Bernoulli **25**(4A), 2019, 3016–3040
https://doi.org/10.3150/18-BEJ1078

Localized Gaussian width of M-convex hulls with applications to Lasso and convex aggregation

PIERRE C. BELLEC[1]

[1]*Department of Statistics and Biostatistics, Rutgers University, 501 Hill Center, 110 Frelinghuysen Road, Piscataway, NJ 08854, USA. E-mail: pcb71@stat.rutgers.edu*

Upper and lower bounds are derived for the Gaussian mean width of a convex hull of M points intersected with a Euclidean ball of a given radius. The upper bound holds for any collection of extreme points bounded in Euclidean norm. The upper bound and the lower bound match up to a multiplicative constant whenever the extreme points satisfy a one sided Restricted Isometry Property.

An appealing aspect of the upper bound is that no assumption on the covariance structure of the extreme points is needed. This aspect is especially useful to study regression problems with anisotropic design distributions. We provide applications of this bound to the Lasso estimator in fixed-design regression, the Empirical Risk Minimizer in the anisotropic persistence problem, and the convex aggregation problem in density estimation.

Keywords: anisotropic design; convex aggregation; convex hull; Gaussian mean width; Lasso; localized Gaussian width

1. Introduction

Let T be a subset of $\mathbf{R}^n$. The Gaussian width of T is defined as

$$\ell(T) := \mathbb{E} \sup_{\boldsymbol{u} \in T} \boldsymbol{u}^T \boldsymbol{g},$$

where $\boldsymbol{g} = (g_1, \dots, g_n)^T$ and $g_1, \dots, g_n$ are i.i.d. standard normal random variables. For any vector $\boldsymbol{u} \in \mathbf{R}^n$, denote by $|\boldsymbol{u}|_2$ its Euclidean norm and define the Euclidean balls

$$B_2 = \{\boldsymbol{u} \in \mathbf{R}^n : |\boldsymbol{u}|_2 \leq 1\}, \qquad sB_2 = \{\boldsymbol{u} \in \mathbf{R}^n : |\boldsymbol{u}|_2 \leq s\}, \qquad \text{for all } s \geq 0.$$

We will also use the notation $S^{n-1} = \{\boldsymbol{u} \in \mathbf{R}^n : |\boldsymbol{u}|_2 = 1\}$. The localized Gaussian width of T with radius $s > 0$ is the quantity $\ell(T \cap sB_2)$. For any $\boldsymbol{u} \in \mathbf{R}^n$, define the ℓ_p norm by $|\boldsymbol{u}|_p = (\sum_{i=1}^n |u_i|^p)^{1/p}$ for any $p \geq 1$, and let $|\boldsymbol{u}|_0$ be the number of nonzero coefficients of $\boldsymbol{u}$.

This paper studies the localized Gaussian width

$$\ell(sB_2 \cap T),$$

where T is the convex hull of M points in $\mathbf{R}^n$. We will refer to $s > 0$ as the localization parameter.

1350-7265

If $T = B_1 = \{\boldsymbol{u} \in \mathbf{R}^n : |\boldsymbol{u}|_1 \le 1\}$, then matching upper and lower bounds are available for the localized Gaussian width:

$$\ell(sB_2 \cap B_1) \asymp \sqrt{\log\big(en\big(s^2 \wedge 1\big)\big)} \wedge (s\sqrt{n}), \tag{1}$$

cf. [23] and [34], Section 4.1. In the above display, $a \asymp b$ means that $a \le Cb$ and $b \le Ca$ for some large enough numerical constant $C \ge 1$.

The first goal of this paper is to generalize this bound to any T that is the convex hull of $M \ge 1$ points in $\mathbf{R}^n$.

Contributions

Section 2 is devoted to the generalization of (1) and provides sharp bounds on the localized Gaussian width of the convex hull of M points in $\mathbf{R}^n$, see Propositions 1 and 2 below. Sections 3 to 5 provide statistical applications of the results of Section 2. Section 3 studies the Lasso estimator and the convex aggregation problem in fixed-design regression. In Section 4, we show that Empirical Risk Minimization achieves the minimax rate for the persistence problem in the anisotropic setting. Finally, Section 5 provides results for bounded empirical processes and for the convex aggregation problem in density estimation.

2. Localized Gaussian width of a M-convex hull

The first contribution of the present paper is the following upper bound on localized Gaussian width of the convex hull of M points in $\mathbf{R}^n$.

Proposition 1. *Let $n \ge 1$ and $M \ge 2$. Let T be the convex hull of M points in $\mathbf{R}^n$ and assume that $T \subset B_2$. Then for all $s > 0$,*

$$\ell(T \cap sB_2) \le \Big(4\sqrt{\log_+\big(4eM\big(s^2 \wedge 1\big)\big)}\Big) \wedge (s\sqrt{n \wedge M}), \tag{2}$$

where $\log_+(a) = \max(1, \log a)$.

Proposition 1 is proved in the next two subsections. Inequality

$$\ell(T \cap sB_2) \le s\sqrt{n \wedge M} \tag{3}$$

is a direct consequence of the Cauchy–Schwarz inequality and $\mathbb{E}_{\boldsymbol{g}\sim N(\mathbf{0}, I_{n\times n})}|P\boldsymbol{g}|_2 \le \sqrt{d}$ where matrix $P \in \mathbf{R}^{n\times n}$ is the orthogonal projection onto the linear span of T and $d \le (n \wedge M)$ is the rank of P. Inequality $\ell(T \cap sB_2) \le \sqrt{2\log M}$ is a direct consequence of the fact that the expectation maximum of M centered normal random variables with variance at most 1 is bounded from above by $\sqrt{2\log M}$. The novelty of (2) is inequality

$$\ell(T \cap sB_2) \le 4\sqrt{\log_+\big(4eMs^2\big)}. \tag{4}$$

Inequality (4) was known for the ℓ_1-ball $T = \{\boldsymbol{u} \in \mathbf{R}^n : |\boldsymbol{u}|_1 \leq 1\}$ [23], but to our knowledge (4) is new for general M-convex hulls.

An appealing aspect of the above result is that it holds with no assumption on the correlation structure of the vertices of T. Indeed, if the vertices of T are $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_M$ and $\Sigma = (\boldsymbol{\mu}_j^T \boldsymbol{\mu}_k)_{j,k=1,\dots,M}$ is the Gram matrix, then (4) holds with no assumption on Σ. Previous results similar (4) only apply to the ℓ_1 ball or the simplex in $\mathbf{R}^M$, where the vertices are orthogonal. The result above shows that no orthogonality assumption is necessary for (4) to hold. This aspect will be especially useful to derive results for regression problems with anisotropic design distributions as in Section 4.

The above result does not assume any type of Restricted Isometry Property (RIP). The following proposition shows that (4) is essentially sharp provided that the vertices of T satisfies a one-sided RIP of order $2/s^2$.

Proposition 2. *Let $n \geq 1$ and $M \geq 2$. Let $\boldsymbol{g}$ be a centered Gaussian random variable with covariance matrix $I_{n\times n}$. Let $s \in (0, 1]$ and assume for simplicity that $m = 1/s^2$ is a positive integer such that $m \leq M/5$. Let T be the convex hull of the $2M$ points $\{\pm\boldsymbol{\mu}_1, \dots, \pm\boldsymbol{\mu}_M\}$ where $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_M \in S^{n-1}$. Assume that for some real number $\kappa \in (0, 1)$ we have*

$$\kappa|\boldsymbol{\theta}|_2 \leq |\boldsymbol{\mu}_{\boldsymbol{\theta}}|_2 \qquad \text{for all } \boldsymbol{\theta} \in \mathbf{R}^M \text{ such that } |\boldsymbol{\theta}|_0 \leq 2m, \tag{5}$$

where $\boldsymbol{\mu}_{\boldsymbol{\theta}} = \sum_{j=1}^M \theta_j \boldsymbol{\mu}_j$ and $|\boldsymbol{\theta}|_0$ denotes the number of nonzero coefficients of $\boldsymbol{\theta}$. Then

$$\ell(T \cap sB_2) \geq (\sqrt{2}/4)\kappa\sqrt{\log\left(\frac{Ms^2}{5}\right)}. \tag{6}$$

The proof of Proposition 2 is given in Section B. The lower bound (6) shows that the upper bound (4) is sharp up to multiplicative constants, not only if T is the ℓ_1-ball (cf. (1)) but also if the extreme points of T satisfy the one-sided RIP (5). This suggests that the quantity $\ell(T \cap sB_2)$ is maximal when the extreme points of T are uncorrelated. We investigate this further with simulations in Appendix A.

2.1. A refinement of Maurey's argument

This subsection provides the main tool to derive the upper bound (4). Define the simplex in $\mathbf{R}^M$ by

$$\Lambda^M = \left\{\boldsymbol{\theta} \in \mathbf{R}^M, \sum_{j=1}^M \theta_j = 1, \forall j = 1, \dots, M, \theta_j \geq 0\right\}. \tag{7}$$

Let $m \geq 1$ be an integer, and let

$$Q(\boldsymbol{\theta}) = \boldsymbol{\theta}^T \Sigma \boldsymbol{\theta},$$

where $\Sigma = (\Sigma_{jj'})_{j,j'=1,\dots,M}$ is a positive semi-definite matrix of size M. Let $\bar{\boldsymbol{\theta}} \in \Lambda^M$ be a vector of interest such that $Q(\bar{\boldsymbol{\theta}})$ is small. Maurey's argument [41] has been used extensively to prove the existence of a sparse vector $\tilde{\boldsymbol{\theta}} \in \Lambda^M$ such that $Q(\tilde{\boldsymbol{\theta}})$ is of the same order as that of $Q(\bar{\boldsymbol{\theta}})$. Maurey's argument uses the probabilistic method to prove the existence of such $\tilde{\boldsymbol{\theta}}$. A sketch of this argument is as follows.

Define the discrete set Λ_m^M as

$$\Lambda_m^M := \left\{ \frac{1}{m} \sum_{k=1}^m \boldsymbol{u}_k, \ \ \boldsymbol{u}_1, \dots, \boldsymbol{u}_m \in \{\boldsymbol{e}_1, \dots, \boldsymbol{e}_M\} \right\}, \tag{8}$$

where $(\boldsymbol{e}_1, \dots, \boldsymbol{e}_M)$ is the canonical basis in $\mathbf{R}^M$. The discrete set Λ_m^M is a subset of the simplex Λ^M that contains only m-sparse vectors.

Let $\Theta_1, \dots, \Theta_m$ be i.i.d. random variables valued in $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_M\}$ with distribution

$$\mathbb{P}(\Theta_k = \boldsymbol{e}_j) = \bar{\theta}_j \qquad \text{for all } k = 1, \dots, m. \tag{9}$$

Next, consider the random variable

$$\hat{\boldsymbol{\theta}} = \frac{1}{m} \sum_{k=1}^m \Theta_k. \tag{10}$$

The random variable $\hat{\boldsymbol{\theta}}$ is valued in Λ_m^M and is such that $\mathsf{E}[\hat{\boldsymbol{\theta}}] = \bar{\boldsymbol{\theta}}$, where E denotes the expectation with respect to $\hat{\boldsymbol{\theta}}$. Then a bias-variance decomposition yields

$$\mathsf{E}\big[Q(\hat{\boldsymbol{\theta}})\big] \le Q(\bar{\boldsymbol{\theta}}) + R^2/m,$$

where $R > 0$ is a constant such that $\max_{j=1,\dots,M} \Sigma_{jj} \le R^2$. As $\min_{\boldsymbol{\theta} \in \Lambda_m^M} Q(\boldsymbol{\theta}) \le \mathsf{E}[Q(\hat{\boldsymbol{\theta}})]$, this yields the existence of $\tilde{\boldsymbol{\theta}} \in \Lambda_m^M$ such that

$$Q(\tilde{\boldsymbol{\theta}}) \le Q(\bar{\boldsymbol{\theta}}) + R^2/m.$$

If m is chosen large enough, the two terms $Q(\bar{\boldsymbol{\theta}})$ and R^2/m are of the same order and we have established the existence of an m-sparse vector $\tilde{\boldsymbol{\theta}}$ so that $Q(\tilde{\boldsymbol{\theta}})$ is not much substantially larger than $Q(\bar{\boldsymbol{\theta}})$.

For our purpose, we need to refine this argument by controlling the deviation of the random variable $Q(\hat{\boldsymbol{\theta}})$. This is done in Lemma 3 below.

Lemma 3. *Let $m \ge 1$ and define Λ_m^M by (8). Let $F : \mathbf{R}^M \to [0, +\infty)$ be a convex function. For all $\boldsymbol{\theta} \in \mathbf{R}^M$, let*

$$Q(\boldsymbol{\theta}) = \boldsymbol{\theta}^T \Sigma \boldsymbol{\theta},$$

where $\Sigma = (\Sigma_{jj'})_{j,j'=1,\dots,M}$ is a positive semi-definite matrix of size M. Assume that the diagonal elements of Σ satisfy $\Sigma_{jj} \le R^2$ for all $j = 1, \dots, M$. Then for all $t > 0$,

$$\sup_{\boldsymbol{\theta} \in \Lambda^M : Q(\boldsymbol{\theta}) \le t^2} F(\boldsymbol{\theta}) \le \int_1^{+\infty} \Big[\max_{\boldsymbol{\theta} \in \Lambda_m^M :\, Q(\boldsymbol{\theta}) \le x(t^2 + R^2/m)} F(\boldsymbol{\theta}) \Big] \frac{dx}{x^2}. \tag{11}$$

In the next sections, it will be useful to bound from above the quantity $F(\boldsymbol{\theta})$ maximized over Λ^M subject to the constraint $Q(\boldsymbol{\theta}) \leq t^2$. An interpretation of (11) is as follows. Consider the two optimization problems

$$\begin{aligned} &\text{maximize } F(\boldsymbol{\theta}) \qquad \text{for } \boldsymbol{\theta} \in \Lambda^M \qquad \text{subject to } Q(\boldsymbol{\theta}) \leq t^2, \\ &\text{maximize } F(\boldsymbol{\theta}) \qquad \text{for } \boldsymbol{\theta} \in \Lambda_m^M \qquad \text{subject to } Q(\boldsymbol{\theta}) \leq Y\big(t^2 + R^2/m\big), \end{aligned}$$

for some $Y \geq 1$. Equation 11 says that the optimal value of the first optimization problem is smaller than the optimal value of the second optimization problem averaged over the distribution of Y given by the density $y \mapsto 1/y^2$ on $[1, +\infty)$. The second optimization problem above is over the *discrete* set Λ_m^M with the relaxed constraint $Q(\boldsymbol{\theta}) \leq Y(t^2 + R^2/m)$, hence we have relaxed the constraint in exchange for discreteness. The discreteness of the set Λ_m^M will be used in the next subsection for the proof of Proposition 1.

Proof of Lemma 3. The set $\{\boldsymbol{\theta} \in \Lambda^M : Q(\boldsymbol{\theta}) \leq t^2\}$ is compact. The function F is convex with domain $\mathbf{R}^M$ and thus continuous. Hence the supremum in the left hand side of (11) is achieved at some $\bar{\boldsymbol{\theta}} \in \Lambda^M$ such that $Q(\bar{\boldsymbol{\theta}}) \leq t^2$. Let $\Theta_1, \dots, \Theta_m, \hat{\boldsymbol{\theta}}$ be the random variables defined in (9) and (10) above. Denote by E the expectation with respect to $\Theta_1, \dots, \Theta_m$. By definition, $\hat{\boldsymbol{\theta}} \in \Lambda_m^M$ and $\mathsf{E}\hat{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}}$. Let $E = \mathsf{E}[Q(\hat{\boldsymbol{\theta}})]$. A bias-variance decomposition and the independence of $\Theta_1, \dots, \Theta_m$ yield

$$\begin{aligned} E &:= \mathsf{E}\big[Q(\hat{\boldsymbol{\theta}})\big] \\ &= Q(\bar{\boldsymbol{\theta}}) + \mathsf{E}(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})^T \Sigma (\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) \\ &= Q(\bar{\boldsymbol{\theta}}) + \frac{1}{m}\mathsf{E}\big[(\Theta_1 - \bar{\boldsymbol{\theta}})^T \Sigma (\Theta_1 - \bar{\boldsymbol{\theta}})\big]. \end{aligned}$$

Another bias-variance decomposition yields

$$\mathsf{E}\big[(\Theta_1 - \bar{\boldsymbol{\theta}})^T \Sigma (\Theta_1 - \bar{\boldsymbol{\theta}})\big] = \mathsf{E}\big[Q(\Theta_1)\big] - Q(\bar{\boldsymbol{\theta}}) \leq \mathsf{E} Q(\Theta_1) \leq R^2,$$

where we used that $Q(\cdot) \geq 0$ and that $\Theta_1 \Sigma \Theta_1 \leq R^2$ almost surely. Thus,

$$E = \mathsf{E}\big[Q(\hat{\boldsymbol{\theta}})\big] \leq Q(\bar{\boldsymbol{\theta}}) + R^2/m \leq t^2 + R^2/m. \tag{12}$$

Define the random variable $X = Q(\hat{\boldsymbol{\theta}})/E$, which is nonnegative and satisfies $\mathsf{E}[X] = 1$. By Markov inequality, it holds that $\mathsf{P}(X > t) \leq 1/t = \int_t^{+\infty} (1/x^2)\, dx$. Define the random variable Y by the density function $x \to 1/x^2$ on $[1, +\infty)$. Then we have $\mathsf{P}(X > t) \leq \mathbb{P}(Y > t)$ for any $t > 0$. By stochastic dominance, there exists a rich enough probability space Ω and random variables $\tilde{X}$ and $\tilde{Y}$ defined on Ω such that

- $\tilde{X}$ and X have the same distribution,
- $\tilde{Y}$ and Y have the same distribution,
- and $\tilde{X} \leq \tilde{Y}$ holds almost surely on Ω,

see, for instance, Theorem 7.1 in [21]. Denote by $\mathbb{E}_\Omega$ the expectation sign on the probability space Ω.

By definition of $\bar{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$, using Jensen's inequality, Fubini's theorem and the fact that $\hat{\boldsymbol{\theta}} \in \Lambda_m^M$ we have

$$\sup_{\boldsymbol{\theta}\in\Lambda^M:Q(\boldsymbol{\theta})\le t^2} F(\boldsymbol{\theta}) = F(\bar{\boldsymbol{\theta}}) = F\big(\mathsf{E}[\hat{\boldsymbol{\theta}}]\big) \le \mathsf{E}\big[F(\hat{\boldsymbol{\theta}})\big] \le \mathsf{E}\big[g\big(Q(\hat{\boldsymbol{\theta}})/E\big)\big],$$

where $g(\cdot)$ is the nondecreasing function $g(x) = \max_{\boldsymbol{\theta}\in\Lambda_m^M:Q(\boldsymbol{\theta})\le xE} F(\boldsymbol{\theta})$. The right-hand side of the previous display is equal to $\mathsf{E}[g(X)]$. Next, we use the random variables $\tilde{X}$ and $\tilde{Y}$ as follows:

$$\mathsf{E}\big[g(X)\big] = \mathbb{E}_\Omega\big[g(\tilde{X})\big] \le \mathbb{E}_\Omega\big[g(\tilde{Y})\big] = \int_1^{+\infty} \frac{g(x)}{x^2}\,dx.$$

Combining the previous display and (12) completes the proof. □

2.2. Proof of (4)

We are now ready to prove Proposition 1. The main ingredients are Lemma 3 and the following upper bound on the cardinality of Λ_m^M

$$\log\big|\Lambda_m^M\big| = \log\binom{M+m-1}{m} \le \log\binom{2M}{m} \le m\log\left(\frac{2eM}{m}\right). \tag{13}$$

Proof of (4). If $s^2 < 1/M$ then by (3) we have $\ell(T\cap sB_2) \le 1$, hence (4) holds. Thus it is enough to focus on the case $s^2 \ge 1/M$.

Let $r = \min(s, 1)$ and set $m = \lfloor 1/r^2\rfloor$, which satisfies $1 \le m \le M$. As T is the convex hull of M points, let $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_M \in \mathbf{R}^n$ be such that

$$T = \text{convex hull of } \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_M\} = \big\{\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Lambda^M\big\},$$

where $\boldsymbol{\mu}_{\boldsymbol{\theta}} = \sum_{j=1}^M \theta_j \boldsymbol{\mu}_j$ for $\boldsymbol{\theta} \in \Lambda^M$.

Let $Q(\boldsymbol{\theta}) = |\boldsymbol{\mu}_{\boldsymbol{\theta}}|_2^2$ for all $\boldsymbol{\theta} \in \mathbf{R}^M$. This is a polynomial of order 2, of the form $Q(\boldsymbol{\theta}) = \boldsymbol{\theta}^T\Sigma\boldsymbol{\theta}$, where Σ is the Gram matrix with $\Sigma_{jk} = \boldsymbol{\mu}_k^T\boldsymbol{\mu}_j$ for all $j, k = 1, \dots, M$. As we assume that $T \subset B_2$, the diagonal elements of Σ satisfy $\Sigma_{jj} \le 1$. For all $\boldsymbol{\theta} \in \mathbf{R}^M$, let $F(\boldsymbol{\theta}) = \boldsymbol{g}^T\boldsymbol{\mu}_{\boldsymbol{\theta}}$. Applying Lemma 3 with the above notation, $R = 1$, $m = \lfloor 1/r^2\rfloor$ and $t = r$, we obtain

$$\mathbb{E} \sup_{\boldsymbol{\theta}\in\Lambda^M:Q(\boldsymbol{\theta})\le r^2} \boldsymbol{g}^T\boldsymbol{\mu}_{\boldsymbol{\theta}} \le \mathbb{E}\int_1^{+\infty}\Big[\max_{\boldsymbol{\theta}\in\Lambda_m^M:\, Q(\boldsymbol{\theta})\le x(r^2+1/m)} F(\boldsymbol{\theta})\Big]\frac{dx}{x^2}.$$

By definition of m, $r^2 \le 1/m$ so that $x(r^2 + 1/m) \le 2x/m$. Using Fubini's theorem and a bound on the expectation of the maximum of $|\Lambda_m^M|$ centered Gaussian random variables with variances bounded from above by $2x/m$, we obtain that the right-hand side of the previous display is bounded from above by

$$\int_1^{+\infty} \frac{1}{x^2}\sqrt{\frac{4x\log|\Lambda_m^M|}{m}}\,dx \le \sqrt{\log(2eM/m)}\int_1^{+\infty}\frac{2}{x^{3/2}}\,dx,$$

where we used the bound (13). To complete the proof of (4), notice that we have $1/m \le 2r^2$ and $\int_1^{+\infty}\frac{2}{x^{3/2}}\,dx = 4$. □

3. Statistical applications in fixed-design regression

Numerous works have established a close relationship between localized Gaussian widths and the performance of statistical and compressed sensing procedures. Some of these works are reviewed below.

- In a regression problem with random design where the design and the target are sub-Gaussian, Lecué and Mendelson [34] established that two quantities govern the performance of empirical risk minimizer over a convex class $\mathcal{F}$. These two quantities are defined using the Gaussian width of the class $\mathcal{F}$ intersected with an L_2 ball [34], Definition 1.3,
- If $p, p' > 1$ are such that $p' \le p \le +\infty$ and $\log(2n)/(\log(2en) \le p'$. Gordon et al. [23] provide precise estimates of $\ell(B_p \cap sB_{p'})$ where $B_p \subset \mathbf{R}^n$ is the unit L_p ball and $sB_{p'}$ is the $L_{p'}$ ball of radius $s > 0$. These estimates are then used to solve the approximate reconstruction problem where one wants to recover an unknown high dimensional vector from a few random measurements [23], Section 7.
- Plan et al. [43] shows that in the semiparametric single index model, if the signal is known to belong to some star-shaped set $T \subset \mathbf{R}^n$, then the Gaussian width of T and its localized version characterize the gain obtained by using the additional information that the signal belongs to T, cf. Theorem 1.3 in [43].
- Finally, Chatterjee [17] exhibits a tight connection between localized Gaussian widths and the performance of the least-squares estimator in shape-constrained regression problems. The machinery developed in [17] is particularly appealing since it provides both a lower bound and an upper bound on the risk of the least-squares estimator. Theorem 4 provides sharp oracle inequalities in the same setting as that of [17]; however Theorem 4 does not provide lower bounds on the risk.

These results are reminiscent of the isomorphic method [2,3,28], where localized expected supremum of empirical processes are used to obtain upper bounds on the performance of Empirical Risk Minimization (ERM) procedures. These results show that Gaussian width estimates are important to understand the statistical properties of estimators in many statistical contexts.

In Proposition 1, we established an upper bound on the Gaussian width of M-convex hulls. We now provide some statistical applications of this result in regression with fixed-design. We will use the following theorem from [9].

Theorem 4 ([9]). *Let K be a closed convex subset of $\mathbf{R}^n$ and $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 I_{n\times n})$. Let $\boldsymbol{f}_0 \in \mathbf{R}^n$ be an unknown vector and let $\mathbf{y} = \boldsymbol{f}_0 + \boldsymbol{\xi}$. Denote by $\boldsymbol{f}_0^*$ the projection of $\boldsymbol{f}_0$ onto K. Assume that for some $t_* > 0$,*

$$\frac{1}{n}\mathbb{E}\Big[\sup_{\boldsymbol{u}\in K:\ \frac{1}{n}|\boldsymbol{f}_0^* - \boldsymbol{u}|_2^2 \le t_*^2} \boldsymbol{\xi}^T\big(\boldsymbol{u} - \boldsymbol{f}_0^*\big)\Big] \le \frac{t_*^2}{2}. \tag{14}$$

Then for any $x > 0$, with probability greater than $1 - e^{-x}$, the Least Squares estimator $\hat{\boldsymbol{f}} = \arg\min_{\boldsymbol{f}\in K} |\mathbf{y} - \boldsymbol{f}|_2^2$ satisfies

$$\frac{1}{n}|\hat{\boldsymbol{f}} - \boldsymbol{f}_0|_2^2 \le \frac{1}{n}\big|\boldsymbol{f}_0^* - \boldsymbol{f}_0\big|_2^2 + 2t_*^2 + \frac{4\sigma^2 x}{n}.$$

Hence, to prove an oracle inequality of the form (4), it is enough to prove the existence of a quantity t_* such that (14) holds. If the convex set K in the above theorem is the convex hull of M points, then a quantity t_* is given by the following proposition.

Proposition 5. *Let $\sigma^2 > 0$, $R > 0$, $n \ge 1$ and $M \ge 2$. Let $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_M \in \mathbf{R}^n$ such that $\frac{1}{n}|\boldsymbol{\mu}_j|_2^2 \le R^2$ for all $j = 1, \dots, M$. For all $\boldsymbol{\theta} \in \Lambda^M$, let $\boldsymbol{\mu}_{\boldsymbol{\theta}} = \sum_{j=1,\dots,M} \theta_j \boldsymbol{\mu}_j$. Let $\boldsymbol{g}$ be a centered Gaussian random variable with covariance matrix $\sigma^2 I_{n\times n}$. If $R\sqrt{n} \le M\sigma$, then the quantity*

$$t_*^2 = 31\sigma R\sqrt{\frac{\log(\frac{eM\sigma}{R\sqrt{n}})}{n}} \quad \textit{satisfies}\ \frac{1}{n}\mathbb{E}\sup_{\boldsymbol{\theta}\in\Lambda^M:\frac{1}{n}|\boldsymbol{\mu}_{\boldsymbol{\theta}}|_2^2 \le t_*^2} \boldsymbol{g}^T\boldsymbol{\mu}_{\boldsymbol{\theta}} \le \frac{t_*^2}{2}, \tag{15}$$

provided that $t_* \le R$.

Proof. Inequality

$$\frac{1}{\sqrt{n}}\mathbb{E}\sup_{\boldsymbol{\theta}\in\Lambda^M:\frac{1}{n}|\boldsymbol{\mu}_{\boldsymbol{\theta}}|_2^2 \le r^2} (\sigma\boldsymbol{g})^T\boldsymbol{\mu}_{\boldsymbol{\theta}} \le 4\sigma R\sqrt{\log\big(4eM\min\big(1, r^2/R^2\big)\big)}$$

is a reformulation of Proposition 1 using the notation of Proposition 5. Thus, in order to prove (15), it is enough to establish that for $\gamma = 31$ we have

$$(*) := 64\log\bigg(\frac{4eM\sigma\gamma\sqrt{\log(eM\sigma/(R\sqrt{n}))}}{R\sqrt{n}}\bigg) \le \frac{\gamma^2}{4}\log\bigg(\frac{eM\sigma}{R\sqrt{n}}\bigg). \tag{16}$$

As $1 \le \log(eM\sigma/(R\sqrt{n}))$ and $\log t \le t$ for all $t > 0$, the left-hand side of the previous display satisfies

$$\begin{aligned}(*) &\le 64\bigg(\log\bigg(\frac{eM\sigma}{R\sqrt{n}}\bigg) + \log(4\gamma) + \frac{1}{2}\log\big(\log\big(eM\sigma/(R\sqrt{n})\big)\big)\bigg) \\ &\le 64\big(3/2 + \log(4\gamma)\big)\log\bigg(\frac{eM\sigma}{R\sqrt{n}}\bigg).\end{aligned}$$

Thus (16) holds if $64(3/2+\log(4\gamma)) \le \gamma^2/4$, which is the case if the absolute constant is $\gamma = 31$. □

Inequality (15) establishes the existence of a quantity t_* such that

$$\frac{1}{n}\mathbb{E} \sup_{\boldsymbol{\mu}\in T:\frac{1}{n}|\boldsymbol{\mu}|_2^2\le t_*^2} \boldsymbol{g}^T\boldsymbol{\mu}_{\boldsymbol{\theta}} \le \frac{t_*^2}{2}, \tag{17}$$

where T is the convex hull of $\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_M$. Consequences of (17) and Theorem 4 are given in the next subsections.

We now introduce two statistical frameworks where the localized Gaussian width of an M-convex hull has applications: the Lasso estimator in high-dimensional statistics and the convex aggregation problem.

3.1. Convex aggregation

Let $\boldsymbol{f}_0 \in \mathbf{R}^n$ be an unknown regression vector and let $\mathbf{y} = \boldsymbol{f}_0 + \boldsymbol{\xi}$ be an observed random vector, where $\boldsymbol{\xi}$ satisfies $\mathbb{E}[\boldsymbol{\xi}] = \mathbf{0}$. Let $M \ge 2$ and let $\boldsymbol{f}_1,\dots,\boldsymbol{f}_M$ be deterministic vectors in $\mathbf{R}^n$. The set $\{\boldsymbol{f}_1,\dots,\boldsymbol{f}_M\}$ will be referred to as the dictionary. For any $\boldsymbol{\theta} = (\theta_1,\dots,\theta_M)^T \in \mathbf{R}^M$, let $\boldsymbol{f}_{\boldsymbol{\theta}} = \sum_{j=1}^M \theta_j \boldsymbol{f}_j$. If a set $\Theta \subset \mathbf{R}^M$ is given, the goal of the aggregation problem induced by Θ is to find an estimator $\hat{\boldsymbol{f}}$ constructed with $\mathbf{y}$ and the dictionary such that

$$\frac{1}{n}|\hat{\boldsymbol{f}} - \boldsymbol{f}_0|_2^2 \le \inf_{\boldsymbol{\theta}\in\Theta}\left(\frac{1}{n}|\boldsymbol{f}_{\boldsymbol{\theta}} - \boldsymbol{f}_0|_2^2\right) + \delta_{n,M,\Theta}, \tag{18}$$

either in expectation or with high probability, where $\delta_{n,M,\Theta}$ is a small quantity. Inequality (18) is called a sharp oracle inequality, where "sharp" means that in the right hand side of (18), the multiplicative constant of the term $\inf_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}|\boldsymbol{f}_{\boldsymbol{\theta}} - \boldsymbol{f}_0|_2^2$ is 1. Similar notations will be defined for regression with random design and density estimation. Define the simplex in $\mathbf{R}^M$ by (7). The following aggregation problems were introduced in [40,50].

- *Model Selection type aggregation* with $\Theta = \{\boldsymbol{e}_1,\dots,\boldsymbol{e}_M\}$, that is, Θ is the canonical basis of $\mathbf{R}^M$. The goal is to construct an estimator whose risk is as close as possible to the best function in the dictionary. Such results can be found in [1,35,50] for random design regression, in [8,18,20,36] for fixed design regression, and in [6,26] for density estimation.
- *Convex aggregation* with $\Theta = \Lambda^M$, that is, Θ is the simplex in $\mathbf{R}^M$. The goal is to construct an estimator whose risk is as close as possible to the best convex combination of the dictionary functions. See [32,33,50,52] for results of this type in the regression framework and [46] for such results in density estimation.
- *Linear aggregation* with $\Theta = \mathbf{R}^M$. The goal is to construct an estimator whose risk is as close as possible to the best linear combination of the dictionary functions, cf. [50,52] for such results in regression and [46] for such results in density estimation.

One may also define the *Sparse* or *Sparse Convex* aggregation problems: construct an estimator whose risk is as close as possible to the best sparse combination of the dictionary functions. Such results can be found in [44,45,52] for fixed design regression and in [37] for regression with random design. These problems are out of the scope of the present paper.

A goal of the present paper is to provide a unified argument that shows that empirical risk minimization is optimal for the convex aggregation problem in density estimation, regression with fixed design and regression with random design.

Theorem 6. *Let $\boldsymbol{f}_0 \in \mathbf{R}^n$, let $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 I_{n\times n})$ and define $\mathbf{y} = \boldsymbol{f}_0 + \boldsymbol{\xi}$. Let $\boldsymbol{f}_1, \dots, \boldsymbol{f}_M \in \mathbf{R}^n$ and let $\boldsymbol{f}_{\boldsymbol{\theta}} = \sum_{j=1}^M \theta_j \boldsymbol{f}_j$ for all $\boldsymbol{\theta} = (\theta_1, \dots, \theta_M)^T \in \mathbf{R}^M$. Let*

$$\hat{\boldsymbol{\theta}} \in \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Lambda^M} |\boldsymbol{f}_{\boldsymbol{\theta}} - \mathbf{y}|_2^2.$$

Then for all $x > 0$, with probability greater than $1 - \exp(-x)$,

$$\frac{1}{n}|\boldsymbol{f}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{f}_0|_2^2 \le \min_{\boldsymbol{\theta}\in\Lambda^M} \frac{1}{n}|\boldsymbol{f}_{\boldsymbol{\theta}} - \boldsymbol{f}_0|_2^2 + 2t_*^2 + \frac{4\sigma^2 x}{n},$$

where $t_*^2 = \min(\frac{4\sigma^2 M}{n}, \frac{31\sigma R\sqrt{\log(\frac{eM\sigma}{R\sqrt{n}})}}{\sqrt{n}})$ and $R^2 = 4\max_{j=1,\dots,M} \frac{1}{n}|\boldsymbol{f}_j|_2^2$.

Proof of Theorem 6. Let V be the linear span of $\boldsymbol{f}_1, \dots, \boldsymbol{f}_M$ and let $P \in \mathbf{R}^{n\times n}$ be the orthogonal projector onto V. If $t_*^2 = 4\sigma^2 M/n$, then

$$\frac{1}{n}\mathbb{E} \sup_{\boldsymbol{v}\in V: \frac{1}{n}|\boldsymbol{v}|_2^2 \le t_*^2} \boldsymbol{\xi}^T \boldsymbol{v} = \sqrt{\frac{t_*^2}{n}}\mathbb{E}|P\boldsymbol{\xi}|_2 \le \sqrt{\frac{t_*^2\sigma^2 M}{n}} = t_*^2/2. \tag{19}$$

Let K be the convex hull of $\boldsymbol{f}_1, \dots, \boldsymbol{f}_M$. Let $\boldsymbol{f}_0^*$ be the convex projection of $\boldsymbol{f}_0$ onto K. We apply Proposition 5 to $K - \boldsymbol{f}_0^*$ which is a convex hull of M points, and for all $\boldsymbol{v} \in K$, $\frac{1}{n}|\boldsymbol{v}|_2^2 \le R^2$. By (19) and (15), the quantity t_* satisfies (14). Applying Theorem 4 completes the proof. □

3.2. Lasso

We consider the following regression model. Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_M \in R^n$ and assume that $\frac{1}{n}|\boldsymbol{x}_j|_2^2 \le 1$ for all $j = 1, \dots, M$. We will refer to $\boldsymbol{x}_1, \dots, \boldsymbol{x}_M$ as the covariates. Let $\mathbf{X}$ be the matrix of dimension $n \times M$ with columns $\boldsymbol{x}_1, \dots, \boldsymbol{x}_M$. We observe

$$\mathbf{y} = \boldsymbol{f}_0 + \boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}\big(0, \sigma^2 I_{n\times n}\big), \tag{20}$$

where $\boldsymbol{f}_0 \in \mathbf{R}^n$ is an unknown mean. The goal is to estimate $\boldsymbol{f}_0$ using the design matrix $\mathbf{X}$.

Let $R > 0$ be a tuning parameter and define the constrained Lasso estimator [49] by

$$\hat{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbf{R}^M: |\boldsymbol{\beta}|_1 \le R} |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|_2^2. \tag{21}$$

Note that, although a different notation is used, this model equivalent to the convex aggregation problem considered in the previous subsection.

Our goal will be to study the performance of the estimator (21) with respect to the prediction loss

$$\frac{1}{n}|\boldsymbol{f}_0 - \mathbf{X}\hat{\boldsymbol{\beta}}|_2^2.$$

Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_M \in R^n$ and assume that $\frac{1}{n}|\boldsymbol{x}_j|_2^2 \leq 1$ for all $j = 1, \dots, M$. Let $\mathbf{X}$ be the matrix of dimension $n \times M$ with columns $\boldsymbol{x}_1, \dots, \boldsymbol{x}_M$.

Theorem 7. *Let $R > 0$ be a tuning parameter and consider the regression model* (20). *Define the Lasso estimator $\hat{\boldsymbol{\beta}}$ by* (21). *Then for all $x > 0$, with probability greater than $1 - \exp(-x)$,*

$$\frac{1}{n}|\mathbf{X}\hat{\boldsymbol{\beta}} - \boldsymbol{f}_0|_2^2 \leq \min_{\boldsymbol{\beta} \in \mathbf{R}^M : |\boldsymbol{\beta}|_1 \leq R} \frac{1}{n}|\mathbf{X}\boldsymbol{\beta} - \boldsymbol{f}_0|_2^2 + 2t_*^2 + \frac{4\sigma^2 x}{n}, \tag{22}$$

where $t_*^2 = \min(\frac{4\sigma^2 \operatorname{rank}(\mathbf{X})}{n}, \frac{62\sigma R\sqrt{\log(\frac{2eM\sigma}{R\sqrt{n}})}}{\sqrt{n}})$.

Proof of Theorem 7. Let V be the linear span of $\boldsymbol{x}_1, \dots, \boldsymbol{x}_M$ and let $P \in \mathbf{R}^{n \times n}$ be the orthogonal projector onto V. If $t_*^2 = 4\sigma^2 \operatorname{rank}(\mathbf{X})/n$, then

$$\frac{1}{n}\mathbb{E} \sup_{\boldsymbol{v} \in V : \frac{1}{n}|\boldsymbol{v}|_2^2 \leq t_*^2} \boldsymbol{\xi}^T \boldsymbol{v} = \sqrt{\frac{t_*^2}{n}}\mathbb{E}|P\boldsymbol{\xi}|_2 \leq \sqrt{\frac{t_*^2 \sigma^2 \operatorname{rank}(\mathbf{X})}{n}} = t_*^2/2. \tag{23}$$

Let K be the convex hull of $\{\pm R\boldsymbol{x}_1, \dots, \pm R\boldsymbol{x}_M\}$, so that $K = \{\mathbf{X}\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbf{R}^M : |\boldsymbol{\beta}|_1 \leq R\}$. Let $\boldsymbol{f}_0^*$ be the convex projection of $\boldsymbol{f}_0$ onto K. We apply Proposition 5 to $K - \boldsymbol{f}_0^*$ which is a convex hull of $2M$ points of empirical norm less or equal to R^2. By (23) and (15), the quantity t_* satisfies (14). Applying Theorem 4 completes the proof. □

The lower bound [44], Theorem 5.4 and (5.25), states that there exists an absolute constant $C_0 > 0$ such that the following holds. If $\log(1 + eM/\sqrt{n}) \leq C_0\sqrt{n}$, then there exists a design matrix $\mathbf{X}$ such that for all estimator $\hat{\boldsymbol{f}}$,

$$\sup_{\boldsymbol{\beta} \in \mathbf{R}^M : |\boldsymbol{\beta}|_1 \leq R} \frac{1}{n}\mathbb{E}_{\mathbf{X}\boldsymbol{\beta}}|\mathbf{X}\boldsymbol{\beta} - \hat{\boldsymbol{f}}|_2^2 \geq \frac{1}{C_0}\min\left(\frac{\sigma^2 \operatorname{rank}(\mathbf{X})}{n}, \sigma R\sqrt{\frac{\log(1 + \frac{eM\sigma}{R\sqrt{n}})}{n}}\right),$$

where for all $\boldsymbol{f}_0 \in \mathbf{R}^n$, $\mathbb{E}_{\boldsymbol{f}_0}$ denotes the expectation with respect to the distribution of $\mathbf{y} \sim \mathcal{N}(\boldsymbol{f}_0, \sigma^2 I_{n\times n})$. Thus, Theorem 7 shows that the Least Squares estimator over the set $\{\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\beta} \in \mathbf{R}^M : |\boldsymbol{\beta}|_1 \leq R\}$ is minimax optimal. In particular, the right-hand side of inequality (22) cannot be improved.

There is an extensive literature on risk bounds on oracle inequalities for the Lasso (cf. [5,11–13,15,16,19,27,30,38,48,49,55] for a non-exhaustive list). However, to our knowledge, the fact that the Lasso enjoys the above upper bound with rate

$$\frac{\sigma R\sqrt{\log(\frac{2eM\sigma}{R\sqrt{n}})}}{\sqrt{n}}$$

holds with no assumption on the correlations of the design matrix, is new. This is an improvement over the rate $\sigma R\sqrt{(\log M)/n}$ which is known and which can be easily obtained by using high probability bounds on the maximum of M normal random variables.

The above result holds in completely anisotropic settings, that is, in situations where the design matrix $\mathbf{X}$ is very far from isotropic. This contrasts with numerous results on the Lasso which holds for sparse target vectors under Restricted Eigenvalue conditions, Compatibility conditions which grant isotropy on restricted cones. We refer the reader to [13] or the books [15,22,25] for more information on these conditions and the risk bounds satisfied by the Lasso in such settings.

Finally, let us note that the risk bounds obtained under such conditions on the design yield "fast" rates of convergence, of the form $O(1/n)$, while the rate obtained above is "slow", of the form $O(1/\sqrt{n})$. Although fast rates are appealing, it was recently shown in [7], Section 4.1, or [10], Section 3, that the compatibility condition is necessary to obtain fast rates of convergence, hence there is no hope to achieve fast rates of convergence with no assumption on the design. This motivates the study of risk bounds on the Lasso that are free of any assumption on the design such as Theorem 7.

4. The anisotropic persistence problem in regression with random design

Consider n i.i.d. observations $(Y_i, X_i)_{i=1,\dots,n}$ where $(Y_i)_{i=1,\dots,n}$ are real valued and the $(X_i)_{i=1,\dots,n}$ are design random variables in $\mathbf{R}^M$ with $\mathbb{E}[X_i X_i^T] = \Sigma$ for some covariance matrix $\Sigma \in \mathbf{R}^{M\times M}$. We consider the learning problem over the function class

$$\left\{f_\beta : f_\beta(x) = x^T\beta \text{ for some } \beta \in \mathbf{R}^M \text{ with } |\beta|_1 \le R\right\}$$

for a given constant $R > 0$. We consider the Emprical Risk Minimizer defined by

$$\hat{\boldsymbol{\beta}} = \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbf{R}^M:|\boldsymbol{\beta}|_1\le R} \sum_{i=1}^n \left(Y_i - \boldsymbol{\beta}^T X_i\right)^2. \tag{24}$$

This problem is sometimes referred to as the persistence problem or the persistence framework [4,24]. The prediction risk of $f_{\hat{\boldsymbol{\beta}}}$ is given by

$$R(f_{\hat{\boldsymbol{\beta}}}) = \mathbb{E}\left[\left(f_{\hat{\boldsymbol{\beta}}}(X) - Y\right)^2 \mid (X_i, Y_i)_{i=1,\dots,n}\right],$$

where (X, Y) is a new observation distributed as (X_1, Y_1) and independent from the data $(X_i, Y_i)_{i=1,\dots,n}$. Define also the oracle $\boldsymbol{\beta}^*$ by

$$\boldsymbol{\beta}^* = \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbf{R}^M:|\boldsymbol{\beta}|_1\le R} R(\boldsymbol{\beta}) \tag{25}$$

and define $\sigma > 0$ by

$$\sigma = \left\| Y - X^T \boldsymbol{\beta}^* \right\|_{\psi_2}, \tag{26}$$

where the sub-Gaussian norm $\|\cdot\|_{\psi_2}$ is defined by $\|Z\|_{\psi_2} = \sup_{p\geq 1} \mathbb{E}[|Z|^p]^{1/p}/\sqrt{p}$ for any random variable Z (see Section 5.2.3 in [53] for equivalent definitions of the ψ_2 norm).

To analyse the above learning problem, we use the machinery developed by Lecué and Mendelson [34] to study learning problems over sub-Gaussian classes. Consider the two quantities

$$r_n(\gamma) = \inf\Big\{ r > 0 : \mathbb{E} \sup_{\boldsymbol{\beta}: |\boldsymbol{\beta}|_1 \leq 2R,\, \mathbb{E}[(G^T \boldsymbol{\beta})^2] \leq s^2} \boldsymbol{\beta}^T G \leq \gamma r \sqrt{n} \Big\}, \tag{27}$$

$$s_n(\gamma) = \inf\Big\{ s > 0 : \mathbb{E} \sup_{\boldsymbol{\beta}: |\boldsymbol{\beta}|_1 \leq 2R,\, \mathbb{E}[(G^T \boldsymbol{\beta})^2] \leq s^2} \boldsymbol{\beta}^T G \leq \gamma s^2 \sqrt{n}/\sigma \Big\}, \tag{28}$$

where $G \sim N(\mathbf{0}, \Sigma)$. In the present setting, Theorem A from Lecué and Mendelson [34] reads as follows.

Theorem 8 (Theorem A in Lecué and Mendelson [34]). *There exist absolute constants $c_1, c_2, c_4 > 0$ such that the following holds. Let $R > 0$. Consider i.i.d. observations (X_i, Y_i) with $\mathbb{E}[X_i X_i^T] = \Sigma$. Assume that the design random vectors X_i are sub-Gaussian with respect to the covariance matrix Σ in the sense that $\|X_i^T \tau\|_{\psi_2} \leq 10|\Sigma^{1/2}\tau|_2$ for any $\tau \in \mathbf{R}^p$. Define $\boldsymbol{\beta}^*$ by (25) and σ by (26). Assume that the diagonal elements of Σ are no larger than 1. Then, there exists absolute constants $c_0, c_1, c_2, c_3 > 0$ such that the estimator $\hat{\boldsymbol{\beta}}$ defined in (24) satisfies*

$$R(f_{\hat{\boldsymbol{\beta}}}) \leq R(f_{\boldsymbol{\beta}^*}) + \max\big(s_n^2(c_1), r_n^2(c_2)\big),$$

with probability at least $1 - 6\exp(-c_4 n \min(c_2, s_n(c_1)))$.

In the isotropic case ($\Sigma = I_M$), [39] proves that

$$r_n^2(\gamma) \leq \begin{cases} \dfrac{c_3 R^2}{n} \log\left(\dfrac{c_3 M}{n}\right) & \text{if } n \leq c_4 M, \\ 0 & \text{if } n > c_4 M, \end{cases} \tag{29}$$

for some constants $c_3, c_4 > 0$ that only depends on γ, while

$$s_n^2(\gamma) \leq \begin{cases} \dfrac{c_5 R\sigma}{\sqrt{n}} \sqrt{\log\left(\dfrac{c_5 M\sigma}{\sqrt{n} R}\right)} & \text{if } n \leq c_6 \sigma^2 M^2/R^2, \\ \dfrac{c_5 \sigma^2 M}{n} & \text{if } n > c_6 \sigma^2 M^2/R^2, \end{cases} \tag{30}$$

for some constants $c_5, c_6 > 0$ that only depend on γ.

Using Proposition 1 and equation 10 above lets us extend these bounds to the anisotropic case where Σ is not proportional to the identity matrix.

Proposition 9. *Let $R > 0$, let $G \sim N(\mathbf{0}, \Sigma)$ and assume that the diagonal elements of Σ are no larger than 1. For any $\gamma > 0$, define $r_n(\gamma)$ and $s_n(\gamma)$ by (28) and (27). Then for any $\gamma > 0$, there exists constants $c_3, c_4, c_5, c_6 > 0$ that depend only on γ such that (30) and (29) hold.*

The proof of Proposition 9 will be given at the end of this subsection. The primary improvement of Proposition 1 over previous results is that this result is agnostic to the underlying covariance structure. This lets us handle the anisotropic case with $\Sigma \neq I_M$ in the above proposition. Proposition 9 combined with Theorem 8 lets us obtain the minimax rate of estimation for the persistence problem in the anisotropic case. Although the minimax rate was previously obtained in the isotropic case, we are not aware of a previous result that yields this rate for general covariance matrices $\Sigma \neq I_M$.

Since it holds for any covariance Σ, the result of the present section holds for covariance structure that are very far from isotropic. This contrasts with previous results of the literature, such as [4,42,47] or [39], Section 4.1. More precisely, [39], Section 4.1, assumes that Σ is proportional to identity, [4] assumes either that Σ is proportional to identity or that $|X_i|_\infty$ is almost surely bounded, [47] assumes Σ satisfies a Restricted Eigenvalue condition which grants isotropy on a restricted cone of $\mathbf{R}^M$, and finally, [42], (1.9) and Section 3, assumes that the condition number of Σ is bounded in order to handle anisotropic covariance structures.

Proof of Proposition 9. In this proof, $c > 0$ is an absolute constant whose value may change from line to line. Let $\gamma > 0$. We first bound $r_n(\gamma)$ from above. Let $r > 0$ and define

$$T_r(R) = \{\boldsymbol{\beta} \in \mathbf{R}^p : |\boldsymbol{\beta}|_1 \leq 2R, \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta} \leq r^2\}.$$

The random variable $G \sim N(\mathbf{0}, \Sigma)$ has the same distribution as $\Sigma^{1/2}\boldsymbol{g}$ where $\boldsymbol{g} \sim N(\mathbf{0}, I_M)$. Thus, the expectation inside the infimum in (27) is equal to

$$\mathbb{E} \sup_{\boldsymbol{\beta} \in T_r(R)} \boldsymbol{\beta}^T \Sigma^{1/2} \boldsymbol{g}. \tag{31}$$

To bound $r_n(\gamma)$ from above, it is enough to find some $r > 0$ such that (31) is bounded from above by $\gamma r \sqrt{n}$.

By the Cauchy–Schwarz inequality, (31) is bounded from above by $r\sqrt{M}$, which is smaller than $\gamma r \sqrt{n}$ for all all $r > 0$ provided that $n > c_4 M$ for some constant c_4 that only depends on γ. Hence, $r_n(\gamma) = 0$ provided that $n > c_4 M$.

We now bound $r_n(\gamma)$ from above in the regime $n \leq c_4 M$. Let $\boldsymbol{u}_1, \dots, \boldsymbol{u}_M$ be the columns of $\Sigma^{1/2}$ and let $\tilde{T}$ be the convex hull of the $2M$ points $\{\pm\boldsymbol{u}_1, \dots, \pm\boldsymbol{u}_M\}$. Using the fact that $T_r(R) = 2RT_{r/(2R)}(1) \subset 2R(\tilde{T} \cap (r/(2R))B_2)$, the right-hand side of the previous display is bounded from above by

$$2R\,\ell\big(\tilde{T} \cap (r/R)\big)B_2) \leq 8R\sqrt{\log_+\left(4eM\left(\frac{r}{2R}\right)^2\right)}, \tag{32}$$

where we used Proposition 1 for the last inequality. By simple algebra, one can show that if $r = c_3(\gamma)\frac{R}{\sqrt{n}}\sqrt{\log(c_3(\gamma)M/n)}$ for some large enough constant $c_3(\gamma)$ that only depends on γ, then the right-hand side of (32) is bounded from above by $\gamma r\sqrt{n}$. This completes the proof of (29).

We now bound $s_n(\gamma)$ from above. Let $s>0$. By definition of $s_n(\gamma)$, to prove that $s_n(\gamma)\le s$, it is enough to show that

$$\sigma \mathbb{E}_\xi \sup_{\boldsymbol{\beta}\in T_s(R)} \boldsymbol{\beta}^T \Sigma^{1/2}\boldsymbol{g}$$

is smaller than $\gamma s^2\sqrt{n}$. We use Proposition 1 to show that the previous display is bounded from above by

$$c\sigma \min\left(s\sqrt{M}, R\sqrt{\log_+\left(4eM\left(\frac{s}{2R}\right)^2\right)}\right).$$

By simple algebra similar to that of the proof of Proposition 5, we obtain that if s^2 equals the right hand side of (30) for large enough $c_5=c_5(\gamma)$ and $c_6=c_6(\gamma)$, then the right-hand side of the previous display is bounded from above by $\gamma s^2\sqrt{n}$. This completes the proof of (30). □

5. Bounded empirical processes and density estimation

We now prove a result similar to Proposition 1 for bounded empirical processes indexed by the convex hull of M points. This will be useful to study the convex aggregation problem for density estimation. Throughout this section, $\varepsilon_1,\dots,\varepsilon_n$ are i.i.d. Rademacher random variables that are independent of all other random variables.

Proposition 10. *There exists an absolute constant $c>0$ such that the following holds. Let $M\ge 2, n\ge 1$ be integers and let $b,R,L>0$ be real numbers. Let $Q(\boldsymbol{\theta})=\boldsymbol{\theta}^T\Sigma\boldsymbol{\theta}$ for some positive semi-definite matrix Σ. Let $Z_1,\dots,Z_n$ be i.i.d. random variables valued in some measurable set $\mathcal{Z}$. Let $h_1,\dots,h_M:\mathcal{Z}\to\mathbf{R}$ be measurable functions. Let $h_{\boldsymbol{\theta}}=\sum_{j=1}^M\theta_j h_j$ for all $\boldsymbol{\theta}=(\theta_1,\dots,\theta_M)^T\in\mathbf{R}^M$. Assume that almost surely*

$$\left|h_j(Z_1)\right|\le b, \qquad Q(\boldsymbol{e}_j)=\Sigma_{jj}\le R^2, \qquad \mathbb{E}\left[h_{\boldsymbol{\theta}}^2(Z_1)\right]\le LQ(\boldsymbol{\theta}), \tag{33}$$

for all $j=1,\dots,M$ and all $\boldsymbol{\theta}\in\Lambda^M$. Then for all $r>0$ such that $R/\sqrt{M}\le r\le R$ we have

$$\mathbb{E}\Big[\sup_{\boldsymbol{\theta}\in\Lambda^M:\,Q(\boldsymbol{\theta})\le r^2} F(\boldsymbol{\theta})\Big]\le c\max\left(\sqrt{L}R\sqrt{\frac{\log(eMr^2/R^2)}{n}}, \frac{bR^2\log(eMr^2/R^2)}{r^2 n}\right), \tag{34}$$

where $F(\boldsymbol{\theta})=\frac{1}{n}|\sum_{i=1}^n\varepsilon_i h_{\boldsymbol{\theta}}(Z_i)|$ for all $\boldsymbol{\theta}\in\mathbf{R}^M$.

Proof of Proposition 10. Let $m=\lfloor R^2/r^2\rfloor\ge 1$. The function F is convex since it can be written as the maximum of two linear functions. Applying Lemma 3 with the above notation and $t=r$ yields

$$\mathbb{E}\sup_{\boldsymbol{\theta}\in\Lambda^M:Q(\boldsymbol{\theta})\le r^2} F(\boldsymbol{\theta})\le \mathbb{E}\int_1^{+\infty} M(x)\frac{dx}{x^2}=\int_1^{+\infty}\mathbb{E}\big[M(x)\big]\frac{dx}{x^2}, \tag{35}$$

where the second inequality is a consequence of Fubini's theorem and for all $x \geq 1$,

$$M(x) = \max_{\boldsymbol{\theta} \in \Lambda_m^M:\, Q(\boldsymbol{\theta}) \leq x(r^2 + R^2/m)} F(\boldsymbol{\theta}).$$

Using (33) and the Rademacher complexity bound for finite classes given in [29], Theorem 3.5, we obtain that for all $x \geq 1$,

$$\mathbb{E}\big[M(x)\big] \leq c' \max\left(\sqrt{\frac{Lx(r^2 + R^2/m)\log|\Lambda_m^M|}{n}}, \frac{b\log|\Lambda_m^M|}{n}\right), \tag{36}$$

where $c' > 0$ is a numerical constant and $|\Lambda_m^M|$ is the cardinality of the set Λ_m^M. By definition of m we have $r^2 \leq R^2/m$. The cardinality $|\Lambda_m^M|$ of the set Λ_m^M is bounded from above by the right-hand side of (13). Combining inequality (35), inequality (36), the fact that the integrals $\int_1^{+\infty} \frac{dx}{x^2}$ and $\int_1^{+\infty} \frac{dx}{x^{3/2}}$ are finite, we obtain

$$\mathbb{E} \sup_{\boldsymbol{\theta} \in \Lambda^M : Q(\boldsymbol{\theta}) \leq r^2} F(\boldsymbol{\theta}) \leq c'' \max\left(\sqrt{L} R \sqrt{\frac{\log(eM/m)}{n}}, \frac{bm\log(eM/m)}{n}\right)$$

for some absolute constant $c'' > 0$. By definition of m, we have $R^2/(2r^2) \leq m \leq R^2/r^2$. A monotonicity argument completes the proof. $\square$

Next, we show that Proposition 10 can be used to derive a condition similar to (15) for bounded empirical processes. To bound from above the performance of ERM procedures in density estimation, Theorem 13 in the Appendix C requires the existence of a quantity $r_* > 0$ such that

$$\mathbb{E}\Big[\sup_{\boldsymbol{\theta} \in \Lambda^M:\, Q(\boldsymbol{\theta}) \leq r_*^2} F(\boldsymbol{\theta})\Big] \leq \frac{r_*^2}{16}, \tag{37}$$

where F is the function defined in Proposition 10 above.

To obtain such quantity $r_* > 0$ under the assumptions of Proposition 10, we proceed as follows. Let $K = \max(b, \sqrt{L})$ and assume that

$$MK > R\sqrt{n}.$$

Define $r^2 = CKR\sqrt{\log(eMK/(R\sqrt{n}))}$ where $C \geq 1$ is a numerical constant that will be chosen later. We now bound from above the right-hand side of (34). We have

$$\begin{aligned}\log\big(eMr^2/R^2\big) &\leq \log(C) + \log\big(eMK/(R\sqrt{n})\big) + (1/2)\log\log\big(eMK/(R\sqrt{n})\big) \\ &\leq \big(\log(C) + 3/2\big)\log\big(eMK/(R\sqrt{n})\big),\end{aligned}$$

where for the last inequality we used that $\log\log(u) \leq \log u$ for all $u > 1$ and that $\log(C) \leq \log(C)\log(eMK/(R\sqrt{n}))$, since $C \geq 1$ and $MK/(R\sqrt{n}) \geq 1$. Thus, the right-hand side of (34) is bounded from above by

$$c \max\left(\frac{\sqrt{\log(C) + 3/2}}{C}, \frac{\log(C) + 3/2}{C^2} \right) r^2.$$

It is clear that the above quantity is bounded from above by $r^2/16$ if the numerical constant C is large enough. Thus, we have proved that as long as $MK > R\sqrt{n}$, inequality (37) holds for

$$r_*^2 = CRK \sqrt{\frac{\log(eMK/(R\sqrt{n}))}{n}},$$

where $C \geq 1$ is a numerical constant.

ERM and convex aggregation in density estimation

We now study density estimation with respect to the squared loss. Consider a measurable space $(\mathcal{Z}, \mu)$ with measure μ and an unknown density p_0 with respect to the measure μ. We observe $Z_1, \dots, Z_n$ distributed according to p_0 and we are given preliminary estimators $p_1, \dots, p_M \in L_2(\mu)$. We assume that the preliminary estimators $p_1, \dots, p_M$ have been constructed with data different than $Z_1, \dots, Z_n$, and for the purpose of the present section, we assume that $p_1, \dots, p_M$ are deterministic functions in $L^2(\mu)$. For instance, $p_1, \dots, p_M$ may be kernel density estimators with different bandwidth parameters, we refer the reader to [51], Section 2.1, and the references therein.

Linear, convex and model-selection type aggregation of density estimators with respect to the L^2 loss is studied in [6,46]. Aggregation of density estimator with respect to the Kullback-Leibler loss is studied in [31,54].

The goal is to construct an estimator that performs almost as well as the best convex combination of $p_1, \dots, p_M$. Since we consider the L^2 loss, the empirical loss function

$$p \to \int p^2 \, d\mu - \frac{2}{n} \sum_{i=1}^{n} p(Z_i)$$

is an unbiased estimator of the L^2 loss $\int (p - p_0)^2 \, d\mu$, up to an additive constant that is independent of p.

The minimax optimal rate for the convex aggregation problem is known to be of order

$$\phi_M^C(n) := \min\left(\frac{M}{n}, \sqrt{\frac{\log(\frac{eM}{\sqrt{n}})}{n}} \right)$$

for regression with fixed design [44] and regression with random design [50] if the integers M and $\sqrt{n}$ satisfy $eM\sigma \leq R\sqrt{n}\exp(\sqrt{n})$ or equivalently $\phi_M^C(n) \leq 1$. The arguments for the convex aggregation lower bound from [50] can be readily applied to density estimation, showing that the rate $\phi_M^C(n)$ is a lower bound on the optimal rate of convex aggregation for density estimation.

We now use the results of the previous sections to show that ERM is optimal for the convex aggregation problem in density estimation.

Theorem 11. *There exists an absolute constant $c > 0$ such that the following holds. Let $(\mathcal{Z}, \mu)$ be a measurable space with measure μ. Let p_0 be an unknown density with respect to the measure μ. Let $Z_1, \dots, Z_n$ be i.i.d. random variables valued in $\mathcal{Z}$ with density p_0. Let $p_1, \dots, p_M \in L_2(\mu)$ and let $p_{\boldsymbol{\theta}} = \sum_{j=1}^M \theta_j p_j$ for all $\boldsymbol{\theta} = (\theta_1, \dots, \theta_M)^T \in \mathbf{R}^M$. Let*

$$\hat{\boldsymbol{\theta}} \in \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \Lambda^M} \left(\int p_{\boldsymbol{\theta}}^2 \, d\mu - \frac{2}{n} \sum_{i=1}^n p_{\boldsymbol{\theta}}(Z_i) \right).$$

Then for all $x > 0$, with probability greater than $1 - \exp(-x)$,

$$\int (p_{\hat{\boldsymbol{\theta}}} - p_0)^2 \, d\mu \le \min_{\boldsymbol{\theta} \in \Lambda^M} \int (p_{\boldsymbol{\theta}} - p_0)^2 \, d\mu + c \max\left(\frac{b_\infty M}{n}, R\sqrt{b_\infty} \sqrt{\frac{\log(\frac{eM\sqrt{b_\infty}}{R\sqrt{n}})}{n}} \right) + \frac{88 b_\infty x}{3n},$$

where $R^2 = 4 \max_{j=1,\dots,M} \int p_j^2 \, d\mu$ and $b_\infty = \max_{j=0,1,\dots,M} \|p_j\|_{L_\infty(\mu)}$.

Proof. It is a direct application of Theorem 13 in the Appendix C. If $M\sqrt{b_\infty} \le R\sqrt{n}$, a fixed point t_* is given by Lemma 14. If $M\sqrt{b_\infty} > R\sqrt{n}$, we use Proposition 10 with $Q(\boldsymbol{\theta}) = \int (p_0^* - p_{\boldsymbol{\theta}})^2$, $L = b_\infty$ and $b = b_\infty$. The bound (37) yields the existence of a fixed point t_* in this regime. □

Appendix A: Simulations: How correlations affect $\ell(T \cap sB_2)$

The upper bound on $\ell(T \cap sB_2)$ given in Proposition 1 applies to any convex set T with M extreme points and this bound is sharp for the ℓ_1-ball. However, the upper bound of Proposition 1 does not involve geometric characteristics of the set T. Since this upper bound is sharp for the ℓ_1-ball, one may conjecture that correlations among the extreme points of T will decrease the value $\ell(T \cap sB_2)$.

We study here with simulations how correlations among the extreme points of the convex set T affect the Gaussian width $\ell(T \cap sB_2)$. For a given correlation parameter $\rho \in (0, 1)$ and a given localization parameter $s \in (0, 1)$, we construct a set T_ρ as follows. First, we construct a matrix $\mathbf{X} \in \mathbf{R}^{n \times p}$ with n i.i.d. rows distributed according to $N(\mathbf{0}, \Sigma_\rho)$, where $\Sigma_\rho \in \mathbf{R}^{M \times M}$ is the matrix with diagonal entries equal to 1 and off-diagonal entries equal to ρ. Then we set T_ρ to be the convex envelope of $\{\pm \boldsymbol{x}_1/|\boldsymbol{x}_1|_2, \dots, \pm \boldsymbol{x}_M/|\boldsymbol{x}_M|_2\}$ where $\boldsymbol{x}_1, \dots, \boldsymbol{x}_M$ are the columns of $\mathbf{X}$. Thus, T_ρ is a convex set with $2M$ extreme points, and any two extreme points have correlation approximately equal to ρ. Next, we compute the Gaussian width $\ell(T_\rho \cap sB_2)$ approximately, by drawing k independent standard normal vectors $\boldsymbol{g}_1, \dots, \boldsymbol{g}_k$ and by taking the average, that is, we report the value $V(s, \rho) = \frac{1}{k} \sum_{l=1}^k \max_{\boldsymbol{u} \in T_\rho \cap sB_2} \boldsymbol{u}^T \boldsymbol{g}_l$. Computing $V(s, \rho)$ can be done by solving k convex programs. The values $V(s, \rho)$ are reported on Figure A.1 with $n = 100$, $M = 300$.

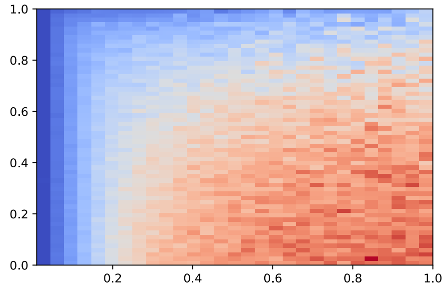

Figure A.1. Heatmap of the approximate values of the Gaussian width $\ell(T_\rho \cap (sB_2))$ for different correlation parameter $\rho \in (0,1)$ and localization parameter $s \in (0,1)$. Red is largest, blue is smallest. The x-axis represents the localization parameter s, the y-axis represents ρ the approximate correlation of two extreme points of T.

We see on Figure A.1 that the Gaussian width $\ell(T_\rho \cap sB_2)$ is monotonic with respect to the correlation parameter ρ for every fixed localization parameter s. This supports the theoretical conclusions of Propositions 1 and 2, which shows that the upper bound of Proposition 1 holds for any set T and that this upper bound is sharp for the ℓ_1-ball or for sets T whose extreme points have small correlations.

However, let us emphasize that the situation highlighted in Figure A.1 is specific to the construction of the sets T_ρ. It is still unclear, for general sets T, whether $\ell(T \cap sB_2)$ is monotonous with respect to correlations among the extreme points of T. More theoretical work is needed to understand the geometric features of the set T that characterize $\ell(T \cap sB_2)$; we leave these questions open for future work.

Appendix B: Proof of the lower bound (6)

Proof of Proposition 2. By the Varshamov–Gilbert extraction lemma [22], Lemma 2.5, there exist a subset Ω of $\{0,1\}^M$ such that

$$|\boldsymbol{\omega}|_0 = m, \qquad \left|\boldsymbol{\omega} - \boldsymbol{\omega}'\right|_0 > m, \qquad \log|\Omega| \geq (m/2)\log\bigl(M/(5m)\bigr)$$

for any distinct $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$.

For each $\boldsymbol{\omega} \in \Omega$, we define $\mathsf{sgn}(\boldsymbol{\omega}) \in \{-1,0,1\}^M$, a signed version of $\boldsymbol{\omega}$, as follows. Let $\varepsilon_1, \dots, \varepsilon_M$ be M i.i.d. Rademacher random variables. Then we have

$$\mathbb{E}\left[\left|\sum_{j=1}^M \omega_j \varepsilon_j \boldsymbol{\mu}_j\right|_2^2\right] = \sum_{j=1}^M \omega_j |\boldsymbol{\mu}_j|_2^2 = m.$$

Hence, there exists some $\mathsf{sgn}(\boldsymbol{\omega}) \in \{-1,0,1\}^M$ with $|\mathsf{sgn}(\boldsymbol{\omega})_j| = \omega_j$ for all $j = 1, \dots, M$ such that $|\boldsymbol{\mu}_{\mathsf{sgn}(\boldsymbol{\omega})}|_2^2 \leq m$.

Define $T_\Omega = \{s^2 \boldsymbol{\mu}_{\mathsf{sgn}(\boldsymbol{\omega})}, \boldsymbol{\omega} \in \Omega\}$. Since $s^2 = 1/m$, each element of T_Ω is of the form $(1/m)(\pm\boldsymbol{\mu}_{j_1} \pm ... \pm \boldsymbol{\mu}_{j_m})$ where $\boldsymbol{\mu}_{j_1}, \dots, \boldsymbol{\mu}_{j_m}$ are m distinct elements of $\{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_M\}$, hence by convexity of T we have $T_\Omega \subset T$. By definition of $\mathsf{sgn}(\boldsymbol{\omega})$, it holds that $T_\Omega \subset sB_2$, and thus $T_\Omega \subset T \cap sB_2$. For any two distinct $\boldsymbol{u}, \boldsymbol{v} \in T_\Omega$,

$$|\boldsymbol{u} - \boldsymbol{v}|_2^2 \geq \kappa^2 s^4 \inf_{\boldsymbol{\omega},\boldsymbol{\omega}'} \left|\mathsf{sgn}(\boldsymbol{\omega}) - \mathsf{sgn}(\boldsymbol{\omega}')\right|_2^2 > \kappa^2 s^4 m = \kappa^2 s^2,$$

where the supremum is taken over any two distinct elements of Ω. By Sudakov's inequality (see for instance [14], Theorem 13.4) we have

$$\ell(T \cap sB_2) \geq \ell(T_\Omega) \geq (1/2)\kappa s\sqrt{\log \Omega} \geq 1/(2\sqrt{2})\kappa s\sqrt{m}\sqrt{\log(M/5m)}.$$

Since $1/m = s^2$, the right hand side of the previous display is equal to the right hand side of (6) and the proof is complete. □

Appendix C: Local Rademacher complexities and density estimation

In the last decade emerged a vast literature on local Rademacher complexities to study the performance of empirical risk minimizers (ERM) for general learning problems, cf. [2,3,28] and the references therein. The following result is given in [2], Theorem 2.1. Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables, that are independent from all other random variables considered in the paper.

Theorem 12 (Bartlett et al. [2]). *Let $Z_1, \dots, Z_n$ be i.i.d. random variables valued in some measurable space $\mathcal{Z}$. Let $\mathcal{H} : \mathcal{Z} \to [-b_\infty, b_\infty]$ be a class of measurable functions. Assume that there is some $v > 0$ such that $\mathbb{E}[h(Z_1)^2] \leq v$ for all $h \in \mathcal{H}$. Then for all $x > 0$, with probability greater than $1 - \exp(-x)$,*

$$\sup_{h\in\mathcal{H}} (P - P_n)h \leq 4\mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i h(Z_i)\right] + \sqrt{\frac{2vx}{n}} + \frac{8b_\infty x}{3n}.$$

Theorem 12 is a straightforward consequence of Talagrand inequality. We now explain how Theorem 12 can be used to derive sharp oracle inequalities in density estimation.

Theorem 13. *Let $(\mathcal{Z}, \mu)$ be a measurable space with measure μ. Let p_0 be an unknown density with respect to the measure μ. Let $Z_1, \dots, Z_n$ be i.i.d. random variables valued in $\mathcal{Z}$ with density p_0. Let $\mathcal{P}$ be a convex subset of $L^2(\mu)$. Assume that there exists $p_0^* \in \mathcal{P}$ such that $\int (p_0 - p_0^*)^2\, d\mu = \inf_{p\in\mathcal{P}} \int (p_0 - p)^2\, d\mu$. Assume that for some $t_* > 0$,*

$$\mathbb{E}\left[\sup_{p\in\mathcal{P}:\, \int (p-p_0^*)^2\, d\mu \leq t_*^2} \frac{1}{n}\sum_{i=1}^n \varepsilon_i\big(p - p_0^*\big)(Z_i)\right] \leq \frac{t_*^2}{16}. \tag{38}$$

Assume that there exists an estimator $\hat{p}$ such that almost surely,

$$\hat{p} \in \arg\min_{p \in \mathcal{P}} \left(\int p^2 \, d\mu - \frac{1}{n} \sum_{i=1}^{n} 2p(Z_i) \right).$$

Then for all $x > 0$, with probability greater than $1 - \exp(-x)$,

$$\int (\hat{p} - p_0)^2 \, d\mu \le \min_{p \in \mathcal{P}} \int (p - p_0)^2 \, d\mu + 2 \max\left(t_*^2, \frac{4(\|p_0\|_{L_\infty(\mu)} + 8b_\infty/3)x}{n} \right),$$

where $b_\infty = \sup_{p \in \mathcal{P}} \|p\|_{L_\infty(\mu)}$.

Proof of Theorem 13. By optimality of $\hat{p}$ we have

$$\int (\hat{p} - p_0)^2 \, d\mu \le \int \left(p_0^* - p_0\right)^2 d\mu + 2\Xi_{\hat{p}},$$

where for all $p \in \mathcal{P}$, Ξ_p is the random variable

$$\Xi_p = (P - P_n)\left(p - p_0^*\right) - \frac{1}{2} \int \left(p_0^* - p\right)^2 d\mu.$$

Let $\rho = \max(t_*^2, 4(\|p_0\|_{L_\infty(\mu)} + 8b_\infty/3)x/n)$ and define

$$\mathcal{H} = \left\{ h = p_0^* - p \text{ for some } p \in \mathcal{P} \text{ such that } \int h^2 \, d\mu \le \rho \right\}.$$

The class $\mathcal{H}$ is convex, $0 \in \mathcal{H}$ and $t_*^2 \le \rho$ so that $h \in \mathcal{H}$ implies $\frac{t_*^2}{\rho} h \in \mathcal{H}$. For any linear form L,

$$\frac{1}{\rho} \sup_{h \in \mathcal{H}: \int h^2 \, d\mu \le \rho} L(h) \le \frac{1}{t_*^2} \sup_{h \in \mathcal{H}: \int h^2 \, d\mu \le \rho} L\left(\frac{t_*^2}{\rho} h \right) \le \frac{1}{t_*^2} \sup_{h \in \mathcal{H}: \int h^2 \, d\mu \le t_*^2} L(h)$$

so that by taking expectations, (38) holds if t_*^2 is replaced by ρ.

For any $h \in \mathcal{H}$, $\mathbb{E}[h(Z_1)^2] \le \|p_0\|_{L_\infty(\mu)} \rho$ and h is valued in $[-2b_\infty, 2b_\infty]$ μ-almost surely. We apply Theorem 12 to the class $\mathcal{H}$. This yields that with probability greater than $1 - e^{-x}$, if $p \in \mathcal{P}$ is such that $p_0^8 - p \in \mathcal{H}$, then

$$\begin{aligned}(P - P_n)\left(p_0^* - p\right) &\le \frac{\rho}{4} + \sqrt{\frac{2\rho \|p_0\|_{L_\infty(\mu)} x}{n}} + \frac{16 b_\infty x}{n} \\ &\le \frac{\rho}{2} + 2\left(\|p_0\|_{L_\infty(\mu)} + 8b_\infty/3\right) \frac{x}{n} \le \rho.\end{aligned}$$

On the same event of probability greater than $1 - e^{-x}$, if $p \in \mathcal{P}$ is such that $\int (p_0^* - p)^2 \, d\mu > \rho$, consider $h = \sqrt{\rho}(p_0^* - p)/\sqrt{\int (p_0^* - p)^2 \, d\mu}$ which belongs to $\mathcal{H}$. We have $(P - P_n)h \le \rho$, which can be rewritten

$$(P - P_n)\big(p_0^* - p\big) \le \sqrt{\rho}\sqrt{\int \big(p_0^* - p\big)^2 \, d\mu} \le \rho/2 + \int \big(p_0^* - p\big) \, d\mu/2,$$

so that $\Xi_p \le \rho/2 \le \rho$. In summary, we have proved that on an event of probability greater than $1 - e^{-x}$, $\sup_{p \in \mathcal{P}} \Xi_p \le \rho$. In particular, this holds for $p = \hat{p}$ which completes the proof. □

Appendix D: A fixed point t_* for finite dimensional classes

Lemma 14. *Consider the notations of Theorem 13 and assume that the linear span of $\mathcal{P}$ is finite dimensional of dimension d. Then (38) is satisfied for $t_*^2 = 256\|p_0\|_{L_\infty(\mu)} d/n$.*

Proof. Let $e_1, \dots, e_d$ be an orthonormal basis of the linear span of $\mathcal{P}$, for the scalar product $\langle p_1, p_2 \rangle = \int p_1 p_2 \, d\mu$. Then

$$\begin{aligned}\mathbb{E}\left[\sup_{p \in \mathcal{P}: \int (p - p_0^*)^2 \, d\mu \le t_*^2} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \big(p - p_0^*\big)(Z_i)\right] &\le \mathbb{E} \sup_{\boldsymbol{\theta} \in \mathbf{R}^d: |\boldsymbol{\theta}|_2^2 \le t_*^2} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \sum_{j=1}^d e_j(X_i) \\ &\le t_* \sqrt{\sum_{j=1}^d \left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i e_j(X_i)\right)^2} \\ &\le \frac{t_* \sqrt{\|p_0\|_{L_\infty(\mu)} d}}{\sqrt{n}} = \frac{t_*^2}{16},\end{aligned}$$

where we have used the Cauchy–Schwarz inequality, Jensen' inequality, and that $\mathbb{E} e_j(X)^2 \le \|p_0\|_{L_\infty(\mu)}$ for all $j = 1, \dots, d$. □

References

[1] Audibert, J.-Y. and Tsybakov, A.B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.* **35** 608–633. MR2336861

[2] Bartlett, P.L., Bousquet, O. and Mendelson, S. (2005). Local Rademacher complexities. *Ann. Statist.* **33** 1497–1537. MR2166554

[3] Bartlett, P.L. and Mendelson, S. (2006). Empirical minimization. *Probab. Theory Related Fields* **135** 311–334. MR2240689

[4] Bartlett, P.L., Mendelson, S. and Neeman, J. (2012). ℓ_1-Regularized linear regression: Persistence and oracle inequalities. *Probab. Theory Related Fields* **154** 193–224. MR2981422

[5] Bellec, P. and Tsybakov, A. (2017). Bounds on the prediction error of penalized least squares estimators with convex penalty. In *Modern Problems of Stochastic Analysis and Statistics*. *Springer Proc. Math. Stat.* **208** (V. Panov, ed.) 315–333. Cham: Springer. MR3747672

[6] Bellec, P.C. (2017). Optimal exponential bounds for aggregation of density estimators. *Bernoulli* **23** 219–248. MR3556772

[7] Bellec, P.C. (2017). Optimistic lower bounds for convex regularized least-squares. ArXiv preprint. Available at arXiv:1703.01332.

[8] Bellec, P.C. (2018). Optimal bounds for aggregation of affine estimators. *Ann. Statist.* **46** 30–59. MR3766945

[9] Bellec, P.C. (2018). Sharp oracle inequalities for least squares estimators in shape restricted regression. *Ann. Statist.* **46** 745–780. MR3782383

[10] Bellec, P.C. (2018). The noise barrier and the large signal bias of the lasso and other convex estimators. ArXiv preprint. Available at arXiv:1804.01230.

[11] Bellec, P.C., Dalalyan, A.S., Grappin, E. and Paris, Q. (2018). On the prediction loss of the lasso in the partially labeled setting. *Electron. J. Stat.* **12** 3443–3472. MR3864589

[12] Bellec, P.C., Lecué, G. and Tsybakov, A.B. (2018). Slope meets Lasso: Improved oracle bounds and optimality. *Ann. Statist.* **46** 3603–3642. MR3852663

[13] Bickel, P.J., Ritov, Y. and Tsybakov, A.B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[14] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford: Oxford Univ. Press. With a foreword by Michel Ledoux. MR3185193

[15] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. *Springer Series in Statistics*. Heidelberg: Springer. MR2807761

[16] Candès, E.J. and Plan, Y. (2009). Near-ideal model selection by ℓ_1 minimization. *Ann. Statist.* **37** 2145–2177. MR2543688

[17] Chatterjee, S. (2014). A new perspective on least squares under convex constraint. *Ann. Statist.* **42** 2340–2381. MR3269982

[18] Dai, D., Rigollet, P. and Zhang, T. (2012). Deviation optimal learning using greedy Q-aggregation. *Ann. Statist.* **40** 1878–1905. MR3015047

[19] Dalalyan, A.S., Hebiri, M. and Lederer, J. (2017). On the prediction performance of the Lasso. *Bernoulli* **23** 552–581. MR3556784

[20] Dalalyan, A.S. and Tsybakov, A.B. (2007). Aggregation by exponential weighting and sharp oracle inequalities. In *Learning Theory*. *Lecture Notes in Computer Science* **4539** 97–111. Berlin: Springer. MR2397581

[21] den Hollander, F. (2012). Probability theory: The coupling method.

[22] Giraud, C. (2014). *Introduction to High-Dimensional Statistics*. *Chapman amd Hall/CRC Monographs on Statistics and Applied Probability* **138**. Boca Raton, FL: CRC Press.

[23] Gordon, Y., Litvak, A.E., Mendelson, S. and Pajor, A. (2007). Gaussian averages of interpolated bodies and applications to approximate reconstruction. *J. Approx. Theory* **149** 59–73. MR2371614

[24] Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971–988. MR2108039

[25] Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. *Monographs on Statistics and Applied Probability* **143**. Boca Raton, FL: CRC Press. MR3616141

[26] Juditsky, A., Rigollet, P. and Tsybakov, A.B. (2008). Learning by mirror averaging. *Ann. Statist.* **36** 2183–2206. MR2458184

[27] Knight, K. and Fu, W. (2000). Asymptotics for Lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787

[28] Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34** 2593–2656. MR2329442

[29] Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. *Lecture Notes in Math*. **2033**. Heidelberg: Springer. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School]. MR2829871

[30] Koltchinskii, V., Lounici, K. and Tsybakov, A.B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann*. *Statist*. **39** 2302–2329. MR2906869

[31] Lecué, G. (2006). Lower bounds and aggregation in density estimation. *J*. *Mach*. *Learn*. *Res*. **7** 971–981. MR2274393

[32] Lecué, G. (2013). Empirical risk minimization is optimal for the convex aggregation problem. *Bernoulli* **19** 2153–2166. MR3160549

[33] Lecué, G. and Mendelson, S. (2009). Aggregation via empirical risk minimization. *Probab. Theory Related Fields* **145** 591–613. MR2529440

[34] Lecué, G. and Mendelson, S. (2013). Learning subGaussian classes: Upper and minimax bounds. ArXiv preprint. Available at arXiv:1305.4825.

[35] Lecué, G. and Rigollet, P. (2014). Optimal learning with Q-aggregation. *Ann*. *Statist*. **42** 211–224. MR3178462

[36] Leung, G. and Barron, A.R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans*. *Inform*. *Theory* **52** 3396–3410. MR2242356

[37] Lounici, K. (2007). Generalized mirror averaging and D-convex aggregation. *Math*. *Methods Statist*. **16** 246–259. MR2356820

[38] Lounici, K., Pontil, M., van de Geer, S. and Tsybakov, A.B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann*. *Statist*. **39** 2164–2204. MR2893865

[39] Mendelson, S. (2015). Learning without concentration. *J. ACM* **62** Article ID 21. MR3367000

[40] Nemirovski, A. (2000). Topics in non-parametric statistics. In *Lectures on Probability Theory and Statistics* (*Saint-Flour*, 1998). *Lecture Notes in Math*. **1738** 85–277. Berlin: Springer. MR1775640

[41] Pisier, G. (1981). Remarques sur un résultat non publié de B. Maurey. In *Seminar on Functional Analysis*, 1980–1981, *Exp*. *No*. 5 1–12. Palaiseau: École Polytech. MR0659306

[42] Plan, Y. and Vershynin, R. (2016). The generalized Lasso with non-linear observations. *IEEE Trans*. *Inform*. *Theory* **62** 1528–1537. MR3472264

[43] Plan, Y., Vershynin, R. and Yudovina, E. (2017). High-dimensional estimation with geometric constraints. *Inf. Inference* **6** 1–40. MR3636866

[44] Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann*. *Statist*. **39** 731–771. MR2816337

[45] Rigollet, P. and Tsybakov, A.B. (2012). Sparse estimation by exponential weighting. *Statist*. *Sci*. **27** 558–575. MR3025134

[46] Rigollet, P. and Tsybakov, A.B. (2007). Linear and convex aggregation of density estimators. *Math*. *Methods Statist*. **16** 260–280. MR2356821

[47] Rudelson, M. and Zhou, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Trans*. *Inform*. *Theory* **59** 3434–3447. MR3061256

[48] Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166

[49] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J*. *Roy*. *Statist*. *Soc*. *Ser. B* **58** 267–288. MR1379242

[50] Tsybakov, A.B. (2003). Optimal rates of aggregation. In *Learning Theory and Kernel Machines* 303–313. Springer.

[51] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. New York: Springer. Revised and extended from the 2004 French original. Translated by Vladimir Zaiats. MR2724359

[52] Tsybakov, A.B. (2014). Aggregation and minimax optimality in high-dimensional estimation. In *Proceedings of the International Congress of Mathematicians—Seoul* 2014. *Vol*. *IV* 225–246. Seoul: Kyung Moon Sa. MR3727610

[53] Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge: Cambridge Univ. Press. MR2963170

[54] Yang, Y. (2000). Mixing strategies for density estimation. *Ann*. *Statist*. **28** 75–87. MR1762904

[55] Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann*. *Statist*. **36** 1567–1594. MR2435448

Received November 2017 and revised June 2018