arXiv: arXiv:1702.01402

SUPPLEMENTARY MATERIAL TO "ESTIMATION BOUNDS AND SHARP ORACLE INEQUALITIES OF REGULARIZED PROCEDURES WITH LIPSCHITZ LOSS FUNCTIONS"

By Pierre Alquier[*,†] and Vincent Cottet[*] and Guillaume Lecué[*]

CREST, CNRS, ENSAE, Université Paris Saclay.
Email: pierre.alquier, vincent.cottet,
guillaume.lecue@ensae.fr

The supplementary material is organized as follows.

- Section 6 proposes an alternating direction method of multiplier algorithm for the matrix completion methods proposed in Section 4. Our estimator is evaluated on simulated and real datasets. Python notebooks can be downloaded: *https://sites.google.com/site/vincentcottet/code*
- Section 7 contains an application of our general results to the case where E is a reproducing kernel Hilbert space (RKHS).
- Section 8 contains a complete discussion of the Bernstein condition. Many examples and sufficient conditions (old and new) are provided.
- Section 9 contains the proofs of the theorems of Section 2.
- Section 10 contains the proofs of the optimality results of Section 4.
- Section 11 contains the proofs of the new results stated in Section 8.
- Finally, Section 12 contains the study of the (non-penalized) ERM. We also provide an application to shape-constrained logistic regression.

6. Simulation study in matrix completion.

6.1. *Algorithm and Simulation Outlines.* Since this part provides new methods and results on matrix completion, we propose an algorithm in order to compute efficiently the RERM using the hinge loss and the quantile loss. This section explains the structure of the algorithm that is used with specific loss functions in next sections. Although many algorithms exist for the least squares matrix completion, to our knowledge many of them treat only the exact recovery such as in [10] and [30], or at least they all deal with

[*] Supported by Labex ECODEC (ANR - 11-LABEX-0047)

[†] Research programme *New Challenges for New Data* from LCL and GENES, hosted by the *Fondation du Risque*

differentiable loss functions, see [22]. On the other hand, the two losses that we mainly consider here are non differentiable because they are piecewise linear (in the case of hinge and $0,5$-quantile loss functions): new algorithms are hence needed. It has been often noted that the RERM with respect to the hinge loss or 0.5-quantile loss can be solved by a semidefinite programming but the cost is prohibitive for large matrices, say dimensions larger than 100. It actually works for small matrices as we ran SDP solver in Python in very small examples.

We propose here an *alternating direction method of multiplier* (ADMM) algorithm. For a clear and self-contained introduction to this class of algorithms, the reader is referred to the very pedagogical introduction [9], and we do not explain all the details here. When the optimization problem is a sum of two parts, the core idea is to split the problem by introducing an extra variable. In our case, the two following problems are equivalent:

$$\underset{M}{\text{minimize}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \ell(\big\langle X_i, M \big\rangle, Y_i) + \lambda \left\| M \right\|_{S_1} \right\},$$

and

$$\underset{M,L}{\text{minimize}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \ell(\big\langle X_i, M \big\rangle, Y_i) + \lambda \left\| L \right\|_{S_1} \right\} \text{subject to } M = L$$

Below, we use the scaled form and the $m \times T$ matrix U is then called the *scaled dual variable*. Note that the S_2 norm is also the Froebenius norm and is thus elementwise. We can now exhibit the *augmented Lagrangian*:

$$L_\alpha(M, L, U) = \frac{1}{N} \sum_{i=1}^{N} \ell(\big\langle X_i, M \big\rangle, Y_i) + \lambda \left\| L \right\|_{S_1} + \frac{\alpha}{2} \left\| M - L + U \right\|_{S_2}^2 - \frac{\alpha}{2} \left\| U \right\|_{S_2}^2,$$

where α is a positive constant, called the *augmented Lagrange parameter*. The ADMM algorithm [9] is then:

$$M^{k+1} = \underset{M}{\text{argmin}} \left(\frac{1}{N} \sum_{i=1}^{N} \ell(\big\langle X_i, M \big\rangle, Y_i) + \frac{\alpha}{2} \left\| M - L^k + U^k \right\|_{S_2}^2 \right) \tag{24}$$

$$L^{k+1} = \underset{L}{\text{argmin}} \left(\lambda \left\| L \right\|_{S_1} + \frac{\alpha}{2} \left\| M^{k+1} - L + U^k \right\|_{S_2}^2 \right) \tag{25}$$

$$U^{k+1} = U^k + M^{k+1} - L^{k+1}$$

The starting point (M^0, L^0, U^0) uses one random matrix with independent Gaussian entries for M^0 and two zero matrices for L^0 and U^0. Another choice of starting point is to use a previous estimator with a larger λ. The stopping criterion is, as explained in [9], $\left\|M^{k+1}-M^{k}\right\|_{S_2}^2+\left\|U^{k+1}-U^{k}\right\|_{S_2}^2 \leq \varepsilon$ for a fixed threshold ε. It means that it stops when both (U_k) and (M_k) start converging.

General considerations. The second step (25) is independent of the loss function. It is well-known that the solution to this problem is $S_{\lambda/\alpha}(M^{k+1}+U^k)$ when $S_a(M)$ is the soft-thresholding operator with magnitude a applied to the singular values of the matrix M. It is defined for a rank r matrix M with SVD $M=U\Sigma V^\top$ where $\Sigma=\operatorname{diag}\left((d_i)_{1\leq i\leq r}\right)$ by $S_a(M)=US_a(\Sigma)V^\top$ where $S_a(\Sigma)=\operatorname{diag}\left((\max(0,d_i-a))_{1\leq i\leq r}\right)$.

It requires the computation of the SVD of a $m\times T$ matrix at each iteration. This is the main bottleneck of this algorithm (the other main step (24) can be performed elementwise since the X_i's take their values in the canonical basis of $\mathbb{R}^{m\times T}$; so it needs only at most N operations). Two methods may be used to speed up the algorithm. First, there are efficient algorithms for computing the n largest singular values and the associate subspaces, such as the Fortran PROPACK. It can be plugged in order to solve (25) by computing the n largest and stop at this stage if the lowest computed singular values is lower than the threshold. It is obviously more relevant when the target is expected to have a very small rank. This method has been implemented in Python and works well in practice even though the parameter n has to be tuned carefully. Second, one can use an approximate SVD such as in [21].

Moreover, the first step (24) (which may be performed elementwise) has a closed form solution for hinge and quantile loss: it is a soft-thresholding applied to a specified quantity.

Simulated observations as well as real-world data (cf. the MovieLens dataset[1]) are considered in the examples below. Finally note that parameter λ is tuned by cross-validation.

6.2. *Simulation study for RERM "Hinge + $S1$".* As the hinge loss has not been often studied in the matrix context, we provide many simulations in order to show the robustness of our method and the opportunity of using the hinge loss rather than the logistic loss. We follow the simulations ran in [14] and compare several methods. An estimator based on the logistic model, studied in [16], is also challenged[2].

[1] available in http://grouplens.org/datasets/movielens/

[2] In the followings, the four estimators will be referred to *Hinge* for estimator given in (18), *Hinge Bayes* and *Logit Bayes* for the two Bayesian estimators from [14] with respectively hinge and logistic loss functions, and *Logit* for the estimator from [16]. The

A first set of simulations. The simulations are all based on a low-rank 200×200 matrix $M^\star$ from which the data are generated and which is the target for the predictions. $M^\star$ is also a minimizer of $R_{0/1}$ so the error criterion that we will report for a matrix M is the difference of the predictions between $M^\star$ and M, which is $\mathbb{P}[\mathrm{sign}(\langle M^\star, X\rangle) \neq \mathrm{sign}(\langle M, X\rangle)]$. The X_i's correspond to 20% of the entries randomly picked so the misclassification rate is also $(1/mT)\sum_{p,q} I\{\mathrm{sign}(M_{p,q}) \neq \mathrm{sign}(M^\star_{p,q})\}$.

Two different scenarios are tested: the first one (called A), involves a matrix $M^\star$ with only entries in $\{-1, +1\}$ so the Bayes classifier is low rank and favors the hinge loss. The second test (called B) involves a matrix $M^\star = LR^\top$ where L, R have i.i.d. Gaussian entries and the rank is the number of columns. In this case, the Bayes matrix contains the signs of a low-rank matrix, but it is not itself low rank in general. We also test the impact of several noise structures on the performance of the estimators:

1. (noiseless) $Y_i = \mathrm{sign}(\langle M^\star, X_i\rangle)$
2. (logistic) $Y_i = \mathrm{sign}(\langle M^\star, X_i\rangle + Z_i)$, where Z_i follows a logistic distribution
3. (switch) $Y_i = \epsilon_i \mathrm{sign}(\langle M^\star, X_i\rangle)$ where $\epsilon_i = (1-p)\delta_1 + p\delta_{-1}$

Finally, we run all the simulations on rank 3 and rank 5 matrices and λ is tuned by cross validation. All the simulations are run one time.

	Model	A1	A2 ($p = .1$)	A3	B1	B2 ($p = .1$)	B3
Rank 3	Hinge	0	0	14.5	6.7	10.9	21.0
	Logit	0	0.5	17.3	5.1	10.7	19.8
	Hinge Bayes	0	0.1	8.5	5.3	10.8	22.1
	Logit Bayes	0	0.5	16.0	4.1	10.1	16.0
Rank 5	Hinge	0	0.8	29.0	11.7	19.3	23.3
	Logit	0	3.1	30.1	9.0	18.3	22.1
	Hinge Bayes	0	0.5	27	9.4	17.9	24.4
	Logit Bayes	0	4.4	32.5	7.8	17.3	21.5

TABLE 1

Misclassification error rates on simulated matrices in various cases. Model $\in \{A, B\}\{1, 2, 3\}$ *refers to scenario* $\in \{A, B\}$ *and noise structure* $\in \{1, 2, 3\}$. *For the noise-free Model* $=$ *A0, the* 0 *column shows the exact reconstruction property of all procedures.*

The results are very similar among the methods, see Table 1. The logistic loss performs better for matrices of type B and especially for high level of noise in the logistic data generation as expected. For type A matrices, the hinge loss performs slightly better. The Bayesian estimators performs as well as the frequentist estimators even though the program solved is not convex.

Bayesian estimators use the Gammma prior distribution.

Impact of the noise level. The second experiment is a focus on the switch noise and matrices that are well separated (as A2 in the previous example). The noise lies between $p = 0$ and almost full noise ($p = .4$). The performance of the RERM with the hinge loss is slightly worse than the Bayesian estimator with hinge loss but always better than the RERM with the logistic loss, see Figure 6.2.

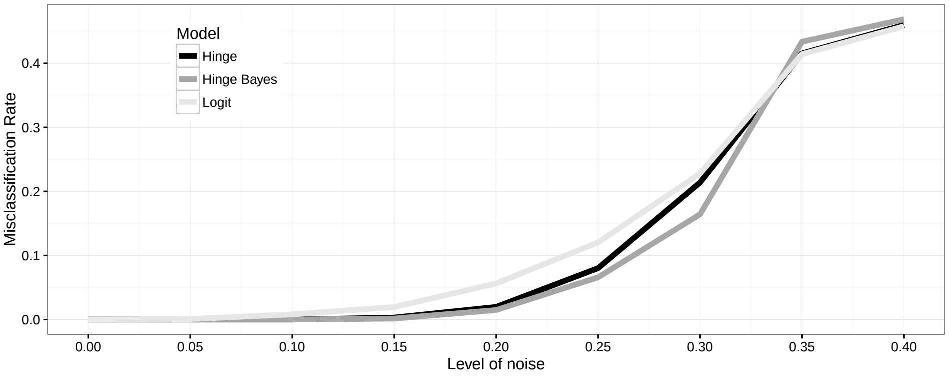

Fig 1. *Misclassification error rates for a large range of switch noise (noise structure number 3).*

Real dataset. We finally run the hinge loss estimator on the MovieLens dataset. The ratings, that lie in $\{1, 2, 3, 4, 5\}$, are split between good ratings $(4, 5)$ and bad ratings (others). The goal is therefore to predict whether the user will like a movie or not. On a test set that contains 20% of the data, the misclassification rate in prediction are almost the same for all the methods (Table 2).

Model	Hinge Bayes	Logit	Hinge
misclassification rate	.28	.27	.28

Table 2
Misclassification Rate on MovieLens 100K dataset

6.3. *Simulation study for quantile matrix completion.* The goal of this section is to challenge the regularized least squares estimator by the RERM with 0.5-quantile loss. The quantile used here is therefore the median. The main conclusion of our study is that median based estimators are more robust to outliers and noise than mean based estimators. We first test them on simulated datasets and then turn to use a real dataset.

Simulated matrices. The observations come from a base matrix $M^\star$ which is a 200×200 low rank matrix. It is built by $M^\star = LR^\top$ where the entries of L, R are i.i.d. gaussian and L, R have 3 columns (and therefore, the rank of $M^\star$ is 3). The X_i's correspond to 20% randomly picked entries. The criterion that we retain is the l_1 reconstruction of $M^\star$ that is: $(1/mT)\sum_{p,q}|M^\star_{p,q} - M_{p,q}|$.

The observations are made according to this flexible model:

$$Y_i = \langle M^\star, X\rangle + z_i + o\zeta_i.$$

z_i is the noise, o is the magnitude of outliers and ζ_i is the outlier indicator parametrized by the share p such that $\zeta_i = p/2\delta_{-1} + (1-p)\delta_0 + p/2\delta_1$. The different parameters for the different scenarios are summarized in Table 3.

On the first experiment, p is fixed to 10% and the magnitude o increases. As expected for least squares, the results are better for low magnitude of outliers (it corresponds to the penalized maximum likelihood estimator), see Figure 2. Quickly, the performance of the least squares estimator is getting worse and when the outliers are large enough, the best least squares predictor is a matrix with null entries. In opposite to this estimator, the median of the distribution is almost not affected by outliers and it is completely in line with the results: the performances are strictly the same for mid-range to high-range magnitude of outliers. The robustness of the quantile reconstruction is totally independent to the magnitude of the outliers.

	z_i	o	ζ_i
Figure 2	$\mathcal{N}(0, 1/4)$	$o = 0..30$	$p = 0.1$
Figure 3	$\mathcal{N}(0, 1/4)$	10	$p = 0.025$
Figure 4	$t_\alpha, \alpha = 1..10$	0	$p = 0$

t_α: t-distribution with α degrees of freedom.

TABLE 3

Parameters and distributions of the simulations

In a second experiment, we fixed the magnitude of outliers but we increase their proportion within the dataset (see Figure 3). The median completion is, as expected, more robust and the results deteriorate less than the ones from least squares. When the outliers ratio is greater than 20%, the least squares estimator completely fails while the median completion still works.

The third simulation involves non gaussian noise without outliers: we use the t-distribution, that has heavy tails. In this challenge, a lower degree of freedom involves heavier tails and the worst case is for Student distribution with degree 1. We can see that the least squares is inadequate for small degrees of freedom (1 to 2) and behaves better than the median completion for larger degrees of freedom, see Figure 4.

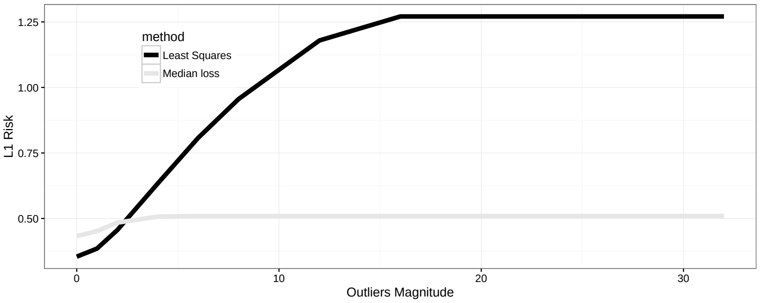

Fig 2. *l_1 reconstruction for different magnitude of outliers*

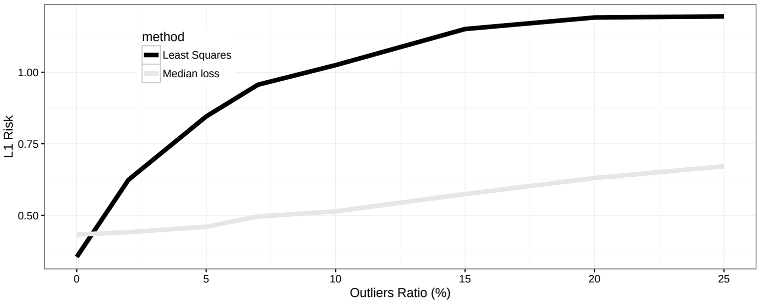

Fig 3. *l_1 reconstruction for different percentage of outliers*

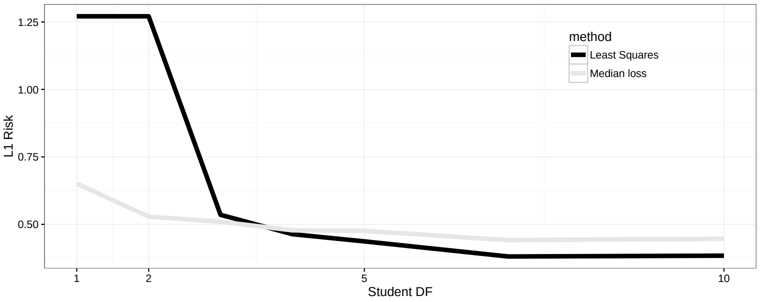

Fig 4. *l_1 reconstruction for student noise with various magnitude degrees of freedom*

Real dataset. The last experiment involves the MovieLens dataset. We keep one fifth of the sample for test set to check the prediction accuracy. Even though the least squares estimator remains very efficient in the standard case, see Table 4, the results are quite similar for the MAE criterion. In a second step, we add artificial outliers. In order to do that, we change 20% of 5 ratings to 1 ratings. It can be seen as malicious users that change ratings in order to distort the perception of some movies. As expected, it depreciates

	MSE	MAE
Raw Data, LS	0.89	0.75
Raw Data, Median	0.93	0.75
Outliers, LS	1.04	0.84
Outliers, Median	0.96	0.78

TABLE 4

Prediction power of Least Squares and Median Loss on MovieLens 100K dataset

the least squares estimator performance but the median estimator returns almost as good performances as in the standard case.

7. Kernel methods via the hinge loss and a RKHS-norm regularization. In this section, we consider regularization methods in some general Reproducing Kernel Hilbert Space (RKHS) (cf. [15], Chapter 4 in [38] or Chapter 3 of [44] for general references on RKHS). Note that RKHS is a vast class of functions space which include non-parametric spaces such as Sobolev spaces (see [38]).

Unlike the previous examples, the regularization norm here, which is the norm $\|\cdot\|_{\mathcal{H}_K}$ of a RKHS $\mathcal{H}_K$, is not associated with some "hidden" concept of sparsity. In particular, RKHS norms have no singularity (except at 0) since they are differentiable at any point except in 0. As a consequence the sparsity parameter $\Delta(\rho)$ cannot be larger than $4\rho/5$, i.e. ρ does not satisfy the sparsity equation, unless the set $\Gamma_{f^*}(\rho)$ contains 0 that is for $\rho \geq 20\,\|f^*\|_{\mathcal{H}_K}$. Indeed, one key observation is that norms are non differentiable at 0 and that its subdifferential at 0 is somehow extremal:

$$\partial\left\|\cdot\right\|(0) = B_* := \{f : \|f\|_* \leq 1\}, \tag{26}$$

where $\|\cdot\|_*$ is the dual norm.

As a consequence, the rates obtained in this section do not depend on some *hidden sparsity parameter* associated with the oracle f^* but on the RKHS norm at f^*, that is $\|f^*\|_{\mathcal{H}_K}$ (such error rates are refer as “complexity bounds” in [28]). The aim of this section is therefore to show that our main results apply beyond “sparsity inducing regularization methods” by showing that “classic” regularization method, inducing smoothness for instance, may also be analyzed the same way and fall into the scope of Theorem 2.1 and Theorem 2.2. This section also shows an explicit expression for the Gaussian mean-width with localization as used in Definition 9.1 (a sharper way to measure statistical complexity via a local $r(\cdot)$ function provided below).

Mathematical background. In this setup, the data are still N i.i.d. pairs $(X_i, Y_i)_{i=1}^N$ where the X_i's take their values in some set $\mathcal{X}$ and $Y_i \in \{-1, +1\}$.

A "similarity measure" is provided over the set $\mathcal{X}$ by means of a kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ so that $x_1, x_2 \in \mathcal{X}$ are "similar" when $K(x_1, x_2)$ is small. One can think, for instance, of $\mathcal{X}$ as the set of all DNA sequences (that is finite words over the alphabet $\{A, T, C, G\}$) and $K(w_1, w_2)$ is the minimal number of changes like insertion, deletion and mutation needed to transform word $w_1 \in \mathcal{X}$ into word $w_2 \in \mathcal{X}$.

The core idea behind kernel methods is to transport the design data X_i's from $\mathcal{X}$ to a Hilbert space via the application $x \to K(x, \cdot)$ and then construct statistical procedures based on the "transported" dataset $(K(X_i, \cdot), Y_i)_{i=1}^N$. The advantage of doing so is that the space where the $K(X_i, \cdot)$'s belong have much structure than the initial set $\mathcal{X}$ which may have no algebraic structure at all. The first thing to set is to define somehow the "smallest" Hilbert space containing all the functions $x \to K(x, \cdot)$. We recall now one classic way of doing so that will be used later to define the objects that need to be considered in order to construct RERM in this setup and to derive their estimation rates via Theorem 2.1 and Theorem 2.2. Note that even though, we derive estimation rates only in the bounded case (because the subgaussian assumption is not natural for RKHS), we provide a computation of the two complexity parameters since their analysis is identical and they yield an example where the two Gaussian width and Rademacher complexities are equal.

Recall that if $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel such that $\|K\|_{L_2} < \infty$, then by Mercer's theorem, there is an orthogonal basis $(\phi_i)_{i\in\mathbb{N}}$ of L_2 such that $\mu \otimes \mu$-almost surely, $K(x, x') = \sum_{i=1}^\infty \lambda_i \phi_i(x)\phi_i(x')$ where $(\lambda_i)_{i\in\mathbb{N}}$ is the sequence of eigenvalues of the positive self-adjoint integral operator T_K (arranged in a non-increasing order) defined for every $f \in L_2$ and μ-almost every $x \in \mathcal{X}$ by

$$(T_K f)(x) = \int_{x'\in\mathcal{X}} K(x, x') f(x') d\mu(x').$$

In particular, for all $i \in \mathbb{N}$, ϕ_i is an eigenvector of T_K corresponding to the eigenvalue λ_i; and $(\phi_i)_i$ is an orthonormal system in $L_2(\mu)$.

The reproducing kernel Hilbert space $\mathcal{H}_K$ is the set of all function series $\sum_{i=1}^\infty a_i K(x_i, \cdot)$ converging in L_2 endowed with the inner product

$$\left\langle \sum a_i K(x_i, \cdot), \sum b_j K(x'_j, \cdot) \right\rangle = \sum_{i,j} a_i b_j K(x_i, x'_j)$$

where a_i, b_j's are any real numbers and the x_i's and x'_j's are any points in $\mathcal{X}$.

Estimator. The RKHS $\mathcal{H}_K$ is therefore a class of functions from $\mathcal{X}$ to $\mathbb{R}$ that can be used as a learning model and the norm naturally associated to its Hilbert structure can be used as a regularization function. Given a Lipschitz loss function ℓ, the oracle is defined as

$$f^* \in \underset{f\in\mathcal{H}_K}{\operatorname{argmin}}\, \mathbb{E}\ell_f(X,Y)$$

and it is believed that $\|f^*\|_{\mathcal{H}_K}$ is small which justified the use of the RERM with regularization function given by the RKHS norm $\|\cdot\|_{\mathcal{H}_K}$:

$$\hat{f} \in \underset{f\in\mathcal{H}_K}{\operatorname{argmin}} \left(\frac{1}{N}\sum_{i=1}^{N} \ell_f(X_i,Y_i) + \lambda \|f\|_{\mathcal{H}_K} \right)$$

Statistical properties of this RERM may be obtained from Theorem 2.1 in the subgaussian case and from Theorem 2.2 in the bounded case. To that end, we only have to compute the Gaussian mean width and/or the Rademacher complexities of $B_{\mathcal{H}_K}$. In this example, we rather compute the localized version of those quantities because it is possible to derive explicit formula. They are obtained by intersecting the ball with $r\mathcal{E}$. In order not to induce any confusion, we still use the global ones in estimation bounds.

Localized complexity parameter. The goal is to compute $w(\rho B_{\mathcal{H}_K} \cap r\mathcal{E})$ and $\mathrm{Rad}(\rho B_{\mathcal{H}_K} \cap r\mathcal{E})$ for all $\rho, r > 0$ where $B_{\mathcal{H}_K} = \{f \in \mathcal{H}_K : \|f\|_{\mathcal{H}_K} \le 1\}$ is the unit ball of the RKHS and $\mathcal{E} = \{f \in \mathcal{H}_K : \mathbb{E}f(X)^2 \le 1\}$ is the ellipsoid associated with X. In the following, we embed the two sets $B_{\mathcal{H}_K}$ and $\mathcal{E}$ in $l_2 = l_2(\mathbb{N})$ so that we simply have to compute the Gaussian mean width and the Rademacher complexities of the intersection of two ellipsoids sharing the same coordinates structures.

The unit ball of $\mathcal{H}_K$ can be constructed from the eigenvalue decomposition of T_K by considering the feature map $\Phi : \mathcal{X} \to l_2$ defined by $\Phi(x) = \big(\sqrt{\lambda_i}\phi_i(x)\big)_{i\in\mathbb{N}}$ and then the unit ball of $\mathcal{H}_K$ is just

$$B_{\mathcal{H}_K} = \big\{ f_\beta(\cdot) = \big\langle \beta, \Phi(\cdot) \big\rangle : \|\beta\|_{l_2} \le 1 \big\}.$$

One can use the feature map Φ to show that there is an isometry between the two Hilbert spaces $\mathcal{H}_K$ and l_2 endowed with the norm $\|\beta\|_K = \big(\sum \beta_i^2/\lambda_i\big)^{1/2}$. The unit ball of l_2 endowed with the norm $\|\cdot\|_K$ is an ellipsoid denoted by $\mathcal{E}_K$.

Let us now determine the ellipsoid in l_2 associated with the design X obtained via this natural isomorphism $\beta \in l_2 \to f_\beta(\cdot) = \big\langle \beta, \Phi(\cdot) \big\rangle \in \mathcal{H}_K$ between l_2 and $\mathcal{H}_K$. Since $(\phi_i)_i$ is an orthonormal system in L_2, the covariance operator of $\Phi(X)$ in l_2 is simply the diagonal operator with diagonal elements $(\lambda_i)_i$. As a consequence the ellipsoid associated with X is isomorphic to $\tilde{\mathcal{E}} = \{\beta \in l_2 : \mathbb{E}\big\langle \beta, \Phi(X)\big\rangle^2 \leq 1\}$; it has the same coordinate structure as the canonical one in l_2 endowed with $\|\cdot\|_K$: $\tilde{\mathcal{E}} = \{\beta \in l_2 : \sum \lambda_i \beta_i^2 \leq 1\}$. So that, we obtain

$$w(K_\rho(f^*) \cap r\mathcal{E}_{f^*}) = w(\rho\mathcal{E}_K \cap r\tilde{\mathcal{E}}) \sim \left(\sum_j (\rho^2\lambda_j) \wedge r^2\right)^{1/2} \tag{27}$$

where the last inequality follows from Proposition 2.2.1 in [39] (note that we defined the Gaussian mean widths in Definition (2.4) depending on the covariance of X). We also get from Theorem 2.1 in [32] that

$$\mathrm{Rad}(K_\rho(f^*) \cap r\mathcal{E}_{f^*}) \sim \left(\sum_j (\rho^2\lambda_j) \wedge r^2\right)^{1/2}. \tag{28}$$

Note that unlike the previous examples, we do not have to assume isotropicity of the design. Indeed, in the RKHS case, the unit ball of the regularization function is isomorphic to the ellipsoid $\mathcal{E}_K$. Since $\mathcal{E}$ is also an ellipsoid having the same coordinates structure as $\mathcal{E}_K$ (cf. paragraph above), for all $\rho, r > 0$, the intersection $\rho B_{\mathcal{H}_K} \cap r\mathcal{E}$ is equivalent to an ellipsoid, meaning that, it contains an ellipsoid and is contained in a multiple of this ellipsoid. Therefore, the Gaussian mean width and the Rademacher complexity of $\rho B_{\mathcal{H}_K} \cap r\mathcal{E}$ has been computed without assuming isotropicity (thanks to general results on the complexity of Ellipsoids from Proposition 2.2.1 in [39] and Theorem 2.1 in [32]).

It follows from (27) and (28) that the Gaussian mean width and the Rademacher complexities are equal. Therefore, up to constant (L in the subgaussian case and b in the bounded case), the two subgaussian and bounded setups may be analyzed at the same time. Nevertheless, since we will only consider in this setting the hinge loss and that the Bernstein condition (cf. Assumption 2.1) with respect to the hinge loss has been studied in Proposition 8.3 only in the bounded case, we continue the analysis only for the bounded framework. Moreover, the subgaussian assumption is not natural for RKHS.

We are now able to identify the complexity parameter of the problem. We actually do not use the localization in this and rather use only the global complexity parameter as defined in Definition 2.7: for all $\rho > 0$:

$$r(\rho) = \left[\frac{\mathbf{C}\rho \left(\sum_j \lambda_j\right)^{1/2}}{\sqrt{N}} \right]^{\frac{1}{2\kappa}} \tag{29}$$

where $\kappa \geq 1$ is the Bernstein parameter.

Results in the bounded setting. Finally, let us discuss about the boundedness assumption. It is known (cf., for instance, Lemma 4.23 in [38]) that if the kernel K is bounded then the functions in the RKHS $\mathcal{H}_K$ are bounded: for any $f \in \mathcal{H}_K$, $\|f\|_{L_\infty} \leq \|K\|_\infty \|f\|_{\mathcal{H}_K}$ where $\|K\|_\infty := \sup_{x\in\mathcal{X}} \sqrt{K(x,x)}$. As a consequence, if one restricts the search space of the RERM to a RKHS ball of radius R, one has $F := RB_{\mathcal{H}_K} \subset R\|K\|_\infty B_{L_\infty}$ and therefore the boundedness assumption is satisfied by F. However, note that a refinement of the proof of Theorem 9.2 using a boundedness parameter b depending on the radius of the RKHS balls used while performing the peeling device yields statistical properties for the RERM with no search space constraint. For the sake of shortness, we do not provide this analysis here.

We are now in a position to provide estimation and prediction results for the RERM

$$\hat{f} \in \underset{f \in RB_{\mathcal{H}_K}}{\operatorname{argmin}} \left(\frac{1}{N} \sum_{i=1}^{N} (1 - Y_i f(X_i))_+ + \frac{\mathbf{C}\left(\sum_j \lambda_j\right)^{1/2}}{\sqrt{N}} \|f\|_{\mathcal{H}_K} \right) \tag{30}$$

where the choice of the regularization parameter λ follows from Theorem (2.2) and (28) (for $r = +\infty$). Note that unlike the examples in the previous sections, we do not have to find some radius ρ^* satisfying the sparsity equation (3) to apply Theorem 2.2 since we simply take $\rho^* = 20\|f^*\|_{\mathcal{H}_K}$ to insure that $0 \in \Gamma_{f^*}(\rho^*)$.

THEOREM 7.1. *Let $\mathcal{X}$ be some space, $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a bounded kernel and denote by $\mathcal{H}_K$ the associated RKHS. Denote by $(\lambda_i)_i$ the sequence of eigenvalues associated to $\mathcal{H}_K$ in L_2. Assume that the Bayes rule $\overline{f}$ from (35) belongs to $RB_{\mathcal{H}_K}$ and that the margin assumption (36) is satisfied for some $\kappa \geq 1$.*

Then the RERM defined in (30) satisfies with probability larger than

$$1 - \mathbf{C}\exp\left(-\mathbf{C}N^{1/2\kappa} \left(\|\overline{f}\|_{\mathcal{H}_K} \left(\sum_j \lambda_j \right)^{1/2} \right)^{(2\kappa-1)/\kappa} \right),$$

that

$$\left\|\hat{f}-\overline{f}\right\|_{L_2} \leq \mathbf{C}\left[\frac{\|\overline{f}\|_{\mathcal{H}_K}\left(\sum_j \lambda_j\right)^{1/2}}{\sqrt{N}}\right]^{1/2\kappa} \quad \text{and } \mathcal{E}_{hinge}(\hat{f}) \leq \mathbf{C}\frac{\|\overline{f}\|_{\mathcal{H}_K}\left(\sum_j \lambda_j\right)^{1/2}}{\sqrt{N}}$$

where $\mathcal{E}_{hinge}(\hat{f})$ is the excess hinge risk of $\hat{f}$.

Note that classic procedures in the literature on RKHS are mostly developed in the classification framework. They are usually based on the hinge loss and the regularization function is the square of the RKHS norm. For such procedures, oracle inequalities have been obtained in Chapter 7 from [38] under the margin assumption (cf. [40]). A result that is close to the one obtained in Theorem 7.1 is Corollary 4.12 in [34]. Assuming that $\|Y\|_\infty \leq \mathbf{C}$, $\mathcal{X} \subset \mathbb{R}^d$, $\|K\|_\infty \leq 1$, that the eigenvalues of the integral operator satisfies

$$\lambda_i \leq c i^{-1/p} \tag{31}$$

for some $0 < p < 1$ and that the eigenvectors (ϕ_i) are such that $\|\phi_i\|_\infty \leq A$ for any i and some constant A then the RERM $\tilde{f}$ over the entire RKHS space, w.r.t. the quadratic loss and for a regularization function of the order of (up to logarithmic terms)

$$f \mapsto \rho(\|f\|_{\mathcal{H}}) := \max\left(\frac{\|f\|_{\mathcal{H}}^{2p/(1+p)}}{N^{1/(1+p)}}, \frac{\|f\|_{\mathcal{H}}^2}{N}\right) \tag{32}$$

satisfies with large probability an oracle inequality like

$$\mathbb{E}(Y-\tilde{f}(X))^2 \leq \inf_{r\geq 1}\left(\inf_{\|f\|_{\mathcal{H}}\leq r} \mathbb{E}(Y-f(X))^2 + \mathbf{C}\rho(r)\right).$$

In particular, an error bound (up to log factors) follows from this result: with high probability,

$$\left\|\tilde{f}-f^*\right\|_{L_2}^2 \leq \mathbf{C}\rho(\|f^*\|_{\mathcal{H}}) = \mathbf{C}\max\left(\frac{\|f^*\|_{\mathcal{H}}^{2p/(1+p)}}{N^{1/(1+p)}}, \frac{\|f^*\|_{\mathcal{H}}^2}{N}\right). \tag{33}$$

One may compare this result to the one from Theorem 7.1 under assumption (31) even though the two procedures $\tilde{f}$ and $\hat{f}$ use different loss functions, regularization function and different search space. If assumption (31) holds then $\left(\sum_j \lambda_j\right)^{1/2} \leq c$ and so, one can take $r(\rho) = \left(\mathbf{C}c\rho/(\theta\sqrt{N})\right)^{1/(2\kappa)}$ and

$\lambda = \mathbf{C}\sqrt{C/N}$. For such a choice of regularization parameter, Theorem 7.1 provides an error bound of the order of

$$\left\|\tilde{f} - \overline{f}\right\|_{L_2(\mu)}^2 \leq \mathbf{C}\left[\frac{\left\|\overline{f}\right\|_{\mathcal{H}_K} C}{\sqrt{N}}\right]^{1/\kappa} \tag{34}$$

which is almost the same as the one obtained in (33) when $\kappa = 1$ and p is close to 1. But our result is worse when $\kappa > 1$ and p is far from 1. This is the price that we pay by using the hinge loss – note that the quadratic loss satisfies the Bernstein condition with $\kappa = 1$ – and by fixing a regularization function which is the norm $\|\cdot\|_{\mathcal{H}_K}$ instead of fitting the regularization function in a "complexity dependent way" as in (32). In the last case, our procedure $\hat{f}$ does not benefit from the "real complexity" of the problem which is localized Rademacher complexities – note that we used global Rademacher complexities to fit λ and construct the complexity function $r(\cdot)$.

8. A review of the Bernstein and margin conditions.

In order to apply the main results from Theorem 2.1 and Theorem 2.2, one has to check the Bernstein condition. This section is devoted to the study of this condition for three loss functions: the hinge loss, the quantile loss and the logistic loss. This condition has been extensively studied in Learning theory (cf. [2, 46, 33, 4, 43, 18]). We can identify mainly two approaches to study this condition: when the class F is convex and the loss function ℓ is "strongly convex", then the risk function inherits this property and automatically satisfies the Bernstein condition (cf. [2]). On the other hand, for loss functions like the hinge or quantile loss, that are affine by parts, one has to use a different path. In such cases, one may go back to a statistical framework and try to check the margin assumption. As a consequence, in the latter case, the Bernstein condition is usually more restrictive and requires strong assumptions on the distribution of the observations.

8.1. *Logistic loss.*

In this section, we study the Bernstein condition of the logistic loss function which is defined for every $f : \mathcal{X} \to \mathbb{R}$, $x \in \mathcal{X}$, $y \in \{-1, 1\}$ and $u \in \mathbb{R}$ by

$$\ell_f(x, y) = \tilde{\ell}(yf(x)) \text{ where } \tilde{\ell}(u) = \log(1 + \exp(-u)).$$

Function $\tilde{\ell}$ is strongly convex on every compact interval in $\mathbb{R}$. As it was first observed in [2, 3], one may use this property to check the Bernstein condition for the loss function ℓ. This approach was extended to the bounded regression problem with respect to L_p loss functions ($1 < p < \infty$) in [31] and to non convex classes in [33].

In the bounded scenario, [3] proved that the logistic loss function satisfies the Bernstein condition for $\kappa = 1$. One may therefore use that result to apply Theorem 2.2. The analysis is pretty straightforward in the bounded case. It becomes more delicate in the subgaussian scenario as considered in Theorem 2.1.

Proposition 8.1 ([2]). *Let F be a convex class of functions from $\mathcal{X}$ to $\mathbb{R}$. Assume that for every $f \in F$, $\|f\|_{L_\infty} \leq b$. Then the class F satisfies the Bernstein condition (for the logistic loss) with Bernstein parameter $\kappa = 1$ and constant $A = 4\exp(2b)$.*

The proof is given in Section 11. This result solves the problem of the Bernstein condition with respect to the logistic loss function over a convex class F of functions as long as all functions in F are uniformly bounded by some constant b. We will therefore use this result only in the bounded framework, for instance, when F is a class of linear functionals indexed by a bounded set of vectors and X takes values in the canonical basis – like in Section 4 where we assumed that $X \in \{E_{1,1}, \cdots, E_{m,T}\}$, the canonical basis of $\mathbb{R}^{m\times T}$.

In the subgaussian framework, one may proceed as in [43] and assume that a statistical model holds. In that case, the Bernstein condition is reduced to the study of the margin assumption since, in that case, the "Bayes rule" $\overline{f}$ (which is called the log-odds ratio in the case of the logistic loss function) is assumed to belong to the class F and so $f^* = \overline{f}$. The margin assumption with respect to the logistic loss function has been studied in Example 1 from [43] but for a slightly different definition of the margin assumption. Indeed, in [43] only functions f in a L_∞ neighborhood of $\overline{f}$ need to satisfy the margin assumption whereas in Assumption 2.1 it has to be satisfied at least in the non-bounded set $\mathcal{C}$ (see Remark 2.1).

From our perspective, we do not want to make no "statistical modeling assumption". In particular, we do not want to assume that $\overline{f}$ belongs to F. We therefore have to prove the Bernstein condition when $\bar{f}$ may not belong to F. We used this result in Section 3 in order to obtain statistical bounds for the logistic LASSO and logistic SLOPE procedures. In those cases, F is a class of linear functionals. We now state that the Bernstein condition is satisfied for a class of linear functional when X is a standard Gaussian vector (the proof has been postponed to Section 11).

Proposition 8.2. *Let $F = \{\langle \cdot, t\rangle : t \in RB_{l_2}\}$ be a class of linear functionals indexed by RB_{l_2} for some radius $R \geq 1$. Let X be a standard Gaussian vector in $\mathbb{R}^d$ and let Y be a $\{-1, 1\}$ random variable. For every*

$f \in F$, the excess logistic risk of f, denoted by $P\mathcal{L}_f$, satisfies

$$\mathcal{E}_{logistic}(f) = P\mathcal{L}_f \geq \frac{c_0}{R^3} \|f - f^*\|_{L_2}^2$$

where c_0 is some absolute constant.

8.2. *Hinge loss.* Unlike the logistic loss function (on a bounded support), both the hinge loss and the quantile losses do not enjoy a strong convexity property. Therefore, one has to turn to a different approach as the one used in the previous section to check the Bernstein condition for those two loss functions.

For the hinge loss function, Bernstein condition is more stringent and is connected to the margin condition in classification. So, let us first introduce some notations specific to classification. In this setup, one is given N labeled pairs $(X_i, Y_i), i = 1, \dots, N$ where X_i takes its values in $\mathcal{X}$ and Y_i is a label taking values in $\{-1, +1\}$. The aim is to predict the label Y associated with X from the data when (X, Y) is distributed like the (X_i, Y_i)'s. The classical loss function considered in this setup is the $0-1$ loss function $\ell_f(x, y) = I(y \neq f(x))$ defined for any $f : \mathcal{X} \to \{-1, +1\}$. The $0-1$ loss function is not convex, this may result in some computational issues when dealing with it. A classic approach is to use a " convex relaxation function" as a surrogate to the $0-1$ loss function: note that this is a way to motivate the introduction of the hinge loss $\ell_f(x, y) = \max(1 - yf(x), 0)$. It is well known that the Bayes rules minimizes both the standard $0-1$ risk as well as the hinge risk: put $\eta(x) := \mathbb{E}[Y|X = x]$ for all $x \in \mathcal{X}$ and define the Bayes rule as

$$\overline{f}(x) = \operatorname{sgn}(\eta(x)), \tag{35}$$

then $\overline{f}$ minimizes $f \to P\ell_f$ over all measurable functions from $\mathcal{X}$ to $\mathbb{R}$ when ℓ_f is the hinge loss of f.

Let F be a class of functions from $\mathcal{X}$ to $[-1, 1]$. Assume that $\overline{f} \in F$ so that $\overline{f}$ is an oracle in F and thus (using the notations from Section 2) $f^* = \overline{f}$. In this situation, the margin assumption with respect to the hinge loss (cf. [40, 25]) restricted to the class F and Bernstein condition (cf. Assumption 2.1) coincide. Therefore, Assumption 2.1 holds when the margin assumption w.r.t. the hinge loss holds. According to Proposition 1 in [25], the margin assumption with respect to the hinge loss is equivalent the margin assumption with respect to the $0-1$ loss for a class F of functions with values in $[-1, 1]$. Then, according to Proposition 1 in [40] and [8] the margin assumption with respect to the $0-1$ loss with parameter κ is equivalent to

$$\left\{\begin{array}{ll} \mathbb{P}(|\eta(X)|\leq t)\leq ct^{\frac{1}{\kappa-1}}, \forall 0\leq t\leq 1 & \text{when } \kappa>1\\ |\eta(X)|\geq\tau \text{ a.s. for some } \tau>0 & \text{when } \kappa=1. \end{array}\right. \tag{36}$$

As a consequence, one can state the following result on the Bernstein condition for the hinge loss in the bounded case scenario.

Proposition 8.3 (Proposition 1, [25]). *Let F be a class of functions from $\mathcal{X}$ to $[-1,1]$. Define $\eta(x)=\mathbb{E}[Y|X=x]$ for all $x\in\mathcal{X}$ and assume that the Bayes rule (35) belongs to F. If (36) is satisfied for some $\kappa\geq 1$ then Assumption 2.1 holds with parameter κ for the hinge loss, and A depending on c, κ and τ (which is explicitly given in the mentioned references). In the special case when $\kappa=1$ then $A=1/(2\tau)$.*

Note that up to a modification of the constant A, the same result holds for functions with values in $[-b,b]$ for $b>0$, a fact we used in Section 7.

8.3. *Quantile loss.* In this section, we study the Bernstein parameter of the quantile loss in the bounded regression model, that is when for all $f\in F, \|f\|_{L_\infty}\leq b$ a.s.. Let $\tau\in(0,1)$ and, for all $x\in\mathcal{X}$, define $\overline{f}(x)$ as the quantile of order τ of $Y|X=x$ and assume that $\bar{f}$ belongs to F, in that case, $\overline{f}=f^*$ and Bernstein condition and margin assumption are the same. Therefore one may follow the study of the margin assumption for the quantile loss in [18] to obtain the following result.

Proposition 8.4 ([18]). *Assume that for any $x\in\mathcal{X}$, it is possible to define a density f_x w.r.t the Lebesgue measure for $Y|X=x$ such that $f_x(u)\geq 1/C$ for some $C>0$ for all $u\in\mathbb{R}$ with $|u-f^*(x)|\leq 2b$. Then the quantile loss satisfies Bernstein's assumption with $\kappa=1$ and $A=2C$ over F.*

9. Proof of Theorem 2.1 and Theorem 2.2.

9.1. *More general statements: Theorems 9.2 and 9.1.* First, we state two theorems: Theorem 9.1 in the subgaussian setting, and Theorem 9.2 in the bounded setting. These two theorems rely on localized versions of the complexity function $r(\cdot)$ that will be defined first. Note that the localized version of $r(\cdot)$ can always be upper bounded by the simpler version used in the core of the paper. Thus, Theorem 2.1 is a direct corollary of Theorem 9.1, and Theorem 2.2 is a direct corollary of Theorem 9.2.

So let us start with a localized complexity parameters. The "statistical size" of the family of "sub-models" $(\rho B)_{\rho > 0}$ is now measured by local Gaussian mean-widths in the subgaussian framework.

Definition 9.1. *Let $\theta > 0$. The **complexity parameter** is a non-decreasing function $r(\cdot)$ such that for every $\rho \geq 0$,*

$$CLw\left(\rho B \cap r(\rho) B_{L_2}\right) \leq \theta r(\rho)^{2\kappa} \sqrt{N}$$

where κ is the Bernstein parameter from Assumption 2.1.

In the boundedness case, it is written as follows.

Definition 9.2. *Let $\theta > 0$. The **complexity parameter** is a non-decreasing function $r(\cdot)$ such that for every $\rho \geq 0$,*

$$48 \mathrm{Rad}(\rho B \cap r(\rho) B_{L_2}) \leq \theta r(\rho)^{2\kappa} \sqrt{N}$$

where κ is the Bernstein parameter from Assumption 2.1.

To obtain the complexity functions from Definition 2.5 and 2.7, we use the fact that $w\left(\rho B \cap r(\rho) B_{L_2}\right) \leq w(\rho B)$ and $\mathrm{Rad}(\rho B \cap r(\rho) B_{L_2}) \leq \mathrm{Rad}(\rho B)$: it indeed does not use the localization. We also set $\theta = 7/40A$ in those definitions because it is the largest value allowed in the following theorems.

Theorem 9.1. *Assume that Assumption 1.1, Assumption 2.1 and Assumption 2.2 hold. Let $r(\cdot)$ be a function as in Definition 9.1 for some θ such that $40A\theta \leq 7$ and assume that $\rho \to r(2\rho)/\rho$ is non-increasing. Let the regularization parameter λ be chosen such that*

$$\frac{10\theta r(2\rho)^{2\kappa}}{7\rho} < \lambda < \frac{r(2\rho)^{2\kappa}}{2A\rho}, \quad \forall \rho \geq \rho^* \tag{37}$$

where ρ^* satisfies (3). Then, with probability larger than

$$1 - \sum_{j=0}^{\infty} \sum_{i \in I_j} \exp\left(- \frac{\theta^2 N (2^{(i-1)\vee 0} r(2^j \rho^*))^{4\kappa - 2}}{4C^2 L^2} \right) \tag{38}$$

where for all $j \in \mathbb{N}$, $I_j = \{1\} \cup \{i \in \mathbb{N}^* : 2^{i-1} r(2^j \rho^*) \leq 2^j \rho^* d_{L_2}(B)\}$, we have

$$\left\| \hat{f} - f^* \right\| \leq \rho^*, \quad \left\| \hat{f} - f^* \right\|_{L_2} \leq r(2\rho^*) \text{ and } \mathcal{E}(\hat{f}) \leq r(2\rho^*)^{2\kappa}/A.$$

Proof of Theorem 2.1: Let $r(\cdot)$ be chosen as in Definition 2.5. For this choice, one can check that the regularization parameter used for the construction of the RERM satisfies (37) with an adequate choice of constants. Moreover, for this choice of function $r(\cdot)$ it is straightforward to lower bound the sum in the probability estimate in (38). □

The bounded case is in the same spirit.

Theorem 9.2. *Assume that Assumption 1.1, Assumption 2.1 and Assumption 2.3 hold. Let $r(\cdot)$ be a function as in Definition 9.2 for some θ such that $40A\theta \leq 7$ and assume that $\rho \to r(2\rho)/\rho$ is non-increasing. Let the regularization parameter λ be chosen such that*

$$\frac{10\theta r(2\rho)^{2\kappa}}{7\rho} < \lambda < \frac{r(2\rho)^{2\kappa}}{2A\rho}, \quad \forall \rho \geq \rho^* \tag{39}$$

where ρ^* satisfies (3). Then, with probability larger than

$$1 - 2\sum_{j=0}^{\infty}\sum_{i\in I_j} \exp\left(-c_0\theta^2 N(2^i r(2^{j+1}\rho^*))^{4\kappa-2}\right) \tag{40}$$

where $c_0 = 1/\max\left(48, 207\theta b^{2\kappa-1}\right)$ and for all $j \in \mathbb{N}$, $I_j := \{1\} \cup \{i \in \mathbb{N}^* : 2^{i-1} r(2^j\rho^*) \leq \min(2^j\rho^* d_{L_2}(B), b)\}$, we have

$$\left\|\hat{f} - f^*\right\| \leq \rho^*, \quad \left\|\hat{f} - f^*\right\|_{L_2} \leq r(2\rho^*) \text{ and } \mathcal{E}(\hat{f}) \leq r(2\rho^*)^{2\kappa}/A.$$

The proof of Theorem 2.2 is straightforward consequence of Theorem 9.2. It is identical to the one of Theorem 2.1 and we do not reproduce it here.

9.2. *Proofs of Theorems 9.2 and 9.1.* Proof of Theorem 9.1 and and Theorem 9.2 follow the same strategy. They are split into two parts. First, we identify an event onto which the statistical behavior of the regularized estimator $\hat{f}$ can be controlled using only deterministic arguments. Then, we prove that this event holds with a probability at least as large as the one in (38) in the case of Theorem 9.1 and as in (40) in the case of Theorem 9.2. We first introduce this event which is common to the subgaussian and the bounded setups:

$$\Omega_0 := \left\{ \begin{array}{cc} & \text{for all } f \in F \\ \left|(P - P_N)\mathcal{L}_f\right| \leq & \theta \max\left(r(2\max(\|f - f^*\|, \rho^*))^{2\kappa}, \|f - f^*\|_{L_2}^{2\kappa} \right) \end{array} \right\}$$

where θ is a parameter appearing in the definition of $r(\cdot)$ in Definition 9.1 and Definition 9.2, $\kappa \geq 1$ is the Bernstein parameter from Definition 2.1 and ρ^* is a radius satisfying the sparsity Equation (3).

Proposition 9.1. *Let λ be as in (37) (or equivalently as in (39)) and let ρ^* satisfy (3), on the event Ω_0, one has*

$$\left\|\hat{f} - f^*\right\| \leq \rho^*, \quad \left\|\hat{f} - f^*\right\|_{L_2} \leq r(2\rho^*) \text{ and } \mathcal{E}(\hat{f}) \leq \theta r(2\rho^*)^{2\kappa}.$$

Proof. Denote $\hat{\rho} = \left\|\hat{f} - f^*\right\|$. We first prove that $\hat{\rho} < \rho^*$. To that end, we assume that the reverse inequality holds and show some contradiction. Assume that $\hat{\rho} \geq \rho^*$. Since $\rho \to r(2\rho)/\rho$ is non-increasing then by Lemma A.1, $\rho \to \Delta(\rho)/\rho$ is non-decreasing and so we have

$$\frac{\Delta(\hat{\rho})}{\hat{\rho}} \geq \frac{\Delta(\rho^*)}{\rho^*} \geq \frac{4}{5}.$$

Now, we consider two cases: either $\left\|\hat{f} - f^*\right\|_{L_2} \leq r(2\hat{\rho})$ or $\left\|\hat{f} - f^*\right\|_{L_2} > r(2\hat{\rho})$.

First assume that $\left\|\hat{f} - f^*\right\|_{L_2} \leq r(2\hat{\rho})$. Since $\Delta(\hat{\rho}) \geq 4\hat{\rho}/5$ and $h = \hat{f} - f^* \in \hat{\rho}S \cap r(2\hat{\rho})B_{L_2}$, it follows from the definition of the sparsity parameter $\Delta(\hat{\rho})$ that there exists some $f \in F$ such that $\|f - f^*\| \leq \hat{\rho}/20$ and for which

$$\|f + h\| - \|f\| \geq \frac{4\hat{\rho}}{5}.$$

It follows that

$$\left\|\hat{f}\right\| - \|f^*\| = \|f^* + h\| - \|f^*\| \geq \|f + h\| - \|f\| - 2\|f - f^*\| \geq \frac{4\hat{\rho}}{5} - \frac{\hat{\rho}}{10} = \frac{7\hat{\rho}}{10}.$$

Let us now introduce the excess regularized loss: for all $f \in F$,

$$\mathcal{L}_f^\lambda = \mathcal{L}_f + \lambda(\|f\| - \|f^*\|) = (\ell_f + \lambda\|f\|) - (\ell_{f^*} + \lambda\|f^*\|).$$

On the event Ω_0, we have

$$\begin{aligned} P_N\mathcal{L}_{\hat{f}}^\lambda &= P_N\mathcal{L}_{\hat{f}} + \lambda\left(\left\|\hat{f}\right\| - \|f^*\|\right) \geq (P_N - P)\mathcal{L}_{\hat{f}} + \lambda\left(\left\|\hat{f}\right\| - \|f^*\|\right) \\ &\geq -\theta\max\left(r(2\hat{\rho})^{2\kappa}, \left\|\hat{f} - f^*\right\|_{L_2}^{2\kappa}\right) + \frac{7\lambda\hat{\rho}}{10} = -\theta r(2\hat{\rho})^{2\kappa} + \frac{7\lambda\hat{\rho}}{10} > 0 \end{aligned}$$

because by definition of λ, $7\lambda\hat{\rho} > 10\theta r(2\hat{\rho})^{2\kappa}$. Therefore, $P_N\mathcal{L}_{\hat{f}}^\lambda > 0$. But, by construction, one has $P_N\mathcal{L}_{\hat{f}}^\lambda \leq 0$.

Then, assume that $\left\|\hat{f} - f^*\right\|_{L_2} > r(2\hat{\rho})$. In particular, $\hat{f} \in \mathcal{C}$ where $\mathcal{C}$ is the set introduced in Equation (2) from Remark 2.1 . By definition of $\hat{f}$ we have

$P_N\mathcal{L}_{\hat{f}}^{\lambda} \leq 0$ so it follows from the Bernstein condition (cf. Assumption 2.1) that

$$\left\|\hat{f}-f^*\right\|_{L_2}^{2\kappa} \leq AP\mathcal{L}_{\hat{f}} = A\left[(P-P_N)\mathcal{L}_{\hat{f}} + P_N\mathcal{L}_{\hat{f}}^{\lambda} + \lambda\left(\|f^*\| - \left\|\hat{f}\right\|\right)\right]$$

(41)

$$\leq A\theta \max\left(r(2\hat{\rho})^{2\kappa}, \left\|\hat{f}-f^*\right\|_{L_2}^{2\kappa}\right) + A\lambda\left\|\hat{f}-f^*\right\| = A\theta\left\|\hat{f}-f^*\right\|_{L_2}^{2\kappa} + A\lambda\hat{\rho}.$$

Hence, if $A\theta \leq 1/2$ then

$$r(2\hat{\rho})^{2\kappa} \leq \left\|\hat{f}-f^*\right\|_{L_2}^{2\kappa} \leq 2A\lambda\hat{\rho}.$$

But, by definition of λ one has $r(2\hat{\rho})^{2\kappa} > 2A\lambda\hat{\rho}$.

Therefore, none of the two cases is possible when one assumes that $\hat{\rho} \geq \rho^*$ and so we necessarily have $\hat{\rho} < \rho^*$.

Now, assuming that $\left\|\hat{f}-f^*\right\|_{L_2} > r(2\rho^*)$ and following (41) step by step also leads to a contradiction, so $\left\|\hat{f}-f^*\right\|_{L_2} \leq r(2\rho^*)$. p Next, we prove the result for the excess risk. One has

$$P_N\mathcal{L}_{\hat{f}}^{\lambda} = P_N\mathcal{L}_{\hat{f}} + \lambda\left(\left\|\hat{f}\right\| - \|f^*\|\right) = (P_N-P)\mathcal{L}_{\hat{f}} + P\mathcal{L}_{\hat{f}} + \lambda\left(\left\|\hat{f}\right\| - \|f^*\|\right)$$

$$\geq -\theta\max\left(r(2\rho^*)^{2\kappa}, \left\|\hat{f}-f^*\right\|_{L_2}^{2\kappa}\right) + P\mathcal{L}_{\hat{f}} - \lambda\hat{\rho} \geq -\theta r(2\rho^*)^{2\kappa} - \lambda\rho^* + P\mathcal{L}_{\hat{f}}$$

$$\geq -\left(\theta + \frac{1}{2A}\right)r(2\rho^*)^{2\kappa} + P\mathcal{L}_{\hat{f}} \geq \frac{-r(2\rho^*)^{2\kappa}}{A} + P\mathcal{L}_{\hat{f}}.$$

In particular, if $P\mathcal{L}_{\hat{f}} > r(2\rho^*)^{2\kappa}/A$ then $P_N\mathcal{L}_{\hat{f}}^{\lambda} > 0$ which is not possible by construction of $\hat{f}$ so we necessarily have $P\mathcal{L}_{\hat{f}} \leq r(2\rho^*)^{2\kappa}/A$. □

Proposition 9.1 shows that $\hat{f}$ satisfies some estimation and prediction properties on the event Ω_0. Next, we prove that Ω_0 holds with large probability in both subgaussian and bounded frameworks. We start with the subgaussian framework. To that end, we introduce several tools.

Recall that the ψ_2-norm of a real valued random variable Z is defined by

$$\|Z\|_{\psi_2} = \inf\left\{c > 0 : \mathbb{E}\psi_2(|Z|/c) \leq \psi_2(1)\right\}$$

where $\psi_2(u) = \exp(u^2) - 1$ for all $u \geq 0$. The space L_{ψ_2} of all real valued random variables with finite ψ_2-norm is called the Orlicz space of subgaussian variables. We refer the reader to [36, 37] for more details on Orlicz spaces.

We recall several facts on the ψ_2-norm and subgaussian processes. First, it follows from Theorem 1.1.5 from [11] that $\|Z\|_{\psi_2} \leq \max(K_0, K_1)$ if

$$\mathbb{E}\exp(\lambda|Z|) \leq \exp\left(\lambda^2 K_1^2\right), \quad \forall \lambda \geq 1/K_0. \tag{42}$$

It follows from Lemma 1.2.2 from [11] that, if Z is a centered ψ_2 random variable then, for all $\lambda > 0$,

$$\mathbb{E}\exp\left(\lambda Z\right) \leq \exp\left(e\lambda^2 \|Z\|_{\psi_2}^2\right). \tag{43}$$

Then, it follows from Theorem 1.2.1 from [11] that if $Z_1, \ldots, Z_N$ are independent centered real valued random variables then

$$\left\|\sum_{i=1}^N Z_i\right\|_{\psi_2} \leq 16 \left(\sum_{i=1}^N \|Z_i\|_{\psi_2}^2\right)^{1/2}. \tag{44}$$

Finally, let us turn to some properties of subgaussian processes. Let (T, d) be a pseudo-metric space. Let $(X_t)_{t\in T}$ be a random process in L_{ψ_2} such that for all $s, t \in T$, $\|X_t - X_s\|_{\psi_2} \leq d(s,t)$. It follows from the comment below Theorem 11.2 p.300 in [29] that for all measurable set A and all $s, t \in T$,

$$\int_A |X_s - X_t| d\mathbb{P} \leq d(s,t)\mathbb{P}(A)\psi_2^{-1}\left(\frac{1}{\mathbb{P}(A)}\right).$$

Therefore, it follows from equation (11.14) in [29] that for every $u > 0$,

$$\mathbb{P}\left(\sup_{s,t\in T} |X_s - X_t| > c_0(\gamma_2 + Du)\right) \leq \psi_2(u)^{-1} \tag{45}$$

where D is the diameter of (T, d), c_0 is an absolute constant and γ_2 is the majorizing measure integral $\gamma(T, d; \psi_2)$ (cf. Chapter 11 in [29]). When T is a subset of L_2 and d is the natural metric of L_2 it follows from the majorizing measure theorem that $\gamma_2 \leq c_1 w(T)$ (cf. Chapter 1 in [39]).

Lemma 9.1. *Assume that Assumption 1.1 and Assumption 2.2 hold. Let $F' \subset F$ then for every $u > 0$, with probability at least $1 - 2\exp(-u^2)$*

$$\sup_{f,g\in F'} |(P - P_N)(\mathcal{L}_f - \mathcal{L}_g)| \leq \frac{c_0 L}{\sqrt{N}}\left(w(F') + u d_{L_2}(F')\right)$$

where d is the L_2 metric and $d_{L_2}(F')$ is the diameter of (F', d).

PROOF. To prove Lemma 9.1, it is enough to show that $((P-P_N)\mathcal{L}_f)_{f\in F'}$ has $(L/\sqrt{N})$-subgaussian increments and then to apply (45) where $\gamma_2 \sim w(F')$ in this case.

Let us prove that for some absolute constant c_0: for all $f,g\in F'$,

$$\left\|(P-P_N)(\mathcal{L}_f-\mathcal{L}_g)\right\|_{\psi_2}\leq c_0(L/\sqrt{N})\left\|f-g\right\|_{L_2}$$

It follows from (44) that

$$\begin{aligned}\left\|(P-P_N)(\mathcal{L}_f-\mathcal{L}_g)\right\|_{\psi_2} &\leq 16\left(\sum_{i=1}^N\frac{\left\|(\mathcal{L}_f-\mathcal{L}_g)(X_i,Y_i)-\mathbb{E}(\mathcal{L}_f-\mathcal{L}_g)\right\|_{\psi_2}^2}{N^2}\right)^{1/2}\\ &=\frac{16}{\sqrt{N}}\left\|\zeta_{f,g}\right\|_{\psi_2}.\end{aligned}$$

where $\zeta_{f,g}=(\mathcal{L}_f-\mathcal{L}_g)(X,Y)-\mathbb{E}(\mathcal{L}_f-\mathcal{L}_g)$. Therefore, it only remains to show that $\left\|\zeta_{f,g}\right\|_{\psi_2}\leq c_1L\left\|f-g\right\|_{L_2}$.

It follows from (42), that the last inequality holds if one proves that for all $\lambda\geq c_1/(L\left\|f-g\right\|_{L_2})$,

$$\mathbb{E}\exp\left(\lambda|\zeta_{f,g}|\right)\leq\exp(c_2\lambda^2L^2\left\|f-g\right\|_{L_2}^2) \tag{46}$$

for some absolute constants c_1 and c_2. To that end, it is enough to prove that, for some absolute constant c_3 – depending only on c_1 and c_2 – and all $\lambda>0$,

$$\mathbb{E}\exp\left(\lambda|\zeta_{f,g}|\right)\leq 2\exp(c_3\lambda^2L^2\left\|f-g\right\|_{L_2}^2).$$

Note that if Z is a real valued random variable and ϵ is a Rademacher variable independent of Z then $\mathbb{E}\exp(|Z|)\leq 2\mathbb{E}\exp(\epsilon Z)$. Hence, it follows from a symmetrization argument (cf. Lemma 6.3 in [29]), (a simple version of) the contraction principle (cf. Theorem 4.4 in [29]) and (43) that, for all $\lambda>0$,

$$\begin{aligned}&\mathbb{E}\exp\left(\lambda|\zeta_{f,g}|\right)\leq 2\mathbb{E}\exp(\lambda\epsilon\zeta_{f,g})\leq 2\mathbb{E}\exp\left(2\lambda\epsilon(\mathcal{L}_f-\mathcal{L}_g)(X,Y)\right)\\ &\leq 2\mathbb{E}\exp\left(2\lambda\epsilon(f-g)(X)\right)\leq 2\mathbb{E}\exp\left(c_4\lambda^2L^2\left\|f-g\right\|_{\psi_2}^2\right)\end{aligned}$$

where ϵ is a Rademacher variable independent of (X,Y) and where we used in the last but one inequality that $|\mathcal{L}_f(X,Y)-\mathcal{L}_g(X,Y)|\leq|f(X)-g(X)|$ a.s.. □

PROPOSITION 9.2. *We assume that Assumption 1.1, 2.2 and 2.1 hold. Then the probability measure of Ω_0 is at least as large as the one in (38).*

Proof. The proof is based on a peeling argument (cf. [42]) with respect to the two distances naturally associated with this problem: the regularization norm $\|\cdot\|$ and the L_2-norm $\|\cdot\|_{L_2}$ associated with the design X. The peeling according to $\|\cdot\|$ is performed along the radii $\rho_j = 2^j\rho^*$ for $j \in \mathbb{N}$ and the peeling according to $\|\cdot\|_{L_2}$ is performed within the class $\{f \in F : \|f - f^*\| \leq \rho_j\} := f^* + \rho_j B$ along the radii $2^i r(\rho_j)$ for all $i = 0, 1, 2, \cdots$ up to a radius such that $2^i r(\rho_j)$ becomes larger than the radius of $f^* + \rho_j B$ in L_2, that is for all $i \in I_j$.

We introduce the following partition of the class F. We first introduce the "true model", i.e. the subset of F where we want to show that $\hat{f}$ belongs to with high probability:

$$F_{0,0} = \big\{f \in F : \|f - f^*\| \leq \rho_0 \text{ and } \|f - f^*\|_{L_2} \leq r(\rho_0)\big\}$$

(note that $\rho_0 = \rho^*$). Then we peel the remaining set $F \backslash F_{0,0}$ according to the two norms: for every $i \in I_0$,

$$F_{0,i} = \big\{f \in F : \|f - f^*\| \leq \rho_0 \text{ and } 2^{i-1} r(\rho_0) < \|f - f^*\|_{L_2} \leq 2^i r(\rho_0)\big\},$$

for all $j \geq 1$,

$$F_{j,0} = \big\{f \in F : \rho_{j-1} < \|f - f^*\| \leq \rho_j \text{ and } \|f - f^*\|_{L_2} \leq r(\rho_j)\big\}$$

and for every integer $i \in I_j$,

$$F_{j,i} = \big\{f \in F : \rho_{j-1} < \|f - f^*\| \leq \rho_j \text{ and } 2^{i-1} r(\rho_j) < \|f - f^*\|_{L_2} \leq 2^i r(\rho_j)\big\}.$$

We also consider the sets $F^*_{j,i} = \rho_j B \cap (2^i r(\rho_j)) B_{L_2}$ for all integers i and j.

Let j and $i \in I_j$ be two integers. It follows from Lemma 9.1 that for any $u > 0$, with probability larger than $1 - 2\exp(-u^2)$,

$$\begin{aligned} \sup_{f \in F_{j,i}} |(P - P_N)\mathcal{L}_f| &\leq \sup_{f,g \in F^*_{j,i} + f^*} |(P - P_N)(\mathcal{L}_f - \mathcal{L}_g)| \\ &\leq \frac{c_0 L}{\sqrt{N}} \big(w(F^*_{j,i}) + u d_{L_2}(F^*_{j,i})\big) \end{aligned} \tag{47}$$

where $d_{L_2}(F^*_{j,i}) \leq 2^{i+1} r(\rho_j)$.

Note that for any $\rho > 0$, $h : r \to w(\rho B \cap r B_{L_2})/r$ is non-increasing (cf. Lemma A.2 in the Appendix) and note that, by definition of $r(\rho)$ (cf. Definition 9.1), $h(r(\rho)) \leq \theta r(\rho)^{2\kappa - 1}\sqrt{N}/(CL)$. Since $h(\cdot)$ is non-increasing, we have $w(F^*_{j,i})/(2^i r(\rho_j)) \leq h(2^i r(\rho_j)) \leq h(r(\rho_j)) \leq \theta r(\rho_j)^{2\kappa-1}\sqrt{N}/(CL)$ and so $w(F^*_{j,i}) \leq \theta 2^i r(\rho_j)^{2\kappa}\sqrt{N}/(CL)$. Therefore, it follows from (47) for

$u = \theta\sqrt{N}(2^{(i-1)\vee 0}r(\rho_j))^{2\kappa-1}/(2CL)$, if $C \geq 4c_0$ then, with probability at least

$$1 - 2\exp\left(-\theta^2 N(2^{(i-1)\vee 0}r(\rho_j))^{4\kappa-2}/(4C^2L^2)\right), \tag{48}$$

for every $f \in F_{j,i}$,

$$\begin{aligned}|(P-P_N)\mathcal{L}_f| &\leq \theta(2^{(i-1)\vee 0}r(\rho_j))^{2\kappa}\\ &\leq \theta\max\left(r(2\max(\|f-f^*\|, \rho^*))^{2\kappa}, \|f-f^*\|_{L_2}^{2\kappa}\right).\end{aligned}$$

The result follows from a union bound. □

Now we turn to the proof of Theorem 9.1 under the boundedness assumption. The proof follows the same strategy as in the "subgaussian case": we first use Proposition 9.1 and then show (under the boundedness assumption) that event Ω_0 holds with probability at least as large as the one in (40).

Similar to Proposition 9.2, we prove the following result under the boundedness assumption.

PROPOSITION 9.3. *We assume that Assumption 1.1, 2.3 and 2.1 hold. Then the probability measure of Ω_0 is at least as large as the one in (40).*

PROOF. Using the same notation as in the proof of Proposition 9.2, we have for any integer j and i such that $2^i r(\rho_j) \leq b$ that by Talagrand's concentration inequality: for any $x > 0$, with probability larger than $1-2e^{-x}$,

$$Z_{j,i} \leq 2\mathbb{E}Z_{j,i} + \sigma(\mathcal{L}_{F_{j,i}})\sqrt{\frac{8x}{N}} + \frac{69\left\|\mathcal{L}_{F_{j,i}}\right\|_\infty x}{2N} \tag{49}$$

where

$$Z_{j,i} = \sup_{f\in F_{j,i}} |(P-P_N)\mathcal{L}_f|, \quad \sigma(\mathcal{L}_{F_{j,i}}) = \sup_{f\in F_{j,i}} \sqrt{\mathbb{E}\mathcal{L}_f^2} \text{ and } \left\|\mathcal{L}_{F_{j,i}}\right\|_\infty = \sup_{f\in F_{j,i}} \|\mathcal{L}_f\|_\infty.$$

By the Lipschitz assumption, one has

$$\sigma(\mathcal{L}_{F_{j,i}}) \leq 2^{i+1}r(\rho_j) \text{ and } \left\|\mathcal{L}_{F_{j,i}}\right\|_\infty \leq 2b.$$

Therefore, it only remains to upper bound the expectation $\mathbb{E}Z_{j,i}$. Let $\epsilon_1, \dots, \epsilon_N$ be a N i.i.d. Rademacher variables independent of the (X_i, Y_i)'s. For all function f, we set

$$P_{N,\epsilon}f = \frac{1}{N}\sum_{i=1}^N \epsilon_i f(X_i)$$

It follows from a symmetrization and a contraction argument (cf. Chapter 4 in [29]) that

$$\mathbb{E}Z_{j,i} \le 4\mathbb{E}\sup_{f\in F_{j,i}} |P_{N,\epsilon}(f-f^*)| \le \frac{4\mathrm{Rad}(\rho_j B \cap (2^i r(\rho_j))B_{L_2})}{\sqrt{N}} \le (\theta/12)2^i r(\rho_j)^{2\kappa}.$$

Now, we take $x = c_2\theta^2 N(2^{i-1}r(\rho_j))^{4\kappa-2}$ in (49) and note that $2^i r(\rho_j) \le b$ and $\kappa \ge 1$: with probability larger than

$$1 - 2\exp(-c_2\theta N(2^i r(\rho_j))^{4\kappa-2}), \tag{50}$$

for any $f \in F_{j,i}$,

$$\begin{aligned}
|(P-P_N)\mathcal{L}_f| &\le \theta 2^{i-1}r(\rho_j)^{2\kappa}/3 + 2\sqrt{8c_2}\theta\left(2^{i-1}r(\rho_j)\right)^{2\kappa} + 69c_2\theta^2 b(2^{i-1}r(\rho_j))^{4\kappa-2}\\
&\le \theta\left(2^{(i-1)\vee 0}r(\rho_j)\right)^{2\kappa}\left[\frac{1}{3} + 2\sqrt{8c_2} + 69c_2\theta b(2^i r(\rho_j))^{2\kappa-2}\right]\\
&\le \theta\left(2^{(i-1)\vee 0}r(\rho_j)\right)^{2\kappa}\left[\frac{1}{3} + 2\sqrt{8c_2} + 69c_2\theta b^{2\kappa-1}\right]\\
&\le \theta\max\left(r(2\max(\|f-f^*\|, \rho^*))^{2\kappa}, \|f-f^*\|_{L_2}^{2\kappa}\right)
\end{aligned}$$

if c_2 is defined by

$$c_2 = \min\left(\frac{1}{48}, \frac{1}{207\theta b^{2\kappa-1}}\right). \tag{51}$$

We conclude with a union bound. □

Remark 9.1 (Technical comments). *The machinery we used here to prove Theorems 9.2 and 9.1 is inspired by the technique from [28, 27]. In the latter papers, estimation properties of regularized ERM are also obtained but for the square loss. Here, we extend those results to Lipschitz loss functions. Going from the square loss to Lipschitz loss functions is not straightforward. It requires another machinery given that the homogeneity argument used in [28, 27] do not apply here. In a nutshell, this argument shows that if $P_N\mathcal{L}_f^\lambda > 0$ (where $\mathcal{L}_f^\lambda$ is the regularized excess loss of f) for all $f \in F \cap (f^* + r_* S_{L_2})$ then it is also the case for all f such that $\|f-f^*\|_{L_2} \ge r_*$ (and therefore the RERM is in $f^* + r_* B_{L_2}$). This argument cannot be used for Lipschitz loss function because there is no homogeneity between $\mathcal{L}_f$ and $\mathcal{L}_g$ when $f - f^* = \lambda(g-f^*)$ for some $\lambda > 0$. As a consequence, one needs to control the oscillation of the process $f \to (P_N - P)\mathcal{L}_f$ on every shelves $F_{i,j}$ in the Lipschitz case. This requires a (double) peeling argument. Peeling arguments* *(see [42]) work only if one can sum all the deviation probabilities of the events where the empirical process is controlled on each shell. This is an infinite sum and therefore, one needs enough concentration to make this sum converging. That is the reason why we assumed subgaussianity or boundedness in order to make the peeling argument work. The weak stochastic framework considered in [28, 27] would not be enough here because of this approach based on a peeling argument.*

10. Proof of the optimality results in Section 4.

10.1. *Proof of Theorem 4.3.* For the sake of simplicity, assume that $m \geq T$ so $\max(m, T) = m$. Fix $r \in \{1, \dots, T\}$. Fix $x > 0$ such that $\exp(x)/[1 + \exp(x)] \leq b$, we define the set of matrices

$$\mathcal{C}_x = \left\{ A \in \mathbb{R}^{m \times r} : \forall (p,q), A_{p,q} \in \{0, x\} \right\}$$

and

$$\mathcal{M}_x = \{ A \in \mathbb{R}^{m \times T} : A = (B| \dots |B|O), B \in \mathcal{C}_x \}$$

where the block B is repeated $\lfloor T/r \rfloor$ times (this construction is taken from [24]). Varshamov-Gilbert bound (Lemma 2.9 in [41]) implies that there is a finite subset $\mathcal{M}_x^0 \subset \mathcal{M}_x$ with $\mathrm{card}(\mathcal{M}_x^0) \geq 2^{rm/8} + 1$ with $0 \in \mathcal{M}_x^0$, and for any distinct $A, B \in \mathcal{M}_x^0$,

$$\|A - B\|_{S_2}^2 \geq \frac{mr\lfloor T/r \rfloor}{8} x^2 \geq \frac{mT}{16} x^2$$

and so

$$\frac{1}{mT} \|A - B\|_{S_2}^2 \geq \frac{x^2}{16}.$$

Then, for $M \in \mathcal{M}_x^0 \setminus \{0\}$,

$$\begin{aligned}
\mathcal{K}(\mathbb{P}_0, \mathbb{P}_M) &= \frac{n}{mT} \sum_{i=1}^{m} \sum_{j=1}^{T} \left[\frac{1}{2} \log \left(\frac{1 + \exp(M_{i,j})}{2 \exp(M_{i,j})} \right) + \frac{1}{2} \log \left(\frac{1 + \exp(M_{i,j})}{2} \right) \right] \\
&= \frac{n}{mT} \sum_{i=1}^{m} \sum_{j=1}^{T} \left[\log \left(\frac{1 + \exp(M_{i,j})}{2} \right) - \frac{1}{2} M_{i,j} \right] \\
&\leq n \left[\log \left(\frac{1 + \exp(x)}{2} \right) - \frac{1}{2} x \right] \\
&\leq c(b) n x^2
\end{aligned}$$

where $c(b) > 0$ is a constant that depends only on b. So:

$$\frac{1}{\mathrm{card}(\mathcal{M}_x^0) - 1} \sum_{A \in \mathcal{M}_x^0} \mathcal{K}(\mathbb{P}_0, \mathbb{P}_A) \leq c(b) n x^2 \leq c(b) \log(\mathrm{card}(\mathcal{M}_x^0) - 1)$$

as soon as we choose

$$x \leq \sqrt{\frac{\log(\mathrm{card}(\mathcal{M}_x^0) - 1)}{n}} \leq \sqrt{\frac{rm \log(2)}{8n}}$$

(note that the condition $n \geq rm \log(2)/(8b^2)$ implies that $\exp(x)/[1 + \exp(x)] \leq b$). Then, Theorem 2.5 in [41] leads to the existence of $\beta, c > 0$ such that

$$\inf_{\widehat{M}} \sup_{A \in \mathcal{M}_x^0} \mathbb{P}_A \left(\frac{1}{mT} \|\widehat{M} - A\|_{S_2}^2 \geq c \frac{mr}{N} \right) \geq \beta.$$

□

10.2. *Proof of Theorem 4.5.* For the sake of simplicity, assume that $m \geq T$ so $\max(m, T) = m$. Fix $r \in \{2, \dots, T\}$ and assume that $rT \leq N \leq mT$.

We recall that $\{E_{p,q} : 1 \leq p \leq m, 1 \leq p \leq T\}$ is the canonical basis of $\mathbb{R}^{m \times T}$. We consider the following "blocks of coordinates": for every $1 \leq k \leq r - 1$ and $1 \leq l \leq T$,

$$B_{kl} = \left\{ E_{p,l} : \frac{(k-1)mT}{N} + 1 \leq p < \frac{kmT}{N} + 1 \right\}$$

(note that $(r-1)mT/N + 1 \leq m$ when $rT \leq N \leq mT$). We also introduce the "blocks" of "remaining" coordinates:

$$B_0 = \left\{ E_{p,q} : \frac{(r-1)mT}{N} + 1 \leq p, 1 \leq q \leq T \right\}$$

For every $\sigma = (\sigma_{kl}) \in \{0, 1\}^{(r-1) \times T}$, we denote by $\mathbb{P}_\sigma$ the probability distribution of a pair (X, Y) taking its values in $\mathbb{R}^{m \times T} \times \{-1, 1\}$ where X is uniformly distributed over the basis $\{E_{p,q} : 1 \leq p \leq m, 1 \leq p \leq T\}$ and for every $(p, q) \in \{1, \dots, m\} \times \{1, \dots, T\}$,

$$\mathbb{P}_\sigma[Y = 1 | X = E_{p,q}] = \begin{cases} \sigma_{kl} & \text{if } E_{p,q} \in B_{kl} \\ 1 & \text{otherwise.} \end{cases}$$

We also introduce $\eta_\sigma(E_{p,q}) = \mathbb{E}[Y = 1 | X = E_{p,q}] = 2\mathbb{P}_\sigma[Y = 1 | X = E_{p,q}] - 1$. It follows from [46] that the Bayes rules minimizes the Hinge risk, that is $f_\sigma^* \in \mathrm{argmin}_f \, \mathbb{E}_\sigma(Y - f(X))_+$, where the minimum runs over all measurable functions and $\mathbb{E}_\sigma$ denotes the expectation w.r.t. (X,Y) when $(X,Y) \sim \mathbb{P}_\sigma$, is achieved by $f_\sigma^* = \operatorname{sgn}(\eta_\sigma(\cdot))$. Therefore, $f_\sigma^*(\cdot) = \left\langle M_\sigma^*, \cdot \right\rangle$ where for every $(p,q) \in \{1, \dots, m\} \times \{1, \dots, T\}$,

$$(M_\sigma^*)_{pq} = \begin{cases} 2\sigma_{kl} - 1 & \text{if } E_{p,q} \in B_{kl} \\ 1 & \text{otherwise.} \end{cases} = \eta_\sigma(E_{p,q}).$$

In particular, M_σ^* has a rank at most equal to r.

Let $\sigma = (\sigma_{p,q}), \sigma' = (\sigma'_{pq})$ be in $\{0,1\}^{(r-1)T}$. We denote by $\rho(\sigma, \sigma')$ the Hamming distance between σ and σ' (i.e. the number of times the coordinates of σ and σ' are different). We denote by $H(\mathbb{P}_\sigma, \mathbb{P}_{\sigma'})$ the Hellinger distance between the probability measures $\mathbb{P}_\sigma$ and $\mathbb{P}_{\sigma'}$. We have

$$H(\mathbb{P}_\sigma, \mathbb{P}_{\sigma'}) = \int \left(\sqrt{d\mathbb{P}_\sigma} - \sqrt{d\mathbb{P}_{\sigma'}} \right)^2 = \frac{2\rho(\sigma, \sigma')}{N}.$$

Then, if $\rho(\sigma, \sigma') = 1$, it follows that (cf. Section 2.4 in [41]),

$$\begin{aligned} H^2(\mathbb{P}_\sigma^{\otimes N}, \mathbb{P}_{\sigma'}^{\otimes N}) &= 2 \left(1 - \left(1 - \frac{H^2(\mathbb{P}_\sigma, \mathbb{P}_{\sigma'})}{2} \right)^N \right) \\ &= 2 \left(1 - \left(1 - \frac{1}{N} \right)^N \right) \leq 2(1 - e^{-2}) := \alpha. \end{aligned}$$

Now, it follows from Theorem 2.12 in [41], that

$$\inf_{\hat\sigma} \max_{\sigma \in \{0,1\}^{(r-1)T}} \mathbb{E}_\sigma^{\otimes N} \left\| \hat\sigma - \sigma \right\|_{l_1} \geq \frac{(r-1)T}{8} \left(1 - \sqrt{\alpha(1 - \alpha/4)} \right) \tag{52}$$

where the infimum $\inf_{\hat\sigma}$ runs over all measurable functions $\hat\sigma$ of the data $(X_i, Y_i)_{i=1}^N$ with values in $\mathbb{R}$ (note that Theorem 2.12 in [41] is stated for functions $\hat\sigma$ taking values in $\{0,1\}^{(r-1)T}$ but its is straightforward to extend this result to any $\hat\sigma$ valued in $\mathbb{R}$) and $\mathbb{E}_\sigma^{\otimes N}$ denotes the expectation w.r.t. those data distributed according to $\mathbb{P}_\sigma^{\otimes N}$.

Now, we lower bound the excess risk of any estimator. Let $\hat f$ be an estimator with values in $\mathbb{R}$. Using a truncation argument it is not hard to see that one can restrict the values of $\hat f$ to $[-1,1]$. In that case, We have

$$\begin{aligned} &\mathcal{E}_{hinge}(\hat f) = \mathbb{E}\left[|2\eta_\sigma(X) - 1| |\hat f(X) - f_\sigma^*(X)| \right] = \mathbb{E}|\hat f(X) - f_\sigma^*(X)| \\ &= \sum_{p,q} |\hat f(E_{p,q}) - f_\sigma^*(E_{p,q})| \mathbb{P}[X = E_{p,q}] \\ &\geq \sum_{kl} \frac{1}{mT} \sum_{E_{p,q} \in B_{kl}} |\hat f(E_{p,q}) - (2\sigma_{pq} - 1)| \geq \frac{2}{N} \sum_{kl} |\hat\sigma_{kl} - \sigma_{pq}| \end{aligned}$$

where $\hat{\sigma}_{kl}$ is the mean of $\{(\hat{f}(E_{p,q})+1)/2 : E_{p,q} \in B_{kl}\}$. Then we obtain,

$$\inf_{\hat{f}} \sup_{\sigma\in\{0,1\}^{(r-1)T}} \mathbb{E}_\sigma^{\otimes N} \mathcal{E}_{hinge}(\hat{f}) \geq \frac{2}{N} \inf_{\hat{\sigma}} \max_{\sigma\in\{0,1\}^{(r-1)T}} \mathbb{E}_\sigma^{\otimes N} \left\|\hat{\sigma}-\sigma\right\|_{l_1}$$

and, using (52), we get

$$\inf_{\hat{f}} \sup_{\sigma\in\{0,1\}^{(r-1)T}} \mathbb{E}_\sigma^{\otimes N} \mathcal{E}_{hinge}(\hat{f}) \geq c_0 \frac{rT}{N}$$

for $c_0 = \left(1-\sqrt{\alpha(1-\alpha/4)}\right)/4$.
□

11. Proofs of Section 8.

The proof of Proposition 8.1 may be found in several papers (cf., for instance, [2]). Let us recall this argument since we will be using it at a starting point to prove the Bernstein condition in the subgaussian case.

Proof of Proposition 8.1. The logistic risk of a function $f : \mathcal{X} \to \mathbb{R}$ can be written as $P\ell_f = \mathbb{E}[g(X, f(X))]$ where for all $x, a \in \mathbb{R}$, $g(x,a) := ((1+\eta(x))/2)\log(1+e^{-a}) + ((1-\eta(x))/2)\log(1+e^{a})$ and $\eta(x) = \mathbb{E}[Y|X = x]$ is the conditional expectation of Y given $X = x$.

Since f^* minimizes $f \to P\ell_f$ over the convex class F, one has by the first order condition that for every $f \in F$, $\mathbb{E}\partial_2 g(X, f^*(X))(f - f^*)(X) \geq 0$. Therefore, it follows from a second order Taylor expansion that the excess logistic loss of every $f \in F$ is such that

$$\begin{aligned} &\mathcal{E}_{logistic}(f) = P\mathcal{L}_f \\ (53)\qquad &\geq \mathbb{E}\left[(f(X)-f^*(X))^2 \int_0^1 (1-u)\delta(f^*(X)+u(f-f^*)(X))du\right] \end{aligned}$$

where $\delta(u) = \partial_2^2 g(x,u) = e^u/(1+e^u)^2$ for every $u \in \mathbb{R}$.

Since $|f^*(X)|, |f(X)| \leq b$ a.s. then for every $u \in [0,1]$, $|f^*(X) + u(f - f^*)(X)| \leq 2b$, a.s. and since $\delta(v) \geq \delta(2b) \geq \exp(-2b)/4$ for every $|v| \leq 2b$, it follows from (53) that $P\mathcal{L}_f \geq \delta(2b)\left\|f - f^*\right\|_{L_2}^2$. □

Proof of Proposition 8.2. Let $t^* \in RB_{l_2}$ be such that $f^* = \left\langle\cdot, t^*\right\rangle$, where f^* is an oracle in $F = \{\left\langle\cdot, t\right\rangle : t \in RB_{l_2}\}$ w.r.t. the logistic loss risk. Let $f = \left\langle\cdot, t\right\rangle \in F$ for some $t \in RB_{l_2}$. It follows from (53) that the excess logistic risk of f satisfies

$$P\mathcal{L}_f \geq \int_0^1 \mathbb{E}\left[\left\langle X, t^*-t\right\rangle^2 \delta\left(\left\langle X, t^*+u(t-t^*)\right\rangle\right)\right] du.$$

The result will follow if one proves that for every $t_0, t \in \mathbb{R}^d$,

$$\mathbb{E}\left[\left\langle X,t\right\rangle^2\delta\left(\left\langle X,t_0\right\rangle\right)\right] \geq \frac{\min\left(\pi, \pi^2\left(\|t_0\|_2\sqrt{2\pi+\|t_0\|_2^2}\right)^{-1}\right)}{\sqrt{2\pi+\|t_0\|_2^2}+(\pi-1)\|t_0\|_2}\frac{\|t\|_2^2}{8\sqrt{2\pi}}. \tag{54}$$

Let us now prove (54). We write $t = t_0^{\perp} + \lambda t_0$ where $t_0^{\perp}$ is a vector orthogonal to t_0 and $\lambda \in \mathbb{R}$. Since $\left\langle X, t_0^{\perp}\right\rangle$ and $\left\langle X, t_0\right\rangle$ are independent random variables, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\langle X,t\right\rangle^2\delta\left(\left\langle X,t_0\right\rangle\right)\right]\\
&= \mathbb{E}\left[\left\langle X,t_0^{\perp}\right\rangle^2\right]\mathbb{E}\left[\delta\left(\left\langle X,t_0\right\rangle\right)\right] + \lambda^2\mathbb{E}\left[\left\langle X,t_0\right\rangle^2\delta\left(\left\langle X,t_0\right\rangle\right)\right],\\
&= \left\|t_0^{\perp}\right\|_2^2\mathbb{E}\delta(\|t_0\|_2\, g) + \lambda^2\|t_0\|_2^2\,\mathbb{E}g^2\delta(\|t_0\|_2\, g)
\end{aligned}$$

where $g \sim \mathcal{N}(0,1)$ is standard Gaussian variable and we recall that $\delta(v) = e^v/(1+e^v)^2$ for all $v \in \mathbb{R}$. Now, it remains to lower bound $\mathbb{E}\delta(\sigma g)$ and $\mathbb{E}g^2\delta(\sigma g)$ for every $\sigma > 0$.

Since $\delta(v) \geq \exp(-|v|)/4$ for all $v \in \mathbb{R}$, one has for all $\sigma > 0$,

$$\mathbb{E}\delta(\sigma g) \geq \mathbb{E}\exp(-\sigma|g|)/4 = \exp(\sigma^2/2)\mathbb{P}[g \geq \sigma]/2$$

and

$$\begin{aligned}
&\mathbb{E}g^2\delta(\sigma g) \geq \mathbb{E}g^2\exp(-\sigma|g|)/4\\
&= (1/2)\exp(\sigma^2/2)\left[(1+\sigma^2)\mathbb{P}[g\geq\sigma] - \frac{\sigma\exp(-\sigma^2/2)}{\sqrt{2\pi}}\right].
\end{aligned}$$

Therefore, for $\sigma = \|t_0\|_2$,

$$\begin{aligned}
&\mathbb{E}\left[\left\langle X,t\right\rangle^2\delta\left(\left\langle X,t_0\right\rangle\right)\right] \geq \exp(\sigma^2/2)\mathbb{P}[g\geq\sigma]\left\|t_0^{\perp}\right\|_2^2\\
&+ 2\lambda^2\|t_0\|_2^2\exp(\sigma^2/2)\left[\left(1+\sigma^2\right)\mathbb{P}[g\geq\sigma] - \frac{\sigma\exp(-\sigma^2/2)}{\sqrt{2\pi}}\right]
\end{aligned}$$

and since $\|t\|_2^2 = \left\|t_0^{\perp}\right\|_2^2 + \lambda^2\|t_0\|_2^2$, one has,

$$\mathbb{E}\left[\left\langle X,t\right\rangle^2\delta\left(\left\langle X,t_0\right\rangle\right)\right] \geq \frac{\|t\|_2^2}{\sqrt{2\pi}}\min\left\{\left(\frac{1-\Phi(\sigma)}{\phi(\sigma)}\right), (1+\sigma^2)\left(\frac{1-\Phi(\sigma)}{\phi(\sigma)}\right) - \sigma\right\} \tag{55}$$

where ϕ and Φ denote the standard Gaussian density and distribution functions, respectively.

We lower bound the right-hand side of (55) using estimates on the Mills ratio $(1-\Phi)/\phi$ that follows from Equation (10) in [17]: for every $\sigma > 0$,

$$\frac{1-\Phi(\sigma)}{\phi(\sigma)} > \frac{\pi}{\sqrt{2\pi+\sigma^2}+(\pi-1)\sigma}.$$

□

Proof of Proposition 8.4. We follow a proof from [18]. We have

$$\begin{aligned} P\mathcal{L}_f &= \mathbb{E}[\rho_\tau(Y-f(X))-\rho_\tau(Y-f^*(X))] \\ &= \mathbb{E}\Big\{\mathbb{E}[\rho_\tau(Y-f(X))-\rho_\tau(Y-f^*(X))|X]\Big\}. \end{aligned}$$

For all $x \in \mathcal{X}$, denote by F_x the c.d.f. associated with f_x. We have

$$\begin{aligned} &\mathbb{E}[\rho_\tau(Y-f(X))|X=x] \\ &= (\tau-1)\int_{y<f(x)}(y-f(x))F_x(\mathrm{d}y)+\tau\int_{y\geq f(x)}(y-f(x))F_x(\mathrm{d}y) \\ &= \int_{y\geq f(x)}(y-f(x))F_x(\mathrm{d}y)+(\tau-1)\int_{\mathbb{R}}(y-f(x))F_x(\mathrm{d}y) \\ &= \int_{y\geq f(x)}(1-F_x(y))\mathrm{d}y+(\tau-1)\left(\int_{\mathbb{R}}yF_x(\mathrm{d}y)-f(x)\right) \\ &= g(x,f(x))+(\tau-1)\int_{\mathbb{R}}yF_x(\mathrm{d}y) \end{aligned}$$

where $g(x,a)=\int_{y\geq a}(1-F_x(y))\mathrm{d}y+(1-\tau)a$. Note that $\partial_2 g(x,f^*(x))=0$ (can be checked by calculations but also obvious from the definition). So

$$\begin{aligned} &\mathbb{E}[\rho_\tau(Y-f(X))-\rho_\tau(Y-f^*(X))|X=x] \\ &= g(x,f(x))-g(x,f^*(x))=\int_{f^*(x)}^{f(x)}(f(x)-u)\partial_2^2 g(x,u)\mathrm{d}u \\ &= \int_{f^*(x)}^{f(x)}(f(x)-u)f_x(u)\mathrm{d}u \geq \frac{1}{C}\int_{f^*(x)}^{f(x)}(f(x)-u)\mathrm{d}u = \frac{(f(x)-f^*(x))^2}{2C^2}. \end{aligned}$$

It follows that

$$\mathcal{E}_{quantile}(f)=P\mathcal{L}_f \geq \mathbb{E}\left\{\frac{(f(X)-f^*(X))^2}{2C}\right\}=\frac{1}{2C}\|f-f^*\|_{L_2}^2.$$

□

12. Empirical risk minimization with Lipschitz loss function. In this section, we consider the case where we have no a priori knowledge on the oracle f^* such as “low dimensional structure” or “smoothness”. In that case, there is no reason to force some structure on the estimators and therefore a natural procedure is the one minimizing the empirical risk itself (without regularization):

$$\tilde{f} \in \underset{f\in F}{\operatorname{argmin}} \left(\frac{1}{N} \sum_{i=1}^{N} \ell_f(X_i, Y_i) \right) \tag{56}$$

where ℓ is a Lipschitz loss function as considered in the previous sections and F is a convex class of functions.

The aim of this section is to obtain estimation and prediction bounds for $\tilde{f}$ in the two subgaussian and bounded settings as introduced in Section 2. As in the regularized case, the statistical performances of $\tilde{f}$ are driven by fixed points, one for each of the two setups.

Definition 12.1. *Let $\theta > 0$. The* ***complexity parameter in the subgaussian case*** *r_* is any point satisfying*

$$Lw\left(F \cap (f^* + r_* B_{L_2})\right) \leq \theta r_*^{2\kappa}\sqrt{N}$$

where κ is the Bernstein parameter from Assumption 2.1.

In the bounded case, it is written as follows.

Definition 12.2. *Let $\theta > 0$. The* ***complexity parameter in the bounded case*** *r^* is any point satisfying*

$$48\text{Rad}(F \cap (f^* + r_* B_{L_2})) \leq \theta r_*^{2\kappa}\sqrt{N}$$

where κ is the Bernstein parameter from Assumption 2.1.

Explicit computations of r_* in both subgaussian and bounded cases are available in the literature for various classes F (see, for instance, [19, 5, 26, 27, 32, 1, 6, 20]). An example in shape constrained regression is provided below. Let us now state the main results of this section (we treat both cases at the same time since they are identical in nature).

Theorem 12.1. *Assume that Assumption 1.1, Assumption 2.1 and Assumption 2.2 (resp. Assumption 2.3) hold. Let r_* be as in Definition 12.1 (resp. Definition 12.2) for some θ such that $A\theta < 1$. There exists a constant*

c_0 depending only on L (resp. on b) such that, with probability larger than $1-\exp(-c_0\theta^2 N r_*^{4\kappa-2})$, the ERM (56) $\tilde f$ satisfies

$$\left\|\tilde f - f^*\right\|_{L_2} \leq r_* \text{ and } \mathcal{E}(\tilde f) \leq \theta r_*^{2\kappa}.$$

Proof of Theorem 12.1: As in Section 9, we introduce an event (which is common to the subgaussian and the bounded setups) on which we can derive the statistical performances of $\tilde f$ using only deterministic arguments:

$$\Omega_0' := \left\{ \text{ for all } f\in F, \left|(P-P_N)\mathcal{L}_f\right| \leq \theta \max\left(r_*^{2\kappa}, \|f-f^*\|_{L_2}^{2\kappa}\right)\right\}$$

where θ is a parameter appearing in the definition of r_* in Definition 12.1 and Definition 12.2 and $\kappa\geq 1$ is the Bernstein parameter from Definition 2.1.

Let us place ourselves on the event Ω_0'. Let $f\in F$ be a function in F such that $\|f-f^*\|_{L_2} > r_*$. We have

$$P_N\mathcal{L}_f = (P_N-P)\mathcal{L}_f + P\mathcal{L}_f \geq -\theta\left\|f-f^*\right\|_{L_2}^{2\kappa} + (1/A)\left\|f-f^*\right\|_{L_2}^{2\kappa} > 0.$$

Since $P_N\mathcal{L}_{\tilde f}\leq 0$, we have $\left\|\tilde f - f^*\right\|_{L_2}\leq r_*$ on the event Ω_0'. Let us now prove the sharp oracle inequality. It follows from

$$\begin{aligned}0 &\geq P_N\mathcal{L}_{\tilde f} = (P_N-P)\mathcal{L}_{\tilde f} + P\mathcal{L}_{\tilde f}\\ &\geq -\theta\max\left(r_*^{2\kappa}, \left\|\tilde f - f^*\right\|_{L_2}^{2\kappa}\right) + P\mathcal{L}_{\tilde f} = -\theta r_*^{2\kappa} + P\mathcal{L}_{\tilde f}\end{aligned}$$

that $\theta r_*^{2\kappa}\geq P\mathcal{L}_{\tilde f}$.

Let us prove that Ω_0' holds at least with the exponential probability from Theorem 12.1. The proof uses a peeling argument of the class F along the shelves

$$F_k = \left\{f\in F: 2^{k-1}r_* < \|f-f^*\|_{L_2}\leq 2^k r_*\right\}$$

for all integer $k\geq 1$ and $F_0 = F\cap(f^*+r_*B_{L_2})$. The peeling argument used here is similar to the one from Proposition 9.2 and 9.3 but it is simpler since we do not have to peel simultaneously along the values of the distance to f^* of the regularization norm.

Let us first consider the subgaussian case. It follows from Lemma 9.1 that for all $k\in\mathbb{N}$ and $u>0$, with probability larger than $1-2\exp(-u^2)$,

$$\sup_{f,g\in F_k}\left|(P-P_N)(\mathcal{L}_f-\mathcal{L}_g)\right| \leq \frac{c_0 L}{\sqrt N}\left(w(F_k) + u d_{L_2}(F_k)\right)$$

where $d_{L_2}(F_k) \leq 2^{k+1} r_*$. Finally the result in the subgaussian case follows from a union bound and the same arguments as in the proof of Proposition 9.2. The result in the bounded case is similar except that we use Talagrand's concentration inequality instead of Lemma 9.1 and the arguments from the proof of Proposition 9.3.

As a proof of concept we apply Theorem 12.1 to the isotonic regression which is an example of a shape constrained regression problem [35, 12, 45, 13, 7, 23]. Unlike the majority of works on shape constrained regression, we consider the classification problem with $\{-1, 1\}$-valued outputs where one wants to fit a logistic function $f_t : x \in \mathbb{R}^p \to \sigma(\langle x, t \rangle)$ (where $\sigma(u) = e^u/(1+e^u), \forall u \in \mathbb{R}$) with a constraint on the shape of the weights t. Note that we do not assume that the log-odds ratio has a particular structure satisfying a shape constraint since we do not make any assumption of the distribution of $Y|X$ but we want to predict Y by the best logistic function f_{t^*} having weights t^* satisfying a constraint.

We consider the class of linear functionals $F = \big\{ f_t(\cdot) = \big\langle \cdot, t \big\rangle : t \in \mathcal{I} \big\}$ indexed by isotonic vectors:

$$\mathcal{I} = \{ t = (t_j)_1^p \in \mathbb{R}^p : t_1 \leq t_2 \leq \cdots \leq t_p \} . \tag{57}$$

We are given a dataset made of N i.i.d. pairs of random variables $(X_i, Y_i)_{i=1}^N$ distributed like (X, Y) where the outputs Y_i's take their values in $\{-1, 1\}$ and the X_i's take their values in $\mathbb{R}^p$. Let $R \geq 1$ and the ERM

$$\tilde{t} \in \underset{t \in \mathcal{I}, \|t\|_2 \leq R}{\operatorname{argmin}} \left(\frac{1}{N} \sum_{i=1}^N \log \big(1 + \exp \big(\big\langle X_i, t \big\rangle \big) \big) \right) . \tag{58}$$

Let us assume that the X_i's are distributed according to a standard Gaussian variable. In that case, it follows from Proposition 8.2 that the Bernstein condition is satisfied with parameter $\kappa = 1$ and the subgaussian Assumption 2.2 is satisfied. Therefore, in order to apply Theorem 12.1, it only remains to compute the complexity parameter r_* in the Gaussian case. To that end, we need to upper bound the local Gaussian mean width of $\mathcal{I}$. This follows from [1, (D.12)] that

$$w(\mathcal{I} \cap r B_{\ell_2}) \leq r \log(ed)$$

and so one can take $r_* = \log(ed)/(\theta\sqrt{N})$ in Definition 12.1 given that $\kappa = 1$. We can therefore apply Theorem 12.1 and obtain the following result.

THEOREM 12.2. *Assume that X is a standard Gaussian vector in $\mathbb{R}^p$. With probability larger than $1 - \mathbf{C}\exp\left(-\mathbf{C}\log^2(ep)\right)$, the isotonic logistic regression estimator $\tilde{t}$ defined in (58) satisfies*

$$\left\|\tilde{t} - t^*\right\|_2 \leq \frac{\log(ed)}{\theta\sqrt{N}} \text{ and } \mathcal{E}_{logistic}(\tilde{t}) = R(\tilde{t}) - R(t^*) \leq \frac{\log^2(ed)}{N}$$

where $t^* \in \mathrm{argmin}_{t\in\mathcal{I},\|t\|_2\leq R}\, \mathbb{E}\log(1+\exp(\langle X, t\rangle))$.

Note that other examples of applications of Theorem 12.1 in shape constrained regression can be obtained using the same strategy for other type of constraints such as convex regression or unimodal regression by using the results on the Gaussian mean widths from [6].

APPENDIX A: TECHNICAL LEMMAS

LEMMA A.1. *If $\rho \to r(2\rho)/\rho$ is non-increasing then $\rho \to \Delta(\rho)/\rho$ is non-decreasing.*

Proof. We have for all $\rho > 0$

$$\frac{\Delta(\rho)}{\rho} = \inf_{H\in S\cap(r(2\rho)/\rho)B_{L_2}} \sup_{G\in\partial\|\cdot\|(M^*)} \langle H, G\rangle.$$

The result follows since $\rho \to S \cap (r(2\rho)/\rho)B_{L_2}$ is non-increasing. □

LEMMA A.2. *Let $\rho > 0$. The function $h : r > 0 \to w(\rho B \cap rB_{L_2})/r$ is non-increasing.*

Proof. Let $r_1 \geq r_2$. By convexity of B and B_{L_2}, we have

$$(\rho B \cap r_1 B_{L_2})/r_1 = (\rho/r_1)B \cap B_{L_2} \subset (\rho/r_2)B \cap B_{L_2} = (\rho B \cap r_2 B_{L_2})/r_2. \tag{59}$$

□

REFERENCES

[1] Dennis Amelunxen, Martin Lotz, Michael B. McCoy, and Joel A. Tropp. Living on the edge: phase transitions in convex programs with random data. *Inf. Inference*, 3(3):224–294, 2014.

[2] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Large margin classifiers: Convex loss, low noise, and convergence rates. In *NIPS*, pages 1173–1180, 2003.

[3] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[4] Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probab. Theory Related Fields*, 135(3):311–334, 2006.

[5] Pierre Bellec. Localized gaussian width of m-convex hulls with applications to lasso and convex aggregation. Technical report, Rutgers University, 2017.

[6] Pierre Bellec. Sharp oracle inequalities for least squares estimators in shape restricted regression. Technical report, ENSAE, 2017. To appear in the annals of statistics.

[7] Pierre C. Bellec and Alexandre B. Tsybakov. Sharp oracle bounds for monotone and convex regression through aggregation. *J. Mach. Learn. Res.*, 16:1879–1892, 2015.

[8] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375, 2005.

[9] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[10] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[11] Djalil Chafaï, Olivier Guédon, Guillaume Lecué, and Alain Pajor. *Interactions between compressed sensing random matrices and high dimensional geometry*, volume 37 of *Panoramas et Synthèses [Panoramas and Syntheses]*. Société Mathématique de France, Paris, 2012.

[12] Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen. On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.*, 43(4):1774–1800, 2015.

[13] Sourav Chatterjee. A new perspective on least squares under convex constraint. *Ann. Statist.*, 42(6):2340–2381, 2014.

[14] V. Cottet and P. Alquier. 1-bit Matrix Completion: PAC-Bayesian Analysis of a Variational Approximation. preprint arXiv:1604.04191, to appear in Machine Learning, 2016.

[15] Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1):1–49 (electronic), 2002.

[16] Mark A Davenport, Yaniv Plan, Ewout van den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, 2014.

[17] Lutz Dümbgen. Bounding standard gaussian tail probabilities. Technical report, University of Bern, 2010.

[18] Andreas Elsener and Sara van de Geer. Robust low-rank matrix estimation. *arXiv preprint arXiv:1603.09071*, 2016.

[19] Y. Gordon, A. E. Litvak, S. Mendelson, and A. Pajor. Gaussian averages of interpolated bodies and applications to approximate reconstruction. *J. Approx. Theory*, 149(1):59–73, 2007.

[20] Adityanand Guntuboyina and Bodhisattva Sen. Global risk bounds and adaptation in univariate convex regression. *Probab. Theory Related Fields*, 163(1-2):379–411, 2015.

[21] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[22] Cho-Jui Hsieh and Peder A Olsen. Nuclear norm minimization via active subspace selection. In *ICML*, pages 575–583, 2014.

[23] Arlene K. H. Kim and Richard J. Samworth. Global rates of convergence in log-concave density estimation. *Ann. Statist.*, 44(6):2756–2779, 2016.

[24] Vladimir Koltchinskii, Karim Lounici, Alexandre B Tsybakov, et al. Nuclear-norm

penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

[25] Guillaume Lecué. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13(4):1000–1022, 2007.

[26] Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes: Upper and minimax bounds. Technical report, CNRS, Ecole polytechnique and Technion - to appear in Topics in Learning Theory - Societe Mathématique de France, (S. Boucheron and N. Vayatis Eds.), 2013.

[27] Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method I: sparse recovery. Technical report, CNRS, Ecole Polytechnique and Technion - to appear in the Annals of Statistics, 2015.

[28] Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method II: complexity dependent error rates. Technical report, CNRS, Ecole Polytechnique and Technion - to appear in Journal of Machine Learnint Research, 2015.

[29] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. Isoperimetry and processes.

[30] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322, 2010.

[31] Shahar Mendelson. Improving the sample complexity using global data. *IEEE transactions on Information Theory*, 48(7):1977–1991, 2002.

[32] Shahar Mendelson. On the performance of kernel classes. *J. Mach. Learn. Res.*, 4(5):759–771, 2004.

[33] Shahar Mendelson. Obtaining fast error rates in nonconvex situations. *J. Complexity*, 24(3):380–397, 2008.

[34] Shahar Mendelson and Joseph Neeman. Regularization in kernel learning. *Ann. Statist.*, 38(1):526–565, 2010.

[35] Mary Meyer and Michael Woodroofe. On the degrees of freedom in shape-restricted regression. *Ann. Statist.*, 28(4):1083–1104, 2000.

[36] M. M. Rao and Z. D. Ren. *Theory of Orlicz spaces*, volume 146 of *Monographs and Textbooks in Pure and Applied Mathematics*. Marcel Dekker, Inc., New York, 1991.

[37] M. M. Rao and Z. D. Ren. *Applications of Orlicz spaces*, volume 250 of *Monographs and Textbooks in Pure and Applied Mathematics*. Marcel Dekker, Inc., New York, 2002.

[38] Ingo Steinwart and Andreas Christmann. *Support vector machines.* Information Science and Statistics. Springer, New York, 2008.

[39] Michel Talagrand. *The generic chaining.* Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2005. Upper and lower bounds of stochastic processes.

[40] Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004.

[41] Alexandre B Tsybakov. *Introduction to nonparametric estimation.* Springer Series in Statistics, 2009.

[42] Sara A. van de Geer. *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press, Cambridge, 2000.

[43] Sara A Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, pages 614–645, 2008.

[44] Vladimir N. Vapnik. *Statistical learning theory.* Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York,

1998. A Wiley-Interscience Publication.

[45] Cun-Hui Zhang. Risk bounds in isotonic regression. *Ann. Statist.*, 30(2):528–555, 2002.

[46] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pages 56–85, 2004.

ENSAE, 3, avenue Pierre Larousse,
92245 MALAKOFF. France.