Robust learning and complexity dependent bounds for regularized problems

Geoffrey Chinot

February 7, 2019

Abstract

We obtain risk bounds for Regularized Empirical Risk Minimizers (RERM) and minmax Median-Of-Means (MOM) estimators where the regularization function $\phi(\cdot)$ is not necessary a norm. It covers for example the Support Vector Machine (SVM) and the Elastic Net procedures. We obtain bounds on the $L_2$ estimation error rate that depend on the complexity of the true model $F^* = \{ f \in F : \phi(f) \leq \phi(f^*) \}$, where $f^*$ is the minimizer of the risk over the class $F$. The estimators are based on loss functions that are both Lipschitz and convex. Results for the RERM are derived without assumptions on the outputs and under subgaussian assumptions on the design. Similar results are shown for minmax MOM estimators in a close setting where outliers may be present in the dataset and where the design is only supposed to satisfy moment assumptions, relaxing the subgaussian and the i.i.d hypothesis necessary for RERM. Unlike alternatives based on MOM’s principle, the analysis of minmax MOM estimators is not based on the small ball assumption (SBA) of [7] but on a weak local Bernstein Assumption.

1 Introduction

Let $\mathcal{X}, \mathcal{Y}$ be two measurable spaces such that $\mathcal{Y} \subset \mathbb{R}$ and $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ be random variables with joint distribution $P$. Let $\mu$ be the marginal distribution of $X$. For $E$ a linear subset of $L_2(X)$, let $F \subset E$ be a class of measurable functions $f: \mathcal{X} \to \mathcal{Y}$ where $\mathcal{Y} \subset \mathbb{R}$ is convex (we do not have necessarily $\mathcal{Y} = \mathcal{Y}$, see below). In the standard learning framework, one would like to identify the best approximation to $Y$ using functions $f$ in the class $F$. To do so, let $\ell$ be a loss function, $\ell: F \times X \times Y \to \mathbb{R}$, $(f, x, y) \mapsto \ell_f(x, y) = \bar{\ell}(f(x), y)$ measuring the error made when predicting $y$ by $f(x)$, for $\bar{\ell}: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Let $f^* \in \arg\min_{f \in F} R(f)$ where $R(f) := P_{\ell_f} := E_{P}[\ell_f(X, Y)]$. Without loss of generality we can assume that the set of risk minimizers is reduced to a singleton. In this case $f^*$ is called the oracle. It provides the prediction of $Y$ with minimal risk among functions in $F$. Obviously, the distribution $P$ is unknown and minimizing the risk $R(f)$ over $f$ in $F$ is impossible in practice. Instead one is given a dataset $D = (X_i, Y_i)_{i=1}^N$ of random variables taking values in $\mathcal{X} \times \mathcal{Y}$. Using $D$, the objective is to construct an estimator $\hat{f}_N$ such that, with high probability (with respect to $D$), the error rate

$$|| \hat{f}_N - f^* ||_{L_2(\mu)}^2 \leq \mathbb{E} \left[ (\hat{f}_N(X) - f^*(X))^2 | D \right]$$

and the excess of risk

$$PL_{\hat{f}_N} := P\ell_{\hat{f}_N} - P\ell_{f^*} = \mathbb{E}_P \left[ \bar{\ell}(\hat{f}_N(X), Y) - \bar{\ell}(f^*(X), Y) \right]$$

$PL_{\hat{f}_N}$ specifies the quality of prediction of the estimator $\hat{f}_N$ when $|| \hat{f}_N - f^* ||_{L_2(\mu)}$ quantifies the $L_2(\mu)$ approximation of the oracle $f^*$ by the estimator $\hat{f}_N$.

In this paper we will consider loss functions being simultaneously Lipschitz and convex.

**Assumption 1.** There exists $L > 0$ such that, for any $y \in \mathcal{Y}$, $\bar{\ell}( \cdot, y )$ is $L$-Lipschitz i.e for every $f$ and $g$ in $F$, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, $| \bar{\ell}(f(x), y) - \bar{\ell}(g(x), y) | \leq L|f(x) - g(x)|$ and convex i.e for all $\alpha \in [0, 1]$, $\bar{\ell}(\alpha f(x) + (1-\alpha)g(x), y) \leq \alpha \bar{\ell}(f(x), y) + (1-\alpha)\bar{\ell}(g(x), y)$
Many classical loss functions satisfy Assumption 1 as shown by the following examples.

- The **logistic loss** defined, for any \( u \in \mathcal{Y} = \mathbb{R} \) and \( y \in \mathcal{Y} = \{-1, 1\} \), by \( \ell(u, y) = \log(1 + \exp(-uy)) \) satisfies Assumption 1 with \( L = 1 \).

- The **hinge loss** defined, for any \( u \in \mathcal{Y} = \mathbb{R} \) and \( y \in \mathcal{Y} = \{-1, 1\} \), by \( \ell(u, y) = \max(1 - uy, 0) \) satisfies Assumption 1 with \( L = 1 \).

Note that in those two examples the set \( \mathcal{Y} \) and \( \mathcal{Y} \) are different. The fact that every functions \( f \) in \( F \) map to the convex set \( \mathcal{Y} \) is crucial for the computation of the estimator \( f_N \) in practice.

- The **Huber loss** defined, for any \( \delta > 0 \), \( u, y \in \mathcal{Y} = \mathcal{Y} = \mathbb{R} \), by

  \[
  \ell(u, y) = \begin{cases} \frac{1}{2}(y - u)^2 & \text{if } |u - y| \leq \delta \\ \delta|y - u| - \delta^2 & \text{if } |u - y| > \delta 
  \end{cases},
  \]

  satisfies Assumption 1 with \( L = \delta \).

- The **quantile loss** is defined, for any \( \tau \in (0, 1) \), \( u, y \in \mathcal{Y} = \mathcal{Y} = \mathbb{R} \), by \( \ell(u, y) = \rho_\tau(u - y) \) where, for any \( z \in \mathbb{R} \), \( \rho_\tau(z) = z(\tau - I\{z \leq 0\}) \). It satisfies Assumption 1 with \( L = 1 \). For \( \tau = 1/2 \), the quantile loss is the \( L_1 \) loss.

All along the paper, the following geometric Assumption is also granted.

**Assumption 2.** The class \( F \) is convex.

For example Assumption 2 holds when \( F \) is an Hilbert space or the set of linear functionals in \( \mathbb{R}^p \), \( F = \{\langle t, \cdot \rangle : t \in \mathbb{R}^p\} \).

The paper focuses on robustness with respect to outliers in the dataset and heavy-tailed data in learning theory. First we will present results for regularized empirical risk minimizers which are robust with respect to the noise of the problem. It serves as a benchmark for more advanced estimators. Then we will study minmax MOM estimators. Such estimators turn out to be robust 1) with respect to the noise, 2) with respect to heavy-tailed data and 3) with respect to possible outliers in the dataset.

**Notations:** In the remaining of the paper, the following notations will be used repeatedly: We will write \( L_2 \) instead of \( L_2(\mu) \), let \( r > 0 \),

\[
rB_{L_2} = \{f \in F : \|f(X)\|_{L_2(\mu)} \leq r\}, \quad rS_{L_2} = \{f \in F : \|f(X)\|_{L_2(\mu)} = r\}.
\]

For any set \( H \) for which it makes sense, let \( H + f^* = \{h + f^* \text{ s.t } h \in H\} \), \( H - f^* = \{h - f^* \text{ s.t } h \in H\} \). The notations \( a \vee b \) and \( a \wedge b \), will denote respectively \( \max(a, b) \) and \( \min(a, b) \). Let \( C > 0 \) denote an absolute constant whose value might change from line to line, we will write \( C(A) \) if the constant depends on the parameter \( A \).

## 2 Regularized Empirical Risk Minimization (RERM)

All along this section, data \( (X_i, Y_i)_{i=1}^N \) are independent and identically distributed with common distribution \( P \). Since the risk is unknown, a simple and first approach is to estimate the risk by its empirical counterpart and minimize it over \( F \). It leads to empirical risk minimizer (ERM) (see [6]). Let

\[
\hat{f}^{ERM} = \arg \min_{f \in F} PN \ell_f
\]

Clearly, if the class \( F \) is too small there is no hope that \( f^*(X) \) is close to \( Y \). One has to consider large classes leading to large error rates. To bypass the fact that \( F \) may be very large, we can use the classical approach of regularization
where the penализation functions emphasizes the believe we may have on the oracle $f^*$. It leads to the Regularized Empirical Risk Minimizer (RERM) defined as

$$ f^\text{RERM}_\lambda = \arg \min_{f \in F} P_N \ell_f + \lambda \| f \| $$

where $\| \cdot \| : E \to \mathbb{R}^+$ is a norm. Those methods of regularization have been used to "smooth" the estimators that have poor generalization capabilities. It decreases the over-fitting phenomenon (see [5] for many examples). This use of regularization corresponds to the "classical" approach of regularization. In "modern" statistics the aim is somehow different. One uses the penalty to enhance an hidden property of $f^*$. In this "modern" approach, the error rates depend on this underlying structure. However, the estimator $f^\text{RERM}_\lambda$ defined in (1) is rather restrictive since it does not cover penализations which are not a norm such as $\| f \|_{\mathcal{H}^2}$ (i.e the square of the norm in a reproducible Kernel Hilbert space) or even the Elastic net (see [24]). To bypass this limitation, the estimator defined in Equation (1) will be replaced by

$$ f^\phi = \arg \min_{f \in F} P_N \ell_f + \lambda \phi(f) $$

where $\phi : E \to \mathbb{R}^+$ is a function verifying the following Assumption

**Assumption 3.**

- $\phi$ is non negative, even, convex and $\phi(0) = 0$
- There exists a constant $\eta > 0$ such that for all $f, g \in F$, 
  $$ \phi(f + g) \leq \eta (\phi(f) + \phi(g)) $$

Note that Assumption 3 holds for any norm.

As surprising as it might seem and as far as we know there exists almost no general results when the regularization function is not a norm. The only article is [9], where they consider the quadratic loss function and suppose that the Small Ball Assumption (SBA) is verified. In this article, we obtain complexity dependent bounds on the $L_2$ error rate, i.e bounds depending on the complexity of the true model $F^* = \{ f \in F : \phi(f) \leq \phi(f^*) \}$ for any Lipschitz and convex loss function under a local Bernstein Assumption which is weaker than the SBA (see discussion in [4] for instance).

To control the $L_2$ error rates for the RERM, it is necessary to put a concentration Assumption on the class $F$ (this Assumption will be relaxed using MOM type estimators in Section 3).

**Definition 1.** A class $F$ is called $B$-subgaussian (with respect to $X$) for some constant $B \geq 1$ when for all $f \in F$ and for all $\lambda > 1$

$$ \mathbb{E}\exp(\lambda |f(X)|/\|f\|_{L_2}) \leq \exp(\lambda^2 B^2/2) $$

**Assumption 4.** The class $F - f^*$ is $B$-subgaussian.

Assumption 4 holds for instance when the class $F$ is bounded. When $F$ is a class of linear functionals $F = \{ \langle \cdot, t \rangle, \ t \in T \}$ for $T \subset \mathbb{R}^p$ and $X$ is a random variable in $\mathbb{R}^p$ then $F - f^*$ is $B$-subgaussian if $X$ is a Gaussian vectors in $\mathbb{R}^p$ or if $X = (x_j)_{j=1}^p$ has independent coordinates that are subgaussian. In the subgaussian framework, a natural way to measure the statistical complexity of the function class $F$ is via the Gaussian mean-width that we introduce now.

**Definition 2.** Let $H \subset L_2$ and $(G_h)_{h \in H}$ be the canonical centered Gaussian process indexed by $H$, with covariance structure

$$ \forall h_1, h_2 \in H, \quad (\mathbb{E}(G_{h_1} - G_{h_2})^2)^{1/2} = (\mathbb{E}(h_1(X) - h_2(X))^2)^{1/2} $$

The Gaussian mean-width of $H$ is $w(H) = \mathbb{E}\sup_{h \in H} G_h$.

The complexity parameter driving the statistical behavior of the estimator $f^\phi$ is presented in the following definition.
**Definition 3.** The complexity is measured via a non-decreasing function \( r(\cdot) \) such that for every \( A > 0 \),

\[
r(A) = \inf \left\{ r > 0 : 32Lw(F \cap B^\phi_{\eta(\lambda(4+2\lambda^{-1})\phi(f^*))}^\phi(0) \cap r B_{L^2} \leq (2A)^{-1} \sqrt{N}r^2 \right\}
\]

where \( B^\phi_{\delta}(g) = \{ f \in F : \phi(f - g) \leq \delta \} \), \( L \) is the Lipschitz constant of Assumption 1 and \( \eta \) is defined in Assumption 3.

Note that when \( \phi \) is a norm, \( B^\phi_{\delta}(g) \) simply corresponds to the ball of regularization centered in \( f^* \) with radius \( \delta \).

We are now in position to introduce the local Bernstein condition.

**Assumption 5.** There exists a constant \( A^* > 0 \) such that for all \( f \in F \) if \( \| f - f^* \|_{L_2} \leq r(A^*) \) then \( \| f - f^* \|^2_{L_2} \leq A^*P L_f \).

In the sequel of this section we will write \( r^* \) for \( r(A^*) \).

Roughly, this condition says that the variance of the problem is not too large in a neighborhood of the oracle \( f^* \).

As explained in [4], this local Bernstein condition holds in examples where \( F \) is not bounded in \( L_2 \)-norm, and therefore, where the global Bernstein condition of [1] \( \| f - f^* \|^2_{L_2} \leq A^*P L_f \) for all \( f \in F \) does not hold.

Somehow, Assumption 5 replaces the small-ball Assumption (see [16] for instance) when learning problems with a Lipschitz and convex loss function are considered. However our condition is local which is much weaker. It allows to cover well known examples where the small ball is not verified (see [18] for different examples).

We are now in position to present the main Theorems of this section.

**Theorem 1.** Grant Assumptions 1, 2, 3, 4 and 5, with probability larger than

\[
1 - 2 \exp \left( - C(A^*, L, \eta) N (r^*)^2 \right)
\]

for all regularization parameters \( \lambda \geq \lambda_0 = (r^*)^2/\phi(f^*) \) we have

\[
\| \hat f^\phi - f^* \|_{L_2} \leq C(A^*) \lambda^{\phi(f^*)/r^*} \quad \text{and} \quad \phi(\hat f^\phi - f^*) \leq C(A^*, \eta) \phi(f^*).
\]

The explicit constants can be found in the proof of the Theorem (see Section A).

**Remark 1.** Theorem 1 holds for an exponentially large probability (4) simultaneously for all \( \lambda \geq \lambda_0 \). As a consequence it can be used with a random choice of regularization parameter \( \lambda \) as long as \( \{ \lambda \geq \lambda_0 \} \) hold with large probability. For example, we could use a cross validation scheme to generate \( \lambda \).

Note that for \( \lambda = \lambda_0 \) we obtain \( \| \hat f^\phi - f^* \|_{L_2} \leq C(A^*) r^* \) known to be minimax into the class\( \{ f \in F : \phi(f) \leq \phi(f^*) \} \) (see [9]). Since we do not have access to \( \phi(f^*) \), taking \( \lambda_0 \) is impossible. To bypass this issue we use a Lepski’s adaption method (see [11, 12, 3]). To do so the following Assumption is required

**Assumption 6.** There exists \( M > 0 \) such that \( \phi(f^*) \leq M \).

Assumption 6 is natural since regularization procedures are used when one believes that \( \phi(f^*) \) is small. Since Theorem 1 holds with the same probability for all \( \lambda \geq \lambda_0 \) one can choose \( M \) very large in the Lepski’s method without deteriorating the probability of the event.

For \( j = 1, \cdots, J = M + \lceil \log_2(M) \rceil \) let us define \( \phi_j = 2^j/2^M, \phi_0 = 0 \) and \( \lambda_j = r^*_j/\phi_j \) where

\[
r_j = \inf \left\{ r > 0 : 32Lw(F \cap B^\phi_{\eta(\lambda(4+2\lambda^{-1})\phi_j)}^\phi(0) \cap r B_{L^2} \leq (2A^*)^{-1} \sqrt{N}r^2 \right\}
\]

Moreover for all \( \lambda > 0 \) let us define

\[
T_\lambda(f) = P_N(\ell_f - \ell_\lambda^\phi) + \lambda(\phi(f) - \phi(\hat f^\phi_j)), \quad \hat R_j = \{ f \in F : T_\lambda_j(f) \leq ((A^*)^{-1} + 2) \lambda_j \phi_j \}
\]

\[
k^* = \inf\{ k \in \{1, \cdots, J \} : \cap_{j \geq k} \hat R_j \neq \emptyset \} \quad \text{and set} \quad \hat f \in \cap_{j \geq k^*} \hat R_j .
\]

Using the Lepski’s method we are in position to state to following Theorem.
Theorem 2. Assumptions 1, 2, 3, 4, 5 and 6, with probability larger than

\[ 1 - 2 \exp \left( -C(A^*, L, \eta) N(r^*)^2 \right) \]

\[ \| \hat{f} - f^* \|_{L_2} \leq C(A^*) r^* \quad \phi(\hat{f} - f^*) \leq C(A^*, \eta) \phi(f^*) \quad \text{and} \quad PL_{\hat{f}} \leq C(A^*)(r^*)^2. \]

Note that such a procedure required the knowledge of \( A^* \) and \( M \). Complete proofs of Theorem 2 and Theorem 1 are presented in Section A in the Appendix.

Here we present a simple sketch of the proof of Theorem 1.

Sketch of the proof: The main arguments are presented up to some constants depending on \( A^* \), \( L \) and \( \eta \). The proof is split into two parts. First we identify a random event onto which the statistical behavior of \( f^*_\Lambda \) can be studied using deterministic arguments. Secondly we prove that this event holds with large probability. Here we will only focus on the deterministic argument (see Section A for the stochastic control).

Let \( B_\Lambda = \{ f \in E : \| f - f^* \|_{L_2} \leq \lambda \phi(f^*)/r^* \text{ and } \phi(f - f^*) \leq \phi(f^*) \} \) and the stochastic event is defined as

\[ \Omega := \left\{ \text{for all } f \in F \cap (f^* + r^* B_{L_2}) \cap B_{\phi(f^*)}^\Lambda \right\} \]

By definition, the estimator \( \hat{f}_\Lambda \) satisfies \( P_N L_{\hat{f}_\Lambda} \leq 0 \). Therefore, to prove Theorem 1 it is sufficient to show that on \( \Omega \), \( P_N L_{\hat{f}_\Lambda} > 0 \) for all function \( f \) in \( F \setminus B_\Lambda \). It turns out that, up to the choice of the constants, that is equivalent to show the following Lemma

**Lemma 1.** Let \( \lambda \geq (r^*)^2/\phi(f^*) \), on the event \( \Omega \) we have

- For all \( f \in F \setminus B_\Lambda \), \( P_N L_{\hat{f}} > \lambda \phi(f^*) \)
- For all \( f \in F \cap B_\Lambda \), \( P_N L_{\hat{f}} \geq -\lambda \phi(f^*) \)

Lemma 1 is explicitly used in the proof of Theorem 2. From Lemma 1 it follows immediately that on \( \Omega \) one has \( \phi(f^*_\Lambda - f^*) \leq \phi(f^*) \) and \( \| f^*_\Lambda - f^* \|_{L_2} \leq \lambda \phi(f^*)/r^* \). The proof of Lemma 1 follows from an homogeneity argument saying that for all functions \( f \in F \setminus B_\Lambda \) there exist \( f_0 \) in the border of \( B_\Lambda \) and \( \alpha \geq 1 \) such that \( P_N L_{\hat{f}_\Lambda} \geq \alpha P_N L_{\hat{f}_0} \).

On the border of \( B_\Lambda \), either we have 1) \( \phi(f_0 - f^*) = \phi(f^*) \) and \( \| f_0 - f^* \|_{L_2} \leq \lambda \phi(f^*)/r^* \) or 2) \( \| f_0 - f^* \|_{L_2} = \lambda \phi(f^*)/r^* \) and \( \phi(f_0 - f^*) \leq \phi(f^*) \).

The homogeneity argument linking the empirical excess risk of \( f \) to the one of \( f_0 \) is the following. For all \( i \in \{1, \cdots, N\} \), let \( \psi_i : \mathbb{R} \to \mathbb{R} \) be defined for all \( u \in \mathbb{R} \) by

\[ \psi_i(u) = \bar{\ell}(u + f^*(X_i), Y_i) - \bar{\ell}(f^*(X_i), Y_i). \]  

(5)

The functions \( \psi_i \) are such that \( \psi_i(0) = 0 \), they are convex because \( \bar{\ell} \) is, in particular \( \alpha \psi_i(u) \leq \psi_i(\alpha u) \) for all \( u \in \mathbb{R} \) and \( \alpha \geq 1 \) and \( \psi_i(f(X_i) - f^*(X_i)) = \bar{\ell}(f(X_i), Y_i) - \bar{\ell}(f^*(X_i), Y_i) \) so that the following holds:

\[ P_N L_{\hat{f}} = \frac{\alpha}{N} \sum_{i=1}^{N} \psi_i(f(X_i) - f^*(X_i)) \leq \frac{\alpha}{N} \sum_{i=1}^{N} \psi_i(\alpha(f_0(X_i) - f^*(X_i))) \]

\[ \geq \frac{\alpha}{N} \sum_{i=1}^{N} \psi_i((f_0(X_i) - f^*(X_i))) = \alpha P_N L_{\hat{f}_0}. \]  

(6)

For the regularization part, since \( \alpha \geq 1 \), the same homogeneity arguments holds.

\[ \phi(f) - \phi(f^*) = \phi(f^* + \alpha(f_0 - f^*)) - \phi(f^*) \geq \alpha(\phi(f_0) - \phi(f^*)) \]

It remains to control \( P_N L_{\hat{f}_0} \) in the two cases 1) and 2). Up to technicalities in the first case 1) we use Assumption 3 to show that \( \phi(f_0) - \phi(f^*) \geq \phi(f^*) \) (up to constants). Using the event \( \Omega \) we show that \( P_N L_{\hat{f}_0} \geq -\theta \lambda \phi(f^*) \) for a well chosen constant \( \theta > 0 \). In the second case 2) we use the fact that \( \phi(f_0) - \phi(f^*) \geq -\phi(f^*) \) and the local Bernstein Assumption 5 to prove that \( P_N L_{\hat{f}_0} \geq \gamma \lambda \phi(f^*) \) for another well chosen constant \( \gamma > 0 \).
3 Robustness to outliers and heavy-tailed data via Minmax MOM estimators

In section 2 we assumed that the class $F - f^*$ is subgaussian and that the data $(X_i, Y_i)_{i=1}^N$ are i.i.d with the same distribution $P$. In this section we relax those two Assumptions using MOM type estimators. Let $P_i$ be the distribution of $(X_i, Y_i)$. To highlight the robustness of those estimators with respect to possible outliers in the dataset, we present here the $\mathcal{I} \cup \mathcal{O}$ framework. Let $\mathcal{I} \cup \mathcal{O}$ denote a partition of $\{1, \ldots, N\}$. The cardinality of $\mathcal{O}$ is denoted $|\mathcal{O}|$. Data $(X_i, Y_i)_{i \in \mathcal{O}}$ are considered as outliers. No assumption on the distribution $P_i$ for $i \in \mathcal{O}$ is made. For instance those data can be dependent and even adversarial data. The random variables $(X_i, Y_i)_{i \in \mathcal{I}}$ are the informative data. This is only on these data that assumptions will be made. Of course no one knows in advance which data is informative or not. In other words, the partition $\mathcal{I} \cup \mathcal{O}$ is unknown.

In the sequel we will need the following Assumption

**Assumption 7.** For all $i \in \mathcal{I}$: $P_i((f - f^*)^2(X_i) = P((f - f^*)^2(X) and P_i\mathcal{L}_f = P\mathcal{L}_f$. 

Assumption 7 holds in the i.i.d framework but it covers other situations such as when informative data $(X_i, Y_i)_{i \in \mathcal{I}}$ may not have the same distribution. It is only required to induce the same $L_2$-structure on the class $F$ and the same risk, which is a minimal Assumption for the problem to make sense.

Let $(B_k)_{k=1,\ldots,K}$ denote a partition of $\{1, \ldots, N\}$ into blocks $B_k$ of equal size $N/K$ (if $N$ is not a multiple of $K$, just remove some data). Following [8] the minmax MOM estimators are defined as

$$\hat{f}_K^\lambda = \arg\min_{f \in F} \sup_{g \in F} MOM_K(\ell_f - \ell_g) + \lambda(\phi(f) - \phi(g)).$$

(7)

where

$$MOM_K(f) = Med(P_{B_1}\ell_f, \ldots, P_{B_K}\ell_f), \text{ with } P_{B_k}\ell_f = \frac{1}{|B_k|} \sum_{i \in B_k} \ell_f(X_i, Y_i).$$

Since we no longer consider the subgaussian-framework, we have to adapt the complexity parameter to this new setup. The complexity is measured via a function $\tilde{r}(\cdot)$ defined as

$$\tilde{r}(A) = \inf \left\{ r > 0 : \forall J \subseteq \mathcal{I} : |J| \geq N/2, \sup_{f \in F \cap (f^* + r B_L)} \sup_{\sigma \in \Omega(4+2A-1)} \left| \sum_{i \in J} \sigma_i(f_f^*(X_i)) \right| \leq (384A L)^{-1} r |J| \right\}.\tag{8}$$

This complexity function is very close to the one in the subgaussian case from Section 2 expect that the Rademacher-complexity replaces the Gaussian mean-width. It is also necessary to adapt the local Bernstein condition from Assumption 5 to the MOM framework

**Assumption 8.** There exists a constant $\hat{A} > 0$ such that, for all $f$ in $F$ satisfying $\|f - f^*\|_{L_2} \leq \sqrt{C_{K,r}(\hat{A})}$, then

$$\|f - f^*\|_{L_2}^2 \leq \hat{A}P\mathcal{L}_f$$

where

$$C_{K,r}(A) = \max \left( \tilde{r}^2(A), 368A^2 L^2 K N \right).\tag{9}$$

As Assumption 5, Assumption 8 is only granted in a neighborhood of the oracle $f^*$ whose radius is proportional to the rate of convergence of the estimators. We are now in position to state our main results for the minmax MOM estimators.

**Theorem 3.** Grant Assumptions 1, 2, 3, 7 and 8. Let $K \geq 7|\mathcal{O}|/3$. Then, with probability larger than $1 - 2\exp(-CK)$, for all regularization parameter $\lambda > C_{K,r}(\hat{A})/\phi(f^*)$

$$\phi(\hat{f}_K^\lambda - f^*) \leq C(\hat{A}, \eta)\phi(f^*), \quad \|\hat{f}_K^\lambda - f^*\|_{L_2} \leq C(\hat{A})\lambda \phi(f^*)$$
It is also possible to use the Lepski’s method to get an adaptive estimator as the one in Theorm 2. For the sake of brevity we do not present the result here. Note that there is a tradeoff between confidence and accuracy and that an optimal choice of $K$ would be $K \propto \tilde{r}(\tilde{A})N$. In that case $C_{K,r}(\tilde{A}) = \tilde{r}(\tilde{A})$. For this value of $K$ the optimal $\lambda$ is $\tilde{r}^2(\tilde{A})/\phi(f^*)$ and we would obtain $\|f^*_K - f^*\|_{L_2}^2 \leq C(\tilde{A})\tilde{r}(\tilde{A})$. With $K \propto \tilde{r}(\tilde{A})N$ and $\lambda = \tilde{r}^2(\tilde{A})/\phi(f^*)$ we recover the same result as the one in the subgaussian setting by replacing the gaussian mean width by the Rademacher complexity. We will see in section 4 that the Rademacher complexity can be equivalent to the Gaussian mean width under weak moments Assumptions. However, using minmax MOM estimators we have relaxed two strong Assumptions 1) the i.i.d setting and 2) the subgaussian Assumption on the class $F$. Moreover the properly calibrated minmax MOM estimators are not affected if the number of outliers is less than number of observations $\times$ optimal rate in the i.i.d setup.

4 Application to Support Vectors Machine

In this section we consider regularization methods in some general Reproducing Kernel Hilbert Space (RKHS) (cf. [19] for a specific analysis on RKHS). The regularization function $\phi(\cdot)$ is defined as $\phi(\cdot) = \| \cdot \|^2_{\mathcal{H}_K}$ where $\| \cdot \|_{\mathcal{H}_K}$ design a norm in a certain RKHS $\mathcal{H}_K$. Using Theorems 2 and 3 we will derive explicit bounds on the error rates depending on $\|f^*\|_{\mathcal{H}_K}$.

4.1 Setting and results

We are given $N$ pairs $(X_i, Y_i)_{i=1}^N$ of random variable where the $X_i$’s take their values in some measurable space $\mathcal{X}$ and $Y_i \in \{-1, 1\}$. We introduce a kernel $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ measuring a similarity between elements of $\mathcal{X}$ i.e $K(x_1,x_2)$ is small if $x_1, x_2 \in \mathcal{X}$ are ”similar”. The main idea of kernel methods is to transport the design data $X_i$’s from the set $\mathcal{X}$ to a certain Hilbert space via the application $x \mapsto K(x, \cdot) := K_x(\cdot)$ and construct statistical procedure in this ”transported” and structured space. The kernel $K$ is used to generate an Hilbert space known as Reproducing Kernel Hilbert Space (RKHS). Recall that if $K$ is a positive definite function $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, then by Mercer’s Theorem there exists an othonormal basis $(\phi_i)_{i=1}^\infty$ of $L_2(\mu)$ such that $\mu \times \mu$ almost surely, $K(x,y) = \sum_{i=1}^\infty \lambda_i \phi_i(x) \phi_i(y)$, where $(\lambda_i)_{i=1}^\infty$ is the sequence of eigenvalues (arranged in a non-increasing order) of $T_K$ and $\phi_i$ is the eigenvector corresponding to $\lambda_i$ where

$$T_K : L_2(\mu) \times L_2(\mu) \mapsto L_2(\mu)$$

$$(T_K f)(x) = \int K(x,y) f(y) d\mu(y) \quad (10)$$

The Reproducible Kernel Hilbert Space $\mathcal{H}_K$ is the set of all functions of the form $\sum_{i=1}^\infty a_i K(x_i, \cdot)$ where $x_i \in \mathcal{X}$ and $a_i \in \mathbb{R}$ converging in $L_2(\mu)$ endowed with the inner product

$$\langle \sum_{i=1}^\infty a_i K(x_i, \cdot), \sum_{i=1}^\infty b_i K(y_i, \cdot) \rangle = \sum_{i,j=1}^\infty a_i b_j K(x_i, y_i)$$

An alternative way to define a RKHS is via the feature map $\Phi : \mathcal{X} \mapsto \ell_2$ such that $\Phi(x) = (\sqrt{\lambda_i} \phi_i(i))_{i=1}^\infty$. Since $(\Phi_k)_{k=1}^\infty$ is an orthogonal basis of $\mathcal{H}_K$, it is easy to see that the unit ball of $\mathcal{H}_K$ can be expressed as

$$B_{\mathcal{H}_K} = \{f_{\beta}(\cdot) = \langle \beta, \Phi(\cdot) \rangle_{\ell_2}, \|\beta\|_{\ell_2} \leq 1\} \quad (11)$$

where $\langle \cdot, \cdot \rangle_{\ell_2}$ is the standard inner product in the Hilbert space $\ell_2$. In other words, the feature map $\Phi$ can be used to define an isometry between the two Hilbert spaces $\mathcal{H}_K$ and $\ell_2$.

The RKHS $\mathcal{H}_K$ is therefore a class of functions from $\mathcal{X}$ to $\mathbb{R}$ that can be used as a learning class $F$. Let the oracle $f^*$ be defined as

$$f^* \in \arg \min_{f \in \mathcal{H}_K} \mathbb{E}[\tilde{r}(f(X),Y)]$$

.

7
Let $f$ be in $\mathcal{H}_K$, by the reproducing property and Cauchy-Schwarz we have for all $x, y$ in $\mathcal{X}$

$$|f(x) - f(y)| = \langle f, K_x - K_y \rangle \leq \|f\|_{\mathcal{H}_K}\|K_x - K_y\|_{\mathcal{H}_K}$$

The norm of a function in the RKHS controls how fast the function varies over $\mathcal{X}$ with respect to the geometry defined by the kernel (Lipschitz with constant $\|f\|_{\mathcal{H}_K}$). As a consequence the norm of regularization $\|\cdot\|_{\mathcal{H}_K}$ is related with its degree of smoothness w.r.t. the metric defined by the kernel on $\mathcal{X}$. By considering the convex function $\phi(\cdot) = \|\cdot\|^2_{\mathcal{H}_K}$ as a regularizer, we are thus in the "classical" setup of regularized problems. Clearly $\phi$ verifies Assumption 3 with $\eta = 2$. by taking the Hinge loss function, the estimator $\hat{f}_\lambda^\phi$ defined in Equation (2) becomes

$$\hat{f}_\lambda^\phi = \arg \min_{f \in \mathcal{H}_K} \frac{1}{N} \sum_{i=1}^{N} \max(1 - f(X_i)Y_i, 0) + \lambda\|f\|_{\mathcal{H}_K}^2$$

(12)

known as Support Vector Machines (see [20]). Note that the minmax MOM estimator defined in (7) can also be considered in this example.

To apply Theorem 1 and 3 it is necessary to verify the Bernstein Assumption in the neighborhood of the oracle $f^*$. Such a result is established in Section 4.2 for the Hinge loss function. Secondly, to obtain explicit bounds, the complexity parameters $r(A^\ast)$ and $\hat{r}(A)$ have to be computed explicitly. To do so we have to compute the gaussian mean-width and the Rademacher complexity of the set $\{f \in \mathcal{H}_K : \|f - f^*\|_{\mathcal{H}_K} \leq \rho, \|f - f^*\|_{L^2} \leq r\}$ for any $\rho, r > 0$.

**Lemma 2.** Let $B_{\mathcal{H}_K}$ denote the unit ball of $\mathcal{H}_K$, for all $\rho, r > 0$

$$w(\mathcal{H}_K \cap rB_{L_2} \cap \rho B_{\mathcal{H}_K}) \leq \left( \sum_{k=1}^{\infty} \left( \rho^2 \lambda_k \land r^2 \right) \right)^{1/2} \leq \left( \sum_{k=1}^{\infty} \rho^2 \lambda_k \right)^{1/2}$$

(13)

where $(\lambda_k)_{k=1}^{\infty}$ are the eigenvalues associated to $T_K$ defined in Equation (10).

The proof is presented in Section D.1. Straightforward computations finally give the following Theorem

**Theorem 4.** Let $\mathcal{X}$ be some measurable space and $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be a positive definite kernel where $\mathcal{H}_K$ denote its associated RKHS. Let $(\lambda_k)_{k=1}^{\infty}$ be the sequence of eigenvalues associated to $T_K$ in $L^2(\mu)$. Grant Assumptions 4, 5 and 6. Then the Lepski’s estimator $\hat{f}$ defined in Section 2 satisfies with probability larger than

$$1 - 2 \exp \left( - C(A^\ast, L)\sqrt{N}\|f^*\|_{\mathcal{H}_K}\left( \sum_{k=1}^{\infty} \lambda_k \right) \right)$$

that

$$\|\hat{f} - f^*\|_{L^2}^2 \leq C(A^\ast, L)\|f^*\|_{\mathcal{H}_K}\left( \sum_{k=1}^{\infty} \lambda_k \right)^{1/2} \quad \mathbb{P}\{\hat{f} \leq C(A^\ast, L)\|f^*\|_{\mathcal{H}_K}\left( \frac{\sum_{k=1}^{\infty} \lambda_k}{N} \right)^{1/2}\}$$

and

$$\|\hat{f} - f^*\|_{\mathcal{H}_K} \leq C(A^\ast, L)\|f^*\|_{\mathcal{H}_K}$$

For bounded kernel $K$, we also get from Theorem 2.1 in [15] that for all $\rho, r > 0$,

$$\mathbb{E} \sup_{f \in \mathcal{H}_K \cap rB_{L_2} \cap \rho B_{\mathcal{H}_K}} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sigma_i(f - f^*)(X_i) \leq C(\|K\|_{\infty}) \left( \sum_{k=1}^{\infty} \left( \rho^2 \lambda_k \land r^2 \right) \right)^{1/2}$$

In this setup, Gaussian mean-width and Rademacher complexities are equivalent. Under the Assumption that the kernel $K$ is bounded, it is thus possible to remove the subgaussian and i.i.d Assumptions without deteriorating the rates of convergence by using minmax MOM estimators.
4.2 Bernstein’s assumption

In this section, we show that the local Bernstein condition holds for various design \( X \) for the Hinge loss function. We obtain the result under the assumption that the oracle \( f^* \) is actually the Bayes rules which is the function minimizing the risk \( f \mapsto R(f) \) over all measurable functions from \( \mathcal{X} \) to \( \mathbb{R} \). It is easy to verify that \( f^*(x) = \text{sign}(2\eta(x) - 1) \) where \( \eta(X) = \mathbb{P}(Y = 1|X) \). In that case, the Bernstein condition (see [2]) coincides with the margin assumption (see [23, 13]). We also assume that the class \( F \) and design \( X \) satisfies a “local \( L_4/L_2 \)-assumption”. Let \( r^2 \in \{r^2(A^*), C_{K,r}(\hat{A})\} \), to verify Assumption 5, choose \( r^2 = r^2(A^*) \) while for Assumption 8, choose \( r^2 = C_{K,r}(\hat{A}) \).

Assumption 9. There exists \( c > 0 \) such that for all \( f \in F \), \( \|f - f^*\|_{L_4} \leq c\|f - f^*\|_{L_2} \) such that \( 2c^2 r \leq 1 \).

Note that the constant \( c \) may depends on the dimension as long as \( 2c^2 r \leq 1 \) (see [4] for examples).

Assumption 10. There exists \( \alpha > 0 \) such that, that for all \( f \in F \) such that \( \|f - f^*\|_{L_2} \leq r \) and for all \( x \in \mathcal{X} \), \( |f(x) - f^*(x)| \leq 2c^2 r \)
\[
\min(\eta(x), 1 - \eta(x), |1 - 2\eta(x)|) \geq \alpha.
\]

Assumption 10 is also local. It excludes trivial cases where deterministic predictors equal to 1 or -1 are optimal in a neighborhood of the oracle \( f^* \). It is also related to the margin condition [14, 22] trough the term \( |1 - 2\eta(x)| \).

Theorem 5. Grant Assumptions 9 and 10. Assume that the oracle \( f^* \) is the Bayes estimator i.e. \( f^*(x) = \text{sign}(2\eta(x) - 1) \) for all \( x \in \mathcal{X} \). Then, the local Bernstein condition holds: for all \( f \in F \) such that \( \|f - f^*\|_{L_2} \leq r \), \( \|f - f^*\|_{L_2} \leq \frac{2}{\alpha} PL_f \).

5 Application to Elastic net

Elastic net is a regularization and variables selection method introduced in [24]. Let \( F \) be the class of linear functionnals in \( \mathbb{R}^p \), \( F = \{\langle \cdot, t \rangle \mid t \in \mathbb{R}^p \} \) and let \( (X_i, Y_i)_{i=1}^N \) be random variables valued in \( \mathbb{R}^p \times \mathcal{Y} \). As the oracle is denoted \( f^* \), we introduce \( t^* \) such that \( f^*(\cdot) = \langle t^*, \cdot \rangle \). Let \( \alpha \in [0, 1] \), for any \( t \in \mathbb{R}^p \), the elastic net penalization is defined as
\[
\phi(t) = (1 - \alpha)\|t\|_1 + \alpha\|t\|_2^2
\]
(14)
where \( \|t\|_1 = \sum_{i=1}^p |t_i| \) and \( \|t\|_2^2 = \sum_{i=1}^p t_i^2 \). For \( \alpha = 1 \) and \( \alpha = 0 \) we recover respectively the Lasso and the ridge penalizations (those cases will not be studied in the sequel). Clearly \( \phi \) defined in Equation (14) satisfies Assumption 3 with \( \eta = 2 \). Let \( \hat{t} \) be a loss function verifying Assumption 1, the estimator \( \hat{t}_\lambda^\phi \) defined in Equation (2) becomes
\[
\hat{t}_\lambda^\phi \in \arg\min_{t \in \mathbb{R}^p} \frac{1}{N} \sum_{i=1}^N \ell(\langle X_i, t \rangle, Y_i) + \lambda((1 - \alpha)\|t\|_1 + \alpha\|t\|_2^2)
\]
(15)
Results on the local Bernstein Assumption (see Assumptions 5 and 8) can be found in [4] for the quantile, the logistic or even the Huber loss functions. For the Hinge loss, see Section 4.2.

To compute the Gaussian mean-width assume that the design \( X \) is isotropic i.e for all \( t \in \mathbb{R}^p \), \( \mathbb{E}\langle X, t \rangle^2_{\mathbb{R}^p} = \|t\|_2^2 \).

It means that the \( L_2(\mu) \) norm coincides with the natural Euclidean structure on the space \( \ell^2_2 \). Thus, for all \( \rho, r > 0 \), under the isotropic Assumption we have
\[
w(B^p_\rho(0) \cap rB^p_2) = \mathbb{E}_{t \in \mathbb{R}^p} \sup_{(1 - \alpha)\|t\|_1 + \alpha\|t\|_2^2 \leq \rho, \|t\|_2 \leq r} \langle G, t \rangle_{\mathbb{R}^p}
\]
(16)
where \( G \) is a standard Gaussian random vector in \( \mathbb{R}^p \) and \( B^p_2 \) denote the unit ball in \( (\mathbb{R}^p, \| \cdot \|_2) \). Let \( \alpha \in (0, 1) \). We have,
\[
w(B^p_\rho(0) \cap rB^p_{L_2}) \leq \min \left( w\left(\frac{\rho}{1 - \alpha}B^p_1 \cap rB^p_2\right), w\left(\min\left(\sqrt{\frac{\rho}{\alpha}}B^p_2\right)\right) \right)
\]
(17)
Computations of \( w(\rho B_1^p \cap rB_2^p) \) for all \( r, \rho > 0 \) have been done in [9]. Let us define,

\[
\begin{aligned}
  r_1^* &= \inf \left\{ r > 0 : w \left( \phi(f^*) \frac{1}{1-\alpha} B_1^p \cap rB_2^p \right) \leq \sqrt{N}r^2 \right\} \\
  r_2^* &= \inf \left\{ r > 0 : w \left( \min (r, \sqrt{\frac{\phi(f^*)}{\alpha}}B_2^p) \right) \leq \sqrt{N}r^2 \right\}
\end{aligned}
\]

From Equation (17) and the definition of \( r^* \) it follows that \( r^* \leq C(A^*, L) \min(r_1, r_2) \). Straightforward computations give,

\[
(r_1^*)^2 = \begin{cases} \\
  \frac{\phi(f^*)}{1-\alpha} \sqrt{\frac{\log(1-\alpha)}{\lambda}} & \text{if } \log(p) \leq \frac{\phi^2(f^*)}{N(1-\alpha)} \leq p^2 \\
  \frac{\phi(f^*)}{1-\alpha} \sqrt{\frac{\log(\exp(p))}{N}} & \text{if } \frac{N\phi^2(f^*)}{(1-\alpha)} \geq \log(p) \\
  \frac{\sqrt{\phi(f^*)p \log(\exp(p))}}{\sqrt{N}} & \text{if } N \geq \frac{\alpha p}{\phi(f^*)} \\
  \frac{\sqrt{\phi(f^*)p}}{\alpha N} & \text{if } N \leq \frac{\alpha p}{\phi(f^*)}
\end{cases}
\]

**Theorem 6.** Assume that the data are i.i.d with the same distribution \( P \) and consider a loss function verifying Assumption 1. Grant Assumptions 4 and 5. Let \( r^* = \min(r_1^*, r_2^*) \) and let \( \lambda = (r^*)^2/\phi(f^*) \). With probability larger than

\[1 - 2 \exp \left( - C(A^*, L) N (r^*)^2 \right)\]

the estimator \( \hat{f}^\phi_\lambda \) defined in Equation (15) satisfies

\[ \| \hat{f}^\phi_\lambda - f^* \|_{L^2} \leq C(A^*) r^* \quad \text{and} \quad \phi(\hat{f} - f^*) \leq C(A^*) \phi(f^*) . \]

In Theorem 6 we set \( \lambda = (r^*)^2/\phi(f^*) \) which is evidently unknown, however it is possible to use Theorem 2 to get an adaptive estimator for the Elastic net. When \( 1 - \alpha \) is close to 1 that is when the penalization \( \ell_1 \) is dominant we have \( r^* = r_1^* \) and we recover the result for the Lasso (see [9]). When \( \alpha \) is close to 1 the elastic net is almost equivalent to ridge regression and \( r^* = r_2^* \). We recover the results for the ridge regression.

To use the minimax MOM estimator for the elastic net procedure it is necessary to compute the rademacher complexities. Under technical Assumptions, it is possible to link Rademacher complexity and Gaussian mean-width (see [17]). For the sake of brevity we do not present the result here. It is possible to relax the i.i.d Assumption and the subgaussian Assumption on the class \( F - f^* \) using MOM types estimators.

### 6 Conclusion

We have presented two general results for the RERM and minimax MOM estimators describing the statistical properties of regularization in learning theory. The results highlight the importance of the calibration of the parameter \( \lambda \). For those two estimators we do not assume that the regularization is a norm which is, as far as we know a new general result for Lipschitz and convex loss functions. Under the local Bernstein Assumption, we can obtain rates of convergence depending on \( \phi(f^*) \). We recover the minimax rates in the true model \( F^* = \{ f : \phi(f) \leq \phi(f^*) \} \).

Results for the RERM have been derived under the i.i.d and the subgaussian Assumptions on the class \( F - f^* \) while no concentration Assumption is required for minimax MOM estimators. For MOM estimators, a number of outliers smaller than the rate of convergence × number of observations does not deteriorate the learning procedure. There are a number of interesting directions in which this work can be extended. One relevant and closely related problem is to obtain sparsity bounds, i.e bounds depending on an underlying structure of the oracle \( f^* \) such as the sparsity or the rank of the oracle \( f^* \). It has been partially done (under really strong Assumption) in [1] when the regularization function if a norm. However without this Assumption, the proofs no longer hold and a new analysis has to be design.

### References

[1] P. Alquier, V. Cottet, and G. Lecué. Estimation bounds and sharp oracle inequalities of regularized procedures with lipschitz loss functions. *arXiv preprint arXiv:1702.01402*, 2017.
is split into two parts. First, we identify an event onto which the statistical behavior of the regularized estimator 
\( \hat{f}_\lambda := \hat{f}_{\lambda}^\phi \) 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 (4). Let us define \( \rho^* = (2 + \gamma)\eta \phi(f^*) \). We first introduce this event:

\[
\Omega := \left\{ \text{for all } f \in F \cap (f^* + r^* B_{L^2}) \cap B^\rho_{\rho^*}(f^*), \quad |(P - P_N)\mathcal{L}_f| \leq \theta(r^*)^2 \right\}
\]

where we recall that \( r^* = r(A^*) \) and \( B^\rho_{\rho^*}(f^*) = \{ f \in F : \phi(f - f^*) \leq \rho^* \} \).

## A Proof of Theorems RERM

In the remaining of the proof we shall use repeatedly the following notations

\[
A = A^*, \quad \theta = \frac{1}{2A}, \quad \delta = \frac{2}{A} + 3 \quad \gamma = \frac{2}{A} + 2.
\]

### A.1 Proof Theorem 1

Proof of Theorem 1 is split into two parts. First, we identify an event onto which the statistical behavior of the regularized estimator 
\( \hat{f}_\lambda := \hat{f}_{\lambda}^\phi \) 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 (4). Let us define \( \rho^* = (2 + \gamma)\eta \phi(f^*) \). We first introduce this event:

\[
\Omega := \left\{ \text{for all } f \in F \cap (f^* + r^* B_{L^2}) \cap B^\rho_{\rho^*}(f^*), \quad |(P - P_N)\mathcal{L}_f| \leq \theta(r^*)^2 \right\}
\]

where we recall that \( r^* = r(A^*) \) and \( B^\rho_{\rho^*}(f^*) = \{ f \in F : \phi(f - f^*) \leq \rho^* \} \).
Proposition 1. Let \( \lambda \geq \lambda_0 := (r^*)^2 / \phi(f^*) \), on the event \( \Omega \), one has

\[
\phi(\hat{f}_\lambda - f^*) \leq \rho^*, \quad \|\hat{f}_\lambda - f^*\|_{L_2} \leq \lambda \frac{\delta\phi(f^*)}{(A^{-1} - \theta)r^*}
\]

Proof. Let \( \lambda \geq \lambda_0 \), we denote \( B_\lambda = \left( f^* + (\lambda\delta\phi(f^*)/((A^{-1} - \theta)r^*))B_{L_2} \right) \cap B^\alpha_{\rho^*}(f^*) \). We want to prove that \( \hat{f}_\lambda \in B_\lambda \). We recall that the regularized empirical excess loss function is defined for all \( f \in F \) by

\[
P_N \mathcal{L}_f^\lambda = P_N \mathcal{L}_f + \lambda(\phi(f) - \phi(f^*))
\]

Since \( \hat{f}_\lambda \) is such that \( P_N \mathcal{L}_f^\lambda \leq 0 \), it is enough to prove that \( P_N \mathcal{L}_f^\lambda > 0 \) for all \( f \in F \setminus B_\lambda \) to get that \( \hat{f}_\lambda \in B_\lambda \). In fact, for the adaptive procedure it will be necessary to use the results from Lemma 1 which is equivalent (up to the choice of the constants) to show that \( P_N \mathcal{L}_f^\lambda > 0 \) for all \( f \in F \setminus B_\lambda \). From Lemma 1 it follows immediately that \( \phi(\hat{f}_\lambda - f^*) \leq \rho^* \) and \( \|\hat{f}_\lambda - f^*\|_{L_2} \leq \lambda \frac{\delta\phi(f^*)}{(A^{-1} - \theta)r^*} \).

Proof. Lemma 1

The proof follows from an homogeneity argument saying that if \( P_N \mathcal{L}_f^\lambda > 2(\theta + 1)\lambda\phi(f^*) \) on the border of \( B_\lambda \), then we also have \( P_N \mathcal{L}_f^\lambda > 2(\theta + 1)\alpha\phi(f^*) \) for all \( f \in F \) outside \( B_\lambda \). Inside \( B_\lambda \), the arguments are similar. Let \( f \in F \) be outside of \( B_\lambda \). By convexity of \( F \), there exists \( f_0 \in F \) and \( \alpha > 1 \) such that \( f - f^* = \alpha(f_0 - f^*) \) and \( f_0 \in \partial B_\lambda \) where we denote by \( \partial B_\lambda \) the border of \( B_\lambda \). By definition, we either have: 1) \( \phi(f_0 - f^*) = \rho^* \) and \( \|f_0 - f^*\|_{L_2} \leq (\lambda\delta\phi(f^*)/((A^{-1} - \theta)r^*)) \) in that case, \( \alpha \) is such that \( 1 \leq \alpha \leq \phi(f - f^*)/\rho^* \) (see Lemma 6 in Section D.2) or 2) \( \|f_0 - f^*\|_{L_2} = (\lambda\delta\phi(f^*)/((A^{-1} - \theta)r^*)) \) and \( \phi(f_0 - f^*) \leq \rho^* \) and, in that case, \( \alpha = \|f - f^*\|_{L_2} / ((\lambda\delta\phi(f^*)/((A^{-1} - \theta)r^*))) \). We will treat the two cases independently.

Let us first explain the role of the convexity of the loss function by writing down an homogeneity argument linking the empirical excess risk of \( f \) to the one of \( f_0 \). For all \( i \in \{1, \cdots, N\} \), let \( \psi_i : \mathbb{R} \to \mathbb{R} \) be defined for all \( u \in \mathbb{R} \) by

\[
\psi_i(u) = \tilde{\ell}(u + f^*(X_i), Y_i) - \tilde{\ell}(f^*(X_i), Y_i).
\]

The functions \( \psi_i \) are such that \( \psi_i(0) = 0 \), they are convex because \( \tilde{\ell} \) is, in particular \( \alpha \psi_i(u) \leq \psi_i(\alpha \rho) \) for all \( u \in \mathbb{R} \) and \( \alpha \geq 1 \) and \( \psi_i(f(X_i) - f^*(X_i)) = \tilde{\ell}(f(X_i), Y_i) - \tilde{\ell}(f^*(X_i), Y_i) \) so that the following holds:

\[
P_N \mathcal{L}_f = \frac{\alpha}{N} \sum_{i=1}^{N} \psi_i(f(X_i) - f^*(X_i)) = \frac{\alpha}{N} \sum_{i=1}^{N} \psi_i(\alpha(f_0(X_i) - f^*(X_i))) \\
\geq \frac{\alpha}{N} \sum_{i=1}^{N} \psi_i((f_0(X_i) - f^*(X_i))) = \alpha P_N \mathcal{L}_{f_0}.
\]

For the regularization part the same homogeneity arguments holds.

\[
\phi(f) - \phi(f^*) = \phi(f^* + \alpha(f_0 - f^*)) - \phi(f^*) \geq \alpha (\phi(f_0) - \phi(f^*))
\]

where we used Lemma 7 (see Section D.2). Therefore

\[
P_N \mathcal{L}_f^\lambda \geq \alpha P_N \mathcal{L}_{f_0}^\lambda
\]

Let us now place ourselves on the event \( \Omega \) up to the end of the proof and let \( f_0 \in F \cap \partial B_\lambda \). We explore two cases depending on the localization of \( f_0 \) on the border of \( B_\lambda \): 1) \( \phi(f_0 - f^*) = \rho^* \) and \( \|f_0 - f^*\|_{L_2} \leq (\lambda\delta\phi(f^*)/((A^{-1} - \theta)r^*)) \) which is the case where the regularization part helps to show that \( P_N \mathcal{L}_{f_0}^\lambda > 2(\theta + 1)\lambda\phi(f^*) \) or 2) \( \|f_0 - f^*\|_{L_2} = (\lambda\delta\phi(f^*)/((A^{-1} - \theta)r^*)) \) and \( \phi(f_0 - f^*) \leq \rho^* \) which is where the Bernstein’s condition helps. We consider the first case which is when \( \phi(f_0 - f^*) = \rho^* \) and \( \|f_0 - f^*\|_{L_2} \leq (\lambda\delta\phi(f^*)/((A^{-1} - \theta)r^*)) \).
There are two cases, either \( \|f_0 - f^*\|_{L_2} \leq r^* \) or \( \|f_0 - f^*\|_{L_2} \geq r^* \). In both cases, from the fact that \( \phi(f_0 - f^*) \leq \eta(\phi(f_0) + \phi(f^*)) \) we have \( \phi(f_0) - \phi(f^*) \geq \gamma \phi(f^*) \). If \( \|f_0 - f^*\|_{L_2} \leq r^* \), on \( \Omega \) we have \( |(P - P_N)\lambda f_0| \leq \theta(r^*)^2 \) and we get

\[
P_N\mathcal{L}_f = P_N\mathcal{L}_f + \lambda(\phi(f) - \phi(f^*)) \geq \alpha(P_N\mathcal{L}_f + \lambda\phi(f^*)) \geq \alpha(-\theta(r^*)^2 + \gamma\lambda\phi(f^*)) \geq (-\theta + \gamma)\lambda\phi(f^*) > 2(\theta + 1)\lambda\phi(f^*)
\]

where we used the facts that \( \lambda \geq (r^*)^2/\phi(f^*) \) and \( P\mathcal{L}_f_0 \geq 0 \). If \( r^* \leq \|f_0 - f^*\|_{L_2} \leq \lambda\delta\phi(f^*)/((A^{-1} - \theta)r^*) \) we use the same projection trick. Let \( \alpha_1 = \|f_0 - f^*\|_{L_2}/r^* \) and set \( f_1 \) in \( F \) be such that \( f_0 - f^* = \alpha_1(f_1 - f^*) \). We have \( \|f_1 - f^*\|_{L_2} = r^* \) and \( \phi(f_1 - f^*) \leq \rho^* \). Therefore on \( \Omega \) we have

\[
P_N\mathcal{L}_f \geq \alpha(P_N\mathcal{L}_f_0 + \gamma\lambda\phi(f^*)) \geq \alpha(\alpha_1 P_N\mathcal{L}_f_1 + \gamma\lambda\phi(f^*)) \geq \gamma\lambda\phi(f^*) > 2(\theta + 1)\lambda\phi(f^*)
\]

Since, on \( \Omega \), \( P_N\mathcal{L}_f_1 \geq P\mathcal{L}_f_1 - \theta(r^*)^2 \geq A^{-1}\|f_1 - f^*\|_{L_2} - \theta(r^*)^2 = (A^{-1} - \theta)(r^*)^2 > 0 \) where we used Assumption 5.

We now turn to the second case where \( \|f_0 - f^*\|_{L_2} = \lambda\delta\phi(f^*)/((A^{-1} - \theta)r^*) \) and \( \phi(f_0 - f^*) \leq \rho^* \). Remember that in this case \( \alpha = \|f - f^*\|_{L_2}/((\lambda\delta\phi(f^*/((A^{-1} - \theta)r^*)) \). The regularization part no longer helps. However, by the Bernstein Assumption 5 and using the same projection trick we get

\[
P_N\mathcal{L}_f \geq \frac{\|f - f^*\|_{L_2}}{(\lambda\delta\phi(f^*))/((A^{-1} - \theta)r^*)} P_N\mathcal{L}_f_0 \geq \frac{\|f - f^*\|_{L_2}}{(\lambda\delta\phi(f^*))/((A^{-1} - \theta)r^*)} \frac{\|f_0 - f^*\|_{L_2}}{r^*} P_N\mathcal{L}_f_1 \geq \frac{\|f - f^*\|_{L_2}}{r^*} (A^{-1} - \theta)(r^*)^2
\]

where \( f_1 \) is such that \( f_0 - f^* = (\|f_0 - f^*\|_{L_2}/(r^*)) (f_1 - f^*) \). We have \( \|f_1 - f^*\|_{L_2} = r^* \) and \( \phi(f_1 - f^*) \leq \rho^* \). Since \( \|f - f^*\|_{L_2} \geq \lambda\delta\phi(f^*)/((A^{-1} - \theta)r^*) \), we finally get

\[
P_N\mathcal{L}_f \geq \frac{\|f - f^*\|_{L_2}}{r^*} (A^{-1} - \theta)(r^*)^2 - \lambda\phi(f^*) \geq (\delta - 1)\lambda\phi(f^*) > 2(\theta + 1)\lambda\phi(f^*)
\]
We conclude the proof by studying $P_N \mathcal{L}^A_f$ for $f \in F \cap B_{\lambda}$. One more time there are two cases, either $\|f - f^*\|_{L_2} \leq r^*$ or $\|f - f^*\|_{L_2} \geq r^*$. In the first case, since $P_L f_0$, on $\Omega$ we get that

$$P_N \mathcal{L}^A_f \geq -\theta (r^*)^2 - \lambda \phi(f^* \geq -(\theta + 1) \lambda \phi(f^*)$$

For $\|f - f^*\|_{L_2} \geq r^*$ using the projection trick, there exists $\alpha \geq 1$ such that $P_N \mathcal{L}_f \geq \alpha P_N \mathcal{L}_{f_0}$ where $f_0$ satisfies $\|f_0 - f^*\|_{L_2} = r^*$ and $\phi(f_0 - f^*) \leq \rho^*$. Therefore on $\Omega$, using Assumption 5, we get $P_N \mathcal{L}_f \geq \alpha (A^{-1} - \theta) (r^*)^2 \geq -\theta \lambda \phi(f^*)$. Finally in that case

$$P_N \mathcal{L}^A_f \geq -(\theta + 1) \lambda \phi(f^*)$$

Next, we prove that $\Omega$ holds with large probability. To that end, we use the results from [1].

**Lemma 3.** [1] Assume that Assumption 1 and Assumption 4 hold. Let $F' \subset F$ then for every $u > 0$, with probability at least $1 - 2 \exp(-u^2)$

$$\sup_{f,g \in F'} |(P - P_N)(\mathcal{L}_f - \mathcal{L}_g)| \leq \frac{16L}{\sqrt{N}}(w(F') + ud_{L_2}(F'))$$

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

It follows from Lemma 3 that for any $u > 0$, with probability larger that $1 - 2 \exp(-u^2),$

$$\sup_{f \in F \cap (f^* + r^* B_{L_2}) \cap B_{\rho^*}(f^*)} |(P - P_N)\mathcal{L}_f| \leq \sup_{f, g \in F \cap (f^* + r^* B_{L_2}) \cap B_{\rho^*}(f^*)} |(P - P_N)(\mathcal{L}_f - \mathcal{L}_g)| \leq \frac{16L}{\sqrt{N}}\left(w(F \cap (f^* + r^* B_{L_2}) \cap B_{\rho^*}(f^*)) + ud_{L_2}(F \cap (f^* + r^* B_{L_2}) \cap B_{\rho^*}(f^*)) \right).$$

We have $d_{L_2}(F \cap (f^* + r^* B_{L_2}) \cap B_{\rho^*}(f^*)) \leq r^*$ and $w(F \cap (f^* + r^* B_{L_2}) \cap B_{\rho^*}(f^*)) = w(F \cap r^* B_{L_2} \cap B_{\rho^*}(0))$. By definition of the complexity parameter (see Equation (3)), for $u = \theta \sqrt{Nr^*/(32L)}$, with probability at least

$$1 - 2 \exp\left(-\frac{\theta^2 N(r^*)^2}{32 L^2}\right)$$

for every $f$ in $F \cap (f^* + r^* B_{L_2}) \cap B_{\rho^*}(f^*)$,\n
$$|(P - P_N)\mathcal{L}_f| \leq \theta (r^*)^2 \tag{20}$$

**A.2 Proof Theorem 2**

In this section we work on the event

$$\tilde{\Omega} := \left\{ \text{for all } f \in F \cap \left(f^* + \frac{2\delta}{A - 1 - \theta} r^* B_{L_2}\right) \cap B_{\rho^*}(f^*), \quad |(P - P_N)\mathcal{L}_f| \leq \theta (r^*)^2 \right\}$$

Using the same proof as the one for $\Omega$, it easy to show that $\tilde{\Omega}$ holds with probability larger than

$$1 - 2 \exp\left(-\frac{(\theta(A^{-1} - \theta))^2 N(r^*)^2}{(64\delta)^2}\right)$$

Note that $\Omega \subset \tilde{\Omega}$ and then Lemma 1 still holds.

Let us assume that $(\lambda_j)_{j=0}^J = (r^2 / \phi_j)_{j=0}^J$ is non increasing. From the choice of $(\phi_j)_{j=0}^J$, there exists $\tilde{k}$ such that $\phi_{\tilde{k}} \leq \phi(f^*) \leq 2 \phi_{\tilde{k}}$. Note that if $(\lambda_j)_{j=0}^J$ is non decreasing, it is enough to use the same proof with $\tilde{k}$ such that $(1/2) \phi_{\tilde{k}} \leq \phi(f^*) \leq \phi_{\tilde{k}}$. 

14
Moreover, from Lemma 1, for all \( \lambda \geq \lambda_0 \), \( T_\lambda(f^*) = -P_NL_{f^\lambda}^{\lambda} \leq (\theta + 1)\lambda \phi(f^*) \leq 2(\theta + 1)\lambda \phi_k \). Since \( \phi_k \leq \phi(f^*) \) it follows that \( \lambda_k \geq \lambda_0 \). And finally

\[
P_NL_{f^\lambda}^{\lambda} \leq 2(\theta + 1)\phi_k \lambda_k \leq 2(\theta + 1)\phi_k \lambda_k \text{ for all } k \geq \tilde{k}
\]

From the definition of \( k^* \) and Equation (22) it follows that \( k^* \leq \tilde{k} \) and thus, \( \tilde{f} \in \tilde{R}_{\tilde{k}} \). As a consequence, \( P_NL_{f^\lambda}^{\lambda} \leq T_{\lambda_k}(\tilde{f}) \) and we get

\[
P_NL_{f^\lambda}^{\lambda} \leq 2(\theta + 1)\lambda_\tilde{k} \phi_k \leq 2(\theta + 1)\lambda_\tilde{k} \phi(f^*)
\]

From Lemma 1 it follows that \( \tilde{f} \) satisfies \( \|\tilde{f} - f^*\|_{L_2} \leq \lambda_\tilde{k} \delta \phi(f^*)/(\alpha^{-1} - \theta) r^* \) \( \leq 2\lambda_\tilde{k} \delta \phi_k/(\alpha^{-1} - \theta) r^* \) \( \leq (2\delta/(\alpha^{-1} - \theta)) r^* \) and \( \phi(\tilde{f} - f^*) \leq \eta(2 + \gamma) \phi(f^*) \).

We finish this section by showing an oracle inequality for \( \tilde{f} \). From the fact that \( \|\tilde{f} - f^*\|_{L_2} \leq (2\delta/(\alpha^{-1} - \theta)) r^* \) and \( \phi(\tilde{f} - f^*) \leq \eta(2 + \gamma) \phi(f^*) \), it follows, on \( \bar{\Omega} \) that \( (P - P_N)\mathcal{L}_{\tilde{f}} \leq \theta(r^*)^2 \). For all \( \lambda > 0 \)

\[
P\mathcal{L}_{\tilde{f}} = P_NL_{\tilde{f}} + (P - P_N)\mathcal{L}_{\tilde{f}} \leq P_NL_{\tilde{f}}^\lambda + \lambda(\phi(f^*) - \phi(\tilde{f})) + \theta(r^*)^2 \leq P_NL_{\tilde{f}}^\lambda + \lambda\phi(f^*) + \theta(r^*)^2.
\]

In particular for \( \lambda = \lambda_\tilde{k} \) one has \( P_NL_{f^\lambda}^{\lambda} \leq 2(\theta + 2)\phi_k \lambda_k \leq 2(\theta + 1)(r^*)^2 \) and \( \lambda_\tilde{k} \phi(f^*) \leq 2(r^*)^2 \). Finally

\[
P\mathcal{L}_{\tilde{f}} \leq (4 + 3\theta)(r^*)^2
\]

**B Proof Theorem 3**

Let \( \tilde{r} \) and \( C_{K,r} \) design respectively \( \tilde{r}(\tilde{A}) \) and \( C_{K,r}(\tilde{A}) \). Moreover, all along the proof, the following notations will be used repeatedly.

\[
A = \tilde{A}, \quad \theta = \frac{1}{2A}, \quad \delta = \frac{2}{A} + 3 \quad \gamma = \frac{2}{A} + 2, \quad \mu = \frac{\theta}{192L}.
\]

The proof is divided into two parts. First, we identify an event where the minimax MOM estimators \( \hat{f}_K \) is controlled. Then, we prove that this event holds with large probability. Let \( K \geq 7|\Omega|/3, \) and

\[
C_{K,r} = \max \left( \frac{96L^2K}{\theta^2 N}, \tilde{r}^2 \right) \text{ and } \rho^* = \eta(2 + \gamma) \phi(f^*)
\]

Let \( B_{\lambda,K} = \{ f \in E : \|f - f^*\|_{L_2} \leq \frac{\delta}{A \sqrt{C_{K,r}}} \lambda_k f^*(f^*) \text{ and } \phi(f^* - f^*) \leq \rho^* \} \). Consider the following event

\[
\Omega_K = \left\{ \forall f \in F \cap \sqrt{C_{K,r}}B_{L_2} \cap B_{\rho^*}f^*(f^*), \quad \sum_{k=1}^K I \left( \left| (P_{B_k} - P)(\ell_f - \ell_{f^*}) \right| \leq \theta C_{K,r} \right) \geq \frac{K}{2} \right\}.
\]

**B.1 Deterministic argument**

**Lemma 4.** \( \hat{f}_K \in B_{\lambda,K} \) if the following inequalities holds

\[
\sup_{f \in F \cap B_{\lambda,K}} \text{MOM}_K(\ell_{f^*} - \ell_f) + \lambda(\phi(f^*) - \phi(f)) \leq -2(\theta + 1)\lambda \phi(f^*) \quad \text{and} \quad (24)
\]

\[
\sup_{f \in F \cap B_{\lambda,K}} \text{MOM}_K(\ell_{f^*} - \ell_f) + \lambda(\phi(f^*) - \phi(f)) \leq (\theta + 1)\lambda \phi(f^*) \quad \text{and} \quad (25)
\]
Proof. For any $f \in F$, denote by $S(f) = \sup_{g \in F} MOM_K(\ell_f - \ell_g) + \lambda(\phi(f) - \phi(g))$. If (24) holds, by homogeneity of $MOM_K$, any $f \in F \setminus B_{\lambda,K}$ satisfies

$$S(f) \geq MOM_K(\ell_f - \ell_{f^*}) + \lambda(\phi(f) - \phi(f^*)) > 2(\theta + 1)\lambda\phi(f^*) .$$

On the other hand, if (25) and (24) hold,

$$S(f^*) = \sup_{f \in F} MOM_K(\ell_{f^*} - \ell_f) + \lambda(\phi(f^*) - \phi(f)) \leq (\theta + 1)\lambda\phi(f^*) .$$

Thus, by definition of $\mathbf{j}_K$ and (25),

$$S(\mathbf{j}_K) \leq S(f^*) \leq (\theta + 1)\lambda\phi(f^*) .$$

Therefore, if (24) and (25) hold, $\mathbf{j}_K \in B_{\lambda,K}$.

Lemma 5. For all $K \geq |\Omega|/3$ and $\lambda \geq C_{K,r}/\phi(f^*)$, inequalities (24) and (25) holds on $\Omega_K$.

Proof. The arguments are exactly the same as the one in the proof of Lemma 1. For all functions $f \in F \setminus B_{\lambda,K}$ and for each block $B_k$, there exist $\alpha \geq 1$ and $f_0 \in F$ in the border of $B_{\lambda,K}$ such that $P_{B_k}L_f \geq \alpha P_{B_k}L_{f_0}$. We present here only one case (the others are trivial applications of the arguments in the proof of Lemma 1). In the case where $\phi(f_0 - f^*) = \rho^*$ and $\sqrt{C_{K,r}} \leq \|f_0 - f^*\|_{L_2} \leq (\lambda\phi(f^*))/(A^{-1} - \theta\sqrt{C_{K,r}})$. We still have $\lambda(\phi(f_0) - \phi(f^*)) \geq \gamma\lambda\phi(f^*)$. Using the projection trick, there exists $\alpha_1 > 1$ such that on each block $B_k$, $P_{B_k}L_{f_0} \geq \alpha_1 P_{B_k}L_{f_1}$ for $f_1$ such that $\|f_1 - f^*\|_{L_2} = \sqrt{C_{K,r}}$ and $\phi(f_1 - f^*) \leq \rho^*$ and then, on the event $\Omega_K$, one more than $K/2$ blocks $B_k$

$$P_{B_k}L_f^L \geq \alpha(P_{B_k}L_{f_0} + \gamma\lambda\phi(f^*)) \geq \alpha(\alpha_1 P_{B_k}L_{f_1} + \gamma\lambda\phi(f^*)) \geq \gamma\lambda\phi(f^*) > 2(\theta + 1)\lambda\phi(f^*)$$

(26)

where we used the fact that on $\Omega_K$, there are at least $K/2$ blocks $B_k$ such that, $P_{B_k}L_{f_1} \geq PL_{f_1} - \theta C_{K,r} \geq A^{-1}\|f_1 - f^*\|_{L_2}^2 - \theta C_{K,r} = (A^{-1} - \theta) C_{K,r} > 0$ and Assumption 8.

As Equation (26) holds on more than $K/2$ blocks we get that

$$MOM_K(\ell_f - \ell_{f^*}) + \lambda(\phi(f) - \phi(f^*)) \geq 2(\theta + 1)\lambda\phi(f^*)$$

From the same arguments as the one in the proof of Lemma 1 we finally obtain

$$\sup_{f \in F \setminus B_{\lambda,K}} MOM_K(\ell_f - \ell_{f^*}) + \lambda(\phi(f^*) - \phi(f)) < -2(\theta + 1)\lambda\phi(f^*) ,$$

$$\sup_{f \in F \setminus B_{\lambda,K}} MOM_K(\ell_f - \ell_{f^*}) + \lambda(\phi(f^*) - \phi(f)) \leq (\theta + 1)\lambda\phi(f^*)$$

which concludes to proof.

B.2 Control of the stochastic event

Contrary to the deterministic argument, the control of the stochastic event is very different from the one for the RERM.

Proposition 2. Grant Assumptions 1, 2, 3, 7 and 8. Let $K \geq |\Omega|/3$. Then $\Omega_K$ holds with probability larger than $1 - 2\exp(-K/504)$.

Proof. Let $F = \{f \in F : \|f - f^*\|_{L_2} \leq \sqrt{C_{K,r}}, \phi(f - f^*) \leq \rho^*\}$ and let $h(t) = I\{t \geq 2\} + (t - 1)I\{1 \leq t \leq 2\}$. This function satisfies $\forall t \in \mathbb{R}^+, I\{t \geq 2\} \leq h(t) \leq I\{t \geq 1\}$. Let $W_k = (X_i, Y_i)_{i \in B_k}$ and, for any $f \in F$, let $G_f(W_k) = (P_{B_k} - P)(\ell_f - \ell_{f^*})$. Let also $C_{K,r} = \max\left(96L^2K/(\theta^2N), r^2\right)$. For any $f \in F$, let

$$z(f) = \sum_{k=1}^{K} I\{|G_f(W_k)| \leq \theta C_{K,r}\} .$$
Proposition 2 will be proved if \( \mathbb{P}(z(f) \geq K/2) \geq 1 - e^{-K/504} \). Let \( \mathcal{K} \) denote the set of indices of blocks which have not been corrupted by outliers, \( \mathcal{K} = \{ k \in \{1, \cdots, K \} : B_k \subset I \} \). Basic algebraic manipulations show that

\[
z(f) \geq |\mathcal{K}| \left( 1 - \frac{1}{24} \right) - \sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \left( h(2(\theta C_{K,r})^{-1}|G_f(W_k)|) - \mathbb{E}h(2(\theta C_{K,r})^{-1}|G_f(W_k)|) \right) - \sum_{k \in \mathcal{K}} \mathbb{E}h(2(\theta C_{K,r})^{-1}|G_f(W_k)|). \tag{27}
\]

The last term in (27) can be bounded from below since for all \( f \in \mathcal{F} \) and \( k \in \mathcal{K} \),

\[
\mathbb{E}h(2(\theta C_{K,r})^{-1}|G_f(W_k)|) \leq \mathbb{P}\left( |G_f(W_k)| \geq \frac{\theta C_{K,r}}{2} \right) \leq \frac{4\mathbb{E}G_f(W_k)^2}{\theta^2 C_{K,r}^2} \leq \frac{4K^2}{\theta^2 C_{K,r}^2 N^2} \sum_{i \in B_k} |\mathbb{E}(\ell_f - \ell_{f^*})^2(X_i, Y_i)| \leq \frac{4L^2 K}{\theta^2 C_{K,r} N} \|f - f^*\|_{L_2}^2.
\]

The last inequality follows from Assumption 7. Since \( \|f - f^*\|_{L_2} \leq \sqrt{C_{K,r}} \),

\[
\mathbb{E}h(2(\theta C_{K,r})^{-1}|G_f(W_k)|) \leq \frac{4L^2 K}{\theta^2 C_{K,r} N}.
\]

As \( C_{K,r} \geq 96L^2 K/(\theta^2 N) \),

\[
\mathbb{E}h(2(\theta C_{K,r})^{-1}|G_f(W_k)|) \leq \frac{1}{24}.
\]

Plugging this inequality in (27) yields

\[
z(f) \geq |\mathcal{K}| \left( 1 - \frac{1}{24} \right) - \sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \left( h(2(\theta C_{K,r})^{-1}|G_f(W_k)|) - \mathbb{E}h(2(\theta C_{K,r})^{-1}|G_f(W_k)|) \right). \tag{28}
\]

Using the Mc Diarmid’s inequality, with probability larger than \( 1 - \exp(-|\mathcal{K}|/288) \) we get

\[
\sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \left( h(2(\theta C_{K,r})^{-1}|G_f(W_k)|) - \mathbb{E}h(2(\theta C_{K,r})^{-1}|G_f(W_k)|) \right) \leq \frac{|\mathcal{K}|}{24} + \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \left( h(2(\theta C_{K,r})^{-1}|G_f(W_k)|) - \mathbb{E}h(2(\theta C_{K,r})^{-1}|G_f(W_k)|) \right).
\]

By the symmetrization lemma, it follows that

\[
\sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \left( h(2(\theta C_{K,r})^{-1}|G_f(W_k)|) - \mathbb{E}h(2(\theta C_{K,r})^{-1}|G_f(W_k)|) \right) \leq \frac{|\mathcal{K}|}{24} + 2\mathbb{E} \sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \sigma_k h(2(\theta C_{K,r})^{-1}|G_f(W_k)|).
\]

As \( \phi \) is 1-Lipschitz with \( \phi(0) = 0 \), the contraction Lemma from [10] and yields

\[
\sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \left( h(2(\theta C_{K,r})^{-1}|G_f(W_k)|) - \mathbb{E}h(2(\theta C_{K,r})^{-1}|G_f(W_k)|) \right) \leq \frac{|\mathcal{K}|}{24} + \frac{4}{\theta} \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \sigma_k \frac{G_f(W_k)}{C_{K,r}} \leq \frac{|\mathcal{K}|}{24} + \frac{4}{\theta} \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \sigma_k \frac{(P_{B_k} - P)(\ell_f - \ell_{f^*})}{C_{K,r}}.
\]
For any \( k \in \mathcal{K} \), let \((\sigma_i)_{i \in B_k}\) independent from \((\sigma_k)_{k \in \mathcal{K}}, (X_i)_{i \in I}\) and \((Y_i)_{i \in I}\). The vectors \((\sigma_i, \sigma_k(\ell_f - \ell_f)) (X_i, Y_i)\) and \((\sigma_i(\ell_f - \ell_f)) (X_i, Y_i)\) have the same distribution. Thus, by the symmetrization and contraction lemmas, with probability larger than \( 1 - \exp(-|\mathcal{K}|/288) \),

\[
\sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \left( h(2C_{K,r}^{-1} G_f(W_k)) - \mathbb{E} h(2C_{K,r}^{-1} G_f(W_k)) \right) \leq \frac{|\mathcal{K}|}{24} + \frac{8}{\theta} \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \sum_{i \in B_k} \sigma_i \left( \frac{\ell_f(\ell_f - \ell_f)}{C_{K,r}} \right) \left( X_i, Y_i \right) \sup_{f \in \mathcal{F}} \sum_{i \in \mathcal{K} \setminus \mathcal{K}} \sigma_i \left( \frac{\ell_f(\ell_f - \ell_f)}{C_{K,r}} \right) \left( X_i, Y_i \right) \leq \frac{|\mathcal{K}|}{24} + \frac{8K}{\theta} N \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i \in \mathcal{K} \setminus \mathcal{K}} \sigma_i \left( \frac{\ell_f(\ell_f - \ell_f)}{C_{K,r}} \right) \left( X_i, Y_i \right). \tag{29}
\]

Now either 1) \( K \leq \theta^2 \tilde{r}^2 N/(96L^2) \) or 2) \( K > \theta^2 \tilde{r}^2 N/(96L^2) \). Assume first that \( K \leq \theta^2 \tilde{r}^2 N/(96L^2) \), so \( C_{K,r} = \tilde{r}^2 \) and by definition of the complexity parameter

\[
\mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i \in \mathcal{K} \setminus \mathcal{K}} \sigma_i \left( \frac{f - f^*}{C_{K,r}} \right) \left( X_i \right) \right| = \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{\tilde{r}^2} \sum_{i \in \mathcal{K} \setminus \mathcal{K}} \sigma_i \left( f - f^* \right) \left( X_i \right) \left| \leq \frac{\mu |\mathcal{K}| N}{K}. \right.
\]

If \( K > \theta^2 \tilde{r}^2 N/(96L^2) \), \( C_{K,r} = 96L^2 K/(\theta^2 N) \). Then,

\[
\mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i \in \mathcal{K} \setminus \mathcal{K}} \sigma_i \left( \frac{f - f^*}{C_{K,r}} \right) \left( X_i \right) \right| \leq \mathbb{E} \left[ \frac{1}{\tilde{r}^2} \sup_{f \in \mathcal{F} \cap |f| \leq \sqrt{96L^2 K/(\theta^2 N)}} \sum_{i \in \mathcal{K} \setminus \mathcal{K}} \sigma_i \left( f - f^* \right) \left( X_i \right) \right] \leq \frac{1}{\tilde{r}^2} \sup_{f \in \mathcal{F} \cap |f| \leq \sqrt{96L^2 K/(\theta^2 N)}} \sum_{i \in \mathcal{K} \setminus \mathcal{K}} \sigma_i \left( f - f^* \right) \left( X_i \right). \]
Plugging this inequality into (28) shows that, with probability at least 1 − e^{−|K|/288},
\[ z(f) \geq \frac{7|K|}{8}. \]
As \( K \geq 7|O|/3, |K| \geq K/4K/7 \), hence, \( z(f) \geq K/2 \) holds with probability at least \( 1 − e^{−K/504} \).

C Proof Bernstein Assumption for the Hinge loss functions

Let \( f \) be in \( F \) such that \( \|f − f^*\|_{L_2} \leq r \). For all \( x \) in \( X \) let us denote \( \eta(x) = \mathbb{P}(Y = 1|X = x) \). It is easy to verify that the Bayes estimator i.e the oracle \( f^* \) is defined as \( f^*(x) = \text{sign}(2\eta(x) − 1) \). Consider the set \( A = \{x \in X, |f(x) − f^*(x)| \leq 2c^2r\} \). Since \( \|f − f^*\|_{L_2} \leq r \), by Markov’s inequality \( \mathbb{P}(X \in A) \geq 1 − 1/(4c^4) \).

Let \( x \) be in \( A \). If \( f^*(x) = −1 \) (i.e \( 2\eta(x) − 1 \geq 0 \)) and \( f(x) \leq f^*(x) = −1 \) we obtain
\[ \mathbb{E}[\ell_f(X, Y)|X = x] − \mathbb{E}[\ell_{f^*}(X, Y)|X = x] = \eta(x)(1 − f(x)) − \eta(x)(1 − f^*(x)) \geq \eta(x)(f(x) − f^*(x))^2 \]
where we used the fact that on \( A, |f(x) − f^*(x)| \leq 2c^2r \leq 1 \). Using the same analysis for the other cases we get that
\[ \mathbb{E}[\ell_f(X, Y)|X = x] − \mathbb{E}[\ell_{f^*}(X, Y)|X = x] \geq \min(\eta(x), 1 − \eta(x), 1 − 2\eta(x))(f(x) − f^*(x))^2 \]
\[ \geq \alpha(f(x) − f^*(x))^2 \]

Therefore,
\[ \frac{P\mathcal{L}_f}{\alpha} \geq \mathbb{E}[I_A(X)(f(X) − f^*(X))^2] = \|f − f^*\|_{L_2}^2 − \mathbb{E}[I_{A^c}(X)(f(X) − f^*(X))^2] . \]
(30)

By Cauchy-Schwarz and Markov’s inequalities,
\[ \mathbb{E}[I_{A^c}(X)(f(X) − f^*(X))^2] \leq \sqrt{\mathbb{E}[I_{A^c}(X)]\mathbb{E}[(f(X) − f^*(X))^4]} \leq \frac{\|f − f^*\|_{L_4}^2}{2c^2} . \]

By Assumption 9, it follows that \( \mathbb{E}[I_{A^c}(X)(f(X) − f^*(X))^2] \leq \frac{\|f − f^*\|_{L_4}^2}{2c^2} \) and we conclude with (30).

D Proofs other Lemmas

D.1 Proof Lemma 2

Proof. The computation of the Gaussian mean-width necessitates to identify the canonical centered Gaussian process indexed by \( H \) defined in Definition 2. Let us consider \( G_h = \sum_{k=1}^{\infty} \xi_k \langle h, \phi_k \rangle_{L_2(X)} \) where \( (\xi_k)_{k=1}^{\infty} \) is a sequence of i.i.d standard gaussian random variables. From the fact that \( (\phi_k)_k \) and \( (\Phi_k)_k \) are orthonormal basis in respectively \( L_2(X) \) and \( \mathcal{H}_K \), it is easy to verify that \( \mathbb{E}(G_{h_1} − G_{h_2})^2 = \mathbb{E}(h_1(X) − h_2(X))^2 \) and thus that it defines the canonical Gaussian process indexed by \( H \). As a consequence, \( w(\mathcal{H}_K \cap B_{\mathcal{H}_K} \cap B_{L_2}) = \sup_{h \in \mathcal{H}_K \cap B_{\mathcal{H}_K} \cap B_{L_2}} \sum_{k=1}^{\infty} \xi_k \langle h, \phi_k \rangle_{L_2(X)} \).

Since \( h \) can be decomposed as \( h = \sum_{k=1}^{\infty} \beta_k \Phi_k = \sum_{k=1}^{\infty} \beta_k \sqrt{\lambda_k} \phi_k \) where \( \beta_k = \langle h, \phi_k \rangle \) we get \( \langle h, \phi_k \rangle_{L_2(X)} = \beta_k \sqrt{\lambda_k} \). Finally, from the fact that \( B_{\mathcal{H}_K} = \{f_3 \beta = \langle \beta, \Phi \rangle_{l_2} : \|\beta\|_{l_2} \leq 1 \} \) and the fact that \( B_{L_2} \cap \mathcal{H}_K = \{f \in \mathcal{H}_K : \|f\|_{L_2} \leq 1 \} \) we finally obtain that
\[ w(\mathcal{H}_K \cap B_{\mathcal{H}_K} \cap B_{L_2}) \leq \mathbb{E} \sup_{\beta \in \ell_2} \sum_{k=1}^{\infty} \beta_k^2(1/\lambda_k) \leq \mathbb{E} \sup_{\beta \in \ell_2} \sum_{k=1}^{\infty} \beta_k^2(1/\lambda_k) \leq \mathbb{E} \sup_{\beta \in \ell_2} \sum_{k=1}^{\infty} \xi_k \beta_k \]
The computation of the gaussian mean-width of the unit ball of \( \mathcal{H}_K \cap B_{\mathcal{H}_K} \cap B_{L_2} \) is thus equivalent to the computation of the gaussian mean width of an ellipsoid in \( \ell_2 \). From Theorem 2.2.1 in [21] we finally get the result.


D.2 Supplementary Lemmas

**Lemma 6.** Let $\gamma > 0$ and $f$ in $F$ such that $\phi(f - f^*) \geq \gamma$. Then, there exist $f_0$ in $F$ and $1 \leq \alpha \leq \phi(f - f^*)/\gamma$ such that $f = f^* + \alpha(f_0 - f^*)$ and $\phi(f_0 - f^*) = \gamma$

**Proof.** Let $\alpha_0 = \sup\{\alpha > 0, \phi(\alpha(f - f^*)) \leq \gamma\}$. For $\alpha = \gamma/\phi(f - f^*) \leq 1$ we have $\phi(\alpha(f - f^*)) \leq \alpha\phi(f - f^*) = \gamma$ so that $\alpha_0 \geq \gamma/\phi(f - f^*)$. By convexity of $F$, $f_0 := f^* + \alpha_0(f - f^*) \in F$ and $\alpha_0 \leq 1$ otherwise, by convexity of $\phi$ we would have $\alpha_0\phi(f - f^*) \leq \phi(\alpha_0(f - f^*)) \leq \gamma$. Moreover, by maximality of $\alpha_0$, $f_0$ is such that $\phi(\alpha(f - f^*)) = \phi(f_0 - f^*) = \gamma$. The result follows for $\alpha = \alpha_0^{-1}$.

**Lemma 7.** Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a convex function. Then for all $\lambda \geq 1$ and $x, y$ in $\mathbb{R}$:

$$f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y)$$  \hspace{1cm} (31)

**Proof.** Let $\lambda \geq 1$, by convexity of $f$, for all $x, y$ in $\mathbb{R}$:

$$f\left(\frac{1}{\lambda}x + (1 - \frac{1}{\lambda})y\right) \leq \frac{1}{\lambda}f(x) + (1 - \frac{1}{\lambda})f(y)$$

It suffice to take $x = \lambda x + (1 - \lambda)y$ to get the result.