to contamination is usually invoked to advocate the removal of outliers, claiming that estimators should be made insensitive to small amounts of spurious data. Our work leads to a different theoretical explanation. The observed points having unusually large

outputs when compared with the (empirical) variance should be down-weighted in the estimation of the mean, since they contain less information than noise. In short, huge outputs should be truncated because of their low signal to noise ratio.

1. VARIANTS OF KNOWN RESULTS

1.1. ORDINARY LEAST SQUARES AND EMPIRICAL RISK MINIMIZATION. The ordinary least squares estimator is the most standard method in this case. It minimizes the empirical risk

$$r(f) = \frac{1}{n}\sum_{i=1}^{n}[Y_i - f(X_i)]^2,$$

among functions in $\mathcal{F}_{\text{lin}}$ and produces

$$\hat{f}^{(\text{ols})} = \sum_{j=1}^{d} \hat{\theta}_j^{(\text{ols})}\varphi_j,$$

with $\hat{\theta}^{(\text{ols})} = [\hat{\theta}_j^{(\text{ols})}]_{j=1}^d$ a column vector satisfying

$$\mathbf{X}^T\,\mathbf{X}\,\hat{\theta}^{(\text{ols})} = \mathbf{X}^T\,\mathbf{Y}, \tag{1.1}$$

where $\mathbf{Y} = [Y_j]_{j=1}^n$ and $\mathbf{X} = (\varphi_j(X_i))_{1\le i\le n, 1\le j\le d}$. It is well-known that

- the linear system (1.1) has at least one solution, and in fact, the set of solutions is exactly $\{\mathbf{X}^+\,\mathbf{Y} + u; u \in \ker \mathbf{X}\}$; where $\mathbf{X}^+$ is the Moore-Penrose pseudoinverse of $\mathbf{X}$ and $\ker \mathbf{X}$ is the kernel of the linear operator $\mathbf{X}$.
- $\mathbf{X}\,\hat{\theta}^{(\text{ols})}$ is the (unique) orthogonal projection of the vector $\mathbf{Y} \in \mathbb{R}^n$ on the image of the linear map $\mathbf{X}$;
- if $\sup_{x\in\mathcal{X}} \mathbb{V}\text{ar}(Y|X=x) = \sigma^2 < +\infty$, we have (see [11, Theorem 11.1]) for any $X_1, \dots, X_n$ in $\mathcal{X}$,

$$\mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}^{(\text{ols})}(X_i) - f^{(\text{reg})}(X_i)\right]^2 \middle| X_1, \dots, X_n\right\} - \min_{f\in\mathcal{F}_{\text{lin}}}\frac{1}{n}\sum_{i=1}^{n}\left[f(X_i) - f^{(\text{reg})}(X_i)\right]^2 \le \sigma^2\frac{\text{rank}(\mathbf{X})}{n} \le \sigma^2\frac{d}{n}, \tag{1.2}$$

  where we recall that $f^{(\text{reg})} : x \mapsto \mathbb{E}[Y|X=x]$ is the optimal regression function, and that when this function belongs to $\mathcal{F}_{\text{lin}}$ (i.e., $f^{(\text{reg})} = f^*_{\text{lin}}$), the minimum term in (1.2) vanishes;

- from Pythagoras' theorem for the (semi)norm $W \mapsto \sqrt{\mathbb{E}W^2}$ on the space of the square integrable random variables,

$$R(\hat{f}^{(\text{ols})}) - R(f^*_{\text{lin}}) \\ = \mathbb{E}\big[\hat{f}^{(\text{ols})}(X) - f^{(\text{reg})}(X)\big|Z_1, \ldots, Z_n\big]^2 - \mathbb{E}\big[f^*_{\text{lin}}(X) - f^{(\text{reg})}(X)\big]^2. \tag{1.3}$$

The analysis of the ordinary least squares often stops at this point in classical statistical textbooks. (Besides, to simplify, the strong assumption $f^{(\text{reg})} = f^*_{\text{lin}}$ is often made.) This can be misleading since Inequality (1.2) does not imply a d/n upper bound on the risk of $\hat{f}^{(\text{ols})}$. Nevertheless the following result holds [11, Theorem 11.3].

THEOREM 1.1 *If* $\sup_{x\in\mathcal{X}} \mathbb{V}\text{ar}(Y|X=x) = \sigma^2 < +\infty$ *and*

$$\|f^{(\text{reg})}\|_\infty = \sup_{x\in\mathcal{X}} |f^{(\text{reg})}(x)| \le H$$

for some $H > 0$*, then the truncated estimator* $\hat{f}^{(\text{ols})}_H = (\hat{f}^{(\text{ols})} \wedge H) \vee -H$ *satisfies*

$$\mathbb{E}R(\hat{f}^{(\text{ols})}_H) \;-\; R(f^{(\text{reg})}) \;\le\; 8[R(f^*_{\text{lin}}) \;-\; R(f^{(\text{reg})})] + \kappa\frac{(\sigma^2 \vee H^2)d\log n}{n} \tag{1.4}$$

for some numerical constant κ*.*

Using PAC-Bayesian inequalities, Catoni [8, Proposition 5.9.1] has proved a different type of results on the generalization ability of $\hat{f}^{(\text{ols})}$.

THEOREM 1.2 *Let* $\mathcal{F}' \subset \mathcal{F}_{\text{lin}}$ *satisfying for some positive constants* a, M, M'*:*

- *there exists* $f_0 \in \mathcal{F}'$ *s.t. for any* $x \in \mathcal{X}$*,*

$$\mathbb{E}\Big\{\exp\Big[a\big|Y - f_0(X)\big|\Big]\;\Big|\; X = x\Big\} \le M.$$

- *for any* $f_1, f_2 \in \mathcal{F}', \sup_{x\in\mathcal{X}} |f_1(x) - f_2(x)| \le M'$*.*

Let $Q = \mathbb{E}\big[\varphi(X)\varphi(X)^T\big]$ *and* $\hat{Q} = \big[\frac{1}{n}\sum_{i=1}^n \varphi(X_i)\varphi(X_i)^T\big]$ *be respectively the expected and empirical Gram matrices. If* $\det Q \neq 0$*, then there exist positive constants* C_1 *and* C_2 *(depending only on* a*,* M *and* M'*) such that with probability at least* $1-\varepsilon$*, as soon as*

$$\left\{f \in \mathcal{F}_{\text{lin}} : r(f) \le r(\hat{f}^{(\text{ols})}) + C_1\frac{d}{n}\right\} \subset \mathcal{F}', \tag{1.5}$$

we have

$$R(\hat{f}^{(\text{ols})}) - R(f^*_{\text{lin}}) \le C_2\frac{d + \log(\varepsilon^{-1}) + \log(\frac{\det\hat{Q}}{\det Q})}{n}.$$

This result can be understood as follows. Let us assume we have some prior knowledge suggesting that f^*_{lin} belongs to the interior of a set $\mathcal{F}' \subset \mathcal{F}_{\text{lin}}$ (e.g., a bound on the coefficients of the expansion of f^*_{lin} as a linear combination of $\varphi_1, \ldots, \varphi_d$). It is likely that (1.5) holds, and it is indeed proved in Catoni [8, section 5.11] that the probability that it does not hold goes to zero exponentially fast with n in the case when $\mathcal{F}'$ is a Euclidean ball. If it is the case, then we know that the excess risk is of order d/n up to the unpleasant ratio of determinants, which, fortunately, almost surely tends to 1 as n goes to infinity.

By using *localized* PAC-Bayes inequalities introduced in Catoni [7, 9], one can derive from Inequality (6.9) and Lemma 4.1 of Alquier [1] the following result.

THEOREM 1.3 *Let $q_{\min}$ be the smallest eigenvalue of the Gram matrix $Q = \mathbb{E}\big[\varphi(X)\varphi(X)^T\big]$. Assume that there exist a function $f_0 \in \mathcal{F}_{\text{lin}}$ and positive constants H and C such that*

$$\|f^*_{\text{lin}} - f_0\|_\infty \leq H.$$

and $|Y| \leq C$ almost surely.

Then for an appropriate randomized estimator requiring the knowledge of f_0, H and C, for any $\varepsilon > 0$ with probability at least $1 - \varepsilon$ w.r.t. the distribution generating the observations $Z_1, \ldots, Z_n$ and the randomized prediction function $\hat{f}$, we have

$$R(\hat{f}) - R(f^*_{\text{lin}}) \leq \kappa(H^2 + C^2)\frac{d\log(3q_{\min}^{-1}) + \log((\log n)\varepsilon^{-1})}{n}, \tag{1.6}$$

for some κ not depending on d and n.

Using the result of [8, Section 5.11], one can prove that Alquier's result still holds for $\hat{f} = \hat{f}^{(\text{ols})}$, but with κ also depending on the determinant of the product matrix Q. The $\log[\log(n)]$ factor is unimportant and could be removed in the special case quoted here (it comes from a union bound on a grid of possible temperature parameters, whereas the temperature could be set here to a fixed value). The result differs from Theorem 1.2 essentially by the fact that the ratio of the determinants of the empirical and expected product matrices has been replaced by the inverse of the smallest eigenvalue of the quadratic form $\theta \mapsto R(\sum_{j=1}^d \theta_j\varphi_j) - R(f^*_{\text{lin}})$. In the case when the expected Gram matrix is known, (e.g., in the case of a fixed design, and also in the slightly different context of transductive inference), this smallest eigenvalue can be set to one by choosing the quadratic form $\theta \mapsto R(f_\theta) - R(f^*_{\text{lin}})$ to define the Euclidean metric on the parameter space.

Localized Rademacher complexities [13, 4] allow to prove the following property of the empirical risk minimizer.

THEOREM 1.4 *Assume that the input representation $\varphi(X)$, the set of parameters and the output Y are almost surely bounded, i.e., for some positive constants H and C,*

$$\sup_{\theta\in\Theta} \|\theta\| \leq 1$$

$$\operatorname{ess\,sup} \|\varphi(X)\| \leq H,$$

and

$$|Y| \leq C \quad \text{a.s..}$$

Let $\nu_1 \geq \cdots \geq \nu_d$ be the eigenvalues of the Gram matrix $Q = \mathbb{E}\big[\varphi(X)\varphi(X)^T\big]$. The empirical risk minimizer satisfies for any $\varepsilon > 0$, with probability at least $1-\varepsilon$:

$$\begin{aligned} R(\hat{f}^{(\text{erm})}) - R(f^*) &\leq \kappa(H+C)^2 \frac{\min_{0\leq h\leq d} \left(h + \sqrt{\frac{n}{(H+C)^2}\sum_{i>h}\nu_i}\right) + \log(\varepsilon^{-1})}{n} \\ &\leq \kappa(H+C)^2 \frac{\operatorname{rank}(Q) + \log(\varepsilon^{-1})}{n}, \end{aligned}$$

where κ is a numerical constant.

PROOF. The result is a modified version of Theorem 6.7 in [4] applied to the linear kernel $k(u,v) = \langle u, v\rangle/(H+C)^2$. Its proof follows the same lines as in Theorem 6.7 *mutatis mutandi*: Corollary 5.3 and Lemma 6.5 should be used as intermediate steps instead of Theorem 5.4 and Lemma 6.6, the nonzero eigenvalues of the integral operator induced by the kernel being the nonzero eigenvalues of Q. □

When we know that the target function f^*_{lin} is inside some L^∞ ball, it is natural to consider the empirical risk minimizer on this ball. This allows to compare Theorem 1.4 to excess risk bounds with respect to f^*_{lin}.

Finally, from the work of Birgé and Massart [5], we may derive the following risk bound for the empirical risk minimizer on a L^∞ ball (see Appendix B).

THEOREM 1.5 *Assume that $\mathcal{F}$ has a diameter H for L^∞-norm, i.e., for any f_1, f_2 in $\mathcal{F}$, $\sup_{x\in\mathcal{X}} |f_1(x) - f_2(x)| \leq H$ and there exists a function $f_0 \in \mathcal{F}$ satisfying the exponential moment condition:*

$$\text{for any } x \in \mathcal{X}, \quad \mathbb{E}\Big\{\exp\Big[A^{-1}\big|Y - f_0(X)\big|\Big] \,\Big|\, X = x\Big\} \leq M, \tag{1.7}$$

for some positive constants A and M. Let

$$\tilde{B} = \inf_{\phi_1,\ldots,\phi_d} \sup_{\theta\in\mathbb{R}^d-\{0\}} \frac{\|\sum_{j=1}^d \theta_j\phi_j\|_\infty^2}{\|\theta\|_\infty^2}$$

where the infimum is taken with respect to all possible orthonormal basis of $\mathcal{F}$ for the dot product $\langle f_1, f_2\rangle = \mathbb{E}f_1(X)f_2(X)$ (when the set $\mathcal{F}$ admits no basis with

exactly d functions, we set $\tilde{B} = +\infty$). Then the empirical risk minimizer satisfies for any $\varepsilon > 0$, with probability at least $1 - \varepsilon$:

$$R(\hat{f}^{(\text{erm})}) - R(f^*) \leq \kappa(A^2 + H^2) \frac{d \log[2 + (\tilde{B}/n) \wedge (n/d)] + \log(\varepsilon^{-1})}{n},$$

where κ is a positive constant depending only on M.

This result comes closer to what we are looking for: it gives exponential deviation inequalities of order at worse $d \log(n/d)/n$. It shows that, even if the Gram matrix Q has a very small eigenvalue, there is an algorithm satisfying a convergence rate of order $d \log(n/d)/n$. With this respect, this result is stronger than Theorem 1.3. However there are cases in which the smallest eigenvalue of Q is of order 1, while $\tilde{B}$ is large (i.e., $\tilde{B} \gg n$). In these cases, Theorem 1.3 does not contain the logarithmic factor which appears in Theorem 1.5.

1.2. PROJECTION ESTIMATOR. When the input distribution is known, an alternative to the ordinary least squares estimator is the following projection estimator. One first finds an orthonormal basis of $\mathcal{F}_{\text{lin}}$ for the dot product $\langle f_1, f_2 \rangle = \mathbb{E} f_1(X) f_2(X)$, and then uses the projection estimator on this basis. Specifically, if $\phi_1, \ldots, \phi_d$ form an orthonormal basis of $\mathcal{F}_{\text{lin}}$, then the projection estimator on this basis is:

$$\hat{f}^{(\text{proj})} = \sum_{j=1}^{d} \hat{\theta}_j^{(\text{proj})} \phi_j,$$

with

$$\hat{\theta}^{(\text{proj})} = \frac{1}{n} \sum_{i=1}^{n} Y_i \phi_j(X_i).$$

Theorem 4 in [20] gives a simple bound of order d/n on the expected excess risk $\mathbb{E} R(\hat{f}^{(\text{proj})}) - R(f^*_{\text{lin}})$.

1.3. PENALIZED LEAST SQUARES ESTIMATOR. It is well established that parameters of the ordinary least squares estimator are numerically unstable, and that the phenomenon can be corrected by adding an L^2 penalty ([15, 18]). This solution has been labeled ridge regression in statistics ([12]), and consists in replacing $\hat{f}^{(\text{ols})}$ by $\hat{f}^{(\text{ridge})} = f_{\hat{\theta}^{(\text{ridge})}}$ with

$$\hat{\theta}^{(\text{ridge})} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ r(f_\theta) + \lambda \sum_{j=1}^{d} \theta_j^2 \right\},$$

where λ is a positive parameter. The typical value of λ should be small to avoid excessive shrinkage of the coefficients, but not too small in order to make the optimization task numerically more stable.

Risk bounds for this estimator can be derived from general results concerning penalized least squares on reproducing kernel Hilbert spaces ([6]), but as it is shown in Appendix C, this ends up with complicated results having the desired d/n rate only under strong assumptions.

Another popular regularizer is the L^1 norm. This procedure is known as Lasso [19] and is defined by

$$\hat{\theta}^{(\text{lasso})} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ r(f_\theta) + \lambda \sum_{j=1}^{d} |\theta_j| \right\}.$$

As the L^2 penalty, the L^1 penalty shrinks the coefficients. The difference is that for coefficients which tend to be close to zero, the shrinkage makes them equal to zero. This allows to select relevant variables (i.e., find the j's such that $\theta_j^* \neq 0$). If we assume that the regression function $f^{(\text{reg})}$ is a linear combination of only $d^* \ll d$ variables/functions φ_j's, the typical result is to prove that the risk of the Lasso estimator for λ of order $\sqrt{(\log d)/n}$ is of order $(d^* \log d)/n$. Since this quantity is much smaller than d/n, this makes a huge improvement (provided that the sparsity assumption is true). This kind of results usually requires strong conditions on the eigenvalues of submatrices of Q, essentially assuming that the functions φ_j are near orthogonal. We do not know to which extent these conditions are required. However, if we do not consider the specific algorithm of Lasso, but the model selection approach developed in [1], one can change these conditions into a single condition concerning only the minimal eigenvalue of the submatrix of Q corresponding to relevant variables. In fact, we will see that even this condition can be removed.

1.4. CONCLUSION OF THE SURVEY. Previous results clearly leave room to improvements. The projection estimator requires the unrealistic assumption that the input distribution is known, and the result holds only in expectation. Results using L^1 or L^2 regularizations require strong assumptions, in particular on the eigenvalues of (submatrices of) Q. Theorem 1.1 provides a $(d \log n)/n$ convergence rate only when the $R(f_{\text{lin}}^*) - R(f^{(\text{reg})})$ is at most of order $(d \log n)/n$. Theorem 1.2 gives a different type of guarantee: the d/n is indeed achieved, but the random ratio of determinants appearing in the bound may raise some eyebrows and forbid an explicit computation of the bound and comparison with other bounds. Theorem 1.3 seems to indicate that the rate of convergence will be degraded when the Gram matrix Q is unknown and ill-conditioned. Theorem 1.4 does not put any assumption on Q to reach the d/n rate, but requires particular boundedness constraints on the parameter set, the input vector $\varphi(X)$ and the output. Finally, Theorem 1.5 comes closer to what we are looking for. Yet there is still an unwanted loga-

rithmic factor, and the result holds only when the output has uniformly bounded conditional exponential moments, which as we will show is not necessary.

2. Ridge regression and empirical risk minimization

We recall the definition

$$\mathcal{F} = \big\{ f_\theta = \sum_{j=1}^{d} \theta_j \varphi_j ; (\theta_1, \dots, \theta_d) \in \Theta \big\},$$

where Θ is a closed convex set, not necessarily bounded (so that $\Theta = \mathbb{R}^d$ is allowed). In this section, we provide exponential deviation inequalities for the empirical risk minimizer and the ridge regression estimator on $\mathcal{F}$ under weak conditions on the tail of the output distribution.

The most general theorem which can be obtained from the route followed in this section is Theorem 6.5 (page 46) stated along with the proof. It is expressed in terms of a series of empirical bounds. The first deduction we can make from this technical result is of asymptotic nature. It is stated under weak hypotheses, taking advantage of the weak law of large numbers.

THEOREM 2.1 *For $\lambda \geq 0$, let $\tilde{f}$ be its associated optimal ridge function (see* (0.4)*). Let us assume that*

$$\mathbb{E}\big[\|\varphi(X)\|^4\big] < +\infty, \tag{2.1}$$

$$\textit{and} \quad \mathbb{E}\Big\{ \|\varphi(X)\|^2 \big[\tilde{f}(X) - Y\big]^2 \Big\} < +\infty. \tag{2.2}$$

Let $\nu_1, \dots, \nu_d$ be the eigenvalues of the Gram matrix $Q = \mathbb{E}\big[\varphi(X)\varphi(X)^T\big]$, and let $Q_\lambda = Q + \lambda I$ be the ridge regularization of Q. Let us define the effective ridge dimension

$$D = \sum_{i=1}^{d} \frac{\nu_i}{\nu_i + \lambda} \mathbb{1}(\nu_i > 0) = \mathrm{Tr}\big[(Q + \lambda I)^{-1} Q\big] = \mathbb{E}\big[\|Q_\lambda^{-1/2} \varphi(X)\|^2\big].$$

When $\lambda = 0$, D is equal to the rank of Q and is otherwise smaller. For any $\varepsilon > 0$, there is n_ε, such that for any $n \geq n_\varepsilon$, with probability at least $1 - \varepsilon$,

$$\begin{aligned} R(\hat{f}^{(\text{ridge})}) + \lambda \|\hat{\theta}^{(\text{ridge})}\|^2 & \\ \leq \min_{\theta \in \Theta} \big\{ R(f_\theta) + \lambda \|\theta\|^2 \big\} & \\ + \frac{30\, \mathbb{E}\big\{ \|Q_\lambda^{-1/2} \varphi(X)\|^2 \big[\tilde{f}(X) - Y\big]^2 \big\}}{\mathbb{E}\big\{ \|Q_\lambda^{-1/2} \varphi(X)\|^2 \big\}} \, \frac{D}{n} & \end{aligned}$$

$$
\begin{aligned}
&\quad + 1000 \sup_{v\in\mathbb{R}^d} \frac{\mathbb{E}\Big[\langle v,\varphi(X)\rangle^2\big[\tilde f(X)-Y\big]^2\Big]}{\mathbb{E}(\langle v,\varphi(X)\rangle^2)+\lambda\|v\|^2}\,\frac{\log(3\varepsilon^{-1})}{n}\\
&\le \min_{\theta\in\Theta}\big\{R(f_\theta)+\lambda\|\theta\|^2\big\}\\
&\qquad + \operatorname{ess\,sup}\mathbb{E}\big\{[Y-\tilde f(X)]^2\big|X\big\}\,\frac{30D+1000\log(3\varepsilon^{-1})}{n}
\end{aligned}
$$

PROOF. See Section 6.2 (page 40). □

This theorem shows that the ordinary least squares estimator (obtained when $\Theta = \mathbb{R}^d$ and $\lambda = 0$), as well as the empirical risk minimizer on any closed convex set, asymptotically reaches a d/n speed of convergence under very weak hypotheses. It shows also the regularization effect of the ridge regression. There emerges an *effective dimension* D, where the ridge penalty has a threshold effect on the eigenvalues of the Gram matrix.

On the other hand, the weakness of this result is its asymptotic nature : n_ε may be arbitrarily large under such weak hypotheses, and this shows even in the simplest case of the estimation of the mean of a real valued random variable by its empirical mean (which is the case when $d = 1$ and $\varphi(X) \equiv 1$).

Let us now give some non asymptotic rate under stronger hypotheses and for the empirical risk minimizer (i.e., $\lambda = 0$).

THEOREM 2.2 *Let $d' = \operatorname{rank}(Q)$. Assume that*

$$\mathbb{E}\big\{[Y-f^*(X)]^4\big\} < +\infty$$

and

$$B = \sup_{f\in\operatorname{span}\{\varphi_1,\dots,\varphi_d\}-\{0\}} \|f\|_\infty^2/\mathbb{E}[f(X)^2] < +\infty.$$

Consider the (unique) empirical risk minimizer $\hat f^{(\mathrm{erm})} = f_{\hat\theta^{(\mathrm{erm})}} : x \mapsto \langle\hat\theta^{(\mathrm{erm})},\varphi(x)\rangle$ on $\mathcal{F}$ for which $\hat\theta^{(\mathrm{erm})} \in \operatorname{span}\{\varphi(X_1),\dots,\varphi(X_n)\}$[4]. For any values of ε and n such that $2/n \le \varepsilon \le 1$ and

$$n > 1280B^2\left[3Bd' + \log(2/\varepsilon) + \frac{16B^2d'^2}{n}\right],$$

with probability at least $1-\varepsilon$,

$$
\begin{aligned}
&R(\hat f^{(\mathrm{erm})}) - R(f^*)\\
&\qquad \le 1920\,B\sqrt{\mathbb{E}[Y-f^*(X)]^4}\left[\frac{3Bd'+\log(2\varepsilon^{-1})}{n} + \left(\frac{4Bd'}{n}\right)^2\right].
\end{aligned}
$$

[4]When $\mathcal{F} = \mathcal{F}_{\mathrm{lin}}$, we have $\hat\theta^{(\mathrm{erm})} = \mathbf{X}^+\,\mathbf{Y}$, with $\mathbf{X} = (\varphi_j(X_i))_{1\le i\le n, 1\le j\le d}$, $\mathbf{Y} = [Y_j]_{j=1}^n$ and $\mathbf{X}^+$ is the Moore-Penrose pseudoinverse of $\mathbf{X}$.

PROOF. See Section 6.2 (page 40). □

It is quite surprising that the traditional assumption of uniform boundedness of the conditional exponential moments of the output can be replaced by a simple moment condition for reasonable confidence levels (i.e., $\varepsilon \geq 2/n$). For highest confidence levels, things are more tricky since we need to control with high probability a term of order $[r(f^*) - R(f^*)]d/n$ (see Theorem 6.6). The cost to pay to get the exponential deviations under only a fourth-order moment condition on the output is the appearance of the geometrical quantity B as a multiplicative factor, as opposed to Theorems 1.3 and 1.5. More precisely, from [5, Inequality (3.2)], we have $B \leq \tilde{B} \leq Bd$, but the quantity $\tilde{B}$ appears inside a logarithm in Theorem 1.5. However, Theorem 1.5 is restricted to the empirical risk minimizer on a L^∞ ball, while the result here is valid for any closed convex set Θ, and in particular applies to the ordinary least squares estimator.

Theorem 2.2 is still limited in at least three ways: it applies only to uniformly bounded $\varphi(X)$, the output needs to have a fourth moment, and the confidence level should be as great as $\varepsilon \geq 2/n$. These limitations will be addressed in the next sections by considering more involved algorithms.

3. A MIN-MAX ESTIMATOR FOR ROBUST ESTIMATION

3.1. THE MIN-MAX ESTIMATOR AND ITS THEORETICAL GUARANTEE. This section provides an alternative to the empirical risk minimizer with non asymptotic exponential risk deviations of order d/n for any confidence level. Moreover, we will assume only a second order moment condition on the output and cover the case of unbounded inputs, the requirement on $\varphi(X)$ being only a finite fourth order moment. On the other hand, we assume that the set Θ of the vectors of coefficients is bounded. The computability of the proposed estimator and numerical experiments are discussed at the end of the section.

Let $\alpha > 0$, $\lambda \geq 0$, and consider the truncation function:

$$\psi(x) = \begin{cases} -\log\bigl(1 - x + x^2/2\bigr) & 0 \leq x \leq 1, \\ \log(2) & x \geq 1, \\ -\psi(-x) & x \leq 0, \end{cases}$$

For any $\theta, \theta' \in \Theta$, introduce

$$\mathcal{D}(\theta, \theta') = n\alpha\lambda(\|\theta\|^2 - \|\theta'\|^2) + \sum_{i=1}^{n} \psi\Bigl(\alpha\bigl[Y_i - f_\theta(X_i)\bigr]^2 - \alpha\bigl[Y_i - f_{\theta'}(X_i)\bigr]^2\Bigr).$$

We recall $\tilde{f} = f_{\tilde{\theta}}$ with $\tilde{\theta} \in \arg\min_{\theta\in\Theta} \bigl\{R(f_\theta) + \lambda\|\theta\|^2\bigr\}$, and the effective ridge

dimension

$$D = \sum_{i=1}^{d} \frac{\nu_i}{\nu_i + \lambda} \mathbb{1}(\nu_i > 0) = \mathrm{Tr}\big[(Q+\lambda I)^{-1} Q\big] = \mathbb{E}\big[\|Q_\lambda^{-1/2}\varphi(X)\|^2\big].$$

Let us assume in this section that for any $j \in \{1, \dots, d\}$,

$$\mathbb{E}\big\{\varphi_j(X)^2 [Y - \tilde{f}(X)]^2\big\} < +\infty, \tag{3.1}$$

and

$$\mathbb{E}\big[\varphi_j^4(X)\big] < +\infty. \tag{3.2}$$

Define

$$\mathcal{S} = \{f \in \mathcal{F}_{\mathrm{lin}} : \mathbb{E}[f(X)^2] = 1\}, \tag{3.3}$$

$$\sigma = \sqrt{\mathbb{E}\big\{[Y - \tilde{f}(X)]^2\big\}} = \sqrt{R(\tilde{f})}, \tag{3.4}$$

$$\chi = \max_{f \in \mathcal{S}} \sqrt{\mathbb{E}[f(X)^4]}, \tag{3.5}$$

$$\kappa = \frac{\sqrt{\mathbb{E}\big\{[\varphi(X)^T Q_\lambda^{-1} \varphi(X)]^2\big\}}}{\mathbb{E}\big[\varphi(X)^T Q_\lambda^{-1} \varphi(X)\big]}, \tag{3.6}$$

$$\kappa' = \frac{\sqrt{\mathbb{E}\big\{[Y - \tilde{f}(X)]^4\big\}}}{\mathbb{E}\big\{[Y - \tilde{f}(X)]^2\big\}} = \frac{\sqrt{\mathbb{E}\big\{[Y - \tilde{f}(X)]^4\big\}}}{\sigma^2}, \tag{3.7}$$

$$T = \max_{\theta \in \Theta, \theta' \in \Theta} \sqrt{\lambda \|\theta - \theta'\|^2 + \mathbb{E}[f_\theta(X) - f_{\theta'}(X)]^2}. \tag{3.8}$$

THEOREM 3.1 *Let us assume that* (3.1) *and* (3.2) *hold. For some numerical constants c and c', for*

$$n > c\kappa\chi D,$$

by taking

$$\alpha = \frac{1}{2\chi\big[2\sqrt{\kappa'}\sigma + \sqrt{\chi}T\big]^2}\left(1 - \frac{c\kappa\chi D}{n}\right), \tag{3.9}$$

for any estimator $f_{\hat{\theta}}$ satisfying $\hat{\theta} \in \Theta$ a.s., for any $\varepsilon > 0$ and any $\lambda \geq 0$, with probability at least $1 - \varepsilon$, we have

$$\begin{aligned} R(f_{\hat{\theta}}) + \lambda\|\hat{\theta}\|^2 \leq \min_{\theta \in \Theta} \big\{R(f_\theta) + \lambda\|\theta\|^2\big\} \\ + \frac{1}{n\alpha}\left(\max_{\theta_1 \in \Theta} \mathcal{D}(\hat{\theta}, \theta_1) - \inf_{\theta \in \Theta} \max_{\theta_1 \in \Theta} \mathcal{D}(\theta, \theta_1)\right) \\ + \frac{c\kappa\kappa' D\sigma^2}{n} + \frac{8\chi\big(\frac{\log(\varepsilon^{-1})}{n} + \frac{c'\kappa^2 D^2}{n^2}\big)\big[2\sqrt{\kappa'}\sigma + \sqrt{\chi}T\big]^2}{1 - \frac{c\kappa\chi D}{n}}. \end{aligned}$$

PROOF. See Section 6.3 (page 50). □

By choosing an estimator such that

$$\max_{\theta_1\in\Theta} \mathcal{D}(\hat{\theta},\theta_1) < \inf_{\theta\in\Theta} \max_{\theta_1\in\Theta} \mathcal{D}(\theta,\theta_1) + \sigma^2\frac{D}{n},$$

Theorem 3.1 provides a non asymptotic bound for the excess (ridge) risk with a D/n convergence rate and an exponential tail even when neither the output Y nor the input vector $\varphi(X)$ has exponential moments. This stronger non asymptotic bound compared to the bounds of the previous section comes at the price of replacing the empirical risk minimizer by a more involved estimator. Section 3.3 provides a way of computing it approximately.

3.2. THE VALUE OF THE UNCENTERED KURTOSIS COEFFICIENT χ. Let us discuss here the value of constant χ, which plays a critical role in the speed of convergence of our bound. With the convention $\frac{0}{0}=0$, we have

$$\chi = \sup_{u\in\mathbb{R}^d} \frac{\mathbb{E}\big(\langle u,\varphi(X)\rangle^4\big)^{1/2}}{\mathbb{E}\big(\langle u,\varphi(X)\rangle^2\big)}.$$

Let us first examine the case when $\varphi_1(X)\equiv 1$ and $\big[\varphi_j(X), j=2,\dots,d\big]$ are independent. To compute χ, we can assume without loss of generality that they are centered and of unit variance, which will be the case after $Q^{-1/2}$ is applied to them. In this situation, introducing

$$\chi_* = \max_{j=1,\dots,d} \frac{\mathbb{E}\big[\varphi_j(X)^4\big]^{1/2}}{\mathbb{E}\big[\varphi_j(X)^2\big]},$$

we see that for any $u\in\mathbb{R}^d$ with $\|u\|=1$, we have

$$\begin{aligned}
\mathbb{E}\big(\langle u,\varphi(X)\rangle^4\big) &= \sum_{i=1}^d u_i^4\mathbb{E}(\varphi_i(X)^4) + 6\sum_{1\le i<j\le d} u_i^2u_j^2\mathbb{E}\big[\varphi_i(X)^2\big]\mathbb{E}\big[\varphi_j(X)^2\big] \\
&\qquad + 4\sum_{i=2}^d u_1u_i^3\mathbb{E}\big[\varphi_i(X)^3\big] \\
&\le \chi_*^2\sum_{i=1}^d u_i^4 + 6\sum_{i<j} u_i^2u_j^2 + 4\chi_*^{3/2}\sum_{i=2}^d|u_1u_i|^3 \\
&\le \sup_{u\in\mathbb{R}_+^d,\|u\|=1} \big(\chi_*^2-3\big)\sum_{i=1}^d u_i^4 + 3\left(\sum_{i=1}^d u_i^2\right)^2 + 4\chi_*^{3/2}u_1\sum_{i=2}^d u_i^3
\end{aligned}$$

$$\leq \frac{3^{3/2}}{4}\chi_*^{3/2} + \begin{cases} \chi_*^2, & \chi_*^2 \geq 3, \\ 3 + \frac{\chi_*^2 - 3}{d}, & 1 \leq \chi_*^2 < 3. \end{cases}$$

Thus in this case

$$\chi \leq \begin{cases} \chi_* \left(1 + \frac{3^{3/2}}{4\sqrt{\chi_*}}\right)^{1/2}, & \chi_* \geq \sqrt{3}, \\ \left(3 + \frac{3^{3/2}}{4}\chi_*^{3/2} + \frac{\chi_*^2 - 3}{d}\right)^{1/2}, & 1 \leq \chi_* < \sqrt{3}. \end{cases}$$

If moreover the random variables $\varphi_j(X)$ are not skewed, in the sense that $\mathbb{E}\big[\varphi_j(X)^3\big] = 0$, $j = 2, \ldots, d$, then

$$\begin{cases} \chi = \chi_*, & \chi_* \geq \sqrt{3}, \\ \chi \leq \left(3 + \frac{\chi_*^2 - 3}{d}\right)^{1/2}, & 1 \leq \chi_* < \sqrt{3}. \end{cases}$$

In particular in the case when $\varphi_j(X)$ are Gaussian variables, $\chi = \chi_* = \sqrt{3}$ (as could be seen in a more straightforward way, since in this case $\langle u, \varphi(X)\rangle$ is also Gaussian !).

In particular, this situation arises in compress sensing using random projections on Gaussian vectors. Specifically, assume that we want to recover a signal $f \in \mathbb{R}^M$ that we know to be well approximated by a linear combination of d basis vectors $f_1, \ldots, f_d$. We measure $n \ll M$ projections of the signal f on i.i.d. M-dimensional standard normal random vectors $X_1, \ldots, X_n$: $Y_i = \big\langle f, X_i\big\rangle$, $i = 1, \ldots, n$. Then, recovering the coefficient $\theta_1, \ldots, \theta_d$ such that $f = \sum_{j=1}^d \theta_j f_j$ is associated to the least squares regression problem $Y \approx \sum_{j=1}^d \theta_j \varphi_j(X)$, with $\varphi_j(x) = \langle f_j, x\rangle$, and X having a M-dimensional standard normal distribution.

Let us discuss now a bound which is suited to the case when we are using a partial basis of regression functions. The functions φ_j are usually bounded (think of the Fourier basis, wavelet bases, histograms, splines ...).

Let us assume that for some positive constant A and any $u \in \mathbb{R}^d$,

$$\|u\| \leq A\mathbb{E}\big[\langle u, \varphi(X)\rangle^2\big]^{1/2}.$$

This appears as some stability property of the partial basis φ_j with respect to the $\mathbb{L}_2$-norm, since it can also be written as

$$\sum_{j=1}^d u_j^2 \leq A^2\mathbb{E}\left[\left(\sum_{j=1}^d u_j\varphi_j(X)\right)^2\right], \qquad u \in \mathbb{R}^d.$$

This will be the case if φ_j is nearly orthogonal in the sense that

$$\mathbb{E}\big[\varphi_j(X)^2\big] \geq 1, \quad \text{and} \quad \Big|\mathbb{E}\big[\varphi_j(X)\varphi_k(X)\big]\Big| \leq \frac{1 - A^2}{d - 1}.$$

In this situation, by using

$$\mathbb{E}\big[\langle u, \varphi(X)\rangle^4\big] \le \|u\|^2 \operatorname{ess\,sup}\|\varphi(X)\|^2 \mathbb{E}\big[\langle u, \varphi(X)\rangle^2\big],$$

one can check that

$$\chi \le A \left\| \left(\sum_{j=1}^{d} \varphi_j^2 \right)^{1/2} \right\|_\infty .$$

Therefore, if X is the uniform random variable on the unit interval and φ_j, $j = 1, \dots, d$ are any functions from the Fourier basis (meaning that they are of the form $\sqrt{2}\cos(2k\pi X)$ or $\sqrt{2}\sin(2k\pi X)$), then $\chi \le \sqrt{2d}$ (because they form an orthogonal system, so that $A = 1$).

On the other hand, a localized basis like the evenly spaced histogram basis of the unit interval

$$\varphi_j(x) = \sqrt{d}\mathbb{1}\Big(x \in \big[(j-1)/d, j/d\big[\Big),$$

will also be such that $\chi \le \sqrt{d}$. Similar computations could be made for other local bases, like wavelet bases. Note that when χ is of order $\sqrt{d}$, Theorem 3.1 means that the excess risk of the min-max truncated estimator $\hat{f}$ is upper bounded by $C\frac{d}{n}$ provided that $n \ge Cd\sqrt{d}$ for a large enough constant C.

Let us discuss the case when X is some observed random variable whose distribution is only approximately known. Namely let us assume that $(\varphi_j)_{j=1}^d$ is some basis of functions in $\mathbb{L}_2\big[\widetilde{\mathbb{P}}\big]$ with some known coefficient $\widetilde{\chi}$, where $\widetilde{\mathbb{P}}$ is an approximation of the true distribution of X in the sense that the density of the true distribution $\mathbb{P}$ of X with respect to the distribution $\widetilde{\mathbb{P}}$ is in the range $(\eta^{-1/2}, \eta)$. In this situation, the coefficient χ satisfies the inequality $\chi \le \eta\widetilde{\chi}$. Indeed

$$\begin{aligned}\mathbb{E}_{X\sim\mathbb{P}}\big[\langle u, \varphi(X)\rangle^4\big] &\le \eta \mathbb{E}_{X\sim\widetilde{\mathbb{P}}}\big[\langle u, \varphi(X)\rangle^4\big] \\ &\le \eta\widetilde{\chi}^2 \mathbb{E}_{X\sim\widetilde{\mathbb{P}}}\big[\langle u, \varphi(X)\rangle^2\big]^2 \le \eta^2\widetilde{\chi}^2 \mathbb{E}_{X\sim\mathbb{P}}\big[\langle u, \varphi(X)\rangle^2\big]^2.\end{aligned}$$

Let us conclude this section with some scenario for the case when X is a real-valued random variable. Let us consider the distribution function of $\widetilde{\mathbb{P}}$

$$\widetilde{F}(x) = \widetilde{\mathbb{P}}(X \le x).$$

Then, if $\widetilde{\mathbb{P}}$ has no atoms, the distribution of $\widetilde{F}(X)$ is uniform in $(0, 1)$. Starting from some suitable partial basis $(\varphi_j)_{j=1}^d$ of $\mathbb{L}_2\big[(0,1), \mathbb{U}\big]$ where $\mathbb{U}$ is the uniform distribution, like the ones discussed above, we can build a basis for our problem as

$$\widetilde{\varphi}_j(X) = \varphi_j\big[\widetilde{F}(X)\big].$$

Moreover, if $\mathbb{P}$ is absolutely continuous with respect to $\widetilde{\mathbb{P}}$ with density g, then $\mathbb{P}\circ\widetilde{F}^{-1}$ is absolutely continuous with respect to $\widetilde{\mathbb{P}}\circ\widetilde{F}^{-1}$, with density $g\circ\widetilde{F}^{-1}$, and

of course, the fact that g takes values in $(\eta^{-1/2}, \eta)$ implies the same property for $g \circ \widetilde{F}^{-1}$. Thus, if $\widetilde{\chi}$ is the coefficient corresponding to $\varphi_j(U)$ when U is the uniform random variable on the unit interval, then the true coefficient χ (corresponding to $\widetilde{\varphi}_j(X)$) will be such that $\chi \leq \eta\widetilde{\chi}$.

3.3. COMPUTATION OF THE ESTIMATOR. For ease of description of the algorithm, we will write X for $\varphi(X)$, which is equivalent to considering without loss of generality that the input space is $\mathbb{R}^d$ and that the functions $\varphi_1, \ldots, \varphi_d$ are the coordinate functions. Therefore, the function f_θ maps an input x to $\langle \theta, x \rangle$.

Let us introduce

$$\overline{L}_i(\theta) = \alpha\big(\langle \theta, X_i \rangle - Y_i\big)^2.$$

For any subset of indices $I \subset \{1, \ldots, n\}$, let us define

$$r_I(\theta) = \lambda \|\theta\|^2 + \frac{1}{\alpha |I|} \sum_{i \in I} \overline{L}_i(\theta).$$

We suggest the following heuristics to compute an approximation of

$$\arg\min_{\theta \in \Theta} \sup_{\theta' \in \Theta} \mathcal{D}(\theta, \theta').$$

- Start from $I_1 = \{1, \ldots, n\}$ with the empirical risk minimizer

$$\widehat{\theta}_1 = \arg\min_{\mathbb{R}^d} r_{I_1} = \hat{\theta}^{(\text{erm})}.$$

- At step number k, compute

$$\widehat{Q}_k = \frac{1}{|I_k|} \sum_{i \in I_k} X_i X_i^T.$$

- Consider the sets

$$J_{k,1}(\eta) = \left\{ i \in I_k : \overline{L}_i(\widehat{\theta}_k) X_i^T \widehat{Q}_k^{-1} X_i \left(1 + \sqrt{1 + \big[\overline{L}_i(\widehat{\theta}_k)\big]^{-1}} \right)^2 < \eta \right\},$$

  where $\widehat{Q}_k^{-1}$ is the (pseudo-)inverse of the matrix $\widehat{Q}_k$.

- Let us define

$$\theta_{k,1}(\eta) = \arg\min_{\mathbb{R}^d} r_{J_{k,1}(\eta)},$$
$$J_{k,2}(\eta) = \Big\{ i \in I_k : \big|\overline{L}_i\big(\theta_{k,1}(\eta)\big) - \overline{L}_i\big(\widehat{\theta}_k\big)\big| \leq 1 \Big\},$$

$$
\begin{aligned}
\theta_{k,2}(\eta) &= \arg\min_{\mathbb{R}^d} r_{J_{k,2}(\eta)},\\
(\eta_k, \ell_k) &= \arg\min_{\eta\in\mathbb{R}_+, \ell\in\{1,2\}} \max_{j=1,\dots,k} \mathcal{D}\big(\theta_{k,\ell}(\eta), \widehat{\theta}_j\big),\\
I_{k+1} &= J_{k,\ell_k}(\eta_k),\\
\widehat{\theta}_{k+1} &= \theta_{k,\ell_k}(\eta_k).
\end{aligned}
$$

- Stop when
$$\max_{j=1,\dots,k} \mathcal{D}(\widehat{\theta}_{k+1}, \widehat{\theta}_j) \geq 0,$$
and set $\widehat{\theta} = \widehat{\theta}_k$ as the final estimator of $\tilde{\theta}$.

Note that there will be at most n steps, since $I_{k+1} \subsetneq I_k$ and in practice much less in this iterative scheme. Let us give some justification for this proposal. Let us notice first that

$$
\begin{aligned}
\mathcal{D}(\theta + h, \theta) = n\alpha\lambda(\|\theta + h\|^2 - \|\theta\|^2)&\\
+ \sum_{i=1}^{n} \psi\Big(\alpha\big[2\langle h, X_i\rangle\big(\langle\theta, X_i\rangle - Y_i\big) + \langle h, X_i\rangle^2\big]\Big).&
\end{aligned}
$$

Hopefully, $\tilde{\theta} = \arg\min_{\theta\in\mathbb{R}^d}\big(R(f_\theta) + \lambda\|\theta\|^2\big)$ is in some small neighbourhood of $\widehat{\theta}_k$ already, according to the distance defined by $Q \simeq \widehat{Q}_k$. So we may try to look for improvements of $\widehat{\theta}_k$ by exploring neighbourhoods of $\widehat{\theta}_k$ of increasing sizes with respect to some approximation of the relevant norm $\|\theta\|_Q^2 = \mathbb{E}\big[\langle\theta, X\rangle^2\big]$.

Since the truncation function ψ is constant on $(-\infty, -1]$ and $[1, +\infty)$, the map $\theta \mapsto \mathcal{D}(\theta, \widehat{\theta}_k)$ induces a decomposition of the parameter space into cells corresponding to different sets I of examples. Indeed, such a set I is associated to the set $\mathcal{C}_I$ of θ such that $\overline{L}_i(\theta) - \overline{L}_i(\widehat{\theta}_k) < 1$ if and only if $i \in I$. Although this may not be the case, we will do as if the map $\theta \mapsto \mathcal{D}(\theta, \widehat{\theta}_k)$ restricted to the cell $\mathcal{C}_I$ reached its minimum at some interior point of $\mathcal{C}_I$, and approximates this minimizer by the minimizer of r_I.

The idea is to remove first the examples which will become inactive in the closest cells to the current estimate $\widehat{\theta}_k$. The cells for which the contribution of example number i is constant are delimited by at most four parallel hyperplanes.

It is easy to see that the square of the inverse of the distance of $\widehat{\theta}_k$ to the closest of these hyperplanes is equal to

$$\frac{1}{\alpha} X_i^T \widehat{Q}_k^{-1} X_i \overline{L}_i(\widehat{\theta}_k)\left(1 + \sqrt{1 + \frac{1}{\overline{L}_i(\widehat{\theta}_k)}}\right)^2.$$

Indeed, this distance is the infimum of $\|\widehat{Q}_k^{1/2}h\|$, where h is a solution of

$$\langle h, X_i\rangle^2 + 2\langle h, X_i\rangle\big(\langle\widehat{\theta}_k, X_i\rangle - Y_i\big) = \frac{1}{\alpha}.$$

It is computed by considering h of the form $h = \xi\|\widehat{Q}_k^{-1/2}X_i\|^{-1}\widehat{Q}_k^{-1}X_i$ and solving an equation of order two in ξ.

This explains the proposed choice of $J_{k,1}(\eta)$. Then a first estimate $\theta_{k,1}(\eta)$ is computed on the basis of this reduced sample, and the sample is readjusted to $J_{k,2}(\eta)$ by checking which constraints are really activated in the computation of $\mathcal{D}(\theta_{k,1}(\eta), \widehat{\theta}_k)$. The estimated parameter is then readjusted taking into account the readjusted sample (this could as a variant be iterated more than once). Now that we have some new candidates $\theta_{k,\ell}(\eta)$, we check the minimax property against them to elect I_{k+1} and $\widehat{\theta}_{k+1}$. Since we did not check the minimax property against the whole parameter set $\Theta = \mathbb{R}^d$, we have no theoretical warranty for this simplified algorithm. Nonetheless, similar computations to what we did could prove that we are close to solving $\min_{j=1,\dots,k} R(f_{\widehat{\theta}_j})$, since we checked the minimax property on the reduced parameter set $\{\widehat{\theta}_j, j = 1, \dots, k\}$. Thus the proposed heuristics is capable of improving on the performance of the ordinary least squares estimator, while being guaranteed not to degrade its performance significantly.

3.4. SYNTHETIC EXPERIMENTS. In Section 3.4.1, we detail the three kinds of noises we work with. Then, Sections 3.4.2, 3.4.3 and 3.4.4 describe the three types of functional relationships between the input, the output and the noise involved in our experiments. A motivation for choosing these input-output distributions was the ability to compute exactly the excess risk, and thus to compare easily estimators. Section 3.4.5 provides details about the implementation, its computational efficiency and the main conclusions of the numerical experiments. Figures and tables are postponed to Appendix E.

3.4.1. Noise distributions. In our experiments, we consider three types of noise that are centered and with unit variance:

- the standard Gaussian noise: $W \sim \mathcal{N}(0, 1)$,
- a heavy-tailed noise defined by: $W = \mathrm{sign}(V)/|V|^{1/q}$, with $V \sim \mathcal{N}(0, 1)$ a standard Gaussian random variable and $q = 2.01$ (the real number q is taken strictly larger than 2 as for $q = 2$, the random variable W would not admit a finite second moment).
- a mixture of a Dirac random variable with a low-variance Gaussian random variable defined by: with probability p, $W = \sqrt{(1-\rho)/p}$, and with

probability $1-p$, W is drawn from

$$\mathcal{N}\Big(-\frac{\sqrt{p(1-\rho)}}{1-p}, \frac{\rho}{1-p}-\frac{p(1-\rho)}{(1-p)^2}\Big).$$

The parameter $\rho \in [p,1]$ characterizes the part of the variance of W explained by the Gaussian part of the mixture. Note that this noise admits exponential moments, but for n of order $1/p$, the Dirac part of the mixture generates low signal to noise points.

3.4.2. Independent normalized covariates (INC(n,d)). In INC(n,d), the input-output pair is such that

$$Y = \langle \theta^*, X\rangle + \sigma W,$$

where the components of X are independent standard normal distributions, $\theta^* = (10,\dots,10)^T \in \mathbb{R}^d$, and $\sigma = 10$.

3.4.3. Highly correlated covariates (HCC(n,d)). In HCC(n,d), the input-output pair is such that

$$Y = \langle \theta^*, X\rangle + \sigma W,$$

where X is a multivariate centered normal Gaussian with covariance matrix Q obtained by drawing a (d,d)-matrix A of uniform random variables in $[0,1]$ and by computing $Q = AA^T$, $\theta^* = (10,\dots,10)^T \in \mathbb{R}^d$, and $\sigma = 10$. So the only difference with the setting of Section 3.4.2 is the correlation between the covariates.

3.4.4. Trigonometric series (TS(n,d)). Let X be a uniform random variable on $[0,1]$. Let d be an even number. Let

$$\varphi(X) = \big(\cos(2\pi X),\dots,\cos(d\pi X),\sin(2\pi X),\dots,\sin(d\pi X)\big)^T.$$

In TS(n,d), the input-output pair is such that

$$Y = 20X^2 - 10X - \frac{5}{3} + \sigma W,$$

with $\sigma = 10$. One can check that this implies

$$\theta^* = \left(\frac{20}{\pi^2},\dots,\frac{20}{\pi^2(\frac{d}{2})^2}, -\frac{10}{\pi},\dots,-\frac{10}{\pi(\frac{d}{2})}\right)^T \in \mathbb{R}^d.$$

3.4.5. Experiments.

Choice of the parameters and implementation details. Our min-max truncated algorithm has two parameters α and λ. In the subsequent experiments, we set the ridge parameter λ to the natural default choice for it: $\lambda = 0$. For the truncation parameter α, according to our analysis (see (3.9)), it roughly should be of order $1/\sigma^2$ up to kurtosis coefficients. By using the ordinary least squares estimator, we roughly estimate this value, and test values of α in a geometric grid (of 8 points) around it (with ratio 3). Cross-validation can be used to select the final α. Nevertheless, it is computationally expensive and is significantly outperformed in our experiments by the following simple procedure: start with the smallest α in the geometric grid and increase it as long as $\hat{\theta} = \theta_1$, that is as long as we stop at the end of the first iteration and output the empirical risk minimizer.

To compute $\theta_{k,1}(\eta)$ or $\theta_{k,2}(\eta)$, one needs to determine a least squares estimate (for a modified sample). To reduce the computational burden, we do not want to test all possible values of η (note that there are at most n values leading to different estimates). Our experiments show that testing only three levels of η is sufficient. Precisely, we sort the quantity

$$\overline{L}_i(\widehat{\theta}_k) X_i^T \widehat{Q}_k^{-1} X_i \left(1 + \sqrt{1 + \left[\overline{L}_i(\widehat{\theta}_k) \right]^{-1}} \right)^2$$

by decreasing order and consider η being the first, 5-th and 25-th value of the ordered list. Overall, in our experiments, the computational complexity is approximately fifty times larger than the one of computing the ordinary least squares estimator.

Results. The tables and figures have been gathered in Appendix E. Tables 1 and 2 give the results for the mixture noise. Tables 3, 4 and 5 provide the results for the heavy-tailed noise and the standard Gaussian noise. Each line of the tables has been obtained after 1000 generations of the training set. These results show that the min-max truncated estimator is often equal to $\hat{f}^{(\mathrm{erm})}$, while it ensures impressive consistent improvements when it differs from $\hat{f}^{(\mathrm{erm})}$. In this latter case, the number of points that are not considered in $\hat{f}$, i.e. the number of points with low signal to noise ratio, varies a lot from 1 to 150 and is often of order 30. Note that not only the points that we expect to be considered as outliers (i.e. very large output points) are erased, and that these points seem to be taken out by local groups: see Figures 1 and 2 in which the erased points are marked by surrounding circles.

Besides, the heavier the noise tail is (and also the larger the variance of the noise is), the more often the truncation modifies the initial ordinary least squares estimator, and the more improvements we get from the min-max truncated estimator, which also becomes much more robust than the ordinary least squares estimator (see the confidence intervals in the tables).

4. A SIMPLE TIGHT RISK BOUND FOR A SOPHISTICATED PAC-BAYES ALGORITHM

A disadvantage of the min-max estimator proposed in the previous section is that its theoretical guarantee depends on kurtosis like coefficients. In this section, we provide a more sophisticated estimator, having a simple theoretical excess risk bound, which is independent of these kurtosis like quantities when we assume L_∞-boundedness of the set $\mathcal{F}$.

We consider that the set Θ is bounded so that we can define the "prior" distribution π as the uniform distribution on $\mathcal{F}$ (i.e., the one induced by the Lebesgue distribution on $\Theta \subset \mathbb{R}^d$ renormalized to get $\pi(\mathcal{F}) = 1$). Let $\lambda > 0$ and

$$W_i(f, f') = \lambda\big\{\big[Y_i - f(X_i)\big]^2 - \big[Y_i - f'(X_i)\big]^2\big\}.$$

Introduce

$$\hat{\mathcal{E}}(f) = \log \int \frac{\pi(df')}{\prod_{i=1}^n [1 - W_i(f, f') + \frac{1}{2} W_i(f, f')^2]}. \tag{4.1}$$

We consider the "posterior" distribution $\hat{\pi}$ on the set $\mathcal{F}$ with density:

$$\frac{d\hat{\pi}}{d\pi}(f) = \frac{\exp[-\hat{\mathcal{E}}(f)]}{\int \exp[-\hat{\mathcal{E}}(f')]\pi(df')}. \tag{4.2}$$

To understand intuitively why this distribution concentrates on functions with low risk, one should think that when λ is small enough, $1 - W_i(f, f') + \frac{1}{2} W_i(f, f')^2$ is close to $e^{-W_i(f,f')}$, and consequently

$$\hat{\mathcal{E}}(f) \approx \lambda \sum_{i=1}^n [Y_i - f(X_i)]^2 + \log \int \pi(df') \exp\Big\{-\lambda \sum_{i=1}^n \big[Y_i - f'(X_i)\big]^2\Big\},$$

and

$$\frac{d\hat{\pi}}{d\pi}(f) \approx \frac{\exp\{-\lambda \sum_{i=1}^n [Y_i - f(X_i)]^2\}}{\int \exp\{-\lambda \sum_{i=1}^n [Y_i - f'(X_i)]^2\}\pi(df')}\,.$$

The following theorem gives a d/n convergence rate for the randomized algorithm which draws the prediction function from $\mathcal{F}$ according to the distribution $\hat{\pi}$.

THEOREM 4.1 *Assume that $\mathcal{F}$ has a diameter H for L^∞-norm:*

$$\sup_{f_1, f_2 \in \mathcal{F}, x \in \mathcal{X}} |f_1(x) - f_2(x)| = H \tag{4.3}$$

and that, for some $\sigma > 0$,

$$\sup_{x \in \mathcal{X}} \mathbb{E}\big\{[Y - f^*(X)]^2 \big| X = x\big\} \le \sigma^2 < +\infty. \tag{4.4}$$

Let $\hat{f}$ be a prediction function drawn from the distribution $\hat{\pi}$ defined in (4.2, page 25) *and depending on the parameter $\lambda > 0$. Then for any $0 < \eta' < 1 - \lambda(2\sigma + H)^2$ and $\varepsilon > 0$, with probability (with respect to the distribution $P^{\otimes n}\hat{\pi}$ generating the observations $Z_1, \dots, Z_n$ and the randomized prediction function $\hat{f}$) at least $1 - \varepsilon$, we have*

$$R(\hat{f}) - R(f^*) \leq (2\sigma + H)^2 \, \frac{C_1 d + C_2 \log(2\varepsilon^{-1})}{n}$$

with

$$C_1 = \frac{\log(\frac{(1+\eta)^2}{\eta'(1-\eta)})}{\eta(1 - \eta - \eta')} \quad \text{and} \quad C_2 = \frac{2}{\eta(1 - \eta - \eta')} \quad \text{and} \quad \eta = \lambda(2\sigma + H)^2.$$

In particular for $\lambda = 0.32(2\sigma + H)^{-2}$ and $\eta' = 0.18$, we get

$$R(\hat{f}) - R(f^*) \leq (2\sigma + H)^2 \, \frac{16.6\, d + 12.5 \log(2\varepsilon^{-1})}{n}.$$

Besides if $f^* \in \operatorname{argmin}_{f \in \mathcal{F}_{\text{lin}}} R(f)$, then with probability at least $1 - \varepsilon$, we have

$$R(\hat{f}) - R(f^*) \leq (2\sigma + H)^2 \, \frac{8.3\, d + 12.5 \log(2\varepsilon^{-1})}{n}.$$

PROOF. This is a direct consequence of Theorem 5.5 (page 33), Lemma 5.3 (page 31) and Lemma 5.6 (page 35). □

If we know that f^*_{lin} belongs to some bounded ball in $\mathcal{F}_{\text{lin}}$, then one can define a bounded $\mathcal{F}$ as this ball, use the previous theorem and obtain an excess risk bound with respect to f^*_{lin}.

REMARK 4.1 Let us discuss this result. On the positive side, we have a d/n convergence rate in expectation and in deviations. It has no extra logarithmic factor. It does not require any particular assumption on the smallest eigenvalue of the covariance matrix. To achieve exponential deviations, a uniformly bounded second moment of the output knowing the input is surprisingly sufficient: we do not require the traditional exponential moment condition on the output. Appendix A (page 64) argues that the uniformly bounded conditional second moment assumption cannot be replaced with just a bounded second moment condition.

On the negative side, the estimator is rather complicated. When the target is to predict as well as the best linear combination f^*_{lin} up to a small additive term, it requires the knowledge of a L^∞-bounded ball in which f^*_{lin} lies and an upper bound on $\sup_{x \in \mathcal{X}} \mathbb{E}\big\{[Y - f^*_{\text{lin}}(X)]^2 \big| X = x\big\}$. The looser this knowledge is, the bigger the constant in front of d/n is.

Finally, we propose a randomized algorithm consisting in drawing the prediction function according to $\hat{\pi}$. As usual, by convexity of the loss function,

the risk of the deterministic estimator $\hat{f}_{\text{determ}} = \int f \hat{\pi}(df)$ satisfies $R(\hat{f}_{\text{determ}}) \leq \int R(f)\hat{\pi}(df)$, so that, after some pretty standard computations, one can prove that for any $\varepsilon > 0$, with probability at least $1 - \varepsilon$:

$$R(\hat{f}_{\text{determ}}) - R(f^*_{\text{lin}}) \leq \kappa(2\sigma + H)^2 \frac{d + \log(\varepsilon^{-1})}{n},$$

for some appropriate numerical constant $\kappa > 0$.

REMARK 4.2 The previous result was expressing boundedness in terms of the L^∞ diameter of the set of functions $\mathcal{F}$. By using Lemma 5.7 (page 35) instead of Lemma 5.6 (page 35), Theorem 4.1 still holds without assuming (4.3) and (4.4), but by replacing $(2\sigma + H)^2$ by

$$V = \Bigg[2 \sqrt{\sup_{f \in \mathcal{F}_{\text{lin}} : \mathbb{E}[f(X)^2]=1} \mathbb{E}\big(f^2(X)[Y - f^*(X)]^2\big)} + \sqrt{\sup_{f', f'' \in \mathcal{F}} \mathbb{E}\big([f'(X) - f''(X)]^2\big)} \sqrt{\sup_{f \in \mathcal{F}_{\text{lin}} : \mathbb{E}[f(X)^2]=1} \mathbb{E}\big[f^4(X)\big]} \Bigg]^2.$$

The quantity V is finite when simultaneously, Θ is bounded, and for any j in $\{1, \ldots, d\}$, the quantities $\mathbb{E}\big[\varphi_j^4(X)\big]$ and $\mathbb{E}\big\{\varphi_j(X)^2[Y - f^*(X)]^2\big\}$ are finite.

5. A GENERIC LOCALIZED PAC-BAYES APPROACH

5.1. NOTATION AND SETTING. In this section, we drop the restrictions of the linear least squares setting considered in the other sections in order to focus on the ideas underlying the estimator and the results presented in Section 4. To do this, we consider that the loss incurred by predicting y' while the correct output is y is $\tilde{\ell}(y, y')$ (and is not necessarily equal to $(y - y')^2$). The quality of a (prediction) function $f : \mathcal{X} \to \mathbb{R}$ is measured by its risk

$$R(f) = \mathbb{E}\big\{\tilde{\ell}\big[Y, f(X)\big]\big\}.$$

We still consider the problem of predicting (at least) as well as the best function in a given set of functions $\mathcal{F}$ (but $\mathcal{F}$ is not necessarily a subset of a finite dimensional linear space). Let f^* still denote a function minimizing the risk among functions in $\mathcal{F}$: $f^* \in \operatorname{argmin}_{f \in \mathcal{F}} R(f)$. For simplicity, we assume that it exists. The excess risk is defined by

$$\bar{R}(f) = R(f) - R(f^*).$$

Let $\ell : \mathcal{Z} \times \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ be a function such that $\ell(Z, f, f')$ represents[5] how worse f predicts than f' on the data Z. Let us introduce the real-valued random processes $L : (f, f') \mapsto \ell(Z, f, f')$ and $L_i : (f, f') \mapsto \ell(Z_i, f, f')$, where $Z, Z_1, \ldots, Z_n$ denote i.i.d. random variables with distribution P.

Let π and π^* be two (prior) probability distributions on $\mathcal{F}$. We assume the following integrability condition.

Condition I. For any $f \in \mathcal{F}$, we have

$$\int \mathbb{E}\big\{\exp[L(f, f')]\big\}^n \pi^*(df') < +\infty, \tag{5.1}$$

$$\text{and} \quad \int \frac{\pi(df)}{\int \mathbb{E}\big\{\exp[L(f, f')]\big\}^n \pi^*(df')} < +\infty. \tag{5.2}$$

We consider the real-valued processes

$$\hat{L}(f, f') = \sum_{i=1}^{n} L_i(f, f'), \tag{5.3}$$

$$\hat{\mathcal{E}}(f) = \log \int \exp[\hat{L}(f, f')]\pi^*(df'), \tag{5.4}$$

$$L^\flat(f, f') = -n \log\big\{\mathbb{E}\big[\exp[-L(f, f')]\big]\big\}, \tag{5.5}$$

$$L^\sharp(f, f') = n \log\big\{\mathbb{E}\big[\exp[L(f, f')]\big]\big\}, \tag{5.6}$$

$$\text{and} \quad \mathcal{E}^\sharp(f) = \log\Big\{\int \exp\big[L^\sharp(f, f')\big]\pi^*(df')\Big\}. \tag{5.7}$$

Essentially, the quantities $\hat{L}(f, f')$, $L^\flat(f, f')$ and $L^\sharp(f, f')$ represent how worse is the prediction from f than from f' with respect to the training data or in expectation. By Jensen's inequality, we have

$$L^\flat \leq n\mathbb{E}(L) = \mathbb{E}(\hat{L}) \leq L^\sharp. \tag{5.8}$$

The quantities $\hat{\mathcal{E}}(f)$ and $\mathcal{E}^\sharp(f)$ should be understood as some kind of (empirical or expected) excess risk of the prediction function f with respect to an implicit reference induced by the integral over $\mathcal{F}$.

For a distribution ρ on $\mathcal{F}$ absolutely continuous w.r.t. π, let $\dfrac{d\rho}{d\pi}$ denote the density of ρ w.r.t. π. For any real-valued (measurable) function h defined on $\mathcal{F}$

[5]While the natural choice in the least squares setting is $\ell((X, Y), f, f') = [Y - f(X)]^2 - [Y - f'(X)]^2$, we will see that for heavy-tailed outputs, it is preferable to consider the following soft-truncated version of it, up to a scaling factor $\lambda > 0$: $\ell((X, Y), f, f') = T\big(\lambda\big[(Y - f(X))^2 - (Y - f'(X))^2\big]\big)$, with $T(x) = -\log(1 - x + x^2/2)$. Equality (5.4, page 28) corresponds to (4.1, page 25) with this choice of function ℓ and for the choice $\pi^* = \pi$.

such that $\int \exp[h(f)]\pi(df) < +\infty$, we define the distribution π_h on $\mathcal{F}$ by its density:

$$\frac{d\pi_h}{d\pi}(f) = \frac{\exp[h(f)]}{\int \exp[h(f')]\pi(df')}. \tag{5.9}$$

We will use the posterior distribution:

$$\frac{d\hat{\pi}}{d\pi}(f) = \frac{d\pi_{-\hat{\mathcal{E}}}}{d\pi}(f) = \frac{\exp[-\hat{\mathcal{E}}(f)]}{\int \exp[-\hat{\mathcal{E}}(f')]\pi(df')}. \tag{5.10}$$

Finally, for any $\beta \geq 0$, we will use the following measures of the size (or complexity) of $\mathcal{F}$ around the target function:

$$\mathcal{I}^*(\beta) = -\log\Big\{\int \exp\big[-\beta\bar{R}(f)\big]\pi^*(df)\Big\}$$

and

$$\mathcal{I}(\beta) = -\log\Big\{\int \exp\big[-\beta\bar{R}(f)\big]\pi(df)\Big\}.$$

5.2. THE LOCALIZED PAC-BAYES BOUND. With the notation introduced in the previous section, we have the following risk bound for any randomized estimator.

THEOREM 5.1 *Assume that π, π^*, $\mathcal{F}$ and ℓ satisfy the integrability conditions* (5.1) *and* (5.2, page 28). *Let ρ be a (posterior) probability distribution on $\mathcal{F}$ admitting a density with respect to π depending on $Z_1, \ldots, Z_n$. Let $\hat{f}$ be a prediction function drawn from the distribution ρ. Then for any $\gamma \geq 0$, $\gamma^* \geq 0$ and $\varepsilon > 0$, with probability (with respect to the distribution $P^{\otimes n}\rho$ generating the observations $Z_1, \ldots, Z_n$ and the randomized prediction function $\hat{f}$) at least $1 - \varepsilon$:*

$$\begin{aligned}\int \big[L^\flat(\hat{f}, f) + \gamma^*\bar{R}(f)\big]\pi^*_{-\gamma^*\bar{R}}(df) - \gamma\bar{R}\big(\hat{f}\,\big) \\ \leq \mathcal{I}^*(\gamma^*) - \mathcal{I}(\gamma) - \log\Big\{\int \exp\big[-\mathcal{E}^\sharp(f)\big]\pi(df)\Big\} \\ + \log\Big[\frac{d\rho}{d\hat{\pi}}\big(\hat{f}\,\big)\Big] + 2\log(2\varepsilon^{-1}).\end{aligned} \tag{5.11}$$

PROOF. See Section 6.4 (page 57). □

Some extra work will be needed to prove that Inequality (5.11) provides an upper bound on the excess risk $\bar{R}(\hat{f})$ of the estimator $\hat{f}$. As we will see in the next sections, despite the $-\gamma\bar{R}(\hat{f})$ term and provided that γ is sufficiently small, the lefthand-side will be essentially lower bounded by $\lambda\bar{R}(\hat{f})$ with $\lambda > 0$, while, by choosing $\rho = \hat{\pi}$, the estimator does not appear in the righthand-side.

5.3. APPLICATION UNDER AN EXPONENTIAL MOMENT CONDITION. The estimator proposed in Section 4 and Theorem 5.1 seems rather unnatural (or at least complicated) at first sight. The goal of this section is twofold. First it shows that under exponential moment conditions (i.e., stronger assumptions than the ones in Theorem 4.1 when the linear least square setting is considered), one can have a much simpler estimator than the one consisting in drawing a function according to the distribution (4.2) with $\hat{\mathcal{E}}$ given by (4.1) and yet still obtain a d/n convergence rate. Secondly it illustrates Theorem 5.1 in a different and simpler way than the one we will use to prove Theorem 4.1.

In this section, we consider the following variance and complexity assumptions.

Condition V1. There exist $\lambda > 0$ and $0 < \eta < 1$ such that for any function $f \in \mathcal{F}$, we have $\mathbb{E}\Big\{\exp\big\{\lambda\,\tilde{\ell}\big[Y, f(X)\big]\big\}\Big\} < +\infty$,

$$\log\Big\{\mathbb{E}\Big\{\exp\Big\{\lambda\,\Big[\tilde{\ell}\big[Y, f(X)\big] - \tilde{\ell}\big[Y, f^*(X)\big]\Big]\Big\}\Big\}\Big\} \leq \lambda(1+\eta)[R(f) - R(f^*)],$$

$$\text{and } \log\Big\{\mathbb{E}\Big\{\exp\Big\{-\lambda\Big[\tilde{\ell}\big[Y, f(X)\big] - \tilde{\ell}\big[Y, f^*(X)\big]\Big]\Big\}\Big\}\Big\} \leq -\lambda(1-\eta)[R(f) - R(f^*)].$$

Condition C. There exist a probability distribution π, and constants $D > 0$ and $G > 0$ such that for any $0 < \alpha < \beta$,

$$\log\left(\frac{\int \exp\{-\alpha[R(f) - R(f^*)]\}\pi(df)}{\int \exp\{-\beta[R(f) - R(f^*)]\}\pi(df)}\right) \leq D\log\left(\frac{G\beta}{\alpha}\right).$$

THEOREM 5.2 *Assume that* V1 *and* C *are satisfied. Let* $\hat{\pi}^{(\text{Gibbs})}$ *be the probability distribution on* $\mathcal{F}$ *defined by its density*

$$\frac{d\hat{\pi}^{(\text{Gibbs})}}{d\pi}(f) = \frac{\exp\{-\lambda\sum_{i=1}^n \tilde{\ell}[Y_i, f(X_i)]\}}{\int \exp\{-\lambda\sum_{i=1}^n \tilde{\ell}[Y_i, f'(X_i)]\}\pi(df')},$$

where $\lambda > 0$ *and the distribution* π *are those appearing respectively in* V1 *and* C*. Let* $\hat{f} \in \mathcal{F}$ *be a function drawn according to this Gibbs distribution. Then for any* η' *such that* $0 < \eta' < 1 - \eta$ *(where* η *is the constant appearing in* V1*) and any* $\varepsilon > 0$*, with probability at least* $1 - \varepsilon$*, we have*

$$R(\hat{f}) - R(f^*) \leq \frac{C_1' D + C_2' \log(2\varepsilon^{-1})}{n}$$

with

$$C_1' = \frac{\log(\frac{G(1+\eta)}{\eta'})}{\lambda(1-\eta-\eta')} \quad \textit{and} \quad C_2' = \frac{2}{\lambda(1-\eta-\eta')}.$$

PROOF. We consider $\ell\big[(X,Y),f,f'\big] = \lambda\big\{\tilde{\ell}\big[Y,f(X)\big] - \tilde{\ell}\big[Y,f'(X)\big]\big\}$, where λ is the constant appearing in the variance assumption. Let us take $\gamma^* = 0$ and let π^* be the Dirac distribution at f^*: $\pi^*(\{f^*\}) = 1$. Then Condition V1 implies Condition I (page 28) and we can apply Theorem 5.1. We have

$$\begin{aligned}
L(f,f') &= \lambda\big\{\tilde{\ell}\big[Y,f(X)\big] - \tilde{\ell}\big[Y,f'(X)\big]\big\},\\
\hat{\mathcal{E}}(f) &= \lambda\sum_{i=1}^{n}\tilde{\ell}\big[Y_i,f(X_i)\big] - \lambda\sum_{i=1}^{n}\tilde{\ell}\big[Y_i,f^*(X_i)\big],\\
\hat{\pi} &= \hat{\pi}^{(\text{Gibbs})},\\
L^\flat(f) &= -n\log\Big\{\mathbb{E}\Big[\exp\big[-L(f,f^*)\big]\Big]\Big\},\\
\mathcal{E}^\sharp(f) &= n\log\Big\{\mathbb{E}\Big[\exp\big[L(f,f^*)\big]\Big]\Big\}
\end{aligned}$$

and Assumption V1 leads to:

$$\begin{aligned}
\log\Big\{\mathbb{E}\Big[\exp\big[L(f,f^*)\big]\Big]\Big\} &\le \lambda(1+\eta)[R(f) - R(f^*)]\\
\text{and } \log\Big\{\mathbb{E}\Big[\exp\big[-L(f,f^*)\big]\Big]\Big\} &\le -\lambda(1-\eta)[R(f) - R(f^*)].
\end{aligned}$$

Thus choosing $\rho = \hat{\pi}$, (5.11) gives

$$[\lambda n(1-\eta) - \gamma]\bar{R}(\hat{f}) \le -\mathcal{I}(\gamma) + \mathcal{I}\big[\lambda n(1+\eta)\big] + 2\log(2\varepsilon^{-1}).$$

Accordingly by the complexity assumption, for $\gamma \le \lambda n(1+\eta)$, we get

$$[\lambda n(1-\eta) - \gamma]\bar{R}(\hat{f}) \le D\log\left(\frac{G\lambda n(1+\eta)}{\gamma}\right) + 2\log(2\varepsilon^{-1}),$$

which implies the announced result. □

Let us conclude this section by mentioning settings in which assumptions V1 and C are satisfied.

LEMMA 5.3 *Let Θ be a bounded convex set of $\mathbb{R}^d$, and $\varphi_1,\dots,\varphi_d$ be d square integrable prediction functions. Assume that*

$$\mathcal{F} = \Big\{f_\theta = \sum_{j=1}^{d}\theta_j\varphi_j;\,(\theta_1,\dots,\theta_d)\in\Theta\Big\},$$

π is the uniform distribution on $\mathcal{F}$ (i.e., the one coming from the uniform distribution on Θ), and that there exist $0 < b_1 \le b_2$ such that for any $y\in\mathbb{R}$, the function $\tilde{\ell}_y : y' \mapsto \tilde{\ell}(y,y')$ admits a second derivative satisfying: for any $y'\in\mathbb{R}$,

$$b_1 \le \tilde{\ell}''_y(y') \le b_2.$$

Then Condition C *holds for the above uniform* π, $G = \sqrt{b_2/b_1}$ *and* $D = d$.

Besides when $f^* = f^*_{\text{lin}}$ *(i.e.,* $\min_{\mathcal{F}} R = \min_{\theta\in\mathbb{R}^d} R(f_\theta)$*), Condition* C *holds for the above uniform* π, $G = b_2/b_1$ *and* $D = d/2$.

PROOF. See Section 6.5 (page 61). □

REMARK 5.1 In particular, for the least squares loss $\tilde{\ell}(y, y') = (y-y')^2$, we have $b_1 = b_2 = 2$ so that condition C holds with π the uniform distribution on $\mathcal{F}$, $D = d$ and $G = 1$, and with $D = d/2$ and $G = 1$ when $f^* = f^*_{\text{lin}}$.

LEMMA 5.4 *Assume that there exist* $0 < b_1 \le b_2$, $A > 0$ *and* $M > 0$ *such that for any* $y \in \mathbb{R}$*, the functions* $\tilde{\ell}_y : y' \mapsto \tilde{\ell}(y, y')$ *are twice differentiable and satisfy:*

$$\text{for any } y' \in \mathbb{R}, \qquad b_1 \le \tilde{\ell}''_y(y') \le b_2, \tag{5.12}$$

$$\text{and for any } x \in \mathcal{X}, \quad \mathbb{E}\Big\{\exp\Big[A^{-1}\big|\tilde{\ell}'_Y[f^*(X)]\big|\Big]\,\Big|\,X = x\Big\} \le M. \tag{5.13}$$

Assume that $\mathcal{F}$ *is convex and has a diameter* H *for* L^∞*-norm:*

$$\sup_{f_1,f_2\in\mathcal{F},x\in\mathcal{X}} |f_1(x) - f_2(x)| = H.$$

In this case Condition V1 *holds for any* (λ, η) *such that*

$$\eta \ge \frac{\lambda A^2}{2b_1}\exp\Big[M^2\exp\big(Hb_2/A\big)\Big].$$

and $0 < \lambda \le (2AH)^{-1}$ *is small enough to ensure* $\eta < 1$.

PROOF. See Section 6.6 (page 62). □

5.4. APPLICATION WITHOUT EXPONENTIAL MOMENT CONDITION. When we do not have finite exponential moments as assumed by Condition V1 (page 30), e.g., when $\mathbb{E}\big\{\exp\big\{\lambda\big\{\tilde{\ell}[Y, f(X)] - \tilde{\ell}[Y, f^*(X)]\big\}\big\}\big\} = +\infty$ for any $\lambda > 0$ and some function f in $\mathcal{F}$, we cannot apply Theorem 5.1 with $\ell\big[(X, Y), f, f'\big] = \lambda\big\{\tilde{\ell}\big[Y, f(X)\big] - \tilde{\ell}\big[Y, f'(X)\big]\big\}$ (because of the $\mathcal{E}^\sharp$ term). However, we can apply it to the soft truncated excess loss

$$\ell\big[(X, Y), f, f'\big] = T\Big(\lambda\big\{\tilde{\ell}\big[Y, f(X)\big] - \tilde{\ell}\big[Y, f'(X)\big]\big\}\Big),$$

with $T(x) = -\log(1-x+x^2/2)$. This section provides a result similar to Theorem 5.2 in which condition V1 is replaced by the following condition.

Condition V2. For any function f, the random variable $\tilde{\ell}\big[Y, f(X)\big] - \tilde{\ell}\big[Y, f^*(X)\big]$ is square integrable and there exists $V > 0$ such that for any function f,

$$\mathbb{E}\Big\{\Big[\tilde{\ell}\big[Y, f(X)\big] - \tilde{\ell}\big[Y, f^*(X)\big]\Big]^2\Big\} \le V[R(f) - R(f^*)].$$

THEOREM 5.5 *Assume that Conditions* V2 *above and* C *(page 30) are satisfied. Let* $0 < \lambda < V^{-1}$ *and*

$$\ell\big[(X,Y),f,f'\big] = T\Big(\lambda\big\{\tilde{\ell}\big[Y,f(X)\big] - \tilde{\ell}\big[Y,f'(X)\big]\big\}\Big), \tag{5.14}$$

with

$$T(x) = -\log(1 - x + x^2/2). \tag{5.15}$$

Let $\hat{f} \in \mathcal{F}$ *be a function drawn according to the distribution* $\hat{\pi}$ *defined in* (5.10, page 29) *with* $\hat{\mathcal{E}}$ *defined in* (5.4, page 28) *and* $\pi^* = \pi$ *the distribution appearing in Condition* C. *Then for any* $0 < \eta' < 1 - \lambda V$ *and* $\varepsilon > 0$, *with probability at least* $1 - \varepsilon$, *we have*

$$R(\hat{f}) - R(f^*) \le V\frac{C_1' D + C_2' \log(2\varepsilon^{-1})}{n}$$

with

$$C_1' = \frac{\log(\frac{G(1+\eta)^2}{\eta'(1-\eta)})}{\eta(1-\eta-\eta')} \quad \textit{and} \quad C_2' = \frac{2}{\eta(1-\eta-\eta')} \quad \textit{and} \quad \eta = \lambda V.$$

In particular, for $\lambda = 0.32V^{-1}$ *and* $\eta' = 0.18$, *we get*

$$R(\hat{f}) - R(f^*) \le V\frac{16.6D + 12.5\log(2\sqrt{G}\varepsilon^{-1})}{n}.$$

PROOF. We apply Theorem 5.1 for ℓ given by (5.14) and $\pi^* = \pi$. Let

$$W(f,f') = \lambda\big\{\tilde{\ell}\big[Y,f(X)\big] - \tilde{\ell}\big[Y,f'(X)\big]\big\} \qquad \text{for any } f, f' \in \mathcal{F}.$$

Since $\log u \le u - 1$ for any $u > 0$, we have

$$L^\flat = -n\log\mathbb{E}(1 - W + W^2/2) \ge n(\mathbb{E}W - \mathbb{E}W^2/2).$$

Moreover, from Assumption V2,

$$\frac{\mathbb{E}W(f,f')^2}{2} \le \mathbb{E}W(f,f^*)^2 + \mathbb{E}W(f',f^*)^2 \le \lambda^2 V\bar{R}(f) + \lambda^2 V\bar{R}(f'), \tag{5.16}$$

hence, by introducing $\eta = \lambda V$,

$$\begin{aligned} L^\flat(f,f') &\ge \lambda n\big[\bar{R}(f) - \bar{R}(f') - \lambda V\bar{R}(f) - \lambda V\bar{R}(f')\big] \\ &= \lambda n\big[(1-\eta)\bar{R}(f) - (1+\eta)\bar{R}(f')\big]. \end{aligned} \tag{5.17}$$

Noting that

$$\exp\bigl[T(u)\bigr] = \frac{1}{1-u+u^2/2} = \frac{1+u+\frac{u^2}{2}}{\bigl(1+\frac{u^2}{2}\bigr)^2-u^2} = \frac{1+u+\frac{u^2}{2}}{1+\frac{u^4}{4}} \le 1+u+\frac{u^2}{2},$$

we see that

$$L^\sharp = n\log\Bigl\{\mathbb{E}\Bigl[\exp\bigl[T(W)\bigr]\Bigr]\Bigr\} \le n\bigl[\mathbb{E}\bigl(W\bigr)+\mathbb{E}\bigl(W^2\bigr)/2\bigr].$$

Using (5.16) and still $\eta = \lambda V$, we get

$$\begin{aligned} L^\sharp(f,f') &\le \lambda n\bigl[\bar R(f)-\bar R(f')+\eta\bar R(f)+\eta\bar R(f')\bigr] \\ &= \lambda n(1+\eta)\bar R(f)-\lambda n(1-\eta)\bar R(f'), \end{aligned}$$

and

$$\mathcal{E}^\sharp(f) \le \lambda n(1+\eta)\bar R(f) - \mathcal{I}(\lambda n(1-\eta)). \tag{5.18}$$

Plugging (5.17) and (5.18) in (5.11) for $\rho = \hat\pi$, we obtain

$$\begin{aligned} &[\lambda n(1-\eta)-\gamma]\bar R(\hat f) + [\gamma^* - \lambda n(1+\eta)]\textstyle\int \bar R(f)\pi_{-\gamma^*\bar R}(df) \\ &\qquad \le \mathcal{I}(\gamma^*) - \mathcal{I}(\gamma) + \mathcal{I}(\lambda n(1+\eta)) - \mathcal{I}(\lambda n(1-\eta)) + 2\log(2\varepsilon^{-1}). \end{aligned}$$

By the complexity assumption, choosing $\gamma^* = \lambda n(1+\eta)$ and $\gamma < \lambda n(1-\eta)$, we get

$$[\lambda n(1-\eta)-\gamma]\bar R(\hat f) \le D\log\left(G\frac{\lambda n(1+\eta)^2}{\gamma(1-\eta)}\right) + 2\log(2\varepsilon^{-1}),$$

hence the desired result by considering $\gamma = \lambda n\eta'$ with $\eta' < 1-\eta$. □

REMARK 5.2 The estimator seems abnormally complicated at first sight. This remark aims at explaining why we were not able to consider a simpler estimator.

In Section 5.3, in which we consider the exponential moment condition V1, we took $\ell\bigl[(X,Y),f,f'\bigr] = \lambda\bigl\{\tilde\ell\bigl[Y,f(X)\bigr] - \tilde\ell\bigl[Y,f'(X)\bigr]\bigr\}$ and π^* as the Dirac distribution at f^*. For these choices, one can easily check that $\hat\pi$ does not depend on f^*.

In the absence of an exponential moment condition, we cannot consider the function $\ell\bigl[(X,Y),f,f'\bigr] = \lambda\bigl\{\tilde\ell\bigl[Y,f(X)\bigr] - \tilde\ell\bigl[Y,f'(X)\bigr]\bigr\}$ but a truncated version of it. The truncation function T we use in Theorem 5.5 can be replaced by the simpler function $u \mapsto (u\vee -M)\wedge M$ for some appropriate constant $M>0$ but this would lead to a bound with worse constants, without really simplifying the algorithm. The precise choice $T(x) = -\log(1-x+x^2/2)$ comes from the remarkable property: there exist second order polynomial $P^\flat$ and $P^\sharp$ such that $\frac{1}{P^\flat(u)} \le \exp\bigl[T(u)\bigr] \le P^\sharp(u)$ and $P^\flat(u)P^\sharp(u) \le 1+\mathrm{O}(u^4)$ for $u\to 0$, which are

reasonable properties to ask in order to ensure that (5.8), and consequently (5.11), are tight.

Besides, if we take ℓ as in (5.14) with T a truncation function and π^* as the Dirac distribution at f^*, then $\hat{\pi}$ would depend on f^*, and is consequently not observable. This is the reason why we do not consider π^* as the Dirac distribution at f^*, but $\pi^* = \pi$. This lead to the estimator considered in Theorems 5.5 and 4.1.

REMARK 5.3 Theorem 5.5 still holds for the same randomized estimator in which (5.15, page 33) is replaced with

$$T(x) = \log(1 + x + x^2/2).$$

Condition V2 holds under weak assumptions as illustrated by the following lemma.

LEMMA 5.6 *Consider the least squares setting: $\tilde{\ell}(y, y') = (y - y')^2$. Assume that $\mathcal{F}$ is convex and has a diameter H for L^∞-norm:*

$$\sup_{f_1, f_2 \in \mathcal{F}, x \in \mathcal{X}} |f_1(x) - f_2(x)| = H$$

and that for some $\sigma > 0$, we have

$$\sup_{x \in \mathcal{X}} \mathbb{E}\big\{[Y - f^*(X)]^2 \big| X = x\big\} \leq \sigma^2 < +\infty. \tag{5.19}$$

Then Condition V2 *holds for $V = (2\sigma + H)^2$.*

PROOF. See Section 6.7 (page 63). □

LEMMA 5.7 *Consider the least squares setting: $\tilde{\ell}(y, y') = (y - y')^2$. Assume that $\mathcal{F}$ (i.e., Θ) is bounded, and that for any $j \in \{1, \ldots, d\}$, we have $\mathbb{E}\big[\varphi_j^4(X)\big] < +\infty$ and $\mathbb{E}\big\{\varphi_j(X)^2 [Y - f^*(X)]^2\big\} < +\infty$. Then Condition* V2 *holds for*

$$V = \Bigg[2 \sqrt{\sup_{f \in \mathcal{F}_{\text{lin}} : \mathbb{E}[f(X)^2] = 1} \mathbb{E}\big(f^2(X)[Y - f^*(X)]^2\big)} + \sqrt{\sup_{f', f'' \in \mathcal{F}} \mathbb{E}\big([f'(X) - f''(X)]^2\big)} \sqrt{\sup_{f \in \mathcal{F}_{\text{lin}} : \mathbb{E}[f(X)^2] = 1} \mathbb{E}\big[f^4(X)\big]} \Bigg]^2 .$$

PROOF. See Section 6.8 (page 64). □

6. PROOFS

6.1. MAIN IDEAS OF THE PROOFS. The goal of this section is to explain the key ingredients appearing in the proofs which both allows to obtain sub-exponential tails for the excess risk under a non-exponential moment assumption and get rid of the logarithmic factor in the excess risk bound.

6.1.1. Sub-exponential tails under a non-exponential moment assumption via truncation. Let us start with the idea allowing us to prove exponential inequalities under just a moment assumption (instead of the traditional exponential moment assumption). To understand it, we can consider the (apparently) simplistic 1-dimensional situation in which we have $\Theta = \mathbb{R}$ and the marginal distribution of $\varphi_1(X)$ is the Dirac distribution at 1. In this case, the risk of the prediction function f_θ is $R(f_\theta) = \mathbb{E}(Y-\theta)^2 = \mathbb{E}(Y-\theta^*)^2 + (\mathbb{E}Y - \theta)^2$, so that the least squares regression problem boils down to the estimation of the mean of the output variable. If we only assume that Y admits a finite second moment, say $\mathbb{E}Y^2 \leq 1$, it is not clear whether for any $\varepsilon > 0$, it is possible to find $\hat{\theta}$ such that with probability at least $1-2\varepsilon$,

$$R(f_{\hat{\theta}}) - R(f^*) = (\mathbb{E}(Y) - \hat{\theta})^2 \leq c\frac{\log(\varepsilon^{-1})}{n}, \tag{6.1}$$

for some numerical constant c. Indeed, from Chebyshev's inequality, the trivial choice $\hat{\theta} = \frac{\sum_{i=1}^n Y_i}{n}$ just satisfies: with probability at least $1-2\varepsilon$,

$$R(f_{\hat{\theta}}) - R(f^*) \leq \frac{1}{n\varepsilon},$$

which is far from the objective (6.1) for small confidence levels (consider $\varepsilon = \exp(-\sqrt{n})$ for instance). The key idea is thus to average (soft) *truncated* values of the outputs. This is performed by taking

$$\hat{\theta} = \frac{1}{n\lambda}\sum_{i=1}^n \log\left(1 + \lambda Y_i + \frac{\lambda^2 Y_i^2}{2}\right),$$

with $\lambda = \sqrt{\frac{2\log(\varepsilon^{-1})}{n}}$. Since we have

$$\log \mathbb{E}\exp(n\lambda\hat{\theta}) = n\log\left(1 + \lambda\mathbb{E}(Y) + \frac{\lambda^2}{2}\mathbb{E}(Y^2)\right) \leq n\lambda\mathbb{E}(Y) + n\frac{\lambda^2}{2},$$

the exponential Chebyshev's inequality (see Lemma 6.10) guarantees that with probability at least $1-\varepsilon$, we have $n\lambda(\hat{\theta} - \mathbb{E}(Y)) \leq n\frac{\lambda^2}{2} + \log(\varepsilon^{-1})$, hence

$$\hat{\theta} - \mathbb{E}(Y) \leq \sqrt{\frac{2\log(\varepsilon^{-1})}{n}}.$$

Replacing Y by $-Y$ in the previous argument, we obtain that with probability at least $1-\varepsilon$, we have

$$n\lambda\left\{\mathbb{E}(Y) + \frac{1}{n\lambda}\sum_{i=1}^n \log\left(1 - \lambda Y_i + \frac{\lambda^2 Y_i^2}{2}\right)\right\} \leq n\frac{\lambda^2}{2} + \log(\varepsilon^{-1}).$$

Since $-\log(1+x+x^2/2) \leq \log(1-x+x^2/2)$, this implies $\mathbb{E}(Y)-\hat{\theta} \leq \sqrt{\frac{2\log(\varepsilon^{-1})}{n}}$. The two previous inequalities imply Inequality (6.1) (for $c = 2$), showing that sub-exponential tails are achievable even when we only assume that the random variable admits a finite second moment (see [10] for more details on the robust estimation of the mean of a random variable).

6.1.2. Localized PAC-Bayesian inequalities to eliminate a logarithm factor.

High level description of the PAC-Bayesian approach and the localization argument. The analysis of statistical inference generally relies on upper bounding the supremum of an empirical process χ indexed by the functions in a model $\mathcal{F}$. One central tool to obtain these bounds is the concentration inequalities. An alternative approach, called the PAC-Bayesian one, consists in using the entropic equality

$$\mathbb{E}\exp\left(\sup_{\rho\in\mathcal{M}}\left\{\int\rho(df)\chi(f)-K(\rho,\pi')\right\}\right)=\int\pi'(df)\mathbb{E}\exp\big(\chi(f)\big). \quad (6.2)$$

where $\mathcal{M}$ is the set of probability distributions on $\mathcal{F}$ and $K(\rho,\pi')$ is the Kullback-Leibler divergence (whose definition is recalled in (6.29)) between ρ and some fixed distribution π'.

Let $\check{r}:\mathcal{F}\to\mathbb{R}$ be an observable process such that for any $f\in\mathcal{F}$, we have

$$\mathbb{E}\exp\big(\chi(f)\big)\leq 1$$

for $\chi(f)=\lambda[R(f)-\check{r}(f)]$ and some $\lambda>0$. Then (6.2) leads to: for any $\varepsilon>0$, with probability at least $1-\varepsilon$, for any distribution ρ on $\mathcal{F}$, we have

$$\int\rho(df)R(f)\leq\int\rho(df)\check{r}(f)+\frac{K(\rho,\pi')+\log(\varepsilon^{-1})}{\lambda}. \quad (6.3)$$

The lefthand-side quantity represents the expected risk with respect to the distribution ρ. To get the smallest upper bound on this quantity, a natural choice of the (posterior) distribution ρ is obtained by minimizing the righthand-side, that is by taking $\rho=\pi'_{-\lambda\check{r}}$ (with the notation introduced in (5.9)). This distribution concentrates on functions $f\in\mathcal{F}$ for which $\check{r}(f)$ is small. Without prior knowledge, one may want to choose a prior distribution $\pi'=\pi$ which is rather "flat" (e.g., the one induced by the Lebesgue measure in the case of a model $\mathcal{F}$ defined by a bounded parameter set in some Euclidean space). Consequently the Kullback-Leibler divergence $K(\rho,\pi')$, which should be seen as the complexity term, might be excessively large.

To overcome the lack of prior information and the resulting high complexity term, one can alternatively use a more "localized" prior distribution $\pi' = \pi_{-\beta R}$ for some $\beta > 0$. Since the righthand-side of (6.3) is then no longer observable, an empirical upper bound on $K(\rho, \pi_{-\beta R})$ is required. It is obtained by writing

$$K(\rho, \pi_{-\beta R}) = K(\rho, \pi) + \log\left(\int \pi(df) \exp[-\beta R(f)]\right) + \beta \int \rho(df) R(f),$$

and by controlling the two non-observable terms by their empirical versions, calling for additional PAC-Bayesian inequalities.

Low level description of localization. To simplify a more detailed presentation of the PAC-Bayesian localization argument, we will consider a setting in which $\mathcal{F}$, φ_1, ..., φ_d and the outputs are bounded almost surely, specifically assume $\mathbb{P}(\text{for any } f \in \mathcal{F}, |Y - f(X)| \leq 1) = 1$.

Introduce $\Psi(u) = [\exp(u) - 1 - u]/u^2$ for any $u > 0$, $\bar{R}(f) = R(f) - R(f^*)$ and $\bar{r}(f) = r(f) - r(f^*)$ for any $f \in \mathcal{F}$. Let π be a distribution on $\mathcal{F}$ and $\Delta(f, f') = \mathbb{E}\big\{[Y - f(X)]^2 - [Y - f^*(X)]^2\big\}^2$. The starting point is the following PAC-Bayesian inequality: for any $\varepsilon > 0$ and $\lambda > 0$, with probability at least $1 - \varepsilon$, for any distribution ρ on $\mathcal{F}$, we have

$$\int \rho(df) \bar{R}(f) \leq \int \rho(df) \bar{r}(f) + \frac{\lambda}{n} \Psi\Big(\frac{2\lambda}{n}\Big) \int \rho(df) \Delta(f, f^*) + \frac{K(\rho, \pi) + \log(\varepsilon^{-1})}{\lambda}. \quad (6.4)$$

This inequality derives from the duality formula given in (6.30), the inequality $\mathbb{E} \exp\Big(\frac{\lambda}{n}\big\{[Y - f^*(X)]^2 - [Y - f(X)]^2 + R(f) - R(f^*)\big\} - \frac{\lambda^2}{n^2}\Psi\big(\frac{2\lambda}{n}\big)\Delta(f, f^*)\Big) \leq 1$, and Lemma 6.10 (see [2, Theorem 8.1]). Since

$$\Delta(f, f^*) = \mathbb{E}\big\{[f(X) - f^*(X)]^2 [2Y - f(X) - f^*(X)]^2\big\} \leq 4\mathbb{E}\big\{[f(X) - f^*(X)]^2\big\} \leq 4\bar{R}(f),$$

by taking $\lambda = n/6$, Inequality (6.4) implies

$$\int \rho(df) \bar{R}(f) \leq 2 \int \rho(df) \bar{r}(f) + 10 \frac{K(\rho, \pi) + \log(\varepsilon^{-1})}{n}. \quad (6.5)$$

The distribution $\hat{\pi}(df) = \frac{\exp[-n\bar{r}(f)/5]}{\int \exp[-n\bar{r}(f')/5]\pi(df')} \cdot \pi(df)$ minimizes the righthand-side, and we have

$$\int \hat{\pi}(df) \bar{R}(f) \leq 10 \frac{-\log\big(\int \pi(df) \exp[-n\bar{r}(f)/5]\big) + \log(\varepsilon^{-1})}{n}.$$

Let π_U be the uniform distribution on $\mathcal{F}$ (i.e., the one coming from the uniform distribution on Θ). For $\pi = \pi_U$, using similar arguments to the ones developed in Section 6.5, it can be shown that $-\log\left(\int \pi(df)\exp[-n\bar{r}(f)/5]\right. \leq cd\log(n)$ for some constant c depending only on $\sup_{f,f'\in\mathcal{F}}\|f-f'\|_\infty$. This implies a $\frac{d\log n}{n}$ convergence rate of the excess risk of the randomized algorithm associated with $\hat{\pi}$.

The localization idea from [7] allows to prove

$$\int \rho(df)\bar{R}(f) \leq 2\int \rho(df)\bar{r}(f) + 10\frac{K(\rho,\hat{\pi}') + \log(\varepsilon^{-1})}{n}, \tag{6.6}$$

with $\hat{\pi}'(df) = \frac{\exp[-\zeta n\bar{r}(f)]}{\int \exp[-\zeta n\bar{r}(f')]\pi(df')}\cdot\pi(df)$ for some $0 < \zeta < 1/5$. The key difference with (6.5) is that the Kullback-Leibler term is now much smaller for the distributions ρ which concentrates on low empirical risk functions, like $\hat{\pi}$. Since $-\log\left(\int \hat{\pi}'(df)\exp[-n\bar{r}(f)/5]\right. \leq cd$ for some constant c depending only on ζ (see Lemma 5.3), this allows to get rid of the $\log n$ factor and obtain a convergence rate of order d/n.

The proof of (6.6) is rather intricate but the central idea is to use (6.5) for $\pi(df) = \frac{\exp[-n\bar{R}(f)/5]}{\int \exp[-n\bar{R}(f')/5]\pi(df')}\cdot\pi_U(df)$, and control the non-observable Kullback-Leibler term by $c\int\rho(df)\bar{R}(f)$ plus $K(\rho,\hat{\pi}')$ up to minor additive terms.

Let us conclude this section by pointing out some difficulties and possibilities when considering unbounded $Y - f_\theta(X)$. The sketches of proof presented hereafter are far from being actual proofs as some technical problems are hidden. Full proofs will be given in the later sections. For unbounded $Y - f_\theta(X)$, Inequality (6.4) no longer holds, but by using the soft truncation argument of the previous section, one can prove a similar inequality in which $\int\rho(df)\bar{r}(f)$ is replaced with $\frac{1}{\lambda}\int\rho(df)\sum_{i=1}^n\log\left(1+W_i(f,f^*)+W_i^2(f,f^*)/2\right)$ for $W_i(f,f^*) = \frac{\lambda}{n}\left\{[Y-f(X_i)]^2-[Y-f^*(X_i)]^2\right\}$ for $\lambda > 0$ a parameter of the bound. One significant difficulty is that the minimizer of this quantity is no longer observable (since f^* is unknown). Nevertheless the quantity can be upper bounded by the observable one:

$$\max_{f'\in\mathcal{F}}\frac{1}{\lambda}\int\rho(df)\sum_{i=1}^n\log\left(1+W_i(f,f')+\frac{W_i^2(f,f')}{2}\right).$$

This explains why the procedures in Section 3 make appear a min-max.

Another interesting idea is to use Gaussian distributions for π and ρ, which are respectively centered at θ^* and $\hat{\theta}$ and with covariance matrix proportional to the identity matrix. The interest of these choices comes essentially from the coexistence of the two following properties: the distribution π concentrates on a neighbourhood of the best prediction function so the complexity term $K(\rho,\pi)$ can be much smaller than the one obtained for π the uniform distribution on $\mathcal{F}$

(this is again the localization idea), and $K(\rho, \pi)$ and, when $\Theta = \mathbb{R}^d$, the integrals with respect to ρ can be explicitly computed in terms of $\bar{R}(\hat{\theta})$ and other rather simple quantities, which implies that the modified inequality (6.4) gets a tractable form for further computations, provided nevertheless some assumptions on the eigenvalues of the matrix Q. The idea of using PAC-Bayesian inequalities with Gaussian prior and posterior distributions has first been proposed by Langford and Shawe-Taylor [14] in the context of linear classification.

6.2. PROOFS OF THEOREMS 2.1 AND 2.2. To shorten the formulae, we will write X for $\varphi(X)$, which is equivalent to considering without loss of generality that the input space is $\mathbb{R}^d$ and that the functions $\varphi_1, \ldots, \varphi_d$ are the coordinate functions. Therefore, the function f_θ maps an input x to $\langle \theta, x \rangle$. With a slight abuse of notation, $R(\theta)$ will denote the risk of this prediction function.

Let us first assume that the matrix $Q_\lambda = Q + \lambda I$ is positive definite. This indeed does not restrict the generality of our study, even in the case when $\lambda = 0$, as we will discuss later (Remark 6.1). Consider the change of coordinates

$$\overline{X} = Q_\lambda^{-1/2} X.$$

Let us introduce

$$\overline{R}(\theta) = \mathbb{E}\big[(\langle \theta, \overline{X} \rangle - Y)^2\big],$$

so that

$$\overline{R}(Q_\lambda^{1/2}\theta) = R(\theta) = \mathbb{E}\big[(\langle \theta, X \rangle - Y)^2\big].$$

Let

$$\overline{\Theta} = \big\{ Q_\lambda^{1/2}\theta ; \theta \in \Theta \big\}.$$

Consider

$$r(\theta) = \frac{1}{n} \sum_{i=1}^n \big(\langle \theta, X_i \rangle - Y_i \big)^2, \tag{6.7}$$

$$\overline{r}(\theta) = \frac{1}{n} \sum_{i=1}^n \big(\langle \theta, \overline{X}_i \rangle - Y_i \big)^2, \tag{6.8}$$

$$\theta_0 = \arg\min_{\theta \in \overline{\Theta}} \overline{R}(\theta) + \lambda \|Q_\lambda^{-1/2}\theta\|^2, \tag{6.9}$$

$$\hat{\theta} \in \arg\min_{\theta \in \Theta} r(\theta) + \lambda \|\theta\|^2, \tag{6.10}$$

$$\theta_1 = Q_\lambda^{1/2}\hat{\theta} \in \arg\min_{\theta \in \overline{\Theta}} \overline{r}(\theta) + \lambda \|Q_\lambda^{-1/2}\theta\|^2. \tag{6.11}$$

For $\alpha > 0$, let us introduce the notation

$$W_i(\theta) = \alpha \Big\{ \big(\langle \theta, \overline{X}_i \rangle - Y_i \big)^2 - \big(\langle \theta_0, \overline{X}_i \rangle - Y_i \big)^2 \Big\},$$

$$W(\theta) = \alpha\Big\{\big(\langle\theta, \overline{X}\rangle - Y\big)^2 - \big(\langle\theta_0, \overline{X}\rangle - Y\big)^2\Big\}.$$

For any $\theta_2 \in \mathbb{R}^d$ and $\beta > 0$, let us consider the Gaussian distribution centered at θ_2

$$\rho_{\theta_2}(d\theta) = \left(\frac{\beta}{2\pi}\right)^{d/2} \exp\left(-\frac{\beta}{2}\|\theta - \theta_2\|^2\right) d\theta.$$

LEMMA 6.1 *For any $\eta > 0$ and $\alpha > 0$, with probability at least $1 - \exp(-\eta)$, for any $\theta_2 \in \mathbb{R}^d$,*

$$\begin{aligned} &- n\int \rho_{\theta_2}(d\theta) \log\Big\{1 - \mathbb{E}\big[W(\theta)\big] + \mathbb{E}\big[W(\theta)^2\big]/2\Big\} \\ &\qquad \le -\sum_{i=1}^n \left(\int \rho_{\theta_2}(d\theta) \log\Big\{1 - W_i(\theta) + W_i(\theta)^2/2\Big\}\right) + \mathcal{K}(\rho_{\theta_2}, \rho_{\theta_0}) + \eta, \end{aligned}$$

where $\mathcal{K}(\rho_{\theta_2}, \rho_{\theta_0})$ is the Kullback-Leibler divergence function :

$$\mathcal{K}(\rho_{\theta_2}, \rho_{\theta_0}) = \int \rho_{\theta_2}(d\theta) \log\left[\frac{d\rho_{\theta_2}}{d\rho_{\theta_0}}(\theta)\right].$$

PROOF.

$$\mathbb{E}\left(\int \rho_{\theta_0}(d\theta) \prod_{i=1}^n \frac{1 - W_i(\theta) + W_i(\theta)^2/2}{1 - \mathbb{E}\big[W(\theta)\big] + \mathbb{E}\big[W(\theta)^2\big]/2}\right) \le 1,$$

thus with probability at least $1 - \exp(-\eta)$

$$\log\left(\int \rho_{\theta_0}(d\theta) \prod_{i=1}^n \frac{1 - W_i(\theta) + W_i(\theta)^2/2}{1 - \mathbb{E}\big[W(\theta)\big] + \mathbb{E}\big[W(\theta)^2\big]/2}\right) \le \eta.$$

We conclude from the convex inequality (see [8, page 159])

$$\log\big(\int \rho_{\theta_0}(d\theta) \exp\big[h(\theta)\big]\big) \ge \int \rho_{\theta_2}(d\theta) h(\theta) - \mathcal{K}(\rho_{\theta_2}, \rho_{\theta_0}).$$

□

Let us compute some useful quantities

$$\mathcal{K}(\rho_{\theta_2}, \rho_{\theta_0}) = \frac{\beta}{2}\|\theta_2 - \theta_0\|^2, \tag{6.12}$$

$$\begin{aligned} \int \rho_{\theta_2}(d\theta)\big[W(\theta)\big] &= \alpha\int \rho_{\theta_2}(d\theta)\langle\theta - \theta_2, \overline{X}\rangle^2 + W(\theta_2) \\ &= W(\theta_2) + \alpha\frac{\|\overline{X}\|^2}{\beta}, \end{aligned} \tag{6.13}$$

$$\int \rho_{\theta_2}(d\theta)\langle \theta-\theta_2, \overline{X}\rangle^4 = \frac{3\|\overline{X}\|^4}{\beta^2}, \tag{6.14}$$

$$\begin{aligned}
\int \rho_{\theta_2}(d\theta)\big[W(\theta)^2\big] &= \alpha^2 \int \rho_{\theta_2}(d\theta)\langle \theta-\theta_0, \overline{X}\rangle^2 \big(\langle \theta+\theta_0, \overline{X}\rangle - 2Y\big)^2 \\
&= \alpha^2 \int \rho_{\theta_2}(d\theta)\Big[\langle \theta-\theta_2+\theta_2-\theta_0, \overline{X}\rangle \big(\langle \theta-\theta_2+\theta_2+\theta_0, \overline{X}\rangle - 2Y\big)\Big]^2 \\
&= \int \rho_{\theta_2}(d\theta)\Big[\alpha\langle \theta-\theta_2, \overline{X}\rangle^2 + 2\alpha\langle \theta-\theta_2, \overline{X}\rangle\big(\langle \theta_2, \overline{X}\rangle - Y\big) + W(\theta_2)\Big]^2 \\
&= \int \rho_{\theta_2}(d\theta)\Big[\alpha^2\langle \theta-\theta_2, \overline{X}\rangle^4 + 4\alpha^2\langle \theta-\theta_2, \overline{X}\rangle^2\big(\langle \theta_2, \overline{X}\rangle - Y\big)^2 + W(\theta_2)^2 \\
&\qquad + 2\alpha\langle \theta-\theta_2, \overline{X}\rangle^2 W(\theta_2)\Big] \\
&= \frac{3\alpha^2\|\overline{X}\|^4}{\beta^2} + \frac{2\alpha\|\overline{X}\|^2}{\beta}\Big[2\alpha\big(\langle \theta_2, \overline{X}\rangle - Y\big)^2 + W(\theta_2)\Big] + W(\theta_2)^2.
\end{aligned} \tag{6.15}$$

Using the fact that

$$2\alpha\big(\langle \theta_2, \overline{X}\rangle - Y\big)^2 + W(\theta_2) = 2\alpha\big(\langle \theta_0, \overline{X}\rangle - Y\big)^2 + 3W(\theta_2),$$

and that for any real numbers a and b, $6ab \le 9a^2 + b^2$, we get

LEMMA 6.2

$$\int \rho_{\theta_2}(d\theta)\big[W(\theta)\big] = W(\theta_2) + \alpha\frac{\|\overline{X}\|^2}{\beta}, \tag{6.16}$$

$$\begin{aligned}
\int \rho_{\theta_2}(d\theta)\big[W(\theta)^2\big] &= W(\theta_2)^2 + \frac{2\alpha\|\overline{X}\|^2}{\beta}\Big[2\alpha\big(\langle \theta_0, \overline{X}\rangle - Y\big)^2 + 3W(\theta_2)\Big] \\
&\qquad + \frac{3\alpha^2\|\overline{X}\|^4}{\beta^2} \qquad (6.17) \\
&\le 10W(\theta_2)^2 + \frac{4\alpha^2\|\overline{X}\|^2}{\beta}\big(\langle \theta_0, \overline{X}\rangle - Y\big)^2 + \frac{4\alpha^2\|\overline{X}\|^4}{\beta^2}, \qquad (6.18)
\end{aligned}$$

and the same holds true when W is replaced with W_i and $(\overline{X}, Y)$ with $(\overline{X}_i, Y_i)$.

Another important thing to realize is that

$$\begin{aligned}
\mathbb{E}\big[\|\overline{X}\|^2\big] &= \mathbb{E}\big[\mathrm{Tr}\big(\overline{X}\,\overline{X}^T\big)\big] &&= \mathbb{E}\big[\mathrm{Tr}\big(Q_\lambda^{-1/2} X X^T Q_\lambda^{-1/2}\big)\big] \\
&= \mathbb{E}\big[\mathrm{Tr}\big(Q_\lambda^{-1} X X^T\big)\big] &&= \mathrm{Tr}\big[Q_\lambda^{-1}\mathbb{E}(X X^T)\big] \\
&= \mathrm{Tr}\big(Q_\lambda^{-1}(Q_\lambda - \lambda I)\big) &&= d - \lambda\,\mathrm{Tr}(Q_\lambda^{-1}) \quad = \; D\,.
\end{aligned} \tag{6.19}$$

We can weaken Lemma 6.1 (page 41) noticing that for any real number x, $x \leq -\log(1-x)$ and

$$-\log\left(1-x+\frac{x^2}{2}\right) = \log\left(\frac{1+x+x^2/2}{1+x^4/4}\right) \leq \log\left(1+x+\frac{x^2}{2}\right) \leq x+\frac{x^2}{2}.$$

We obtain with probability at least $1-\exp(-\eta)$

$$\begin{aligned} n\mathbb{E}\big[W(\theta_2)\big] + \frac{n\alpha}{\beta}\mathbb{E}\big[\|\overline{X}\|^2\big] - 5n\mathbb{E}\big[W(\theta_2)^2\big] \\ - \mathbb{E}\left\{\frac{2n\alpha^2\|\overline{X}\|^2}{\beta}\big(\langle\theta_0,\overline{X}\rangle - Y\big)^2 + \frac{2n\alpha^2\|\overline{X}\|^4}{\beta^2}\right\} \\ \leq \sum_{i=1}^n \left\{ W_i(\theta_2) + 5W_i(\theta_2)^2 \right. \\ \left. + \frac{\alpha\|\overline{X}_i\|^2}{\beta} + \frac{2\alpha^2\|\overline{X}_i\|^2}{\beta}\big(\langle\theta_0,\overline{X}_i\rangle - Y\big)^2 + \frac{2\alpha^2\|\overline{X}_i\|^4}{\beta^2}\right\} \\ + \frac{\beta}{2}\|\theta_2-\theta_0\|^2 + \eta. \end{aligned}$$

Noticing that for any real numbers a and b, $4ab \leq a^2+4b^2$, we can then bound

$$\begin{aligned} \alpha^{-2}W(\theta_2)^2 &= \langle\theta_2-\theta_0,\overline{X}\rangle^2\big(\langle\theta_2+\theta_0,\overline{X}\rangle - 2Y\big)^2 \\ &= \langle\theta_2-\theta_0,\overline{X}\rangle^2\Big[\langle\theta_2-\theta_0,\overline{X}\rangle + 2\big(\langle\theta_0,\overline{X}\rangle - Y\big)\Big]^2 \\ &= \langle\theta_2-\theta_0,\overline{X}\rangle^4 + 4\langle\theta_2-\theta_0,\overline{X}\rangle^3\big(\langle\theta_0,\overline{X}\rangle - Y\big) \\ &\quad + 4\langle\theta_2-\theta_0,\overline{X}\rangle^2\big(\langle\theta_0,\overline{X}\rangle - Y\big)^2 \\ &\leq 2\langle\theta_2-\theta_0,\overline{X}\rangle^4 + 8\langle\theta_2-\theta_0,\overline{X}\rangle^2\big(\langle\theta_0,\overline{X}\rangle - Y\big)^2. \end{aligned}$$

THEOREM 6.3 *Let us put*

$$\begin{aligned} \widehat{D} &= \frac{1}{n}\sum_{i=1}^n \|\overline{X}_i\|^2 \qquad \textit{(let us remind that } D = \mathbb{E}\big[\|\overline{X}\|^2\big] \textit{ from (6.19))}, \\ B_1 &= 2\mathbb{E}\Big[\|\overline{X}\|^2\big(\langle\theta_0,\overline{X}\rangle - Y\big)^2\Big], \\ \widehat{B}_1 &= \frac{2}{n}\sum_{i=1}^n\Big[\|\overline{X}_i\|^2\big(\langle\theta_0,\overline{X}_i\rangle - Y_i\big)^2\Big], \end{aligned}$$

$$B_2 = 2\mathbb{E}\Big[\|\overline{X}\|^4\Big],$$
$$\widehat{B}_2 = \frac{2}{n}\sum_{i=1}^n \|\overline{X}_i\|^4,$$
$$B_3 = 40\sup\Big\{\mathbb{E}\big[\langle u, \overline{X}\rangle^2\big(\langle\theta_0, \overline{X}\rangle - Y\big)^2\big] : u \in \mathbb{R}^d, \|u\| = 1\Big\},$$
$$\widehat{B}_3 = \sup\Big\{\frac{40}{n}\sum_{i=1}^n \langle u, \overline{X}_i\rangle^2\big(\langle\theta_0, \overline{X}_i\rangle - Y_i\big)^2 : u \in \mathbb{R}^d, \|u\| = 1\Big\},$$
$$B_4 = 10\sup\Big\{\mathbb{E}\Big[\langle u, \overline{X}\rangle^4\Big] : u \in \mathbb{R}^d, \|u\| = 1\Big\},$$
$$\widehat{B}_4 = \sup\Big\{\frac{10}{n}\sum_{i=1}^n \langle u, \overline{X}_i\rangle^4 : u \in \mathbb{R}^d, \|u\| = 1\Big\}.$$

With probability at least $1 - \exp(-\eta)$*, for any* $\theta_2 \in \mathbb{R}^d$*,*

$$n\mathbb{E}\big[W(\theta_2)\big] - \left[n\alpha^2(B_3 + \widehat{B}_3) + \frac{\beta}{2}\right]\|\theta_2 - \theta_0\|^2 - n\alpha^2(B_4 + \widehat{B}_4)\|\theta_2 - \theta_0\|^4 \le \sum_{i=1}^n W_i(\theta_2) + \frac{n\alpha}{\beta}(\widehat{D} - D) + \frac{n\alpha^2}{\beta}(B_1 + \widehat{B}_1) + \frac{n\alpha^2}{\beta^2}(B_2 + \widehat{B}_2) + \eta.$$

Let us now assume that $\theta_2 \in \overline{\Theta}$ and let us use the fact that $\overline{\Theta}$ is a convex set and that $\theta_0 = \arg\min_{\theta\in\overline{\Theta}} \overline{R}(\theta) + \lambda\|Q_\lambda^{-1/2}\theta\|^2$. Introduce $\theta_* = \arg\min_{\theta\in\mathbb{R}^d} \overline{R}(\theta) + \lambda\|Q_\lambda^{-1/2}\theta\|^2$. As we have

$$\overline{R}(\theta) + \lambda\|Q_\lambda^{-1/2}\theta\|^2 = \|\theta - \theta_*\|^2 + \overline{R}(\theta_*) + \lambda\|Q_\lambda^{-1/2}\theta_*\|^2,$$

the vector θ_0 is uniquely defined as the projection of θ_* on $\overline{\Theta}$ for the Euclidean distance, and for any $\theta_2 \in \overline{\Theta}$

$$\begin{aligned}\alpha^{-1}\mathbb{E}\big[W(\theta_2)\big] + \lambda\|Q_\lambda^{-1/2}\theta_2\|^2 - \lambda\|Q_\lambda^{-1/2}\theta_0\|^2 &= \overline{R}(\theta_2) - \overline{R}(\theta_0) + \lambda\|Q_\lambda^{-1/2}\theta_2\|^2 - \lambda\|Q_\lambda^{-1/2}\theta_0\|^2 \\ &= \|\theta_2 - \theta_*\|^2 - \|\theta_0 - \theta_*\|^2 \\ &= \|\theta_2 - \theta_0\|^2 + 2\langle\theta_2 - \theta_0, \theta_0 - \theta_*\rangle \ge \|\theta_2 - \theta_0\|^2. \quad (6.20)\end{aligned}$$

This and the inequality

$$\alpha^{-1}\sum_{i=1}^n W_i(\theta_1) + n\lambda\|Q_\lambda^{-1/2}\theta_1\|^2 - n\lambda\|Q_\lambda^{-1/2}\theta_0\|^2 \le 0$$

leads to the following result.

THEOREM 6.4 *With probability at least* $1 - \exp(-\eta)$,

$$R(\hat{\theta}) + \lambda\|\hat{\theta}\|^2 - \inf_{\theta\in\Theta}\big[R(\theta) + \lambda\|\theta\|^2\big] \\ = \alpha^{-1}\mathbb{E}\big[W(\theta_1)\big] + \lambda\|Q_\lambda^{-1/2}\theta_1\|^2 - \lambda\|Q_\lambda^{-1/2}\theta_0\|^2$$

is not greater than the smallest positive non degenerate root of the following polynomial equation as soon as it has one

$$\Big\{1 - \big[\alpha(B_3 + \widehat{B}_3) + \tfrac{\beta}{2n\alpha}\big]\Big\}x - \alpha(B_4 + \widehat{B}_4)x^2 \\ = \frac{1}{\beta}\max(\widehat{D} - D, 0) + \frac{\alpha}{\beta}(B_1 + \widehat{B}_1) + \frac{\alpha}{\beta^2}(B_2 + \widehat{B}_2) + \frac{\eta}{n\alpha}.$$

PROOF. Let us remark first that when the polynomial appearing in the theorem has two distinct roots, they are of the same sign, due to the sign of its constant coefficient. Let $\widehat{\Omega}$ be the event of probability at least $1 - \exp(-\eta)$ described in Theorem 6.3 (page 43). For any realization of this event for which the polynomial described in Theorem 6.4 does not have two distinct positive roots, the statement of Theorem 6.4 is void, and therefore fulfilled. Let us consider now the case when the polynomial in question has two distinct positive roots $x_1 < x_2$. Consider in this case the random (trivially nonempty) closed convex set

$$\widehat{\Theta} = \big\{\theta \in \Theta : R(\theta) + \lambda\|\theta\|^2 \leq \inf_{\theta'\in\Theta}\big[R(\theta') + \lambda\|\theta'\|^2\big] + \tfrac{x_1+x_2}{2}\big\}.$$

Let $\theta_3 \in \arg\min_{\theta\in\widehat{\Theta}} r(\theta) + \lambda\|\theta\|^2$ and $\theta_4 \in \arg\min_{\theta\in\Theta} r(\theta) + \lambda\|\theta\|^2$. We see from Theorem 6.3 that

$$R(\theta_3) + \lambda\|\theta_3\|^2 < R(\theta_0) + \lambda\|\theta_0\|^2 + \frac{x_1 + x_2}{2}, \tag{6.21}$$

because it cannot be larger from the construction of $\widehat{\Theta}$. On the other hand, since $\widehat{\Theta} \subset \Theta$, the line segment $[\theta_3, \theta_4]$ is such that $[\theta_3, \theta_4] \cap \widehat{\Theta} \subset \arg\min_{\theta\in\widehat{\Theta}} r(\theta) + \lambda\|\theta\|^2$. We can therefore apply equation (6.21) to any point of $[\theta_3, \theta_4] \cap \widehat{\Theta}$, which proves that $[\theta_3, \theta_4] \cap \widehat{\Theta}$ is an open subset of $[\theta_3, \theta_4]$. But it is also a closed subset by construction, and therefore, as it is non empty and $[\theta_3, \theta_4]$ is connected, it proves that $[\theta_3, \theta_4] \cap \widehat{\Theta} = [\theta_3, \theta_4]$, and thus that $\theta_4 \in \widehat{\Theta}$. This can be applied to any choice of $\theta_3 \in \arg\min_{\theta\in\widehat{\Theta}} r(\theta) + \lambda\|\theta\|^2$ and $\theta_4 \in \arg\min_{\theta\in\Theta} r(\theta) + \lambda\|\theta\|^2$, proving that $\arg\min_{\theta\in\Theta} r(\theta) + \lambda\|\theta\|^2 \subset \arg\min_{\theta\in\widehat{\Theta}} r(\theta) + \lambda\|\theta\|^2$ and therefore that any $\theta_4 \in \arg\min_{\theta\in\Theta} r(\theta) + \lambda\|\theta\|^2$ is such that

$$R(\theta_4) + \lambda\|\theta_4\|^2 \leq \inf_{\theta\in\Theta}\big[R(\theta) + \lambda\|\theta\|^2\big] + x_1.$$

because the values between x_1 and x_2 are excluded by Theorem 6.3. $\square$

The actual convergence speed of the least squares estimator $\hat{\theta}$ on Θ will depend on the speed of convergence of the "empirical bounds" $\widehat{B}_k$ towards their expectations. We can rephrase the previous theorem in the following more practical way:

THEOREM 6.5 *Let $\eta_0, \eta_1, \ldots, \eta_5$ be positive real numbers. With probability at least*

$$1 - \mathbb{P}(\widehat{D} > D + \eta_0) - \sum_{k=1}^{4} \mathbb{P}(\widehat{B}_k - B_k > \eta_k) - \exp(-\eta_5),$$

$R(\hat{\theta}) + \lambda\|\hat{\theta}\|^2 - \inf_{\theta\in\Theta}\left[R(\theta) + \lambda\|\theta\|^2\right]$ is smaller than the smallest non degenerate positive root of

$$\left\{1 - \left[\alpha(2B_3 + \eta_3) + \tfrac{\beta}{2n\alpha}\right]\right\}x - \alpha(2B_4 + \eta_4)x^2 = \frac{\eta_0}{\beta} + \frac{\alpha}{\beta}(2B_1 + \eta_1) + \frac{\alpha}{\beta^2}(2B_2 + \eta_2) + \frac{\eta_5}{n\alpha}, \quad (6.22)$$

where we can optimize the values of $\alpha > 0$ and $\beta > 0$, since this equation has non random coefficients. For example, taking for simplicity

$$\alpha = \frac{1}{8B_3 + 4\eta_3},$$
$$\beta = \frac{n\alpha}{2},$$

we obtain

$$x - \frac{2B_4 + \eta_4}{4B_3 + 2\eta_3}x^2 = \frac{16\eta_0(2B_3 + \eta_3)}{n} + \frac{8B_1 + 4\eta_1}{n} + \frac{32(2B_3 + \eta_3)(2B_2 + \eta_2)}{n^2} + \frac{8\eta_5(2B_3 + \eta_3)}{n}.$$

6.2.1. Proof of Theorem 2.1. Let us now deduce Theorem 2.1 (page 13) from Theorem 6.5. Let us first remark that with probability at least $1 - \varepsilon/2$

$$\widehat{D} \leq D + \sqrt{\frac{B_2}{\varepsilon n}},$$

because the variance of $\widehat{D}$ is less than $\frac{B_2}{2n}$. For a given $\varepsilon > 0$, let us take $\eta_0 = \sqrt{\frac{B_2}{\varepsilon n}}$, $\eta_1 = B_1$, $\eta_2 = B_2$, $\eta_3 = B_3$ and $\eta_4 = B_4$. We get that $R_\lambda(\hat{\theta}) - \inf_{\theta\in\Theta} R_\lambda(\theta)$ is smaller than the smallest positive non degenerate root of

$$x - \frac{B_4}{2B_3}x^2 = \frac{48B_3}{n}\sqrt{\frac{B_2}{n\varepsilon}} + \frac{12B_1}{n} + \frac{288B_2B_3}{n^2} + \frac{24\log(3/\varepsilon)B_3}{n},$$

with probability at least

$$1 - \frac{5\,\varepsilon}{6} - \sum_{k=1}^{4} \mathbb{P}(\widehat{B}_k > B_k + \eta_k).$$

According to the weak law of large numbers, there is n_ε such that for any $n \geq n_\varepsilon$,

$$\sum_{k=1}^{4} \mathbb{P}(\widehat{B}_k > B_k + \eta_k) \leq \varepsilon/6.$$

Thus, increasing n_ε and the constants to absorb the second order terms, we see that for some n_ε and any $n \geq n_\varepsilon$, with probability at least $1 - \varepsilon$, the excess risk is less than the smallest positive root of

$$x - \frac{B_4}{2B_3}x^2 = \frac{13B_1}{n} + \frac{24\log(3/\varepsilon)B_3}{n}.$$

Now, as soon as $ac < 1/4$, the smallest positive root of $x - ax^2 = c$ is $\frac{2c}{1+\sqrt{1-4ac}}$. This means that for n large enough, with probability at least $1 - \varepsilon$,

$$R_\lambda(\hat{\theta}) - \inf_\theta R_\lambda(\theta) \leq \frac{15B_1}{n} + \frac{25\log(3/\varepsilon)B_3}{n},$$

which is precisely the statement of Theorem 2.1 (page 13), up to some change of notation.

6.2.2. Proof of Theorem 2.2. Let us now weaken Theorem 6.4 in order to make a more explicit non asymptotic result and obtain Theorem 2.2. From now on, we will assume that $\lambda = 0$. We start by giving bounds on the quantity defined in Theorem 6.3 in terms of

$$B = \sup_{f \in \mathrm{span}\{\varphi_1,\ldots,\varphi_d\} - \{0\}} \|f\|_\infty^2 / \mathbb{E}[f(X)]^2.$$

Since we have

$$\|\overline{X}\|^2 = \|Q_\lambda^{-1/2}X\|^2 \leq dB,$$

we get

$$\widehat{d} = \frac{1}{n}\sum_{i=1}^{n} \|\overline{X}_i\|^2 \leq dB,$$

$$B_1 = 2\mathbb{E}\Big[\|\overline{X}\|^2\big(\langle\theta_0, \overline{X}\rangle - Y\big)^2\Big] \leq 2dB\,R(f^*),$$

$$\widehat{B}_1 = \frac{2}{n}\sum_{i=1}^{n}\Big[\|\overline{X}_i\|^2\big(\langle\theta_0,\overline{X}_i\rangle - Y_i\big)^2\Big] \le 2dB\, r(f^*),$$

$$B_2 = 2\mathbb{E}\Big[\|\overline{X}\|^4\Big] \le 2d^2B^2,$$

$$\widehat{B}_2 = \frac{2}{n}\sum_{i=1}^{n}\|\overline{X}_i\|^4 \le 2d^2B^2,$$

$$B_3 = 40\sup\Big\{\mathbb{E}\big[\langle u,\overline{X}\rangle^2\big(\langle\theta_0,\overline{X}\rangle - Y\big)^2\big] : u\in\mathbb{R}^d, \|u\| = 1\Big\} \le 40B\, R(f^*),$$

$$\widehat{B}_3 = \sup\Big\{\frac{40}{n}\sum_{i=1}^{n}\langle u,\overline{X}_i\rangle^2\big(\langle\theta_0,\overline{X}_i\rangle - Y_i\big)^2 : u\in\mathbb{R}^d, \|u\| = 1\Big\} \le 40B\, r(f^*),$$

$$B_4 = 10\sup\Big\{\mathbb{E}\Big[\langle u,\overline{X}\rangle^4\Big] : u\in\mathbb{R}^d, \|u\| = 1\Big\} \le 10B^2,$$

$$\widehat{B}_4 = \sup\Big\{\frac{10}{n}\sum_{i=1}^{n}\langle u,\overline{X}_i\rangle^4 : u\in\mathbb{R}^d, \|u\| = 1\Big\} \le 10B^2.$$

Let us put

$$a_0 = \frac{2dB + 4dB\alpha[R(f^*) + r(f^*)] + \eta}{\alpha n} + \frac{16B^2d^2}{\alpha n^2},$$

$$a_1 = 3/4 - 40\alpha B[R(f^*) + r(f^*)],$$

and

$$a_2 = 20\alpha B^2.$$

Theorem 6.4 applied with $\beta = n\alpha/2$ implies that with probability at least $1-\eta$ the excess risk $R(\hat{f}^{(\mathrm{erm})}) - R(f^*)$ is upper bounded by the smallest positive root of $a_1x - a_2x^2 = a_0$ as soon as $a_1^2 > 4a_0a_2$. In particular, setting $\varepsilon = \exp(-\eta)$ when (6.23) holds, we have

$$R(\hat{f}^{(\mathrm{erm})}) - R(f^*) \le \frac{2a_0}{a_1 + \sqrt{a_1^2 - 4a_0a_2}} \le \frac{2a_0}{a_1}.$$

We conclude that

THEOREM 6.6 *For any $\alpha > 0$ and $\varepsilon > 0$, with probability at least $1-\varepsilon$, if the inequality*

$$80\left(\frac{(2 + 4\alpha[R(f^*) + r(f^*)])Bd + \log(\varepsilon^{-1})}{n} + \left(\frac{4Bd}{n}\right)^2\right) < \left(\frac{3}{4B} - 40\alpha[R(f^*) + r(f^*)]\right)^2 \quad (6.23)$$

holds, then we have

$$R(\hat{f}^{(\text{erm})}) - R(f^*) \leq \mathcal{J}\left(\frac{(2 + 4\alpha[R(f^*) + r(f^*)])Bd + \log(\varepsilon^{-1})}{n} + \left(\frac{4Bd}{n} \right)^2 \right), \tag{6.24}$$

where $\mathcal{J} = 8/(3\alpha - 160\alpha^2 B[R(f^*) + r(f^*)])$

Now, the Bienaymé-Chebyshev inequality implies

$$\mathbb{P}\big(r(f^*) - R(f^*) \geq t\big) \leq \frac{\mathbb{E}\big(r(f^*) - R(f^*)\big)^2}{t^2} \leq \mathbb{E}[Y - f^*(X)]^4/nt^2.$$

Under the finite moment assumption of Theorem 2.2, we obtain that for any $\varepsilon \geq 1/n$, with probability at least $1 - \varepsilon$,

$$r(f^*) < R(f^*) + \sqrt{\mathbb{E}[Y - f^*(X)]^4}.$$

From Theorem 6.6 and a union bound, by taking

$$\alpha = \left(80B[2R(f^*) + \sqrt{\mathbb{E}[Y - f^*(X)]^4} \right)^{-1},$$

we get that with probability $1 - 2\varepsilon$,

$$R(\hat{f}^{(\text{erm})}) - R(f^*) \leq \mathcal{J}_1 B \left(\frac{3Bd' + \log(\varepsilon^{-1})}{n} + \left(\frac{4Bd'}{n} \right)^2 \right),$$

with $\mathcal{J}_1 = 640 \left(2R(f^*) + \sqrt{\mathbb{E}\big\{[Y - f^*(X)]^4\big\}} \right)$. This concludes the proof of Theorem 2.2.

REMARK 6.1 Let us indicate now how to handle the case when Q is degenerate. Let us consider the linear subspace S of $\mathbb{R}^d$ spanned by the eigenvectors of Q corresponding to positive eigenvalues. Then almost surely $\text{Span}\{X_i, i = 1, \ldots, n\} \subset S$. Indeed for any θ in the kernel of Q, $\mathbb{E}\big(\langle \theta, X \rangle^2\big) = 0$ implies that $\langle \theta, X \rangle = 0$ almost surely, and considering a basis of the kernel, we see that $X \in S$ almost surely, S being orthogonal to the kernel of Q. Thus we can restrict the problem to S, as soon as we choose

$$\hat{\theta} \in \text{span}\big\{X_1, \ldots, X_n\big\} \cap \arg\min_{\theta} \sum_{i=1}^{n} \big(\langle \theta, X_i \rangle - Y_i\big)^2,$$

or equivalently with the notation $\mathbf{X} = (\varphi_j(X_i))_{1 \leq i \leq n, 1 \leq j \leq d}$ and $\mathbf{Y} = [Y_j]_{j=1}^n$,

$$\hat{\theta} \in \text{im}\, \mathbf{X}^T \cap \arg\min_{\theta} \| \mathbf{X}\theta - Y \|^2$$

This proves that the results of this section apply to this special choice of the empirical least squares estimator. Since we have $\mathbb{R}^d = \ker \mathbf{X} \oplus \text{im}\, \mathbf{X}^T$, this choice is unique.

6.3. PROOF OF THEOREM 3.1. We use a similar notation as in Section 6.2: we write X for $\varphi(X)$. Therefore, the function f_θ maps an input x to $\langle\theta, x\rangle$. We consider the change of coordinates

$$\overline{X} = Q_\lambda^{-1/2} X.$$

Thus, from (6.19), we have $\mathbb{E}\big[\|\overline{X}\|^2\big] = D$. We will use

$$\overline{R}(\theta) = \mathbb{E}\big[(\langle\theta, \overline{X}\rangle - Y)^2\big],$$

so that $\overline{R}(Q^{1/2}\theta) = \mathbb{E}\big[(\langle\theta, X\rangle - Y)^2\big] = R(f_\theta)$. Let

$$\overline{\Theta} = \big\{Q_\lambda^{1/2}\theta; \theta \in \Theta\big\}.$$

Consider

$$\theta_0 = \arg\min_{\theta\in\overline{\Theta}} \Big\{\overline{R}(\theta) + \lambda\|Q_\lambda^{-1/2}\theta\|^2\Big\}.$$

We thus have $\tilde{\theta} = Q_\lambda^{-1/2}\theta_0$, and

$$\begin{aligned}
\sigma &= \sqrt{\mathbb{E}\big[\big(\langle\theta_0, \overline{X}\rangle - Y\big)^2\big]},\\
\chi &= \sup_{u\in\mathbb{R}^d} \frac{\mathbb{E}\big(\langle u, \overline{X}\rangle^4\big)^{1/2}}{\mathbb{E}\big(\langle u, \overline{X}\rangle^2\big)},\\
\kappa &= \frac{\mathbb{E}\big(\|\overline{X}\|^4\big)^{1/2}}{\mathbb{E}\big(\|\overline{X}\|^2\big)} = \frac{\mathbb{E}\big(\|\overline{X}\|^4\big)^{1/2}}{D},\\
\kappa' &= \frac{\mathbb{E}\big[\big(\langle\theta_0, \overline{X}\rangle - Y\big)^4\big]^{1/2}}{\sigma^2},\\
T &= \|\overline{\Theta}\| = \max_{\theta,\theta'\in\overline{\Theta}} \|\theta - \theta'\|.
\end{aligned}$$

For $\alpha > 0$, we introduce

$$\begin{aligned}
J_i(\theta) &= \langle\theta, \overline{X}_i\rangle - Y_i, & J(\theta) &= \langle\theta, \overline{X}\rangle - Y\\
\overline{L}_i(\theta) &= \alpha\big(\langle\theta, \overline{X}_i\rangle - Y_i\big)^2, & \overline{L}(\theta) &= \alpha\big(\langle\theta, \overline{X}\rangle - Y\big)^2\\
W_i(\theta) &= \overline{L}_i(\theta) - \overline{L}_i(\theta_0), & W(\theta) &= \overline{L}(\theta) - \overline{L}(\theta_0),
\end{aligned}$$

and

$$r'(\theta, \theta') = \lambda(\|Q_\lambda^{-1/2}\theta\|^2 - \|Q_\lambda^{-1/2}\theta'\|^2) + \frac{1}{n\alpha}\sum_{i=1}^n \psi\big(\overline{L}(\theta) - \overline{L}(\theta')\big).$$

Let $\bar{\theta} = Q_\lambda^{1/2}\hat{\theta} \in \overline{\Theta}$. We have

$$-r'(\theta_0, \bar{\theta}) = r'(\bar{\theta}, \theta_0) \quad \leq \max_{\theta_1 \in \overline{\Theta}} r'(\bar{\theta}, \theta_1) \quad \leq \gamma + \max_{\theta_1 \in \overline{\Theta}} r'(\theta_0, \theta_1), \tag{6.25}$$

where $\gamma = \max_{\theta_1 \in \overline{\Theta}} r'(\bar{\theta}, \theta_1) - \inf_{\theta \in \overline{\Theta}} \max_{\theta_1 \in \overline{\Theta}} r'(\theta, \theta_1)$ is a quantity which can be made arbitrary small by choice of the estimator. By using an upper bound $r'(\theta_0, \theta_1)$ that holds uniformly in θ_1, we will control both left and right hand sides of (6.25).

To achieve this, we will upper bound

$$r'(\theta_0, \theta_1) = \lambda(\|Q_\lambda^{-1/2}\theta_0\|^2 - \|Q_\lambda^{-1/2}\theta_1\|^2) + \frac{1}{n\alpha}\sum_{i=1}^{n} \psi\big[-W_i(\theta_1)\big] \tag{6.26}$$

by the expectation of a distribution depending on θ_1 of a *quantity that does not depend on* θ_1, and then use the PAC-Bayesian argument to control this expectation uniformly in θ_1. The distribution depending on θ_1 should therefore be taken such that for any $\theta_1 \in \overline{\Theta}$, its Kullback-Leibler divergence with respect to some fixed distribution is small (at least when θ_1 is close to θ_0).

Let us start with the following result.

LEMMA 6.7 *Let $f, g : \mathbb{R} \to \mathbb{R}$ be two Lebesgue measurable functions such that $f(x) \leq g(x)$, $x \in \mathbb{R}$. Let us assume that there exists $h \in \mathbb{R}$ such that $x \mapsto g(x) + h\frac{x^2}{2}$ is convex. Then for any probability distribution μ on the real line,*

$$f\left(\int x\mu(dx)\right) \leq \int g(x)\mu(dx) + \min\Big\{\sup f - \inf f, \frac{h}{2}\,\mathbb{V}\mathrm{ar}(\mu)\Big\}.$$

PROOF. Let us put $x_0 = \int x\mu(dx)$ The function

$$x \mapsto g(x) + \frac{h}{2}(x - x_0)^2$$

is convex. Thus, by Jensen's inequality

$$f(x_0) \leq g(x_0) \leq \int \mu(dx)\left[g(x) + \frac{h}{2}(x - x_0)^2\right] = \int g(x)\mu(dx) + \frac{h}{2}\,\mathbb{V}\mathrm{ar}(\mu).$$

On the other hand

$$f(x_0) \leq \sup f \leq \sup f + \int \big[g(x) - \inf f\big]\mu(dx)$$
$$= \int g(x)\mu(dx) + \sup f - \inf f.$$

The lemma is a combination of these two inequalities. $\square$

The above lemma will be used with $f = g = \psi$, where ψ is the increasing influence function

$$\psi(x) = \begin{cases} -\log(2), & x \le -1, \\ \log(1 + x + x^2/2), & -1 \le x \le 0, \\ -\log(1 - x + x^2/2), & 0 \le x \le 1, \\ \log(2), & x \ge 1. \end{cases}$$

Since we have for any $x \in \mathbb{R}$

$$-\log\left(1 - x + \frac{x^2}{2}\right) = \log\left(\frac{1 + x + \frac{x^2}{2}}{1 + \frac{x^4}{4}}\right) < \log\left(1 + x + \frac{x^2}{2}\right),$$

the function ψ satisfies for any $x \in \mathbb{R}$

$$-\log\left(1 - x + \frac{x^2}{2}\right) < \psi(x) < \log\left(1 + x + \frac{x^2}{2}\right).$$

Moreover

$$\psi'(x) = \frac{1 - x}{1 - x + \frac{x^2}{2}}, \quad \psi''(x) = \frac{x(x-2)}{2\left(1 - x + \frac{x^2}{2}\right)^2} \ge -2, \quad 0 \le x \le 1,$$

showing (by symmetry) that the function $x \mapsto \psi(x) + 2x^2$ is convex on the real line.

For any $\theta' \in \mathbb{R}^d$ and $\beta > 0$, we consider the Gaussian distribution with mena θ' and covariance $\beta^{-1} I$:

$$\rho_{\theta'}(d\theta) = \left(\frac{\beta}{2\pi}\right)^{d/2} \exp\left(-\frac{\beta}{2}\|\theta - \theta'\|^2\right) d\theta.$$

From Lemmas 6.2 and 6.7 (with μ the distribution of $-W_i(\theta) + \frac{\alpha\|\overline{X}_i\|^2}{\beta}$ when θ is drawn from ρ_{θ_1} and for a fixed pair (X_i, Y_i)), we can see that

$$\begin{aligned} \psi\big[-W_i(\theta_1)\big] &= \psi\left\{\int \rho_{\theta_1}(d\theta)\left[-W_i(\theta) + \frac{\alpha\|\overline{X}_i\|^2}{\beta}\right]\right\} \\ &\le \int \rho_{\theta_1}(d\theta)\psi\left[-W_i(\theta) + \frac{\alpha\|\overline{X}_i\|^2}{\beta}\right] \\ &\qquad + \min\left\{\log(4), \mathbb{V}\mathrm{ar}_{\rho_{\theta_1}}\left[\overline{L}_i(\theta)\right]\right\}. \end{aligned}$$

Let us compute

$$\begin{aligned}\frac{1}{\alpha^2}\,\mathbb{V}\mathrm{ar}_{\rho_{\theta_1}}\big[\overline{L}_i(\theta)\big] &= \mathbb{V}\mathrm{ar}_{\rho_{\theta_1}}\big[J_i^2(\theta) - J_i^2(\theta_1)\big] \\ &= \int \rho_{\theta_1}(d\theta)\big[J_i^2(\theta) - J_i^2(\theta_1)\big]^2 - \frac{\|\overline{X}_i\|^4}{\beta^2} \\ &= \int \rho_{\theta_1}(d\theta)\Big[\langle\theta-\theta_1,\overline{X}_i\rangle^2 + 2\langle\theta-\theta_1,\overline{X}_i\rangle J_i(\theta_1)\Big]^2 - \frac{\|\overline{X}_i\|^4}{\beta^2} \\ &= \frac{2\|\overline{X}_i\|^4}{\beta^2} + \frac{4\overline{L}_i(\theta_1)\|\overline{X}_i\|^2}{\alpha\beta}. \qquad (6.27)\end{aligned}$$

Let $\xi \in (0,1)$. Now we can remark that

$$\overline{L}_i(\theta_1) \le \frac{\overline{L}_i(\theta)}{\xi} + \frac{\alpha\langle\theta-\theta_1,\overline{X}_i\rangle^2}{1-\xi}.$$

We get

$$\begin{aligned}&\min\Big\{\log(4), \mathbb{V}\mathrm{ar}_{\rho_{\theta_1}}\big[\overline{L}_i(\theta)\big]\Big\} \\ &\qquad = \min\Big\{\log(4), \frac{4\alpha\|\overline{X}_i\|^2\overline{L}_i(\theta_1)}{\beta} + \frac{2\alpha^2\|\overline{X}_i\|^4}{\beta^2}\Big\} \\ &\le \int \rho_{\theta_1}(d\theta)\min\Big\{\log(4), \\ &\qquad\qquad \frac{4\alpha\|\overline{X}_i\|^2\overline{L}_i(\theta)}{\beta\xi} + \frac{2\alpha^2\|\overline{X}_i\|^4}{\beta^2} + \frac{4\alpha^2\|\overline{X}_i\|^2\langle\theta-\theta_1,\overline{X}_i\rangle^2}{\beta(1-\xi)}\Big\} \\ &\qquad \le \int \rho_{\theta_1}(d\theta)\min\Big\{\log(4), \frac{4\alpha\|\overline{X}_i\|^2\overline{L}_i(\theta)}{\beta\xi} + \frac{2\alpha^2\|\overline{X}_i\|^4}{\beta^2}\Big\} \\ &\qquad\qquad + \min\Big\{\log(4), \frac{4\alpha^2\|\overline{X}_i\|^4}{\beta^2(1-\xi)}\Big\}.\end{aligned}$$

Let us now put $a = \frac{3}{\log(4)} < 2.17$, $b = a + a^2\log(4) < 8.7$ and let us remark that

$$\begin{aligned}&\min\big\{\log(4),x\big\} + \min\big\{\log(4),y\big\} \\ &\qquad \le \log\big[1 + a\min\{\log(4),x\}\big] + \log(1+ay) \\ &\qquad\qquad \le \log\big(1+ax+by\big), \qquad x,y\in\mathbb{R}_+.\end{aligned}$$

Thus

$$\min\Big\{\log(4), \mathbb{V}\mathrm{ar}_{\rho_{\theta_1}}\big[\overline{L}_i(\theta)\big]\Big\}$$

$$\leq \int \rho_{\theta_1}(d\theta) \log\biggl[1 + \frac{4a\alpha\|\overline{X}_i\|^2\overline{L}_i(\theta)}{\beta\xi} + \frac{2\alpha^2\|\overline{X}_i\|^4}{\beta^2}\biggl(a + \frac{2b}{1-\xi}\biggr)\biggr].$$

We can then remark that

$$\begin{aligned}\psi(x) + \log(1+y) &= \log\bigl[\exp[\psi(x)] + y\exp[\psi(x)]\bigr] \\ &\leq \log\bigl[\exp[\psi(x)] + 2y\bigr] \leq \log\biggl(1 + x + \frac{x^2}{2} + 2y\biggr), \qquad x \in \mathbb{R}, y \in \mathbb{R}_+.\end{aligned}$$

Thus, putting $c_0 = a + \dfrac{2b}{1-\xi}$, we get

$$\psi\bigl[-W_i(\theta_1)\bigr] \leq \int \rho_{\theta_1}(d\theta) \log[A_i(\theta)], \tag{6.28}$$

with

$$\begin{aligned}A_i(\theta) = 1 - W_i(\theta) + \frac{\alpha\|\overline{X}_i\|^2}{\beta} + \frac{1}{2}\biggl(-W_i(\theta) + \frac{\alpha\|\overline{X}_i\|^2}{\beta}\biggr)^2 \\ + \frac{8a\alpha\|\overline{X}_i\|^2\overline{L}_i(\theta)}{\beta\xi} + \frac{4c_0\alpha^2\|\overline{X}_i\|^4}{\beta^2}.\end{aligned}$$

Similarly, we define $A(\theta)$ by replacing (X_i, Y_i) by (X, Y). Since we have

$$\mathbb{E}\exp\biggl(\sum_{i=1}^n \log[A_i(\theta)] - n\log[\mathbb{E}A(\theta)]\biggr) = 1,$$

from the usual PAC-Bayesian argument, we have with probability at least $1-\varepsilon$, for any $\theta_1 \in \mathbb{R}^d$,

$$\begin{aligned}\int \rho_{\theta_1}(d\theta)\biggl(\sum_{i=1}^n \log[A_i(\theta)]\biggr) - n\int \rho_{\theta_1}(d\theta)\log[A(\theta)] &\leq K(\rho_{\theta_1}, \rho_{\theta_0}) + \log(\varepsilon^{-1}) \\ &\leq \frac{\beta\|\theta_1 - \theta_0\|^2}{2} + \log(\varepsilon^{-1})\end{aligned}$$

From (6.26) and (6.28), with probability at least $1-\varepsilon$, for any $\theta_1 \in \mathbb{R}^d$, we get

$$\begin{aligned}r'(\theta_0, \theta_1) \leq \frac{1}{\alpha}\log\biggl\{1 + \mathbb{E}\biggl[\int \rho_{\theta_1}(d\theta)\biggl(-W(\theta) + \frac{\alpha\|\overline{X}\|^2}{\beta} \\ + \frac{1}{2}\biggl(-W(\theta) + \frac{\alpha\|\overline{X}\|^2}{\beta}\biggr)^2 + \frac{8a\alpha\|\overline{X}\|^2\overline{L}(\theta)}{\beta\xi} + \frac{4c_0\alpha^2\|\overline{X}\|^4}{\beta^2}\biggr)\biggr]\biggr\}\end{aligned}$$

$$+\frac{\beta\|\theta_1-\theta_0\|^2}{2n\alpha}+\frac{\log(\varepsilon^{-1})}{n\alpha}+\lambda(\|Q_\lambda^{-1/2}\theta_0\|^2-\|Q_\lambda^{-1/2}\theta_1\|^2).$$

Now from (6.27) and $\frac{\alpha\|\overline{X}\|^2}{\beta}=-\overline{L}(\theta_1)+\int\rho_{\theta_1}(d\theta)\overline{L}(\theta)$, we have

$$\begin{aligned}\int\rho_{\theta_1}(d\theta)\Big(-W(\theta)+\frac{\alpha\|\overline{X}\|^2}{\beta}\Big)^2&=\mathbb{V}\mathrm{ar}_{\rho_{\theta_1}}\big[\overline{L}(\theta)\big]+W(\theta_1)^2\\&=W(\theta_1)^2+\frac{4\alpha\overline{L}(\theta_1)\|\overline{X}\|^2}{\beta}+\frac{2\alpha^2\|\overline{X}\|^4}{\beta^2}.\end{aligned}$$

PROPOSITION 6.8 *With probability at least* $1-\varepsilon$*, for any* $\theta_1\in\mathbb{R}^d$*,*

$$\begin{aligned}r'(\theta_0,\theta_1)&\le\frac{1}{\alpha}\log\Big\{1+\mathbb{E}\Big[-W(\theta_1)+\frac{W(\theta_1)^2}{2}+\frac{\big(2+8a/\xi\big)\alpha\|\overline{X}\|^2\overline{L}(\theta_1)}{\beta}\\&\qquad+\frac{\big(1+8a/\xi+4c_0\big)\alpha^2\|\overline{X}\|^4}{\beta^2}\Big]\Big\}+\frac{\beta\|\theta_1-\theta_0\|^2}{2n\alpha}+\frac{\log(\varepsilon^{-1})}{n\alpha}\\&\qquad+\lambda(\|Q_\lambda^{-1/2}\theta_0\|^2-\|Q_\lambda^{-1/2}\theta_1\|^2)\\&\le\mathbb{E}\Big[J(\theta_0)^2-J(\theta_1)^2+\frac{1}{2\alpha}W(\theta_1)^2+\frac{(2+8a/\xi)\|\overline{X}\|^2\overline{L}(\theta_1)}{\beta}\\&\qquad+\frac{(1+8a/\xi+4c_0)\alpha\|\overline{X}\|^4}{\beta^2}\Big]+\frac{\beta\|\theta_1-\theta_0\|^2}{2n\alpha}+\frac{\log(\varepsilon^{-1})}{n\alpha}\\&\qquad+\lambda(\|Q_\lambda^{-1/2}\theta_0\|^2-\|Q_\lambda^{-1/2}\theta_1\|^2).\end{aligned}$$

By using the triangular inequality and Cauchy-Scwarz's inequality, we get

$$\begin{aligned}\frac{1}{\alpha^2}\mathbb{E}\big[W(\theta_1)^2\big]&=\mathbb{E}\Big\{\big[\langle\theta_1-\theta_0,\overline{X}\rangle^2+2\langle\theta_1-\theta_0,\overline{X}\rangle J(\theta_0)\big]^2\Big\}\\&\le\Big\{\mathbb{E}\big[\langle\theta_1-\theta_0,\overline{X}\rangle^4\big]^{1/2}+2\mathbb{E}\big[\langle\theta_1-\theta_0,\overline{X}\rangle^4\big]^{1/4}\mathbb{E}\big[J(\theta_0)^4\big]^{1/4}\Big\}^2\\&\le\Bigg\{\chi\|\theta_1-\theta_0\|^2\mathbb{E}\Big[\Big\langle\frac{\theta_1-\theta_0}{\|\theta_1-\theta_0\|},\overline{X}\Big\rangle^2\Big]\\&\qquad+2\|\theta_1-\theta_0\|\sigma\sqrt{\kappa'\chi}\sqrt{\mathbb{E}\Big[\Big\langle\frac{\theta_1-\theta_0}{\|\theta_1-\theta_0\|},\overline{X}\Big\rangle^2\Big]}\Bigg\}^2\\&\le\frac{\chi q_{\max}}{q_{\max}+\lambda}\|\theta_1-\theta_0\|^2\Big\{\|\theta_1-\theta_0\|\sqrt{\frac{\chi q_{\max}}{q_{\max}+\lambda}}+2\sigma\sqrt{\kappa'}\Big\}^2,\end{aligned}$$

and

$$\frac{1}{\alpha}\mathbb{E}\big[\|\overline{X}\|^2\overline{L}(\theta_1)\big] = \mathbb{E}\Big\{\big[\|\overline{X}\|\langle\theta_1-\theta_0,\overline{X}\rangle + \|\overline{X}\|J(\theta_0)\big]^2\Big\}$$
$$\leq \mathbb{E}\big[\|\overline{X}\|^4\big]^{1/2}\Big\{\mathbb{E}\big[\langle\theta_1-\theta_0,\overline{X}\rangle^4\big]^{1/4} + \mathbb{E}\big[J(\theta_0)^4\big]^{1/4}\Big\}^2$$
$$\leq \kappa D\left\{\|\theta_1-\theta_0\|\sqrt{\frac{\chi q_{\max}}{q_{\max}+\lambda}} + 2\sigma\sqrt{\kappa'}\right\}^2,$$

Let us put

$$\widetilde{R}(\theta) = \overline{R}(\theta) + \lambda\|Q_\lambda^{-1/2}\theta\|^2,$$
$$c_1 = 4(2+8a/\xi),$$
$$c_2 = 4(1+8a/\xi+4c_0),$$
$$\delta = \frac{c_1\kappa\kappa' D\sigma^2}{n} + \frac{2\chi\big(\frac{\log(\varepsilon^{-1})}{n} + \frac{c_2\kappa^2D^2}{n^2}\big)\big[2\sqrt{\kappa'}\sigma + \|\overline{\Theta}\|\sqrt{\chi}\big]^2}{1-\frac{4c_1\kappa\chi D}{n}}.$$

We have proved the following result.

PROPOSITION 6.9 *With probability at least* $1-\varepsilon$*, for any* $\theta_1\in\mathbb{R}^d$*,*

$$r'(\theta_0,\theta_1) \leq \widetilde{R}(\theta_0) - \widetilde{R}(\theta_1) + \frac{\alpha}{2}\chi\|\theta_1-\theta_0\|^2\big[2\sqrt{\kappa'}\sigma + \|\theta_1-\theta_0\|\sqrt{\chi}\big]^2$$
$$+ \frac{c_1\alpha}{4\beta}\kappa D\big[\sqrt{\kappa'}\sigma + \|\theta_1-\theta_0\|\sqrt{\chi}\big]^2 + \frac{c_2\alpha\kappa^2D^2}{4\beta^2}$$
$$+ \frac{\beta\|\theta_1-\theta_0\|^2}{2n\alpha} + \frac{\log(\varepsilon^{-1})}{n\alpha}.$$

Let us assume from now on that $\theta_1\in\overline{\Theta}$, our convex bounded parameter set. In this case, as seen in (6.20), we have $\|\theta_0-\theta_1\|^2\leq\widetilde{R}(\theta_1)-\widetilde{R}(\theta_0)$. We can also use the fact that

$$\big[\sqrt{\kappa'}\sigma + \|\theta_1-\theta_0\|\sqrt{\chi}\big]^2 \leq 2\kappa'\sigma^2 + 2\chi\|\theta_1-\theta_0\|^2.$$

We deduce from these remarks that with probability at least $1-\varepsilon$,

$$r'(\theta_0,\theta_1) \leq \left\{-1+\frac{\alpha\chi}{2}\big[2\sqrt{\kappa'}\sigma+\|\overline{\Theta}\|\sqrt{\chi}\big]^2+\frac{\beta}{2n\alpha}+\frac{c_1\alpha\kappa D\chi}{2\beta}\right\}\big[\widetilde{R}(\theta_1)-\widetilde{R}(\theta_0)\big]$$
$$+ \frac{c_1\alpha\kappa D\kappa'\sigma^2}{2\beta} + \frac{c_2\alpha\kappa^2D^2}{4\beta^2} + \frac{\log(\varepsilon^{-1})}{n\alpha}.$$

Let us assume that $n>4c_1\kappa\chi D$ and let us choose

$$\beta = \frac{n\alpha}{2},$$

$$\alpha = \frac{1}{2\chi\big[2\sqrt{\kappa'}\sigma + \|\overline{\Theta}\|\sqrt{\chi}\big]^2}\left(1 - \frac{4c_1\kappa\chi D}{n}\right),$$

to get

$$r'(\theta_0, \theta_1) \le -\frac{\widetilde{R}(\theta_1) - \widetilde{R}(\theta_0)}{2} + \delta.$$

Plugging this into (6.25), we get

$$\frac{\widetilde{R}(\bar\theta) - \widetilde{R}(\theta_0)}{2} - \delta \le r'(\bar\theta, \theta_0) \le \max_{\theta_1 \in \overline{\Theta}} \left(\frac{\widetilde{R}(\theta_0) - \widetilde{R}(\theta_1)}{2}\right) + \gamma + \delta = \gamma + \delta,$$

hence

$$\widetilde{R}(\bar\theta) - \widetilde{R}(\theta_0) \le 2\gamma + 4\delta.$$

Computing the numerical values of the constants when $\xi = 0.8$ gives $c_1 < 95$ and $c_2 < 1511$.

6.4. PROOF OF THEOREM 5.1. We use the standard way of obtaining PAC bounds through upper bounds on Laplace transform of appropriate random variables. This argument is synthetized in the following result.

LEMMA 6.10 *For any $\varepsilon > 0$ and any real-valued random variable V such that $\mathbb{E}\big[\exp(V)\big] \le 1$, with probability at least $1 - \varepsilon$, we have*

$$V \le \log(\varepsilon^{-1}).$$

$$\text{Let } V_1(\hat f) = \int \big[L^\flat(\hat f, f) + \gamma^* \bar R(f)\big]\pi^*_{-\gamma^*\bar R}(df) - \gamma \bar R(\hat f)$$
$$- \mathcal{I}^*(\gamma^*) + \mathcal{I}(\gamma) + \log\left(\int \exp\big[-\hat{\mathcal{E}}(f)\big]\pi(df)\right) - \log\left[\frac{d\rho}{d\hat\pi}(\hat f)\right],$$

$$\text{and } V_2 = -\log\left(\int \exp\big[-\hat{\mathcal{E}}(f)\big]\pi(df)\right) + \log\left(\int \exp\big[-\mathcal{E}^\sharp(f)\big]\pi(df)\right)$$

To prove the theorem, according to Lemma 6.10, it suffices to prove that

$$\mathbb{E}\Big\{\int \exp\big[V_1(\hat f)\big]\rho(d\hat f)\Big\} \le 1 \quad \text{and} \quad \mathbb{E}\Big[\int \exp(V_2)\rho(d\hat f)\Big] \le 1.$$

These two inequalities are proved in the following two sections.

6.4.1. Proof of $\mathbb{E}\Big\{\int \exp\big[V_1(\hat{f})\big]\rho(d\hat{f})\Big\} \leq 1$. From Jensen's inequality, we have

$$\begin{aligned}
&\int \big[L^\flat(\hat{f},f) + \gamma^*\bar{R}(f)\big]\pi^*_{-\gamma^*\bar{R}}(df) \\
&\quad = \int \big[\hat{L}(\hat{f},f) + \gamma^*\bar{R}(f)\big]\pi^*_{-\gamma^*\bar{R}}(df) + \int \big[L^\flat(\hat{f},f) - \hat{L}(\hat{f},f)\big]\pi^*_{-\gamma^*\bar{R}}(df) \\
&\quad \leq \int \big[\hat{L}(\hat{f},f) + \gamma^*\bar{R}(f)\big]\pi^*_{-\gamma^*\bar{R}}(df) + \log \int \exp\big[L^\flat(\hat{f},f) - \hat{L}(\hat{f},f)\big]\pi^*_{-\gamma^*\bar{R}}(df).
\end{aligned}$$

From Jensen's inequality again,

$$\begin{aligned}
-\hat{\mathcal{E}}(\hat{f}) &= -\log \int \exp\big[\hat{L}(\hat{f},f)\big]\pi^*(df) \\
&= -\log \int \exp\big[\hat{L}(\hat{f},f) + \gamma^*\bar{R}(f)\big]\pi^*_{-\gamma^*\bar{R}}(df) - \log \int \exp\big[-\gamma^*\bar{R}(f)\big]\pi^*(df) \\
&\leq -\int [\hat{L}(\hat{f},f) + \gamma^*\bar{R}(f)]\pi^*_{-\gamma^*\bar{R}}(df) + \mathcal{I}^*(\gamma^*).
\end{aligned}$$

From the two previous inequalities, we get

$$\begin{aligned}
V_1(\hat{f}) &\leq \int \big[\hat{L}(\hat{f},f) + \gamma^*\bar{R}(f)\big]\pi^*_{-\gamma^*\bar{R}}(df) \\
&\qquad + \log \int \exp\big[L^\flat(\hat{f},f) - \hat{L}(\hat{f},f)\big]\pi^*(df) - \gamma\bar{R}(\hat{f}) \\
&\qquad - \mathcal{I}^*(\gamma^*) + \mathcal{I}(\gamma) + \log\bigg(\int \exp\big[-\hat{\mathcal{E}}(f)\big]\pi(df)\bigg) - \log\bigg[\frac{d\rho}{d\hat{\pi}}(\hat{f})\bigg], \\
&= \int \big[\hat{L}(\hat{f},f) + \gamma^*\bar{R}(f)\big]\pi^*_{-\gamma^*\bar{R}}(df) \\
&\qquad + \log \int \exp\big[L^\flat(\hat{f},f) - \hat{L}(\hat{f},f)\big]\pi^*(df) - \gamma\bar{R}(\hat{f}) \\
&\qquad - \mathcal{I}^*(\gamma^*) + \mathcal{I}(\gamma) - \hat{\mathcal{E}}(\hat{f}) - \log\bigg[\frac{d\rho}{d\pi}(\hat{f})\bigg], \\
&\leq \log \int \exp\big[L^\flat(\hat{f},f) - \hat{L}(\hat{f},f)\big]\pi^*_{-\gamma^*\bar{R}}(df)(df) \\
&\qquad - \gamma\bar{R}(\hat{f}) + \mathcal{I}(\gamma) - \log\bigg[\frac{d\rho}{d\pi}(\hat{f})\bigg] \\
&= \log \int \exp\big[L^\flat(\hat{f},f) - \hat{L}(\hat{f},f)\big]\pi^*_{-\gamma^*\bar{R}}(df) + \log\bigg[\frac{d\pi_{-\gamma\bar{R}}}{d\rho}(\hat{f})\bigg],
\end{aligned}$$

hence, by using Fubini's inequality and the equality

$$\mathbb{E}\Big\{\exp\big[-\hat{L}(\hat{f},f)\big]\Big\} = \exp\big[-L^\flat(\hat{f},f)\big],$$

we obtain $\mathbb{E}\int \exp\big[V_1(\hat{f})\big]\rho(\hat{f})$

$$\leq \mathbb{E}\int\bigg(\int \exp\big[L^\flat(\hat{f},f)-\hat{L}(\hat{f},f)\big]\pi^*_{-\gamma^*\bar{R}}(df)\bigg)\pi_{-\gamma\bar{R}}(d\hat{f})$$
$$= \int\bigg(\int \mathbb{E}\exp\big[L^\flat(\hat{f},f)-\hat{L}(\hat{f},f)\big]\pi^*_{-\gamma^*\bar{R}}(df)\bigg)\pi_{-\gamma\bar{R}}(d\hat{f}) = 1.$$

6.4.2. Proof of $\mathbb{E}\Big[\int \exp(V_2)\rho(d\hat{f})\Big] \leq 1$. It relies on the following result.

LEMMA 6.11 *Let $\mathcal{W}$ be a real-valued measurable function defined on a product space $\mathcal{A}_1 \times \mathcal{A}_2$ and let μ_1 and μ_2 be probability distributions on respectively $\mathcal{A}_1$ and $\mathcal{A}_2$.*

- *if* $\mathbb{E}_{a_1\sim\mu_1}\Big\{\log\Big[\mathbb{E}_{a_2\sim\mu_2}\big\{\exp\big[-\mathcal{W}(a_1,a_2)\big]\big\}\Big]\Big\} < +\infty$*, then we have*

$$-\mathbb{E}_{a_1\sim\mu_1}\Big\{\log\Big[\mathbb{E}_{a_2\sim\mu_2}\big\{\exp\big[-\mathcal{W}(a_1,a_2)\big]\big\}\Big]\Big\} \leq -\log\Big\{\mathbb{E}_{a_2\sim\mu_2}\Big[\exp\big[-\mathbb{E}_{a_1\sim\mu_1}\,\mathcal{W}(a_1,a_2)\big]\Big]\Big\}.$$

- *if* $\mathcal{W} > 0$ *on* $\mathcal{A}_1\times\mathcal{A}_2$ *and* $\mathbb{E}_{a_2\sim\mu_2}\Big\{\mathbb{E}_{a_1\sim\mu_1}\big[\mathcal{W}(a_1,a_2)\big]^{-1}\Big\}^{-1} < +\infty$*, then*

$$\mathbb{E}_{a_1\sim\mu_1}\Big\{\mathbb{E}_{a_2\sim\mu_2}\Big[\mathcal{W}(a_1,a_2)^{-1}\Big]^{-1}\Big\} \leq \mathbb{E}_{a_2\sim\mu_2}\Big\{\mathbb{E}_{a_1\sim\mu_1}\big[\mathcal{W}(a_1,a_2)\big]^{-1}\Big\}^{-1}.$$

PROOF.

- Let $\mathcal{A}$ be a measurable space and $\mathcal{M}$ denote the set of probability distributions on $\mathcal{A}$. The Kullback-Leibler divergence between a distribution ρ and a distribution μ is

$$K(\rho,\mu) \triangleq \begin{cases}\mathbb{E}_{a\sim\rho}\,\log\left[\dfrac{d\rho}{d\mu}(a)\right] & \text{if } \rho \ll \mu,\\ +\infty & \text{otherwise,}\end{cases} \tag{6.29}$$

  where $\dfrac{d\rho}{d\mu}$ denotes as usual the density of ρ w.r.t. μ. The Kullback-Leibler divergence satisfies the duality formula (see, e.g., [8, page 159]): for any real-valued measurable function h defined on $\mathcal{A}$,

$$\inf_{\rho\in\mathcal{M}}\big\{\mathbb{E}_{a\sim\rho}\,h(a)+K(\rho,\mu)\big\} = -\log\mathbb{E}_{a\sim\mu}\Big\{\exp\big[-h(a)\big]\Big\}. \tag{6.30}$$

By using twice (6.30) and Fubini's theorem, we have

$$\begin{aligned}
&-\mathbb{E}_{a_1\sim\mu_1}\Big\{\log\Big\{\mathbb{E}_{a_2\sim\mu_2}\Big[\exp\big[-\mathcal{W}(a_1,a_2)\big]\Big]\Big\}\Big\}\\
&\qquad=\mathbb{E}_{a_1\sim\mu_1}\Big\{\inf_{\rho}\big\{\mathbb{E}_{a_2\sim\rho}\big[\mathcal{W}(a_1,a_2)\big]+K(\rho,\mu_2)\big\}\Big\}\\
&\qquad\le\inf_{\rho}\Big\{\mathbb{E}_{a_1\sim\mu_1}\Big[\mathbb{E}_{a_2\sim\rho}\big[\mathcal{W}(a_1,a_2)\big]+K(\rho,\mu_2)\Big]\Big\}\\
&\qquad=-\log\Big\{\mathbb{E}_{a_2\sim\mu_2}\Big[\exp\big\{-\mathbb{E}_{a_1\sim\mu_1}\big[\mathcal{W}(a_1,a_2)\big]\big\}\Big]\Big\}.
\end{aligned}$$

- By using twice (6.30) and the first assertion of Lemma 6.11, we have

$$\begin{aligned}
&\mathbb{E}_{a_1\sim\mu_1}\Big\{\mathbb{E}_{a_2\sim\mu_2}\Big[\mathcal{W}(a_1,a_2)^{-1}\Big]^{-1}\Big\}\\
&\quad=\mathbb{E}_{a_1\sim\mu_1}\Big\{\exp\Big\{-\log\Big[\mathbb{E}_{a_2\sim\mu_2}\big\{\exp\big[-\log\mathcal{W}(a_1,a_2)\big]\big\}\Big]\Big\}\Big\}\\
&\quad=\mathbb{E}_{a_1\sim\mu_1}\Big\{\exp\Big\{\inf_{\rho}\Big[\mathbb{E}_{a_2\sim\rho}\big\{\log\big[\mathcal{W}(a_1,a_2)\big]\big\}+K(\rho,\mu_2)\Big]\Big\}\Big\}\\
&\quad\le\inf_{\rho}\Big\{\exp\big[K(\rho,\mu_2)\big]\mathbb{E}_{a_1\sim\mu_1}\Big\{\exp\Big\{\mathbb{E}_{a_2\sim\rho}\Big[\log\big[\mathcal{W}(a_1,a_2)\big]\Big]\Big\}\Big\}\\
&\quad\le\inf_{\rho}\Big\{\exp\big[K(\rho,\mu_2)\big]\exp\Big\{\mathbb{E}_{a_2\sim\rho}\Big\{\log\Big[\mathbb{E}_{a_1\sim\mu_1}\big[\mathcal{W}(a_1,a_2)\big]\Big]\Big\}\Big\}\\
&\quad=\exp\Big\{\inf_{\rho}\Big\{\mathbb{E}_{a_2\sim\rho}\Big[\log\big\{\mathbb{E}_{a_1\sim\mu_1}\big[\mathcal{W}(a_1,a_2)\big]\big\}\Big]+K(\rho,\mu_2)\Big\}\Big\}\\
&\quad=\exp\Big\{-\log\Big\{\mathbb{E}_{a_2\sim\mu_2}\Big\{\exp\Big[-\log\big\{\mathbb{E}_{a_1\sim\mu_1}\big[\mathcal{W}(a_1,a_2)\big]\big\}\Big]\Big\}\Big\}\Big\}\\
&\quad=\mathbb{E}_{a_2\sim\mu_2}\Big\{\mathbb{E}_{a_1\sim\mu_1}\big[\mathcal{W}(a_1,a_2)\big]^{-1}\Big\}^{-1}.\quad\square
\end{aligned}$$

From Lemma 6.11 and Fubini's theorem, since V_2 does not depend on $\hat{f}$, we have

$$\begin{aligned}
\mathbb{E}\Big[\int\exp(V_2)\rho(d\hat{f})\Big]&=\mathbb{E}\big[\exp(V_2)\big]\\
&=\int\exp\big[-\mathcal{E}^\sharp(f)\big]\pi(df)\mathbb{E}\Big\{\Big[\int\exp\big[-\hat{\mathcal{E}}(f)\big]\pi(df)\Big]^{-1}\Big\}\\
&\le\int\exp\big[-\mathcal{E}^\sharp(f)\big]\pi(df)\Big\{\int\mathbb{E}\Big[\exp\big[\hat{\mathcal{E}}(f)\big]\Big]^{-1}\pi(df)\Big\}^{-1}\\
&=\int\exp\big[-\mathcal{E}^\sharp(f)\big]\pi(df)\Big\{\int\mathbb{E}\Big[\int\exp\big[\hat{L}(f,f')\big]\pi^*(df')\Big]^{-1}\pi(df)\Big\}^{-1}\\
&=\int\exp\big[-\mathcal{E}^\sharp(f)\big]\pi(df)\Big\{\int\Big[\int\exp\big[L^\sharp(f,f')\big]\pi^*(df')\Big]^{-1}\pi(df)\Big\}^{-1}=1.
\end{aligned}$$

This concludes the proof that for any $\gamma \geq 0$, $\gamma^* \geq 0$ and $\varepsilon > 0$, with probability (with respect to the distribution $P^{\otimes n}\rho$ generating the observations $Z_1, \ldots, Z_n$ and the randomized prediction function $\hat{f}$) at least $1 - 2\varepsilon$:

$$V_1(\hat{f}) + V_2 \leq 2\log(\varepsilon^{-1}).$$

6.5. PROOF OF LEMMA 5.3. Let us look at $\mathcal{F}$ from the point of view of f^*. Precisely let $\mathcal{S}_{\mathbb{R}^d}(O,1)$ be the sphere of $\mathbb{R}^d$ centered at the origin and with radius 1 and

$$\mathcal{S} = \big\{ \sum_{j=1}^{d} \theta_j \varphi_j ; (\theta_1, \ldots, \theta_d) \in \mathcal{S}_{\mathbb{R}^d}(O,1) \big\}.$$

Introduce

$$\Omega = \big\{ \phi \in \mathcal{S} ; \exists u > 0 \text{ s.t. } f^* + u\phi \in \mathcal{F} \big\}.$$

For any $\phi \in \Omega$, let $u_\phi = \sup\{u > 0 : f^* + u\phi \in \mathcal{F}\}$. Since π is the uniform distribution on the convex set $\mathcal{F}$ (i.e., the one coming from the uniform distribution on Θ), we have

$$\int \exp\big\{-\alpha[R(f) - R(f^*)]\big\}\pi(df) = \int_{\phi\in\Omega} \int_0^{u_\phi} \exp\big\{-\alpha[R(f^* + u\phi) - R(f^*)]\big\} u^{d-1} du d\phi.$$

Let $c_\phi = \mathbb{E}[\phi(X)\tilde{\ell}'_Y(f^*(X))]$ and $a_\phi = \mathbb{E}\big[\phi^2(X)\big]$. Since

$$f^* \in \operatorname{argmin}_{f\in\mathcal{F}} \mathbb{E}\big\{\tilde{\ell}_Y\big[f(X)\big]\big\},$$

we have $c_\phi \geq 0$ (and $c_\phi = 0$ if both $-\phi$ and ϕ belong to Ω). Moreover from Taylor's expansion,

$$\frac{b_1 a_\phi u^2}{2} \leq R(f^* + u\phi) - R(f^*) - uc_\phi \leq \frac{b_2 a_\phi u^2}{2}.$$

Introduce

$$\psi_\phi = \frac{\int_0^{u_\phi} \exp\big\{-\alpha[uc_\phi + \frac{1}{2}b_1 a_\phi u^2]\big\} u^{d-1} du}{\int_0^{u_\phi} \exp\big\{-\beta[uc_\phi + \frac{1}{2}b_2 a_\phi u^2]\big\} u^{d-1} du}.$$

For any $0 < \alpha < \beta$, we have

$$\frac{\int \exp\big\{-\alpha[R(f) - R(f^*)]\big\}\pi(df)}{\int \exp\big\{-\beta[R(f) - R(f^*)]\big\}\pi(df)} \leq \inf_{\phi\in\mathcal{S}} \psi_\phi.$$

For any $\zeta > 1$, by a change of variable,

$$\begin{aligned}\psi_\phi &< \zeta^d \frac{\int_0^{u_\phi} \exp\{-\alpha[\zeta u c_\phi + \frac{1}{2} b_1 a_\phi \zeta^2 u^2]\} u^{d-1} du}{\int_0^{u_\phi} \exp\{-\beta[u c_\phi + \frac{1}{2} b_2 a_\phi u^2]\} u^{d-1} du} \\ &\leq \zeta^d \sup_{u>0} \exp\{\beta[u c_\phi + \tfrac{1}{2} b_2 a_\phi u^2] - \alpha[\zeta u c_\phi + \tfrac{1}{2} b_1 a_\phi \zeta^2 u^2]\}.\end{aligned}$$

By taking $\zeta = \sqrt{(b_2\beta)/(b_1\alpha)}$ when $c_\phi = 0$ and $\zeta = \sqrt{(b_2\beta)/(b_1\alpha)} \vee (\beta/\alpha)$ otherwise, we obtain $\psi_\phi < \zeta^d$, hence

$$\log\left(\frac{\int \exp\{-\alpha[R(f) - R(f^*)]\}\pi(df)}{\int \exp\{-\beta[R(f) - R(f^*)]\}\pi(df)}\right) \leq \begin{cases} \dfrac{d}{2} \log\big(\dfrac{b_2\beta}{b_1\alpha}\big) \text{ when } \sup\limits_{\phi\in\Omega} c_\phi = 0, \\ d \log\Big(\sqrt{\dfrac{b_2\beta}{b_1\alpha}} \vee \dfrac{\beta}{\alpha}\Big) \text{ otherwise,} \end{cases}$$

which proves the announced result.

6.6. PROOF OF LEMMA 5.4. For $-(2AH)^{-1} \leq \lambda \leq (2AH)^{-1}$, introduce the random variables

$$F = f(X) \qquad F^* = f^*(X),$$

$$\Omega = \tilde{\ell}'_Y(F^*) + (F - F^*) \int_0^1 (1-t) \tilde{\ell}''_Y(F^* + t(F - F^*)) dt,$$

$$L = \lambda[\tilde{\ell}(Y, F) - \tilde{\ell}(Y, F^*)],$$

and the quantities

$$a(\lambda) = \frac{M^2 A^2 \exp(H b_2/A)}{2\sqrt{\pi}(1 - |\lambda| AH)}$$

and

$$\tilde{A} = H b_2/2 + A \log(M) = \frac{A}{2} \log\big\{M^2 \exp\big[H b_2/(2A)\big]\big\}.$$

From Taylor-Lagrange formula, we have

$$L = \lambda(F - F^*)\Omega.$$

Since $\mathbb{E}\big[\exp\big(|\Omega|/A\big) \,|\, X\big] \leq M \exp\big[H b_2/(2A)\big]$, Lemma D.2 gives

$$\log\Big\{\mathbb{E}\Big[\exp\big\{\alpha[\Omega - \mathbb{E}(\Omega|X)]/A\big\} \,|\, X\Big]\Big\} \leq \frac{M^2 \alpha^2 \exp\big(H b_2/A\big)}{2\sqrt{\pi}(1 - |\alpha|)}$$

for any $-1 < \alpha < 1$, and

$$\big|\mathbb{E}(\Omega|X)\big| \leq \tilde{A}. \tag{6.31}$$

By considering $\alpha = A\lambda[f(x) - f^*(x)] \in [-1/2; 1/2]$ for fixed $x \in \mathcal{X}$, we get

$$\log\Big\{\mathbb{E}\Big[\exp\big[L - \mathbb{E}(L|X)\big] \,|\, X\Big]\Big\} \leq \lambda^2 (F - F^*)^2 a(\lambda). \tag{6.32}$$

Let us put moreover

$$\tilde{L} = \mathbb{E}(L|X) + a(\lambda)\lambda^2 (F - F^*)^2.$$

Since $-(2AH)^{-1} \leq \lambda \leq (2AH)^{-1}$, we have $\tilde{L} \leq |\lambda| H\tilde{A} + a(\lambda)\lambda^2 H^2 \leq b'$ with $b' = \tilde{A}/(2A) + M^2 \exp\big(Hb_2/A\big)/(4\sqrt{\pi})$. Since $L - \mathbb{E}(L) = L - \mathbb{E}(L|X) + \mathbb{E}(L|X) - \mathbb{E}(L)$, by using Lemma D.1, (6.32) and (6.31), we obtain

$$\begin{aligned}
\log\Big\{\mathbb{E}\Big[\exp\big[L - \mathbb{E}(L)\big]\Big]\Big\} &\leq \log\Big\{\mathbb{E}\Big[\exp\big[\tilde{L} - \mathbb{E}(\tilde{L})\big]\Big]\Big\} + \lambda^2 a(\lambda)\mathbb{E}\big[(F - F^*)^2\big] \\
&\leq \mathbb{E}\big(\tilde{L}^2\big) g(b') + \lambda^2 a(\lambda)\mathbb{E}\big[(F - F^*)^2\big] \\
&\leq \lambda^2 \mathbb{E}\big[(F - F^*)^2\big]\big[\tilde{A}^2 g(b') + a(\lambda)\big],
\end{aligned}$$

with $g(u) = \big[\exp(u) - 1 - u\big]/u^2$. Computations show that for any $-(2AH)^{-1} \leq \lambda \leq (2AH)^{-1}$,

$$\tilde{A}^2 g(b') + a(\lambda) \leq \frac{A^2}{4} \exp\Big[M^2 \exp\big(Hb_2/A\big)\Big].$$

Consequently, for any $-(2AH)^{-1} \leq \lambda \leq (2AH)^{-1}$, we have

$$\begin{aligned}
&\log\Big\{\mathbb{E}\Big[\exp\big\{\lambda[\tilde{\ell}(Y, F) - \tilde{\ell}(Y, F^*)]\big\}\Big]\Big\} \\
&\qquad \leq \lambda[R(f) - R(f^*)] + \lambda^2 \mathbb{E}\big[(F - F^*)^2\big] \frac{A^2}{4} \exp\Big[M^2 \exp\big(Hb_2/A\big)\Big].
\end{aligned}$$

Now it remains to notice that $\mathbb{E}\big[(F - F^*)^2\big] \leq 2[R(f) - R(f^*)]/b_1$. Indeed consider the function $\phi(t) = R(f^* + t(f - f^*)) - R(f^*)$, where $f \in \mathcal{F}$ and $t \in [0; 1]$. From the definition of f^* and the convexity of $\mathcal{F}$, we have $\phi \geq 0$ on $[0; 1]$. Besides we have $\phi(t) = \phi(0) + t\phi'(0) + \frac{t^2}{2}\phi''(\zeta_t)$ for some $\zeta_t \in]0; 1[$. So we have $\phi'(0) \geq 0$, and using the lower bound on the convexity, we obtain for $t = 1$

$$\frac{b_1}{2}\mathbb{E}(F - F^*)^2 \leq R(f) - R(f^*). \tag{6.33}$$

6.7. PROOF OF LEMMA 5.6. We have

$$\begin{aligned}
&\mathbb{E}\Big(\big\{[Y - f(X)]^2 - [Y - f^*(X)]^2\big\}^2\Big) \\
&= \mathbb{E}\Big([f^* - f(X)]^2\big\{2[Y - f^*(X)] + [f^* - f(X)]\big\}^2\Big)
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\Big([f^* - f(X)]^2 \big\{ 4\mathbb{E}\big([Y - f^*(X)]^2 \big| X \big) \\
&\qquad\qquad + 4\mathbb{E}(Y - f^*(X)|X)[f^*(X) - f(X)] + [f^*(X) - f(X)]^2 \big\} \Big) \\
&\leq \mathbb{E}\Big([f^* - f(X)]^2 \big\{ 4\sigma^2 + 4\sigma |f^*(X) - f(X)| + [f^*(X) - f(X)]^2 \big\} \Big) \\
&\leq \mathbb{E}\Big([f^* - f(X)]^2 (2\sigma + H)^2 \Big) \\
&\leq (2\sigma + H)^2 [R(f) - R(f^*)],
\end{aligned}
$$

where the last inequality is the usual relation between excess risk and L^2 distance using the convexity of $\mathcal{F}$ (see above (6.33) for a proof).

6.8. PROOF OF LEMMA 5.7. Let $\mathcal{S} = \{s \in \mathcal{F}_{\text{lin}} : \mathbb{E}[s(X)^2] = 1\}$. Using the triangular inequality in $\mathbb{L}^2$, we get

$$
\begin{aligned}
&\mathbb{E}\Big(\big\{ [Y - f(X)]^2 - [Y - f^*(X)]^2 \big\}^2 \Big) \\
&= \mathbb{E}\Big(\big\{ 2[f^* - f(X)][Y - f^*(X)] + [f^*(X) - f(X)]^2 \big\}^2 \Big) \\
&\leq \Big(2\sqrt{\mathbb{E}\big\{ [f^*(X) - f(X)]^2 [Y - f^*(X)]^2 \big\}} + \sqrt{\mathbb{E}\big\{ [f^*(X) - f(X)]^4 \big\}} \Big)^2 \\
&\leq \Bigg[2\sqrt{\mathbb{E}\big([f^*(X) - f(X)]^2 \big)} \sqrt{\sup_{s \in \mathcal{S}} \mathbb{E}\big(s^2(X) [Y - f^*(X)]^2 \big)} \\
&\qquad + \mathbb{E}\big([f^*(X) - f(X)]^2 \big) \sqrt{\sup_{s \in \mathcal{S}} \mathbb{E}\big[s^4(X) \big]} \Bigg]^2 \\
&\leq V[R(f) - R(f^*)],
\end{aligned}
$$

with

$$
\begin{aligned}
V = \Bigg[& 2\sqrt{\sup_{s \in \mathcal{S}} \mathbb{E}\big(s^2(X) [Y - f^*(X)]^2 \big)} \\
& + \sqrt{\sup_{f', f'' \in \mathcal{F}} \mathbb{E}\big([f'(X) - f''(X)]^2 \big)} \sqrt{\sup_{s \in \mathcal{S}} \mathbb{E}\big[s^4(X) \big]} \Bigg]^2,
\end{aligned}
$$

where the last inequality is the usual relation between excess risk and L^2 distance using the convexity of $\mathcal{F}$ (see above (6.33) for a proof).

A. UNIFORMLY BOUNDED CONDITIONAL VARIANCE IS NECESSARY TO REACH d/n RATE

In this section, we will see that the target (0.2) cannot be reached if we just assume that Y has a finite variance and that the functions in $\mathcal{F}$ are bounded.

For this, consider an input space $\mathcal{X}$ partitioned into two sets $\mathcal{X}_1$ and $\mathcal{X}_2$: $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ and $\mathcal{X}_1 \cap \mathcal{X}_2 = \emptyset$. Let $\varphi_1(x) = 1_{x \in \mathcal{X}_1}$ and $\varphi_2(x) = 1_{x \in \mathcal{X}_2}$. Let $\mathcal{F} = \big\{\theta_1\varphi_1 + \theta_2\varphi_2; (\theta_1, \theta_2) \in [-1, 1]^2\big\}$.

THEOREM A.1 *For any estimator $\hat{f}$ and any training set size $n \geq 1$, we have*

$$\sup_P \big\{\mathbb{E}R(\hat{f}) - R(f^*)\big\} \geq \frac{1}{4\sqrt{n}}, \tag{A.1}$$

where the supremum is taken with respect to all probability distributions such that $f^{(\text{reg})} \in \mathcal{F}$ and $\mathbb{V}\text{ar}\, Y \leq 1$.

PROOF. Let β satisfying $0 < \beta \leq 1$ be some parameter to be chosen later. Let P_σ, $\sigma \in \{-, +\}$, be two probability distributions on $\mathcal{X} \times \mathbb{R}$ such that for any $\sigma \in \{-, +\}$,

$$P_\sigma(\mathcal{X}_1) = 1 - \beta,$$
$$P_\sigma(Y = 0 | X = x) = 1 \qquad \text{for any } x \in \mathcal{X}_1,$$

and

$$P_\sigma\Big(Y = \frac{1}{\sqrt{\beta}} \,|\, X = x\Big) = \frac{1 + \sigma\sqrt{\beta}}{2}$$
$$= 1 - P_\sigma\Big(Y = -\frac{1}{\sqrt{\beta}} \,|\, X = x\Big) \quad \text{for any } x \in \mathcal{X}_2.$$

One can easily check that for any $\sigma \in \{-, +\}$, $\mathbb{V}\text{ar}_{P_\sigma}(Y) = 1 - \beta^2 \leq 1$ and $f^{(\text{reg})}(x) = \sigma\varphi_2 \in \mathcal{F}$. To prove Theorem A.1, it suffices to prove (A.1) when the supremum is taken among $P \in \{P_-, P_+\}$. This is done by applying Theorem 8.2 of [3]. Indeed, the pair (P_-, P_+) forms a $(1, \beta, \beta)$-hypercube in the sense of Definition 8.2 with edge discrepancy of type I (see (8.5), (8.11) and (10.20) for $q = 2$): $d_I = 1$. We obtain

$$\sup_{P \in \{P_-, P_+\}} \big\{\mathbb{E}R(\hat{f}) - R(f^*)\big\} \geq \beta(1 - \beta\sqrt{n}),$$

which gives the desired result by taking $\beta = 1/(2\sqrt{n})$. □

B. EMPIRICAL RISK MINIMIZATION ON A BALL: ANALYSIS DERIVED FROM THE WORK OF BIRGÉ AND MASSART

We will use the following covering number upper bound [16, Lemma 1]

LEMMA B.1 *If $\mathcal{F}$ has a diameter $H > 0$ for L^∞-norm (i.e., $\sup_{f_1,f_2\in\mathcal{F},x\in\mathcal{X}} |f_1(x)-f_2(x)| = H$), then for any $0 < \delta \leq H$, there exists a set $\mathcal{F}^\# \subset \mathcal{F}$, of cardinality $|\mathcal{F}^\#| \leq (3H/\delta)^d$ such that for any $f \in \mathcal{F}$ there exists $g \in \mathcal{F}^\#$ such that $\|f-g\|_\infty \leq \delta$.*

We apply a slightly improved version of Theorem 5 in Birgé and Massart [5]. First for homogeneity purpose, we modify Assumption M2 by replacing the condition "$\sigma^2 \geq D/n$" by "$\sigma^2 \geq B^2 D/n$" where the constant B is the one appearing in (5.3) of [5]. This modifies Theorem 5 of [5] to the extent that "$\vee 1$" should be replaced with "$\vee B^2$". Our second modification is to remove the assumption that W_i and X_i are independent. A careful look at the proof shows that the result still holds when (5.2) is replaced by: for any $x \in \mathcal{X}$, and $m \geq 2$

$$\mathrm{E}_s[M^m(W_i)|X_i = x] \leq a_m A^m, \qquad \text{for all } i = 1,\dots,n$$

We consider $W = Y - f^*(X)$, $\gamma(z,f) = (y-f(x))^2$, $\Delta(x,u,v) = |u(x)-v(x)|$, and $M(w) = 2(|w| + H)$. From (1.7), for all $m \geq 2$, we have $\mathbb{E}\big\{[(2(|W| + H)]^m|X = x\big] \leq \frac{m!}{2}[4M(A+H)]^m$. Now consider B' and r such that Assumption M2 of [5] holds for $D = d$. Inequality (5.8) for $\tau = 1/2$ of [5] implies that for any $v \geq \kappa\frac{d}{n}(A^2 + H^2)\log(2B' + B'r\sqrt{d/n})$, with probability at least $1 - \kappa\exp\Big[\dfrac{-nv}{\kappa(A^2+H^2)}\Big]$,

$$R(\hat{f}^{(\mathrm{erm})}) - R(f^*) + r(f^*) - r(\hat{f}^{(\mathrm{erm})}) \leq \big(\mathbb{E}\big\{\big[\hat{f}^{(\mathrm{erm})}(X) - f^*(X)\big]^2\big\} \vee v\big)/2$$

for some large enough constant κ depending on M. Now from Proposition 1 of [5] and Lemma B.1, one can take either $B' = 6$ and $r\sqrt{d} = \sqrt{\tilde{B}}$ or $B' = 3\sqrt{n/d}$ and $r = 1$. By using $\mathbb{E}\big\{\big[\hat{f}^{(\mathrm{erm})}(X) - f^*(X)\big]^2\big\} \leq R(\hat{f}^{(\mathrm{erm})}) - R(f^*)$ (since $\mathcal{F}$ is convex and f^* is the orthogonal projection of Y on $\mathcal{F}$), and $r(f^*) - r(\hat{f}^{(\mathrm{erm})}) \geq 0$ (by definition of $\hat{f}^{(\mathrm{erm})}$), the desired result can be derived.

Theorem 1.5 provides a d/n rate provided that the geometrical quantity $\tilde{B}$ is at most of order n. Inequality (3.2) of [5] allows to bracket $\tilde{B}$ in terms of $B = \sup_{f\in\mathrm{span}\{\varphi_1,\dots,\varphi_d\}} \|f\|_\infty^2/\mathbb{E}[f(X)]^2$, namely $B \leq \tilde{B} \leq Bd$. To understand better how this quantity behaves and to illustrate some of the presented results, let us give the following simple example.

Example 1. Let $A_1,\dots,A_d$ be a partition of $\mathcal{X}$, i.e., $\mathcal{X} = \sqcup_{j=1}^d A_j$. Now consider the indicator functions $\varphi_j = 1_{A_j}, j = 1,\dots,d$: φ_j is equal to 1 on A_j and zero elsewhere. Consider that X and Y are independent and that Y is a Gaussian random variable with mean θ and variance σ^2. In this situation: $f^*_{\mathrm{lin}} = f^{(\mathrm{reg})} = \sum_{j=1}^d \theta\varphi_j$. According to Theorem 1.1, if we know an upper bound H on $\|f^{(\mathrm{reg})}\|_\infty = \theta$, we have that the truncated estimator $(\hat{f}^{(\mathrm{ols})} \wedge H) \vee -H$ satisfies

$$\mathbb{E}R(\hat{f}_H^{(\mathrm{ols})}) - R(f^*_{\mathrm{lin}}) \leq \kappa\frac{(\sigma^2 \vee H^2)d\log n}{n}$$

for some numerical constant κ. Let us now apply Theorem C.1. Introduce $p_j = \mathbb{P}(X \in A_j)$ and $p_{\min} = \min_j p_j$. We have $Q = \big(\mathbb{E}\varphi_j(X)\varphi_k(X)\big)_{j,k} = \mathrm{Diag}(p_j)$, $\mathcal{K} = 1$ and $\|\theta^*\| = \theta\sqrt{d}$. We can take $A = \sigma$ and $M = 2$. From Theorem C.1, for $\lambda = d\mathcal{L}_\varepsilon/n$, as soon as $\lambda \le p_{\min}$, the ridge regression estimator satisfies with probability at least $1-\varepsilon$:

$$R(\hat{f}^{(\text{ridge})}) - R(f^*_{\text{lin}}) \le \kappa\mathcal{L}_\varepsilon \frac{d}{n}\left(\sigma^2 + \frac{\theta^2 d^2 \mathcal{L}_\varepsilon^2}{n p_{\min}}\right) \tag{B.1}$$

for some numerical constant κ. When d is large, the term $(d^2\mathcal{L}_\varepsilon^2)/(np_{\min})$ is felt, and leads to suboptimal rates. Specifically, since $p_{\min} \le 1/d$, the r.h.s. of (B.1) is greater than d^4/n^2, which is much larger than d/n when d is much larger than $n^{1/3}$. If Y is not Gaussian but almost surely uniformly bounded by $C < +\infty$, then the randomized estimator proposed in Theorem 1.3 satisfies the nicer property: with probability at least $1-\varepsilon$,

$$R(\hat{f}) - R(f^*_{\text{lin}}) \le \kappa(H^2 + C^2)\frac{d\log(3p_{\min}^{-1}) + \log((\log n)\varepsilon^{-1})}{n},$$

for some numerical constant κ. In this example, one can check that $\tilde{B} = \tilde{B}' = 1/p_{\min}$ where $p_{\min} = \min_j \mathbb{P}(X \in A_j)$. As long as $p_{\min} \ge 1/n$, the target (0.1) is reached from Corollary 1.5. Otherwise, without this assumption, the rate is in $(d\log(n/d))/n$. ■

C. Ridge regression analysis from the work of Caponnetto and De Vito

From [6], one can derive the following risk bound for the ridge estimator.

THEOREM C.1 *Let $q_{\min}$ be the smallest eigenvalue of the $d \times d$-product matrix $Q = \big(\mathbb{E}\varphi_j(X)\varphi_k(X)\big)_{j,k}$. Let $\mathcal{K} = \sup_{x\in\mathcal{X}} \sum_{j=1}^d \varphi_j(x)^2$. Let $\|\theta^*\|$ be the Euclidean norm of the vector of parameters of $f^*_{\text{lin}} = \sum_{j=1}^d \theta^*_j\varphi_j$. Let $0 < \varepsilon < 1/2$ and $\mathcal{L}_\varepsilon = \log^2(\varepsilon^{-1})$. Assume that for any $x \in \mathcal{X}$,*

$$\mathbb{E}\Big\{\exp\big[|Y - f^*_{\text{lin}}(X)|/A\big] \,|\, X = x\Big\} \le M.$$

For $\lambda = (\mathcal{K}d\mathcal{L}_\varepsilon)/n$, if $\lambda \le q_{\min}$, the ridge regression estimator satisfies with probability at least $1-\varepsilon$:

$$R(\hat{f}^{(\text{ridge})}) - R(f^*_{\text{lin}}) \le \frac{\kappa\mathcal{L}_\varepsilon d}{n}\left(A^2 + \frac{\lambda}{q_{\min}}\mathcal{K}\mathcal{L}_\varepsilon\|\theta^*\|^2\right) \tag{C.1}$$

for some positive constant κ depending only on M.

PROOF. One can check that $\hat{f}^{(\text{ridge})} \in \operatorname{argmin}_{f\in\mathcal{H}} r(f) + \lambda \sum_{j=1}^{d} \|f\|_{\mathcal{H}}^2$, where $\mathcal{H}$ is the reproducing kernel Hilbert space associated with the kernel $K : (x, x') \mapsto \sum_{j=1}^{d} \varphi_j(x)\varphi_k(x')$. Introduce $f^{(\lambda)} \in \operatorname{argmin}_{f\in\mathcal{H}} R(f) + \lambda \sum_{j=1}^{d} \|f\|_{\mathcal{H}}^2$. Let us use Theorem 4 in [6] and the notation defined in their Section 5.2. Let φ be the column vector of functions $[\varphi_j]_{j=1}^d$, $\text{Diag}(a_j)$ denote the diagonal $d \times d$-matrix whose j-th element on the diagonal is a_j, and I_d be the $d \times d$-identity matrix. Let U and $q_1, \ldots, q_d$ be such that $UU^T = I$ and $Q = U\text{Diag}(q_j)U^T$. We have $f^*_{\text{lin}} = \varphi^T\theta^*$ and $f^{(\lambda)} = \varphi^T(Q + \lambda I)^{-1}Q\theta^*$, hence

$$f^*_{\text{lin}} - f^{(\lambda)} = \varphi^T U\text{Diag}(\lambda/(q_j + \lambda))U^T\theta^*.$$

After some computations, we obtain that the residual, reconstruction error and effective dimension respectively satisfy $\mathcal{A}(\lambda) \leq \frac{\lambda^2}{q_{\min}}\|\theta^*\|^2$, $\mathcal{B}(\lambda) \leq \frac{\lambda^2}{q_{\min}^2}\|\theta^*\|^2$, and $\mathcal{N}(\lambda) \leq d$. The result is obtained by noticing that the leading terms in (34) of [6] are $\mathcal{A}(\lambda)$ and the term with the effective dimension $\mathcal{N}(\lambda)$. □

The dependence in the sample size n is correct since $1/n$ is known to be minimax optimal. The dependence on the dimension d is not optimal, as it is observed in the example given page 66. Besides the high probability bound (C.1) holds only for a regularization parameter λ depending on the confidence level ε. So we do not have a single estimator satisfying a PAC bound for every confidence level. Finally the dependence on the confidence level is larger than expected. It contains an unusual square. The example given page 66 illustrates Theorem C.1.

D. SOME STANDARD UPPER BOUNDS ON LOG-LAPLACE TRANSFORMS

LEMMA D.1 *Let V be a random variable almost surely bounded by $b \in \mathbb{R}$. Let $g : u \mapsto \big[\exp(u) - 1 - u\big]/u^2$.*

$$\log\Big\{\mathbb{E}\Big[\exp\big[V - \mathbb{E}(V)\big]\Big]\Big\} \leq \mathbb{E}\big(V^2\big)g(b).$$

PROOF. Since g is an increasing function, we have $g(V) \leq g(b)$. By using the inequality $\log(1 + u) \leq u$, we obtain

$$\begin{aligned}\log\Big\{\mathbb{E}\Big[\exp\big[V - \mathbb{E}(V)\big]\Big]\Big\} &= -\mathbb{E}(V) + \log\big\{\mathbb{E}\big[1 + V + V^2g(V)\big]\big\} \\ &\leq \mathbb{E}\big[V^2g(V)\big] \leq \mathbb{E}\big(V^2\big)g(b).\end{aligned}$$

□

LEMMA D.2 *Let V be a real-valued random variable such that $\mathbb{E}\big[\exp\big(|V|\big)\big] \leq M$ for some $M > 0$. Then we have $|\mathbb{E}(V)| \leq \log M$, and for any $-1 < \alpha < 1$,*

$$\log\Big\{\mathbb{E}\Big[\exp\big\{\alpha\big[V - \mathbb{E}(V)\big]\big\}\Big]\Big\} \leq \frac{\alpha^2M^2}{2\sqrt{\pi}(1 - |\alpha|)}.$$

PROOF. First note that by Jensen's inequality, we have $|\mathbb{E}(V)| \leq \log(M)$. By using $\log(u) \leq u - 1$ and Stirling's formula, for any $-1 < \alpha < 1$, we have

$$\begin{aligned}
\log\Big\{\mathbb{E}\Big[\exp\big\{\alpha\big[V - \mathbb{E}(V)\big]\big\}\Big]\Big\} &\leq \mathbb{E}\Big[\exp\big\{\alpha\big[V - \mathbb{E}(V)\big]\big\}\Big]\Big\} - 1 \\
&= \mathbb{E}\Big\{\exp\big\{\alpha\big[V - \mathbb{E}(V)\big]\big\} - 1 - \alpha\big[V - \mathbb{E}(V)\big]\Big\} \\
&\leq \mathbb{E}\Big\{\exp\big[|\alpha||V - \mathbb{E}(V)|\big] - 1 - |\alpha||V - \mathbb{E}(V)|\Big\} \\
&\leq \mathbb{E}\Big\{\exp\big[|V - \mathbb{E}(V)|\big]\Big\} \sup_{u\geq 0}\Big\{\big[\exp(|\alpha|u) - 1 - |\alpha|u\big]\exp(-u)\Big\} \\
&\leq \mathbb{E}\Big[\exp\big(|V| + |\mathbb{E}(V)|\big)\Big] \sup_{u\geq 0}\sum_{m\geq 2}\frac{|\alpha|^m u^m}{m!}\exp(-u) \\
&\leq M^2\sum_{m\geq 2}\frac{|\alpha|^m}{m!}\sup_{u\geq 0}u^m\exp(-u) = \alpha^2M^2\sum_{m\geq 2}\frac{|\alpha|^{m-2}}{m!}m^m\exp(-m) \\
&\leq \alpha^2M^2\sum_{m\geq 2}\frac{|\alpha|^{m-2}}{\sqrt{2\pi m}} \leq \frac{\alpha^2M^2}{2\sqrt{\pi}(1-|\alpha|)}.
\end{aligned}$$

□

E. EXPERIMENTAL RESULTS FOR THE MIN-MAX TRUNCATED ESTIMATOR DEFINED IN SECTION 3.3

Table 1: Comparison of the min-max truncated estimator $\hat{f}$ with the ordinary least squares estimator $\hat{f}^{(\text{ols})}$ for the mixture noise (see Section 3.4.1) with $\rho = 0.1$ and $p = 0.005$. In parenthesis, the 95%-confidence intervals for the estimated quantities.

	nb of iterations	nb of iter. with $R(\hat{f}) \neq R(\hat{f}^{(\text{ols})})$	nb of iter. with $R(\hat{f}) < R(\hat{f}^{(\text{ols})})$	$\mathbb{E}R(\hat{f}^{(\text{ols})}) - R(f^*)$	$\mathbb{E}R(\hat{f}) - R(f^*)$	$\mathbb{E}R[(\hat{f}^{(\text{ols})})\|\hat{f} \neq \hat{f}^{(\text{ols})}] - R(f^*)$	$\mathbb{E}[R(\hat{f})\|\hat{f} \neq \hat{f}^{(\text{ols})}] - R(f^*)$
INC(n=200,d=1)	1000	419	405	$0.567(\pm 0.083)$	$0.178(\pm 0.025)$	$1.191(\pm 0.178)$	$0.262(\pm 0.052)$
INC(n=200,d=2)	1000	506	498	$1.055(\pm 0.112)$	$0.271(\pm 0.030)$	$1.884(\pm 0.193)$	$0.334(\pm 0.050)$
HCC(n=200,d=2)	1000	502	494	$1.045(\pm 0.103)$	$0.267(\pm 0.024)$	$1.866(\pm 0.174)$	$0.316(\pm 0.032)$
TS(n=200,d=2)	1000	561	554	$1.069(\pm 0.089)$	$0.310(\pm 0.027)$	$1.720(\pm 0.132)$	$0.367(\pm 0.036)$
INC(n=1000,d=2)	1000	402	392	$0.204(\pm 0.015)$	$0.109(\pm 0.008)$	$0.316(\pm 0.029)$	$0.081(\pm 0.011)$
INC(n=1000,d=10)	1000	950	946	$1.030(\pm 0.041)$	$0.228(\pm 0.016)$	$1.051(\pm 0.042)$	$0.207(\pm 0.014)$
HCC(n=1000,d=10)	1000	942	942	$0.980(\pm 0.038)$	$0.222(\pm 0.015)$	$1.008(\pm 0.039)$	$0.203(\pm 0.015)$
TS(n=1000,d=10)	1000	976	973	$1.009(\pm 0.037)$	$0.228(\pm 0.017)$	$1.018(\pm 0.038)$	$0.217(\pm 0.016)$
INC(n=2000,d=2)	1000	209	207	$0.104(\pm 0.007)$	$0.078(\pm 0.005)$	$0.206(\pm 0.021)$	$0.082(\pm 0.012)$
HCC(n=2000,d=2)	1000	184	183	$0.099(\pm 0.007)$	$0.076(\pm 0.005)$	$0.196(\pm 0.023)$	$0.070(\pm 0.010)$
TS(n=2000,d=2)	1000	172	171	$0.101(\pm 0.007)$	$0.080(\pm 0.005)$	$0.206(\pm 0.020)$	$0.083(\pm 0.012)$
INC(n=2000,d=10)	1000	669	669	$0.510(\pm 0.018)$	$0.206(\pm 0.012)$	$0.572(\pm 0.023)$	$0.117(\pm 0.009)$
HCC(n=2000,d=10)	1000	669	669	$0.499(\pm 0.018)$	$0.207(\pm 0.013)$	$0.561(\pm 0.023)$	$0.125(\pm 0.011)$
TS(n=2000,d=10)	1000	754	753	$0.516(\pm 0.018)$	$0.195(\pm 0.013)$	$0.558(\pm 0.022)$	$0.131(\pm 0.011)$

Table 2: Comparison of the min-max truncated estimator $\hat{f}$ with the ordinary least squares estimator $\hat{f}^{(\text{ols})}$ for the mixture noise (see Section 3.4.1) with $\rho = 0.4$ and $p = 0.005$. In parenthesis, the 95%-confidence intervals for the estimated quantities.

	nb of iterations	nb of iter. with $R(\hat{f}) \neq R(\hat{f}^{(\text{ols})})$	nb of iter. with $R(\hat{f}) < R(\hat{f}^{(\text{ols})})$	$\mathbb{E}R(\hat{f}^{(\text{ols})}) - R(f^*)$	$\mathbb{E}R(\hat{f}) - R(f^*)$	$\mathbb{E}R[(\hat{f}^{(\text{ols})})\|\hat{f} \neq \hat{f}^{(\text{ols})}] - R(f^*)$	$\mathbb{E}[R(\hat{f})\|\hat{f} \neq \hat{f}^{(\text{ols})}] - R(f^*)$
INC(n=200,d=1)	1000	234	211	$0.551(\pm 0.063)$	$0.409(\pm 0.042)$	$1.211(\pm 0.210)$	$0.606(\pm 0.110)$
INC(n=200,d=2)	1000	195	186	$1.046(\pm 0.088)$	$0.788(\pm 0.061)$	$2.174(\pm 0.293)$	$0.848(\pm 0.118)$
HCC(n=200,d=2)	1000	222	215	$1.028(\pm 0.079)$	$0.748(\pm 0.051)$	$2.157(\pm 0.243)$	$0.897(\pm 0.112)$
TS(n=200,d=2)	1000	291	268	$1.053(\pm 0.079)$	$0.805(\pm 0.058)$	$1.701(\pm 0.186)$	$0.851(\pm 0.093)$
INC(n=1000,d=2)	1000	127	117	$0.201(\pm 0.013)$	$0.181(\pm 0.012)$	$0.366(\pm 0.053)$	$0.207(\pm 0.035)$
INC(n=1000,d=10)	1000	262	249	$1.023(\pm 0.035)$	$0.902(\pm 0.030)$	$1.238(\pm 0.081)$	$0.777(\pm 0.054)$
HCC(n=1000,d=10)	1000	201	192	$0.991(\pm 0.033)$	$0.902(\pm 0.031)$	$1.235(\pm 0.088)$	$0.790(\pm 0.067)$
TS(n=1000,d=10)	1000	171	162	$1.009(\pm 0.033)$	$0.951(\pm 0.031)$	$1.166(\pm 0.098)$	$0.825(\pm 0.071)$
INC(n=2000,d=2)	1000	80	77	$0.105(\pm 0.007)$	$0.099(\pm 0.006)$	$0.214(\pm 0.042)$	$0.135(\pm 0.029)$
HCC(n=2000,d=2)	1000	44	42	$0.102(\pm 0.007)$	$0.099(\pm 0.007)$	$0.187(\pm 0.050)$	$0.120(\pm 0.034)$
TS(n=2000,d=2)	1000	47	47	$0.101(\pm 0.007)$	$0.099(\pm 0.007)$	$0.147(\pm 0.032)$	$0.103(\pm 0.026)$
INC(n=2000,d=10)	1000	116	113	$0.511(\pm 0.016)$	$0.491(\pm 0.016)$	$0.611(\pm 0.052)$	$0.437(\pm 0.042)$
HCC(n=2000,d=10)	1000	110	105	$0.500(\pm 0.016)$	$0.481(\pm 0.015)$	$0.602(\pm 0.056)$	$0.430(\pm 0.044)$
TS(n=2000,d=10)	1000	101	98	$0.511(\pm 0.016)$	$0.499(\pm 0.016)$	$0.601(\pm 0.054)$	$0.486(\pm 0.051)$

Table 3: Comparison of the min-max truncated estimator $\hat{f}$ with the ordinary least squares estimator $\hat{f}^{(\text{ols})}$ with the heavy-tailed noise (see Section 3.4.1).

	nb of iterations	nb of iter. with $R(\hat{f}) \neq R(\hat{f}^{(\text{ols})})$	nb of iter. with $R(\hat{f}) < R(\hat{f}^{(\text{ols})})$	$\mathbb{E}R(\hat{f}^{(\text{ols})}) - R(f^*)$	$\mathbb{E}R(\hat{f}) - R(f^*)$	$\mathbb{E}R[(\hat{f}^{(\text{ols})}) \mid \hat{f} \neq \hat{f}^{(\text{ols})}] - R(f^*)$	$\mathbb{E}[R(\hat{f}) \mid \hat{f} \neq \hat{f}^{(\text{ols})}] - R(f^*)$
INC(n=200,d=1)	1000	163	145	$7.72(\pm 3.46)$	$3.92(\pm 0.409)$	$30.52(\pm 20.8)$	$7.20(\pm 1.61)$
INC(n=200,d=2)	1000	104	98	$22.69(\pm 23.14)$	$19.18(\pm 23.09)$	$45.36(\pm 14.1)$	$11.63(\pm 2.19)$
HCC(n=200,d=2)	1000	120	117	$18.16(\pm 12.68)$	$8.07(\pm 0.718)$	$99.39(\pm 105)$	$15.34(\pm 4.41)$
TS(n=200,d=2)	1000	110	105	$43.89(\pm 63.79)$	$39.71(\pm 63.76)$	$48.55(\pm 18.4)$	$10.59(\pm 2.01)$
INC(n=1000,d=2)	1000	104	100	$3.98(\pm 2.25)$	$1.78(\pm 0.128)$	$23.18(\pm 21.3)$	$2.03(\pm 0.56)$
INC(n=1000,d=10)	1000	253	242	$16.36(\pm 5.10)$	$7.90(\pm 0.278)$	$41.25(\pm 19.8)$	$7.81(\pm 0.69)$
HCC(n=1000,d=10)	1000	220	211	$13.57(\pm 1.93)$	$7.88(\pm 0.255)$	$33.13(\pm 8.2)$	$7.28(\pm 0.59)$
TS(n=1000,d=10)	1000	214	211	$18.67(\pm 11.62)$	$13.79(\pm 11.52)$	$30.34(\pm 7.2)$	$7.53(\pm 0.58)$
INC(n=2000,d=2)	1000	113	103	$1.56(\pm 0.41)$	$0.89(\pm 0.059)$	$6.74(\pm 3.4)$	$0.86(\pm 0.18)$
HCC(n=2000,d=2)	1000	105	97	$1.66(\pm 0.43)$	$0.95(\pm 0.062)$	$7.87(\pm 3.8)$	$1.13(\pm 0.23)$
TS(n=2000,d=2)	1000	101	95	$1.59(\pm 0.64)$	$0.88(\pm 0.058)$	$8.03(\pm 6.2)$	$1.04(\pm 0.22)$
INC(n=2000,d=10)	1000	259	255	$8.77(\pm 4.02)$	$4.23(\pm 0.154)$	$21.54(\pm 15.4)$	$4.03(\pm 0.39)$
HCC(n=2000,d=10)	1000	250	242	$6.98(\pm 1.17)$	$4.13(\pm 0.127)$	$15.35(\pm 4.5)$	$3.94(\pm 0.25)$
TS(n=2000,d=10)	1000	238	233	$8.49(\pm 3.61)$	$5.95(\pm 3.486)$	$14.82(\pm 3.8)$	$4.17(\pm 0.30)$

Table 4: Comparison of the min-max truncated estimator $\hat{f}$ with the ordinary least squares estimator $\hat{f}^{(\text{ols})}$ with an asymetric variant of the heavy-tailed noise.

	nb of iterations	nb of iter. with $R(\hat{f}) \neq R(\hat{f}^{(\text{ols})})$	nb of iter. with $R(\hat{f}) < R(\hat{f}^{(\text{ols})})$	$\mathbb{E}R(\hat{f}^{(\text{ols})}) - R(f^*)$	$\mathbb{E}R(\hat{f}) - R(f^*)$	$\mathbb{E}R[(\hat{f}^{(\text{ols})}) \vert \hat{f} \neq \hat{f}^{(\text{ols})}] - R(f^*)$	$\mathbb{E}[R(\hat{f}) \vert \hat{f} \neq \hat{f}^{(\text{ols})}] - R(f^*)$
INC(n=200,d=1)	1000	87	77	$5.49(\pm 3.07)$	$3.00(\pm 0.330)$	$35.44(\pm 34.7)$	$6.85(\pm 2.48)$
INC(n=200,d=2)	1000	70	66	$19.25(\pm 23.23)$	$17.4(\pm 23.2)$	$37.95(\pm 13.1)$	$11.05(\pm 2.87)$
HCC(n=200,d=2)	1000	67	66	$7.19(\pm 0.88)$	$5.81(\pm 0.397)$	$31.52(\pm 10.5)$	$10.87(\pm 2.64)$
TS(n=200,d=2)	1000	76	68	$39.80(\pm 64.09)$	$37.9(\pm 64.1)$	$34.28(\pm 14.8)$	$9.21(\pm 2.05)$
INC(n=1000,d=2)	1000	101	92	$2.81(\pm 2.21)$	$1.31(\pm 0.106)$	$16.76(\pm 21.8)$	$1.88(\pm 0.69)$
INC(n=1000,d=10)	1000	211	195	$10.71(\pm 4.53)$	$5.86(\pm 0.222)$	$29.00(\pm 21.3)$	$6.03(\pm 0.71)$
HCC(n=1000,d=10)	1000	197	185	$8.67(\pm 1.16)$	$5.81(\pm 0.177)$	$20.31(\pm 5.59)$	$5.79(\pm 0.43)$
TS(n=1000,d=10)	1000	258	233	$13.62(\pm 11.27)$	$11.3(\pm 11.2)$	$14.68(\pm 2.45)$	$5.60(\pm 0.36)$
INC(n=2000,d=2)	1000	106	92	$1.04(\pm 0.37)$	$0.64(\pm 0.042)$	$4.54(\pm 3.45)$	$0.79(\pm 0.16)$
HCC(n=2000,d=2)	1000	99	90	$0.90(\pm 0.11)$	$0.66(\pm 0.042)$	$3.23(\pm 0.93)$	$0.82(\pm 0.16)$
TS(n=2000,d=2)	1000	84	81	$1.11(\pm 0.66)$	$0.60(\pm 0.042)$	$6.80(\pm 7.79)$	$0.69(\pm 0.17)$
INC(n=2000,d=10)	1000	238	222	$6.32(\pm 4.18)$	$3.07(\pm 0.147)$	$16.84(\pm 17.5)$	$3.18(\pm 0.51)$
HCC(n=2000,d=10)	1000	221	203	$4.49(\pm 0.98)$	$2.98(\pm 0.091)$	$9.76(\pm 4.39)$	$2.93(\pm 0.22)$
TS(n=2000,d=10)	1000	412	350	$5.93(\pm 3.51)$	$4.59(\pm 3.44)$	$6.07(\pm 1.76)$	$2.84(\pm 0.16)$

Table 5: Comparison of the min-max truncated estimator $\hat{f}$ with the ordinary least squares estimator $\hat{f}^{(\text{ols})}$ for standard Gaussian noise.

	nb of iter.	nb of iter. with $R(\hat{f}) \neq R(\hat{f}^{(\text{ols})})$	nb of iter. with $R(\hat{f}) < R(\hat{f}^{(\text{ols})})$	$\mathbb{E}R(\hat{f}^{(\text{ols})}) - R(f^*)$	$\mathbb{E}R(\hat{f}) - R(f^*)$	$\mathbb{E}R[(\hat{f}^{(\text{ols})}) \vert \hat{f} \neq \hat{f}^{(\text{ols})}] - R(f^*)$	$\mathbb{E}[R(\hat{f}) \vert \hat{f} \neq \hat{f}^{(\text{ols})}] - R(f^*)$
INC(n=200,d=1)	1000	20	8	$0.541(\pm 0.048)$	$0.541(\pm 0.048)$	$0.401(\pm 0.168)$	$0.397(\pm 0.167)$
INC(n=200,d=2)	1000	1	0	$1.051(\pm 0.067)$	$1.051(\pm 0.067)$	2.566	2.757
HCC(n=200,d=2)	1000	1	0	$1.051(\pm 0.067)$	$1.051(\pm 0.067)$	2.566	2.757
TS(n=200,d=2)	1000	0	0	$1.068(\pm 0.067)$	$1.068(\pm 0.067)$	–	–
INC(n=1000,d=2)	1000	0	0	$0.203(\pm 0.013)$	$0.203(\pm 0.013)$	–	–
INC(n=1000,d=10)	1000	0	0	$1.023(\pm 0.029)$	$1.023(\pm 0.029)$	–	–
HCC(n=1000,d=10)	1000	0	0	$1.023(\pm 0.029)$	$1.023(\pm 0.029)$	–	–
TS(n=1000,d=10)	1000	0	0	$0.997(\pm 0.028)$	$0.997(\pm 0.028)$	–	–
INC(n=2000,d=2)	1000	0	0	$0.112(\pm 0.007)$	$0.112(\pm 0.007)$	–	–
HCC(n=2000,d=2)	1000	0	0	$0.112(\pm 0.007)$	$0.112(\pm 0.007)$	–	–
TS(n=2000,d=2)	1000	0	0	$0.098(\pm 0.006)$	$0.098(\pm 0.006)$	–	–
INC(n=2000,d=10)	1000	0	0	$0.517(\pm 0.015)$	$0.517(\pm 0.015)$	–	–
HCC(n=2000,d=10)	1000	0	0	$0.517(\pm 0.015)$	$0.517(\pm 0.015)$	–	–
TS(n=2000,d=10)	1000	0	0	$0.501(\pm 0.015)$	$0.501(\pm 0.015)$	–	–

Figure 1: Surrounding points are the points of the training set generated several times from $TS(1000, 10)$ (with the mixture noise with $p = 0.005$ and $\rho = 0.4$) that are not taken into account in the min-max truncated estimator (to the extent that the estimator would not change by removing simultaneously all these points). The min-max truncated estimator $x \mapsto \hat{f}(x)$ appears in dash-dot line, while $x \mapsto \mathbb{E}(Y|X = x)$ is in solid line. In these six simulations, it outperforms the ordinary least squares estimator.

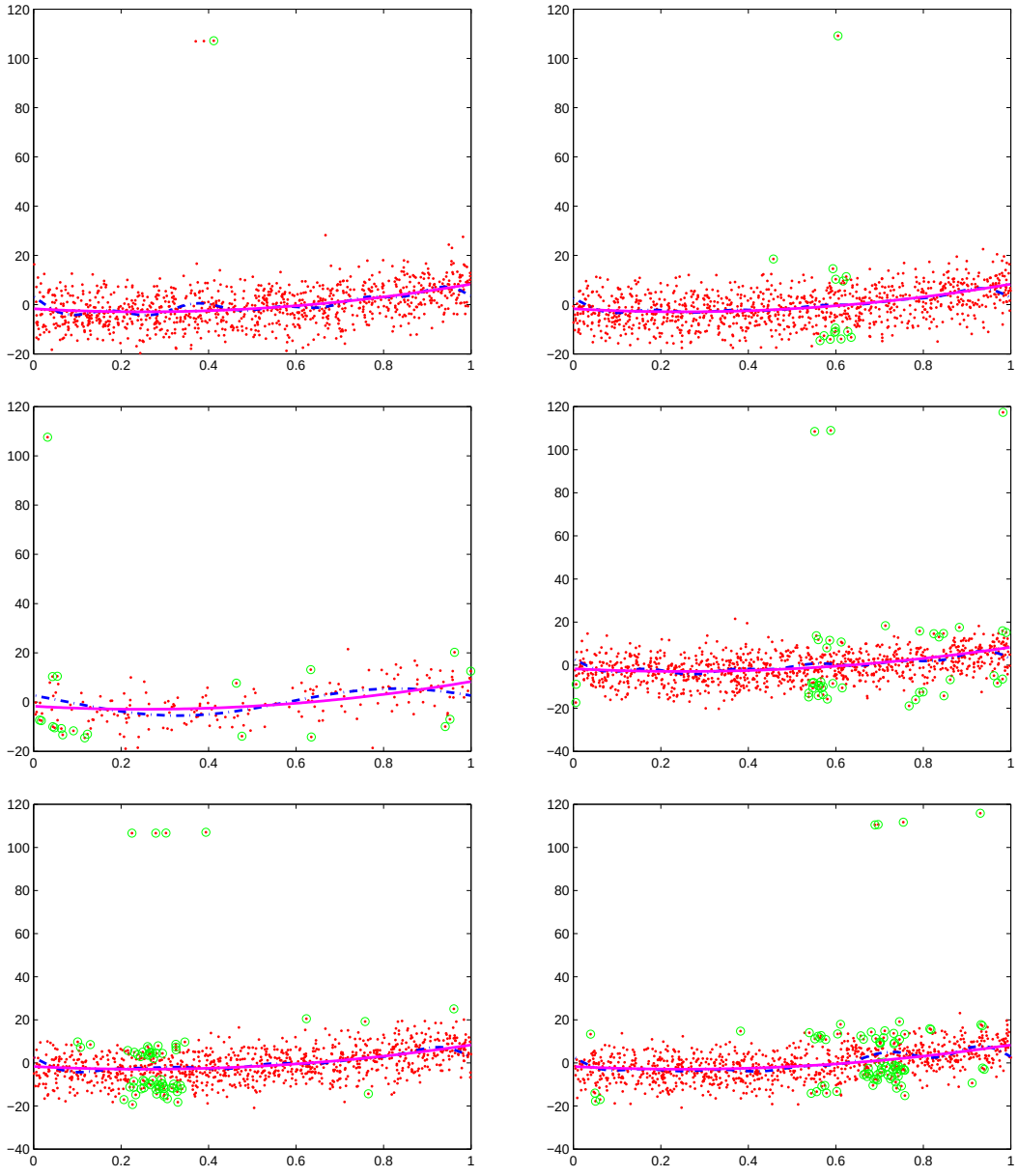
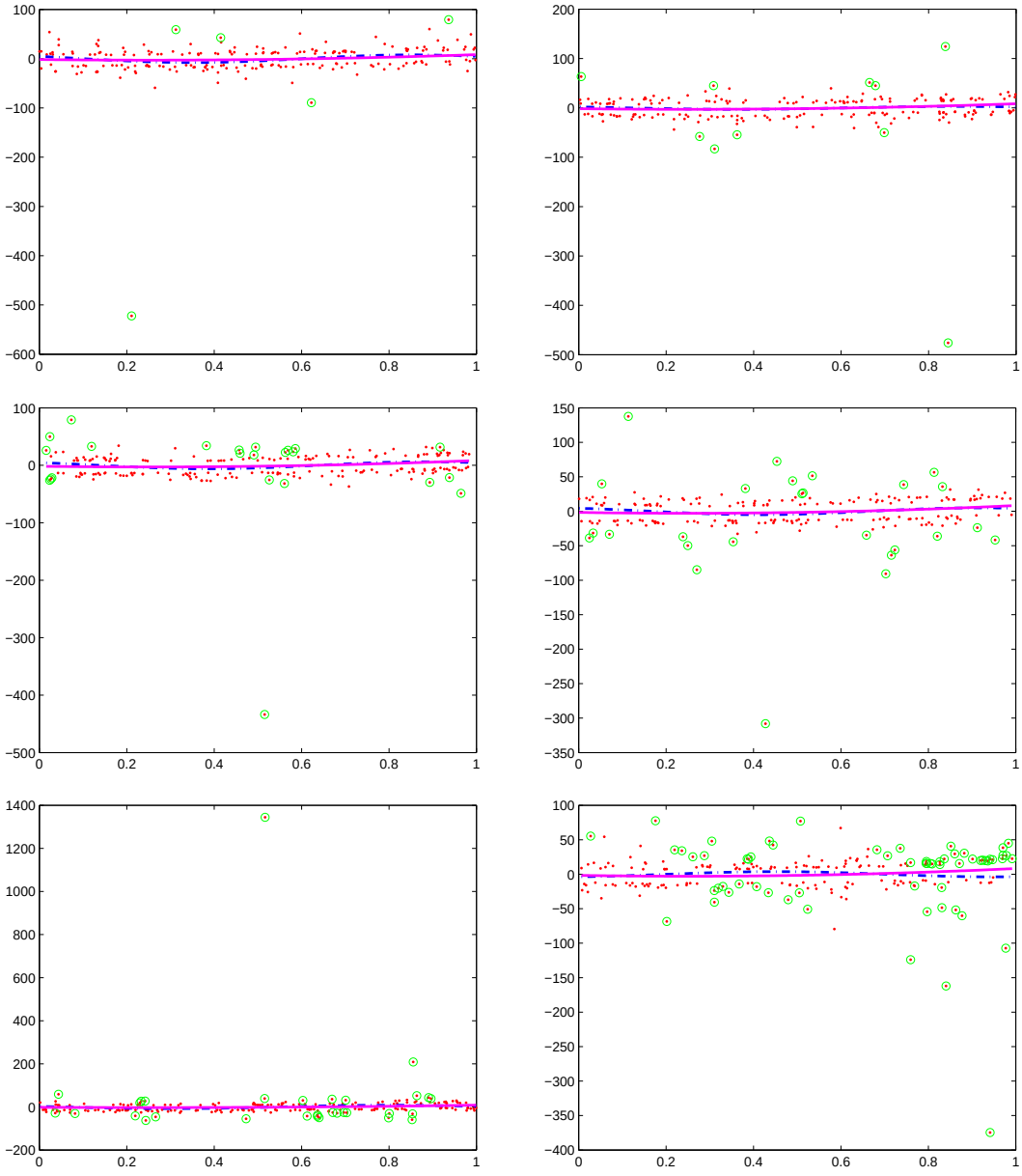


Figure 2: Surrounding points are the points of the training set generated several times from $TS(200, 2)$ (with the heavy-tailed noise) that are not taken into account in the min-max truncated estimator (to the extent that the estimator would not change by removing these points). The min-max truncated estimator $x \mapsto \hat{f}(x)$ appears in dash-dot line, while $x \mapsto \mathbb{E}(Y|X = x)$ is in solid line. In these six simulations, it outperforms the ordinary least squares estimator. Note that in the last figure, it does not consider 64 points among the 200 training points.



REFERENCES

[1] P. Alquier. PAC-bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods of Statistics*, 17(4):279–304, 2008.

[2] J.-Y. Audibert. A better variance control for PAC-Bayesian classification. Preprint n.905, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, 2004.

[3] J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. *Annals of Statistics*, 2009.

[4] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

[5] L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

[6] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, pages 331–368, 2007.

[7] O. Catoni. A PAC-Bayesian approach to adaptive classification. Technical report, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, 2003.

[8] O. Catoni. *Statistical Learning Theory and Stochastic Optimization, Lectures on Probability Theory and Statistics, École d'Été de Probabilités de Saint-Flour XXXI – 2001*, volume 1851 of *Lecture Notes in Mathematics*. Springer, 2004. Pages 1–269.

[9] O. Catoni. *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes Monograph Series*. Institute of Mathematical Statistics, 2007. Pages i-xii, 1-163.

[10] O. Catoni. High confidence estimates of the mean of heavy-tailed real random variables. *submitted to Ann. Inst. Henri Poincaré, Probab. Stat.*, 2009.

[11] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2004.

[12] A.E. Hoerl. Application of ridge analysis to regression problems. *Chem. Eng. Prog.*, 58:54–59, 1962.

[13] V. Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 34(6):2593–2656, 2006.

[14] J. Langford and J. Shawe-Taylor. PAC-bayes & margins. In *Advances in Neural Information Processing Systems*, pages 423–430, 2002.

[15] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quart. Appl. Math.*, pages 164–168, 1944.

[16] G. G. Lorentz. Metric entropy and approximation. *Bull. Amer. Math. Soc.*, 72(6):903–937, 1966.

[17] A. Nemirovski. *Lectures on probability theory and statistics. Part II: topics in Non-parametric statistics*. Springer-Verlag. Probability summer school, Saint Flour, 1998.

[18] J. Riley. Solving systems of linear equations with a positive definite, symmetric but possibly ill-conditioned matrix. *Math. Tables Aids Comput.*, 9:96–101, 1955.

[19] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B*, 58:267–288, 1994.

[20] A.B. Tsybakov. Optimal rates of aggregation. In B.Scholkopf and M.Warmuth, editors, *Computational Learning Theory and Kernel Machines, Lecture Notes in Artificial Intelligence*, volume 2777, pages 303–313. Springer, Heidelberg, 2003.

[21] Y. Yang. Aggregating regression procedures for a better performance. *Bernoulli*, 10:25–47, 2004.