arXiv: arXiv:1711.10306

SUPPLEMENTARY MATERIAL TO "ROBUST MACHINE LEARNING BY MEDIAN-OF-MEANS : THEORY AND PRACTICE"

By Guillaume Lecué[*],[†] and Matthieu Lerasle[‡]

CREST, CNRS, ENSAE[†] *and CNRS, Université Paris Sud Orsay* [‡]

The supplementary material is organized as follows:

- In Section 6, we provide the proofs of Theorem 1 and Theorem 2.
- In Section 7, we introduce minmax and maxmin MOM estimators for the problem of learning without a priori regularization. We study its statistical properties such as estimation bounds and sharp oracle inequalities. We apply these results to the example of Ordinary least squares.
- In Section 8, we state a minimax optimality of our results.

6. Proofs of the main results. Recall the quadratic / multiplier decomposition of the difference of losses: for all $f, g \in F$, $x \in \mathcal{X}$ and $y \in \mathbb{R}$,

$$\begin{aligned} \ell_f(x,y) - \ell_g(x,y) &= (y - f(x))^2 - (y - g(x))^2 \\ &= (f(x) - g(x))^2 + 2(y - g(x))(g(x) - f(x)). \end{aligned} \tag{15}$$

Upper and lower bounds on $T_K(\cdot,\cdot)$ follow from a study of "quadratic" and "multiplier" quantiles of means processes. As no assumption is granted on the outliers, any block of data containing one or more of these outliers is "lost" from our perspective meaning that empirical means over these blocks cannot be controlled. Let $\mathcal{K}$ denote the set of blocks which have not been corrupted by outliers:

$$\mathcal{K} = \{k \in [K] : B_k \subset \mathcal{I}\}\,. \tag{16}$$

If $k \in \mathcal{K}$, all data indexed by B_k are informative. We will show that controls on the blocks indexed by $\mathcal{K}$ are sufficient to demonstrate statistical performance of MOM estimators.

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

6.1. *Bounding quadratic and multiplier processes.* The following lemmas are the only two "stochastic tools" needed to control the performance of minmax MOM estimators. There is in particular no need to estimate the L_P^2 geometry over F to study minmax MOM estimators. The two following lemmas have already been proved in Lemma 1 and Lemma 2 in [3] and can also be obtained in the i.i.d. setup under similar assumptions using Lemmas 5.1 and 5.5 in [8], see [7]. We reproduce here the proof of these technical lemmas for the sake of completeness. The first result is a lower bound on the quantiles of means quadratic processes.

LEMMA 3. *Grant Assumptions 1 and 3. Fix $\eta \in (0,1)$, $\rho \in (0,+\infty]$ and let $\alpha, \gamma, \gamma_Q, x$ be positive numbers such that $\gamma\left(1-\alpha-x-16\gamma_Q\theta_0\right) \geqslant 1-\eta$. Assume that $K \in [|\mathcal{O}|/(1-\gamma), N\alpha/4\theta_0^2]$. Then there exists an event $\Omega_Q(K)$ such that $\mathbb{P}\left(\Omega_Q(K)\right) \geqslant 1-\exp\left(-K\gamma x^2/2\right)$ and, on $\Omega_Q(K)$: for all $f \in F$ such that $\|f-f^*\| \leqslant \rho$, if $\|f-f^*\|_{L_P^2} \geqslant r_Q(\rho, \gamma_Q)$ then*

$$\left|\left\{k \in [K] : P_{B_k}(f-f^*)^2 \geqslant (4\theta_0)^{-2}\left\|f-f^*\right\|_{L_P^2}^2\right\}\right| \geqslant (1-\eta)K \enspace .$$

In particular, $Q_{\eta,K}((f-f^*)^2) \geqslant (4\theta_0)^{-2}\left\|f-f^*\right\|_{L_P^2}^2$.

PROOF. Define $F_\rho^* = B(f^*,\rho) = \{f \in F : \|f-f^*\| \leqslant \rho\}$. For all $f \in F_\rho^*$, let $n_f = (f-f^*)/\left\|f-f^*\right\|_{L_P^2}$ and note that for all $i \in \mathcal{I}$, $P_i|n_f| \geqslant \theta_0^{-1}$ by Assumption 3 and $P_i n_f^2 = 1$ by Assumption 1. It follows from Markov's inequality that, for all $k \in \mathcal{K}$ ($\mathcal{K}$ is defined in (16)),

$$\mathbb{P}\left(|(P_{B_k}-P)|n_f||>\frac{1}{\sqrt{\alpha|B_k|}}\right) \leqslant \alpha \enspace .$$

As $P|n_f| \geqslant \theta_0^{-1}$,

$$\mathbb{P}\left(P_{B_k}|n_f| \geqslant \frac{1}{\theta_0}-\frac{1}{\sqrt{\alpha|B_k|}}\right) \geqslant 1-\alpha \enspace .$$

Since $K \leqslant N\alpha/4\theta_0^2$, $|B_k| = N/K \geqslant 4\theta_0^2/\alpha$ and so

$$\mathbb{P}\left(2\theta_0 P_{B_k}|n_f| \geqslant 1\right) \geqslant 1-\alpha \enspace . \tag{17}$$

Let ϕ be the function defined on $\mathbb{R}_+$ by $\phi(t) = (t-1)I(1 \leqslant t \leqslant 2) + I(t \geqslant 2)$, and, for all $f \in F_\rho^*$ let $Z(f) = \sum_{k\in[K]} I(4\theta_0 P_{B_k}|n_f| \geqslant 1)$. Since for all $x \in \mathbb{R}$, $I(x \geqslant 1) \geqslant \phi(x)$,

$$Z(f) \geqslant \sum_{k\in\mathcal{K}} \phi\left(4\theta_0 P_{B_k}|n_f|\right) \enspace .$$

Now, for any $x \in \mathbb{R}_+$, $\phi(x) \geqslant I(x \geqslant 2)$, thus, according to (17),

$$\mathbb{E}\left[\sum_{k\in\mathcal{K}} \phi\left(4\theta_0 P_{B_k}|n_f|\right)\right] \geqslant \sum_{k\in\mathcal{K}} \mathbb{P}\left(4\theta_0 P_{B_k}|n_f| \geqslant 2\right) \geqslant |\mathcal{K}|(1-\alpha) \ .$$

Therefore,

$$Z(f) \geqslant |\mathcal{K}|(1-\alpha) + \sum_{k\in\mathcal{K}} \left(\phi\left(4\theta_0 P_{B_k}|n_f|\right) - \mathbb{E}\left[\phi\left(4\theta_0 P_{B_k}|n_f|\right)\right]\right) \ .$$

Denote $\mathcal{F} = \{f \in F : \|f - f^*\| \leqslant \rho, \ \|f - f^*\|_{L_P^2} \geqslant r_Q(\rho, \gamma_Q)\}$. By the bounded difference inequality (see, for instance [1, Theorem 6.2]), there exists an event $\Omega_Q(K)$ with probability larger than $1 - \exp(-x^2|\mathcal{K}|/2)$, on which, for all $f \in \mathcal{F}$,

$$\begin{aligned}&\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}} \left(\phi\left(4\theta_0 P_{B_k}|n_f|\right) - \mathbb{E}\left[\phi\left(4\theta_0 P_{B_k}|n_f|\right)\right]\right)\right| \\ &\leqslant \mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}} \left(\phi\left(4\theta_0 P_{B_k}|n_f|\right) - \mathbb{E}\left[\phi\left(4\theta_0 P_{B_k}|n_f|\right)\right]\right)\right| + |\mathcal{K}|x \ .\end{aligned}$$

By the symmetrization argument,

$$\begin{aligned}&\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}} \left(\phi\left(4\theta_0 P_{B_k}|n_f|\right) - \mathbb{E}\left[\phi\left(4\theta_0 P_{B_k}|n_f|\right)\right]\right)\right| \\ &\leq 2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}} \epsilon_k \phi\left(4\theta_0 P_{B_k}|n_f|\right)\right| \ .\end{aligned}$$

Since the function ϕ is 1-Lipschitz and $\phi(0) = 0$, by the contraction principle (see, for example [6, Chapter 4] or [1, Theorem 11.6]), we have

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}} \epsilon_k \phi\left(4\theta_0 P_{B_k}|n_f|\right)\right| \leqslant 4\theta_0 \mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}} \epsilon_k P_{B_k}|n_f|\right| .$$

The family $(\epsilon_{[i]}|n_f(X_i)| : i \in \cup_{k\in\mathcal{K}} B_k)$, where $[i] = \lceil i/K \rceil$ for all $i \in \mathcal{I}$, is a collection of centered random variables. Therefore, if $(\epsilon'_k)_{k\in\mathcal{K}}$ and $(X'_i)_{i\in\mathcal{I}}$ denote independent copies of $(\epsilon_k)_{k\in\mathcal{K}}$ and $(X_i)_{i\in\mathcal{I}}$ then

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}} \epsilon_k P_{B_k}|n_f|\right| \leqslant \mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}} \frac{1}{|B_k|}\sum_{i\in B_k} \epsilon_k|n_f(X_i)| - \epsilon'_k|n_f(X'_i)|\right| .$$

Then, as $(X_i)_{i\in\mathcal{I}}$ and $(X'_i)_{i\in\mathcal{I}}$ are two independent families of independent variables therefore, if $(\epsilon''_i)_{i\in\mathcal{I}}$ denote a family of i.i.d. Rademacher variables independent of $(\epsilon_i), (\epsilon'_i), (X_i)_{i\in\mathcal{I}}, (X'_i)_{i\in\mathcal{I}}$ then $(\epsilon_k|n_f(X_i)| - \epsilon'_k|n_f(X'_i)|)$ and $(\epsilon''_i\,(\epsilon_k|n_f(X_i)| - \epsilon'_k|n_f(X'_i)|))$ have the same distribution. Therefore,

$$\begin{aligned}
&\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}}\frac{1}{|B_k|}\sum_{i\in B_k}\epsilon_k|n_f(X_i)| - \epsilon'_k|n_f(X'_i)|\right|\\
&\leqslant \mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}}\frac{1}{|B_k|}\sum_{i\in B_k}\epsilon''_i\left(\epsilon_k|n_f(X_i)| - \epsilon'_k|n_f(X'_i)|\right)\right|\\
&= \mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}}\frac{1}{|B_k|}\sum_{i\in B_k}\epsilon''_i\left(|n_f(X_i)| - |n_f(X'_i)|\right)\right|\\
&\leqslant \frac{2K}{N}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\epsilon_i n_f(X_i)\right| .
\end{aligned}$$

By the contraction principle, on $\Omega_Q(K)$,

$$Z(f) \geqslant |\mathcal{K}|\left(1-\alpha-x-\frac{16\theta_0 K}{|\mathcal{K}|N}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\epsilon_i n_f(X_i)\right|\right) . \tag{18}$$

For any $f\in\mathcal{F}$, $r_Q(\rho,\gamma_Q)n_f + f^*\in F$ because F is convex. Moreover, $\|r_Q(\rho,\gamma_Q)n_f\|_{L^2_P} = r_Q(\rho,\gamma_Q)$ and

$$\|r_Q(\rho,\gamma_Q)n_f\| = [r_Q(\rho,\gamma_Q)/\left\|f-f^*\right\|_{L^2_P}]\left\|f-f^*\right\| \leqslant \rho.$$

Therefore, $r_Q(\rho,\gamma_Q)n_f + f^*\in\mathcal{F}$ and by definition of $r_Q(\rho,\gamma_Q)$,

$$\begin{aligned}
&\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\epsilon_i n_f(X_i)\right|\\
&= \frac{1}{r_Q(\rho,\gamma_Q)}\mathbb{E}\sup_{f\in F:\|f-f^*\|\leqslant\rho,\ \|f-f^*\|_{L^2_P}=r_Q(\rho,\gamma_Q)}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\epsilon_i(f-f^*)(X_i)\right|\\
&\leqslant \gamma_Q\frac{|\mathcal{K}|N}{K} .
\end{aligned}$$

Using the last inequality together with (18) and the assumption $K\geqslant|\mathcal{O}|/(1-\gamma)$ (so that $|\mathcal{K}|\geqslant K-|\mathcal{O}|\geqslant\gamma K$), we get that, on the event $\Omega_Q(K)$, for any $f \in \mathcal{F}$,

$$Z(f) \geqslant |\mathcal{K}| \left(1-\alpha-x-16\theta_0\gamma_Q\right) \geqslant (1-\eta)K \ .$$

Hence, on $\Omega_Q(K)$, for any $f \in \mathcal{F}$, there exists at least $(1-\eta)K$ blocks B_k for which $P_{B_k}|n_f| \geqslant (4\theta_0)^{-1}$. On these blocks, $P_{B_k}n_f^2 \geqslant (P_{B_k}|n_f|)^2 \geqslant (4\theta_0)^{-2}$, therefore, on $\Omega_Q(K)$, $Q_{\eta,K}[n_f^2] \geqslant (4\theta_0)^{-2}$. □

Now, let us turn to a control of the multiplier process.

Lemma 4. *Grant Assumption 2. Fix $\eta \in (0,1)$, $\rho \in (0,+\infty]$, and let $\alpha, \gamma_M, \gamma, x$ and ϵ be positive absolute constants such that $\gamma\left(1-\alpha-x-8\gamma_M/\epsilon\right) \geqslant 1-\eta$. Let $K \in [|\mathcal{O}|/(1-\gamma), N]$. There exists an event $\Omega_M(K)$ such that $\mathbb{P}(\Omega_M(K)) \geqslant 1-\exp(-\gamma K x^2/2)$ and on the event $\Omega_M(K)$: if $f \in F$ is such that $\|f-f^*\| \leqslant \rho$ then the number of elements $k \in \mathcal{K}$ such that*

$$|2(P_{B_k}-P)(\zeta(f-f^*))| \leqslant \epsilon \max\left(\frac{16\theta_m^2}{\epsilon^2\alpha}\frac{K}{N}, r_M^2(\rho,\gamma_M), \|f-f^*\|_{L_P^2}^2\right)$$

is at least $(1-\eta)K$.

Proof. For all $k \in [K]$ and $f \in F$, set $W_k = ((X_i,Y_i))_{i\in B_k}$ and define

$$g_f(W_k) = 2(P_{B_k}-P)\left(\zeta(f-f^*)\right)$$

and

$$\gamma_k(f) = \epsilon \max\left(\frac{16\theta_m^2}{\epsilon^2\alpha}\frac{K}{N}, r_M^2(\rho,\gamma_M), \|f-f^*\|_{L_P^2}^2\right) \ .$$

Let $f \in F$ and $k \in \mathcal{K}$. It follows from Markov's inequality that

$$\mathbb{P}\left[2\left|g_f(W_k)\right| \geqslant \gamma_k(f)\right] \leqslant \frac{4\mathbb{E}\left[\left(2(P_{B_k}-P)(\zeta(f-f^*))\right)^2\right]}{\frac{16\theta_m^2}{\alpha}\|f-f^*\|_{L_P^2}^2\frac{K}{N}}$$

$$\leqslant \frac{\alpha\sum_{i\in B_k}\mathrm{var}_{P_i}(\zeta(f-f^*))}{|B_k|^2\theta_m^2\|f-f^*\|_{L_P^2}^2\frac{K}{N}} \leqslant \frac{\alpha\theta_m^2\|f-f^*\|_{L_P^2}^2}{|B_k|\theta_m^2\|f-f^*\|_{L_P^2}^2\frac{K}{N}} = \alpha \ . \tag{19}$$

Let $J = \cup_{k\in\mathcal{K}} B_k$ and let $r_M(\rho) = r_M(\rho, \gamma_M)$. We have

$$
\begin{aligned}
&\mathbb{E}\sup_{f\in B(f^*,\rho)}\sum_{k\in\mathcal{K}}\epsilon_k\frac{g_f(W_k)}{\gamma_k(f)} \leqslant 2\mathbb{E}\sup_{f\in B(f^*,\rho)}\left|\sum_{k\in\mathcal{K}}\frac{\epsilon_k(P_{B_k}-P)(\zeta(f-f^*))}{\epsilon\max(r_M^2(\rho),\|f-f^*\|_{L_P^2}^2)}\right|\\
&\leqslant \frac{2}{\epsilon r_M^2(\rho)}\mathbb{E}\left[\sup_{f\in B(f^*,\rho):\|f-f^*\|_{L_P^2}\geqslant r_M(\rho)}\left|\sum_{k\in\mathcal{K}}\epsilon_k(P_{B_k}-P)\left(\zeta r_M(\rho)\frac{f-f^*}{\|f-f^*\|_{L_P^2}}\right)\right|\right.\\
&\qquad\qquad\left.\vee\sup_{f\in B(f^*,\rho):\|f-f^*\|_{L_P^2}\leqslant r_M(\rho)}\left|\sum_{k\in\mathcal{K}}\epsilon_k(P_{B_k}-P)\left(\zeta(f-f^*)\right)\right|\right]\\
&\leqslant \frac{2}{\epsilon r_M^2(\rho)}\mathbb{E}\sup_{f\in B(f^*,\rho):\|f-f^*\|_{L_P^2}\leqslant r_M(\rho)}\left|\sum_{k\in\mathcal{K}}\epsilon_k(P_{B_k}-P)\left(\zeta(f-f^*)\right)\right| \ ,
\end{aligned}
$$

where in the last but one inequality, we used that the class F is convex and the same argument as in the proof of Lemma 3. Since $(\epsilon_{[i]}(\zeta_i(f-f^*)(X_i) - P_i\zeta_i(f-f^*)) : i \in \mathcal{I})$ is a family of centered random variables, one can use the symmetrization argument to get

$$
\begin{aligned}
&\mathbb{E}\sup_{f\in B(f^*,\rho)}\sum_{k\in\mathcal{K}}\epsilon_k\frac{g_f(W_k)}{\gamma_k(f)}\\
&\leqslant \frac{4K}{\epsilon r_M^2(\rho)N}\mathbb{E}\sup_{f\in B(f^*,\rho):\|f-f^*\|_{L_P^2}\leqslant r_M(\rho)}\left|\sum_{i\in J}\epsilon_i\zeta_i(f-f^*)(X_i)\right|\\
&\leqslant \frac{4K}{\epsilon N}\gamma_M|\mathcal{K}|\frac{N}{K} = \frac{4\gamma_M}{\epsilon}|\mathcal{K}| \ , \qquad (20)
\end{aligned}
$$

where the definition of $r_M(\rho)$ has been used in the last but one inequality.

Let $\psi(t) = (2t-1)I(1/2 \leqslant t \leqslant 1) + I(t \geqslant 1)$. The function ψ is 2-Lipschitz and satisfies $I(t \geqslant 1) \leqslant \psi(t) \leqslant I(t \geqslant 1/2)$, for all $t \in \mathbb{R}$. Therefore, all

$f \in B(f^*, \rho)$ satisfies

$$\begin{aligned}
&\sum_{k\in\mathcal{K}} I\left(|g_f(W_k)| < \gamma_k(f)\right) \\
&= |\mathcal{K}| - \sum_{k\in\mathcal{K}} I\left(\frac{|g_f(W_k)|}{\gamma_k(f)} \geqslant 1\right) \geqslant |\mathcal{K}| - \sum_{k\in\mathcal{K}} \psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right) \\
&= |\mathcal{K}| - \sum_{k\in\mathcal{K}} \mathbb{E}\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right) - \sum_{k\in\mathcal{K}} \left[\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right) - \mathbb{E}\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right)\right] \\
&\geqslant |\mathcal{K}| - \sum_{k\in\mathcal{K}} \mathbb{E}I\left(\frac{|g_f(W_k)|}{\gamma_k(f)} \geqslant \frac{1}{2}\right) - \sum_{k\in\mathcal{K}} \left[\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right) - \mathbb{P}\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right)\right] \\
&\geqslant (1-\alpha)|\mathcal{K}| - \sup_{f\in B(f^*,\rho)} \left|\sum_{k\in\mathcal{K}} \left[\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right) - \mathbb{E}\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right)\right]\right|
\end{aligned}$$

where we used (19) in the last inequality.

The bounded difference inequality ensures that there exists an event $\Omega_M(K)$ satisfying $\mathbb{P}(\Omega_M(K)) \geqslant 1 - \exp(-x^2|\mathcal{K}|/2)$, where

$$\begin{aligned}
&\sup_{f\in B(f^*,\rho)} \left|\sum_{k\in\mathcal{K}} \left[\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right) - \mathbb{E}\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right)\right]\right| \\
&\qquad \leqslant \mathbb{E} \sup_{f\in B(f^*,\rho)} \left|\sum_{k\in\mathcal{K}} \left[\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right) - \mathbb{E}\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right)\right]\right| + |\mathcal{K}|x \ .
\end{aligned}$$

Furthermore, it follows from by the symmetrization argument that

$$\begin{aligned}
&\mathbb{E} \sup_{f\in B(f^*,\rho)} \left|\sum_{k\in\mathcal{K}} \left[\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right) - \mathbb{E}\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right)\right]\right| \\
&\leqslant 2\mathbb{E} \sup_{f\in B(f^*,\rho)} \left|\sum_{k\in\mathcal{K}} \epsilon_k \psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right)\right|
\end{aligned}$$

and, from the contraction principle and (20), that

$$\mathbb{E} \sup_{f\in B(f^*,\rho)} \left|\sum_{k\in\mathcal{K}} \epsilon_k \psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right)\right| \leqslant 2\mathbb{E} \sup_{f\in B(f^*,\rho)} \left|\sum_{k\in\mathcal{K}} \epsilon_k \frac{|g_f(W_k)|}{\gamma_k(f)}\right| \leqslant \frac{8\gamma_M}{\epsilon}|\mathcal{K}| \ .$$

In conclusion, on $\Omega_M(K)$, for all $f \in B(f^*, \rho)$,

$$\begin{aligned}
&\sum_{k\in\mathcal{K}} I\left(|g_f(W_k)| < \gamma_k(f)\right) \geqslant \left(1 - \alpha - x - 8\gamma_M/\epsilon\right)|\mathcal{K}| \\
&\geqslant K\gamma\left(1 - \alpha - x - 8\gamma_M/\epsilon\right) \geqslant (1-\eta)K \ .
\end{aligned}$$

□

6.2. *Bounding the empirical criterion $\mathcal{C}_{K,\lambda}(f^*)$.* Let us first introduce the event on which the statement of Theorem 1 holds. Denote by $\Omega(K)$ the intersection of the events $\Omega_Q(K)$, $\Omega_M(K)$ defined respectively in Lemmas 3 and 4 for $\rho \in \{\kappa\rho_K : \kappa \in \{1,2\}\}$ and

$$\eta = \frac{1}{4}, \gamma = \frac{7}{8}, \alpha = \frac{1}{24}, x = \frac{1}{24}, \gamma_Q = \frac{1}{384\theta_0}, \epsilon = \frac{1}{c\theta_0^2} \text{ and } \gamma_M = \frac{\epsilon}{192} \tag{21}$$

for some absolute constants $c > 0$ to be specified later. For these values, conditions in both Lemmas 3 and 4 are satisfied:

$$\gamma(1-\alpha-x-16\gamma_Q\theta_0) \geqslant 1-\eta = \frac{3}{4} \text{ and } \gamma(1-\alpha-x-8\gamma_M/\epsilon) \geqslant 1-\eta = \frac{3}{4}.$$

According to Lemmas 3 and 4, the event $\Omega(K)$ satisfies $\mathbb{P}(\Omega(K)) \geqslant 1 - 4\exp(-7K/9216)$. On $\Omega(K)$, the following holds for all $\rho \in \{\kappa\rho_K : \kappa \in \{1,2\}\}$ and $f \in F$ such that $\|f-f^*\| \leqslant \rho$,

1. if $\|f-f^*\|_{L_P^2} \geqslant r_Q(\rho,\gamma_Q)$ then
$$Q_{1/4,K}((f-f^*)^2) \geqslant \frac{1}{(4\theta_0)^2}\|f-f^*\|_{L_P^2}^2 \quad , \tag{22}$$
2. there exists $3K/4$ block B_k with $k \in \mathcal{K}$, for which
(23)
$$|(P_{B_k}-P)[2\zeta(f-f^*)]| \leqslant \epsilon \max\left(r_M^2(\rho,\gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f-f^*\|_{L_P^2}^2\right) \quad .$$

Moreover, on the blocks B_k where (23) holds, it follows that all $f \in F$ such that $\|f-f^*\| \leqslant \rho$ satisfies

$$P_{B_k}[2\zeta(f-f^*)]| \leqslant P[2\zeta(f-f^*)] + \epsilon \max\left(r_M^2(\rho,\gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f-f^*\|_{L_P^2}^2\right) \quad .$$

It follows from the convexity of F and the nearest point theorem that $P[2\zeta(f-f^*)] \leqslant 0$ for all $f \in F$, therefore, for all $f \in F$ such that $\|f-f^*\| \leqslant \rho$,

$$Q_{3/4,K}(2\zeta(f-f^*)) \leqslant \epsilon \max\left(r_M^2(\rho,\gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f-f^*\|_{L_P^2}^2\right) \quad . \tag{24}$$

Moreover, still on the blocks B_k where (23) holds, one also has that for all $f \in F$ such that $\|f-f^*\| \leqslant \rho$,

$$P[-2\zeta(f-f^*)] \leqslant P_{B_k}[-2\zeta(f-f^*)] + \epsilon \max\left(r_M^2(\rho,\gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f-f^*\|_{L_P^2}^2\right) \quad .$$

It follows that, for all $f \in F$ such that $\|f - f^*\| \leqslant \rho$,

$$
\begin{aligned}
&P[-2\zeta(f-f^*)] \leqslant Q_{1/4,K}[-2\zeta(f-f^*)] + \epsilon \max\left(r_M^2(\rho,\gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f-f^*\|_{L_p^2}^2\right)\\
&\leqslant Q_{1/4,K}[(f-f^*)^2 - 2\zeta(f-f^*)] + \lambda(\|f\| - \|f^*\|)\\
&\quad + \epsilon \max\left(r_M^2(\rho,\gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f-f^*\|_{L_p^2}^2\right) + \lambda\rho
\end{aligned}
$$

(25)

$$
\leqslant T_{K,\lambda}(f^*, f) + \epsilon \max\left(r_M^2(\rho,\gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f-f^*\|_{L_p^2}^2\right) + \lambda\rho \ .
$$

The main result of this section is Lemma 5. It will be used to bound from above the criterion $\mathcal{C}_{K,\lambda}(f^*) = \sup_{g\in F} T_{K,\lambda}(g, f^*)$. Recall that ρ_K and λ are defined as

$$
r^2(\rho_K) = \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N} \text{ and } \lambda = \frac{c'\epsilon r^2(\rho_K)}{\rho_K} \tag{26}
$$

where $\epsilon = (c\theta_0^2)^{-1}$ and $c, c' \geqslant$ are absolute constants. We also need to consider a partition of the space F according to the distance between g and f^* w.r.t. $\|\cdot\|$ and $\|\cdot\|_{L_P^2}$ as in Figure 2: define for all $\kappa \geqslant 1$,

$$
\begin{aligned}
F_1^{(\kappa)} &= \left\{ g \in F : \|g - f^*\| \leqslant \kappa\rho_K \text{ and } \|g - f^*\|_{L_P^2} \leqslant r(\kappa\rho_K)\right\} \ ,\\
F_2^{(\kappa)} &= \left\{ g \in F : \|g - f^*\| \leqslant \kappa\rho_K \text{ and } \|g - f^*\|_{L_P^2} > r(\kappa\rho_K)\right\} \ ,\\
F_3^{(\kappa)} &= \{ g \in F : \|g - f^*\| > \kappa\rho_K\} \ .
\end{aligned}
$$

Lemma 5. *On the event $\Omega(K)$, it holds for all $\kappa \in \{1, 2\}$,*

$$
\sup_{g\in F_1^{(\kappa)}} T_{K,\lambda}(g, f^*) \leqslant (1 + c'\kappa)\epsilon r^2(\kappa\rho_K),
$$

$$
\sup_{g\in F_2^{(\kappa)}} T_{K,\lambda}(g, f^*) \leqslant \left((1 + c'\kappa)\epsilon - \frac{1}{16\theta_0^2}\right) r^2(\kappa\rho_K)
$$

and

$$
\sup_{g\in F_3^{(\kappa)}} T_{K,\lambda}(g, f^*) \leqslant \kappa \max\left(\epsilon - \frac{1}{16\theta_0^2} + \frac{11c'\epsilon}{10}, \epsilon - \frac{7c'\epsilon}{10}\right) r^2(\rho_K)
$$

when $c \geqslant 32$ and $10\epsilon/4 \leqslant c'\epsilon \leqslant ((4\theta_0)^{-2} - \epsilon)$.

Proof of Lemma 5. Recall that, for all $g \in F$, $\ell_{f^*} - \ell_g = 2\zeta(g - f^*) - (g - f^*)^2$ where $\zeta(x, y) = y - f^*(x)$. Let us now place ourself on the event $\Omega(K)$ up to the end of proof.

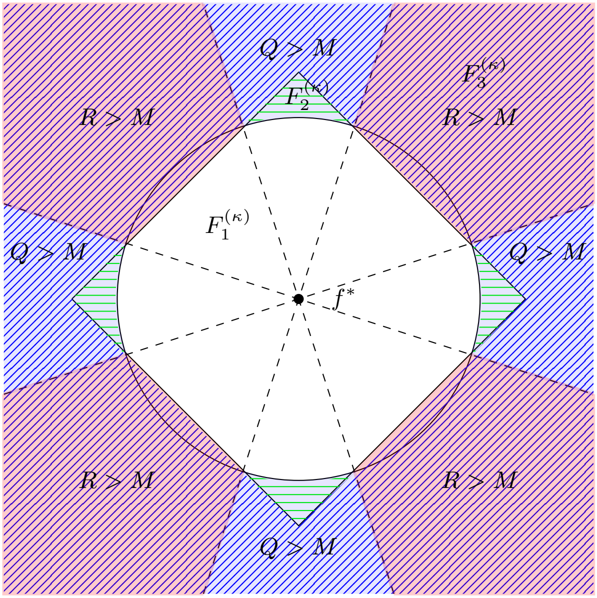

FIG 2. *Partition $\{F_1^{(\kappa)}, F_2^{(\kappa)}, F_3^{(\kappa)}\}$ of F and the control of the multiplier MOM process by either the quadratic MOM process (the "$Q > M$" part) or the regularization term (the "$R > M$" part).*

Bounding $\sup_{\mathbf{g}\in\mathbf{F_1^{(\kappa)}}} \mathbf{T_{K,\lambda}(g, f^*)}$. Let $g \in F_1^{(\kappa)}$. Since the quadratic process is non negative,

$$\begin{aligned}T_{K,\lambda}(g, f^*) &= \mathrm{MOM}_K\big(2\zeta(g-f^*) - (g-f^*)^2\big) - \lambda\left(\|g\| - \|f^*\|\right)\\ &\leqslant Q_{3/4,K}(2\zeta(g-f^*)) + \lambda\left\|f^* - g\right\| \ .\end{aligned}$$

Therefore, applying (24) for $\rho = \kappa\rho_K$ and the choice of ρ_K and λ as in (26), we get

$$\begin{aligned}T_{K,\lambda}(g, f^*) &\le \epsilon \max\left(r_M^2(\kappa\rho_K, \gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f - f^*\|_{L_p^2}^2\right) + \lambda\kappa\rho_K\\ &\leqslant \epsilon r^2(\kappa\rho_K) + c'\kappa\epsilon r^2(\rho_K) \leqslant (1 + c'\kappa)\epsilon r^2(\kappa\rho_K) \ .\end{aligned}$$

Bounding $\sup_{\mathbf{g}\in\mathbf{F_2^{(\kappa)}}} \mathbf{T_{K,\lambda}(g, f^*)}$. Let $g \in F_2^{(\kappa)}$. Given that $Q_{1/2}(x-y) \leqslant Q_{3/4}(x) - Q_{1/4}(y)$ for any vector x and y, we have

$$\begin{aligned}&\mathrm{MOM}_K\big(2\zeta(g-f^*) - (g-f^*)^2\big) + \lambda\left(\|f^*\| - \|g\|\right)\\ &\leqslant Q_{3/4,K}(2\zeta(g-f^*)) - Q_{1/4,K}((f^*-g)^2) + \lambda\kappa\rho_K \ .\end{aligned}$$

Moreover $2\epsilon \leqslant (4\theta_0)^{-2}$ when $c \geqslant 32$, so it follows from (22) and (24) for $\rho = \kappa\rho_K$ that

$$\begin{aligned}
&Q_{3/4,K}(2\zeta(f^*-g)) - Q_{1/4,K}((f^*-g)^2)\\
&\leqslant \epsilon \max\left(r_M^2(\kappa\rho_K,\gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f-f^*\|_{L_P^2}^2\right) - \frac{\|f-f^*\|_{L_P^2}^2}{(4\theta_0)^2}\\
&\leqslant \left(\epsilon - \frac{1}{(4\theta_0)^2}\right)\|f-f^*\|_{L_P^2}^2 \leqslant \left(\epsilon - \frac{1}{16\theta_0^2}\right) r^2(\kappa\rho_K) \ .
\end{aligned}$$

Putting both inequalities together and using that $\lambda\kappa\rho_K = c'\kappa\epsilon r^2(\rho_K)$, we get

$$T_{K,\lambda}(g,f^*) \leqslant \left((1+c'\kappa)\epsilon - \frac{1}{16\theta_0^2}\right) r^2(\kappa\rho_K) \ .$$

Bounding $\sup_{\mathbf{g}\in\mathbf{F}_3^{(\kappa)}} \mathbf{T}_{\mathbf{K},\lambda}(\mathbf{g},\mathbf{f}^*)$ *via an homogeneity argument.* Start with two lemmas.

Lemma 6. *Let $\rho \geqslant 0$, $\Gamma_{f^*}(\rho) = \cup_{f\in f^*+(\rho/20)B}(\partial\|\cdot\|)_f$ (cf.) section 3.3). For all $g \in F$,*

$$\|g\| - \|f^*\| \geqslant \sup_{z^*\in\Gamma_{f^*}(\rho)} z^*(g-f^*) - \frac{\rho}{10} \ .$$

Proof. Let $g \in F$, $f^{**} \in f^* + (\rho/20)B$ and $z^* \in (\partial\|\cdot\|)_{f^{**}}$. We have

$$\begin{aligned}
&\|g\| - \|f^*\| \geqslant \|g\| - \|f^{**}\| - \|f^{**}-f^*\| \geqslant z^*(g-f^{**}) - \frac{\rho}{20}\\
&= z^*(g-f^*) - z^*(f^{**}-f^*) - \frac{\rho}{20} \geqslant z^*(g-f^*) - \frac{\rho}{10} \ ,
\end{aligned}$$

where the last inequality follows from $z^*(f^{**}-f^*) \leqslant \|f^{**}-f^*\|$. The result follows by taking supremum over $z^* \in \Gamma_{f^*}(\rho)$. □

Lemma 7. *Let $\rho \geqslant 0$. Let $g \in F$ be such that $\|g-f^*\| \geqslant \rho$. Define $f = f^* + \rho(g-f^*)/\|g-f^*\|$. Then $f \in F$, $\|f-f^*\| = \rho$ and,*

$$\begin{aligned}
&MOM_K\big((g-f^*)^2 - 2\zeta(g-f^*)\big) + \lambda \sup_{z^*\in\Gamma_{f^*}(\rho)} z^*(g-f^*)\\
&\geqslant \frac{\|g-f^*\|_{L_P^2}}{\rho}\left(MOM_K\big((f-f^*)^2 - 2\zeta(f-f^*)\big) + \lambda \sup_{z^*\in\Gamma_{f^*}(\rho)} z^*(f-f^*)\right) \ .
\end{aligned}$$

Proof. The first conclusion holds by convexity of F, the second statement is obvious. For the last one, let $\Upsilon = \|g-f^*\|/\rho$ and note that $\Upsilon \geqslant 1$ and $g-f^* = \Upsilon(f-f^*)$, so we have

$$\begin{aligned}
&\mathrm{MOM}_K\big((g-f^*)^2 - 2\zeta(g-f^*)\big) + \lambda \sup_{z^*\in\Gamma_{f^*}(\rho)} z^*(g-f^*)\\
&= \mathrm{MOM}_K\big(\Upsilon^2(f-f^*)^2 - 2\Upsilon\zeta(f-f^*)\big) + \lambda\Upsilon \sup_{z^*\in\Gamma_{f^*}(\rho)} z^*(f-f^*)\\
&\geqslant \Upsilon\left(\mathrm{MOM}_K\big((f-f^*)^2 - 2\zeta(f-f^*)\big) + \lambda \sup_{z^*\in\Gamma_{f^*}(\rho)} z^*(f-f^*)\right) \enspace .
\end{aligned}$$

□

Now, let us bound $\sup_{g\in F_3^{(\kappa)}} T_{K,\lambda}(g,f^*)$. Let $g\in F_3^{(\kappa)}$. Apply Lemma 6 and Lemma 7 to $\rho=\rho_K$: there exists $f\in F$ such that $\|f-f^*\|=\rho_K$ and

$$\begin{aligned}
&T_{K,\lambda}(g,f^*) = \mathrm{MOM}_K\big(2\zeta(g-f^*) - (g-f^*)^2\big) - \lambda\left(\|g\| - \|f^*\|\right)\\
&\leqslant \mathrm{MOM}_K\big(2\zeta(g-f^*) - (g-f^*)^2\big) - \lambda \sup_{z^*\in\Gamma_{f^*}(\rho_K)} z^*(g-f^*) + \lambda\frac{\kappa\rho_K}{10}
\end{aligned}$$

(27)

$$\leqslant \frac{\|g-f^*\|}{\rho_K}\left(\mathrm{MOM}_K\big(2\zeta(f-f^*) - (f-f^*)^2\big) - \lambda \sup_{z^*\in\Gamma_{f^*}(\rho_K)} z^*(f-f^*)\right) + \lambda\frac{\kappa\rho_K}{10} \enspace .$$

First assume that $\|f-f^*\|_{L_P^2}\leqslant r(\rho_K)$. In that case, $\|f-f^*\|=\rho_K$ and $\|f-f^*\|_{L_P^2}\leqslant r(\rho_K)$ therefore, $f\in H_{\rho_K}$. Moreover, by definition of K^* and since $K\geqslant K^*$, we have $\rho_K\geqslant\rho^*$ which implies that ρ_K satisfies the sparsity equation from Definition 4. Therefore, $\sup_{z^*\in\Gamma_{f^*}(\rho_K)} z^*(f-f^*)\geqslant\Delta(\rho_K)\geqslant 4\rho_K/5$. Now, it follows from the definition of λ in (26) that

$$-\lambda \sup_{z^*\in\Gamma_{f^*}(\rho_K)} z^*(f-f^*) \leqslant -\frac{4c'\epsilon r^2(\rho_K)}{5} \enspace .$$

Moreover, since the quadratic process is non-negative, by (24) applied to $\rho=\rho_K$,

$$\begin{aligned}
&\mathrm{MOM}_K\big(2\zeta(f-f^*) - (f-f^*)^2\big) \leqslant Q_{3/4,K}[2\zeta(f-f^*)]\\
&\leqslant \epsilon\max\left(r_M^2(\rho_K,\gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f-f^*\|_{L_P^2}^2\right) \leqslant 2\epsilon r^2(\rho_K) \enspace .
\end{aligned}$$

Finally, noting that $\epsilon - 4c'\epsilon/5 \leqslant 0$ when $c' \geqslant 10/4$, binding all the pieces together in (27) yields

$$T_{K,\lambda}(g, f^*) \leqslant \kappa\epsilon\left(1 - 4c'/5\right) r^2(\rho_K) + \lambda\frac{\kappa\rho_K}{10} = \kappa\epsilon\left(1 - \frac{7c'}{10}\right) r^2(\rho_K) \enspace .$$

Second, assume that $\|f - f^*\|_{L_P^2} \geqslant r(\rho_K)$. Since $\|f - f^*\| = \rho_K$, it follows from (22) and (23) for $\rho = \rho_K$ that

$$\begin{aligned}
&\mathrm{MOM}_K\big(2\zeta(f-f^*) - (f-f^*)^2\big) \leqslant Q_{3/4,K}(2\zeta(f-f^*)) - Q_{1/4,K}((f^*-f)^2)\\
&\leqslant \epsilon \max\left(r_M^2(\rho_K,\gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f-f^*\|_{L_P^2}^2\right) - \frac{\|f-f^*\|_{L_P^2}^2}{(4\theta_0)^2}\\
&\leqslant \left(\epsilon - \frac{1}{16\theta_0^2}\right)\|f-f^*\|_{L_P^2}^2 \leqslant \left(2\epsilon - \frac{1}{16\theta_0^2}\right) r^2(\rho_K) \enspace ,
\end{aligned}$$

where we used that $\epsilon \leqslant (16\theta_0)^{-2}$ when $c \geqslant 32$ in the last inequality. Plugging the last result in (27) we get

$$\begin{aligned}
T_{K,\lambda}(g, f^*) &\leqslant \frac{\|g-f^*\|}{\rho_K}\left(\left(\epsilon - \frac{1}{16\theta_0^2}\right) r^2(\rho_K) + \lambda\rho_K\right) + \lambda\frac{\kappa\rho_K}{10}\\
&\leqslant \frac{\|g-f^*\|}{\rho_K}\left((1+c')\epsilon - \frac{1}{16\theta_0^2}\right) r^2(\rho_K) + \frac{c'\kappa\epsilon}{10} r^2(\rho_K)\\
&\leqslant \kappa\left(\left(1 + \frac{11c'}{10}\right)\epsilon - \frac{1}{16\theta_0^2}\right) r^2(\rho_K)
\end{aligned}$$

when $16(1+c')\epsilon \leqslant \theta_0^{-2}$.

6.3. *From a control of $\mathcal{C}_{K,\lambda}(\hat{f})$ to statistical performance.*

The proof follows essentially the one of [5, Theorem 3.2] or [3, Lemma 2].

LEMMA 8. *Let $\hat{f} \in F$ be such that, on $\Omega(K)$, $\mathcal{C}_{K,\lambda}(\hat{f}) \leqslant (1+c')\epsilon r^2(\rho_K)$. Then, on $\Omega(K)$, $\hat{f}$ satisfies*

$$\left\|\hat{f} - f^*\right\| \leqslant 2\rho_K, \quad \left\|\hat{f} - f^*\right\|_{L_P^2} \leqslant r(2\rho_K) \quad \textit{and} \quad R(\hat{f}) \leqslant R(f^*) + (1 + (2+3c')\epsilon) r^2(2\rho_K) \enspace ,$$

when $c' = 16$ and $c > 832$.

PROOF. Recall that for any $x \in \mathbb{R}^K$, $Q_{1/2}(x) \geqslant -Q_{1/2}(-x)$. Therefore,

$$\mathcal{C}_{K,\lambda}(\hat{f}) = \sup_{g\in F} T_{K,\lambda}(g, \hat{f}) \geqslant T_{K,\lambda}(f^*, \hat{f}) \geqslant -T_{K,\lambda}(\hat{f}, f^*) \enspace .$$

Thus, on $\Omega(K)$, $\hat{f} \in \big\{g \in F : T_{K,\lambda}(g, f^*) \geqslant -(1+c')\epsilon r^2(\rho_K)\big\}$. When $c' = 16$ and $c > 832$,

$$-(1+c')\epsilon > 2(1+c')\epsilon - \frac{1}{16\theta_0^2} \text{ and } -(1+c')\epsilon > 2\max\left(\epsilon - \frac{1}{16\theta_0^2} + \frac{11c'\epsilon}{10}, \epsilon - \frac{7c'\epsilon}{10}\right)$$

therefore, $\hat{f} \in F_1^{(2)}$ on $\Omega(K)$. This yields the results for both the regularization and the L_P^2-norm.

Finally, let us turn to the control on the excess risk. It follows from (25) for $\rho = \kappa\rho_K$ that

$$\begin{aligned}
&R(\hat{f}) - R(f^*) = \left\|\hat{f} - f^*\right\|_{L_P^2}^2 + P[-2\zeta(\hat{f} - f^*)]\\
&\leqslant r^2(2\rho_K) + T_{K,\lambda}(f^*, \hat{f}) + \epsilon\max\left(r_M^2(2\rho_K, \gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \left\|\hat{f} - f^*\right\|_{L_p^2}^2\right) + 2\lambda\rho_K\\
&\leqslant r^2(2\rho_K) + \mathcal{C}_{K,\lambda}\big(\hat{f}\big) + \epsilon r^2(2\rho_K) + c'\epsilon r^2(\rho_K) = (1 + (2+3c')\epsilon)r^2(2\rho_K)\ .
\end{aligned}$$

□

6.4. *End of the proof of Theorem 1.* By definition of $\widehat{f}_{K,\lambda}$,

$$\mathcal{C}_{K,\lambda}\big(\widehat{f}_{K,\lambda}\big) \leq \mathcal{C}_{K,\lambda}\big(f^*\big) = \sup_{g\in F} T_{K,\lambda}(g, f^*) \leq \max_{i\in[3]}\sup_{g\in F_i^{(1)}} T_{K,\lambda}(g, f^*),$$

where $\{F_1^{(1)}, F_2^{(1)}, F_3^{(1)}\}$ is the decomposition of F as in Figure 2. It follows from Lemma 5 (for $\kappa = 1$) that on the event $\Omega(K)$,

$$\mathcal{C}_{K,\lambda}\big(\widehat{f}_{K,\lambda}\big) \leqslant (1+c')\epsilon r^2(\rho_K)\ .$$

Therefore, for $c' = 16$ and $c = 833$ the conclusion of the proof of Theorem 1 follows from Lemma 8.

6.5. *Proof of Theorem 2.* Define

$$K_1 = \frac{|\mathcal{O}|}{1-\gamma} = 8|\mathcal{O}| \text{ and } K_2 = \frac{N\alpha}{2\theta_0^2} = \frac{N}{96\theta_0^2}.$$

Let $K \in [K_1, K_2]$ and let $\Omega_{K,c_{ad}} = \{f^* \in \cap_{J=K}^{K_2} \hat{R}_{J,c_{ad}}\}$ where we recall that $\hat{R}_{J,c_{ad}} = \{f \in F : \mathcal{C}_{J,\lambda}(f) \leqslant (c_{ad}/\theta_0^2)r^2(\rho_J)\}$. Lemma 5 (for $\kappa = 1$) shows that, for $c_{ad} = (1+c')/c$, $\Omega_{K,c_{ad}} \supset \cap_{J=K}^{K_2}\Omega(J)$. Therefore, on $\cap_{J=K}^{K_2}\Omega(J)$, $\hat{K}_{c_{ad}} \leqslant K$ which implies that $\widehat{f}_{c_{ad}} \in \hat{R}_{K,c_{ad}}$. By Lemma 8 (for $c' = 16$ and $c = 833$), this implies that

$$\left\|\widehat{f}_{c_{ad}} - f^*\right\| \leqslant 2\rho_K, \qquad \left\|\widehat{f}_{c_{ad}} - f^*\right\|_{L_P^2} \leqslant r(2\rho_K)$$

and

$$R(\widehat{f}_{c_{ad}}) \leqslant R(f^*) + (1 + (2 + 3c')\epsilon) r^2(2\rho_K) \ .$$

7. Learning without regularization: minmax and maxmin MOM procedures.

All the results from the previous sections also apply in the setup of learning with no regularization which is the framework one should consider when there is no a priori known structure on the oracle.

We consider the learning problem with no regularization. In this setup, we may use both minmaximization or maxminimization estimators

$$\widehat{f}_K \in \operatorname*{argmin}_{f \in F} \sup_{g \in F} T_K(g, f) \text{ and } \widehat{g}_K \in \operatorname*{argmax}_{g \in F} \inf_{f \in F} T_K(g, f) \tag{28}$$

where $T_K(g, f) = \mathrm{MOM}_K\big(\ell_f - \ell_g\big)$.

We show below that $\widehat{f}_K$ and $\widehat{g}_K$ are efficient procedures even in situations where the dataset is corrupted by outliers. The case $K = 1$ corresponds to the classical ERM: $\widehat{f}_1 = \widehat{g}_1 \in \operatorname{argmin}_{f \in F} P_N \ell_f$ which can only be trusted when used with a "clean dataset".

Indeed, the ideal setup for ERM is the subgaussian (and convex) framework: that is for a convex class F of functions, i.i.d. data $(X_i, Y_i)_{i=1}^N$ having the same distribution as (X, Y) and such that for some $L > 0$ and all $f, g \in F$,

$$\|Y\|_{\psi_2} < \infty \text{ and } \|g(X) - f(X)\|_{\psi_2} \leqslant L \|g(X) - f(X)\|_{L_2} . \tag{29}$$

When F satisfies the right-hand side of (29), we say that F is a L-subgaussian class. It is proved in [4] that in this setup the ERM is an optimal minimax procedure (cf. Theorem A′ from [4] recalled in Theorem 9 below).

But first, we need a version of the two theorems 1 and 2 valid for $\widehat{f}_K$ and $\widehat{g}_K$ (that is for the learning problem with no regularization). Let us first introduce the set of assumptions we use. Then, we will introduce the two fixed points driving the statistical properties of $\widehat{f}_K$ and $\widehat{g}_K$.

Assumption 8. *For all $i \in \mathcal{I}$ and $f \in F$, we have*

- $\|f(X_i) - f^*(X_i)\|_{L^2} = \|f(X) - f^*(X)\|_{L_2}$,
- $\|Y_i - f(X_i)\|_{L^2} = \|Y - f(X)\|_{L_2}$,
- $\mathrm{var}((Y - f^*(X))(f(X) - f^*(X))) \leqslant \theta_m^2 \|f(X) - f^*(X)\|_{L^2}^2$
- $\|f(X_i) - f^*(X_i)\|_{L^2} \leqslant \theta_0 \|f(X_i) - f^*(X_i)\|_{L^1}$.

The two fixed points associated to this problem are $r_Q(\rho, \gamma_Q)$ and $r_M(\rho, \gamma_M)$

as in Definition 3 for $\rho = \infty$:

$$r_Q(\gamma_Q) = \inf\left\{ r > 0 : \sup_{J \subset \mathcal{I}, |J| \geqslant \frac{N}{2},} \mathbb{E} \sup_{f \in F : \|f - f^*\|_{L_P^2} \leqslant r} \left| \frac{1}{|J|} \sum_{i \in J} \epsilon_i (f - f^*)(X_i) \right| \leqslant \gamma_Q r \right\} \quad ,$$

$$r_M(\gamma_M) = \inf\left\{ r > 0 : \sup_{J \subset \mathcal{I}, |J| \geqslant \frac{N}{2}} \mathbb{E} \sup_{f \in F : \|f - f^*\|_{L_P^2} \leqslant r} \left| \frac{1}{|J|} \sum_{i \in J} \epsilon_i \zeta_i (f - f^*)(X_i) \right| \leqslant \gamma_M r^2 \right\} \quad ,$$

and let $r^* = r^*(\gamma_Q, \gamma_M) = \max\{r_Q(\gamma_Q), r_M(\gamma_M)\}$.

THEOREM 7. *Grant Assumptions 8 and let $r_Q(\gamma_Q)$, $r_M(\gamma_M)$ and r^* be defined as above for $\gamma_Q = (384\theta_0)^{-1}$, $\gamma_M = \epsilon/192$ and $\epsilon = 1/(32\theta_0^2)$. Assume that $N \geqslant 384\theta_0^2$ and $|\mathcal{O}| \leqslant N/(768\theta_0^2)$. Let K^* denote the smallest integer such that $K^* \geqslant N\epsilon^2(r^*)^2/(384\theta_m^2)$. Then, for all $K \in [\max(K^*, 8|\mathcal{O}|), N/(96\theta_0^2)]$, with probability larger than $1 - 2\exp(-7K/9216)$, the estimators $\widehat{f}_K$ and $\widehat{g}_K$ defined in (28) satisfy*

$$\|\widehat{g}_K - f^*\|_{L_P^2}, \left\|\widehat{f}_K - f^*\right\|_{L_P^2} \leqslant \frac{\theta_m}{\epsilon}\sqrt{\frac{384K}{N}}$$

and

$$R(\widehat{g}_K), R(\widehat{f}_K) \leqslant R(f^*) + (1 + 2\epsilon)\frac{384\theta_m^2 K}{\epsilon^2 N} \quad .$$

Moreover, one can choose adaptively K via Lepski's method. We will do it only for the maxmin estimators $\widehat{g}_K$. Similar result hold for the minmax estimators $\widehat{f}_K$ from straightforward modifications (the same as in Section 3.4.1). Define the confidence regions: for all $J \in [K]$ and $g \in F$,

$$\hat{R}_J = \left\{ g \in F : \mathfrak{C}_J(g) \geqslant \frac{-384\theta_m^2 J}{\epsilon N} \right\} \text{ where } \mathfrak{C}_J(g) = \inf_{f \in F} T_J(g, f)$$

and $T_J(g, f) = \mathrm{MOM}_J\big(\ell_f - \ell_g\big)$ for all $f, g \in F$. Next, let

$$\hat{K} = \inf\left\{ K \in \left[\max(K^*, 8|\mathcal{O}|), \frac{N}{96\theta_0^2}\right] : \bigcap_{J=K}^{K_2} \hat{R}_J \neq \emptyset \right\} \text{ and } \widehat{g} \in \bigcap_{J=\hat{K}}^{K_2} \hat{R}_J \quad .$$

The following theorem shows the performance of the resulting estimator.

THEOREM 8. *Grant Assumption 8. For $\epsilon = 1/(32\theta_0^2)$ and all $K \in [\max(K^*, 8|\mathcal{O}|), N/(96\theta_0^2)]$, with probability larger than $1 - 2\exp(-K/2304)$,*

$$\|\widehat{g} - f^*\|_{L_P^2} \leqslant \frac{\theta_m}{\epsilon}\sqrt{\frac{384K}{N}}, \qquad R(\widehat{g}) \leqslant R(f^*) + (1 + 2\epsilon)\frac{384\theta_m^2 K}{\epsilon^2 N} \quad .$$

The proofs of Theorem 7 and 8 essentially follow the one of Theorem 1 and 2. We will only sketch the proof for the maxmin estimator $\widehat{g}_K$ given that we already studied the minmax estimators in the regularized setup in Section 6.

Proof of Theorem 7. It follows from Lemma 3 and Lemma 4 for $\rho = \infty$ that there exists an event $\Omega(K)$ such that $\mathbb{P}(\Omega(K)) \geqslant 1 - 2\exp\left(-7K/9216\right)$ and, on $\Omega(K)$, for all $f \in F$,

1. if $\|f - f^*\|_{L_P^2} \geqslant r_Q(\gamma_Q)$ then

$$Q_{1/4,K}((f-f^*)^2) \geqslant \frac{1}{(4\theta_0)^2}\|f-f^*\|_{L_P^2}^2 \quad , \tag{30}$$

2. there exists $3K/4$ block B_k with $k \in \mathcal{K}$, for which

$$|(P_{B_k} - P)[2\zeta(f-f^*)]| \leqslant \epsilon \max\left(r_M^2(\gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f-f^*\|_{L_p^2}^2\right) \quad . \tag{31}$$

Moreover, it follows from Assumption 8 that for all $k \in \mathcal{K}$, $\overline{P}_{B_k}[\zeta(f-f^*)] = P[\zeta(f-f^*)]$ and $P[2\zeta(f-f^*)] \leqslant 0$ because of the convexity of F and the nearest point theorem. Therefore, on the event $\Omega(K)$, for all $f \in F$,

$$Q_{3/4,K}(2\zeta(f-f^*)) \leqslant \epsilon \max\left(r_M^2(\gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f-f^*\|_{L_p^2}^2\right) \tag{32}$$

and

$$\begin{aligned}
&P[-2\zeta(f-f^*)] \leqslant P_{B_k}[-2\zeta(f-f^*)] + \epsilon \max\left(r_M^2(\gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f-f^*\|_{L_p^2}^2\right)\\
&\leqslant Q_{1/4,K}[(f-f^*)^2 - 2\zeta(f-f^*)] + \epsilon \max\left(r_M^2(\gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f-f^*\|_{L_p^2}^2\right)\\
&\leqslant T_K(f^*, f) + \epsilon \max\left(r_M^2(\gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f-f^*\|_{L_p^2}^2\right) \quad .
\end{aligned} \tag{33}$$

Let us place ourself on the event $\Omega(K)$ and let r_K be such that $r_K^2 = 384\theta_m^2 K/(\epsilon^2 N)$. Given that $r_K \geqslant r^*$, it follows from (30) and (32) that if $f \in F$ is such that $\|f-f^*\|_{L_P^2} \geqslant r_K$ then

$$\begin{aligned}
&T_K(f, f^*) \leqslant Q_{3/4,K}(2\zeta(f-f^*)) - Q_{1/4}((f-f^*))\\
&\leqslant \left(\epsilon - \frac{1}{16\theta_0^2}\right)\|f-f^*\|_{L_P^2}^2 \leqslant \left(\frac{-1}{32\theta_0^2}\right)\|f-f^*\|_{L_P^2}^2
\end{aligned} \tag{34}$$

for $\epsilon = 1/(32\theta_0^2)$ and if $\|f - f^*\|_{L_P^2} \leqslant r_K$ then $T_K(f, f^*) \leqslant Q_{3/4,K}(2\zeta(f - f^*)) \leqslant \epsilon r_K^2$. In particular,

$$\mathfrak{C}_K(f^*) = \inf_{f\in F} T_K(f^*, f) = -\sup_{f\in F} T_K(f, f^*) \geqslant -\epsilon r_K^2$$

and since $\mathfrak{C}_K(\hat{g}_K) \geqslant \mathfrak{C}_K(f^*)$ one has $\mathfrak{C}_K(\hat{g}_K) \geqslant -\epsilon r_K^2$. On the other hand, we have $\mathfrak{C}_K(\hat{g}_K) = \inf_{f\in F} T_K(\hat{g}_K, f) \leqslant T_K(\hat{g}_K, f^*)$. Therefore, $T_K(\hat{g}_K, f^*) \geqslant -\epsilon r_K^2$. But, we know from (34) that if $g \in F$ is such that $\|g - f^*\|_{L_P^2} > \sqrt{32\epsilon}\theta_0 r_K$ then $T_K(g, f^*) \leqslant (-1/(32\theta_0^2)) \|g - f^*\|_{L_P^2}^2 < -\epsilon r_K^2$. Therefore, one necessarily have $\|\hat{g}_K - f^*\|_{L_P^2} \leqslant \sqrt{32\epsilon}\theta_0 r_K = r_K$.

The oracle inequality now follows from (33):

$$\begin{aligned} R(\hat{g}_K) - R(f^*) &= \|\hat{g}_K - f^*\|_{L_P^2}^2 + P[-2\zeta(\hat{g}_K - f^*)] \\ &\leqslant r_K^2 + T_K(f^*, \hat{g}_K) + \epsilon r_K^2 \leqslant (1 + 2\epsilon) r_K^2 \ . \end{aligned}$$

Proof of Theorem 8. Consider the same notations as in the proof of Theorem 7 and denote $K_2 = N/(96\theta_0^2)$. It follows from the proof of Theorem 7, that with probability larger than $1 - 2\sum_{J=K}^{K_2} \exp(-7J/9216)$, for all $J \in [K, K_2]$, $\mathfrak{C}_J(f^*) \geqslant -\epsilon r_J^2$ therefore, $f^* \in \hat{R}_J$ and so $\hat{K} \leqslant K$. The latter implies that $\widehat{g} \in \hat{R}_K$which, by using the same argument as in the end of the proof of Theorem 7 implies that $\|\widehat{g} - f^*\|_{L_P^2} \leqslant r_K$ and then $R(\widehat{g}) - R(f^*) \leqslant (1 + 2\epsilon) r_K$.

Example: Ordinary least squares. Let us consider the case where $F = \{\langle \cdot, t\rangle : t \in \mathbb{R}^d\}$ is the set of all linear functionals indexed by $\mathbb{R}^d$. We assume that for all $i \in \mathcal{I}$ and $t \in \mathbb{R}^d$,

1. $\mathbb{E}\langle X_i, t\rangle^2 = \mathbb{E}\langle X, t\rangle^2$,
2. $\mathbb{E}(Y_i - \langle X_i, t\rangle)^2 = \mathbb{E}(Y - \langle X, t\rangle)^2$,
3. $\mathbb{E}(Y - \langle X, t^*\rangle)^2 \langle X, t\rangle^2 \leqslant \theta_m^2 \mathbb{E}\langle X, t\rangle^2$,
4. $\sqrt{\mathbb{E}\langle X, t\rangle^2} \leqslant \theta_0 \mathbb{E}|\langle X, t\rangle|$.

Let us now compute the fixed points $r_Q(\gamma_Q)$ and $r_M(\gamma_M)$. The proof essentially follows from Example 1 in [2]. Let $J \subset \mathcal{I}$ be such that $|J| \geqslant N/2$. Denote by $V \subset \mathbb{R}^d$ the smallest linear span containing almost surely X. Let $\varphi_1, \cdots, \varphi_D$ be an orthonormal basis of V with respect to the Hilbert norm

$\|t\| = \mathbb{E}\langle X, t\rangle^2$. It follows from Cauchy-Schwartz inequality that

$$\mathbb{E} \sup_{f \in F: \|f-f^*\|_{L_P^2} \leqslant r} \left| \sum_{i \in J} \epsilon_i (f - f^*)(X_i) \right| = \mathbb{E} \sup_{\sum_{j=1}^D \theta_j^2 \leqslant r^2} \left| \sum_{j=1}^D \theta_j \sum_{i \in J} \epsilon_i \langle X_i, \varphi_j \rangle \right|$$

$$\leqslant r\mathbb{E} \left(\sum_{j=1}^D \left(\sum_{i \in J} \epsilon_i \langle X_i, \varphi_j \rangle \right)^2 \right)^{1/2} \leqslant r \sqrt{\sum_{j=1}^D \sum_{i \in J} \mathbb{E}\langle X_i, \varphi_j \rangle^2} = r\sqrt{D|J|}.$$

As a consequence, $r_Q(\gamma_Q) = 0$ if $\gamma_Q |J| \geqslant \sqrt{D|J|}$, i.e. if $\gamma_Q \geqslant \sqrt{D/|J|}$. Using the same arguments as above, we have

$$\mathbb{E} \sup_{f \in F: \|f-f^*\|_{L_P^2} \leqslant r} \left| \sum_{i \in J} \epsilon_i \zeta_i (f - f^*)(X_i) \right| \leqslant r \sqrt{\sum_{j=1}^D \sum_{i \in J} \mathbb{E}\zeta_i \langle X_i, \varphi_j \rangle^2} \leqslant r\theta_m \sqrt{D|J|}.$$

Therefore, $r_M(\gamma_M) \leqslant (\theta_m/\gamma_M)\sqrt{D/|J|} \leqslant (\theta_m/\gamma_M)\sqrt{2D/N}$ and $K^* = D$.

Now, it follows from Theorem 8, that if $N \geqslant 2(384\theta_0)^2 D$ and $|\mathcal{O}| \leqslant N/(768\theta_0^2)$ then the MOM OLS with adaptively chosen number of blocks K is such that for all $K \in \left[\max\left(D, 8|\mathcal{O}|\right), N/(96\theta_0^2)\right]$, with probability at least $1 - 2\exp(-K/2304)$,

$$\sqrt{\mathbb{E}\langle \hat{t} - t^*, X\rangle^2} \leqslant \frac{\theta_m}{\epsilon}\sqrt{\frac{384K}{N}}. \tag{35}$$

A consequence of (35), is that if the number of outliers is less than $D/8$ then the MOM OLS recovers the classical D/N rate of convergence for the means square error. This happens with probability at least $1 - 2\exp(-D/2304)$, that is with an exponentially large probability. This is a remarkable fact given that we only made assumptions on the L^2 moments of the design X. Moreover, this result is obtained under the only assumption on the informative data that they have equivalent L^2 moments to the one of the distribution of interest P. Therefore, only very little information on P needs to be brought to the statistician via the data; moreover those data can be corrupted up to $D/8$ complete outliers. Finally, note that we did not assume isotropicity of the design X to obtain (35). Therefore, (35) holds even for very degenerate design X and the price we pay is the true dimension of X that is of the dimension of the smallest linear span containing almost surely X not the one of the whole space $\mathbb{R}^d$.

8. Minimax optimality of Theorem 1, 2, 7 and 8. The aim of this section is to show that the rates obtained in Theorems 1, 2, 7 and 8 are optimal in a minimax sense. To that end we recall a minimax lower bound result from [4].

THEOREM 9 (Theorem A$'$ in [4]). *There exists an absolute constant c_0 for which the following holds. Let X be a random variable taking values in $\mathcal{X}$. Let F be a class of functions such that $\mathbb{E}f^2(X) < \infty$. Assume that F is star-shaped around one of its point (i.e. there exists $f_0 \in F$ such that for all $f \in F$ the segment $[f_0, f]$ belongs to F). Let ζ be a centered real-valued Gaussian variable with variance σ independent of X and for all $f^* \in F$ denote by Y^{f^*} the target variable*

$$Y^{f^*} = f^*(X) + \zeta. \tag{36}$$

Let $0 < \delta_N < 1$ and $r_N^2 > 0$. Let $\hat{f}_N$ be a statistics (i.e. a measurable function from $(\mathcal{X} \times \mathbb{R})^N$ to $L^2(P_X)$ where P_X is the probability distribution of X). Assume that $\hat{f}_N$ is such that for all $f^* \in F$, with probability at least $1 - \delta_N$,

$$\left\| \hat{f}_N(\mathcal{D}) - f^* \right\|_{L_P^2}^2 = R(\hat{f}_N(\mathcal{D})) - R(f^*) \leqslant r_N^2$$

where $\mathcal{D} = \{(X_i, Y_i) : i \in [N]\}$ is a set of N i.i.d. copies of (X, Y^{f^*}). Then, necessarily, one has

$$r_N^2 \geqslant \min\left(c_0 \sigma^2 \frac{\log(1/\delta_N)}{N}, \frac{1}{4}\mathrm{diam}(F, L^2(P_X)) \right)$$

where $\mathrm{diam}(F, L^2(P_X))$ denotes the $L^2(P_X)$ diameter of F.

Theorem 9 proves that if the statistical model (36) holds then there is a strong connexion between the deviation parameter δ_N and the uniform rate of convergence r_N^2 over F: the smaller δ_N, the larger r_N^2. We now use this result to prove that Theorems 1, 2, 7 and 8 are essentially optimal.

In Theorems 7 and 8, the deviation bounds are $1 - c_1 \exp(-c_2 K)$ and the residual terms in the L_P^2 (to the square) estimation rates are like $c_3 K/N$. Therefore, setting $\delta_N = c_1 \exp(-c_2 K)$ then Theorem 9 proves that no procedure can do better than

$$\min\left(c_0 \sigma^2 \frac{\log(1/\delta_N)}{N}, \frac{1}{4}\mathrm{diam}(F, L^2(P_X)) \right) = \min\left(c_4 \sigma^2 \frac{K}{N}, \frac{1}{4}\mathrm{diam}(F, L^2(P_X)) \right).$$

Given that one can obviously bound from above the performance of $\widehat{f}_K$ and $\widehat{g}_K$ as well as those of $\widehat{f}$ and $\widehat{g}$ in Theorems 7 and 8 by the L_P^2-diameter of F (because f^* and those estimators are in F), then the result of Theorem 7 and 8 are optimal even in the very strong Gaussian setup with i.i.d. data satisfying a Gaussian regression model like (36). The remarkable point is that Theorem 7 and 8 have been obtained under much weaker assumptions than those considered in Theorem 9 since outliers may corrupt the dataset, the noise and the design do not have to be independent, the informative data are only assumed to have a L^2 norm equivalent to the one of P and may therefore be heavy tailed.

Given the form of the deviation bounds in Theorems 1 and 2 and given that $r(\rho_K) \sim K/N$ and that $r(2\rho_K) \sim K/N$ (if one assumes a weak regularity assumption on the class F) then the same conclusions hold for Theorems 1 and 2: there is no procedure doing better than the MOM estimators even in the very good framework of Theorem 9.

References.

[1] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. ISBN 978-0-19-953525-5.

[2] Vladimir Koltchinskii. Local Rademacher Complexities and Oracle Inequalities in Risk Minimization. *Ann. Statist.*, 34(6):1–50, December 2006. 2004 IMS Medallion Lecture.

[3] G. Lecué and M. Lerasle. Learning from mom's principle : Le cam's approach. Technical report, CNRS, ENSAE, Paris-sud, 2017. To appear in Stochastic Processes and their applications.

[4] Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes: Upper and minimax bounds. Technical report, CNRS, Ecole polytechnique and Technion, 2013.

[5] Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method i: sparse recovery. Technical report, CNRS, ENSAE and Technion, I.I.T., 2016.

[6] Michel Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.

[7] Gabor Lugosi and Shahar Mendelson. A remark on "robust machine learning by median-of-means. *Preprint available on ArXive:1712.06788.*

[8] Gabor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *To appear in JEMS.*

ENSAE
5 avenue Henry Le Chatelier
91120 Palaiseau, France
E-mail: guillaume.lecue@ensae.fr
URL: http://lecueguillaume.github.io

University Paris Sud Orsay
Mathematics department
91405 Orsay
E-mail: matthieu.lerasle@math.u-psud.fr
URL: http://lerasle.perso.math.cnrs.fr