

OPTIMAL BOUNDS FOR AGGREGATION OF AFFINE ESTIMATORS

BY PIERRE C. BELLEC¹

ENSAE and Rutgers University

We study the problem of aggregation of estimators when the estimators are not independent of the data used for aggregation and no sample splitting is allowed. If the estimators are deterministic vectors, it is well known that the minimax rate of aggregation is of order $\log(M)$, where M is the number of estimators to aggregate. It is proved that for affine estimators, the minimax rate of aggregation is unchanged: it is possible to handle the linear dependence between the affine estimators and the data used for aggregation at no extra cost. The minimax rate is not impacted either by the variance of the affine estimators, or any other measure of their statistical complexity. The minimax rate is attained with a penalized procedure over the convex hull of the estimators, for a penalty that is inspired from the Q -aggregation procedure. The results follow from the interplay between the penalty, strong convexity and concentration.

1. Introduction. We study the problem of recovering an unknown vector $\mathbf{f} = (f_1, \dots, f_n)^T \in \mathbf{R}^n$ from noisy observations:

$$(1.1) \quad Y_i = f_i + \xi_i, \quad i = 1, \dots, n,$$

where the noise random variables ξ_1, \dots, ξ_n are i.i.d. $\mathcal{N}(0, \sigma^2)$ or i.i.d. sub-Gaussian random variables. We measure the quality of estimation of the unknown vector \mathbf{f} with the squared Euclidean norm in \mathbf{R}^n :

$$\|\mathbf{f} - \hat{\boldsymbol{\mu}}\|_2^2,$$

for any estimator $\hat{\boldsymbol{\mu}}$ of \mathbf{f} . When the noise random variables are normal, (1.1) is the Gaussian sequence model, which has been extensively studied; see, for example, [22] and the references therein. Several estimators have been proposed to recover the unknown vector \mathbf{f} from the observations: the ordinary least squares, the ridge estimator, the Stein estimator and the procedures based on shrinkage, to name a few. Several of these estimators depend on a parameter that must be chosen carefully to obtain satisfying error bounds. These available estimators have different

Received September 2015; revised December 2016.

¹Supported by GENES and by the grant Investissements d’Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047).

MSC2010 subject classifications. Primary 62G05; secondary 62J07.

Key words and phrases. Affine estimator, aggregation, sequence model, sharp oracle inequality, concentration inequality, Hanson–Wright.

strengths and weaknesses in different scenarios, so it is important to be able to mimic the best among a given family of estimators, without any assumption on the unknown \mathbf{f} . The problem of mimicking the best estimator in a given finite set is the problem of model-selection type aggregation, which was introduced in [30, 37]. More precisely, let $\hat{\boldsymbol{\mu}}_1, \dots, \hat{\boldsymbol{\mu}}_M$ be M estimators of \mathbf{f} based on the data $\mathbf{y} = (Y_1, \dots, Y_n)^T$. The goal is to construct with the same data $\mathbf{y} = (Y_1, \dots, Y_n)^T$ a new estimator $\hat{\boldsymbol{\mu}}$ called the aggregate, which satisfies with probability greater than $1 - \delta$ the sharp oracle inequality²

$$(1.2) \quad \|\hat{\boldsymbol{\mu}} - \mathbf{f}\|_2^2 \leq \min_{j=1, \dots, M} \|\hat{\boldsymbol{\mu}}_j - \mathbf{f}\|_2^2 + \text{PRICE}_M(\delta),$$

where $\text{PRICE}_M(\cdot)$ is a function of δ that should be small. The term $\text{PRICE}_M(\cdot)$ will be referred to as the price to pay for aggregating the estimators $\hat{\boldsymbol{\mu}}_1, \dots, \hat{\boldsymbol{\mu}}_M$. If the estimators $\hat{\boldsymbol{\mu}}_1, \dots, \hat{\boldsymbol{\mu}}_M$ are deterministic vectors, the price to pay for aggregating these estimators is of order $\sigma^2 \log(M/\delta)$ and (1.2) is satisfied for an estimator $\hat{\boldsymbol{\mu}}$ based on Q -aggregation [11]. Considering deterministic estimators is of interest if two independent samples are available, so that $\hat{\boldsymbol{\mu}}_1, \dots, \hat{\boldsymbol{\mu}}_M$ are based on the first sample while aggregation is performed using the second sample. Then the first sample can be considered as frozen at the aggregation step (for more details see [39]). If the estimators are random (dependent on the data \mathbf{y} used for aggregation), two natural questions arise:

1. Does the price to pay for aggregation increase because of the dependence between $\hat{\boldsymbol{\mu}}_1, \dots, \hat{\boldsymbol{\mu}}_M$ and the data \mathbf{y} , or is it still of order $\sigma^2 \log(M/\delta)$? Is there an extra price to pay to handle the dependence?
2. A natural quantity that captures the statistical complexity of a given estimator $\hat{\boldsymbol{\mu}}_j$ is the variance defined by $\mathbb{E}\|\hat{\boldsymbol{\mu}}_j - \mathbb{E}\hat{\boldsymbol{\mu}}_j\|_2^2$. When the estimators are deterministic, their variances are all zero. Now that the estimators are random, does the price to pay for aggregation depend on the statistical complexities of the estimators $\hat{\boldsymbol{\mu}}_1, \dots, \hat{\boldsymbol{\mu}}_M$, for example, through their variances? Is it harder to aggregate estimators with large statistical complexities?

The goal of this paper is to answer these questions for affine estimators.

Among the procedures available to estimate \mathbf{f} , several are linear in the observations Y_1, \dots, Y_n . It is the case for the least squares and the ridge estimators, whereas the shrinkage estimators and the Stein estimator are nonlinear functions of the observations. Examples of estimators that are linear or affine in the observations is given in [12], Section 1.2, [1] and references therein. An affine estimator is of the form $\hat{\boldsymbol{\mu}}_j = A_j \mathbf{y} + \mathbf{b}_j$ for a deterministic matrix A_j of size $n \times n$ and a deterministic vector $\mathbf{b}_j \in \mathbf{R}^n$. The linearity of the estimators $\hat{\boldsymbol{\mu}}_1, \dots, \hat{\boldsymbol{\mu}}_M$ makes it possible to explicitly treat the dependence between the estimators $\hat{\boldsymbol{\mu}}_1, \dots, \hat{\boldsymbol{\mu}}_M$ and the data $\mathbf{y} = (Y_1, \dots, Y_n)^T$ used to aggregate them. Donoho et al. [14] proved

²By sharp, we mean that the constant in front of the term $\min_{j=1, \dots, M} \|\hat{\boldsymbol{\mu}}_j - \mathbf{f}\|_2^2$ is 1.

that for orthosymmetric quadratically convex sets (which include all ellipsoids and hyperrectangles), the minimax risk among all linear estimators is within 25% of the minimax risk among all estimators.

The papers [10, 12, 27] derived different procedures that satisfy sharp oracle inequalities for the problem of aggregation of affine estimators when the noise random variables are Gaussian. Leung and Barron [27], Dalalyan and Salmon [12] proposed an estimator $\hat{\boldsymbol{\mu}}^{\text{EW}}$ based on exponential weights, for which the following sharp oracle inequality holds in expectation:

$$\mathbb{E}\|\mathbf{f} - \hat{\boldsymbol{\mu}}^{\text{EW}}\|_2^2 \leq \min_{j=1,\dots,M} \mathbb{E}\|\hat{\boldsymbol{\mu}}_j - \mathbf{f}\|_2^2 + 8\sigma^2 \log M,$$

under the assumption that all A_j are orthoprojectors [orthogonal projection matrices; cf. (1.4)], or under a strong commutativity assumption on the matrices A_j . The constant 8 can be reduced to 4 if all A_j are orthoprojectors. If the matrices A_j are not symmetric, [12] achieved a similar oracle inequality by symmetrizing the affine estimators before the aggregation step, which suggests that the symmetry assumption can be relaxed. Although the estimator $\hat{\boldsymbol{\mu}}^{\text{EW}}$ achieves this inequality in expectation, it was shown in [2, 11] that it cannot achieve a similar result in deviation, with an unavoidable error term of order \sqrt{n} . In Dai et al. [10], a sharp oracle inequality in deviation is derived for an estimator $\hat{\boldsymbol{\mu}}^{\mathcal{Q}}$ based on \mathcal{Q} -aggregation [11, 31]. Namely, [10] proves that if the matrices A_1, \dots, A_M are symmetric and positive semidefinite, the estimator $\hat{\boldsymbol{\mu}}^{\mathcal{Q}}$ satisfies with probability greater than $1 - \delta$:

$$(1.3) \quad \|\mathbf{f} - \hat{\boldsymbol{\mu}}^{\mathcal{Q}}\|_2^2 \leq \min_{j=1,\dots,M} (\|\hat{\boldsymbol{\mu}}_j - \mathbf{f}\|_2^2 + 4\sigma^2 \text{Tr}(A_j)) + C\sigma^2 \log(M/\delta),$$

where the constant C is proportional to the largest operator norm of the matrices A_1, \dots, A_M . The term $4\sigma^2 \text{Tr}(A_j)$ is intimately linked to the statistical complexity of the estimator $\hat{\boldsymbol{\mu}}_j = A_j \mathbf{y} + \mathbf{b}_j$. For instance, the variance of $\hat{\boldsymbol{\mu}}_j$ is $\mathbb{E}\|\hat{\boldsymbol{\mu}}_j - \mathbb{E}\hat{\boldsymbol{\mu}}_j\|_2^2 = \sigma^2 \text{Tr}(A_j^T A_j)$. If $\hat{\boldsymbol{\mu}}_j$ is a least squares estimator, A_j is an orthoprojector, and the variance becomes $\sigma^2 \text{Tr} A_j$. Thus, the statistical complexity of the estimator $\hat{\boldsymbol{\mu}}_j$ clearly appears in the right-hand side of the oracle inequality (1.3) proved in [10]. Thus, one may think that the price to pay for aggregating affine estimators, that is, the function $\text{PRICE}_M(\delta)$ in (1.2), depends on the statistical complexity of the estimators to aggregate.

The bound (1.3) may lead to the conclusion that the price to pay for aggregation of affine estimators can be substantially larger than $\sigma^2 \log(M/\delta)$ which is the price for aggregating deterministic vectors. Indeed, the extra term $4\sigma^2 \text{Tr}(A_j)$ may be large in common situation where the trace of some matrices A_j is large. For example, if one aggregates the estimators $\hat{\boldsymbol{\mu}}_1 = \lambda_1 \mathbf{y}, \dots, \hat{\boldsymbol{\mu}}_M = \lambda_M \mathbf{y}$, for some positive real numbers $\lambda_1, \dots, \lambda_M$, then the term $4\sigma^2 \text{Tr}(A_j)$ in the above oracle inequality is of order $\sigma^2 n \lambda_j$ for each $j = 1, \dots, M$, which can be greater than the optimal rate $\sigma^2 \log M$. This term $4\sigma^2 \text{Tr}(A_j)$ makes the oracle inequality (1.3) suitable only for scenarios where the matrices A_j have small trace. But more importantly, the term

$\sigma^2 \text{Tr } A_j$ suggests that the price to pay for aggregating affine estimators increases with the statistical complexities of the estimators to aggregate.

The results discussed above rely on specific assumptions on the matrices A_1, \dots, A_M [10, 12, 27]. This raises a third question, although not as important as the two questions above:

3. Does the nature of the matrices A_1, \dots, A_M have an impact on the price to pay to aggregate these affine estimators? Is the price in (1.2) substantially smaller if the matrices are orthoprojectors, semipositive definite or symmetric?

The main contribution of the present paper is to answer the three questions raised above:

1. It is proved in Theorem 2.1 that a penalized procedure over the simplex satisfies the sharp oracle inequality (1.2) with $\text{PRICE}_M(\delta) = c\sigma^2 \log(M/\delta)$ for some absolute constant $c > 0$. This price is of the same order as for the problem of aggregation of deterministic vectors. Thus, the dependence between the estimators and the data used to aggregate them induces no extra cost.

2. The form of the affine estimators to aggregate has no impact on the price to pay for aggregation. In particular, the sharp oracle inequalities of the present paper do not involve quantities dependent on A_j such as $\sigma^2 \text{Tr } A_j$.

3. The only assumption made on the matrices A_1, \dots, A_M is that $\max_{j \neq k} \|A_j - A_k\|_2$ is bounded from above, where $\|\cdot\|_2$ is the operator norm. All other assumptions on the matrices A_1, \dots, A_M can be dropped, in particular the matrices can be nonsymmetric and have negative eigenvalues.

The paper is organized as follows. In Section 1.1, we define the notation used throughout the paper. Section 2 defines a penalized procedure over the simplex and shows that it achieves sharp oracle inequalities in deviation for aggregation of affine estimators. The role of the penalty is studied in Section 3 and Section 4. Prior weights are considered in Section 5. Section 6 shows that the estimator is robust to variance misspecification and to non-Gaussianity of the noise. Some examples are given in Section 7. Section 8 is devoted to the proofs.

1.1. *Notation.* Let $\mathbf{f} = (f_1, \dots, f_n)^T \in \mathbf{R}^n$ be an unknown regression vector. We observe n random variables (1.1) where ξ_1, \dots, ξ_n are sub-Gaussian random variables, with $\mathbb{E}[\xi_i] = 0$ and $\mathbb{E}[\xi_i^2] = \sigma^2$. It can be rewritten in the vector form $\mathbf{y} = \mathbf{f} + \boldsymbol{\xi}$ where $\mathbf{y} = (Y_1, \dots, Y_n)^T$, $\mathbf{f} = (f_1, \dots, f_n)^T$ and $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^T$.

For any estimator $\hat{\boldsymbol{\mu}}$ of \mathbf{f} , we measure the quality of estimation of \mathbf{f} with the loss $\|\hat{\boldsymbol{\mu}} - \mathbf{f}\|_2^2$, where $\|\cdot\|_2$ is the Euclidean norm in \mathbf{R}^n . Let $M \geq 2$. We consider M affine estimators of the form

$$\hat{\boldsymbol{\mu}}_j = A_j \mathbf{y} + \mathbf{b}_j, \quad j = 1, \dots, M.$$

The matrices A_1, \dots, A_M and the vectors $\mathbf{b}_1, \dots, \mathbf{b}_M \in \mathbf{R}^n$ are deterministic. Define the simplex in \mathbf{R}^M :

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

For any $\boldsymbol{\theta} \in \Lambda^M$, let

$$A_{\boldsymbol{\theta}} = \sum_{j=1}^M \theta_j A_j, \quad \mathbf{b}_{\boldsymbol{\theta}} = \sum_{j=1}^M \theta_j \mathbf{b}_j, \quad \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} = A_{\boldsymbol{\theta}} \mathbf{y} + \mathbf{b}_{\boldsymbol{\theta}}.$$

Let $\mathbf{e}_1, \dots, \mathbf{e}_M$ be the vectors of the canonical basis in \mathbf{R}^M . Then $\hat{\boldsymbol{\mu}}_{\mathbf{e}_j} = \hat{\boldsymbol{\mu}}_{\mathbf{e}_j}$ for all $j = 1, \dots, M$.

An orthoprojector is an $n \times n$ matrix P such that

$$(1.4) \quad P = P^T = P^2.$$

Denote by $I_{n \times n}$ the $n \times n$ -identity matrix. For any $n \times n$ real matrix $A = (a_{i,j})_{i,j=1,\dots,n}$, define the operator norm of A , the Frobenius (or Hilbert–Schmidt) norm of A and the nuclear norm of A respectively by

$$\|A\|_2 = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}, \quad \|A\|_F = \sqrt{\sum_{i,j=1,\dots,n} a_{i,j}^2}.$$

The following inequalities hold for any two squared matrices M, M' :

$$(1.5) \quad \|MM'\|_2 \leq \|M\|_2 \|M'\|_2, \quad \|MM'\|_F \leq \|M\|_2 \|M'\|_F.$$

Finally, denote by \log the natural logarithm with $\log(e) = 1$.

2. A penalized procedure on the simplex. For any $\boldsymbol{\theta} \in \Lambda^M$, define

$$(2.1) \quad C_p(\boldsymbol{\theta}) := \|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}\|_2^2 - 2\mathbf{y}^T \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} + 2\sigma^2 \text{Tr}(A_{\boldsymbol{\theta}}),$$

which is Mallows [28] C_p -criterion. Next, define

$$(2.2) \quad H_{\text{pen}}(\boldsymbol{\theta}) = C_p(\boldsymbol{\theta}) + \frac{1}{2} \text{pen}(\boldsymbol{\theta}),$$

where

$$(2.3) \quad \text{pen}(\boldsymbol{\theta}) = \sum_{j=1}^M \theta_j \|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \hat{\boldsymbol{\mu}}_j\|_2^2.$$

We consider the estimator $\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}}$ where

$$(2.4) \quad \hat{\boldsymbol{\theta}}_{\text{pen}} \in \underset{\boldsymbol{\theta} \in \Lambda^M}{\text{argmin}} H_{\text{pen}}(\boldsymbol{\theta}).$$

The function H_{pen} is quadratic and convex (cf. Lemma 8.2). Minimizing H_{pen} over the simplex is a convex quadratic program for which efficient algorithms are available. The convexity of H_{pen} also proves that $\hat{\boldsymbol{\theta}}_{\text{pen}}$ is well defined, although it may not be unique (e.g., if all $\hat{\boldsymbol{\mu}}_j$ are the same then H_{pen} is constant on the simplex).

We now explain the meaning of the terms that appear in (2.2). If $\boldsymbol{\theta}$ is fixed, $C_p(\boldsymbol{\theta})$ is an unbiased estimate of the quantity

$$(2.5) \quad R(\boldsymbol{\theta}) := \|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}\|_2^2 - 2\mathbf{f}^T \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} = \|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \mathbf{f}\|_2^2 - \|\mathbf{f}\|_2^2,$$

which is the quantity of interest $\|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \mathbf{f}\|_2^2$ up to the additive constant $\|\mathbf{f}\|_2^2$.

The penalty (2.3) is borrowed from the Q -aggregation procedure, which is a powerful tool to derive sharp oracle inequalities in deviation when the loss is strongly convex [4, 11, 26, 31]. Since the estimators $\hat{\boldsymbol{\mu}}_1, \dots, \hat{\boldsymbol{\mu}}_M$ depend on the data, the penalty (2.3) is data-driven, which is not the case if $\hat{\boldsymbol{\mu}}_1, \dots, \hat{\boldsymbol{\mu}}_M$ are deterministic vectors as in [11]. In order to give some geometric insights on the penalty (2.3), let $c \in \mathbf{R}^n$ be a solution of M linear equations $2c^T \hat{\boldsymbol{\mu}}_j = \|\hat{\boldsymbol{\mu}}_j\|_2^2$, $j = 1, \dots, M$, and assume only in the rest of this paragraph that such a solution exists, even though this assumption cannot be fulfilled for $M > n$. Then

$$(2.6) \quad \text{pen}(\boldsymbol{\theta}) = \sum_{j=1}^M \theta_j \|\hat{\boldsymbol{\mu}}_j\|_2^2 - \|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}\|_2^2 = 2c^T \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}\|_2^2 = \|c\|_2^2 - \|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - c\|_2^2.$$

We can write $\text{pen}(\boldsymbol{\theta}) = g(\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}})$ for some function g defined on the convex hull of $\{\hat{\boldsymbol{\mu}}_1, \dots, \hat{\boldsymbol{\mu}}_M\}$. Equation (2.6) shows that the level sets of the function g are Euclidean balls centered at c . The function g is nonnegative. It is minimal at the extreme points $\hat{\boldsymbol{\mu}}_1, \dots, \hat{\boldsymbol{\mu}}_M$ since $g(\hat{\boldsymbol{\mu}}_j) = 0$ for all $j = 1, \dots, M$ and g is maximal at the projection of c on the convex hull of $\{\hat{\boldsymbol{\mu}}_1, \dots, \hat{\boldsymbol{\mu}}_M\}$. Intuitively, the penalty (2.3) pushes $\boldsymbol{\theta}$ away from the center of the simplex towards the vertices. Thus, the level sets of the function $\boldsymbol{\theta} \rightarrow \text{pen}(\boldsymbol{\theta})$ in \mathbf{R}^M are ellipsoids centered at $\boldsymbol{\theta}_c$, where $\boldsymbol{\theta}_c$ is the unique point in \mathbf{R}^M such that $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}_c} = c$. If $M > n$ or if the vector c is not well defined, the level sets of $\text{pen}(\cdot)$ are more intricate and cannot be described in such a simple way.

THEOREM 2.1 (Main result). *Let $M \geq 2$. For $j = 1, \dots, M$, consider the affine estimators $\hat{\boldsymbol{\mu}}_j = A_j \mathbf{y} + \mathbf{b}_j$ and let*

$$(2.7) \quad \phi := \max\left(1, \max_{j,k=1,\dots,M:j \neq k} \frac{1}{2} \|A_j - A_k\|_2\right).$$

Assume that the noise random variables ξ_1, \dots, ξ_n are i.i.d. $\mathcal{N}(0, \sigma^2)$. Let $\hat{\boldsymbol{\theta}}_{\text{pen}}$ be the estimator defined in (2.4). Then for all $x > 0$ the estimator $\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}}$ satisfies with probability greater than $1 - \exp(-x)$,

$$(2.8) \quad \|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}} - \mathbf{f}\|_2^2 \leq \min_{j=1,\dots,M} \|\hat{\boldsymbol{\mu}}_j - \mathbf{f}\|_2^2 + 30\phi^2 \sigma^2 (x + 2 \log M).$$

Furthermore,

$$(2.9) \quad \mathbb{E}[\|\hat{\boldsymbol{\mu}}_{\hat{\theta}_{\text{pen}}} - \mathbf{f}\|_2^2] \leq \mathbb{E}\left[\min_{j=1,\dots,M} \|\hat{\boldsymbol{\mu}}_j - \mathbf{f}\|_2^2\right] + 60\phi^2\sigma^2 \log(M).$$

The sharp oracle inequality in deviation given in [10] presents an additive term proportional to $\sigma^2 \text{Tr}(A_j)$, as in (1.3). An improvement of the present paper is the absence of this additive term which can be large for matrices A_j with large trace. Our analysis shows that the quantities $\sigma^2 \text{Tr}(A_j)$ are not meaningful for the problem of aggregation of affine estimators, and Theorem 2.1 improves upon the earlier result of [10].

The quantity ϕ defined in (2.7) appears in the right-hand side of the oracle inequalities of Theorem 2.1. Cohen [9] established that estimators of the form $\hat{\boldsymbol{\mu}}_j = A_j \mathbf{y}$ with $\|A_j\|_2 > 1$ are inadmissible. Thus, if $\hat{\boldsymbol{\mu}}_1, \dots, \hat{\boldsymbol{\mu}}_M$ are admissible affine estimators, then $\|A_j\|_2 \leq 1$ and the quantity (2.7) is equal to 1.

We relax all assumptions on the matrices A_1, \dots, A_M , for instance, they may be nonsymmetric and have negative eigenvalues. The above result shows that the restrictions on the matrices A_1, \dots, A_M introduced in [10, 12, 27] are not intrinsic to the problem of aggregation of affine estimators.

The next proposition shows that the bounds of Theorem 2.1 are optimal in a minimax sense. For any $\mathbf{f} \in \mathbf{R}^n$, we denote by $\mathbb{P}_{\mathbf{f}}$ the probability measure of the random variable $\mathbf{y} = \mathbf{f} + \boldsymbol{\xi}$. A lower bound for aggregation of deterministic vectors was proved in [32], Theorem 5.4 with $S = 1$. This lower bound implies the following result.

PROPOSITION 2.1. *There exist absolute constants $c^*, C^*, p^* > 0$ such that the following holds. For all $M, n \geq C^*$, there exist $\mathbf{b}_1, \dots, \mathbf{b}_M \in \mathbf{R}^n$ and orthoprojectors A_1, \dots, A_M of rank one such that*

$$(2.10) \quad \inf_{\hat{\boldsymbol{\mu}}} \sup_{\mathbf{f} \in \mathbf{R}^n} \mathbb{P}_{\mathbf{f}}\left(\|\hat{\boldsymbol{\mu}} - \mathbf{f}\|_2^2 - \min_{k=1,\dots,M} \|\mathbf{b}_k - \mathbf{f}\|_2^2 \geq c^* \sigma^2 \log(M)\right) \geq p^*,$$

$$(2.11) \quad \inf_{\hat{\boldsymbol{\mu}}} \sup_{\mathbf{f} \in \mathbf{R}^n} \mathbb{P}_{\mathbf{f}}\left(\|\hat{\boldsymbol{\mu}} - \mathbf{f}\|_2^2 - \min_{k=1,\dots,M} \|A_k \mathbf{y} - \mathbf{f}\|_2^2 \geq c^* \sigma^2 \log(M)\right) \geq p^*,$$

where the infima are taken over all estimators $\hat{\boldsymbol{\mu}}$.

This implies that the bounds of Theorem 2.1 are rate minimax in terms of the aggregation price. The lower bound can be constructed either with a dictionary of deterministic vectors [cf. (2.10)], or with a dictionary of orthoprojectors of rank one [cf. (2.11)].

3. The penalty (2.3) improves upon model selection based on C_p . In order to explain the role of the penalty (2.3) for the problem of aggregation of affine

estimators, consider first the standard empirical risk minimization scheme based on the C_p criterion. Define \hat{J} as

$$(3.1) \quad \hat{J} \in \underset{j=1, \dots, M}{\operatorname{argmin}} C_p(\mathbf{e}_j),$$

where $C_p(\cdot)$ is defined in (2.1). Using that $C_p(\mathbf{e}_j) \leq C_p(\mathbf{e}_k)$ for all $k = 1, \dots, M$ together with the definition of $C_p(\cdot)$ and $R(\cdot)$ given in (2.1) and (2.5), the following holds almost surely:

$$(3.2) \quad \|\hat{\boldsymbol{\mu}}_{\hat{J}} - \mathbf{f}\|_2^2 \leq \min_{k=1, \dots, M} \|\hat{\boldsymbol{\mu}}_k - \mathbf{f}\|_2^2 + \max_{j, k=1, \dots, M} \Delta_{jk},$$

where $\Delta_{jk} := C_p(\mathbf{e}_k) - C_p(\mathbf{e}_j) - (R(\mathbf{e}_k) - R(\mathbf{e}_j))$. Thus, it is possible to prove an oracle inequality for the estimator $\hat{\boldsymbol{\mu}}_{\hat{J}}$ if we can control the quantities Δ_{jk} uniformly over all pairs $j, k = 1, \dots, M$. These quantities can be rewritten as

$$(3.3) \quad \begin{aligned} \Delta_{jk} &= 2\boldsymbol{\xi}^T ((A_j - A_k)\mathbf{f} + \mathbf{b}_j - \mathbf{b}_k) \\ &\quad + 2(\boldsymbol{\xi}^T (A_j - A_k)\boldsymbol{\xi} - \sigma^2 \operatorname{Tr}(A_j - A_k)). \end{aligned}$$

Two stochastic terms appear in Δ_{jk} . The first is a centered Gaussian random variable with variance $4\sigma^2 \|(A_j - A_k)\mathbf{f} + \mathbf{b}_j - \mathbf{b}_k\|_2^2$. The second is a centered quadratic form in $\boldsymbol{\xi}$, and it can be shown that its variance is of order $\sigma^4 \|A_j - A_k\|_{\mathbb{F}}^2$. This quadratic term is sometimes called a Gaussian chaos of order 2. The deviations of these two terms are governed by the following concentration inequalities. For any vector $\mathbf{v} \in \mathbf{R}^n$, a standard Gaussian tail bound gives

$$(3.4) \quad \mathbb{P}(\mathbf{v}^T \boldsymbol{\xi} > \sigma \|\mathbf{v}\|_2 \sqrt{2x}) \leq \exp(-x) \quad \forall x > 0.$$

For the Gaussian chaos of order 2, the following is proved in [7], Example 2.12.

LEMMA 3.1. *Assume that $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 I_{n \times n})$. For any squared matrix B of size n ,*

$$(3.5) \quad \mathbb{P}(\boldsymbol{\xi}^T B \boldsymbol{\xi} - \sigma^2 \operatorname{Tr} B > 2\sigma^2 \|B\|_{\mathbb{F}} \sqrt{x} + 2\sigma^2 \|B\|_2 x) \leq \exp(-x),$$

where $\sigma^2 \operatorname{Tr} B = \mathbb{E}[\boldsymbol{\xi}^T B \boldsymbol{\xi}]$.

We set $\mathbf{v} = 2((A_j - A_k)\mathbf{f} + \mathbf{b}_j - \mathbf{b}_k)$ and $B = 2(A_k - A_j)$ to study the deviations of the random variable Δ_{jk} . If $\|A_j - A_k\|_2$ is small, (3.4) and (3.5) yield that the deviations of Δ_{jk} are of order of the two quantities

$$(3.6) \quad \sigma \|(A_j - A_k)\mathbf{f} + \mathbf{b}_j - \mathbf{b}_k\|_2, \quad \sigma^2 \|A_j - A_k\|_{\mathbb{F}},$$

that is, the standard deviations of the two terms in Δ_{jk} . The concentration inequalities (3.4) and (3.5) are known to be tight [23], thus there is little hope to bound the deviations of Δ_{jk} independently of \mathbf{f} , A_j and A_k in order to prove a sharp oracle inequality. It is possible to refine the above analysis and to prove the following oracle inequality, though with a leading constant greater than 1.

PROPOSITION 3.1. *There exist absolute constants $c, C > 0$ such that the following holds. Let $0 < \varepsilon < c$ and let \hat{J} be the estimator defined in (3.1). For all $x > 0$, the estimator $\hat{\boldsymbol{\mu}}_{\hat{J}}$ satisfies with probability greater than $1 - 2 \exp(-x)$:*

$$(3.7) \quad \|\hat{\boldsymbol{\mu}}_{\hat{J}} - \mathbf{f}\|_2^2 \leq (1 + \varepsilon) \min_{k=1, \dots, M} \|\hat{\boldsymbol{\mu}}_k - \mathbf{f}\|_2^2 + C\phi^2\sigma^2(x + 2 \log M)/\varepsilon,$$

where ϕ is defined in (2.7).

The proof of Proposition 3.1 is given in Section 8.6. The estimator $\hat{\boldsymbol{\mu}}_{\hat{J}}$ fails to achieve an oracle inequality with leading constant 1 [the leading constant in (3.7) is $1 + \varepsilon$] and with an error term of order $\sigma^2 \log M$. This drawback cannot be repaired for all procedures of the form $\hat{\boldsymbol{\mu}}_{\hat{K}}$ where \hat{K} is an estimator valued in $\{1, \dots, M\}$. Indeed, it is proved in [16], Section 6.4.2 and Proposition 6.1, that there exist $\mathbf{f}_1, \mathbf{f}_2 \in \mathbf{R}^n$ and orthoprojectors A_1, A_2 such that for any estimator \hat{K} valued in $\{1, 2\}$

$$\sup_{\mathbf{f} \in \{\mathbf{f}_1, \mathbf{f}_2\}} \left(\mathbb{E} \|A_{\hat{K}} \mathbf{y} - \mathbf{f}\|_2^2 - \min_{j=1,2} \mathbb{E} \|A_j \mathbf{y} - \mathbf{f}\|_2^2 \right) \geq \sigma^2 \sqrt{n}/4,$$

provided that n is larger than some absolute constant. Inspection of the proof of this result reveals that

$$\sigma \|(A_2 - A_1)\mathbf{f} + \mathbf{b}_2 - \mathbf{b}_1\|_2 \geq \sigma^2 \sqrt{n} \quad \forall \mathbf{f} \in \{\mathbf{f}_1, \mathbf{f}_2\},$$

where we set $\mathbf{b}_1 = \mathbf{b}_2 = 0$. Thus, this lower bound of order \sqrt{n} is related to the Gaussian component of the random variable Δ_{12} , that is, to the term $\boldsymbol{\xi}^T ((A_1 - A_2)\mathbf{f} + \mathbf{b}_1 - \mathbf{b}_2)$; cf. (3.3).

The procedure $\hat{\boldsymbol{\mu}}_{\hat{J}}$ fails to achieve a sharp oracle inequality because the variances of the two components of Δ_{jk} may be large and cannot be controlled. The role of the penalty (2.3) is exactly to control the deviations of Δ_{jk} by controlling the terms (3.6). The following proposition makes this precise.

PROPOSITION 3.2. *Let $\hat{\boldsymbol{\theta}}_{\text{pen}}$ be the estimator (2.4). Then almost surely,*

$$(3.8) \quad \|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}} - \mathbf{f}\|_2^2 \leq \min_{q=1, \dots, M} (\|\hat{\boldsymbol{\mu}}_q - \mathbf{f}\|_2^2) + \max_{j, k=1, \dots, M} \left(\Delta_{jk} - \frac{1}{2} \|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k\|_2^2 \right),$$

where Δ_{jk} is the quantity (3.3). Furthermore, for all $j, k = 1, \dots, M$,

$$(3.9) \quad \mathbb{E} \left[\frac{1}{2} \|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k\|_2^2 \right] = \frac{1}{2} \|(A_j - A_k)\mathbf{f} + \mathbf{b}_j - \mathbf{b}_k\|_2^2 + \frac{\sigma^2}{2} \|A_j - A_k\|_{\mathbb{F}}^2.$$

The proof of (3.8) is given in Section 4 below. A bias-variance decomposition directly yields (3.9), since $\mathbb{E}[\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k] = (A_j - A_k)\mathbf{f} + \mathbf{b}_j - \mathbf{b}_k$ and $\mathbb{E} \|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k\|_2^2 - \mathbb{E}[\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k] \|\mathbf{f}\|_2^2 = \mathbb{E} \|(A_j - A_k)\boldsymbol{\xi}\|_2^2 = \sigma^2 \|A_j - A_k\|_{\mathbb{F}}^2$.

Compared with (3.2), the right-hand side of (3.8) presents the quantities $-\frac{1}{2}\|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k\|_2^2$. We will explain below that these quantities appear because of the interplay between the penalty (2.3) and the strong convexity of H_{pen} .

From (3.8), an outline of the proof of Theorem 2.1 is as follows. By combining the simple inequality (8.5) and Proposition 8.1 below, we will prove that for any pair (j, k) we have

$$\mathbb{E} \exp\left(\lambda_0 \left(\Delta_{jk} - \frac{1}{2}\|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k\|_2^2\right)\right) \leq 1$$

for $\lambda_0 = (30\phi^2\sigma^2)^{-1}$ if the noise $\boldsymbol{\xi}$ has distribution $\mathcal{N}(0, \sigma^2 I_{n \times n})$. Thus, one has

$$\mathbb{E} \exp\left(\lambda_0 \max_{j,k=1,\dots,M} \left(\Delta_{jk} - \frac{1}{2}\|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k\|_2^2\right)\right) \leq M^2.$$

Then Jensen's inequality yields (2.9) while a Chernoff bound yields (2.8). This explains the success of the penalty (2.3) for the problem of model selection type aggregation: the penalty and the strong convexity of H_{pen} provide the quantity $-\frac{1}{2}\|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k\|_2^2$, and this quantity is exactly what is needed to control the deviations of the random variable Δ_{jk} .

4. Strong convexity and the penalty (2.3). To further understand the interplay between the penalty (2.3) and the strong convexity of H_{pen} , we now give the proof of (3.8).

PROOF OF (3.8). Let $k = 1, \dots, M$ be fixed. The simplex Λ^M is a convex set and the function H_{pen} is convex, hence we have

$$\nabla H_{\text{pen}}(\hat{\boldsymbol{\theta}}_{\text{pen}})^T (\mathbf{e}_k - \hat{\boldsymbol{\theta}}_{\text{pen}}) \geq 0;$$

cf. [8], Section 4.2.3, equation (4.21). Inequality (3.8) follows from

$$\begin{aligned} & \|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}} - \mathbf{f}\|_2^2 - \|\hat{\boldsymbol{\mu}}_k - \mathbf{f}\|_2^2 \\ (4.1) \quad & \leq \|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}} - \mathbf{f}\|_2^2 - \|\hat{\boldsymbol{\mu}}_k - \mathbf{f}\|_2^2 + \nabla H_{\text{pen}}(\hat{\boldsymbol{\theta}}_{\text{pen}})^T (\mathbf{e}_k - \hat{\boldsymbol{\theta}}_{\text{pen}}), \end{aligned}$$

$$(4.2) \quad = \sum_{j=1}^M \hat{\theta}_{\text{pen},j} \left(\Delta_{jk} - \frac{1}{2}\|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k\|_2^2 \right),$$

$$(4.3) \quad \leq \max_{j=1,\dots,M} \left(\Delta_{jk} - \frac{1}{2}\|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k\|_2^2 \right).$$

Equality (4.2) is obtained by simple algebra while (4.3) is a consequence of $\sum_{j=1}^M \hat{\theta}_{\text{pen},j} = 1$ and $\hat{\theta}_{\text{pen},j} \geq 0$ for all $j = 1, \dots, M$. \square

It is possible to interpret this argument in light of the interplay between strong convexity and the penalty (2.3). The right-hand side of (4.1) satisfies

$$\begin{aligned} & \|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}} - \mathbf{f}\|_2^2 - \|\hat{\boldsymbol{\mu}}_k - \mathbf{f}\|_2^2 + \nabla H_{\text{pen}}(\hat{\boldsymbol{\theta}}_{\text{pen}})^T (\mathbf{e}_k - \hat{\boldsymbol{\theta}}_{\text{pen}}) \\ &= \sum_{j=1}^M \hat{\theta}_{\text{pen},j} \Delta_{jk} - \frac{1}{2} [\text{pen}(\hat{\boldsymbol{\theta}}_{\text{pen}}) + \|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}} - \hat{\boldsymbol{\mu}}_k\|_2^2]. \end{aligned}$$

The term $\|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}} - \hat{\boldsymbol{\mu}}_k\|_2^2$ comes from the strong convexity of the function H_{pen} . By simple algebra or using (8.10) with $\mathbf{g} = \hat{\boldsymbol{\mu}}_k$, we have

$$(4.4) \quad \text{pen}(\hat{\boldsymbol{\theta}}_{\text{pen}}) + \underbrace{\|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}} - \hat{\boldsymbol{\mu}}_k\|_2^2}_{\text{Term given by the strong convexity of } H_{\text{pen}}} = \sum_{j=1}^M \hat{\theta}_{\text{pen},j} \underbrace{\|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k\|_2^2}_{\text{Term that controls the deviations of } \Delta_{jk}}$$

Formula (4.4) highlights a feature of the penalty (2.3): the penalty transforms the quadratic term given by strong convexity into the linear term given by the right-hand side of (4.4).

The strong convexity of $C_p(\cdot)$ and $H_{\text{pen}}(\cdot)$ is understood with respect to the pseudometric

$$\|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}'}\|_2, \quad \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbf{R}^M,$$

so it is not the strong convexity in the Euclidean norm. We say that a function $V(\cdot)$ is strongly convex with coefficient $\gamma > 0$ over the simplex if for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Lambda^M$,

$$V(\boldsymbol{\theta}) \geq V(\boldsymbol{\theta}') + \nabla V(\boldsymbol{\theta}')^T (\boldsymbol{\theta} - \boldsymbol{\theta}') + \gamma \|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}'}\|_2^2.$$

The strong convexity of H_{pen} could be used because H_{pen} is minimized over the simplex and not just over the vertices. Indeed, minimizing a strongly convex function over a discrete set, as in the definition of \hat{J} , only grants the inequalities

$$C_p(\mathbf{e}_j) \leq C_p(\mathbf{e}_k) \quad \text{for all } k = 1, \dots, M.$$

Because the simplex is a convex set, minimizing the strongly convex function H_{pen} over the simplex grants the inequalities

$$H_{\text{pen}}(\hat{\boldsymbol{\theta}}_{\text{pen}}) \leq H_{\text{pen}}(\boldsymbol{\theta}) - \frac{1}{2} \|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}}\|_2^2 \quad \text{for all } \boldsymbol{\theta} \in \Lambda^M.$$

One could also consider the estimator $\hat{\boldsymbol{\theta}}_C \in \arg\min_{\boldsymbol{\theta} \in \Lambda^M} C_p(\boldsymbol{\theta})$. Because of the strong convexity of $C_p(\cdot)$, this estimator enjoys the inequalities:

$$C_p(\hat{\boldsymbol{\theta}}_C) \leq C_p(\boldsymbol{\theta}) - \|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}}\|_2^2 \quad \text{for all } \boldsymbol{\theta} \in \Lambda^M.$$

The above displays highlight the fact that $C_p(\cdot)$ and $H_{\text{pen}}(\cdot)$ have different strong convexity coefficients. This is because $H_{\text{pen}}(\cdot) = C_p(\cdot) + (1/2)\text{pen}(\cdot)$ and

(1/2) $\text{pen}(\cdot)$ is strongly concave with coefficient 1/2, thus the strong convexity coefficient of $H_{\text{pen}}(\cdot)$ is less than that of $C_p(\cdot)$. We refer to Lemma 8.2 for a rigorous proof of the strong convexity of H_{pen} and C_p .

The estimator $\hat{\boldsymbol{\mu}}_{\hat{\theta}_C}$ is another candidate for the problem of aggregation of affine estimators. It is close to the estimator $\hat{\boldsymbol{\mu}}_{\hat{\theta}_{\text{pen}}}$, except that the penalty (2.3) has been removed from the function to minimize. It was proved in [11], Section 2.2, that the estimator $\hat{\boldsymbol{\mu}}_{\hat{\theta}_C}$ performs poorly: for large enough M and n , there exist \mathbf{f} and $\mathbf{b}_1, \dots, \mathbf{b}_M \in \mathbf{R}^n$ such that with probability greater than 1/4,

$$\|\hat{\boldsymbol{\mu}}_{\hat{\theta}_C} - \mathbf{f}\|_2^2 \geq \min_{j=1, \dots, M} \|\hat{\boldsymbol{\mu}}_j - \mathbf{f}\|_2^2 + \frac{\sigma^2 \sqrt{n}}{48},$$

where $\hat{\boldsymbol{\mu}}_j = \mathbf{b}_j$ for all $j = 1, \dots, M$.

5. Prior weights. We consider now the problem of aggregation of M affine estimators with a prior probability distribution $\boldsymbol{\pi} = (\pi_1, \dots, \pi_M)^T$ on the finite set of indices $\{1, \dots, M\}$.

THEOREM 5.1. *Let $M \geq 2$. For $j = 1, \dots, M$, consider the estimator $\hat{\boldsymbol{\mu}}_j = A_j \mathbf{y} + \mathbf{b}_j$ and let ϕ be defined in (2.7). Let $\boldsymbol{\pi} = (\pi_1, \dots, \pi_M)^T \in \Lambda^M$. Assume that the noise $\boldsymbol{\xi}$ has distribution $\mathcal{N}(0, \sigma^2 I_{n \times n})$. Let $\hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}} \in \arg\min_{\boldsymbol{\theta} \in \Lambda^M} V_{\text{pen}}(\boldsymbol{\theta})$ where*

$$(5.1) \quad V_{\text{pen}}(\boldsymbol{\theta}) := H_{\text{pen}}(\boldsymbol{\theta}) + 30\phi^2\sigma^2 \sum_{j=1}^M \theta_j \log \frac{1}{\pi_j}.$$

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

$$(5.2) \quad \|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}}} - \mathbf{f}\|_2^2 \leq \min_{j=1, \dots, M} \left(\|\hat{\boldsymbol{\mu}}_j - \mathbf{f}\|_2^2 + 60\phi^2\sigma^2 \log \frac{1}{\pi_j} \right) + 30\phi^2\sigma^2 x.$$

Furthermore,

$$(5.3) \quad \mathbb{E} \|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}}} - \mathbf{f}\|_2^2 \leq \mathbb{E} \min_{j=1, \dots, M} \left(\|\hat{\boldsymbol{\mu}}_j - \mathbf{f}\|_2^2 + 60\phi^2\sigma^2 \log \frac{1}{\pi_j} \right).$$

The prior probability distribution $\boldsymbol{\pi} = (\pi_j)_{j=1, \dots, M}$ is deterministic and does not depend on the data $\mathbf{y} = (Y_1, \dots, Y_n)^T$. The only difference between the function (2.2) and the function minimized in (5.1) is the term proportional to

$$(5.4) \quad \phi^2\sigma^2 \sum_{j=1}^M \theta_j \log \frac{1}{\pi_j}.$$

This term allows us to weight the candidates $\hat{\boldsymbol{\mu}}_1, \dots, \hat{\boldsymbol{\mu}}_M$ with the prior probability distribution $(\pi_j)_{j=1, \dots, M}$ based on some prior knowledge about the estimators $\hat{\boldsymbol{\mu}}_1, \dots, \hat{\boldsymbol{\mu}}_M$. For example, if the estimators are orthoprojectors, one can set prior

weights that decrease with the rank the orthoprojectors [32, 34]. The same term is used in [26] whereas [10] uses the Kullback–Leibler divergence of $\boldsymbol{\theta}$ from $\boldsymbol{\pi}$. It is shown in [11] that for aggregation of deterministic vectors, one may use a quantity of the form $\sum_{j=1}^M \theta_j \log(\rho(\theta_j)/\pi_j)$ where $\rho(\cdot)$ satisfies $\rho(t) \geq t$ and $t \rightarrow t \log(\rho(t))$ is convex. This suggests that we could use the Kullback–Leibler divergence of $\boldsymbol{\theta}$ from $\boldsymbol{\pi}$ instead of (5.4), but in their current form our proofs only hold with the “linear entropy” (5.4).

6. Robustness of the estimator $\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}}$. We prove in this section that the procedure (2.4) is robust to non-Gaussian noise distributions and to variance misspecification.

6.1. *Robustness to non-Gaussian noise.* The following result shows that the penalized procedure (2.4) is robust to non-Gaussian noise distributions.

THEOREM 6.1. *Let $M \geq 2$. Let $\bar{\sigma} > 0$. For $j = 1, \dots, M$, consider the estimator $\hat{\boldsymbol{\mu}}_j = A_j \mathbf{y} + \mathbf{b}_j$ and assume that $\|A_j\|_2 \leq 1$ for all $j = 1, \dots, M$. Assume that the noise components ξ_1, \dots, ξ_n are i.i.d., centered with variance σ^2 and satisfy for all $\mathbf{b} \in \mathbf{R}^n$, all matrices B and all $x > 0$:*

$$(6.1) \quad \mathbb{P}(\boldsymbol{\xi}^T \mathbf{b} > \bar{\sigma} \sqrt{2x}) \leq \exp(-x),$$

$$(6.2) \quad \mathbb{P}(\boldsymbol{\xi}^T B \boldsymbol{\xi} - \sigma^2 \text{Tr} B > 2\sigma \bar{\sigma} \|B\|_F \sqrt{x} + 2\bar{\sigma}^2 \|B\|_2 x) \leq \exp(-x).$$

Let $\hat{\boldsymbol{\theta}}_{\text{pen}}$ be the estimator defined in (2.4). Then for all $x > 0$, the estimator $\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}}$ satisfies with probability greater than $1 - 2\exp(-x)$,

$$(6.3) \quad \|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}} - \mathbf{f}\|_2^2 \leq \min_{j=1, \dots, M} \|\hat{\boldsymbol{\mu}}_j - \mathbf{f}\|_2^2 + 46\bar{\sigma}^2 (2 \log M + x).$$

Let $K > 0$. If the random variables ξ_1, \dots, ξ_n are i.i.d., centered with variance σ^2 and K -sub-Gaussian in the sense that $\log \mathbb{E}[e^{t\xi_i}] \leq K^2 t^2 / 2$ for all $t \in \mathbf{R}$ and all $i = 1, \dots, n$, then (6.1) is satisfied with $\bar{\sigma} = cK$ for some absolute constant $c > 0$ [40], Section 5.2.3. As $\sigma \leq K$, (6.1) is also satisfied with $\bar{\sigma} = cK^2/\sigma$. By the Hanson–Wright inequality [20, 35, 41], (6.2) also holds with $\bar{\sigma} = cK^2/\sigma$ for another absolute constant $c > 0$. Thus, for i.i.d. K -sub-Gaussian random variables with variance σ^2 , (6.3) yields

$$(6.4) \quad \|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}} - \mathbf{f}\|_2^2 \leq \min_{j=1, \dots, M} \|\hat{\boldsymbol{\mu}}_j - \mathbf{f}\|_2^2 + C(K^4/\sigma^2)(2 \log M + x),$$

for some absolute constant $C > 0$. For most common examples of sub-Gaussian random variables, the standard deviation σ is of the same order as the sub-Gaussian norm K , so the bound (6.4) is satisfying. This bound may not be tight if the standard deviation is pathologically small compared to the sub-Gaussian norm.

6.2. *Robustness to variance misspecification.* In order to construct the estimator (2.4) by minimizing (2.2), the knowledge of the variance of the noise is needed. However, the following proposition shows that the procedure (2.4) is robust to variance misspecification, that is, the result still holds if the variance is replaced by an estimator $\hat{\sigma}^2$ as soon as $\hat{\sigma}^2$ is consistent in a weak sense defined below.

THEOREM 6.2 (Aggregation under variance misspecification). *Let $M \geq 2$. For $j = 1, \dots, M$, consider the estimator $\hat{\boldsymbol{\mu}}_j = A_j \mathbf{y} + \mathbf{b}_j$. Assume that the noise random variables ξ_1, \dots, ξ_n are i.i.d. $\mathcal{N}(0, \sigma^2)$. Let $\hat{\sigma}^2$ be an estimator and assume that*

$$(6.5) \quad \forall j = 1, \dots, M, \quad A_j = A_j^T = A_j^2, \quad \delta := \mathbb{P}(|\sigma^2 - \hat{\sigma}^2| > \sigma^2/8) < 1.$$

Let $\hat{\boldsymbol{\theta}}_{\hat{\sigma}} = \operatorname{argmin}_{\boldsymbol{\theta} \in \Lambda^M} W_{\text{pen}}(\boldsymbol{\theta})$ where

$$(6.6) \quad W_{\text{pen}}(\boldsymbol{\theta}) := \|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}\|_2^2 - 2\mathbf{y}^T \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} + 2\hat{\sigma}^2 \operatorname{Tr}(A_{\boldsymbol{\theta}}) + \frac{1}{2} \operatorname{pen}(\boldsymbol{\theta}).$$

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

$$\|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\hat{\sigma}}} - \mathbf{f}\|_2^2 \leq \min_{j=1, \dots, M} \|\hat{\boldsymbol{\mu}}_j - \mathbf{f}\|_2^2 + 48\sigma^2(x + 2 \log M).$$

The proof of Theorem 6.2 is given in Section 8.3. In (6.5), the matrices A_1, \dots, A_M are assumed to be orthoprojectors, so Theorem 6.2 is a result for aggregation of least squares estimators. As soon as an estimator $\hat{\sigma}^2$ satisfies with high probability $|\hat{\sigma}^2 - \sigma^2| \leq \sigma^2/8$, optimal aggregation of least squares estimators is possible. This condition is weaker than consistency, as any estimator $\hat{\sigma}^2$ that converges to σ^2 in probability satisfies this condition for n large enough.

The proof of Theorem 6.2 exploits the form of the penalty (2.3) and the strong convexity of the function (6.6). Similar to Proposition 3.2, we will prove that almost surely

$$(6.7) \quad \begin{aligned} \|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\hat{\sigma}}} - \mathbf{f}\|_2^2 &\leq \min_{q=1, \dots, M} \|\hat{\boldsymbol{\mu}}_q - \mathbf{f}\|_2^2 \\ &+ \max_{j, k=1, \dots, M} \left(\Delta_{jk} - \frac{1}{2} \|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k\|_2^2 \right. \\ &\left. + 2(\sigma^2 - \hat{\sigma}^2) \operatorname{Tr}(A_j - A_k) \right), \end{aligned}$$

where Δ_{jk} is the quantity (3.3). The only difference from (3.8) is in the extra term $2(\sigma^2 - \hat{\sigma}^2) \operatorname{Tr}(A_j - A_k)$ that appears because we used $\hat{\sigma}^2$ instead of σ^2 in the definition of $W_{\text{pen}}(\cdot)$. On the event $|\hat{\sigma}^2 - \sigma^2| \leq \sigma^2/8$, it is easy to check that (cf. Lemma 8.1)

$$2(\sigma^2 - \hat{\sigma}^2) \operatorname{Tr}(A_j - A_k) \leq \frac{\sigma^2}{4} \|A_j - A_k\|_{\mathbb{F}}^2.$$

As explained in the discussion that follows Proposition 3.2, the quantity $\frac{1}{2}\|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k\|_2^2$ is given by the interplay between the penalty (2.3) and the strong convexity of the function that is minimized. By (3.9), the expectation of this quantity is greater than $(\sigma^2/2)\|A_j - A_k\|_F^2$. Thus, the penalty (2.3) and the strong convexity of W_{pen} provide exactly what is needed to compensate the difference between $\hat{\sigma}^2$ and σ^2 . Hence, the proof of Theorem 6.2 reveals that the robustness to variance misspecification is in fact due to the interplay between the penalty (2.3) and the strong convexity of W_{pen} .

The papers [3, 17, 18] aim at performing aggregation of least squares estimators when σ^2 is unknown, but unlike Theorem 6.2 the oracle inequalities that they established have a leading constant greater than 1. To our knowledge, Theorem 6.2 is the first aggregation result, with leading constant 1, that is robust to variance misspecification.

In the following, we describe several situations where the suitable estimator $\hat{\sigma}^2$ is available.

EXAMPLE 6.1 (An estimator $\hat{\sigma}^2$ that does not depend on \mathbf{y}). In [12], Section 3.1, two contexts are given where an unbiased estimator of the covariance matrix, independent from \mathbf{y} , is available. For example, the noise level can be estimated independently if the signal is captured multiple times by a single device, or if several identical devices capture the same signal.

EXAMPLE 6.2 (Difference based estimators). In nonparametric regression where the nonrandom design points are equi-spaced in $[0, 1]$, a well-known estimator of the noise level is the difference based estimator $1/(2n-2)\sum_{i=1}^{n-1}(y_{i+1} - y_i)^2$. This technique can be refined with more complex difference sequences [13, 19], and extends to design points in a multidimensional space [29]. For images, where the underlying space is 2-dimensional, there exist efficient methods which require no multiplication [21].

EXAMPLE 6.3 (Consistent estimation of σ^2 in high-dimensional linear regression). In a high-dimensional setting, it is possible to estimate σ^2 under classical assumptions in high-dimensional regression. First, the scaled Lasso [36] allows a joint estimation of the regression coefficients and of the noise level σ^2 . The estimator $\hat{\sigma}^2$ of the scaled Lasso converges in probability to the true noise level σ^2 [36], Theorem 1, and $\hat{\sigma}^2/\sigma^2$ is asymptotically normal [36], equation (19). Second, [5] proposes to estimate σ^2 with a recursive procedure that uses Lasso residuals, and nonasymptotic guarantees are proved in the supplementary material [5]. Third, [6] provides nonasymptotic bounds on the estimation of σ^2 by the residuals of the square-root Lasso ([6], Theorem 2), and these bounds imply consistency. In Theorem 6.2, we require that $|\hat{\sigma}^2/\sigma^2 - 1| \leq 1/8$ with high-probability and this requirement is far weaker than the guarantees obtained in [5, 36].

7. Examples.

7.1. *Adaptation to the smoothness.* For all $n \geq 1$, given continuous parameters $\beta \geq 1$ and $L > 0$, we consider subsets $\Theta(\beta, L) \subset \mathbf{R}^n$. We assume that for each $\beta \geq 1$, there exists a squared matrix A_β of size n with $\|A_\beta\|_2 \leq 1$ such that for all $L > 0$, as $n \rightarrow +\infty$,

$$(7.1) \quad \inf_{\hat{\mathbf{f}}} \sup_{\mathbf{f} \in \Theta(\beta, L)} \frac{1}{n} \mathbb{E} \|\mathbf{f} - \hat{\mathbf{f}}\|_2^2 \sim \sup_{\mathbf{f} \in \Theta(\beta, L)} \frac{1}{n} \mathbb{E} \|\mathbf{f} - A_\beta \mathbf{y}\|_2^2 \sim C^* n^{-\frac{2\beta}{2\beta+1}},$$

where $a_n \sim b_n$ if and only if $a_n/b_n \rightarrow 1$ as $n \rightarrow +\infty$, the infimum is taken over all estimators and the constant $C^* > 0$ may depend on β, L and σ . The above assumption holds for Sobolev ellipsoids in nonparametric regression, and in this case one can choose the Pinsker filters for the matrices A_β (cf. [38], Theorem 3.2). For Sobolev ellipsoids, there exist different estimators that adapt to the unknown smoothness [12, 15, 38].

Consider the following aggregation procedure. Assume that $n \geq 3$ and let $M = \lceil 120 \log(n)(\log \log n)^2 \rceil$. For all $j = 1, \dots, M$, let

$$\beta_j = (1 + 1/(\log(n) \log \log n))^{j-1}.$$

We aggregate the linear estimators $(\hat{\boldsymbol{\mu}}_j = A_{\beta_j} \mathbf{y})_{j=1, \dots, M}$ using the procedure (2.4) of Theorem 2.1, and denote by $\tilde{\boldsymbol{\mu}}$ the resulting estimator. The following adaptation result is a direct consequence of Theorem 2.1.

PROPOSITION 7.1. *For all $n \geq 3$, $\beta \geq 1$ and $L > 0$, let $\Theta(\beta, L) \subset \mathbf{R}^n$ such that as $n \rightarrow +\infty$, (7.1) is satisfied for some matrices A_β with $\|A_\beta\|_2 \leq 1$. Assume that the sets $\Theta(\beta, L)$ are ordered, that is, $\Theta(\beta, L) \subset \Theta(\beta', L)$ for any $\beta > \beta'$ and any $L > 0$. For all $\beta \geq 1$ and $L > 0$, the estimator $\tilde{\boldsymbol{\mu}}$ defined above satisfies as $n \rightarrow +\infty$*

$$\lim_{n \rightarrow +\infty} \sup_{\mathbf{f} \in \Theta(\beta, L)} \frac{1}{n} \mathbb{E} \|\mathbf{f} - \tilde{\boldsymbol{\mu}}\|_2^2 n^{\frac{2\beta}{2\beta+1}} = C^*.$$

Proposition 7.1 is proved in Section 8.7. The above procedure adapts to the unknown smoothness in exact asymptotic sense by aggregating only

$$\log(n)(\log \log n)^2$$

estimators so its computational complexity is small. Another feature is that the minimax rate and the minimax constant C^* are not altered by the aggregation step.

7.2. *The best convex combination as a benchmark.* We consider convex combinations of the estimators $\hat{\boldsymbol{\mu}}_1, \dots, \hat{\boldsymbol{\mu}}_M$ to construct the estimator (2.4). The goal of this section is to study the performance of the estimator (2.4) if the benchmark is $\min_{\boldsymbol{\theta} \in \Delta^M} \|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \mathbf{f}\|_2^2$ instead of $\min_{k=1, \dots, M} \|\hat{\boldsymbol{\mu}}_k - \mathbf{f}\|_2^2$.

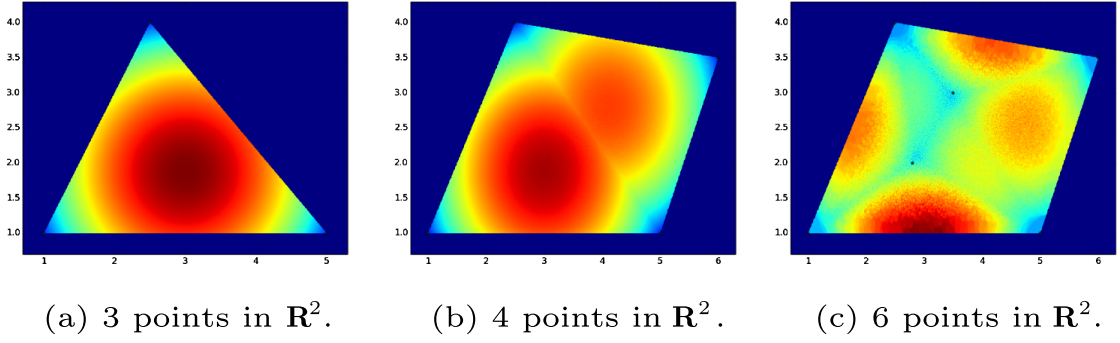


FIG. 1. Penalty (2.3) heatmaps. Largest penalty in red, smallest in blue.

The penalty (2.3) vanishes at the extreme points: $\text{pen}(\mathbf{e}_j) = 0$ for all $j = 1, \dots, M$, and it pushes $\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}}$ towards the points $\{\hat{\boldsymbol{\mu}}_1, \dots, \hat{\boldsymbol{\mu}}_M\}$. This can be seen in Figure 1. Consider a noise-free problem where $\sigma = 0$. Let $\mathbf{f} \in \mathbf{R}^n$ and let $\mathbf{v} \in \mathbf{R}^n$ be such that $\mathbf{v}^T \mathbf{f} = 0$ and $\|\mathbf{v}\|_2^2 = 3\|\mathbf{f}\|_2^2 > 0$. Let also $\rho = 4\|\mathbf{v}\|_2^2$. Consider estimators $\hat{\boldsymbol{\mu}}_1 = 2\mathbf{f}$, $\hat{\boldsymbol{\mu}}_2 = \mathbf{f} + \mathbf{v}$ and $\hat{\boldsymbol{\mu}}_3 = \mathbf{f} - \mathbf{v}$. These estimators are such that $\|\hat{\boldsymbol{\mu}}_j\|_2^2 = \rho > 0$ for all $j = 1, \dots, 3$ (here $M = 3$ and the estimators are deterministic because $\sigma = 0$). Then by simple algebra we have $\text{pen}(\boldsymbol{\theta}) = \rho - \|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}\|_2^2$ and $H_{\text{pen}}(\boldsymbol{\theta}) = (1/2)\|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - 2\mathbf{f}\|_2^2 + c$ where c is constant that depends on \mathbf{f} but not on $\boldsymbol{\theta}$. For this example, although \mathbf{f} lies in the convex hull of $\{\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\mu}}_3\}$, the estimator $\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}}$ defined in (2.4) will be equal to $2\mathbf{f}$ instead of \mathbf{f} and is likely to be a bad procedure with respect to the benchmark $\min_{\boldsymbol{\theta} \in \Lambda^M} \|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \mathbf{f}\|_2^2$. This fact is not surprising since the penalty penalizes heavily some regions of the convex hull of the estimators. Furthermore, this procedure is tailored for the benchmark $\min_{k=1, \dots, M} \|\hat{\boldsymbol{\mu}}_k - \mathbf{f}\|_2^2$ and its goal is not to mimic the best convex combination of the estimators.

It is possible to modify the procedure (2.4) to construct an estimator that performs well with respect to the best convex combination of M linear estimators. Let

$$(7.2) \quad m := \left\lfloor \sqrt{\frac{n}{\log(1 + M/\sqrt{n})}} \right\rfloor.$$

If $m \geq 1$, define the set $\Lambda_m^M \subset \Lambda^M$ as

$$(7.3) \quad \Lambda_m^M := \left\{ \frac{1}{m} \sum_{q=1}^m \mathbf{u}_q, \mathbf{u}_1, \dots, \mathbf{u}_m \in \{\mathbf{e}_1, \dots, \mathbf{e}_M\} \right\}.$$

Denote by $|\Lambda_m^M|$ the cardinality of Λ_m^M . We aggregate the affine estimators $(\hat{\boldsymbol{\mu}}_{\mathbf{u}})_{\mathbf{u} \in \Lambda_m^M}$ using the procedure (2.4) and denote by $\hat{\boldsymbol{\mu}}_{\Lambda_m^M}$ the resulting estimator.

PROPOSITION 7.2. *Let $M, n \geq 1$. For $j = 1, \dots, M$, consider the estimator $\hat{\boldsymbol{\mu}}_j = A_j \mathbf{y} + \mathbf{b}_j$ for any $n \times n$ matrix A_j and vector $\mathbf{b}_j \in \mathbf{R}^n$. Assume that $\boldsymbol{\xi} \sim$*

$\mathcal{N}(0, \sigma^2 I_{n \times n})$ and that for some constant $R > 0$,

$$\frac{1}{n} \|\mathbf{f}\|_2^2 \leq R^2, \quad \frac{1}{n} \|\mathbf{b}_j\|_2^2 \leq R^2, \quad \|A_j\|_2 \leq 1 \quad \forall j = 1, \dots, M.$$

For all $x > 0$, the estimator $\hat{\boldsymbol{\theta}}_C \in \operatorname{argmin}_{\boldsymbol{\theta} \in \Lambda^M} C_p(\boldsymbol{\theta})$ satisfies with probability greater than $1 - 2 \exp(-x)$,

$$(7.4) \quad \frac{1}{n} \|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_C} - \mathbf{f}\|_2^2 \leq \min_{\boldsymbol{\theta} \in \Lambda^M} \frac{1}{n} \|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \mathbf{f}\|_2^2 + 8(\sigma^2 + \sigma R \sqrt{2}) \sqrt{\frac{x + 2 \log M}{n}} + \frac{8\sigma^2(x + 2 \log M)}{n}.$$

If $M \leq \sqrt{n}(\exp(n) - 1)$, then for all $x > 0$, the estimator $\hat{\boldsymbol{\mu}}_{\Lambda_m^M}$ defined above satisfies with probability greater than $1 - 3 \exp(-x)$,

$$(7.5) \quad \frac{1}{n} \|\hat{\boldsymbol{\mu}}_{\Lambda_m^M} - \mathbf{f}\|_2^2 \leq \min_{\boldsymbol{\theta} \in \Lambda^M} \frac{1}{n} \|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \mathbf{f}\|_2^2 + C \max(R^2, \sigma^2) \sqrt{\frac{\log(1 + M/\sqrt{n})}{n}} + \frac{C\sigma^2 x}{n}.$$

Proposition 7.2 is proved in Section 8.8. To our knowledge, this is the first result that provides a sharp oracle inequality for the problem of aggregation of affine estimators with respect to the convex oracle. However, there is a large literature on convex aggregation when the estimators to aggregate are deterministic, which corresponds to the particular case $A_j = 0$ for all $j = 1, \dots, M$. When the error is measured with the scaled squared norm $\frac{1}{n} \|\cdot\|_2^2$, the minimax rate of convex aggregation is known to be of order M/n if $M \leq \sqrt{n}$ and $\sqrt{\log(1 + M/\sqrt{n})}/n$ if $M > \sqrt{n}$. For our setting, this is proved in [32]. This elbow effect was first established for regression with random design [37] and then extended to other settings in [31, 33]. All these results assume that the estimators to aggregate are deterministic or independent of the data used for aggregation. The lower bound [32], Theorem 5.3 with $S = M$, $\delta = \sigma$ and $R = \log(1 + eM)$, yields that there exist absolute constants $c, C > 0$ such that if $\log(1 + eM)^2 \leq Cn$, there exist deterministic vectors $\hat{\boldsymbol{\mu}}_1 = \mathbf{b}_1, \dots, \hat{\boldsymbol{\mu}}_M = \mathbf{b}_M$ such that for all estimators $\hat{\boldsymbol{\mu}}$,

$$\sup_{\mathbf{f} \in \mathbf{R}^n} \mathbb{P}_{\mathbf{f}} \left(\frac{1}{n} \|\hat{\boldsymbol{\mu}} - \mathbf{f}\|_2^2 - \min_{\boldsymbol{\theta} \in \Lambda^M} \frac{1}{n} \|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \mathbf{f}\|_2^2 \geq c\sigma^2 \left(\frac{M}{n} \wedge \sqrt{\frac{\log(1 + M/\sqrt{n})}{n}} \right) \right) \geq c.$$

Thus, if $M \geq \sqrt{n}$, (7.5) is optimal in a minimax sense up to absolute constants, and (7.4) is optimal up to logarithmic factors. However, we do not know whether the minimax rate is M/n when $M < \sqrt{n}$, as in the case of aggregation of deterministic vectors.

The problem of linear aggregation of affine estimators remains open. It is only known that for linear aggregation of deterministic vectors, the least squares estimator on a linear space of dimension M achieves the rate $\sigma^2 M/n$, which is optimal in a minimax sense [31–33, 37].

A bound similar to (7.4) can be obtained by the Lasso estimator for aggregation of deterministic vectors, that is, in the case $A_j = 0$ and $\hat{\boldsymbol{\mu}}_j = \mathbf{b}_j$ for all $j = 1, \dots, M$. Let $\mathbf{x}_j = (\sqrt{n}/\|\mathbf{b}_j\|_2)\mathbf{b}_j$ for all $j = 1, \dots, M$ and let \mathbb{X} be the $n \times M$ matrix with columns $\mathbf{x}_1, \dots, \mathbf{x}_M$. The Lasso estimator is defined by

$$\hat{\boldsymbol{\beta}} \in \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbf{R}^M} \frac{1}{2n} \|\mathbf{y} - \mathbb{X}\boldsymbol{\beta}\|_2^2 + 4\sigma \sqrt{\log(M)/n} \|\boldsymbol{\beta}\|_1,$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^M |\beta_j|$. This estimator satisfies with probability at least $1 - 1/M$ the oracle inequality

$$\frac{1}{n} \|\mathbf{f} - \mathbb{X}\hat{\boldsymbol{\beta}}\|_2^2 \leq \min_{\boldsymbol{\beta} \in \mathbf{R}^M} \left(\frac{1}{n} \|\mathbf{f} - \mathbb{X}\boldsymbol{\beta}\|_2^2 + c\sigma \|\boldsymbol{\beta}\|_1 \sqrt{\log(M)/n} \right),$$

where $c > 0$ is a numerical constant. This oracle inequality is proved in Theorem 4 in Sun and Zhang [36]. If we assume $\frac{1}{n} \|\mathbf{b}_j\|^2 \leq R^2$ for all $j = 1, \dots, M$ as in Proposition 7.2 above, then the right-hand side of the previous display is bounded from above by

$$\min_{\boldsymbol{\theta} \in \Lambda^M} \left(\frac{1}{n} \|\mathbf{f} - \mathbf{b}_{\boldsymbol{\theta}}\|_2^2 + c\sigma R \sqrt{\log(M)/n} \right).$$

Hence, the Lasso estimator satisfies a bound similar to the oracle inequalities of Proposition 7.2 for aggregation of deterministic vectors, that is, if $A_j = 0$ and $\hat{\boldsymbol{\mu}}_j = \mathbf{b}_j$ for all $j = 1, \dots, M$.

8. Proofs.

8.1. *Preliminaries.* The following notation will be useful. Define for all $j, k = 1, \dots, M$:

$$(8.1) \quad Q_{j,k} := \left(-2I_{n \times n} - \frac{1}{2}(A_k - A_j)^T \right) (A_k - A_j),$$

$$(8.2) \quad \mathbf{v}_{j,k} := (-2I_{n \times n} - (A_k - A_j)^T) ((A_k - A_j)\mathbf{f} + \mathbf{b}_k - \mathbf{b}_j).$$

Let $B_{jk} = A_k - A_j$, so that $\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_j = B_{jk}\boldsymbol{\xi} + (B_{jk}\mathbf{f} + \mathbf{b}_k - \mathbf{b}_j)$. Then

$$\|\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_j\|_2^2 = \|B_{jk}\boldsymbol{\xi}\|_2^2 + \|B_{jk}\mathbf{f} + \mathbf{b}_k - \mathbf{b}_j\|_2^2 + 2\boldsymbol{\xi}^T B_{jk}^T (B_{jk}\mathbf{f} + \mathbf{b}_k - \mathbf{b}_j).$$

Thus, simple algebra yields that the quantity Δ_{jk} defined in (3.3) satisfies

$$(8.3) \quad \begin{aligned} \Delta_{jk} - \frac{1}{2} \|\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_j\|_2^2 &= \boldsymbol{\xi}^T Q_{j,k} \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q_{j,k} \boldsymbol{\xi}] + \boldsymbol{\xi}^T \mathbf{v}_{j,k} \\ &\quad - \frac{\sigma^2}{2} \|A_j - A_k\|_{\mathbb{F}}^2 - \frac{1}{2} \|(A_k - A_j)\mathbf{f} + \mathbf{b}_k - \mathbf{b}_j\|_2^2, \end{aligned}$$

where we used the equality $\sigma^2 \|A_j - A_k\|_F^2 = \mathbb{E}[\|(A_j - A_k)\boldsymbol{\xi}\|_2^2]$ and the above definitions of $Q_{j,k}$ and $\mathbf{v}_{j,k}$. Furthermore, using (1.5) and (2.7) we have

$$(8.4) \quad \begin{aligned} \|\|Q_{j,k}\|\|_2 &\leq 6\phi^2, & \|Q_{j,k}\|_F &\leq 3\phi \|A_k - A_j\|_F, \\ \|\mathbf{v}_{j,k}\|_2 &\leq 4\phi \|(A_k - A_j)\mathbf{f} + \mathbf{b}_k - \mathbf{b}_j\|_2 \end{aligned}$$

for all $j, k = 1, \dots, M$. This yields that

$$(8.5) \quad \begin{aligned} \Delta_{jk} - \frac{1}{2} \|\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_j\|_2^2 &\leq \boldsymbol{\xi}^T Q_{j,k} \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q_{j,k} \boldsymbol{\xi}] + \boldsymbol{\xi}^T \mathbf{v}_{j,k} \\ &\quad - \frac{\sigma^2}{18\phi^2} \|Q_{j,k}\|_F^2 - \frac{1}{32\phi^2} \|\mathbf{v}_{j,k}\|_2^2. \end{aligned}$$

PROPOSITION 8.1. *Let $\mathbf{v} \in \mathbf{R}^n$ and let Q be any squared matrix of size n . Assume that ξ_1, \dots, ξ_n are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables. Then for all $u > 0$ such that $2u\sigma^2 \|\|Q\|\|_2 < 1$ we have*

$$(8.6) \quad \mathbb{E}[e^{u(\boldsymbol{\xi}^T Q \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q \boldsymbol{\xi}] + \boldsymbol{\xi}^T \mathbf{v})}] \leq \exp\left(u^2 \sigma^2 \left(\frac{\sigma^2 \|Q\|_F^2 + \frac{\|\mathbf{v}\|_2^2}{2}}{1 - 2\sigma^2 \|\|Q\|\|_2 u}\right)\right).$$

Furthermore, for some $\phi \geq 1$, define

$$\begin{aligned} Z_{Q,\mathbf{v}} &:= \boldsymbol{\xi}^T Q \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q \boldsymbol{\xi}] + \boldsymbol{\xi}^T \mathbf{v} - \frac{\sigma^2}{18\phi^2} \|Q\|_F^2 - \frac{1}{32\phi^2} \|\mathbf{v}\|_2^2, \\ Y_{Q,\mathbf{v}} &:= \boldsymbol{\xi}^T Q \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q \boldsymbol{\xi}] + \boldsymbol{\xi}^T \mathbf{v} - \frac{\sigma^2}{36\phi^2} \|Q\|_F^2 - \frac{1}{32\phi^2} \|\mathbf{v}\|_2^2. \end{aligned}$$

If $\|\|Q\|\|_2 \leq 6\phi^2$, then for $u = 1/(30\phi^2\sigma^2)$ and $u' = 1/(48\phi^2\sigma^2)$ we have

$$\mathbb{E}[e^{uZ_{Q,\mathbf{v}}}] \leq 1, \quad \mathbb{E}[e^{u'Y_{Q,\mathbf{v}}}] \leq 1.$$

The proof relies on an argument similar to that of [24], Lemma 1.

PROOF. If Q is not symmetric, let $Q_s = (Q + Q^T)/2$. We have $\|Q_s\|_F \leq \|Q\|_F$, $\|\|Q_s\|\|_2 \leq \|\|Q\|\|_2$ and almost surely $\boldsymbol{\xi}^T Q \boldsymbol{\xi} = \boldsymbol{\xi}^T Q_s \boldsymbol{\xi}$ so that if (8.6) holds for Q_s then

$$\mathbb{E}[e^{u(\boldsymbol{\xi}^T Q \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q \boldsymbol{\xi}] + \boldsymbol{\xi}^T \mathbf{v})}] \leq e^{u^2 \sigma^2 \left(\frac{\sigma^2 \|Q_s\|_F^2 + \frac{\|\mathbf{v}\|_2^2}{2}}{1 - 2\sigma^2 \|\|Q_s\|\|_2 u}\right)} \leq e^{u^2 \sigma^2 \left(\frac{\sigma^2 \|Q\|_F^2 + \frac{\|\mathbf{v}\|_2^2}{2}}{1 - 2\sigma^2 \|\|Q\|\|_2 u}\right)}.$$

Thus, the result for the symmetric matrix Q_s implies the result for Q .

We now assume that Q is symmetric. There exists a matrix P with $P^T P = P P^T = I_{n \times n}$ such that $Q = P^T \text{diag}(\lambda_1, \dots, \lambda_n) P$, where $\lambda_1, \dots, \lambda_n$ are the eigenvalues of Q . Let $\mathbf{w} = (1/\sigma) P \mathbf{v}$ and define the random variables g_1, \dots, g_n

by $(g_1, \dots, g_n)^T = (1/\sigma)P\xi$. By the rotational invariance of the Gaussian distribution, g_1, \dots, g_n are i.i.d. $\mathcal{N}(0, 1)$ random variables. Thus, the random variable $\xi^T Q \xi - \mathbb{E}[\xi^T Q \xi] + \xi^T \mathbf{v}$ has the same distribution as

$$\sigma^2 \sum_{i=1}^n W_i \quad \text{where } W_i := \lambda_i (g_i^2 - 1) + g_i w_i.$$

For all $i = 1, \dots, n$ and for all $t > 0$ such that $\max_{i=1, \dots, n} 2t|\lambda_i| < 1$, integration using the probability density function of g_i yields

$$\mathbb{E}[e^{tW_i}] = \frac{1}{\sqrt{1-2\lambda_i t}} e^{\frac{t^2 w_i^2}{2(1-2\lambda_i t)} - t\lambda_i} \leq e^{\frac{\lambda_i^2 t^2}{1-2|\lambda_i|t} + \frac{t^2 w_i^2}{2(1-2\lambda_i t)}},$$

where we used the inequalities

$$\log\left(\frac{1}{\sqrt{1-2v}}\right) \leq v + \frac{v^2}{1-2v} = v + \frac{v^2}{1-2|v|} \quad \text{for all } v \in [0, 1/2),$$

$$\log\left(\frac{1}{\sqrt{1-2v}}\right) \leq v + v^2 \leq v + \frac{v^2}{1-2|v|} \quad \text{for all } v \in (-1/2, 0].$$

This can be shown by comparing the power series expansions. As $|\lambda_i| \leq \|Q\|_2$ for all $i = 1, \dots, n$, by independence of W_1, \dots, W_n we obtain

$$\mathbb{E}[e^{t \sum_{i=1}^n W_i}] \leq \exp\left(t^2 \left(\frac{\|Q\|_F^2 + \frac{\|\mathbf{w}\|_2^2}{2}}{1-2\|Q\|_2 t}\right)\right).$$

By definition of \mathbf{w} , we have $\|\mathbf{v}\|_2 = \sigma \|\mathbf{w}\|_2$, so setting $t = u\sigma^2$ completes the proof of (8.6).

The claims about $Z_{Q,\mathbf{v}}$ and $Y_{Q,\mathbf{v}}$ are direct consequences of (8.6). \square

8.2. Proof of the main results.

PROOF OF THEOREM 2.1. By Proposition 3.2, it is enough to prove that

$$(8.7) \quad D_1 := \mathbb{E}[e^{u \max_{j,k=1, \dots, M} (\Delta_{jk} - \frac{1}{2} \|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k\|_2^2)}] \leq M^2$$

for $u = 1/(30\phi^2\sigma^2)$. Then, Jensen's inequality yields (2.9) and a Chernoff bound yields (2.8).

We now prove (8.7). By (8.4), for all $j, k = 1, \dots, M$ we have $\|Q_{j,k}\|_2 \leq 6\phi^2$. Using (8.5) and Proposition 8.1, we have

$$D_1 \leq \sum_{j=1}^M \sum_{k=1}^M \mathbb{E}[e^{u(\Delta_{jk} - \frac{1}{2} \|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k\|_2^2)}] \leq \sum_{j=1}^M \sum_{k=1}^M \mathbb{E}[e^{uZ_{Q_{j,k}, \mathbf{v}_{j,k}}}] \leq M^2,$$

where for any matrix Q and any $\mathbf{v} \in \mathbf{R}^n$, the random variable $Z_{Q,\mathbf{v}}$ is defined in Proposition 8.1. \square

PROOF OF THEOREM 5.1. Let $\beta = 30\phi^2\sigma^2$. Let $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}}$ for notational simplicity. The only difference between H_{pen} and V_{pen} is the linear term (5.4). As in the proof of (3.8) in Section 4, by convexity of V_{pen} we have that for all $k = 1, \dots, M$,

$$\begin{aligned} & \|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}} - \mathbf{f}\|_2^2 - \|\hat{\boldsymbol{\mu}}_k - \mathbf{f}\|_2^2 \\ & \leq \|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}} - \mathbf{f}\|_2^2 - \|\hat{\boldsymbol{\mu}}_k - \mathbf{f}\|_2^2 + \nabla V_{\text{pen}}(\hat{\boldsymbol{\theta}})^T (\mathbf{e}_k - \hat{\boldsymbol{\theta}}) \\ & = 2\beta \log \frac{1}{\pi_k} + \sum_{j=1}^M \hat{\theta}_j \left(\Delta_{jk} - \frac{1}{2} \|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k\|_2^2 - \beta \log \frac{1}{\pi_j \pi_k} \right) \\ & \leq 2\beta \log \frac{1}{\pi_k} + \max_{j=1, \dots, M} \left(\Delta_{jk} - \frac{1}{2} \|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k\|_2^2 - \beta \log \frac{1}{\pi_j \pi_k} \right), \end{aligned}$$

where Δ_{jk} is defined in (3.3). For all $u > 0$, let

$$D_2 := \mathbb{E} \left[\exp \left(u \max_{j, k=1, \dots, M} \left(\Delta_{jk} - \frac{1}{2} \|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k\|_2^2 - \beta \log \frac{1}{\pi_j \pi_k} \right) \right) \right].$$

We now bound from above this moment generating function using (8.5) and Proposition 8.1. If $u = 1/\beta = 1/(30\phi^2\sigma^2)$, then

$$\begin{aligned} (8.8) \quad D_2 & \leq \sum_{j=1}^M \sum_{k=1}^M \pi_j \pi_k \mathbb{E} \left[e^{u(\Delta_{jk} - \frac{1}{2} \|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k\|_2^2)} \right] \\ & \leq \sum_{j=1}^M \sum_{k=1}^M \pi_j \pi_k \mathbb{E} \left[e^{uZ_{Q_{j,k}, v_{j,k}}} \right] \leq \sum_{j=1}^M \sum_{k=1}^M \pi_j \pi_k = 1. \end{aligned}$$

As in the proof of Theorem 2.1, Jensen's inequality yields (5.2) while a Chernoff bound completes the proof of (5.2). \square

PROOF OF THEOREM 6.1. For a fixed pair (j, k) , we apply (6.1) to the vector $\mathbf{v}_{j,k}$ and (6.2) to the matrix $Q_{j,k}$. Using (8.4),

$$\begin{aligned} \boldsymbol{\xi}^T Q_{j,k} \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q_{j,k} \boldsymbol{\xi}] & \leq \bar{\sigma}^2 12x + 6\sigma \bar{\sigma} \|A_k - A_j\|_{\mathbb{F}} \sqrt{x} \\ & \leq 30\bar{\sigma}^2 x + \frac{\sigma^2}{2} \|A_k - A_j\|_{\mathbb{F}}^2, \\ \boldsymbol{\xi}^T \mathbf{v}_{j,k} & \leq \bar{\sigma} 4 \|(A_k - A_j)\mathbf{f} + \mathbf{b}_k - \mathbf{b}_j\|_2 \sqrt{2x} \\ & \leq 16\bar{\sigma}^2 x + \frac{1}{2} \|(A_k - A_j)\mathbf{f} + \mathbf{b}_k - \mathbf{b}_j\|_2^2. \end{aligned}$$

Combining this bound with (6.7), (8.3) and the union bound completes the proof. \square

8.3. *Proof of Theorem 6.2.* The following inequality will be useful.

LEMMA 8.1 (Projection matrices). *Let A, B be two squared matrices of size n with $A^T = A = A^2$ and $B^T = B = B^2$. Then*

$$(8.9) \quad |\mathrm{Tr}(A - B)| \leq \|A - B\|_{\mathbb{F}}^2.$$

PROOF. Without loss of generality, assume that $\mathrm{Tr} A \geq \mathrm{Tr} B$. As $\|A - B\|_{\mathbb{F}}^2 = \|A\|_{\mathbb{F}}^2 + \|B\|_{\mathbb{F}}^2 - 2\mathrm{Tr}(AB)$ and $\|A\|_{\mathbb{F}}^2 = \mathrm{Tr} A$, (8.9) is equivalent to $2\mathrm{Tr}(AB) \leq 2\mathrm{Tr}(B)$. Notice that for projection matrices, $\mathrm{Tr}(AB) = \|AB\|_{\mathbb{F}}^2 \leq \|A\|_2^2 \|B\|_{\mathbb{F}}^2 \leq \|B\|_{\mathbb{F}}^2 = \mathrm{Tr}(B)$ and the proof is complete. \square

PROOF OF THEOREM 6.2. Let $\hat{\theta} = \hat{\theta}_{\hat{\sigma}}$ for notational simplicity. Since the matrices A_j are orthoprojectors, the quantity ϕ defined in (2.7) is equal to 1. As in the proof of (3.8) in Section 4, by convexity of W_{pen} we have that for all $k = 1, \dots, M$,

$$\begin{aligned} & \|\hat{\mu}_{\hat{\theta}} - \mathbf{f}\|_2^2 - \|\hat{\mu}_k - \mathbf{f}\|_2^2 \\ & \leq \|\hat{\mu}_{\hat{\theta}} - \mathbf{f}\|_2^2 - \|\hat{\mu}_k - \mathbf{f}\|_2^2 + \nabla W_{\mathrm{pen}}(\hat{\theta})^T (\mathbf{e}_k - \hat{\theta}) \\ & = \sum_{j=1}^M \hat{\theta}_j \left(\Delta_{jk} + 2(\hat{\sigma}^2 - \sigma^2) \mathrm{Tr}(A_j - A_k) - \frac{1}{2} \|\hat{\mu}_j - \hat{\mu}_k\|_2^2 \right) \\ & \leq \max_{j=1, \dots, M} \left(\Delta_{jk} + 2(\hat{\sigma}^2 - \sigma^2) \mathrm{Tr}(A_j - A_k) - \frac{1}{2} \|\hat{\mu}_j - \hat{\mu}_k\|_2^2 \right) =: D_3, \end{aligned}$$

where Δ_{jk} is defined in (3.3). The assumption on $\hat{\sigma}^2$ and (8.9) yield that on an event Ω_0 of probability greater than $1 - \delta$,

$$2|(\hat{\sigma}^2 - \sigma^2) \mathrm{Tr}(A_j - A_k)| \leq \frac{\sigma^2}{4} \|A_j - A_k\|_{\mathbb{F}}^2 \quad \text{for all } j, k = 1, \dots, M.$$

Using (8.3) and (8.4), we obtain that on the event Ω_0 ,

$$\begin{aligned} D_3 & \leq \max_{j,k=1, \dots, M} \left(\boldsymbol{\xi}^T Q_{j,k} \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q_{j,k} \boldsymbol{\xi}] + \boldsymbol{\xi}^T \mathbf{v}_{j,k} - \frac{\sigma^2}{36} \|Q_{j,k}\|_{\mathbb{F}}^2 - \frac{1}{32} \|\mathbf{v}_{j,k}\|_2^2 \right) \\ & = \max_{j,k=1, \dots, M} Y_{Q_{j,k}, \mathbf{v}_{j,k}}, \end{aligned}$$

where $Q_{j,k}$ and $\mathbf{v}_{j,k}$ are defined in (8.1) and (8.2) while $Y_{Q,v}$ is defined in Proposition 8.1 for any matrix Q and any $\mathbf{v} \in \mathbf{R}^n$. Using Proposition 8.1, for $u = 1/(48\sigma^2)$ we have

$$\mathbb{E} \left[\exp \left(u \max_{j,k=1, \dots, M} Y_{Q_{j,k}, \mathbf{v}_{j,k}} \right) \right] \leq \sum_{j=1}^M \sum_{k=1}^M \mathbb{E} [\exp(u Y_{Q_{j,k}, \mathbf{v}_{j,k}})] \leq M^2.$$

By a Chernoff bound, this proves that on an event Ω_1 of probability greater than $1 - e^{-x}$, we have $\max_{j,k=1,\dots,M} Y_{Q_{j,k}, v_{j,k}} \leq 48\sigma^2(x + 2 \log M)$. On the event $\Omega_0 \cap \Omega_1$, we have $D_3 \leq 48\sigma^2(x + 2 \log M)$ and the union bound yields that $\mathbb{P}(\Omega_0 \cap \Omega_1) \geq 1 - e^{-x} - \delta$. \square

8.4. *Strong convexity.* The penalty (2.3) satisfies for any $\mathbf{g} \in \mathbf{R}^n$ and any $\boldsymbol{\theta} \in \Lambda^M$:

$$(8.10) \quad \sum_{k=1}^M \theta_k \|\hat{\boldsymbol{\mu}}_k - \mathbf{g}\|_2^2 = \|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \mathbf{g}\|_2^2 + \text{pen}(\boldsymbol{\theta}).$$

This can be shown by using simple properties of the Euclidean norm, or by noting that the equality above is a bias-variance decomposition. For $\mathbf{g} = 0$, (8.10) yields $\text{pen}(\boldsymbol{\theta}) = -\|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}\|_2^2 + \sum_{k=1}^M \theta_k \|\hat{\boldsymbol{\mu}}_k\|_2^2$.

LEMMA 8.2. *Let F be any one of the functions H_{pen} , V_{pen} or W_{pen} defined in (2.2), (5.1) and (6.6), respectively. Then F is convex, differentiable and satisfies for all $\boldsymbol{\theta}, \boldsymbol{\theta}_0 \in \Lambda^M$,*

$$(8.11) \quad F(\boldsymbol{\theta}) = F(\boldsymbol{\theta}_0) + \nabla F(\boldsymbol{\theta}_0)^T (\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \frac{1}{2} \|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}_0}\|_2^2.$$

Furthermore, if $\hat{\boldsymbol{\theta}}$ is a minimizer of F over the simplex then for all $\boldsymbol{\theta} \in \Lambda^M$,

$$(8.12) \quad F(\boldsymbol{\theta}) \geq F(\hat{\boldsymbol{\theta}}) + \frac{1}{2} \|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}}\|_2^2.$$

PROOF. Using (8.10) with $\mathbf{g} = 0$, we obtain that the function F is a polynomial of degree 2, of the form $F(\boldsymbol{\theta}) = \text{affine}(\boldsymbol{\theta}) + \frac{1}{2} \|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}\|_2^2$ where $\text{affine}(\cdot)$ is an affine function of $\boldsymbol{\theta}$. This shows that F is convex and differentiable. The result (8.11) follows by uniqueness of the Taylor expansion of F [or by an explicit calculation of $\nabla F(\boldsymbol{\theta}_0)$]. Inequality (8.12) is a consequence of [8], 4.2.3, equation (4.21). \square

8.5. Lower bound.

PROOF OF PROPOSITION 2.1. The lower bounds of [32], Theorem 5.4, are stated in expectation, but inspection of the proof of [32], Theorem 5.3 with $S = 1$, $\delta = \infty$ and $R = \log(1 + eM)$, reveals that the lower bound holds also in probability since it is an application of [38], Theorem 2.7. This result yields that there exist absolute constants $p, c, C > 0$ and $\mathbf{f}_1, \dots, \mathbf{f}_M \in \mathbf{R}^n$ such that for any estimator $\hat{\boldsymbol{\mu}}$,

$$\sup_{j=1,\dots,M} \mathbb{P}_{\mathbf{f}_j}(\Omega_j) \geq p, \quad \Omega_j := \{\|\hat{\boldsymbol{\mu}} - \mathbf{f}_j\|_2^2 \geq c\sigma^2 \log(M)\},$$

provided that $\log(M) \leq cn$ and $n, M > C$. Set $\mathbf{b}_j = \mathbf{f}_j$ for all $j = 1, \dots, M$. This lower bound implies that for any estimator $\hat{\boldsymbol{\mu}}$

$$\sup_{\mathbf{f} \in \mathbf{R}^n} \mathbb{P}_{\mathbf{f}} \left(\|\hat{\boldsymbol{\mu}} - \mathbf{f}\|_2^2 - \min_{k=1, \dots, M} \|\mathbf{b}_k - \mathbf{f}\|_2^2 \geq c\sigma^2 \log(M) \right) \geq p.$$

For all $j = 1, \dots, M$, let $A_j = (1/\|\mathbf{f}_j\|_2^2)\mathbf{f}_j\mathbf{f}_j^T$ so that A_j is the orthogonal projection on the linear span of \mathbf{f}_j . The orthoprojector A_j has rank one so under $\mathbb{P}_{\mathbf{f}_j}$, $\|A_j\mathbf{y} - \mathbf{f}_j\|_2^2/\sigma^2$ is a χ^2 random variable with one degree of freedom. Let $\bar{\Omega}'_j$ be the event $\{\|A_j\mathbf{y} - \mathbf{f}_j\|_2^2 \leq c\sigma^2 \log(M)/2\}$ and let $\bar{\Omega}'_j$ be its complementary event. A two-sided bound on the Gaussian tail implies that $\mathbb{P}_{\mathbf{f}_j}(\bar{\Omega}'_j) \leq 2/(M^{c/4})$, which is smaller than $p/2$ if M is larger than some absolute constant, so that we have $\mathbb{P}_{\mathbf{f}_j}(\bar{\Omega}_j \cup \bar{\Omega}'_j) \leq 1 - p + p/2$ where $\bar{\Omega}_j$ is the complementary of Ω_j , which implies $\mathbb{P}_{\mathbf{f}_j}(\Omega_j \cap \Omega'_j) \geq p/2$. Thus, for any estimator $\hat{\boldsymbol{\mu}}$ and M large enough,

$$\begin{aligned} & \sup_{\mathbf{f} \in \{\mathbf{f}_1, \dots, \mathbf{f}_M\}} \mathbb{P}_{\mathbf{f}} \left(\|\hat{\boldsymbol{\mu}} - \mathbf{f}\|_2^2 - \min_{k=1, \dots, M} \|A_k\mathbf{y} - \mathbf{f}\|_2^2 \geq c\sigma^2 \log(M)/2 \right) \\ & \geq p/2 =: p^*. \end{aligned} \quad \square$$

8.6. Proof of Proposition 3.1.

PROOF OF PROPOSITION 3.1. Let $a \in (0, 1)$. By definition of \hat{J} , we have for all $k = 1, \dots, M$,

$$\begin{aligned} \|\hat{\boldsymbol{\mu}}_{\hat{J}} - \mathbf{f}\|_2^2 & \leq \|\hat{\boldsymbol{\mu}}_k - \mathbf{f}\|_2^2 + \Delta_{\hat{J}k} - \frac{a}{2} \|\hat{\boldsymbol{\mu}}_{\hat{J}} - \hat{\boldsymbol{\mu}}_k\|_2^2 + \frac{a}{2} \|\hat{\boldsymbol{\mu}}_{\hat{J}} - \hat{\boldsymbol{\mu}}_k\|_2^2 \\ & \leq \|\hat{\boldsymbol{\mu}}_k - \mathbf{f}\|_2^2 + \frac{1}{a} \max_{j, k=1, \dots, M} \left(a\Delta_{jk} - \frac{1}{2} \|a\hat{\boldsymbol{\mu}}_{\hat{J}} - a\hat{\boldsymbol{\mu}}_k\|_2^2 \right) \\ & \quad + a(\|\hat{\boldsymbol{\mu}}_{\hat{J}} - \mathbf{f}\|_2^2 + \|\mathbf{f} - \hat{\boldsymbol{\mu}}_k\|_2^2). \end{aligned}$$

By rearranging, we have almost surely

$$(8.13) \quad \|\hat{\boldsymbol{\mu}}_{\hat{J}} - \mathbf{f}\|_2^2 \leq \frac{1+a}{1-a} \min_{k=1, \dots, M} \|\hat{\boldsymbol{\mu}}_k - \mathbf{f}\|_2^2 + \frac{\Xi}{a(1-a)},$$

where

$$\Xi := \max_{j, k=1, \dots, M} \left(2\xi^T(\hat{\boldsymbol{\mu}}'_j - \hat{\boldsymbol{\mu}}'_k) - 2\sigma^2 \text{Tr}(A'_j - A'_k) - \frac{1}{2} \|\hat{\boldsymbol{\mu}}'_j - \hat{\boldsymbol{\mu}}'_k\|_2^2 \right),$$

and for all $j = 1, \dots, M$, $\hat{\boldsymbol{\mu}}'_j := a\hat{\boldsymbol{\mu}}_j = A'_j\mathbf{y} + \mathbf{b}'_j$, $A'_j := aA_j$, $\mathbf{b}'_j := a\mathbf{b}_j$. Simple algebra yields that

$$\frac{1}{2} \max_{j \neq k} \|A'_j - A'_k\|_2 \leq a\phi \leq \phi,$$

where ϕ is defined in (2.7). By Proposition 8.1, as in the proof of Theorem 2.1, we have $\Xi \leq 30\phi^2\sigma^2(x + 2\log M)$ with probability greater than $1 - \exp(-x)$.

Set $\varepsilon = 3a$ and choose the absolute constant $c > 0$ such that for all $\varepsilon < c$, $(1 + a)/(1 - a) \leq 1 + \varepsilon$ and $1/(1 - a) \leq 2$. \square

8.7. Smoothness adaptation.

PROOF OF PROPOSITION 7.1. Because the ellipsoids are ordered, if $\mathbf{f} \in \Theta(\beta, L)$ then

$$\mathbb{E}\|\mathbf{f} - \tilde{\boldsymbol{\mu}}\|_2^2 \leq \min_{j:\beta_j \leq \beta} \mathbb{E}\|\mathbf{f} - A_{\beta_j}\mathbf{y}\|_2^2 + 60\sigma^2 \log M \leq \min_{j:\beta_j \leq \beta} C^* n^{\frac{1}{2\beta_j+1}} (1 + o(1)).$$

If $\beta \in [\beta_j, \beta_{j+1})$ for some j , then $\beta_{j+1} - \beta_j = \beta_j/(\log(n) \log \log n)$ and simple algebra yields

$$\begin{aligned} n^{\frac{1}{2\beta_{j+1}} - \frac{1}{2\beta+1}} &\leq n^{\frac{2\beta_{j+1}-2\beta_j}{(2\beta+1)(2\beta_{j+1})}} = n^{\frac{2\beta_j}{(2\beta+1)(2\beta_{j+1})\log(n)\log \log n}} \\ &\leq n^{\frac{1}{(2\beta+1)\log(n)\log \log n}} \leq e^{\frac{1}{3\log \log n}}, \end{aligned}$$

where we used that $\beta \geq 1$ for the last inequality.

Now assume that $\beta \geq \beta_M$. Let $\varepsilon_n = 1/(\log(n) \log \log n)$, and let $c = 120 \log(1 + \varepsilon_3)/\varepsilon_3$. By definition of M ,

$$\begin{aligned} \beta_M &= e^{M \log(1 + \frac{1}{\log(n)\log \log n})} \geq e^{120 \log \log(n) \frac{\log(1+\varepsilon_n)}{\varepsilon_n}} \\ &\geq e^{c \log \log(n)} = \log(n)^c, \end{aligned}$$

since the function $t \rightarrow \log(1 + t)/t$ is decreasing and $n \geq 3$. A numerical approximation gives $c \geq 1.01$. Thus,

$$n^{\frac{1}{2\beta_M+1}} n^{-\frac{1}{2\beta+1}} \leq n^{\frac{1}{2\beta_M+1}} \leq e^{\frac{\log n}{2\beta_M}} \leq e^{\frac{1}{2\log(n)^{c-1}}}.$$

In summary, we have proved that $\min_{j:\beta_j \leq \beta} n^{\frac{1}{2\beta_j+1}} \leq n^{\frac{1}{2\beta+1}} (1 + o(1))$, thus

$$\sup_{\mathbf{f} \in \Theta(\beta, L)} \mathbb{E}\|\mathbf{f} - \tilde{\boldsymbol{\mu}}\|_2^2 \leq C^* n^{\frac{1}{2\beta+1}} (1 + o(1)). \quad \square$$

8.8. Convex aggregation.

LEMMA 8.3 (Maurey argument). *Let m and Λ_m^M be defined in (7.2) and (7.3). Let $Q(\boldsymbol{\theta}) = \boldsymbol{\theta}^T \Sigma \boldsymbol{\theta} + \mathbf{v}^T \boldsymbol{\theta} + a$ for some semidefinite matrix Σ , $\mathbf{v} \in \mathbf{R}^M$ and $a \in \mathbf{R}$. Then*

$$(8.14) \quad \min_{\boldsymbol{\theta} \in \Lambda_m^M} Q(\boldsymbol{\theta}) \leq \min_{\boldsymbol{\theta} \in \Lambda^M} Q(\boldsymbol{\theta}) + \frac{4 \max_{j=1, \dots, M} \Sigma_{jj}}{m}.$$

PROOF. Let $\theta^* \in \Lambda^M \in \operatorname{argmin}_{\theta \in \Lambda^M} Q(\theta)$. Let η be a random variable valued in $\{e_1, \dots, e_M\}$ such that $\mathbb{P}(\eta = e_j) = \theta_j^*$ for all $j = 1, \dots, M$, and let η_1, \dots, η_m be m i.i.d. copies of η . The random variable $\bar{\eta} = \frac{1}{m} \sum_{q=1}^m \eta_q$ is valued in Λ_m^M and $\mathbb{E}\bar{\eta} = \theta^*$. A bias variance decomposition and the independence of η_1, \dots, η_m yield

$$\mathbb{E}Q(\bar{\eta}) = Q(\theta^*) + \frac{\mathbb{E}[(\eta_1 - \theta^*)^T \Sigma (\eta_1 - \theta^*)]}{m}.$$

Using the triangle inequality,

$$\mathbb{E}[(\eta_1 - \theta^*)^T \Sigma (\eta_1 - \theta^*)] \leq 2(\theta^*)^T \Sigma \theta^* + 2\mathbb{E}[\eta_1^T \Sigma \eta_1] \leq 4 \max_{j=1, \dots, M} \Sigma_{jj}.$$

Since $\bar{\eta}$ is valued in Λ_m^M , $\min_{\theta \in \Lambda_m^M} Q(\theta) \leq \mathbb{E}Q(\bar{\eta})$ and the proof is complete. \square

PROOF OF (7.5) OF PROPOSITION 7.2. The condition on M, n implies that $m \geq 1$ where m is defined in (7.2). Let $C > 0$ be an absolute constant whose value may change from line to line. Applying Theorem 2.1 yields that on an event of probability greater than $1 - 2\exp(-x)$,

$$(8.15) \quad \frac{1}{n} \|\hat{\mu}_{\Lambda_m^M} - \mathbf{f}\|_2^2 \leq \min_{\theta \in \Lambda_m^M} \frac{1}{n} \|\hat{\mu}_\theta - \mathbf{f}\|_2^2 + \frac{C\sigma^2(\log(|\Lambda_m^M|) + x)}{n}.$$

By [25], page 8, we have $\log |\Lambda_m^M| \leq m \log \frac{2eM}{m}$. We use (8.14) with $Q(\theta) = \|\hat{\mu}_\theta - \mathbf{f}\|_2^2$ to get

$$\min_{\theta \in \Lambda_m^M} \frac{1}{n} \|\hat{\mu}_\theta - \mathbf{f}\|_2^2 \leq \min_{\theta \in \Lambda^M} \frac{1}{n} \|\hat{\mu}_\theta - \mathbf{f}\|_2^2 + \frac{4}{nm} \max_{j=1, \dots, M} \|\hat{\mu}_j\|_2^2.$$

We have $(1/n) \max_{j=1, \dots, M} \|\hat{\mu}_j\|_2^2 \leq C(\|\xi\|_2^2/n + R^2) \leq C(\sigma^2(2 + 3x) + R^2)$ on an event of probability at least $1 - \exp(-x)$, where for the second inequality we used (3.5) with $B = I_{n \times n}$. Thus, with probability greater than $1 - e^{-x}$,

$$(8.16) \quad \min_{\theta \in \Lambda_m^M} \frac{1}{n} \|\hat{\mu}_\theta - \mathbf{f}\|_2^2 \leq \min_{\theta \in \Lambda^M} \frac{1}{n} \|\hat{\mu}_\theta - \mathbf{f}\|_2^2 + \frac{C(\sigma^2(2 + 3x) + R^2)}{m}.$$

Simple algebra yields that

$$(8.17) \quad \frac{1}{m} \leq C \sqrt{\frac{\log(1 + M/\sqrt{n})}{n}},$$

$$\frac{m \log(2eM/m)}{n} \leq C \sqrt{\frac{\log(1 + M/\sqrt{n})}{n}}.$$

Combining (8.15), (8.16) and (8.17) with the union bound completes the proof. \square

PROOF OF (7.4) OF PROPOSITION 7.2. Let $\theta \in \Lambda^M$. By definition of $\hat{\theta}_C$, $C_p(\hat{\theta}) \leq C_p(\theta)$. This can be rewritten as

$$\|\hat{\mu}_{\hat{\theta}_C} - \mathbf{f}\|_2^2 \leq \|\hat{\mu}_{\hat{\theta}} - \mathbf{f}\|_2^2 + 2\xi^T(\hat{\mu}_{\hat{\theta}} - \hat{\mu}_{\theta}) - 2\sigma^2 \text{Tr}(A_{\hat{\theta}} - A_{\theta}).$$

The function $(\theta', \theta) \rightarrow 2\xi^T(\hat{\mu}_{\hat{\theta}} - \hat{\mu}_{\theta}) - 2\sigma^2 \text{Tr}(A_{\hat{\theta}} - A_{\theta})$ is linear in θ and linear in $\hat{\theta}$, thus it is maximized at vertices of the simplex and

$$\begin{aligned} 2\xi^T(\hat{\mu}_{\hat{\theta}} - \hat{\mu}_{\theta}) - 2\sigma^2 \text{Tr}(A_{\hat{\theta}} - A_{\theta}) &\leq \max_{j,k=1,\dots,M} 2\xi^T(\hat{\mu}_k - \hat{\mu}_j) - 2\text{Tr}(A_j - A_k) \\ &= \max_{j,k=1,\dots,M} \Delta_{jk}, \end{aligned}$$

where Δ_{jk} is defined in (3.3). Fix some pair (j, k) . Let $B = A_j - A_k$ and $\mathbf{b} = (A_j - A_k)\mathbf{f} + \mathbf{b}_j - \mathbf{b}_k$. We have $\|B\|_2 \leq 2$, $\|B\|_F \leq \|B\|_2 \|I_{n \times n}\|_F \leq 2\sqrt{n}$ and $\|\mathbf{b}\|_2 \leq 4R\sqrt{n}$. We apply (3.5) to the matrix B and (3.4) to the vector \mathbf{b} , which yields that with probability greater than $1 - 2\exp(-x)$,

$$\Delta_{jk} \leq 8(\sigma^2 + \sigma R\sqrt{2})\sqrt{nx} + 8\sigma^2 x.$$

The union bound over all pairs $j, k = 1, \dots, M$ completes the proof. \square

Acknowledgements. The author would like to thank the Associate Editor for valuable comments that improved the paper substantially. The author also thanks Alexandre Tsybakov for some helpful discussions on early versions of this manuscript.

REFERENCES

- [1] ARLOT, S. and BACH, F. R. (2009). Data-driven calibration of linear estimators with minimal penalties. In *Advances in Neural Information Processing Systems* 46–54.
- [2] AUDIBERT, J.-Y. (2007). No fast exponential deviation inequalities for the progressive mixture rule. Preprint. Available at [arXiv:math/0703848](https://arxiv.org/abs/math/0703848).
- [3] BARAUD, Y., GIRAUD, C. and HUET, S. (2014). Estimator selection in the Gaussian setting. *Ann. Inst. Henri Poincaré Probab. Stat.* **50** 1092–1119. [MR3224300](https://doi.org/10.1214/13-AIHP630)
- [4] BELLEC, P. C. (2017). Optimal exponential bounds for aggregation of density estimators. *Bernoulli* **23** 219–248. [MR3556772](https://doi.org/10.1080/10236192.2017.1358772)
- [5] BELLONI, A. and CHERNOZHUKOV, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** 521–547. [MR3037163](https://doi.org/10.1080/10236192.2013.823163)
- [6] BELLONI, A., CHERNOZHUKOV, V. and WANG, L. (2014). Pivotal estimation via square-root lasso in nonparametric regression. *Ann. Statist.* **42** 757–788. [MR3210986](https://doi.org/10.1214/13-AOS1198)
- [7] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford Univ. Press, London.
- [8] BOYD, S. and VANDENBERGHE, L. (2009). *Convex Optimization*. Cambridge Univ. Press, Cambridge. [MR2061575](https://doi.org/10.1017/CBO9780511526454)
- [9] COHEN, A. (1966). All admissible linear estimates of the mean vector. *Ann. Math. Stat.* **37** 458–463. [MR0189164](https://doi.org/10.2307/2333164)
- [10] DAI, D., RIGOLLET, P., XIA, L. and ZHANG, T. (2014). Aggregation of affine estimators. *Electron. J. Stat.* **8** 302–327. [MR3192554](https://doi.org/10.1214/13-EJS925)

- [11] DAI, D., RIGOLLET, P. and ZHANG, T. (2012). Deviation optimal learning using greedy Q -aggregation. *Ann. Statist.* **40** 1878–1905. [MR3015047](#)
- [12] DALALYAN, A. S. and SALMON, J. (2012). Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.* **40** 2327–2355. [MR3059085](#)
- [13] DETTE, H., MUNK, A. and WAGNER, T. (1998). Estimating the variance in nonparametric regression—what is a reasonable choice? *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 751–764. [MR1649480](#)
- [14] DONOHO, D. L., LIU, R. C. and MACGIBBON, B. (1990). Minimax risk over hyperrectangles, and implications. *Ann. Statist.* **18** 1416–1437. [MR1062717](#)
- [15] EFROÍMOVICH, S. Y. and PINSKER, M. S. (1984). A self-training algorithm for nonparametric filtering. *Avtomat. i Telemekh.* **11** 58–65. [MR0797991](#)
- [16] GERCHINOVITZ, S. (2011). Prediction of individual sequences and prediction in the statistical framework: Some links around sparse regression and aggregation techniques. Ph.D. thesis, Univ. Paris Sud-Paris XI. Available at <https://tel.archives-ouvertes.fr/tel-00653550>.
- [17] GIRAUD, C. (2008). Mixing least-squares estimators when the variance is unknown. *Bernoulli* **14** 1089–1107. [MR2543587](#)
- [18] GIRAUD, C., HUET, S. and VERZELEN, N. (2012). High-dimensional regression with unknown variance. *Statist. Sci.* **27** 500–518. [MR3025131](#)
- [19] HALL, P., KAY, J. W. and TITTERINGTON, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77** 521–528. [MR1087842](#)
- [20] HANSON, D. L. and WRIGHT, F. T. (1971). A bound on tail probabilities for quadratic forms in independent random variables. *Ann. Math. Stat.* **42** 1079–1083.
- [21] IMMERKAER, J. (1996). Fast noise variance estimation. *Comput. Vis. Image Underst.* **64** 300–302.
- [22] JOHNSTONE, I. M. (2002). Function estimation and gaussian sequence models. Unpublished manuscript 2 (5.3): 2.
- [23] LATAŁA, R. (1999). Tail and moment estimates for some types of chaos. *Studia Math.* **1** 39–53.
- [24] LAURENT, B. and MASSART, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** 1302–1338. [MR1805785](#)
- [25] LECUÉ, G. (2013). Empirical risk minimization is optimal for the convex aggregation problem. *Bernoulli* **19** 2153–2166.
- [26] LECUÉ, G. and RIGOLLET, P. (2014). Optimal learning with Q -aggregation. *Ann. Statist.* **42** 211–224. [MR3178462](#)
- [27] LEUNG, G. and BARRON, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory* **52** 3396–3410.
- [28] MALLOWS, C. L. (1973). Some comments on C_p . *Technometrics* **15** 661–675.
- [29] MUNK, A., BISSANTZ, N., WAGNER, T. and FREITAG, G. (2005). On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 19–41. [MR2136637](#)
- [30] NEMIROVSKI, A. (2000). Topics in non-parametric statistics. In *Lectures on Probability Theory and Statistics (Saint-Flour, 1998)*. *Lecture Notes in Math.* **1738**. Springer, Berlin.
- [31] RIGOLLET, P. (2012). Kullback–Leibler aggregation and misspecified generalized linear models. *Ann. Statist.* **40** 639–665. [MR2933661](#)
- [32] RIGOLLET, P. and TSYBAKOV, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39** 731–771. [MR2816337](#)
- [33] RIGOLLET, P. and TSYBAKOV, A. B. (2007). Linear and convex aggregation of density estimators. *Math. Methods Statist.* **16** 260–280. [MR2356821](#)
- [34] RIGOLLET, P. and TSYBAKOV, A. B. (2012). Sparse estimation by exponential weighting. *Statist. Sci.* **27** 558–575. [MR3025134](#)

- [35] RUDELSON, M. and VERSHYNIN, R. (2013). Hanson–Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.* **18** 1–9.
- [36] SUN, T. and ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. [MR2999166](#)
- [37] TSYBAKOV, A. B. (2003). Optimal rates of aggregation. In *Learning Theory and Kernel Machines* 303–313. Springer, Berlin.
- [38] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York.
- [39] TSYBAKOV, A. B. (2014). Aggregation and minimax optimality in high-dimensional estimation. In *Proceedings of the International Congress of Mathematicians, Vol. 3* 225–246, Seoul.
- [40] VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. [MR2963170](#)
- [41] WRIGHT, F. T. (1973). A bound on tail probabilities for quadratic forms in independent random variables whose distributions are not necessarily symmetric. *Ann. Probab.* **1** 1068–1070. [MR353419](#)

DEPARTMENT OF STATISTICS & BIostatISTICS
RUTGERS, THE STATE UNIVERSITY OF NEW JERSEY
501 HILL CENTER, BUSCH CAMPUS
110 FRELINGHUYSEN ROAD
PISCATAWAY, NEW JERSEY 08854
USA
E-MAIL: pcb71@stat.rutgers.edu