Risk bounds in linear regression through PAC-Bayesian truncation

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

February 10, 2009

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, CERTIS, 6 avenue Blaise Pascal, 77455 Marne-la-Vallée, France, audibert@certis.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 [17], and [18] 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 [9] 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 $\varphi_1, \ldots, \varphi_d$:

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

The Euclidean norm will simply be written as $\|\cdot\|$, and $\langle \cdot, \cdot \rangle$ will be its associated dot 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 symbol κ will be used to denote constants, which means here deterministic quantities not depending on d and n but possibly depending on other constants of the problem. Its value may differ from line to line.

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 [14] and optimal rates of aggregation are given in [17] 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 $\mathbf{Y} = [Y_j]_{j=1}^d$ 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 [9, Theorem 11.1]) for any $X_1, \ldots, X_n$ in $\mathcal{X}$,

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

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

$$R(\hat{f}^{(\mathrm{ols})})-R(f^*_{\mathrm{lin}})$$
$$=\mathbb{E}\big[\hat{f}^{(\mathrm{ols})}(X)-f^{(\mathrm{reg})}(X)\big]^2-\mathbb{E}\big[f^*_{\mathrm{lin}}(X)-f^{(\mathrm{reg})}(X)\big]^2.\quad(1.3)$$

The analysis of the ordinary least squares often stops at this point in classical statistical textbooks. (Besides, to simplify, the strong assumption $f^{(\mathrm{reg})}=f^*_{\mathrm{lin}}$ is often made.) This can be misleading since Inequality (1.2) does not imply a d/n upper bound on the risk of $\hat{f}^{(\mathrm{ols})}$. Nevertheless the following result holds [9, Theorem 11.3].

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

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

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

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

for some numerical constant κ*.*

Using PAC-Bayesian inequalities, Catoni [7, Proposition 5.9.1] has proved a different type of results on the generalization ability of $\hat{f}^{(\mathrm{ols})}$.

THEOREM 1.2 *Let* $\mathcal{F}'\subset\mathcal{F}_{\mathrm{lin}}$ *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*

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

we have

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

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 [7, 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 [6, 8], 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*

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

Using the result of [7, 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 [11, 3] 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 [3] 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 [4], 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 $\mathcal{F}$, $\sup_{x \in \mathcal{X}} |f_1(x) - f_2(x)|_\infty \leq H$ and there exists a function $f_0 \in \mathcal{F}$ satisfying*

the exponential moment condition:

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

for some positive constants A and M. Let

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

where the infimum is taken with respect to all possible orthonormal basis of $\mathcal{F}$ for the dot product $\langle f_1, f_2\rangle = \mathbb{E} f_1(X) f_2(X)$. Then the empirical risk minimizer satisfies for any $\varepsilon > 0$, with probability at least $1 - \varepsilon$:

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

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 $\phi_1, \dots, \phi_d$ 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 [17] 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 ([12, 15]). This solution has been labeled ridge regression in statistics ([10]), and consists in replacing $\hat{f}^{(\text{ols})}$ by

$$\hat{f}^{(\text{ridge})} \in \operatorname*{argmin}_{\{f_\theta;\, \theta\in\mathbb{R}^d\}} \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - f_\theta(X_i)\big)^2 + \lambda\sum_{j=1}^{d}\theta_j^2,$$

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 ([5]), 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 [16] and is defined as

$$\hat{f}^{(\text{lasso})} \in \operatorname*{argmin}_{\{f_\theta;\, \theta\in\mathbb{R}^d\}} \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - f_\theta(X_i)\big)^2 + \lambda\sum_{j=1}^{d}|\theta_j|.$$

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 logarithmic 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

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

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 on $\mathcal{F}$ under weak conditions on the tail of the output distribution. 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 the ridge regression estimator on $\mathcal{F}$ is defined by

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

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 consider the optimal ridge function $\tilde{f}$ optimizing the expected ridge risk

$$\tilde{f} \in \arg\min_{f_\theta \in \mathcal{F}} \big\{ R(f_\theta) + \lambda \|\theta\|^2 \big\}.$$

The most general theorem which can be obtained from the route followed in this section is Theorem 6.5 (page 33) 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 *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 $1-\varepsilon$,

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

PROOF. See Section 6.1 (page 27). □

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 $d' = \mathrm{rank}(Q)$. Assume that*

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

and

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

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 $\hat{\theta}^{(\mathrm{erm})} \in \mathrm{span}\{\varphi(X_1), \ldots, \varphi(X_n)\}$[4]. For any values of ϵ and n such that $2/n \le \varepsilon \le 1$ and

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

with probability at least $1 - \varepsilon$,

$$R(\hat{f}^{(\mathrm{erm})}) - R(f^*) \le 1920\, B \sqrt{\mathbb{E}[Y - f^*(X)]^4} \left[\frac{3Bd' + \log(2\varepsilon^{-1})}{n} + \left(\frac{4Bd'}{n} \right)^2 \right]. \quad (2.3)$$

PROOF. See Section 6.1 (page 27). □

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., $\varepsilon \ge 2/n$). 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 [4, Inequality (3.2)], we have $B \le \tilde{B} \le Bd$, 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 $\epsilon \ge 2/n$.

These limitations will be addressed in the next sections by considering algorithms explicitly based on PAC-Bayesian truncation.

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

3. AN EASILY COMPUTABLE ALGORITHM USING PAC-BAYESIAN TRUNCATION

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. We give two results, the first covering the case of unbounded input vectors, the requirement on $\varphi(X)$ being only a finite fourth order moment. The computability of the proposed estimator is discussed at the end of the section.

We still consider the function $\tilde{f} = f_{\tilde{\theta}}$ optimizing the expected ridge risk

$$\tilde{f} \in \arg\min_{f_\theta \in \mathcal{F}} \left\{ R(f_\theta) + \lambda \|\theta\|^2 \right\}$$

for a fixed nonnegative real parameter λ.

THEOREM 3.1 *Assume, for some positive constants σ and γ_2, that*

$$\mathbb{E}\big[\big(Y - \tilde{f}(X)\big)^2 \,|\, X\big] \le \sigma^2,$$
$$\textit{and} \quad \sup_{\theta \in \mathbb{R}^d} \frac{\mathbb{E}\big[f_\theta^4(X)\big]}{\big(\mathbb{E}\big[f_\theta^2(X)\big] + \lambda\|\theta\|^2\big)^2} \le \gamma_2.$$

Let ρ_θ be the Gaussian distribution on $\mathbb{R}^d$ with mean θ and diagonal covariance matrix ξI where

$$\xi = \frac{24\big[4\sigma^2 + \gamma_2(q_{\max} + \lambda)\|\Theta\|^2\big]}{n(q_{\min} + \lambda)},$$

$\|\Theta\|$ being the Euclidean diameter $\sup_{\theta,\theta' \in \Theta} \|\theta - \theta'\|$ of the convex closed parameter set Θ. Let us put

$$\alpha = \frac{1}{12\big[4\sigma^2 + \gamma_2(q_{\max} + \lambda)\|\Theta\|^2\big]}$$

and

$$W_i(f, f') = \alpha\Big(\big[Y_i - f(X_i)\big]^2 - \big[Y_i - f'(X_i)\big]^2\Big),$$

and consider some estimator $\hat{\theta}$ satisfying

$$\hat{\theta} \in \arg\min_{\theta_1 \in \Theta} \sup_{\theta_2 \in \Theta} -\frac{1}{n}\sum_{i=1}^n \int \rho_{\theta_1}(d\theta') \log\Bigg\{ \int \rho_{\theta_2}(d\theta) \Big[1 - W_i(f_{\theta'}, f_\theta) + \frac{1}{2} W_i^2(f_{\theta'}, f_\theta)\Big] \Bigg\} + \alpha\lambda\big[\|\theta_1\|^2 - \|\theta_2\|^2\big].$$

For any $\epsilon > 0$, with probability at least $1 - \epsilon$,

$$R(f_{\hat\theta}) + \lambda\|\hat\theta\|^2 \le \min_{f_\theta \in \mathcal{F}} \left\{ R(f_\theta) + \lambda\|\theta\|^2 \right\} + \frac{32\sigma^2 q_{\max} d}{n(q_{\min} + \lambda)} + \left[4\sigma^2 + \gamma_2(q_{\max} + \lambda)\|\Theta\|^2\right] \left\{ \frac{576\gamma_2(q_{\max} + \lambda)^2 d^2}{(q_{\min} + \lambda)^2 n^2} + \frac{48\log(2/\epsilon)}{n} \right\}.$$

PROOF. See Section 6.2 (page 37). □

Theorem 3.1 provides a non asymptotic bound for the excess risk with a d/n speed of convergence and an exponential tail even when the output Y has no exponential moment. It is even possible to assume on the output Y nothing more than the sheer existence of the risk function, in the case when the input X is bounded, as stated in the following theorem. Here we assume for simplicity that $\lambda = 0$, so that $\tilde f = f^* = \text{argmin}_{f \in \mathcal{F}} R(f)$.

THEOREM 3.2 *Assume that $\mathcal{F}$ has a diameter H for the L^∞-norm:*

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

Consider again

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

Assume that we know $\sigma \ge 0$ such that $R(f^*) \le \sigma^2$. Let $\alpha = [12B(4\sigma^2 + BH^2)]^{-1}$ and

$$W_i(f, f') = \alpha[(Y_i - f(X_i))^2 - (Y_i - f'(X_i))^2].$$

Let ρ_θ be the Gaussian distribution on $\mathbb{R}^d$ with mean θ and diagonal covariance matrix ξI with

$$\xi = \frac{12B(4\sigma^2 + BH^2)}{n q_{\min}}.$$

Consider the estimator $\hat f = f_{\hat\theta}$ where

$$\hat\theta \in \operatorname*{argmin}_{\theta_1 \in \mathcal{F}} \max_{\theta_2 \in \mathcal{F}} -\frac{1}{n} \sum_{i=1}^n \int \rho_{\theta_1}(d\theta') \log \left\{ \int \rho_{\theta_2}(d\theta) \left[1 - W_i(f_{\theta'}, f_\theta) + W_i^2(f_{\theta'}, f_\theta)/2 \right] \right\}.$$

For any $\epsilon > 0$, with probability at least $1 - \varepsilon$,

$$R(\hat{f}) - R(f^*) \leq 32\sigma^2 \frac{H^2 d}{q_{\min} n} + 96B(4\sigma^2 + BH^2)\left\{3\left(\frac{H^2 d}{q_{\min} n}\right)^2 + \frac{\log(2\varepsilon^{-1})}{n}\right\}.$$

PROOF. See Section 6.2 (page 37). □

We obtained here stronger results than the non asymptotic bound of Section 2, at the price of replacing the empirical risk minimizer by a more involved estimator.

Section 6.2.3 (page 44) addresses the question of computing this estimator. It shows that an approximation can be made which involves optimizing explicit quantities given in closed form without the help of Gaussian integrals. Some upper bound of the precision of this approximation is itself computable in closed form from observations. It adds, as described in Theorem 6.9 (page 42), to the bound on the excess risk, but should not change its order of magnitude.

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

We recall the definition

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

In this section, we consider that the convex 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:*

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

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

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

Let $\hat{f}$ be a prediction function drawn from the distribution $\hat{\pi}$ defined in (4.2, page 17) *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 $Z_1, \ldots, Z_n$ 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 25), Lemma 5.3 (page 23) and Lemma 5.6 (page 27). □

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 52) 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 and 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 $R(\hat{f}_{\text{determ}}) \le \int R(f)\hat{\pi}(df)$, so that, after some pretty standard computations, one can prove that for any $\varepsilon > 0$, with probability at least $1 - \varepsilon$:

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

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

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 $\mathcal{F}$ such

[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 20) corresponds to (4.1, page 17) with this choice of function ℓ and for the choice $\pi^* = \pi$.

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

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

We will use the posterior distribution:

$$\frac{d\hat{\pi}}{d\pi}(f) = \frac{d\pi_{-\hat{\mathcal{E}}}}{d\pi}(f) = \frac{\exp[-\hat{\mathcal{E}}(f)]}{\int \exp[-\hat{\mathcal{E}}(f')]\pi(df')}. \tag{5.9}$$

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 20). *Let ρ be a (posterior) probability distribution on $\mathcal{F}$ admitting a density with respect to π depending on $Z_1, \dots, Z_n$. 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 $Z_1, \dots, Z_n$ and the randomized prediction function $\hat{f}$) at least $1 - \varepsilon$:*

$$\int \big[L^\flat(\hat{f}, f) + \gamma^*\bar{R}(f)\big]\pi^*_{-\gamma^*\bar{R}}(df) - \gamma\bar{R}\big(\hat{f}\,\big)$$
$$\leq \mathcal{I}^*(\gamma^*) - \mathcal{I}(\gamma) - \log\Big\{\int \exp\big[-\mathcal{E}^\sharp(f)\big]\pi(df)\Big\}$$
$$+ \log\Big[\frac{d\rho}{d\hat{\pi}}\big(\hat{f}\,\big)\Big] + 2\log(2\varepsilon^{-1}). \tag{5.10}$$

PROOF. See Section 6.3 (page 46). □

Some extra work will be needed to prove that Inequality (5.10) 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 20) 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 $\rho = \hat{\pi}$, (5.10) 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*

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

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 6.4 (page 49). □

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 $0 < b_1 \leq b_2$, $A > 0$ 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:*

$$\text{for any } y' \in \mathbb{R}, \qquad b_1 \leq \tilde{\ell}''_y(y') \leq b_2, \tag{5.11}$$

$$\text{and for any } x \in \mathcal{X}, \quad \mathbb{E}\Big\{\exp\Big[A^{-1}\big|\tilde{\ell}'_Y[f^*(X)]\big|\Big] \,\Big|\, X = x\Big\} \leq M. \tag{5.12}$$

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*

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

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

PROOF. See Section 6.5 (page 51). □

5.4. APPLICATION WITHOUT EXPONENTIAL MOMENT CONDITION. When we do not have finite exponential moments as assumed by Condition V1 (page 22), 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,

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

THEOREM 5.5 *Assume that Conditions* V2 *above and* C *(page 22) are satisfied. Let* $0 < \lambda < V^{-1}$ *and*

$$\ell\big[(X,Y),f,f'\big] = T\Big(\lambda\big\{\tilde{\ell}\big[Y,f(X)\big] - \tilde{\ell}\big[Y,f'(X)\big]\big\}\Big), \tag{5.13}$$

with

$$T(x) = -\log(1 - x + x^2/2). \tag{5.14}$$

Let $\hat{f} \in \mathcal{F}$ *be a function drawn according to the distribution* $\hat{\pi}$ *defined in* (5.9, page 21) *with* $\hat{\mathcal{E}}$ *defined in* (5.4, page 20) *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 (5.13) 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,

$$\frac{\mathbb{E}W(f,f')^2}{2} \le \mathbb{E}W(f,f^*)^2 + \mathbb{E}W(f',f^*)^2 \le \lambda^2 V\bar{R}(f) + \lambda^2 V\bar{R}(f'), \tag{5.15}$$

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

$$\begin{aligned} L^\flat(f,f') &\ge \lambda n\big[\bar{R}(f) - \bar{R}(f') - \lambda V\bar{R}(f) - \lambda V\bar{R}(f')\big] \\ &= \lambda n\big[(1-\eta)\bar{R}(f) - (1+\eta)\bar{R}(f')\big]. \end{aligned} \tag{5.16}$$

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 (5.15) 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

$$\mathcal{E}^\sharp(f) \le \lambda n(1+\eta)\bar R(f) - \mathcal{I}(\lambda n(1-\eta)). \tag{5.17}$$

Plugging (5.16) and (5.17) in (5.10) 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 (5.10), are tight.

Besides, if we take ℓ as in (5.13) 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 (5.14, page 25) 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

$$\sup_{x \in \mathcal{X}} \mathbb{E}\big\{[Y - f^*(X)]^2 \big| X = x\big\} \leq \sigma^2 < +\infty. \tag{5.18}$$

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

PROOF. See Section 6.6 (page 52) □

6. PROOFS

6.1. PROOFS OF THEOREMS 2.1 AND 2.2. The proofs rely on the use of PAC Bayesian inequalities. 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

$$r(\theta) = \frac{1}{n}\sum_{i=1}^n \big(\langle \theta, X_i\rangle - Y_i\big)^2, \tag{6.1}$$

$$\overline{r}(\theta) = \frac{1}{n}\sum_{i=1}^n \big(\langle \theta, \overline{X}_i\rangle - Y_i\big)^2, \tag{6.2}$$

$$\theta_0 = \arg\min_{\theta\in\overline{\Theta}} \overline{R}(\theta) + \lambda\|Q_\lambda^{-1/2}\theta\|^2, \tag{6.3}$$

$$\hat{\theta} \in \arg\min_{\theta\in\Theta} r(\theta) + \lambda\|\theta\|^2, \tag{6.4}$$

$$\theta_1 = Q_\lambda^{1/2}\hat{\theta} \in \arg\min_{\theta\in\overline{\Theta}} \overline{r}(\theta) + \lambda\|Q_\lambda^{-1/2}\theta\|^2. \tag{6.5}$$

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 $\theta_2 \in \mathbb{R}^d$,*

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

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.

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

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

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

We conclude from the convex inequality (see [7, page 159])

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

□

Let us compute some useful quantities

$$\mathcal{K}(\rho_{\theta_2},\rho_{\theta_0})=\frac{\beta}{2}\|\theta_2-\theta_0\|^2, \tag{6.6}$$

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

$$\int \rho_{\theta^2}(d\theta)\langle\theta-\theta_2,\overline{X}\rangle^4=\frac{3\|\overline{X}\|^4}{\beta^2}, \tag{6.7}$$

$$\begin{aligned}
\int \rho_{\theta_2}(d\theta)\big[W(\theta)^2\big] &= \alpha^2\int \rho_{\theta_2}(d\theta)\langle\theta-\theta_0,\overline{X}\rangle^2\big(\langle\theta+\theta_0,\overline{X}\rangle-2Y\big)^2\\
&= \alpha^2\int \rho_{\theta_2}(d\theta)\Big[\langle\theta-\theta_2+\theta_2-\theta_0,\overline{X}\rangle\big(\langle\theta-\theta_2+\theta_2+\theta_0,\overline{X}\rangle-2Y\big)\Big]^2\\
&= \int \rho_{\theta_2}(d\theta)\Big[\alpha\langle\theta-\theta_2,\overline{X}\rangle^2+2\alpha\langle\theta-\theta_2,\overline{X}\rangle\big(\langle\theta_2,\overline{X}\rangle-Y\big)+W(\theta_2)\Big]^2\\
&= \int \rho_{\theta_2}(d\theta)\Big[\alpha^2\langle\theta-\theta_2,\overline{X}\rangle^4+4\alpha^2\langle\theta-\theta_2,\overline{X}\rangle^2\big(\langle\theta_2,\overline{X}\rangle-Y\big)^2+W(\theta_2)^2\\
&\qquad +2\alpha\langle\theta-\theta_2,\overline{X}\rangle^2W(\theta_2)\Big]\\
&= \frac{3\alpha^2\|\overline{X}\|^4}{\beta^2}+\frac{2\alpha\|\overline{X}\|^2}{\beta}\Big[2\alpha\big(\langle\theta_2,\overline{X}\rangle-Y\big)^2+W(\theta_2)\Big]+W(\theta_2)^2.
\end{aligned} \tag{6.8}$$

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, $6ab\leq 9a^2+b^2$, we get

LEMMA 6.2

$$\int \rho_{\theta_2}(d\theta)\big[W(\theta)\big] = W(\theta_2) + \alpha\frac{\|\overline{X}\|^2}{\beta}, \tag{6.9}$$

$$\begin{aligned}\int \rho_{\theta_2}(d\theta)\big[W(\theta)^2\big] &= W(\theta_2)^2 + \frac{2\alpha\|\overline{X}\|^2}{\beta}\Big[2\alpha\big(\langle\theta_0,\overline{X}\rangle - Y\big)^2 + 3W(\theta_2)\Big] \\ &\quad + \frac{3\alpha^2\|\overline{X}\|^4}{\beta^2} \qquad (6.10)\\ &\le 10W(\theta_2)^2 + \frac{4\alpha^2\|\overline{X}\|^2}{\beta}\big(\langle\theta_0,\overline{X}\rangle - Y\big)^2 + \frac{4\alpha^2\|\overline{X}\|^4}{\beta^2}, \qquad (6.11)\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}\big[\|\overline{X}\|^2\big] &= \mathbb{E}\big[\mathrm{Tr}\big(\overline{X}\,\overline{X}^T\big)\big] &&= \mathbb{E}\big[\mathrm{Tr}\big(Q_\lambda^{-1/2}XX^TQ_\lambda^{-1/2}\big)\big]\\ &= \mathbb{E}\big[\mathrm{Tr}\big(Q_\lambda^{-1}XX^T\big)\big] &&= \mathrm{Tr}\big[Q_\lambda^{-1}\mathbb{E}(XX^T)\big]\\ &= \mathrm{Tr}\big(Q_\lambda^{-1}(Q_\lambda - \lambda I)\big) &&= d - \lambda\,\mathrm{Tr}(Q_\lambda^{-1}) \quad = \; D\,. \end{aligned} \tag{6.12}$$

We can weaken Lemma 6.1 (page 28) noticing that for any real number x, $x \le -\log(1-x)$ and

$$\begin{aligned}-\log\left(1 - x + \frac{x^2}{2}\right) &= \log\left(\frac{1+x+x^2/2}{1+x^4/4}\right)\\ &\le \log\left(1 + x + \frac{x^2}{2}\right) \le x + \frac{x^2}{2}.\end{aligned}$$

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

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

$$+\frac{\beta}{2}\|\theta_2-\theta_0\|^2+\eta.$$

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

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

THEOREM 6.3 *Let us put*

$$\begin{aligned}\widehat{D}&=\frac{1}{n}\sum_{i=1}^n\|\overline{X}_i\|^2\qquad\textit{(let us remind that}D=\mathbb{E}\big[\|\overline{X}\|^2\big]\textit{ from (6.12))},\\B_1&=2\mathbb{E}\Big[\|\overline{X}\|^2\big(\langle\theta_0,\overline{X}\rangle-Y\big)^2\Big],\\\widehat{B}_1&=\frac{2}{n}\sum_{i=1}^n\Big[\|\overline{X}_i\|^2\big(\langle\theta_0,\overline{X}_i\rangle-Y_i\big)^2\Big],\\B_2&=2\mathbb{E}\Big[\|\overline{X}\|^4\Big],\\\widehat{B}_2&=\frac{2}{n}\sum_{i=1}^n\|\overline{X}_i\|^4,\\B_3&=40\sup\Big\{\mathbb{E}\big[\langle u,\overline{X}\rangle^2\big(\langle\theta_0,\overline{X}\rangle-Y\big)^2\big]:u\in\mathbb{R}^d,\|u\|=1\Big\},\\\widehat{B}_3&=\sup\Big\{\frac{40}{n}\sum_{i=1}^n\langle u,\overline{X}_i\rangle^2\big(\langle\theta_0,\overline{X}_i\rangle-Y_i\big)^2:u\in\mathbb{R}^d,\|u\|=1\Big\},\\B_4&=10\sup\Big\{\mathbb{E}\Big[\langle u,\overline{X}\rangle^4\Big]:u\in\mathbb{R}^d,\|u\|=1\Big\},\\\widehat{B}_4&=\sup\Big\{\frac{10}{n}\sum_{i=1}^n\langle u,\overline{X}_i\rangle^4:u\in\mathbb{R}^d,\|u\|=1\Big\}.\end{aligned}$$

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

$$\begin{aligned}n\mathbb{E}\big[W(\theta_2)\big]-\Big[n\alpha^2(B_3+\widehat{B}_3)+\frac{\beta}{2}\Big]\|\theta_2-\theta_0\|^2\\-n\alpha^2(B_4+\widehat{B}_4)\|\theta_2-\theta_0\|^4\end{aligned}$$

$$\leq \sum_{i=1}^{n} W_i(\theta_2) + \frac{n\alpha}{\beta}(\widehat{D} - D) + \frac{n\alpha^2}{\beta}(B_1 + \widehat{B}_1) + \frac{n\alpha^2}{\beta^2}(B_2 + \widehat{B}_2) + \eta.$$

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 $\theta_* = \arg\min_{\theta\in\mathbb{R}^d} \overline{R} + \lambda\|Q_\lambda^{-1/2}\theta\|^2$. 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 $\theta_2 \in \overline{\Theta}$

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

This and the inequality

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

proves

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

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

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

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

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 31). 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 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

$$R(\theta_3) + \lambda\|\theta_3\|^2 < R(\theta_0) + \lambda\|\theta_0\|^2 + \frac{x_1 + x_2}{2}, \tag{6.13}$$

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 (6.13) to any point of $[\theta_3, \theta_4] \cap \widehat{\Theta}$, which proves that it 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

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

because the values between x_1 and x_2 are excluded by Theorem 6.3. □

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 *With probability at least*

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

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

$$\Big\{1 - \big[\alpha(2B_3 + \eta_3) + \tfrac{\beta}{2n\alpha}\big]\Big\}x - \alpha(2B_4 + \eta_4)x^2 = \frac{\eta_0}{\beta} + \frac{\alpha}{\beta}(2B_1 + \eta_1) + \frac{\alpha}{\beta^2}(2B_2 + \eta_2) + \frac{\eta_5}{n\alpha}, \tag{6.14}$$

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

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

6.1.1. 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 B_2. 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.

6.1.2. 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

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

Since we have

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

we get

$$\begin{aligned}
\widehat{d} &= \frac{1}{n}\sum_{i=1}^n \|\overline{X}_i\|^2 \leq dB,\\
B_1 &= 2\mathbb{E}\Big[\|\overline{X}\|^2\big(\langle\theta_0,\overline{X}\rangle - Y\big)^2\Big] \leq 2dB\, R(f^*),\\
\widehat{B}_1 &= \frac{2}{n}\sum_{i=1}^n \Big[\|\overline{X}_i\|^2\big(\langle\theta_0,\overline{X}_i\rangle - Y_i\big)^2\Big] \leq 2dB\, r(f^*),\\
B_2 &= 2\mathbb{E}\Big[\|\overline{X}\|^4\Big] \leq 2d^2B^2,\\
\widehat{B}_2 &= \frac{2}{n}\sum_{i=1}^n \|\overline{X}_i\|^4 \leq 2d^2B^2,\\
B_3 &= 40\sup\Big\{\mathbb{E}\big[\langle u,\overline{X}\rangle^2\big(\langle\theta_0,\overline{X}\rangle - Y\big)^2\big] : u\in\mathbb{R}^d, \|u\|=1\Big\} \leq 40B\, R(f^*),\\
\widehat{B}_3 &= \sup\Big\{\frac{40}{n}\sum_{i=1}^n \langle u,\overline{X}_i\rangle^2\big(\langle\theta_0,\overline{X}_i\rangle - Y_i\big)^2 : u\in\mathbb{R}^d, \|u\|=1\Big\} \leq 40B\, r(f^*),\\
B_4 &= 10\sup\Big\{\mathbb{E}\Big[\langle u,\overline{X}\rangle^4\Big] : u\in\mathbb{R}^d, \|u\|=1\Big\} \leq 10B^2,\\
\widehat{B}_4 &= \sup\Big\{\frac{10}{n}\sum_{i=1}^n \langle u,\overline{X}_i\rangle^4 : u\in\mathbb{R}^d, \|u\|=1\Big\} \leq 10B^2.
\end{aligned}$$

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 $R(\hat{f}^{(\text{erm})}) - R(f^*)$ is upper bounded by the smallest positive root of $a_1 x - a_2 x^2 = a_0$ as soon as $a_1^2 > 4a_0a_2$. In particular, setting $\varepsilon = \exp(-\eta)$ when (6.15) holds, we have

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

We conclude that

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

$$80\left(\frac{(2 + 4\alpha[R(f^*) + r(f^*)])Bd + \log(\varepsilon^{-1})}{n} + \left(\frac{4Bd}{n}\right)^2\right) < \left(\frac{3}{4B} - 40\alpha[R(f^*) + r(f^*)]\right)^2 \quad (6.15)$$

holds, then we have

$$R(\hat{f}^{(\text{erm})}) - R(f^*) \le \mathcal{J}\left(\frac{(2 + 4\alpha[R(f^*) + r(f^*)])Bd + \log(\varepsilon^{-1})}{n} + \left(\frac{4Bd}{n}\right)^2\right), \quad (6.16)$$

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

Now, the Bienaymé-Chebyshev inequality implies

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

Under the finite moment assumption of Theorem 2.2, we obtain that for any $\varepsilon \ge 1/n$, 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

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

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

$$R(\hat{f}^{(\text{erm})}) - R(f^*) \leq \mathcal{J}_1 B\left(\frac{3Bd' + \log(\varepsilon^{-1})}{n} + \left(\frac{4Bd'}{n}\right)^2\right), \tag{6.17}$$

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 $\mathbf{Y} = [Y_j]_{j=1}^d$,

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

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.

6.2. PROOF OF THEOREMS 3.1 AND 3.2. As in Section 6.1, to shorten the formulae and without loss of generality, we consider that the input space is $\mathbb{R}^d$ and that the functions $\varphi_1, \ldots, \varphi_d$ are the component functions, so that $x = [\varphi_j(x)]_{j=1}^d$. Therefore, the function f_θ maps an input x to $\langle\theta, x\rangle$. With a slight abuse of notation, $R(\theta) = \mathbb{E}\big[(\langle\theta, X\rangle - Y)^2\big]$ will denote the risk of this prediction function. Without loss of generality, we may also assume that the null vector belongs to Θ (otherwise it suffices to replace Y with $Y - \langle\theta_0, X\rangle$ for some arbitrar $\theta_0 \in \Theta$).

Let us come back to the general setting of ridge regression as in the previous section. The vector of coefficients of $\tilde{f}$, the minimizer of the expected ridge risk, is $\tilde{\theta} \in \text{argmin}_{\theta\in\Theta} R(\theta) + \lambda\|\theta\|^2$. Let $\alpha > 0, \beta > 0$ and

$$W(\theta, \theta') = \alpha\Big[\big(\langle\theta, X\rangle - Y\big)^2 - \big(\langle\theta', X\rangle - Y\big)^2\Big]$$

$$= \alpha\langle\theta - \theta', X\rangle\big(\langle\theta + \theta', X\rangle - 2Y\big),$$
$$W_i(\theta, \theta') = \alpha\Big[\big(\langle\theta, X_i\rangle - Y_i\big)^2 - \big(\langle\theta', X_i\rangle - Y_i\big)^2\Big].$$

As in Section 6.1, we consider the change of coordinates

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

with $Q_\lambda = Q + \lambda I$. Let us put for short $R_\lambda(\theta) = R(\theta) + \lambda\|\theta\|^2$. We will use some constants from the previous section, namely

$$\begin{aligned}
B_3 &= 40\sup\Big\{\mathbb{E}\big[\langle u, \overline{X}\rangle^2\big(\langle\tilde{\theta}, X\rangle - Y\big)^2\big] : u \in \mathbb{R}^d, \|u\| = 1\Big\},\\
&= 40\sup\Big\{\mathbb{E}\big[\langle v, X\rangle^2\big(\langle\tilde{\theta}, X\rangle - Y\big)^2\big] : v \in \mathbb{R}^d, \mathbb{E}\big[\langle v, X\rangle^2\big] + \lambda\|v\|^2 = 1\Big\},\\
B_4 &= 10\sup\Big\{\mathbb{E}\Big[\langle u, \overline{X}\rangle^4\Big] : u \in \mathbb{R}^d, \|u\| = 1\Big\}\\
&= 10\sup\Big\{\mathbb{E}\big[\langle v, X\rangle^4\big] : v \in \mathbb{R}^d, \mathbb{E}\big[\langle v, X\rangle^2\big] + \lambda\|v\|^2 = 1\Big\}.
\end{aligned}$$

For any $\theta \in \Theta$, we have

$$\begin{aligned}
5\mathbb{E}\big[W^2(\theta, \tilde{\theta})\big] &\le \alpha^2\Big(10\mathbb{E}\langle\theta - \tilde{\theta}, X\rangle^4 + 40\alpha^2\mathbb{E}\langle\theta - \tilde{\theta}, X\rangle^2(\langle\tilde{\theta}, X\rangle - Y)^2\Big)\\
&\le \alpha^2\Big(B_4\|Q_\lambda^{1/2}(\theta - \tilde{\theta})\|^4 + B_3\|Q_\lambda^{1/2}(\theta - \tilde{\theta})\|^2\Big)
\end{aligned}$$

and

$$\|\theta - \tilde{\theta}\|^2 \le \frac{1}{q_{\min} + \lambda}\|Q_\lambda^{1/2}(\theta - \tilde{\theta})\|^2 = \frac{R_\lambda(\theta) - R_\lambda(\tilde{\theta})}{q_{\min} + \lambda}.$$

The computations done to obtain Lemma 6.2 are still valid if $\overline{X}$ and θ_0 are respectively replaced by X and $\tilde{\theta}$. So we have

$$\begin{aligned}
\int\rho_{\theta_2}(d\theta)\big[W_i(\theta, \tilde{\theta})\big] &= W_i(\theta_2, \tilde{\theta}) + \alpha\frac{\|X_i\|^2}{\beta},\\
\int\rho_{\theta_2}(d\theta)\big[W_i(\theta, \tilde{\theta})^2\big] &= W_i(\theta_2, \tilde{\theta})^2 + \frac{2\alpha\|X_i\|^2}{\beta}\Big[2\alpha\big(\langle\theta_0, X_i\rangle - Y\big)^2 + 3W_i(\theta_2, \tilde{\theta})\Big]\\
&\quad + \frac{3\alpha^2\|X_i\|^4}{\beta^2}.
\end{aligned}$$

By Jensen's inequality, we get

$$\sum_{i=1}^n -\log\Big\{1 + W_i(\theta_2, \tilde{\theta}) + W_i^2(\theta_2, \tilde{\theta})/2 + \frac{\alpha}{\beta}\|X_i\|^2$$

$$
\begin{aligned}
&+ \frac{\alpha}{\beta}\|X_i\|^2\big[2\alpha\big(\langle\tilde\theta, X_i\rangle - Y_i\big)^2 + 3W_i(\theta_2,\tilde\theta)\big] + \frac{3\alpha^2}{2\beta^2}\|X_i\|^4\Big\}\\
&= \sum_{i=1}^n -\log\Big\{\int\rho_{\theta_2}(d\theta)\Big[1 + W_i(\theta,\tilde\theta) + W_i^2(\theta,\tilde\theta)/2\Big]\Big\}\\
&\le \sum_{i=1}^n -\int\rho_{\theta_2}(d\theta)\log\Big[1 + W_i(\theta,\tilde\theta) + W_i^2(\theta,\tilde\theta)/2\Big]\\
&= \sum_{i=1}^n \int\rho_{\theta_2}(d\theta)\log\left(\frac{1 - W_i(\theta,\tilde\theta) + W_i^2(\theta,\tilde\theta)/2}{1 + W_i^4(\theta,\tilde\theta)/4}\right)\\
&\le \sum_{i=1}^n \int\rho_{\theta_2}(d\theta)\log\Big[1 - W_i(\theta,\tilde\theta) + W_i^2(\theta,\tilde\theta)/2\Big].
\end{aligned}
$$

Now, by using Lemma 6.1 and Inequality (6.6) (up to appropriate minor changes), with probability at least $1-\varepsilon$, for any $\theta_2 \in \Theta$,

$$
\begin{aligned}
&\sum_{i=1}^n \int\rho_{\theta_2}(d\theta)\log\Big[1 - W_i(\theta,\tilde\theta) + W_i^2(\theta,\tilde\theta)/2\Big]\\
&\le n\int\rho_{\theta_2}(d\theta)\log\Big[1 - \mathbb{E}\big[W(\theta,\tilde\theta)\big] + \mathbb{E}\big[W^2(\theta,\tilde\theta)\big]/2\Big] + \frac{\beta}{2}\|\theta_2 - \tilde\theta\|^2 - \log(\varepsilon).
\end{aligned}
$$

Therefore, with probability at least $1-\varepsilon$, the following holds for any $\theta_2 \in \Theta$,

$$
\begin{aligned}
&\sum_{i=1}^n -\log\Big\{1 + W_i(\theta_2,\tilde\theta) + W_i^2(\theta_2,\tilde\theta)/2\\
&\qquad + \frac{\alpha}{\beta}\|X_i\|^2 + \frac{\alpha}{\beta}\|X_i\|^2\big[2\alpha\big(\langle\tilde\theta, X_i\rangle - Y_i\big)^2 + 3W_i(\theta_2,\tilde\theta)\big]\\
&\qquad\qquad + \frac{3\alpha^2}{2\beta^2}\|X_i\|^4\Big\} + n\alpha\lambda\big(\|\tilde\theta\|^2 - \|\theta_2\|^2\big)\\
&\le n\int\rho_{\theta_2}(d\theta)\log\Big[1 - \mathbb{E}\big[W(\theta,\tilde\theta)\big] + \mathbb{E}\big[W^2(\theta,\tilde\theta)\big]/2\Big]\\
&\qquad + n\alpha\lambda\big(\|\tilde\theta\|^2 - \|\theta_2\|^2\big) + \frac{\beta}{2}\|\theta_2 - \tilde\theta\|^2 - \log(\varepsilon)\\
&\le -n\log\int\rho_{\theta_2}(d\theta)\mathbb{E}\big[W(\theta,\tilde\theta)\big] + \frac{n}{2}\int\rho_{\theta_2}(d\theta)\mathbb{E}\big[W^2(\theta,\tilde\theta)\big]\\
&\qquad + n\alpha\lambda\big(\|\tilde\theta\|^2 - \|\theta_2\|^2\big) + \frac{\beta}{2}\|\theta_2 - \tilde\theta\|^2 - \log(\epsilon)\\
&= -n\mathbb{E}\big[W(\theta_2,\tilde\theta)\big] + 5n\mathbb{E}\big[W^2(\theta_2,\tilde\theta)\big] + n\alpha\lambda\big(\|\tilde\theta\|^2 - \|\theta_2\|^2\big)\\
&\quad + \frac{n\alpha}{\beta}\mathbb{E}\big[\|X\|^2\big] + \frac{2n\alpha^2}{\beta}\mathbb{E}\Big[\|X\|^2\big(\langle\tilde\theta, X\rangle - Y\big)^2 + \|X\|^4\Big]
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{\beta}{2}\|\theta_2-\tilde\theta\|^2-\log(\varepsilon)\\
&\le n(B_3\alpha^2-\alpha)\|Q_\lambda^{1/2}(\theta_2-\tilde\theta)\|^2+nB_4\alpha^2\|Q_\lambda^{1/2}(\theta_2-\tilde\theta)\|^4\\
&+\frac{n\alpha}{\beta}\mathbb{E}\big[\|X\|^2\big]+\frac{2n\alpha^2}{\beta}\mathbb{E}\Big[\|X\|^2\big(\langle\tilde\theta,X\rangle-Y\big)^2+\|X\|^4\Big]\\
&+\frac{\beta}{2}\|\theta_2-\tilde\theta\|^2-\log(\varepsilon)\\
&\le\Big[\frac{\beta}{2(q_{\min}+\lambda)}+nB_3\alpha^2-n\alpha\Big]\|Q_\lambda^{1/2}(\theta_2-\tilde\theta)\|^2\\
&+nB_4\alpha^2\|Q_\lambda^{1/2}(\theta_2-\tilde\theta)\|^4+\frac{n\alpha}{\beta}\mathbb{E}\big[\|X\|^2\big]\\
&+\frac{2n\alpha^2}{\beta}\mathbb{E}\Big[\|X\|^2\big(\langle\tilde\theta,X\rangle-Y\big)^2+\|X\|^4\Big]-\log(\varepsilon).
\end{aligned}
$$

Assuming all the necessary constants are known, we get a confidence region for $\tilde\theta$:

LEMMA 6.7 *With probability at least* $1-\varepsilon$*, for any* $\theta_2\in\Theta$*, we have*

$$
\begin{aligned}
&\sum_{i=1}^n-\log\Big\{1+W_i(\theta_2,\tilde\theta)+W_i^2(\theta_2,\tilde\theta)/2\\
&+\frac{\alpha}{\beta}\|X_i\|^2+\frac{\alpha}{\beta}\|X_i\|^2\big[2\alpha\big(\langle\tilde\theta,X_i\rangle-Y_i\big)^2+3W_i(\theta_2,\tilde\theta)\big]\\
&+\frac{3\alpha^2}{2\beta^2}\|X_i\|^4\Big\}+n\alpha\lambda\big(\|\tilde\theta\|^2-\|\theta_2\|^2\big)\\
&\le\Big[\frac{\beta}{2(q_{\min}+\lambda)}+nB_3\alpha^2-n\alpha\Big]\|Q_\lambda^{1/2}(\theta_2-\tilde\theta)\|^2\\
&+nB_4\alpha^2\|Q_\lambda^{1/2}(\theta_2-\tilde\theta)\|^4+\frac{n\alpha}{\beta}\mathbb{E}\big[\|X\|^2\big]\\
&+\frac{2n\alpha^2}{\beta}\mathbb{E}\Big[\|X\|^2\big(\langle\tilde\theta,X\rangle-Y\big)^2+\|X\|^4\Big]-\log(\varepsilon).
\end{aligned}
$$

In order to get an estimator with some known generalization bound, let us minimize the lefthand-side of the lemma in $\tilde\theta$. For this we need to substitute $\tilde\theta$ with some randomly chosen θ according to ρ_{θ_1}. Unfortunately, we cannot make things completely explicit.

Let us remark first that

LEMMA 6.8 *For any* $\theta_1\in\Theta$*, we have*

$$
\int\rho_{\theta_1}(d\theta')\int\rho_{\theta_2}(d\theta)\big[W(\theta,\theta')\big]=W(\theta_2,\theta_1),
$$

$$\int \rho_{\theta_1}(d\theta') \int \rho_{\theta_2}(d\theta) \big[W^2(\theta,\theta')\big] = W^2(\theta_2,\theta_1) + \frac{2\alpha^2\|X\|^2}{\beta}\Big[2\big(\langle\theta_1,X\rangle - Y\big)^2 + 2\big(\langle\theta_2,X\rangle - Y\big)^2\Big] + \frac{4\alpha^2\|X\|^4}{\beta^2}.$$

PROOF.

$$\begin{aligned}
\int \rho_{\theta_1}(d\theta') \int \rho_{\theta_2}(d\theta) \big[W^2(\theta,\theta')\big] &= \int \rho_{\theta_1}(d\theta')\Big\{W^2(\theta_2,\theta') \\
&\quad + \alpha\frac{2\|X\|^2}{\beta}\Big[2\alpha\big(\langle\theta_2,X\rangle - Y\big)^2 + W(\theta_2,\theta')\Big]\Big\} + \frac{3\alpha^2\|X\|^4}{\beta^2} \\
&= W^2(\theta_2,\theta_1) + \frac{2\alpha\|X\|^2}{\beta}\Big[2\alpha\big(\langle\theta_2,X\rangle - Y\big)^2 + W(\theta_2,\theta_1) \\
&\quad - \alpha\frac{\|X\|^2}{\beta} + 2\alpha\big(\langle\theta_1,X\rangle - Y\big)^2 + W(\theta_1,\theta_2)\Big] + \frac{6\alpha^2\|X\|^4}{\beta^2} \\
&= W^2(\theta_2,\theta_1) + \frac{2\alpha^2\|X\|^2}{\beta}\Big[2\big(\langle\theta_1,X\rangle - Y\big)^2 \\
&\quad + 2\big(\langle\theta_2,X\rangle - Y\big)^2\Big] + \frac{4\alpha^2\|X\|^4}{\beta^2}.
\end{aligned}$$

□

We see that with probability at least $1-\varepsilon$, for any $\theta_2 \in \Theta$,

$$\begin{aligned}
&-\sum_{i=1}^n \int \rho_{\tilde\theta}(d\theta') \log\Big\{\int \rho_{\theta_2}(d\theta)\big[1 + W_i(\theta,\theta') + W_i^2(\theta,\theta')/2\big]\Big\} + n\alpha\lambda\big(\|\tilde\theta\|^2 - \|\theta_2\|^2\big) \\
&\quad \le n\int \rho_{\tilde\theta}(d\theta')\int \rho_{\theta_2}(d\theta)\log\Big[1 - \mathbb{E}\big[W(\theta,\theta')\big] + \mathbb{E}\big[W^2(\theta,\theta')/2\big]\Big] \\
&\quad\quad + n\alpha\lambda\big(\|\tilde\theta\|^2 - \|\theta_2\|^2\big) + \frac{\beta}{2}\|\theta_2 - \tilde\theta\|^2 - \log(\varepsilon) \\
&\quad \le n\mathbb{E}\big[W(\tilde\theta,\theta_2)\big] + n\alpha\lambda\big(\|\tilde\theta\|^2 - \theta_2\|^2\big) + \frac{n}{2}\mathbb{E}\big[W^2(\theta_2,\tilde\theta)\big] \\
&\quad + \mathbb{E}\left\{\frac{2n\alpha\|X\|^2}{\beta}\Big[2\alpha\big(\langle\tilde\theta,X\rangle - Y\big)^2 + W(\theta_2,\tilde\theta)\Big]\right\} + \frac{2n\alpha^2}{\beta^2}\mathbb{E}\big[\|X\|^4\big] \\
&\quad\quad + \frac{\beta}{2}\|\theta_2 - \tilde\theta\|^2 - \log(\varepsilon) \\
&\quad \le n\alpha\big[R_\lambda(\tilde\theta) - R_\lambda(\theta_2)\big] + \frac{3n}{2}\mathbb{E}\big[W^2(\theta_2,\tilde\theta)\big] \\
&\quad + \frac{4n\alpha^2}{\beta}\mathbb{E}\Big[\|X\|^2\big(\langle\tilde\theta,X\rangle - Y\big)^2\Big] + \frac{3n\alpha^2}{\beta^2}\mathbb{E}\big[\|X\|^4\big] \\
&\quad\quad + \frac{\beta}{2}\|\theta_2 - \tilde\theta\|^2 - \log(\varepsilon)
\end{aligned}$$

$$\leq n\Big\{\frac{\beta}{2n(q_{\min}+\lambda)}+\frac{3}{10}\alpha^2\Big[B_3+B_4\|Q_\lambda^{1/2}(\theta_2-\tilde\theta)\|^2\Big]-\alpha\Big\}\|Q_\lambda^{1/2}(\theta_2-\tilde\theta)\|^2$$
$$+\frac{4n\alpha^2}{\beta}\mathbb{E}\Big[\|X\|^2\big(\langle\tilde\theta,X\rangle-Y\big)^2\Big]+\frac{3n\alpha^2}{\beta^2}\mathbb{E}\big[\|X\|^4\big]-\log(\varepsilon).$$

On the other hand, with probability at least $1-\varepsilon$, for any $\theta_1\in\Theta$,

$$-\sum_{i=1}^n\int\rho_{\theta_1}(d\theta')\log\Big\{\int\rho_{\tilde\theta}(d\theta)\Big[1+W_i(\theta,\theta')+W_i^2(\theta,\theta')/2\Big]\Big\}$$
$$+n\alpha\lambda\big[\|\theta_1\|^2-\|\tilde\theta\|^2\big]$$
$$\geq -n\int\rho_{\theta_1}(d\theta')\log\Big\{\int\rho_{\tilde\theta}(d\theta)\Big[1+\mathbb{E}\big[W(\theta,\theta')\big]+\mathbb{E}\big[W^2(\theta,\theta')/2\big]\Big]\Big\}$$
$$+n\alpha\lambda\big[\|\theta_1\|^2-\|\tilde\theta\|^2\big]-\frac{\beta}{2}\|\theta_1-\tilde\theta\|^2+\log(\varepsilon)$$
$$\geq -n\mathbb{E}\big[W(\tilde\theta,\theta_1)\big]-\frac{n}{2}\mathbb{E}\big[W^2(\tilde\theta,\theta_1)\big]$$
$$-\mathbb{E}\Big\{\frac{2n\alpha\|X\|^2}{\beta}\Big[2\alpha\big(\langle\tilde\theta,X\rangle-Y\big)^2+W(\theta_1,\tilde\theta)\Big]\Big\}-\frac{2n\alpha^2}{\beta^2}\mathbb{E}\big[\|X\|^4\big]$$
$$+n\alpha\lambda\big[\|\theta_1\|^2-\|\tilde\theta\|^2\big]-\frac{\beta}{2}\|\theta_1-\tilde\theta\|^2+\log(\epsilon)$$
$$\geq n\alpha\big[R_\lambda(\theta_1)-R_\lambda(\tilde\theta)\big]-\frac{3n}{2}\mathbb{E}\big[W^2(\theta_1,\tilde\theta)\big]$$
$$-\frac{4n\alpha^2}{\beta}\mathbb{E}\Big[\|X\|^2\big(\langle\tilde\theta,X\rangle-Y\big)^2\Big]-\frac{3n\alpha^2}{\beta^2}\mathbb{E}\big[\|X\|^4\big]$$
$$-\frac{\beta}{2}\|\theta_1-\tilde\theta\|+\log(\varepsilon)$$
$$\geq n\Big\{\alpha-\frac{\beta}{2n(q_{\min}+\lambda)}-\frac{3}{10}\alpha^2\Big[B_3+B_4\|Q_\lambda^{1/2}(\theta_1-\tilde\theta)\|^2\Big]\Big\}\big[R_\lambda(\theta_1)-R_\lambda(\tilde\theta)\big]$$
$$-\frac{4n\alpha^2}{\beta}\mathbb{E}\Big[\|X\|^2\big(\langle\tilde\theta,X\rangle-Y\big)^2\Big]+\frac{3n\alpha^2}{\beta^2}\mathbb{E}\big[\|X\|^4\big]+\log(\varepsilon).$$

We have proved

THEOREM 6.9 *Let Θ be some closed convex set of $\mathbb{R}^d$. Let $\tilde\theta=\arg\min_\Theta R_\lambda=\arg\min_{\theta\in\Theta}R(\theta)+\lambda\|\theta\|^2$. Let $\eta\geq 0$ be a quantity that will characterize the precision in minimizing the empirical criterion. Let us use the notation*

$$\arg\,(\min_{\theta\in\Theta}+\eta)f(\theta)=\big\{\theta\in\Theta:f(\theta)\leq\inf f+\eta\big\}.$$

With probability at least $1-\varepsilon$, for any $\eta\in\mathbb{R}_+$, any estimator satisfying

$$\hat\theta \in \arg\,(\min_{\theta_1\in\Theta}+\eta)\sup_{\theta_2\in\Theta} -\frac{1}{n}\sum_{i=1}^n \int \rho_{\theta_1}(d\theta')\log\Big\{\int\rho_{\theta_2}(d\theta)\Big[1 + W_i(\theta,\theta') + W_i^2(\theta,\theta')/2\Big]\Big\} + \alpha\lambda\big[\|\theta_1\|^2-\|\theta_2\|^2\big]$$

$$= \arg\,(\min_{\theta_1\in\Theta}+\eta)\sup_{\theta_2\in\Theta} -\frac{1}{n}\sum_{i=1}^n \int \rho_{\theta_1}(d\theta')\log\Big\{1 + W_i(\theta_2,\theta') + W_i^2(\theta_2,\theta')/2 + \frac{\alpha}{\beta}\|X_i\|^2 + \frac{\alpha}{\beta}\|X_i\|^2\big[2\alpha\big(\langle\theta',X_i\rangle - Y_i\big)^2 + 3W_i(\theta_2,\theta')\big] + \frac{3\alpha^2}{2\beta^2}\|X_i\|^4\Big\} + \alpha\lambda\big[\|\theta_1\|^2-\|\theta_2\|^2\big]$$

is such that

$$\Big\{\alpha - \frac{\beta}{2n(q_{\min}+\lambda)} - \frac{3}{10}\alpha^2\Big[B_3 + B_4\|Q_\lambda^{1/2}(\hat\theta-\tilde\theta)\|^2\Big]\Big\}\big[R_\lambda(\hat\theta)-R_\lambda(\tilde\theta)\big]$$
$$\le \sup_{\theta_2\in\Theta}\Big\{\frac{\beta}{2n(q_{\min}+\lambda)} + \frac{3}{10}\alpha^2\Big[B_3 + B_4\|Q_\lambda^{1/2}(\theta_2-\tilde\theta)\|^2\Big] - \alpha\Big\}\|Q_\lambda^{1/2}(\theta_2-\tilde\theta)\|^2 + \frac{8\alpha^2}{\beta}\mathbb{E}\Big[\|X\|^2\big(\langle\tilde\theta,X\rangle - Y\big)^2\Big] + \frac{6\alpha^2}{\beta^2}\mathbb{E}\big[\|X\|^4\big] + \frac{2}{n}\log(2\varepsilon^{-1}) + \eta.$$

Moreover

$$\|Q_\lambda^{1/2}(\hat\theta-\tilde\theta)\|^2 = \mathbb{E}\big(\langle\hat\theta-\tilde\theta,X\rangle^2\big) + \lambda\|\hat\theta-\tilde\theta\|^2 \le \|\hat\theta-\tilde\theta\|^2\Big[\sup_{v\in\mathbb{R}^d,\|v\|=1}\mathbb{E}\big(\langle v,X\rangle^2\big) + \lambda\Big].$$

6.2.1. Proof of Theorem 3.1. From Theorem 6.9, it is enough to remark from the definitions that $B_3 \le 40\sigma^2$, $B_4 \le 10\gamma_2$ and to bound

$$\mathbb{E}\big(\|X\|^4\big) = \sum_{k=1}^d\sum_{\ell=1}^d \mathbb{E}\big(X_k^2X_\ell^2\big) \le \sum_{k=1}^d\sum_{\ell=1}^d \mathbb{E}\big(X_k^4\big)^{1/2}\mathbb{E}\big(X_\ell^4\big)^{1/2} \le d^2\gamma_2(q_{\max}+\lambda)^2$$

6.2.2. Proof of Theorem 3.2. Let us now concentrate on the case when $\lambda = 0$, meaning that we do not use a ridge penalty. Under the assumptions of Theorem 3.2, we have

$$B_3 \le 40B\,R(f^*) \le 40B\,\sigma^2,$$

and

$$B_4 \le 10B^2.$$

Besides we have

$$\|Q^{1/2}(\theta-\theta^*)\|^2 = \mathbb{E}\big[\langle\theta,X\rangle - \langle\theta^*,X\rangle\big]^2 \le H^2$$

and

$$\|X\|^2 = \sum_{j=1}^{d} \varphi_j^2(X) \le dH^2.$$

Now let us choose $\alpha = [12B(4\sigma^2+BH^2)]^{-1}$ and $\beta = n\alpha q_{\min}$ so that we have

$$\alpha - \frac{\beta}{2nq_{\min}} - \frac{3}{10}\alpha^2\Big[B_3 + B_4\|Q^{1/2}(\hat\theta-\theta^*)\|^2\Big] \ge \alpha/4$$

From Theorem 6.9, we obtain that with probability $1-\varepsilon$,

$$R(\hat\theta) - R(\theta^*) \le \frac{32}{nq_{\min}}(dH^2\sigma^2) + \frac{24d^2H^4}{\alpha n^2 q_{\min}^2} + \frac{8}{n\alpha}\log(2\varepsilon^{-1}) + \frac{4\eta}{\alpha},$$

which is the desired result (when applied to $\eta = 0$).

6.2.3. Computation of the estimator. From Theorem 6.9, we see that we do not need to minimize exactly the empirical quantity appearing in Theorem 3.2. Let us show how to make an approximate optimization. With the notation of Section 6.2, let us introduce

$$F(\theta_1,\theta_2) = -\frac{1}{n}\sum_{i=1}^{n}\int\rho_{\theta_1}(d\theta')\log\bigg\{\int\rho_{\theta_2}(d\theta)\Big[1+W_i(\theta,\theta')+W_i^2(\theta,\theta')/2\Big]\bigg\} + n\alpha\lambda\big[\|\theta_1\|^2 - \|\theta_2\|^2\big].$$

The computation of F involves computing some expectation with respect to θ_1. As we need only to compute approximately the minimum of $\sup_{\theta_2\in\Theta} F(\theta_1,\theta_2)$, we can use the obvious lower bound :

$$f(\theta_1,\theta_2) = -\frac{1}{n}\sum_{i=1}^{n}\log\bigg\{\int\rho_{\theta_1}(d\theta')\int\rho_{\theta_2}(d\theta)\Big[1+W_i(\theta,\theta')+W_i^2(\theta,\theta')/2\Big]\bigg\} + n\alpha\lambda\big[\|\theta_1\|^2 - \|\theta_2\|^2\big].$$

This auxiliary function f can be computed explicitely, using Lemma 6.8 (page 40):

$$f(\theta_1,\theta_2) = -\frac{1}{n}\sum_{i=1}^{n}\log\bigg\{1 - W_i(\theta_1,\theta_2) + W_i^2(\theta_1,\theta_2)/2$$

$$+\frac{2\alpha^2\|X_i\|^2}{\beta}\Big[\big(\langle\theta_1,X_i\rangle-Y_i\big)^2+\big(\langle\theta_2,X_i\rangle-Y_i\big)^2\Big]+\frac{2\alpha^2\|X_i\|^4}{\beta^2}\Bigg\}$$
$$+n\alpha\lambda\big[\|\theta_1\|^2-\|\theta_2\|^2\big].$$

The convexity of $x\mapsto-\log(x)$ ensures that for any $\theta_1,\theta_2\in\mathbb{R}^d$,

$$f(\theta_1,\theta_2)\le F(\theta_1,\theta_2).$$

Therefore, for any $\hat\theta_1,\hat\theta_2\in\Theta$

$$\begin{aligned}\sup_{\theta_2\in\Theta}F(\hat\theta_1,\theta_2)-\inf_{\theta_1\in\Theta}\sup_{\theta_2\in\Theta}F(\theta_1,\theta_2)&\\ \le\sup_{\theta_2\in\Theta}F(\hat\theta_1,\theta_2)-\inf_{\theta_1\in\Theta}\sup_{\theta_2\in\Theta}f(\theta_1,\theta_2)&\\ \le\sup_{\theta_2\in\Theta}F(\hat\theta_1,\theta_2)-\inf_{\theta_1\in\Theta}f(\theta_1,\hat\theta_2).&\end{aligned}$$

Since we do not want to compute $\sup_{\theta_2\in\Theta}F(\hat\theta_1,\theta_2)$ either, we may now introduce the upper bound

$$F(\theta_1,\theta_2)\le g(\theta_1,\theta_2),$$

where

$$g(\theta_1,\theta_2)=\frac{1}{n}\sum_{i=1}^n\log\Bigg\{\int\rho_{\theta_1}(d\theta')\int\rho_{\theta_2}(d\theta)\Big[1+W_i(\theta',\theta)+W_i^2(\theta,\theta')/2\Big]\Bigg\}$$
$$+n\alpha\lambda\big[\|\theta_1\|^2-\|\theta_2\|^2\big].$$

Of course, g can be computed explicitely, similarly to f:

$$g(\theta_1,\theta_2)=\frac{1}{n}\sum_{i=1}^n\log\Bigg\{1+W_i(\theta_1,\theta_2)+W_i^2(\theta_1,\theta_2)/2$$
$$+\frac{2\alpha^2\|X_i\|^2}{\beta}\Big[\big(\langle\theta_1,X_i\rangle-Y_i\big)^2+\big(\langle\theta_2,X_i\rangle-Y_i\big)^2\Big]+\frac{2\alpha^2\|X_i\|^4}{\beta^2}\Bigg\}$$
$$+n\alpha\lambda\big[\|\theta_1\|^2-\|\theta_2\|^2\big].$$

PROPOSITION 6.10 *For any* $\hat\theta_1,\hat\theta_2\in\Theta$

$$\begin{aligned}\sup_{\theta_2\in\Theta}F(\hat\theta_1,\theta_2)-\inf_{\theta_1\in\Theta}\sup_{\theta_2\in\Theta}F(\theta_1,\theta_2)&\\ \le\sup_{\theta_2\in\Theta}g(\hat\theta_1,\theta_2)-\inf_{\theta_1\in\Theta}\sup_{\theta_2\in\Theta}f(\theta_1,\theta_2)&\\ \le\sup_{\theta_2\in\Theta}g(\hat\theta_1,\theta_2)-\inf_{\theta_1\in\Theta}f(\theta_1,\hat\theta_2).&\end{aligned}$$

This results shows how to obtain an empirical estimate of the default of optimality η in Theorem 6.9 (page 42). In view of this, a sensible choice of estimator is to take

$$\widehat{\theta}_1 = \widehat{\theta}_2 \in \arg\min_{\theta_1\in\Theta} \sup_{\theta_2\in\Theta} g(\theta_1,\theta_2).$$

6.3. 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 6.11 *For any real-valued random variable V such that $\mathbb{E}\big[\exp(V)\big] \le 1$, with probability at least $1-\varepsilon$, we have*

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

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

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

To prove the theorem, according to Lemma 6.11, 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.

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

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

From Jensen's inequality again,

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

$$= -\log \int \exp\bigl[\hat{L}(\hat{f},f) + \gamma^* \bar{R}(f)\bigr] \pi^*_{-\gamma^*\bar{R}}(df) - \log \int \exp\bigl[-\gamma^* \bar{R}(f)\bigr] \pi^*(df)$$
$$\leq -\int [\hat{L}(\hat{f},f) + \gamma^* \bar{R}(f)] \pi^*_{-\gamma^*\bar{R}}(df) + \mathcal{I}^*(\gamma^*).$$

From the two previous inequalities, we get

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

hence, by using Fubini's inequality and the equality

$$\mathbb{E}\Bigl\{ \exp\bigl[-\hat{L}(\hat{f},f)\bigr] \Bigr\} = \exp\bigl[-L^\flat(\hat{f},f)\bigr],$$

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

$$\begin{aligned}
& \leq \mathbb{E} \int \left(\int \exp\bigl[L^\flat(\hat{f},f) - \hat{L}(\hat{f},f)\bigr] \pi^*_{-\gamma^*\bar{R}}(df) \right) \pi_{-\gamma\bar{R}}(d\hat{f}) \\
& = \int \left(\int \mathbb{E} \exp\bigl[L^\flat(\hat{f},f) - \hat{L}(\hat{f},f)\bigr] \pi^*_{-\gamma^*\bar{R}}(df) \right) \pi_{-\gamma\bar{R}}(d\hat{f}) = 1.
\end{aligned}$$

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

LEMMA 6.12 *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

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

  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. [7, page 159]): for any real-valued measurable function h defined on $\mathcal{A}$,

$$\inf_{\rho\in\mathcal{M}}\big\{\mathbb{E}_{a\sim\rho}\,h(a)+K(\rho,\mu)\big\} = -\log\mathbb{E}_{a\sim\mu}\Big\{\exp\big[-h(a)\big]\Big\}. \qquad (6.18)$$

  By using twice (6.18) and Fubini's theorem, we have

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

- By using twice (6.18) and the first assertion of Lemma 6.12, we have

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

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

From Lemma 6.12 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 $\gamma \ge 0$, $\gamma^* \ge 0$ 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 $1 - 2\varepsilon$:

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

6.4. 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

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

$c_\phi \geq 0$ (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

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

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

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

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

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

which proves the announced result.

6.5. 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

$$a(\lambda) = \frac{M^2 A^2 \exp(Hb_2/A)}{2\sqrt{\pi}(1 - |\lambda|AH)}$$

and

$$\tilde{A} = Hb_2/2 + A\log(M) = \frac{A}{2}\log\big\{M^2 \exp\big[Hb_2/(2A)\big]\big\}.$$

From Taylor-Lagrange formula, we have

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

Since $\mathbb{E}\big[\exp\big(|\Omega|/A\big) \,|\, X\big] \leq M \exp\big[Hb_2/(2A)\big]$, Lemma D.2 gives

$$\log\Big\{\mathbb{E}\Big[\exp\big\{\alpha[\Omega - \mathbb{E}(\Omega|X)]/A\big\} \,|\, X\Big]\Big\} \leq \frac{M^2\alpha^2 \exp\big(Hb_2/A\big)}{2\sqrt{\pi}(1 - |\alpha|)}$$

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

$$\big|\mathbb{E}(\Omega|X)\big| \leq \tilde{A}. \tag{6.19}$$

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

$$\log\Big\{\mathbb{E}\Big[\exp\big[L - \mathbb{E}(L|X)\big] \,|\, X\Big]\Big\} \leq \lambda^2(F - F^*)^2 a(\lambda). \tag{6.20}$$

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, (6.20) and (6.19), 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

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

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 $t = 1$

$$\frac{b_1}{2} \mathbb{E}(F - F^*)^2 \leq R(f) - R(f^*). \tag{6.21}$$

6.6. PROOF OF LEMMA 5.6. We have

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

where the last inequality is the usual relation between excess risk and L^2 distance using the convexity of $\mathcal{F}$ (see above (6.21) 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 $n \in \mathbb{N}^*$, 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 [2]. 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 [13, 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 [4]. 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 [4]. This modifies Theorem 5 of [4] 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 $m \geq 2$

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

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 [4] holds for $D = d$. Inequality (5.8) for $\tau = 1/2$ of [4] 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 [4] 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 [4] allows to bracket $\tilde{B}$ in terms of $B = \sup_{f\in\mathrm{span}\{\varphi_1,\ldots,\varphi_d\}} \|f\|_\infty^2/\mathbb{E}[f(X)]^2$, 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 $A_1, \ldots, A_d$ be a partition of $\mathcal{X}$, i.e., $\mathcal{X} = \sqcup_{j=1}^d A_j$. Now consider the indicator functions $\varphi_j = 1_{A_j}, j = 1, \ldots, d$: 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 [5], 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 [5] 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 $\mathcal{A}(\lambda) \leq \frac{\lambda^2}{q_{\min}}\|\theta^*\|^2$, $\mathcal{B}(\lambda) \leq \frac{\lambda^2}{q^2_{\min}}\|\theta^*\|^2$, and $\mathcal{N}(\lambda) \leq d$. The result is obtained by noticing that the leading terms in (34) of [5] 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 54. 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 54 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

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

□

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 $-1 < \alpha < 1$,*

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

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

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

□

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. Fast learning rates in statistical inference through aggregation. *Annals of Statistics*, 2009.

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

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

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

[6] O. Catoni. A PAC-Bayesian approach to adaptive classification. Preprint n.840, `http://www.dma.ens.fr/~catoni/homepage/dea2005.pdf`, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, 2003.

[7] 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.

[8] 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.

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

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

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

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

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

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

[15] 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.

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

[17] 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.

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