Linear regression through PAC-Bayesian truncation

JEAN-YVES AUDIBERT[1,2], OLIVIER CATONI [3,4]

September 11, 2011

ABSTRACT : We consider the problem of predicting as well as the best linear combination of d given functions in least squares regression under L^∞ constraints on 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, complex constants involving the conditioning of the Gram matrix of the covariates, kurtosis coefficients or some geometric quantity characterizing the relation between L^2 and L^∞-balls and require some additional assumptions like exponential moments of the output.

This work provides a PAC-Bayesian shrinkage procedure with a simple excess risk bound of order d/n holding in expectation and in deviations, under various assumptions. The common surprising factor of these results is their simplicity and the absence of exponential moment condition on the output distribution while achieving exponential deviations. The risk bounds are obtained through a PAC-Bayesian analysis on truncated differences of losses. We also show that these results can be generalized to other 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

OUR STATISTICAL TASK 3
OUTLINE AND CONTRIBUTIONS 5
1. VARIANTS OF KNOWN RESULTS 5
1.1. ORDINARY LEAST SQUARES AND EMPIRICAL RISK MINIMIZATION 5

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

[2]Sierra, CNRS/ENS/INRIA — UMR 8548, 45 rue d'Ulm, 75230 Paris cedex 05, France

[3]Département de Mathématiques et Applications, CNRS – UMR 8553, École Normale Supérieure, 45 rue d'Ulm, 75230 Paris cedex 05, France, olivier.catoni@ens.fr

[4]INRIA Paris-Rocquencourt - CLASSIC team.

1.2. Projection estimator . . . 10
1.3. Penalized least squares estimator . . . 10
1.4. Conclusion of the survey . . . 11
2. A simple tight risk bound for a sophisticated PAC-Bayes algorithm . . . 12
3. A generic localized PAC-Bayes approach . . . 15
3.1. Notation and setting . . . 15
3.2. The localized PAC-Bayes bound . . . 17
3.3. Application under an exponential moment condition . . . 18
3.4. Application without exponential moment condition . . . 20
4. Proofs . . . 24
4.1. Main ideas of the proofs . . . 24
4.1.1. Sub-exponential tails under a non-exponential moment assumption via truncation . . . 24
4.1.2. Localized PAC-Bayesian inequalities to eliminate a logarithm factor . . . 25
4.2. Proof of Theorem 3.1 . . . 26
4.2.1. Proof of $\mathbb{E}\Big\{\int \exp\big[V_1(\hat{f})\big]\rho(d\hat{f})\Big\} \leq 1$ . . . 27
4.2.2. Proof of $\mathbb{E}\Big[\int \exp(V_2)\rho(d\hat{f})\Big] \leq 1$ . . . 28
4.3. Proof of Lemma 3.3 . . . 30
4.4. Proof of Lemma 3.4 . . . 31
4.5. Proof of Lemma 3.6 . . . 33
4.6. Proof of Lemma 3.7 . . . 33
A. Uniformly bounded conditional variance is necessary to reach d/n rate . . . 34
B. Empirical risk minimization on a ball: analysis derived from the work of Birgé and Massart . . . 35
C. Ridge regression analysis from the work of Caponnetto and De Vito . . . 37
D. Some standard upper bounds on log-Laplace transforms . . . 38

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

$$f^* = \sum_{j=1}^{d} \theta_j^* \varphi_j \in \operatorname*{argmin}_{f \in \mathcal{F}} R(f). \tag{0.1}$$

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. Except in particular settings (e.g., when Θ is a probability simplex[5] and $d \geq \sqrt{n}$), it is known that the convergence rate d/n cannot be improved in a minimax sense (see [25], and [27] 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

[5]This corresponds to the convex aggregation problem, which has been widely studied by several authors since the work of Nemirovski and Judisky [22, 18]. This particular setting is not the topic of this paper, but our results apply to it, and correspond to the minimax optimal rate for $d \leq \sqrt{n}$. For $d > \sqrt{n}$, the minimax optimal rate of convex aggregation is $\sqrt{\log(1 + d/\sqrt{n})/n}$, which is not achieved by our procedure.

$\kappa > 0$, for any $\varepsilon > 0$, with probability at least $1 - \varepsilon$,

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

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 (0.2) implies

$$\mathbb{E}R(\hat{f}) - R(f^*) \leq \kappa \frac{d+1}{n}. \tag{0.3}$$

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

$$\mathbb{E}R(\hat{f}) - R(f^{(\text{reg})}) \leq C[R(f^*) - R(f^{(\text{reg})})] + \kappa \frac{d}{n}, \tag{0.4}$$

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 [16] and references within). Here we intend to assess the generalization ability of the estimator even when the model is misspecified (namely when $R(f^*) > R(f^{(\text{reg})})$). Moreover we do not assume either that $Y - f^{(\text{reg})}(X)$ and X are independent.

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

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

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

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

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

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

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

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

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

$$\tilde{\theta} \in \arg\min_{\theta\in\Theta} \big\{R(f_\theta) + \lambda\|\theta\|^2\big\}. \tag{0.5}$$

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

OUTLINE AND CONTRIBUTIONS. The paper is organized as follows. Section 1 is a survey on risk bounds in linear least squares regression. Theorems 1.3 and 1.5 are the results which come closer to our target. Section 2 presents our main result on linear least squares regression. Section 3 gives risk bounds for general loss functions from which the results of Section 2 are derived. Appendix A shows that (0.2) cannot hold under the only assumption that the variance of Y is finite, even in the favorable situation where $f^{(reg)}$ belongs to $\mathcal{F}$.

The main contribution of this paper is to show that an appropriate shrinkage estimator involving truncated differences of losses has an excess risk of order d/n (without a logarithmic factor as it appears in numerous works), concentrating exponentially, which does not degrade when the matrix Q is ill-conditioned or when some ratio of L^2 and L^∞ norms behaves badly or when the output distribution is heavy-tailed. Our results tend to say that shrinkage and truncation lead to more robust algorithms when we consider robustness with respect to the distribution of the noise, and not to a potential contamination of the training data by input-output pairs not generated by P.

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 linear least squares regression. It minimizes the empirical risk

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

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

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

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

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

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

- the linear system (1.1) has at least one solution, and in fact, the set of solutions is exactly $\{\mathbf{X}^+ \mathbf{Y} + u; u \in \ker \mathbf{X}\}$; where $\mathbf{X}^+$ is the Moore-Penrose pseudoinverse of $\mathbf{X}$ and $\ker \mathbf{X}$ is the kernel of the linear operator $\mathbf{X}$.
- $\mathbf{X}\, \hat{\theta}^{(\text{ols})}$ is the (unique) orthogonal projection of the vector $\mathbf{Y} \in \mathbb{R}^n$ on the image of the linear map $\mathbf{X}$;
- if $\sup_{x \in \mathcal{X}} \mathbb{V}\text{ar}(Y|X = x) = \sigma^2 < +\infty$, we have (see [16, Theorem 11.1]) for any $X_1, \dots, X_n$ in $\mathcal{X}$,

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

  where we recall that $f^{(\text{reg})} : x \mapsto \mathbb{E}[Y|X = x]$ is the optimal regression function, and that when this function belongs to $\mathcal{F}_{\text{lin}}$ (i.e., $f^{(\text{reg})} = f^*_{\text{lin}}$), the minimum term in (1.2) vanishes;
- from Pythagoras' theorem for the (semi)norm $W \mapsto \sqrt{\mathbb{E}W^2}$ on the space of the square integrable random variables,

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

The analysis of the ordinary least squares often stops at this point in classical statistical textbooks. (Besides, to simplify, the strong assumption $f^{(\text{reg})} = f^*_{\text{lin}}$ is often made.) This can be misleading since Inequality (1.2) does not imply a d/n upper bound on the risk of $\hat{f}^{(\text{ols})}$. Nevertheless the following result holds [16, Theorem 11.3].

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

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

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

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

for some numerical constant κ*.*

Using PAC-Bayesian inequalities, Catoni [10, Proposition 5.9.1] has proved a different type of results on the generalization ability of $\hat{f}^{(\mathrm{ols})}$.

THEOREM 1.2 *Let* $\mathcal{F}' \subset \mathcal{F}_{\mathrm{lin}}$ *be such that for some positive constants* a, M, M'*:*

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

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

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

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*

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

we have

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

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 $\mathcal{F}' \subset \mathcal{F}_{\mathrm{lin}}$ (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 [10, 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 [9, 11], one can derive from Inequality (6.9) and Lemma 4.1 of Alquier [1] the following result.

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

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

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

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

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

for some κ not depending on d and n.

Using the result of [10, 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 [19, 6] 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 $1-\varepsilon$:

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

where κ is a numerical constant.

PROOF. The result is a modified version of Theorem 6.7 in [6] 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 [7], 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 upper bounded by H for the L^∞-norm, i.e., for any f_1, f_2 in $\mathcal{F}$, $\sup_{x\in\mathcal{X}}|f_1(x) - f_2(x)| \leq H$ and there exists a function $f_0 \in \mathcal{F}$ satisfying the exponential moment condition:*

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

for some positive constants A and M. Let

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

where the infimum is taken with respect to all possible orthonormal basis of $\mathcal{F}$ for the dot product $\langle f_1, f_2\rangle = \mathbb{E}f_1(X)f_2(X)$ (when the set $\mathcal{F}$ admits no basis with exactly d functions, we set $\tilde{B} = +\infty$). Then the empirical risk minimizer satisfies for any $\varepsilon > 0$, with probability at least $1-\varepsilon$:

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

where κ is a positive constant depending only on M.

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

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

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

with

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

The following excess risk bound of order d/n for this estimator is Theorem 4 in [25] up to minor changes in the assumptions.

THEOREM 1.6 *If* $\sup_{x\in\mathcal{X}} \mathbb{V}\text{ar}(Y|X=x) = \sigma^2 < +\infty$ *and*

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

then we have

$$\mathbb{E}R(\hat{f}^{(\text{proj})}) - R(f^*_{\text{lin}}) \leq (\sigma^2 + H^2)\frac{d}{n}. \tag{1.8}$$

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 ([20, 23]). This solution has been labeled ridge regression in statistics ([17]), and consists in replacing $\hat{f}^{(\text{ols})}$ by $\hat{f}^{(\text{ridge})} = f_{\hat{\theta}^{(\text{ridge})}}$ with

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

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

Risk bounds for this estimator can be derived from general results concerning penalized least squares on reproducing kernel Hilbert spaces ([8]), 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 [24] and is defined by

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

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

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

on the 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.

Our recent work [4] provides a risk bound for ridge regression showing the benefit on the effective dimension of the shrinkage parameter λ and being of order d/n (without logarithmic factor). The work [4] also proposes a robust estimator for linear least squares, which satisfies a d/n excess risk bound without logarithmic factor, but with constants involving several kurtosis coefficients. As discussed in Section 3.2 of [4], depending on the basis functions and the distribution P, these kurtosis coefficients typically behave either as numerical constants or $\sqrt{d}$ (but worse non-asymptotic behaviors of these constants can also occur).

Finally, several works, and in particular those cited in Section 1.1, have considered the problem of model selection where several linear spaces are simultaneously considered, and the goal is to predict as well as the best function in the union of the linear spaces. Only a few of them considered the case of outputs having only finite conditional moments (and not finite conditional exponential moments). This is the case of [5] in the fixed design setting and [26] in the random design setting. The excess risk bounds there are typically of order d/n with d the dimension of the "best" linear space, but holds in expectation and essentially when the optimal regression function $f^{(\mathrm{reg})}$ belongs to the union of linear spaces.

2. A simple tight risk bound for a sophisticated PAC-Bayes algorithm

In this section, we provide a sophisticated estimator, having a simple theoretical excess risk bound, with neither a logarithmic factor, nor complex constants involving the conditioning of Q, kurtosis coefficients or some geometric quantity characterizing the relation between L^2 and L^∞-balls.

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

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

Introduce

$$\hat{\mathcal{E}}(f) = \log \int \frac{\pi(df')}{\prod_{i=1}^n [1 - W_i(f, f') + \frac{1}{2} W_i(f, f')^2]}. \tag{2.1}$$

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

$$\frac{d\hat{\pi}}{d\pi}(f) = \frac{\exp[-\hat{\mathcal{E}}(f)]}{\int \exp[-\hat{\mathcal{E}}(f')]\pi(df')}. \tag{2.2}$$

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 2.1 *Assume that $\mathcal{F}$ has a diameter upper bounded by H for the L^∞-norm:*

$$\sup_{f_1, f_2 \in \mathcal{F}, x \in \mathcal{X}} |f_1(x) - f_2(x)| \le H \tag{2.3}$$

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

$$\sup_{x \in \mathcal{X}} \mathbb{E}\big\{[Y - f^*(X)]^2 \big| X = x\big\} \le \sigma^2 < +\infty. \tag{2.4}$$

Let $\hat{f}$ be a prediction function drawn from the distribution $\hat{\pi}$ defined in (2.2) *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*

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

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

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

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

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

PROOF. This is a direct consequence of Theorem 3.5 (page 21), Lemma 3.3 (page 19) and Lemma 3.6 (page 23). □

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 2.1 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 34) 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. With nowadays computers and numerical methods, it seems impossible to get a good approximation of it even when the dimension d is small. Nevertheless, in presence of a heavy-tailed noise distribution, it can be a way to move from the empirical risk minimizer (which is the baseline estimator for linear regression) in the right direction (that is in a direction in which one can find an estimator having a smaller risk than the one of the empirical risk minimizer). When the target is to predict as well as the best linear combination f^*_{lin} up to a small additive term, the estimator 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. Note that the possible lack of knowledge of H and σ call for a model selection algorithm, which goes beyond the scope of this work. In practice, a careful application of (cross-)validation ideas would probably be sufficient to select these parameters.

REMARK 2.2 The proposed randomized estimator is more complex than the classical Gibbs estimator (that is the one with exponential weights involving the empirical risk). Even if the paper does not prove it, (we believe that) the classical Gibbs estimator cannot be robust to heavy-tailed noise. This belief is motivated by the same arguments as the ones used in [12] to show the absence of robustness of the empirical mean estimator. In absence of heavy-tailed noise, the classical Gibbs estimator satisfies a similar result to Theorem 2.1, given in Theorem 3.2.

Our randomized algorithm consists in drawing the prediction function according to $\hat{\pi}$. As usual, by convexity of the loss function, the risk of the deterministic estimator $\hat{f}_{\text{determ}} = \int f \hat{\pi}(df)$ satisfies $R(\hat{f}_{\text{determ}}) \leq \int R(f)\hat{\pi}(df)$, so that, after some computations, one can prove that for any $\varepsilon > 0$, with probability at least

$1-\varepsilon$:

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

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

REMARK 2.3 We consider a "prior" distribution π, which is a uniform distribution on $\mathcal{F}$. In presence of sparsity (when only a small number of the coefficients θ^*_j in (0.1) are nonzero), alternative prior distributions (of Laplace form) are useful in fixed design regression [13, 14, 2] and in the random design scenario [15, 2]. When the coefficient vector θ^* is non-sparse (which is not the focus of these works), the latter papers prove a $\frac{d \log n}{n}$ risk bound when the noise distribution admits at least sub-exponential tails.

REMARK 2.4 Theorem 2.1 expresses boundedness in terms of the L^∞ diameter of the set of functions $\mathcal{F}$. Besides, (2.4) implies that the function $f^{(\text{reg})} : x \mapsto \mathbb{E}[Y|X = x]$ satisfies $f^{(\text{reg})}(X) - f^*(X) \leq \sigma$ almost surely. By using Lemma 3.7 (page 23) instead of Lemma 3.6 (page 23), Theorem 2.1 still holds without assuming (2.3) and (2.4), when replacing $(2\sigma + H)^2$ with

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

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

3. A GENERIC LOCALIZED PAC-BAYES APPROACH

3.1. NOTATION AND SETTING. In this section, we drop the restrictions of the linear least squares setting considered so far in order to focus on the ideas underlying the estimator and the results presented in Section 2. 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 as

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

Let $\ell : \mathcal{Z}\times\mathcal{F}\times\mathcal{F} \to \mathbb{R}$ be a function such that $\ell(Z, f, f')$ represents[6] 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

$$\int \mathbb{E}\big\{\exp[L(f, f')]\big\}^n \pi^*(df') < +\infty, \tag{3.1}$$

$$\text{and} \quad \int \frac{\pi(df)}{\int \mathbb{E}\big\{\exp[L(f, f')]\big\}^n \pi^*(df')} < +\infty. \tag{3.2}$$

We consider the real-valued processes

$$\hat{L}(f, f') = \sum_{i=1}^n L_i(f, f'), \tag{3.3}$$

$$\hat{\mathcal{E}}(f) = \log \int \exp\big[\hat{L}(f, f')\big]\, \pi^*(df'), \tag{3.4}$$

$$L^\flat(f, f') = -n \log\Big\{\mathbb{E}\Big[\exp\big(-L(f, f')\big)\Big]\Big\}, \tag{3.5}$$

$$L^\sharp(f, f') = n \log\Big\{\mathbb{E}\Big[\exp\big(L(f, f')\big)\Big]\Big\}, \tag{3.6}$$

$$\text{and} \quad \mathcal{E}^\sharp(f) = \log\bigg\{\int \exp\big[L^\sharp(f, f')\big]\, \pi^*(df')\bigg\}. \tag{3.7}$$

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

$$L^\flat \le n\mathbb{E}(L) = \mathbb{E}(\hat{L}) \le L^\sharp. \tag{3.8}$$

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 $\mathcal{F}$.

[6] 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 (3.4, page 16) corresponds to (2.1, page 12) with this choice of function ℓ and for the choice $\pi^* = \pi$.

For a distribution ρ on $\mathcal{F}$ absolutely continuous w.r.t. π, let $\frac{d\rho}{d\pi}$ denote the density of ρ w.r.t. π. For any real-valued (measurable) function h defined on $\mathcal{F}$ such that $\int \exp[h(f)]\pi(df) < +\infty$, we define the distribution π_h on $\mathcal{F}$ by its density:

$$\frac{d\pi_h}{d\pi}(f) = \frac{\exp[h(f)]}{\int \exp[h(f')]\pi(df')}. \tag{3.9}$$

We will use the posterior distribution:

$$\frac{d\hat{\pi}}{d\pi}(f) = \frac{d\pi_{-\hat{\mathcal{E}}}}{d\pi}(f) = \frac{\exp[-\hat{\mathcal{E}}(f)]}{\int \exp[-\hat{\mathcal{E}}(f')]\pi(df')}. \tag{3.10}$$

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

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

and

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

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

THEOREM 3.1 *Assume that π, π^*, $\mathcal{F}$ and ℓ satisfy the integrability conditions* (3.1) *and* (3.2, page 16)*. Let ρ be a (posterior) probability distribution on $\mathcal{F}$ admitting a density with respect to π depending on $Z_1, \ldots, Z_n$. Let $\hat{f}$ be a prediction function drawn from the distribution ρ. Then for any $\gamma \geq 0$, $\gamma^* \geq 0$ and $\varepsilon > 0$, with probability (with respect to the distribution $P^{\otimes n}\rho$ generating the observations $Z_1, \ldots, Z_n$ and the randomized prediction function $\hat{f}$) at least $1 - \varepsilon$:*

$$\begin{aligned}\int \big[L^\flat(\hat{f}, f) + \gamma^*\bar{R}(f)\big]\pi^*_{-\gamma^*\bar{R}}(df) - \gamma\bar{R}\big(\hat{f}\,\big) \\ \leq \mathcal{I}^*(\gamma^*) - \mathcal{I}(\gamma) - \log\Big\{\int \exp\big[-\mathcal{E}^\sharp(f)\big]\pi(df)\Big\} \\ + \log\left[\frac{d\rho}{d\hat{\pi}}\big(\hat{f}\,\big)\right] + 2\log\big(2\varepsilon^{-1}\big).\end{aligned} \tag{3.11}$$

PROOF. See Section 4.2 (page 26). □

Some extra work will be needed to prove that Inequality (3.11) provides an upper bound on the excess risk $\bar{R}(\hat{f})$ of the estimator $\hat{f}$. As we will see in the next sections, despite the $-\gamma\bar{R}(\hat{f})$ term and provided that γ is sufficiently small, the left-hand side will be essentially lower bounded by $\lambda n\bar{R}(\hat{f})$, while, by choosing $\rho = \hat{\pi}$, the estimator does not appear in the right-hand side.

3.3. APPLICATION UNDER AN EXPONENTIAL MOMENT CONDITION. The estimator proposed in Section 2 and Theorem 3.1 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 2.1 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 (2.2) with $\hat{\mathcal{E}}$ given by (2.1) and yet still obtain a d/n convergence rate. Secondly it illustrates Theorem 3.1 in a different and simpler way than the one we will use to prove Theorem 2.1.

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 3.2 *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

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

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 16) and we can apply Theorem 3.1. 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 $\rho = \hat{\pi}$, (3.11) gives

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

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

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

which implies the announced result by reparameterization (taking $\gamma = \lambda n\eta'$). □

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

LEMMA 3.3 *Let Θ be a bounded convex set of $\mathbb{R}^d$, and $\varphi_1,\dots,\varphi_d$ be d square integrable prediction functions. Assume that*

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

π is the uniform distribution on $\mathcal{F}$ (i.e., the one coming from the uniform distribution on Θ), and that there exist $0 < b_1 \le b_2$ such that for any $y \in \mathbb{R}$, the

function $\tilde{\ell}_y : y' \mapsto \tilde{\ell}(y, y')$ admits almost everywhere a second derivative such that, $(y, y') \mapsto \tilde{\ell}''_y(y')$ is measurable, for any $y, y' \in \mathbb{R}$, $b_1 \leq \tilde{\ell}''_y(y') \leq b_2$, and

$$\tilde{\ell}(y, y') = \tilde{\ell}(y, y) + (y' - y)\tilde{\ell}'_y(y) + \int_y^{y'} (y' - y'')\tilde{\ell}''_y(y'')dy''.$$

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 4.3 (page 30). □

REMARK 3.1 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 3.4 *Assume that the loss function $\tilde{\ell}$ satisfies the conditions stated in Lemma 3.3. Assume moreover that there exist $A > 0$ and $M > 0$ such that for any $x \in \mathcal{X}$,*

$$\mathbb{E}\Big\{\exp\Big[A^{-1}\big|\tilde{\ell}'_Y\big[f^*(X)\big]\big|\Big] \,\Big|\, X = x\Big\} \leq M.$$

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

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

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 4.4 (page 31). □

3.4. APPLICATION WITHOUT EXPONENTIAL MOMENT CONDITION. When we do not have finite exponential moments as assumed by Condition V1 (page 18), 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 3.1 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 3.2 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 3.5 *Assume that Conditions* V2 *above and* C *(page 18) are satisfied. Let* $0 < \lambda < V^{-1}$ *and*

$$\ell\big[(X,Y), f, f'\big] = T\Big(\lambda\big\{\tilde{\ell}\big[Y, f(X)\big] - \tilde{\ell}\big[Y, f'(X)\big]\big\}\Big), \tag{3.12}$$

with

$$T(x) = -\log(1 - x + x^2/2). \tag{3.13}$$

Let $\hat{f} \in \mathcal{F}$ *be a function drawn according to the distribution* $\hat{\pi}$ *defined in* (3.10, page 17) *with* $\hat{\mathcal{E}}$ *defined in* (3.4, page 16) *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*

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

with

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

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

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

PROOF. We apply Theorem 3.1 for ℓ given by (3.12) and $\pi^* = \pi$. Let us define, for any $f, f' \in \mathcal{F}$, $W(f, f') = \lambda\Big\{\tilde{\ell}\big[Y, f(X)\big] - \tilde{\ell}\big[Y, f'(X)\big]\Big\}$. Since $\log u \leq u - 1$ for any $u > 0$, we have

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

Moreover, from Assumption V2,

$$\frac{\mathbb{E}\big[W(f,f')^2\big]}{2} \leq \mathbb{E}\big[W(f,f^*)^2\big] + \mathbb{E}\big[W(f',f^*)^2\big] \leq \lambda^2 V\bar{R}(f) + \lambda^2 V\bar{R}(f'), \tag{3.14}$$

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

$$L^\flat(f, f') \geq \lambda n \Big[\bar{R}(f) - \bar{R}(f') - \lambda V \bar{R}(f) - \lambda V \bar{R}(f') \Big] \\ = \lambda n \Big[(1-\eta)\bar{R}(f) - (1+\eta)\bar{R}(f') \Big]. \tag{3.15}$$

Noting that

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

we see that

$$L^\sharp = n \log\Big\{ \mathbb{E}\Big[\exp\big[T(W)\big] \Big] \Big\} \leq n \Big[\mathbb{E}\big(W\big) + \mathbb{E}\big(W^2\big)/2 \Big].$$

Using (3.14) and still $\eta = \lambda V$, we get

$$L^\sharp(f, f') \leq \lambda n \Big[\bar{R}(f) - \bar{R}(f') + \eta \bar{R}(f) + \eta \bar{R}(f') \Big] \\ = \lambda n (1+\eta) \bar{R}(f) - \lambda n (1-\eta) \bar{R}(f'),$$

and

$$\mathcal{E}^\sharp(f) \leq \lambda n (1+\eta) \bar{R}(f) - \mathcal{I}\big(\lambda n (1-\eta)\big). \tag{3.16}$$

Plugging (3.15) and (3.16) in (3.11) for $\rho = \hat{\pi}$, we obtain

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

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

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

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

REMARK 3.2 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 3.3, in which we consider the exponential moment condition V1, we took $\ell\big[(X,Y), f, f'\big] = \lambda\big\{\tilde{\ell}\big[Y, f(X)\big] - \tilde{\ell}\big[Y, f'(X)\big]\big\}$ 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 $\ell\big[(X,Y),f,f'\big] = \lambda\big\{\tilde{\ell}\big[Y,f(X)\big] - \tilde{\ell}\big[Y,f'(X)\big]\big\}$ but have instead to use a truncated version. The truncation function T of Theorem 3.5 can be replaced by the simpler function $u \mapsto (u \vee -M) \wedge M$ for some appropriate constant $M > 0$ but this leads 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 polynomials $P^\flat$ and $P^\sharp$ such that $\frac{1}{P^\flat(u)} \leq \exp\big[T(u)\big] \leq P^\sharp(u)$ and $P^\flat(u)P^\sharp(u) \leq 1 + \mathrm{O}(u^4)$ for $u \to 0$, which are reasonable properties to ask in order to ensure that (3.8), and consequently (3.11), are tight.

Besides, if we take ℓ as in (3.12) 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 leads to the estimator considered in Theorems 3.5 and 2.1.

REMARK 3.3 Theorem 3.5 still holds for the same randomized estimator in which (3.13, page 21) is replaced with

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

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

LEMMA 3.6 *Consider the least squares setting: $\tilde{\ell}(y,y') = (y-y')^2$. Assume that $\mathcal{F}$ is convex and has a diameter upper bounded by H for the L^∞-norm:*

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

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

$$\sup_{x\in\mathcal{X}} \mathbb{E}\big\{[Y - f^*(X)]^2 \big| X = x\big\} \leq \sigma^2 < +\infty. \tag{3.17}$$

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

PROOF. See Section 4.5 (page 33). □

LEMMA 3.7 *Consider the least squares setting: $\tilde{\ell}(y,y') = (y-y')^2$. Assume that $\mathcal{F}$ (i.e., Θ) is bounded, and that for any $j \in \{1,\dots,d\}$, $\mathbb{E}\big[\varphi_j(X)^4\big] < +\infty$ and $\mathbb{E}\big\{\varphi_j(X)^2[Y - f^*(X)]^2\big\} < +\infty$. Then Condition* V2 *holds for*

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

PROOF. See Section 4.6 (page 33). □

4. PROOFS

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

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

$$R(f_{\hat{\theta}}) - R(f^*) = \big(\mathbb{E}(Y) - \hat{\theta}\,\big)^2 \leq \frac{c\,\log(\varepsilon^{-1})}{n}, \tag{4.1}$$

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

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

which is far from the objective (4.1) for small confidence levels (consider $\varepsilon = \exp(-\sqrt{n})$ for instance). The key idea is thus to average (soft) *truncated* values of the outputs. This is performed by taking

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

with $\lambda = \sqrt{\frac{2\log(\varepsilon^{-1})}{n}}$ (this mean estimator thus depends on the confidence level parameter ε). Since we have

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

the exponential Chebyshev's inequality (see Lemma 4.1) guarantees that with probability at least $1-\varepsilon$, we have $n\lambda(\hat{\theta}-\mathbb{E}(Y)) \leq n\frac{\lambda^2}{2}+\log(\varepsilon^{-1})$, hence

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

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

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

Since $-\log(1+x+x^2/2) \leq \log(1-x+x^2/2)$, this implies

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

The two previous inequalities imply Inequality (4.1) (for $c=2$), showing that sub-exponential tails are achievable even when we only assume that the random variable admits a finite second moment (see [12] for more details on the robust estimation of the mean of a random variable).

4.1.2. Localized PAC-Bayesian inequalities to eliminate a logarithm factor. The analysis of statistical inference generally relies on upper bounding the supremum of an empirical process χ indexed by the functions in a model $\mathcal{F}$. One central tool to obtain these bounds are the concentration inequalities. An alternative approach, called the PAC-Bayesian one, consists in using the entropic equality

$$\mathbb{E}\exp\left(\sup_{\rho\in\mathcal{M}}\left\{\int\rho(df)\chi(f)-K(\rho,\pi')\right\}\right)=\int\pi'(df)\mathbb{E}\exp\big(\chi(f)\big). \quad (4.2)$$

where $\mathcal{M}$ is the set of probability distributions on $\mathcal{F}$ and $K(\rho,\pi')$ is the Kullback-Leibler divergence (whose definition is recalled in (4.4, page 29)) between ρ and some fixed distribution π'.

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

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

for $\chi(f)=\lambda[R(f)-\check{r}(f)]$ and some $\lambda>0$. Then, as a consequence of (4.2), for any $\varepsilon>0$, with probability at least $1-\varepsilon$, for any distribution ρ on $\mathcal{F}$,

$$\int\rho(df)R(f)\leq\int\rho(df)\check{r}(f)+\frac{K(\rho,\pi')+\log(\varepsilon^{-1})}{\lambda}. \quad (4.3)$$

The left-hand side quantity represents the expected risk with respect to the distribution ρ. The question is now how to use (4.3) to design a posterior distribution ρ for which $\int \rho(df)R(f)$ is guaranteed to be small. The constraint on the choice of (ρ, π') is that ρ should be computable from the data (e.g., it cannot depend on R) and π' should not depend on the data: it may depend on R (in contrast with Bayesian prior distributions!) but not on $\check{r}$. Simple choices like $(\rho, \pi') = (\delta_{f^*}, \delta_{f^*})$ or $(\rho, \pi') = (\delta_{\check{f}}, \delta_{\check{f}})$ for $\check{f} \in \operatorname{argmin}_{f\in\mathcal{F}} \check{r}(f)$, where δ_a denotes the Dirac distribution at the function f, are thus forbidden (while they would have led to small right-hand side of (4.3)).

For fixed π', the posterior distribution minimizing the right-hand side of (4.3) is $\rho = \pi'_{-\lambda\check{r}}$. It is computable from the data if π' is. Without prior knowledge, this would lead to take a "flat" distribution for π' (e.g., the one induced by the Lebesgue measure in the case of a model $\mathcal{F}$ defined by a bounded parameter set in some Euclidean space). The resulting Kullback-Leibler divergence might be very large as it compares a distribution with a sharp peak (concentrated on functions $f \in \mathcal{F}$ for which $\check{r}(f)$) with a flat one.

To get a smaller Kullback-Leibler divergence, we can take posterior and prior distributions which are peaked around almost the same function. This can be done by taking π and ρ respectively concentrated around f^* and $\check{f}$. More precisely, one can take posterior distributions of the form $\rho = \pi_{-\lambda\check{r}}$ for some $\lambda > 0$ and a "flat" distribution π computable without knowing neither the distribution P generating the data nor the training data (in particular, π must not depend on R or $\check{r}$), and a "localized" prior distribution $\pi' = \pi_{-\beta R}$ for some $\beta > 0$. The parameters λ and β controlling the sharpness of the peaks at $\operatorname{argmin}_{f\in\mathcal{F}} R(f)*$ and $\operatorname{argmin}_{f\in\mathcal{F}} \check{r}(f)$ should be taken such that the peaks overlap (to ensure that the Kullback-Leibler divergence is small) and are in the same time sharp enough (to ensure that $\int \rho(df)\check{r}(f)$ is small). The use of the "localized" prior distribution $\pi' = \pi_{-\beta R}$ implies an additional technical difficulty as one needs to control the divergence $K(\rho, \pi_{-\beta R})$. This is achieved by writing

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

and controlling the new logarithmic term through PAC-Bayesian inequalities.

4.2. PROOF OF THEOREM 3.1. We use the standard way of obtaining PAC bounds through upper bounds on Laplace transforms of appropriate random variables. This argument is synthesized in the following result.

LEMMA 4.1 *For any $\varepsilon > 0$ and any real-valued random variable V such that $\mathbb{E}\big[\exp(V)\big] \leq 1$, with probability at least $1 - \varepsilon$, we have*

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

Let $V_1(\hat{f}) = \int \big[L^\flat(\hat{f}, f) + \gamma^* \bar{R}(f)\big] \pi^*_{-\gamma^* \bar{R}}(df) - \gamma \bar{R}(\hat{f})$

$$- \mathcal{I}^*(\gamma^*) + \mathcal{I}(\gamma) + \log\bigg(\int \exp\big[-\hat{\mathcal{E}}(f)\big] \pi(df) \bigg) - \log\bigg[\frac{d\rho}{d\hat{\pi}}(\hat{f}) \bigg],$$

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 4.1, it suffices to prove that

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

These two inequalities are proved in the following two sections.

4.2.1. 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

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

From Jensen's inequality again,

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

From the two previous inequalities, we get

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

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

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

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

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

LEMMA 4.2 *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* $\mathbb{E}_{a_1\sim\mu_1}\Bigl\{\log\Bigl[\mathbb{E}_{a_2\sim\mu_2}\bigl\{\exp\bigl[-\mathcal{W}(a_1,a_2)\bigr]\bigr\}\Bigr]\Bigr\} < +\infty$, *then we have*

$$
\begin{aligned}
&-\mathbb{E}_{a_1\sim\mu_1}\Bigl\{\log\Bigl[\mathbb{E}_{a_2\sim\mu_2}\bigl\{\exp\bigl[-\mathcal{W}(a_1,a_2)\bigr]\bigr\}\Bigr]\Bigr\}\\
&\qquad\qquad \leq -\log\Bigl\{\mathbb{E}_{a_2\sim\mu_2}\Bigl[\exp\bigl[-\mathbb{E}_{a_1\sim\mu_1}\,\mathcal{W}(a_1,a_2)\bigr]\Bigr]\Bigr\}.
\end{aligned}
$$

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

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

PROOF.

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

$$K(\rho,\mu) \triangleq \begin{cases} \mathbb{E}_{a\sim\rho}\ \log\left[\dfrac{d\rho}{d\mu}(a)\right] & \text{if } \rho \ll \mu, \\ +\infty & \text{otherwise,} \end{cases} \tag{4.4}$$

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

$$\inf_{\rho\in\mathcal{M}} \left\{\mathbb{E}_{a\sim\rho}\, h(a) + K(\rho,\mu)\right\} = -\log \mathbb{E}_{a\sim\mu}\left\{\exp\bigl[-h(a)\bigr]\right\}. \tag{4.5}$$

  By using twice (4.5) and Fubini's theorem, we have

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

- By using twice (4.5) and the first assertion of Lemma 4.2, we have

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

From Lemma 4.2 and Fubini's theorem, since V_2 does not depend on $\hat{f}$, we have

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

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 $Z_1, \dots, Z_n$ and the randomized prediction function $\hat{f}$) at least $1 - 2\varepsilon$:

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

4.3. PROOF OF LEMMA 3.3. 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

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

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

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

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

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

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

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

Introduce

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

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

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

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

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

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

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

which proves the announced result.

4.4. PROOF OF LEMMA 3.4. 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

$$\big|\mathbb{E}(\Omega|X)\big| \leq \tilde{A}. \tag{4.6}$$

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

$$\log\Big\{\mathbb{E}\Big[\exp\big[L - \mathbb{E}(L|X)\big] \,|\, X\Big]\Big\} \leq \lambda^2(F - F^*)^2 a(\lambda). \tag{4.7}$$

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 $\tilde{L} \leq |\lambda|H\tilde{A} + a(\lambda)\lambda^2H^2 \leq b'$ 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, (4.7) and (4.6), we obtain

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

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

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

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

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

Now it remains to notice that $\mathbb{E}\big[(F-F^*)^2\big] \le 2[R(f)-R(f^*)]/b_1$. 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 $\phi \ge 0$ on $[0;1]$, implying that $\phi'(0) \ge 0$. Besides $\phi(1) = \phi(0) + \phi'(0) + \int_0^1 (1-t)\phi''(t)dt$, where $\phi''(t)$ is defined as

$$\begin{aligned}\phi''(t) &= \mathbb{E}\Big\{\big[f(X)-f^*(X)\big]^2 \tilde{\ell}''_Y\big[[(1-t)f^*+f](X)\big]\Big\}\\ &\ge b_1 \mathbb{E}\big\{\big[f(X)-f^*(X)\big]^2\big\},\end{aligned}$$

implying that

$$\frac{b_1}{2}\mathbb{E}(F-F^*)^2 \le R(f)-R(f^*). \tag{4.8}$$

4.5. PROOF OF LEMMA 3.6. We have

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

where the last inequality is the usual relation between excess risk and L^2 distance using the convexity of $\mathcal{F}$ (see above (4.8) for a proof).

4.6. PROOF OF LEMMA 3.7. Let $\mathcal{S} = \{s \in \mathcal{F}_{\text{lin}} : \mathbb{E}[s(X)^2] = 1\}$. Using the triangular inequality in $\mathbb{L}^2$, we get

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

$$+ \mathbb{E}\big([f^*(X) - f(X)]^2\big) \sqrt{\sup_{s \in \mathcal{S}} \mathbb{E}\big[s(X)^4\big]} \Bigg]^2$$
$$\leq V[R(f) - R(f^*)],$$

with

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

where the last inequality is the usual relation between excess risk and L^2 distance using the convexity of $\mathcal{F}$ (see above (4.8) for a proof).

A. Uniformly bounded conditional variance is necessary to reach d/n rate

In this section, we show that the target (0.3) cannot be reached if we just assume that Y has a finite variance and that the functions in $\mathcal{F}$ are bounded. For this purpose, the following result gives a $1/\sqrt{n}$ lower bound when $d = 2$. (Note that it is not implied by the $\sqrt{\log(1 + d/\sqrt{n})/n}$ lower bound for convex aggregation, proved in [25], and in slightly weaker forms in [18, 27], since the latter bound is shown for $d \geq \sqrt{n}$.)

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 $\varphi_1(x) = \mathbb{1}_{x \in \mathcal{X}_1}$ and $\varphi_2(x) = \mathbb{1}_{x \in \mathcal{X}_2}$. Let $\mathcal{F} = \big\{\theta_1 \varphi_1 + \theta_2 \varphi_2; (\theta_1, \theta_2) \in [-1, 1]^2\big\}$.

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

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

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

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

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

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 [21, Lemma 1]

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

We apply a slightly improved version of Theorem 5 in Birgé and Massart [7]. First for homogeneity purpose, we modify Assumption M2 by replacing the condition "$\sigma^2 \ge D/n$" by "$\sigma^2 \ge B^2D/n$" where the constant B is the one appearing in (5.3) of [7]. This modifies Theorem 5 of [7] 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 $m \ge 2$

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

We consider $W = Y - f^*(X)$, $\gamma(z,f) = (y-f(x))^2$, $\Delta(x,u,v) = |u(x)-v(x)|$, and $M(w) = 2(|w| + H)$. From (1.7), for all $m \ge 2$, we have $\mathbb{E}\big\{[(2(|W| + H)]^m|X = x\big] \le \frac{m!}{2}[4M(A+H)]^m$. Now consider B' and r such that Assumption

M2 of [7] holds for $D = d$. Inequality (5.8) for $\tau = 1/2$ of [7] 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 $1 - \kappa \exp\Big[\frac{-nv}{\kappa(A^2+H^2)}\Big]$,

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

for some large enough constant κ depending on M. Now from Proposition 1 of [7] 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 $\mathbb{E}\big\{\big[\hat{f}^{(\text{erm})}(X) - f^*(X)\big]^2\big\} \leq R(\hat{f}^{(\text{erm})}) - R(f^*)$ (since $\mathcal{F}$ is convex and f^* is the orthogonal projection of Y on $\mathcal{F}$), and $r(f^*) - r(\hat{f}^{(\text{erm})}) \geq 0$ (by definition of $\hat{f}^{(\text{erm})}$), 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 [7] allows to bracket $\tilde{B}$ in terms of $B = \sup_{f \in \text{span}\{\varphi_1,\ldots,\varphi_d\}} \|f\|_\infty^2/\mathbb{E}[f(X)]^2$, 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 $\varphi_j = \mathbb{1}_{A_j}, j = 1, \ldots, d$: 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: $f^*_{\text{lin}} = f^{(\text{reg})} = \sum_{j=1}^d \theta\varphi_j$. According to Theorem 1.1, if we know an upper bound H on $\|f^{(\text{reg})}\|_\infty = \theta$, we have that the truncated estimator $(\hat{f}^{(\text{ols})} \wedge H) \vee -H$ satisfies

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

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 $Q = \big(\mathbb{E}\varphi_j(X)\varphi_k(X)\big)_{j,k} = \text{Diag}(p_j)$, $\mathcal{K} = 1$ 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 $\lambda \leq p_{\min}$, the ridge regression estimator satisfies with probability at least $1 - \varepsilon$:

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

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 $p_{\min} \leq 1/d$, 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 $1-\varepsilon$,

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

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 $p_{\min} \geq 1/n$, the target (0.2) 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 [8], 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 $x \in \mathcal{X}$,*

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

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

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

for some positive constant κ depending only on M.

PROOF. One can check that $\hat{f}^{(\text{ridge})} \in \operatorname{argmin}_{f\in\mathcal{H}} r(f) + \lambda\sum_{j=1}^d \|f\|^2_{\mathcal{H}}$, 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 $f^{(\lambda)} \in \operatorname{argmin}_{f\in\mathcal{H}} R(f) + \lambda\sum_{j=1}^d \|f\|^2_{\mathcal{H}}$. Let us use Theorem 4 in [8] 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 [8] are $\mathcal{A}(\lambda)$ and the term with the effective dimension $\mathcal{N}(\lambda)$. $\square$

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 36. 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 36 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

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

$\square$

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

$$\begin{aligned}\log\Big\{\mathbb{E}\Big[\exp\big\{\alpha\big[V - \mathbb{E}(V)\big]\big\}\Big]\Big\} &\leq \mathbb{E}\Big[\exp\big\{\alpha\big[V - \mathbb{E}(V)\big]\big\}\Big]\Big\} - 1 \\ &= \mathbb{E}\Big\{\exp\big\{\alpha\big[V - \mathbb{E}(V)\big]\big\} - 1 - \alpha\big[V - \mathbb{E}(V)\big]\Big\} \\ &\leq \mathbb{E}\Big\{\exp\big[|\alpha||V - \mathbb{E}(V)|\big] - 1 - |\alpha||V - \mathbb{E}(V)|\Big\} \\ &\leq \mathbb{E}\Big\{\exp\big[|V - \mathbb{E}(V)|\big]\Big\} \sup_{u \geq 0}\Big\{\big[\exp(|\alpha|u) - 1 - |\alpha|u\big]\exp(-u)\Big\}\end{aligned}$$

$$\leq \mathbb{E}\Big[\exp\big(|V| + |\mathbb{E}(V)|\big)\Big] \sup_{u\geq 0} \sum_{m\geq 2} \frac{|\alpha|^m u^m}{m!} \exp(-u)$$

$$\leq M^2 \sum_{m\geq 2} \frac{|\alpha|^m}{m!} \sup_{u\geq 0} u^m \exp(-u) = \alpha^2 M^2 \sum_{m\geq 2} \frac{|\alpha|^{m-2}}{m!} m^m \exp(-m)$$

$$\leq \alpha^2 M^2 \sum_{m\geq 2} \frac{|\alpha|^{m-2}}{\sqrt{2\pi m}} \leq \frac{\alpha^2 M^2}{2\sqrt{\pi}(1-|\alpha|)}.$$

□

REFERENCES

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

[2] P. Alquier and K. Lounici. PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electronic Journal of Statistics*, 5:127–145, 2011.

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

[4] J.-Y. Audibert and O. Catoni. Robust linear least squares regression, 2010. arXiv.

[5] Y. Baraud. Model selection for regression on a fixed design. *Probability Theory and Related Fields*, 117(4):467–493, 2000.

[6] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

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

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

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

[10] 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.

[11] 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.

[12] O. Catoni. Challenging the empirical mean and empirical variance: a deviation study, 2010. arXiv:1009.2048v1.

[13] A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Machine Learning*, 72(1-2):39–61, 2008.

[14] A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *technical report*, page arXiv:0903.1223v3, 2010.

[15] A. S. Dalalyan and A. B. Tsybakov. Mirror averaging with sparsity priors. *Bernoulli*, to appear(arXiv:1003.1189v2 [math.ST]), 2011.

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

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

[18] A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric estimation. *Ann. Stat.*, 28:681–712, 2000.

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

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

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

[22] A. Nemirovski. *Lectures on probability theory and statistics. Topics in nonparametric statistics. Ecole d'Eté de Probabilités de Saint-Flour XXVIII-1998*. Springer-Verlag, 2000.

[23] 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.

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

[25] 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, 2003.

[26] M. Wegkamp. Model selection in nonparametric regression. *Annals of Statistics*, 31(1):252–273, 2003.

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