Electronic Journal of Statistics
ISSN: 1935-7524

Aggregation of Affine Estimators

Dong Dai, Philippe Rigollet*, Lucy Xia and Tong Zhang†

Dong Dai
Statistics department
Rutgers
Piscataway,NJ 08854, USA
dongdai@stat.rutgers.edu

Philippe Rigollet
Department of Operations Research
and Financial Engineering
Princeton University
Princeton, NJ 08544, USA
rigollet@princeton.edu

Lucy Xia
Department of Operations Research
and Financial Engineering
Princeton University
Princeton, NJ 08544, USA
lxia@princeton.edu

Tong Zhang
Statistics department
Rutgers
Piscataway,NJ 08854, USA
tzhang@stat.rutgers.edu

Abstract: We consider the problem of aggregating a general collection of affine estimators for fixed design regression. Relevant examples include some commonly used statistical estimators such as least squares, ridge and robust least squares estimators. Dalalyan and Salmon [DS12] have established that, for this problem, exponentially weighted (EW) model selection aggregation leads to sharp oracle inequalities in expectation, but similar bounds in deviation were not previously known. While results [DRZ12] indicate that the same aggregation scheme may not satisfy sharp oracle inequalities with high probability, we prove that a weaker notion of oracle inequality for EW that holds with high probability. Moreover, using a generalization of the newly introduced Q-aggregation scheme we also prove sharp oracle inequalities that hold with high probability. Finally, we apply our results to universal aggregation and show that our proposed estimator leads simultaneously to all the best known bounds for aggregation, including ℓ_q-aggregation, $q \in (0,1)$, with high probability.

AMS 2000 subject classifications: Primary 62G08; secondary 62C20, 62G05, 62G20.
Keywords and phrases: Aggregation, Affine estimators, Gaussian mean, Oracle inequalities, Maurey's argument.

1. Introduction

In the Gaussian Mean Model (GMM), we observe a Gaussian random vector $Y \in \mathbb{R}^n$ such that $Y \sim \mathcal{N}(\mu, \sigma^2 I_n)$ where the mean $\mu \in \mathbb{R}^n$ is unknown and the variance parameter σ^2 is known. For the purpose of discussion, we assume that $\sigma^2 = 1$ throughout this introduction but our main subsequent results explicitly depend on σ^2.

This apparently simple model introduced in a notorious paper [Ste56] by Stein, was the starting point of a vast literature on shrinkage [Gru98] that later evolved in the Gaussian sequence model. This literature is much too vast to explore here but we refer the reader to the excellent manuscript by Johnstone [Joh11] for both motivation and partial literature review.

Independently of the variety of methods and results dedicated to the GMM, Nemirovski [JN00, Nem00] introduced *aggregation theory* as a versatile tool for adaptation in nonparametric estimation [Lec07, RT07,

*Philippe Rigollet is supported by the following grants: DMS-1317308 and CAREER-DMS-1053987.
†Tong Zhang is supported by the following grants: NSF IIS-1016061, NSF DMS-1007527, and NSF IIS-1250985.

Yan04], but also more recently in high dimensional regression [LB06, RT11, DS12]. In all these results, exponential weights have played a key role (see [RT12] for a recent survey). Specifically, we focus here on *model selection aggregation* where, given a family of estimators $\hat{\mu}_1, \dots, \hat{\mu}_M$, the goal is to mimic the best of them. Originally, aggregation was accompanied with a sample splitting scheme in which the sample was split into two parts: the first one to construct various estimators and the second to aggregate them. For example, this approach was practically implemented in [RT07] for density estimation and in [Lec07] for classification. The advantage of sample splitting is that it allows to *freeze* the first sample and therefore treat the estimators to be aggregated as deterministic functions that only satisfy mild boundedness assumption. This is the framework of *pure aggregation* under which most of the developments have been made starting from the seminal works on aggregation [JN00, Nem00, Tsy03]. Pure model selection aggregation in the GMM can be described as follows. Given $M \geq 2$ vectors $\mu_1, \dots, \mu_M$, the goal is to construct an estimator $\hat{\mu}$ called *aggregate*, using the observation Y and such that

$$\|\hat{\mu} - \mu\|^2 - \min_{1 \leq j \leq M} \|\mu_j - \mu\|^2$$

is as small as possible, where $\|\cdot\|$ denotes the Euclidean distance on $\mathbb{R}^n$. Bounds on this quantity are called *sharp oracle inequalities*. While not directly connected to Stein's original result on admissibility [Ste56], it turns out that for aggregation too, the most natural choice $\hat{\mu} = \mu_{\hat{\jmath}}$ where $\hat{\jmath} = \operatorname{argmin}_{1 \leq j \leq M} \|\mu_j - Y\|^2$ is suboptimal. Nevertheless, this problem is by now well understood and various optimal choices for $\hat{\mu}$ relying on model averaging rather than model selection were proposed and proved to be optimal (see [RT12] and references therein). Two approaches have been employed successfully. The first family of methods is based on exponential weights [DT07, DT08]. Following original ideas of Catoni [Cat99] and Yang [Yan99], it can be proved that for any *prior* probability distribution $\pi = (\pi_1, \dots, \pi_M)$ on $[M] = \{1, \dots, M\}$, there exists an aggregate $\hat{\mu}^{\mathsf{EW}}$ based on exponential weights that satisfies the following *sharp oracle inequality*:

$$\mathbb{E}\|\hat{\mu}^{\mathsf{EW}} - \mu\|^2 \leq \min_{1 \leq j \leq M} \left\{ \|\mu_j - \mu\|^2 + C \log\left(\frac{1}{\pi_j}\right) \right\}, \tag{1.1}$$

where here and in what follows $C > 0$ is a numerical constant that may change from line to line. In particular, if π is chosen to be the uniform distribution, this estimator attains the optimal rate $C \log(M)$ [RT11] that is independent of the dimension n. Nevertheless, it was observed in [DRZ12] that the random quantity $\|\hat{\mu}^{\mathsf{EW}} - \mu\|^2$ may have fluctuation of order $\sqrt{n}$ around its expectation so that the bound (1.1) may fail to accurately describe the risk of $\hat{\mu}^{\mathsf{EW}}$, especially for large dimension n. To overcome this limitation, several methods have been proposed in the literature [Aud08, LM09]. More recently, a new and flexible method called Q-aggregation was proposed and studied in several settings [Rig12, DRZ12, LR14]. It enjoys the following property. For any prior π on $[M]$, it yields an aggregate $\hat{\mu}^{Q}$ that satisfies not only a sharp oracle inequality *in expectation* of form (1.1) but also one that holds *with high probability*:

$$\|\hat{\mu}^{Q} - \mu\|^2 \leq \min_{1 \leq j \leq M} \left\{ \|\mu_j - \mu\|^2 + C \log\left(\frac{1}{\delta \pi_j}\right) \right\}, \tag{1.2}$$

with probability $1 - \delta$.

In this paper, we extend this work to the aggregation of not fixed vectors $\mu_1, \dots, \mu_M$ but of affine estimators $\hat{\mu}_1, \dots, \hat{\mu}_M$ that are of the form $\hat{\mu}_j = A_j Y + b_j$ for some deterministic matrix-vector pair (A_j, b_j). Note that these estimators are constructed using the same observations Y as the ones employed for aggregation. In particular, no sample splitting scheme is needed.

A canonical example of affine estimators where A_j are projection matrices, was first introduced in [LB06] and further studied in [RT11] under the light of high-dimensional linear regression. In a remarkable paper, Dalalayan and Salmon [DS12] recently extended these setups to a more general family of affine estimators, under mild conditions on matrices A_j. Nevertheless, all these previous papers are limited to deriving sharp oracle inequalities in expectation of the same type as (1.1). Moreover, the lower bounds of [DRZ12] indicate that the estimators based on exponential weights that are employed in [LB06, RT11, DS12] are unlikely to satisfy sharp oracle inequalities with high probability. In this paper, akin to [Rig12, DRZ12, LR14], we demonstrate that Q-aggregation succeeds where exponential weights have failed by proving a sharp oracle inequality that holds with high probability in Section 2.1. Yet, the situation regarding exponential weights is not desperate as we show in Section 2.2 that it still leads to a weaker notion of oracle inequalities.

The rest of this paper is organized as follows. In the next section, we give a precise description of the problem of *model selection aggregation of affine estimators* and give a solution to this problem using Q-aggregation. Specifically, in Section 2.1, we show that for any *prior* probability distribution $\pi = (\pi_1, \dots, \pi_M)$ on $[M] = \{1, \dots, M\}$, there exists an aggregate $\hat{\mu}^Q$ based on Q-aggregation that satisfies a sharp oracle inequality of the form

$$\|\hat{\mu}^Q - \mu\|^2 \le \min_{j\in[M]} \left\{ \|\hat{\mu}_j - \mu\|^2 + C \log\left(\frac{1}{\pi_j}\right) + 4\sigma^2 \,\mathsf{Tr}(A_j) \right\}, \tag{1.3}$$

that holds both in expectation and with high probability, where $\mathsf{Tr}(A_j)$ denotes the trace of A_j. We continue by proving in Section 2.2 that for any $\varepsilon > 0$, there exists a choice of the temperature parameter for which the better known aggregate $\hat{\mu}^{\mathsf{EW}}$ based on exponential weights satisfies a *weak oracle inequality* that holds with high probability

$$\|\hat{\mu}^{\mathsf{EW}} - \mu\|^2 \le \min_{j\in[M]} \left\{ (1+\varepsilon)\|\hat{\mu}_j - \mu\|^2 + \frac{C}{\varepsilon} \log\left(\frac{1}{\pi_j}\right) + 8\sigma^2 \,\mathsf{Tr}(A_j) \right\}. \tag{1.4}$$

Such an inequality completes the sharp oracle inequality of [DS12] that holds in expectation.

We give applications of these oracle inequalities to sparsity pattern aggregation and universal aggregation in Section 3. In particular, we show that Q-aggregation of projection estimators leads to the first sharp oracle inequalities *that hold with high probability* for these two problems. By "high probability", we mean a statement that holds with probability at least $1 - \delta$, $0 < \delta < 1/2$. Our results below exhibit explicit dependence on δ.

NOTATION: For any integer n, the set of integers $\{1, \dots, n\}$ is denoted by $[n]$. We denote by $\mathsf{Tr}(A)$ and $\mathsf{Rk}(A)$ respectively the trace and the rank of a square matrix A. We denote by $\|\cdot\|$ the Euclidean norm of $\mathbb{R}^n$ and by $|J|$ the cardinality of a finite set J. For any real numbers $a_1, \dots, a_n$, $\mathsf{diag}(a_1, \dots, a_n)$ denotes the $n \times n$ diagonal matrix with $a_1, \dots, a_n$, on the diagonal. The indicator function is denoted by $\mathbb{1}(\cdot)$ and for any integer n, $K \subset [n]$, $\mathbb{1}_K$ denotes the vector $v \in \{0,1\}^n$ with jth coordinate given by $v_j = 1$ iff $j \in K$. For any matrix B, $B^\dagger$ denotes the Moore-Penrose pseudoinverse of B. The operator norm of a matrix is denoted by $\|\cdot\|_{\mathsf{op}}$. The cone of $n \times n$ symmetric positive semidefinite matrices is denoted by $\mathcal{S}_n$. The flat simplex of $\mathbb{R}^M$ is denoted by Λ_M and is defined by

$$\Lambda_M = \left\{ \theta \in \mathbb{R}^M \,:\, \theta_j \ge 0\,, \sum_{j=1}^M \theta_j = 1 \right\}$$

The set Λ_M can be identified to the set of probability measures on $[M]$ and for any $\theta, \pi \in \Lambda_M$, we define the Kullback-Leibler divergence between these two measures by

$$\mathcal{K}(\theta, \pi) = \sum_{j=1}^M \theta_j \log\left(\frac{\theta_j}{\pi_j}\right),$$

with the usual convention that $0\log(0) = 0, 0\log(0/0) = 0$ and $\theta\log(\theta/0) = +\infty$, $\forall\, \theta > 0$. Finally, throughout the paper, we use the notation $\overline{\log}(x)$ to denote the function $\overline{\log}(x) = (\log x) \vee 1$.

2. Aggregation of affine estimators

Recall that the Gaussian Mean Model (GMM) can be written as follows. One observes $Y \in \mathbb{R}^n$ such that

$$Y = \mu + \xi\,, \qquad \xi \sim \mathcal{N}(0, \sigma^2 I_n)\,. \tag{2.1}$$

Throughout this paper and in accordance with [DS12], we call an affine estimator of μ any estimator $\hat{\mu}$ of the form

$$\hat{\mu} = AY + b\,, \tag{2.2}$$

where $A \in \mathcal{S}_n$ is a $n \times n$ matrix and $b \in \mathbb{R}^n$ is a n-dimensional vector. Note that unlike [DS12], we impose that that A be symmetric. While such an assumption can be relaxed, all examples presented in [DS12] involve symmetric matrices. Both A and b are deterministic.

Given a family of affine estimators $\hat{\mu}_1, \dots, \hat{\mu}_M$, where $\hat{\mu}_j = A_j Y + b_j$ and a prior probability measure $\pi = (\pi_1, \dots, \pi_M)$ on these estimators, our goal is to construct an aggregate $\tilde{\mu}$ such that

$$\|\tilde{\mu} - \mu\|^2 \le (1+\varepsilon)\|\hat{\mu}_j - \mu\|^2 + C\big[\log\big(\frac{1}{\delta\pi_j}\big) + T_j\big]\,, \tag{2.3}$$

with probability $1-\delta$ for any $j \in [M]$, where $T_j > 0$ is as small as possible and $\varepsilon \ge 0$. As we will see, we can achieve $\varepsilon = 0$ using Q-aggregation but only prove a weak oracle inequality with $\varepsilon > 0$ in Section 2.2 using exponential weights.

Inequalities of the form (2.3) with $\varepsilon > 0$ can be of interest as long as there exists a candidate affine estimator $\hat{\mu}_j$ that is close to μ with high probability. Several examples where it is the case are described in [DS12].

Our results below hold under the following general condition on the family of matrices $\{A_j\}_{j\in[M]}$.

Condition 1. *There exists a finite $V > 0$ such that $\max_{j\in[M]} \|A_j\|_{\mathsf{op}} = V$.*

Note that Condition 1 excludes matrices A_j that lead to inadmissible estimators of the Gaussian mean μ [Coh66]. To illustrate the purpose of aggregating affine estimators and the relevance of Condition 1, observe that a large body of the literature on the GMM studies estimators of the form AY, where $A = \mathsf{diag}(a_1, \dots, a_n)$ is a diagonal matrix with elements $a_j \in [0,1]$ for all $j = 1, \dots, n$. If μ is assumed to belong to some family of regularity classes such as Sobolev ellipsoids, Besov classes, tail classes, it has been proved that such estimators are minimax optimal (see [CT01, Tsy09, Joh11]). Commonly used examples include ordered projection estimators, spline estimators and Pinsker estimators (see [DS12] for a detailed description). These estimators are known to be minimax optimal over Sobolev ellipsoids [Pin80, GN92, Tsy09]. Diagonal filters trivially satisfy Condition 1 with $V = 1$.

We give details of a specific application to sparsity pattern aggregation and its consequences on universal aggregation in Section 3.

2.1. Sharp oracle inequalities using Q-Aggregation

In this section, we state our main result: a sharp oracle inequality for an aggregate of affine estimators based on Q-aggregation. Specifically, we consider the problem of aggregating general affine estimators $\hat{\mu}_j = A_j Y + b_j, j \in [M]$ that satisfy Condition 1. Note that unlike [DS12], we do not require that matrices $A_j, j \in [M]$ commute and we make no assumption on the vectors $b_j, j = 1, \dots, M$. Moreover, our results can be extended to an infinite family $\{(A_\lambda, b_\lambda), \lambda \in \mathcal{L}\}$ as in [DS12] but we prefer to present our result in the discrete case for the sake of clarity.

For any $\theta \in \mathbb{R}^M$, let μ_θ denote the linear combination of some given affine estimators $\hat{\mu}_1, \dots, \hat{\mu}_M$ that is defined by

$$\mu_\theta = \sum_{j=1}^M \theta_j \hat{\mu}_j\,.$$

Our goal is to find a vector $\hat{\theta} \in \mathbb{R}^M$ such that the aggregate $\mu_{\hat{\theta}}$ mimics the affine estimator $\hat{\mu}_j$ that is the closest to the true mean μ.

In this paper, we consider a generalization of the Q-aggregation scheme of static models that was developed in [Rig12, DRZ12]. To that end, fix a prior probability distribution $\pi \in \Lambda_M$ and for any $\theta \in \Lambda_M$, define

$$Q(\theta) = \nu \sum_{j=1}^M \theta_j \|Y - \hat{\mu}_j\|^2 + (1-\nu)\|Y - \mu_\theta\|^2 + \sum_{j=1}^M \theta_j C_j + \lambda \mathcal{K}(\theta, \pi), \tag{2.4}$$

where $\nu \in (0,1)$ and $\lambda > 0$ are tuning parameters, and C_j is set to be

$$C_j = 4\sigma^2\, \mathsf{Tr}(A_j)\,. \tag{2.5}$$

Let now $\hat{\theta}$ be defined as

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \Lambda_M} Q(\theta)\,. \tag{2.6}$$

The resulting estimator $\mu_{\hat{\theta}}$ is called Q-*aggregate* estimator of μ. Theorem 1 is our main result.

Theorem 1. *Consider the GMM (2.1) and let $\hat{\mu}_j = A_j Y + b_j, j \in [M]$ be affine estimators of μ together with a prior distribution $\pi = (\pi_1, \dots, \pi_M)$ on these estimators and let $V = \max_{j\in[M]} \|A_j\|_{\mathsf{op}}$. Let $\hat{\mu}^Q = \mu_{\hat{\theta}}$ be the Q-aggregate estimator with $\hat{\theta}$ defined in (2.6) with tuning parameters $\nu \in (0,1)$ and $\lambda \geq \frac{8\sigma^2}{\min(\nu, 1-\nu, 2/(3V))}$. Then for any $\delta > 0$, any $j \in [M]$, with probability at least $1-\delta$, we have*

$$\|\hat{\mu}^Q - \mu\|^2 \leq \min_{j\in[M]} \Big\{ \|\hat{\mu}_j - \mu\|^2 + C_j + 2\lambda \log(\frac{1}{\pi_j \delta}) \Big\}\,. \tag{2.7}$$

Moreover, the same Q-aggregate estimator $\hat{\mu}^Q$ satisfies

$$\mathbb{E}\|\hat{\mu}^Q - \mu\|^2 \leq \min_{j\in[M]} \Big\{ \mathbb{E}\|\hat{\mu}_j - \mu\|^2 + C_j + \lambda \log(\frac{1}{\pi_j}) \Big\}. \tag{2.8}$$

A few remarks are in order. Note that the oracle inequality of Theorem 1 is *sharp* since the leading term $\|\hat{\mu}_j - \mu\|^2$ has multiplicative constant 1. A similar oracle inequality was obtained in [DS12] bit our main theorem above presents significant differences. First, and this is the main contribution of this paper, the above oracle inequality holds with high probability whereas the ones in [DS12] only hold in expectation. Nevertheless, our model is simpler than the one studied in [DS12] who study heteroskedastic regression. Moreover, the bound in [DS12, Theorem 2] is "scale-free" whereas ours depends critically on the size of the matrices A_j via C_j and V. We believe that this dependence cannot be avoided in high probability bounds as such quantities essentially control the deviations of estimators. As we will see, the bounds of Theorem 1 are sufficient to perform sparsity pattern aggregation and universal aggregation optimally.

2.2. Weak oracle inequality using exponential weights

The oracle inequalities (1.1) and (1.2) are *sharp* in contrast to *weak* oracle inequalities where the right-hand side of (1.1) or (1.2) is replaced by

$$\min_{1\leq j\leq M} \{(1+\varepsilon)\|\mu_j - \mu\|^2 + C_j + \frac{C}{\varepsilon} \log\big(\frac{1}{\delta \pi_j}\big)\}\,,$$

for some $\varepsilon > 0$ (see [LM12] for a discussion on the difference between sharp and weak oracle inequalities). While they appear to be quite similar, some estimators do satisfy weak oracle inequalities while they do not satisfy sharp ones. This is the case of the aggregate with exponential weight that provably fails to satisfy a sharp oracle inequality with high probability in a certain setups [DRZ12, Proposition 2.1].

To prove weak oracle inequalities that hold with high probability, we modify the aggregate studied in [DS12]. Recall that $\hat{\mu}_j = A_j Y + b_j, j \in [M]$ is a family of affine estimators equipped with a prior probability distribution $\pi \in \Lambda_M$ and that C_j is defined in (2.5). Let $\hat{\theta} \in \Lambda_M$ be the vector of exponential weights defined by

$$\hat{\theta}_j \propto \pi_j \exp\big(-\frac{\|Y - \hat{\mu}_j\|^2 + C_j}{\lambda}\big)\,, \quad j \in [M]\,. \tag{2.9}$$

The parameter $\lambda > 0$ is often referred to as *temperature parameter*. It is not hard to show (see, e.g., [Cat04, p. 160]) that $\hat{\theta}$ is the solution of the following optimization problem:

$$\hat{\theta} \in \operatorname*{argmin}_{\theta\in\Lambda_M} \Big\{ \sum_{j=1}^M \theta_j \|Y - \hat{\mu}_j\|^2 + \sum_{j=1}^M \theta_j C_j + \lambda \mathcal{K}(\theta, \pi) \Big\}\,. \tag{2.10}$$

Observe that the above criterion corresponds to Q defined in (2.4) with $\nu = 1$, that is without the quadratic term in θ. We believe that this quadratic term is key in obtaining sharp oracle inequalities that hold with high

probability. We already know from previous work [LB06, RT11, RT12, DS12] that this term is not necessary to obtain sharp oracle inequalities that hold in expectation. As illustrated below, it is also not required to get weak oracle inequalities, even with high probability.

Denote by $\hat{\mu}^{\mathsf{EW}} = \sum_{j=1}^{M} \hat{\theta}_j \hat{\mu}_j$ the aggregate with exponential weights $\hat{\theta}_j\,, j \in [M]$ defined in (2.9).

Theorem 2. *Let the conditions of Theorem 1 hold. Let $\hat{\mu}^{\mathsf{EW}} = \mu_{\hat{\theta}}$ be the aggregate with exponential weights $\hat{\theta}$ defined in (2.9) with tuning parameter $\lambda \geq 4\sigma^2(16 \vee 3V)$. Then for any $\delta > 0$, with probability at least $1-\delta$, we have*

$$\|\mu - \hat{\mu}^{\mathsf{EW}}\|^2 \leq \min_{j\in[M]} \Big\{ \big(1 + \frac{128\sigma^2}{3\lambda}\big)\|\mu - \hat{\mu}_j\|^2 + 2C_j + 3\lambda \log\big(\frac{1}{\delta\pi_j}\big) \Big\}. \tag{2.11}$$

Note that unlike Theorem 1, the right-hand side of the above oracle inequality is multiplied by a factor $1+\varepsilon > 1$: it is a weak oracle inequality but it holds with high probability and thus complements the results of [DS12] on aggregation of affine estimators using exponential weights. Note that the inequality (A.8) proved in Theorem 2 gives a slightly stronger bound than (2.11), especially in terms of constants. Alquier and Lounci [AL11, Theorem 3.1] prove the first oracle inequality with high probability using exponential weights. However, their result does not balance the the approximation error and the complexity term which leads to a weaker result than Theorem 3 below.

3. Sparsity pattern aggregation

In this section, we illustrate the power of the two oracle inequalities stated in the previous section. Indeed, carefully selecting the affine estimators $\hat{\mu}_1, \dots, \hat{\mu}_M$, as well as the the prior probability distribution π leads to various optimal results. Some results for diagonal filters can be found in [DS12] and we focus here on sparsity pattern aggregation.

Recall the results we have proved in the previous section. With probability at least $1-\delta$, for λ large enough,

$$\|\mu_{\hat{\theta}} - \mu\|^2 \leq \min_{j\in[M]} \Big\{ \big(1 + \frac{t_1}{\lambda}\big)\|\hat{\mu}_j - \mu\|^2 + t_2 C_j + t_3 \lambda \log\big(\frac{1}{\pi_j \delta}\big) \Big\}\,.$$

where $(t_1, t_2, t_3) = (0, 1, 2)$ if $\hat{\theta}$ is computed according to (2.6) and $(t_1, t_2, t_3) = (128\sigma^2/3, 2, 3)$ if $\hat{\theta}$ is computed according to (2.9).

In the sequel, we fix $\nu = 1/2$ in Q-aggregation since this choice leads to the sharpest bounds.

3.1. Sparsity pattern aggregation

Let $X_1, \dots, X_p \in \mathbb{R}^n$ be given vectors and assume that $\mu \in \mathbb{R}^n$ can be well approximated by a linear combination of $X_j, j \in J^*$ for some unknown sparsity pattern $J^* \subset [p]$. More precisely, we are interested in *sparse* linear regression, where the goal is to find a sparse $\beta \in \mathbb{R}^p$ such that $\|\mathbb{X}\beta - \mu\|^2$ is small, where $\mathbb{X} = [X_1, \dots, X_p]$ is the $n \times p$ design matrix obtained by concatenating the X_j's. Akin to [BRT09, RT11], we do not assume that there exists a sparse β^* such that $\mathbb{X}\beta^* = \mu$ but rather that there may be a systematic error. Oracle inequalities such as the ones described below in Theorems 3 and 4 capture the statistical precision of fitting possibly misspecified sparse linear models in the GMM.

To achieve our goal, we follow the same idea as in [RT11, RT12] and employ *sparsity pattern aggregation*. The idea can be summarized as follows. For each sparsity pattern of β, compute the least squares estimator and then aggregate these (projection) estimators. Specifically, for each sparsity pattern $J \subset [p]$ define $\mathbb{X}_J$ to be the $n \times |J|$ matrix obtained by concatenating $X_j, j \in J$ and let $A_J = \mathbb{X}_J(\mathbb{X}_J^\top \mathbb{X}_J)^\dagger \mathbb{X}_J^\top$ denote the projection matrix onto the linear span $\mathsf{span}(\mathbb{X}_J)$ of $X_j, j \in J$. If J^* was known, a good candidate to estimate μ would be the least squares estimator $\hat{\mu}_{J^*} = A_{J^*}Y$. Since J^* is unknown, we propose to aggregate the affine (actually linear) estimators $\hat{\mu}_J = A_J Y, J \subset [p]$. This approach is called *sparsity pattern aggregation* [RT11] and can be extended to more general notions of sparsity such as group sparsity or fused sparsity [RT12]. It yields a family of affine estimators $\hat{\mu}_J = A_J Y$ such that $C_J = 4\sigma^2\,\mathsf{Tr}(A_J) = 4\sigma^2\,\mathsf{Rk}(\mathbb{X}_J)$ and $V = \max_J \|A_J\|_{\mathsf{op}} = 1$.

Sparsity pattern aggregation has been shown to attain the best available sharp oracle inequalities in expectation [RT11, RT12] and one of the main contribution of this paper is to extend these results to results with high probability. Moreover, it leads to universal aggregation with high probability (see Section 3.3).

The key to sparsity pattern aggregation is to employ a correct prior probability distribution. Rigollet and Tsybakov [RT12], following [LB06, Gir08] suggest to use

$$\pi_J \propto \frac{e^{-|J|}}{\binom{p}{|J|}}. \tag{3.1}$$

In particular, it exponentially downweights patterns J according to their cardinality.

For any $\beta \in \mathbb{R}^p \setminus \{0\}$, let $|\beta|_0$ denote the number of nonzero coefficients of β and, by convention, let $|0|_0 = 1$.

Theorem 3. *Let $\hat{\mu}_J, J \subset [p]$ be the least squares estimator defined as above, let π be the sparsity prior defined in (3.1) and fix $\delta > 0$. Then the following statements hold:*

(i) For $\lambda \geq 12\sigma^2$, with probability at least $1-\delta$, the Q-aggregate estimator $\hat{\mu}^Q$ satisfies

$$\|\hat{\mu}^Q - \mu\|^2 \leq \min_{\beta \in \mathbb{R}^p} \Big\{ \|\mathbb{X}\beta - \mu\|^2 + 6(\lambda + \sigma^2)|\beta|_0 \log\Big(\frac{2ep}{|\beta|_0 \delta}\Big) \Big\}. \tag{3.2}$$

(ii) For $\lambda \geq 64\sigma^2$, with probability at least $1-\delta$, the aggregate with exponential weights $\hat{\mu}^{\mathsf{EW}}$ satisfies

$$\|\hat{\mu}^{\mathsf{EW}} - \mu\|^2 \leq \min_{\beta \in \mathbb{R}^p} \Big\{ \Big(1 + \frac{128\sigma^2}{3\lambda}\Big) \|\mathbb{X}\beta - \mu\|^2 + 6(\lambda + 2\sigma^2)|\beta|_0 \log\Big(\frac{2ep}{|\beta|_0 \delta}\Big) \Big\}. \tag{3.3}$$

Corollary 1. *Taking $\lambda = 12\sigma^2$ and $\lambda = 64\sigma^2$ for $\hat{\mu}^Q$ and $\hat{\mu}^{\mathsf{EW}}$ respectively, with probability at least $1-\delta$, we have:*

$$\|\hat{\mu}^Q - \mu\|^2 \leq \min_{\beta \in \mathbb{R}^p} \Big\{ \|\mathbb{X}\beta - \mu\|^2 + 78\sigma^2|\beta|_0 \log\Big(\frac{2ep}{|\beta|_0 \delta}\Big) \Big\}, \tag{3.4}$$

and

$$\|\hat{\mu}^{\mathsf{EW}} - \mu\|^2 \leq \min_{\beta \in \mathbb{R}^p} \Big\{ \frac{5}{3} \|\mathbb{X}\beta - \mu\|^2 + 396\sigma^2|\beta|_0 \log\Big(\frac{2ep}{|\beta|_0 \delta}\Big) \Big\}. \tag{3.5}$$

The novelty of this result is twofold. First, we use Q-aggregation to obtain the first sharp sparsity oracle inequalities that hold with high probability under no additional condition on the problem. Second, we prove a weak sparsity oracle inequality for the aggregate based on exponential weights that holds with high probability. While it is only a weak oracle inequality, it extends the results of Rigollet and Tsybakov [RT11, RT12] that hold only in expectation and the results of [AL11] that hold with high probability but under additional conditions.

3.2. ℓ_q-aggregation

Recently, Rigollet and Tsybakov [RT11] observed that any estimator that satisfies an oracle inequality such as (2.8) also adapts to sparsity when measured in terms of ℓ_1 norm. Specifically, their result [RT11, Lemma A.2] implies that if $\max_j \|\mu_j\| \leq \sqrt{n}$, then for any constant $\nu > 0$, it holds

$$\min_{\theta \in \mathbb{R}^M} \Big\{ \|\mu_\theta - \mu\|^2 + \nu^2|\theta|_0 \log\Big(1 + \frac{eM}{|\theta|_0}\Big) \Big\} \leq \min_{\theta \in \mathsf{B}_1(1)} \|\mu_\theta - \mu\|^2 + \bar{c}\nu \sqrt{n \log\Big(1 + \frac{eM\nu}{\sqrt{n}}\Big)}, \tag{3.6}$$

where $\bar{c}$ is an absolute constant. The above bound hinges on a Maurey argument, which, as noticed by Wang *et al.* [WPGY11], can be extended from ℓ_1 balls to ℓ_q balls for $q \in (0,1]$. It has been argued that ℓ_q-balls $(0 < q \leq 1)$ describe vectors that are "almost sparse" [FPRU10, Joh11].

For any $q \in (0,1], \theta \in \mathbb{R}^M$, let $|\theta|_q$ denote the ℓ_q-"norm" of $\mathbb{R}^M$ of θ defined by

$$|\theta|_q = \Big(\sum_{j\in[M]} |\theta_j|^q \Big)^{\frac{1}{q}}.$$

Moreover, for a given radius $R > 0$ and any $q \in [0,1]$, define the ℓ_q-ball of radius R by

$$\mathsf{B}_q(R) = \big\{\theta \in \mathbb{R}^M \,:\, |\theta|_q \le R\big\}.$$

Not surprisingly, these almost sparse vectors can be well approximated by sparse vectors as illustrated in the following lemma that generalized (3.6)

Lemma 1. *Fix $\nu > 0$, $M \ge 3$ and let and $\mu_j, j \in [M]$ such that $\max_j \|\mu_j\|^2 \le B^2$. Then*

$$\min_{\theta\in\mathbb{R}^M} \Big\{\|\mu_\theta - \mu\|^2 + \nu^2|\theta|_0 \log\big(1 + \frac{eM}{|\theta|_0}\big)\Big\} \le \inf_{0\le q\le 1} \min_{\theta\in\mathbb{R}^M} \Big\{\|\mu_\theta - \mu\|^2 + \varphi_{q,M}(\theta;\nu,B)\Big\},$$

where

$$\varphi_{q,M}(\theta;\nu,B) = 9\nu^{2-q}|\theta|_q^q B^q \Big[\overline{\log}\Big(\frac{eM\nu^q}{B^q|\bar\theta|_q^q\delta}\Big)\Big]^{1-\frac{q}{2}} \vee 3\nu^2\,\overline{\log}\big(\frac{eM}{\delta}\big). \tag{3.7}$$

with the convention $|\theta|_0^0 = |\theta|_0$.

We postpone the proof to Appendix B where further results on the approximation of vectors with small ℓ_q norm by sparse vectors, can be found. We are now in a position to state the main result of this subsection. Its proof follows directly from the above lemma by rounding $\sqrt{78}$ up to 9 and $\sqrt{5\cdot 396/3}$ to 16.

Theorem 4. *Let $\hat\mu_J, J \subset [p]$ be defined as in subSection 3.1 with π being the sparsity prior defined in (3.1). Moreover, assume that $\max_j \|X_j\|^2 \le B^2$ for some $B > 0$ and assume that $M \ge 3$. Then, the following statements hold with probability at least $1-\delta$:*

(i) The Q-aggregate estimator $\hat\mu^Q$ with $\lambda = 12\sigma^2$ satisfies

$$\|\hat\mu^Q - \mu\|^2 \le \inf_{0\le q\le 1} \min_{\beta\in\mathbb{R}^p} \big\{\|\mathbb{X}\beta - \mu\|^2 + \varphi_{q,p}(\beta; 9\sigma, B)\big\}. \tag{3.8}$$

(ii) The aggregate with exponential weights $\hat\mu^{\mathsf{EW}}$ with $\lambda = 64\sigma^2$ satisfies

$$\|\hat\mu^{\mathsf{EW}} - \mu\|^2 \le \frac{5}{3} \inf_{0\le q\le 1} \min_{\beta\in\mathbb{R}^p} \big\{\|\mathbb{X}\beta - \mu\|^2 + \varphi_{q,p}(\beta; 16\sigma, B)\big\}, \tag{3.9}$$

where, in both cases, $\varphi_{q,p}$ is defined in (3.7).

Both (3.8) and (3.9) can be compared to the prediction rates over ℓ_q balls that were derived in [RWY11] where the setup is the following. First, it is assumed that the true mean μ in (2.1) is of the form $\mu = \mathbb{X}\beta^*$ for some $\beta^* \in \mathsf{B}_q(R), R > 0$ and that $B = \kappa\sqrt{n}$. In this case, it follows from Theorem 4 that with probability at least $1-\delta$, we have for any $\tilde\mu \in \{\hat\mu^Q, \hat\mu^{\mathsf{EW}}\}$ that

$$\max_{\beta^*\in\mathsf{B}_q(R)} \frac{1}{n}\|\tilde\mu - \mathbb{X}\beta^*\|^2 \le C_1\kappa^2 R^q \Big[\frac{\sigma^2}{\kappa^2}\frac{\log\Big(\frac{ep}{\delta}\big(\frac{\sigma}{R\kappa\sqrt{n}}\big)^q\Big)}{n}\Big]^{1-\frac{q}{2}} \vee C_2\sigma^2\log\big(\frac{ep}{\delta}\big),$$

for some numerical constants C_1, C_2. In their specific regime of parameters, our rates are of the same order as [RWY11, Theorem 4] and are therefore optimal in that range. However, we provide a better finite sample performance and explicit dependence in the confidence parameter δ as well as explicit constants that do not depend on q. In particular, our bounds are continuous functions of q on the whole closed interval $[0,1]$. More strikingly, unlike [RWY11] neither of the estimators $\hat\mu^Q, \hat\mu^{\mathsf{EW}}$ depends on q or R and yet they optimally adapt to these parameters. This remarkable phenomenon is even better illustrated in the context of *universal aggregation*.

3.3. Universal aggregation

In his original description of aggregation, Nemirovski [Nem00] introduced three types of aggregation to which three new types were added later [BTW07, Lou07, WPGY11]. All of these aggregation problems can be described in the following unified way. Given $M \geq 2$ deterministic vectors $\mu_1, \dots, \mu_M \in \mathbb{R}^n$ and a set $\Theta \subset \mathbb{R}^M$, the goal is to construct an aggregate $\tilde{\mu}$ such that

$$\|\tilde{\mu} - \mu\|^2 \leq \inf_{\theta \in \Theta} \|\mu_\theta - \mu\|^2 + C\Upsilon_{n,M}(\Theta), \quad C > 0 \tag{3.10}$$

with high probability and where the remainder term $\Upsilon_{n,M}(\Theta) > 0$ is as small as possible. To each of the six types of aggregation, corresponds a unique $\Theta \subset \mathbb{R}^M$ and a smallest possible $\Upsilon_{n,M}(\Theta)$ for which (3.10) holds. Such a $\Upsilon_{n,M}(\Theta)$ is called the *optimal rate of aggregation* (over Θ) [Tsy03]. The six types of aggregation all correspond to choices of Θ that are intersections of balls $\mathsf{B}_q(R)$ for various choices of q and R. They are summarized in Table 1. We add a new natural type of aggregation that we call D-ℓ_q *aggregation*, where, by analogy to D-linear and D-convex aggregation, we add to ℓ_q aggregation the constraint that θ must be D-sparse. In particular, D-convex aggregation introduced in [Lou07] can be identified to D-ℓ_1 aggregation.

Type of aggregation	Θ	Optimal rate
Model Selection [Nem00]	$\mathsf{B}_0(1) \cap \mathsf{B}_1(1)$	$\sigma^2 \log(\frac{M}{\delta})$
Convex [Nem00]	$\mathsf{B}_1(1)$	$\Big[\sigma B\sqrt{\overline{\log}(\frac{\sigma M}{\delta B})} \vee \sigma^2\,\overline{\log}(\frac{M}{\delta})\Big] \wedge \sigma^2 M \log(\frac{1}{\delta})$
Linear [Nem00]	$\mathsf{B}_0(M) = \mathbb{R}^M$	$\sigma^2 M \log(\frac{1}{\delta})$
D-linear [BTW07]	$\mathsf{B}_0(D)$	$\sigma^2 D \log(\frac{M}{\delta D})$
D-convex [Lou07]	$\mathsf{B}_0(D) \cap \mathsf{B}_1(1)$	$[\sigma B\sqrt{\overline{\log}(\frac{\sigma M}{\delta B})} \vee \sigma^2\,\overline{\log}(\frac{M}{\delta})] \wedge \sigma^2 D \log(\frac{M}{\delta D})$
ℓ_q [WPGY11]	$\mathsf{B}_q(R)$	$\Big[\sigma^{2-q} R^q B^q \big[\overline{\log}\,(\frac{M}{\delta}(\frac{\sigma}{BR})^q)\big]^{1-\frac{q}{2}} \vee \sigma^2\,\overline{\log}(\frac{M}{\delta})\Big] \wedge \sigma^2 M \log(\frac{1}{\delta})$
D-ℓ_q	$\mathsf{B}_0(D) \cap \mathsf{B}_q(R)$	$\Big[\sigma^{2-q} R^q B^q \big[\overline{\log}\,(\frac{M}{\delta}(\frac{\sigma}{BR})^q)\big]^{1-\frac{q}{2}} \vee \sigma^2\,\overline{\log}(\frac{M}{\delta})\Big] \wedge \sigma^2 D \log(\frac{M}{\delta D})$

TABLE 1

The seven types of aggregation and the corresponding choice of Θ. The range of parameters is $q \in (0,1), D \in [M], R > 0$. All numerical constants have been removed for clarity.

While most papers on the subject use different estimators for different aggregation problems [Nem00, Tsy03, RT07, Rig12], Bunea *et al.* [BTW07] were the first to suggest that one single estimator could solve several aggregation problems all at once and used the BIC estimator to obtain partial results in the form of weak oracle inequalities. More recently, Rigollet and Tsybakov [RT11] showed that the exponential screening estimator solved the first five types of aggregation all at once, without the knowledge of Θ. Using similar arguments, we now show that the Q-aggregate solves at once, all seven problems of aggregation described in Table 1, not only in expectation, but also with high probability.

Theorem 5. *Fix, $M \geq 3, n \geq 1, D \in [M], B \geq 1, q \in (0,1), R > 0$ and $\delta \in (0,1)$. Moreover, fix $\mu_1, \dots, \mu_M \in \mathbb{R}^n$ such that $\max_j \|\mu_j\|^2 \leq B^2$. Then, for $\lambda = 20\sigma^2$, the Q-aggregate estimator $\hat{\mu}^Q$ satisfies the following oracle inequalities simultaneously with probability at least $1-\delta$. For any $\Theta \in \{\mathsf{B}_0(1) \cap \mathsf{B}_1(1), \mathsf{B}_1(1), \mathbb{R}^M, \mathsf{B}_0(D), \mathsf{B}_0(D) \cap \mathsf{B}_1(1), \mathsf{B}_q(R), \mathsf{B}_0(D) \cap \mathsf{B}_q(R)\}$, it holds*

$$\|\hat{\mu}^Q - \mu\|^2 \leq \min_{\theta \in \Theta} \|\mu_\theta - \mu\|^2 + C\Upsilon_{n,M}(\Theta)\,, \quad C > 0\,. \tag{3.11}$$

where $\Upsilon_{n,M}(\Theta)$ is defined in Table 1.

Note that the rates in Table 1 are optimal in the sense of [Tsy03] for the most interesting ranges of parameters. Indeed, they match the most general lower bounds of [RT11, RWY11, WPGY11] apart from minor discrepancies that can be erased by placing appropriate assumptions on the range of parameters considered. It is not hard to see from our proofs where the ambiant dimension M can be replaced by the dimension of the linear span of $\mu_1, \dots \mu_M$ should appear in these bounds [RT11]. Since this is not the main focus of our paper, we choose not to have this dependence explicit in our bounds but in view of the similarity of our proof techniques and that of [RT11], it is clear that it can be made explicit whenever appropriate by a simple modification of the prior π.

Appendix A: Proofs of the main theorems

The following lemma is key to both of our theorems. It allows us to control the deviation of the empirical risk of any aggregate $\mu_{\hat\theta}$ around its true risk.

Lemma A.1. *Fix $\lambda \geq 12V\sigma^2$. Let $\mu_{\hat\theta} = \sum_{k\in[M]} \hat\theta_k \hat\mu_k$, where $\hat\theta \in \Lambda_M$ is any measurable function of Y. Then, for any $j \in [M]$ we have the following inequality with probability at least $1-\delta$,*

$$2\langle \xi, \mu_{\hat\theta} - \hat\mu_j\rangle - \lambda\mathcal{K}(\hat\theta,\pi) - \sum_{k\in[M]} \hat\theta_k C_k - \frac{8\sigma^2}{\lambda}\sum_{k\in[M]} \hat\theta_k \|\hat\mu_k - \hat\mu_j\|^2 \leq \lambda \log\big(\frac{1}{\delta}\big)\,.$$

Moreover,

$$\mathbb{E}\Big[2\langle \xi, \mu_{\hat\theta} - \hat\mu_j\rangle - \lambda\mathcal{K}(\hat\theta,\pi) - \sum_{k\in[M]} \hat\theta_k C_k - \frac{8\sigma^2}{\lambda}\sum_{k\in[M]} \hat\theta_k \|\hat\mu_k - \hat\mu_j\|^2\Big] \leq 0\,.$$

Proof. Let $\Delta_j = 2\langle \xi, \mu_{\hat\theta} - \hat\mu_j\rangle - \lambda\mathcal{K}(\hat\theta,\pi) - \sum_{k\in[M]} \hat\theta_k C_k$. Then we have

$$\begin{aligned}
&\mathbb{E}\Big[\exp\Big(\frac{\Delta_j}{\lambda} - \frac{8\sigma^2}{\lambda^2}\sum_{k\in[M]}\hat\theta_k\|\hat\mu_k-\hat\mu_j\|^2\Big)\Big]\\
=&\mathbb{E}\Big[\exp\Big(\sum_{k\in[M]}\hat\theta_k\big(\frac{2}{\lambda}\langle\xi,\hat\mu_k-\hat\mu_j\rangle - \log\big(\frac{\hat\theta_k}{\pi_k}\big) - \frac{C_k}{\lambda} - \frac{8\sigma^2}{\lambda^2}\|\hat\mu_k-\hat\mu_j\|^2\big)\Big)\Big]\\
\leq&\mathbb{E}\Big[\sum_{k\in[M]}\hat\theta_k\exp\Big(\frac{2}{\lambda}\langle\xi,\hat\mu_k-\hat\mu_j\rangle - \log\big(\frac{\hat\theta_k}{\pi_k}\big) - \frac{C_k}{\lambda} - \frac{8\sigma^2}{\lambda^2}\|\hat\mu_k-\hat\mu_j\|^2\big)\Big)\Big] \quad \text{(Jensen's ineq.)}\\
=&\mathbb{E}\Big[\sum_{k\in[M]}\pi_k\exp\Big(\frac{2}{\lambda}\langle\xi,\hat\mu_k-\hat\mu_j\rangle - \frac{C_k}{\lambda} - \frac{8\sigma^2}{\lambda^2}\|\hat\mu_k-\hat\mu_j\|^2\Big)\Big] \qquad \text{(A.1)}
\end{aligned}$$

Observe now that the decomposition (2.1) implies that $\hat\mu_k - \hat\mu_j = B_k\xi + v_k$ where $B_k = A_k - A_j$ and $v_k = B_k\mu + b_k - b_j$. It yields

$$\frac{2}{\lambda}\langle\xi,\hat\mu_k-\hat\mu_j\rangle - \frac{8\sigma^2}{\lambda^2}\|\hat\mu_k-\hat\mu_j\|^2 = \xi^\top\Big[\frac{2}{\lambda}B_k - \frac{8\sigma^2}{\lambda^2}B_k^\top B_k\Big]\xi + \xi^\top\Big[\frac{2}{\lambda}I_n - \frac{16\sigma^2}{\lambda^2}B_k^\top\Big]v_k - \frac{8\sigma^2}{\lambda^2}\|v_k\|^2\,, \quad \text{(A.2)}$$

where I_n denotes the identity matrix of $\mathbb{R}^n$. Next, using the symmetry of B_k together with the Cauchy-Schwarz inequality, we get

$$\mathbb{E}\Big[\exp\Big(\xi^\top\Big[\frac{2}{\lambda}B_k - \frac{8\sigma^2}{\lambda^2}B_k^2\Big]\xi + \xi^\top\Big[\frac{2}{\lambda}I_n - \frac{16\sigma^2}{\lambda^2}B_k\Big]v_k\Big)\Big] \leq \sqrt{P_1\cdot P_2}\,,$$

where,

$$P_1 = \mathbb{E}\Big[\exp\Big(\xi^\top\Big[\frac{4}{\lambda}B_k - \frac{16\sigma^2}{\lambda^2}B_k^2\Big]\xi\Big)\Big]\,, \quad P_2 = \mathbb{E}\Big[\exp\Big(\xi^\top\Big[\frac{4}{\lambda}I_n - \frac{32\sigma^2}{\lambda^2}B_k\Big]v_k\Big)\Big]\,.$$

To bound P_1, observe that since A_j and B_k^2 both have nonnegative eigenvalues, it holds

$$\xi^\top\Big[\frac{4}{\lambda}B_k - \frac{16\sigma^2}{\lambda^2}B_k^2\Big]\xi \leq \frac{4}{\lambda}\xi^\top A_k\xi = \frac{4}{\lambda}(U_k\xi)^\top D_k(U_k\xi)$$

where $A_k = U_k^\top D_k U_k$ is the singular value decomposition of A_k. In particular, the matrix U_k is orthogonal so that $Z_k = U_k\xi/\sigma \sim \mathcal{N}(0, I_n)$. Applying now Lemma C.1 yields

$$\mathbb{E}\Big[\exp\big(\frac{4\sigma^2}{\lambda}Z_k^\top D_k Z_k\big)\Big] \leq \exp\big(\frac{4\sigma^2}{\lambda}\,\mathsf{Tr}(D_k) + \frac{16\sigma^4\,\mathsf{Tr}(D_k^2)}{\lambda^2 - 8\lambda\sigma^2\|D_k\|_{\mathsf{op}}}\big) \leq \exp\Big(\frac{4\sigma^2}{\lambda}\,\mathsf{Tr}(D_k)\big(1 + \frac{4\sigma^2 V}{\lambda - 8\sigma^2 V}\big)\Big)\,,$$

where in the second inequality, we used the following inequalities: $\mathsf{Tr}(D_k^2) \le \mathsf{Tr}(D_k)\|D_k\|_{\mathsf{op}}$, $\|D_k\|_{\mathsf{op}} \le V$. Taking now $\lambda \ge 12\sigma^2 V$ yields

$$1 + \frac{4\sigma^2 V}{\lambda - 8\sigma^2 V} \le 2$$

so that $\sqrt{P_1} \le \exp(C_k/\lambda)$, where we recall that $C_k = 4\sigma^2\,\mathsf{Tr}(A_k)$ is defined in (2.5).

We now bound P_2. To that end, observe that it follows from standard properties of the moment generating function of ξ, (see, e.g., [Rig12, Lemma 6.1]) that

$$P_2 \le \exp\Big(\frac{8\sigma^2}{\lambda^2}\Big\|\Big(I_n - \frac{8\sigma^2}{\lambda}B_k\Big)v_k\Big\|^2\Big)$$

Observe now that the eigenvalues of B_k belong to $[-V, V]$ the eigenvalues of $I_n - \frac{8\sigma^2}{\lambda}B_k$ are in $[0, 2]$ as long as $\lambda \ge 8\sigma^2 V$. In particular, for $\lambda \ge 12\sigma^2 V$ as above, we get

$$\sqrt{P_2} \le \exp\Big(\frac{8\sigma^2}{\lambda^2}\|v_k\|^2\Big)$$

The bounds on $\sqrt{P_1}$ and $\sqrt{P_2}$ together with (A.1) and (A.2) yield

$$\mathbb{E}\Big[\exp\Big(\frac{\Delta_j}{\lambda} - \frac{8\sigma^2}{\lambda^2}\sum_{k\in[M]}\hat\theta_k\|\hat\mu_k - \hat\mu_j\|^2\Big)\Big] \le 1$$

The two statements of the lemma follow easily from this bound on the moment generating function using the same arguments as in [Rig12, Theorem 3.1]. Specifically, the statement with high probability follows from a Chernoff bound and the statement in expectation follows from the inequality $t \le e^t - 1$. □

A.1. Proof of Theorem 1

For any $\theta \in \Lambda_M$, define

$$\hat S(\theta) = \nu\sum_{k\in[M]}\theta_k\|Y - \hat\mu_k\|_2^2 + (1-\nu)\|Y - \mu_\theta\|_2^2\,,$$
$$S(\theta) = \nu\sum_{k\in[M]}\theta_k\|\mu - \hat\mu_k\|_2^2 + (1-\nu)\|\mu - \mu_\theta\|_2^2\,.$$

and observe that

$$\hat S(\theta) - S(\theta) = \|Y\|_2^2 - \|\mu\|_2^2 - 2\langle\xi, \mu_\theta\rangle\,.$$

It follows from the definition (2.4) of $\hat\theta$, that for any $\theta \in \Lambda_M$, it holds

$$\hat S(\hat\theta) + \sum_{k\in[M]}\hat\theta_k C_k + \lambda\mathcal{K}(\hat\theta, \pi) \le \hat S(\theta) + \sum_{k\in[M]}\theta_k C_k + \lambda\mathcal{K}(\theta, \pi)\,.$$

The above two displays yield that

$$S(\hat\theta) - S(\theta) \le \sum_{k\in[M]}(\theta_k - \hat\theta_k)C_k + 2\langle\xi, \hat\mu^Q - \mu_\theta\rangle + \lambda\mathcal{K}(\theta, \pi) - \lambda\mathcal{K}(\hat\theta, \pi)\,. \tag{A.3}$$

Observe first that

$$S(\hat\theta) - S(\theta) = (1-\nu)\big[\|\mu - \hat\mu^Q\|^2 - \|\mu - \mu_\theta\|^2\big] + \nu\sum_{k\in[M]}(\hat\theta_k - \theta_k)\|\mu - \hat\mu_k\|^2\,.$$

Fix $\beta \in (0,1)$ and take $\theta = (1-\beta)\hat{\theta} + \beta e_j$ where e_j denotes the jth vector of the canonical basis of $\mathbb{R}^M$. It yields

$$\|\mu - \hat{\mu}^Q\|^2 - \|\mu - \mu_\theta\|^2 = \beta\big[\|\mu - \hat{\mu}^Q\|^2 - \|\mu - \hat{\mu}_j\|^2\big] + \beta(1-\beta)\|\hat{\mu}^Q - \hat{\mu}_j\|^2 .$$

so that

$$\begin{aligned}\frac{1}{\beta}\big[S(\hat{\theta}) - S(\theta)\big] =& (1-\nu)\big[\|\mu - \hat{\mu}^Q\|^2 - \|\mu - \hat{\mu}_j\|^2\big] + (1-\nu)(1-\beta)\|\hat{\mu}^Q - \hat{\mu}_j\|^2 \\ &+ \nu \sum_{k\in[M]} \hat{\theta}_k \|\mu - \hat{\mu}_k\|^2 - \nu\|\mu - \hat{\mu}_j\|^2 .\end{aligned}$$

Together with the identity

$$\sum_{k\in[M]} \hat{\theta}_k \|m - \hat{\mu}_k\|^2 = \sum_{k\in[M]} \hat{\theta}_k \|\hat{\mu}^Q - \hat{\mu}_k\|^2 + \|\hat{\mu}^Q - m\|^2 , \tag{A.4}$$

applied for $m = \hat{\mu}_j$ and $m = \mu$ respectively, it yields

$$\begin{aligned}\frac{1}{\beta}\big[S(\hat{\theta}) - S(\theta)\big] = \|\mu - \hat{\mu}^Q\|^2 - \|\mu - \hat{\mu}_j\|^2 + (1-\nu)(1-\beta) \sum_{k\in[M]} \hat{\theta}_k \|\hat{\mu}_j - \hat{\mu}_k\|^2 \\ + \big(\nu - (1-\nu)(1-\beta)\big) \sum_{k\in[M]} \hat{\theta}_k \|\hat{\mu}^Q - \hat{\mu}_k\|^2 .\end{aligned} \tag{A.5}$$

Next, observe that,

$$2\langle \xi, \hat{\mu}^Q - \mu_\theta\rangle = 2\beta\langle \xi, \hat{\mu}^Q - \hat{\mu}_j\rangle , \quad \sum_{k\in[M]} (\theta_k - \hat{\theta}_k) C_k = \beta\Big[C_j - \sum_{k\in[M]} \hat{\theta}_k C_k\Big] ,$$

and by convexity,

$$\mathcal{K}(\theta, \pi) \le (1-\beta)\mathcal{K}(\hat{\theta}, \pi) + \beta \log\big(\frac{1}{\pi_j}\big) .$$

Substituting the above expressions into (A.3), together with (A.5) yields that

$$\begin{aligned}&\|\mu - \hat{\mu}^Q\|^2 - \|\mu - \hat{\mu}_j\|^2 \\ &\quad\le \Delta_j + C_j + \lambda \log\big(\frac{1}{\pi_j}\big) \\ &\qquad - (1-\nu)(1-\beta) \sum_{k\in[M]} \hat{\theta}_k \|\hat{\mu}_j - \hat{\mu}_k\|^2 - \big(\nu - (1-\nu)(1-\beta)\big) \sum_{k\in[M]} \hat{\theta}_k \|\hat{\mu}^Q - \hat{\mu}_k\|^2\end{aligned}$$

where

$$\Delta_j = 2\langle \xi, \hat{\mu}^Q - \hat{\mu}_j\rangle - \lambda\mathcal{K}(\hat{\theta}, \pi) - \sum_{k\in[M]} \hat{\theta}_k C_k$$

as in the proof of Lemma A.1. Letting $\beta \to 0$ yields

$$\begin{aligned}&\|\mu - \hat{\mu}^Q\|^2 - \|\mu - \hat{\mu}_j\|^2 \\ &\quad\le \Delta_j + C_j + \lambda \log\big(\frac{1}{\pi_j}\big) - (1-\nu) \sum_{k\in[M]} \hat{\theta}_k \|\hat{\mu}_j - \hat{\mu}_k\|^2 + (1-2\nu) \sum_{k\in[M]} \hat{\theta}_k \|\hat{\mu}^Q - \hat{\mu}_k\|^2 \\ &\quad\le \Delta_j + C_j + \lambda \log\big(\frac{1}{\pi_j}\big) - \min(\nu, 1-\nu) \sum_{k\in[M]} \hat{\theta}_k \|\hat{\mu}_j - \hat{\mu}_k\|^2 ,\end{aligned}$$

where the second inequality comes from (A.4) with $m = \hat{\mu}_j$ when $\nu \le 1-\nu$ (the case $\nu \ge 1-\nu$ is trivial). It follows from Lemma A.1 that

$$\Delta_j \le \frac{8\sigma^2}{\lambda} \sum_{k\in[M]} \hat{\theta}_k \|\hat{\mu}_k - \hat{\mu}_j\|^2 + \lambda \log\big(\frac{1}{\delta\pi_j}\big)$$

with probability at least $1-\delta$ simultaneously for all j when $\lambda \geq 12V\sigma^2$. Thus, taking $\lambda \geq \frac{8\sigma^2}{\min(\nu,1-\nu,2/(3V))}$, completes the proof of (2.7). The proof of (2.8) follows by replacing the last display with the corresponding bound in expectation from Lemma A.1.

A.2. Proof of Theorem 2

For any $\theta \in \Lambda_M$, define

$$\hat{S}(\theta) = \sum_{k\in[M]} \theta_k \|Y-\hat{\mu}_k\|^2, \quad S(\theta) = \sum_{k\in[M]} \theta_k \|\mu-\hat{\mu}_k\|^2,$$

and observe that

$$\hat{S}(\theta) - S(\theta) = \|Y\|^2 - \|\mu\|^2 - 2\langle \xi, \sum_{k\in[M]} \theta_k \hat{\mu}_k \rangle. \tag{A.6}$$

It follows from the definition (2.9) of $\hat{\theta}$, that for any $j \in [M]$, it holds

$$\hat{S}(\hat{\theta}) + \lambda \mathcal{K}(\hat{\theta}, \pi) \leq \hat{S}(e_j) + \lambda \log(\frac{1}{\pi_j}) + C_j - \sum_{k\in[M]} \hat{\theta}_k C_k \,,$$

where e_j denotes the jth vector of the canonical basis of $\mathbb{R}^M$. Together with (A.6) applied with $\theta = \hat{\theta}$ and $\theta = e_j$ respectively, and the identity

$$\sum_{k\in[M]} \hat{\theta}_k \|\mu-\hat{\mu}_k\|^2 = \|\mu-\hat{\mu}^{\mathsf{EW}}\|^2 + \sum_{k\in[M]} \hat{\theta}_k \|\hat{\mu}_k - \hat{\mu}^{\mathsf{EW}}\|^2,$$

it yields that for any $j \in [M]$, we have

$$\|\mu-\hat{\mu}^{\mathsf{EW}}\|^2 \leq \|\mu-\hat{\mu}_j\|^2 + \lambda \log\big(\frac{1}{\pi_j}\big) + C_j - \sum_{k\in[M]} \hat{\theta}_k \|\hat{\mu}_k - \hat{\mu}^{\mathsf{EW}}\|^2 + \Delta_j, \tag{A.7}$$

where $\Delta_j = 2\langle \xi, \hat{\mu}^{\mathsf{EW}} - \hat{\mu}_j \rangle - \lambda \mathcal{K}(\hat{\theta}, \pi) - \sum_{k\in[M]} \hat{\theta}_k C_k$.

For any $\lambda \geq 12V\sigma^2$, $\delta > 0$, Lemma A.1 yields that

$$\Delta_j \leq \frac{8\sigma^2}{\lambda} \sum_{k\in[M]} \hat{\theta}_k \|\hat{\mu}_k - \hat{\mu}_j\|^2 + \lambda \log\big(\frac{1}{\delta\pi_j}\big)\,,$$

with probability at least $1-\delta\pi_j$. Together with (A.7) the identity

$$\sum_{k\in[M]} \hat{\theta}_k \|\hat{\mu}_k - \hat{\mu}_j\|^2 = \sum_{k\in[M]} \hat{\theta}_k \|\hat{\mu}_k - \hat{\mu}^{\mathsf{EW}}\|^2 + \|\hat{\mu}^{\mathsf{EW}} - \hat{\mu}_j\|^2,$$

it yields

$$\|\mu-\hat{\mu}^{\mathsf{EW}}\|^2 \leq \|\mu-\hat{\mu}_j\|^2 + \frac{8\sigma^2}{\lambda}\|\hat{\mu}_j - \mu^{\mathsf{EW}}\|^2 + \lambda \log\big(\frac{1}{\delta\pi_j^2}\big) + C_j + \big(\frac{8\sigma^2}{\lambda} - 1\big) \sum_{k\in[M]} \hat{\theta}_k \|\hat{\mu}_k - \hat{\mu}^{\mathsf{EW}}\|^2$$

Recall that our assumptions imply that $\lambda > 16\sigma^2$ so that

$$\big(1 - \frac{16\sigma^2}{\lambda}\big)\|\mu-\hat{\mu}^{\mathsf{EW}}\|^2 \leq (1 + \frac{16\sigma^2}{\lambda})\|\mu-\hat{\mu}_j\|^2 + \lambda \log\big(\frac{1}{\delta\pi_j^2}\big) + C_j\,. \tag{A.8}$$

Next, observe that $(1-x)^{-1} = 1 + x(1-x)^{-1} \leq 1 + 4x/3$ for $x \in (0, 1/4)$ so, for $\lambda \geq 64\sigma^2$, we get

$$\|\mu-\hat{\mu}^{\mathsf{EW}}\|^2 \leq \big(1 + \frac{128\sigma^2}{3\lambda}\big)\|\mu-\hat{\mu}_j\|^2 + \frac{8\lambda}{3}\log\big(\frac{1}{\delta\pi_j}\big) + \frac{4C_j}{3}.$$

The proof is concluded by a union bound.

A.3. Proof of Theorem 3

Let $\bar{\beta} \in \mathbb{R}^p$ realize the minimum in the right-hand side of (3.2) and let $\bar{J} \subset [p]$ denote the support of $\bar{\beta}$. On the one hand, it follows from the Pythagorean identity that

$$\|\hat{\mu}_{\bar{J}} - \mu\|^2 = \|A_{\bar{J}}Y - \mu\|^2 = \|A_{\bar{J}}\mu - \mu\|^2 + \|A_{\bar{J}}\xi\|^2$$

Next, since $\|A_{\bar{J}}\xi\|^2 \sim \sigma^2\chi^2_{\mathsf{Rk}(\mathbb{X}_{\bar{J}})}$, if follows from Lemma C.1 together with the inequality $2\sqrt{ab} \le a+b$ valid for $a, b > 0$ that with probability at least $1-\delta/2$, we have

$$\|A_{\bar{J}}\xi\|^2 \le 2\sigma^2\,\mathsf{Rk}(\mathbb{X}_{\bar{J}}) + 3\sigma^2\log(2/\delta) \le 2\sigma^2|\bar{\beta}|_0 + 3\sigma^2\log(2/\delta)\,.$$

On the other hand, we get from Theorem 1 that with probability at least $1-\delta/2$, it holds

$$\|\hat{\mu}^Q - \mu\|^2 \le \|\hat{\mu}_{\bar{J}} - \mu\|^2 + C_{\bar{J}} + 2\lambda\log\big(\frac{2}{\pi_{\bar{J}}\delta}\big).$$

It can be shown [RT12] that

$$\log(\pi_{\bar{J}}^{-1}) \le 2|\bar{J}|\log\big(\frac{ep}{|\bar{J}|}\big) + \frac{1}{2} \le 2|\bar{\beta}|_0\log\big(\frac{ep}{|\bar{\beta}|_0}\big) + \frac{1}{2}\,.$$

and we also have that $C_{\bar{J}} = 4\sigma^2\,\mathsf{Rk}(\mathbb{X}_{\bar{J}}) \le 4\sigma^2|\bar{\beta}|_0$.

Putting everything together yields that with probability at least $1-\delta$, it holds

$$\begin{aligned}\|\hat{\mu}^Q - \mu\|^2 &\le \|A_{\bar{J}}\mu - \mu\|^2 + 6\sigma^2|\bar{\beta}|_0 + 4\lambda|\bar{\beta}|_0\log\big(\frac{ep}{|\bar{\beta}|_0}\big) + \lambda + (3\sigma^2+2\lambda)\log(2/\delta)\\ &\le \|A_{\bar{J}}\mu - \mu\|^2 + \big(5\lambda + 6\sigma^2\big)|\bar{\beta}|_0\log\big(\frac{2ep}{|\bar{\beta}|_0\delta}\big)\,.\end{aligned}$$

To conclude the proof of (3.2), it suffices to observe that $\|A_{\bar{J}}\mu - \mu\|^2 \le \|\mathbb{X}\bar{\beta} - \mu\|^2$.

The proof of (3.3) follows along the same lines.

A.4. Proof of Theorem 5

Replacing β by θ and $\mathbb{X}_j$ by μ_j in the proof of Theorem 3 leads to

$$\|\hat{\mu}^Q - \mu\|^2 \le \min_{\theta\in\mathbb{R}^M}\big\{\|\mu_\theta - \mu\|^2 + 78\sigma^2|\theta|_0\log\big(\frac{2eM}{|\theta|_0\delta}\big)\big\}\,.$$

The above display combined with Lemma 1 yields that for any $q \in (0,1), R > 0$,

$$\|\hat{\mu}^Q - \mu\|^2 \le \min_{\theta\in\mathbb{R}^M}\big\{\|\mu_\theta - \mu\|^2 + 78\sigma^2|\theta|_0\log\big(\frac{2eM}{|\theta|_0\delta}\big) \wedge \varphi_{q,M}(\theta; 9\sigma, B)\big\}\,,$$

where the function $\varphi_{q,M}$ is defined in (3.7). To complete the proof, if suffices that for any $\Theta \in \{\mathsf{B}_0(1)\cap\mathsf{B}_1(1), \mathsf{B}_1(1), \mathbb{R}^M, \mathsf{B}_0(D), \mathsf{B}_0(D)\cap\mathsf{B}_1(1), \mathsf{B}_q(R), \mathsf{B}_0(D)\cap\mathsf{B}_q(R)\}$, there exists $q \in (0,1)$ and $R > 0$ such that

$$\sup_{\theta\in\Theta}\big\{66\sigma^2|\theta|_0\log\big(\frac{2eM}{|\theta|_0\delta}\big) \wedge \varphi_{q,M}(\theta; 9\sigma, B)\big\} \le C\Delta_{n,M}(\Theta)\,.$$

In the rest of the proof, we treat each case separately. To that ends, write

$$\psi(\theta) = 78\sigma^2|\theta|_0\log\big(\frac{2eM}{|\theta|_0\delta}\big) \wedge \varphi_{q,M}(\theta; 9\sigma, B)\,.$$

MODEL SELECTION AGGREGATION. If $\Theta = \mathsf{B}_0(1)$, observe that for any $\theta \in \Theta$,

$$\psi(\theta) \le 78\sigma^2|\theta|_0 \log\big(\frac{2eM}{|\theta|_0\delta}\big) \le 78\sigma^2 \log(\frac{2eM}{\delta})\,.$$

ℓ_q AND CONVEX AGGREGATION. If $\Theta = \mathsf{B}_q(R)$ (and in particular, $\Theta = \mathsf{B}_1(1)$ for convex aggregation), observe that for any $\theta \in \Theta$,

$$\begin{aligned}\psi(\theta) &\le 78\sigma^2 M\log(2e/\delta) \wedge \varphi_{q,M}(\theta; 9\sigma, B)\\ &\le \Big[17(9\sigma)^{2-q}R^qB^q\Big[\overline{\log}\Big(\frac{eM}{\delta}\big(\frac{9\sigma}{BR}\big)^q\Big)\Big]^{1-\frac{q}{2}} \vee 198\sigma^2\,\overline{\log}\,\big(\frac{eM}{\delta}\big)\Big] \wedge 78\sigma^2 M\log\big(\frac{2e}{\delta}\big)\,.\end{aligned}$$

LINEAR AGGREGATION. If $\Theta = \mathbb{R}^M$, observe that for any $\theta \in \Theta$,

$$\psi(\theta) \le 78\sigma^2|\theta|_0 \log\big(\frac{2eM}{|\theta|_0\delta}\big) \le 78\sigma^2 M\log\big(\frac{2e}{\delta}\big)\,.$$

D-LINEAR AGGREGATION. If $\Theta = \mathsf{B}_0(D)$, observe that for any $\theta \in \Theta$,

$$\psi(\theta) \le 78\sigma^2|\theta|_0 \log\big(\frac{2eM}{|\theta|_0\delta}\big) \le 78\sigma^2 D\log\big(\frac{2eM}{D\delta}\big)\,.$$

D-CONVEX AGGREGATION. If $\Theta = \mathsf{B}_0(D) \cap \mathsf{B}_1(1)$, observe that for any $\theta \in \Theta$,

$$\begin{aligned}\psi(\theta) &\le 78\sigma^2|\theta|_0 \log\big(\frac{2eM}{|\theta|_0\delta}\big) \wedge \varphi_{1,M}(\theta; 9\sigma, B)\\ &\le \Big[153\sigma B\Big[\overline{\log}\Big(\frac{9eM\sigma}{\delta B}\Big)\Big]^{\frac{1}{2}} \vee 198\sigma^2\,\overline{\log}\,\big(\frac{eM}{\delta}\big)\Big] \wedge 78\sigma^2 D\log\big(\frac{2eM}{D\delta}\big).\end{aligned}$$

D-ℓ_q AND D-CONVEX AGGREGATION. If $\Theta = \mathsf{B}_0(D) \cap \mathsf{B}_q(R)$ (and in particular, $q = 1, R = 1$ for convex aggregation), observe that for any $\theta \in \Theta$,

$$\begin{aligned}\psi(\theta) &\le 78\sigma^2|\theta|_0 \log\big(\frac{2eM}{|\theta|_0\delta}\big) \wedge \varphi_{q,M}(\theta; 9\sigma, B)\\ &\le \Big[17(9\sigma)^{2-q}R^qB^q\Big[\overline{\log}\Big(\frac{eM}{\delta}\big(\frac{9\sigma}{BR}\big)^q\Big)\Big]^{1-\frac{q}{2}} \vee 198\sigma^2\,\overline{\log}\,\big(\frac{eM}{\delta}\big)\Big] \wedge 78\sigma^2 D\log\big(\frac{2eM}{D\delta}\big).\end{aligned}$$

Appendix B: A generalized Maurey argument

B.1. Decay of coefficients on ℓ_q-balls

For any $q > 0, \theta \in \mathbb{R}^M$, recall that $|\theta|_q$ denotes the ℓ_q-norm of θ and is defined by

$$|\theta|_q = \Big(\sum_{j\in[M]} |\theta_j|^q\Big)^{\frac{1}{q}}\,.$$

It is known [Joh11] that if $q < 1$, such balls contain sparse signals, in the sense that their coefficients decay at a certain polynomial rate. This is quantified by the following lemma that yields a much sharper result than the one obtained using weak ℓ_q-balls, especially for q close to 1.

Lemma B.1. *Fix $R > 0$ and $q \in (0, 1)$. For any $\theta \in \mathsf{B}_q(R)$, let $|\theta_{(1)}| \ge \ldots \ge |\theta_{(M)}|$ denote a non-increasing rearrangement of the absolute values of the coefficients of θ. Then for any integer m such that $1 \le m \le M$, it holds*

$$\sum_{j=m+1}^{M} |\theta_{(j)}| \le |\theta|_q m^{1-\frac{1}{q}}\,.$$

Proof. Let $\{v_j\}_{j\geq 1}$ be an infinite sequence such that $v_j = |\theta_{(j)}|$ for $j \in [M]$ and $v_j = 0$ for $j \geq M+1$. Next for any $k \geq 0$, let B_k denote the block of m consecutive integers defined by $B_k = \{km+1, \ldots, (k+1)m\}$ and observe that

$$\begin{aligned}
\sum_{j=m+1}^{M} |\theta_{(j)}| = \sum_{j\geq m+1} v_j &= \sum_{k\geq 1}\sum_{j\in B_k} v_j = \sum_{k\geq 1}\sum_{j\in B_k} (v_j^q)^{\frac{1}{q}} \\
&\leq \sum_{k\geq 1}\sum_{j\in B_k} \Big(\frac{1}{m}\sum_{i\in B_{k-1}} v_i^q\Big)^{\frac{1}{q}} \\
&= m^{1-\frac{1}{q}} \sum_{k\geq 1} \Big(\sum_{i\in B_{k-1}} v_i^q\Big)^{\frac{1}{q}} \\
&\leq m^{1-\frac{1}{q}} \Big(\sum_{k\geq 1}\sum_{i\in B_{k-1}} v_i^q\Big)^{\frac{1}{q}} \\
&= |\theta|_q m^{1-\frac{1}{q}}\,,
\end{aligned}$$

where in the last inequality, we use the fact that $a^p + b^p \leq (a+b)^p$ for any $a, b > 0, p \geq 1$. □

B.2. Proof of Lemma 1

We begin by an approximation bound *a la* Maurey on ℓ_q balls.

Lemma B.2. *Let $\mu_1, \ldots, \mu_M \in \mathbb{R}^M$ be such that $\max_j \|\mu_j\|^2 \leq B^2$. Then for any $\mu \in \mathbb{R}^M$, any q, θ, and any positive integer $m \leq M/2$, there exists $\theta^m \in \mathbb{R}^M$ such that $|\theta^m|_0 \leq 2m$ and*

$$\|\mu_{\theta^m} - \mu\|^2 \leq \|\mu_\theta - \mu\|^2 + B^2 |\theta|_q^2 m^{1-\frac{2}{q}}\,. \tag{B.9}$$

Proof. Fix $q \in (0,1]$ and $\theta \in \mathbb{R}^M$. Denote by $|\theta_{(1)}| \geq \ldots \geq |\theta_{(M)}| \geq 0$ a non-decreasing rearrangement of the absolute value of the coordinates of θ. Next, decompose the vector θ into $\theta = \alpha + \beta$ so that $\mu_\theta = \mu_\alpha + \mu_\beta$, where α and β have disjoint support and $\alpha \in \mathsf{B}_0(m)$ is supported by the m indices with the largest absolute coordinates of θ. Since $\theta \in \mathsf{B}_q$, it follows form Lemma B.1 that the ℓ_1-norm of $\beta = \theta - \alpha$ satisfies

$$|\beta|_1 = \sum_{j=m+1}^{M} |\theta_{(j)}| \leq |\theta|_q m^{1-\frac{1}{q}} =: r\,.$$

Therefore, $\beta \in r\mathsf{B}_1$. We now use Maurey's empirical method [Pis81] to find a m-sparse approximate of μ_β. Define a random vector $U \in \mathbb{R}^M$ with values in $\{0, \pm r\mu_1, \ldots, \pm r\mu_M\}$ by $P[U = r\mathrm{sign}(\beta_i)\mu_i] = |\beta_i|/r$ and $P[U = 0] = 1 - |\beta|_1/r$. Let $U_1, \ldots, U_m$ be *i.i.d.* copies of U and notice that $\mathbb{E}[U] = \mu_\beta$ and $\|U\| \leq r\max_j \|\mu_j\| \leq rB$. It yields,

$$\mathbb{E}\|\mu - \mu_\alpha - \frac{1}{m}\sum_{i=1}^{m} U_i\|^2 = \|\mu - \mu_\theta\|^2 + \frac{\mathbb{E}\|U - \mathbb{E}U\|^2}{m} \leq \|\mu - \mu_\theta\|^2 + \frac{(rB)^2}{m}.$$

Therefore there exists some realization μ_{θ^m} of the random vector $\mu_\alpha + \frac{1}{m}\sum_{i=1}^m U_i$ for which (B.9) holds and $|\theta^m|_0 \leq 2m$. □

We now return to the proof of Lemma 1. Define

$$A = \min_{\theta\in\mathbb{R}^M} \Big\{\|\mu_\theta - \mu\|^2 + \nu^2|\theta|_0 \log\big(\frac{2eM}{|\theta|_0\delta}\big)\Big\}\,.$$

Fix $\theta \in \mathbb{R}^M$ and define $m = \lceil x \rceil$ where

$$x = \frac{B^q|\theta|_q^q}{\nu^q}\big[\overline{\log}\big(\frac{eM\nu^q}{B^q|\theta|_q^q\delta}\big)\big]^{-\frac{q}{2}} > 0,$$

First, if $m > |\theta|_0/2$, we use the simple bound

$$A \le \|\mu_\theta - \mu\|^2 + \nu^2|\theta|_0 \log\big(\frac{2eM}{|\theta|_0\delta}\big) \ \le \|\mu_\theta - \mu\|^2 + 2\nu^2 m \overline{\log}\big(\frac{eM}{m\delta}\big).$$

Next, if $m \le |\theta|_0/2$, it follows from Lemma B.2 that there exists θ^m such that $|\theta^m|_0 \le 2m$ and

$$\begin{aligned} A &\le \|\mu_{\theta^m} - \mu\|^2 + 2\nu^2 m \log\big(\frac{eM}{m\delta}\big) \\ &\le \|\mu_\theta - \mu\|^2 + 2\nu^2 m \log\big(\frac{eM}{m\delta}\big) + B^2|\theta|_q^2 m^{1-\frac{2}{q}}\,. \end{aligned}$$

Therefore, whether $m > |\theta|_0/2$ or $m \le |\theta|_0/2$, it holds for any $\theta \in \mathbb{R}^M$,

$$A \le \|\mu_\theta - \mu\|^2 + 2\nu^2 m \log\big(\frac{eM}{m\delta}\big) + B^2|\theta|_q^2 m^{1-\frac{2}{q}}\,. \tag{B.10}$$

To control the right-hand side of (B.10), consider two cases for the value of x.
Case 1: If $x < 1$, we have $m = 1$ and we will show $B^2|\theta|_q^2 \le \nu^2\overline{\log}(eM/\delta)$. Indeed, if $B|\theta|_q \le \nu$, then this bound holds trivially and if $B|\theta|_q \ge \nu$, then $x \ge (B|\theta|_q/\nu)^q[\overline{\log}(eM/\delta)]^{-\frac{q}{2}}$. Together with $x < 1$, the last inequality implies that $B^2|\theta|_q^2 \le \nu^2\overline{\log}(eM/\delta)$. Therefore, in Case 1, we have

$$A \le \|\mu_\theta - \mu\|^2 + 3\nu^2(eM/\delta).$$

Case 2:If $x \ge 1$, then $x \le m \le 2x$. Together with the fact that $\overline{\log}\big(t\,\overline{\log}(t)\big) \le 2\,\overline{\log}(t)$, for any $t > 0$, it yields

$$\begin{aligned} 2\nu^2 m \overline{\log}\big(\frac{eM}{m\delta}\big) + B^2|\theta|_q^2 m^{1-\frac{2}{q}} &\le 4\nu^2 x \overline{\log}\big(\frac{eM}{2x\delta}\big) + B^2|\theta|_q^2 x^{1-\frac{2}{q}} \\ &\le 16\nu^{2-q}B^q|\theta|_q^q\Big[\overline{\log}\big(\frac{eM\nu^q}{B^q|\theta|_q^q\delta}\big)\Big]^{1-\frac{q}{2}} + \nu^{2-q}B^q|\theta|_q^q\Big[\overline{\log}\big(\frac{eM\nu^q}{B^q|\theta|_q^q\delta}\big)\Big]^{1-\frac{q}{2}} \\ &\le 17\nu^{2-q}B^q|\theta|_q^q\Big[\overline{\log}\big(\frac{eM\nu^q}{B^q|\theta|_q^q\delta}\big)\Big]^{1-\frac{q}{2}}\,. \end{aligned}$$

Putting the two cases together with (B.10), we get that

$$A \le \min_{\theta\in\mathbb{R}^M}\Big\{\|\mu_\theta - \mu\|^2 + \varphi_{q,M}(\theta;\nu,B)\Big\},$$

where

$$\varphi_{q,M}(\theta;\nu,B) \ = 3\nu^2\,\overline{\log}\big(\frac{eM}{\delta}\big) \vee 17\nu^{2-q}B^q|\theta|_q^q\Big[\overline{\log}\big(\frac{eM\nu^q}{B^q|\theta|_q^q\delta}\big)\Big]^{1-\frac{q}{2}}.$$

Appendix C: Technical lemmas

C.1. Deviations of a χ^2 distribution

Let us first recall Lemma 1 of [LM00] in a form that is adapted to our purpose. We omit its proof.

Lemma C.1. *Suppose $(Z_1, \cdots, Z_k)$ are i.i.d. standard Gaussian random variables. Let $a_1, \cdots, a_k$ be non-negative numbers and define $|a|_\infty = \max_{i\in[k]} a_i$, $|a|_2^2 = \sum_{i=1}^k a_i^2$. Let*

$$S = \sum_{i=1}^k a_i(Z_i^2 - 1).$$

Then for any u such that $0 < 2|a|_\infty u < 1$, it holds

$$\mathbb{E}\left[\exp\left(uS\right)\right] \le \exp\Big(\frac{|a|_2^2u^2}{1-2|a|_\infty u}\Big).$$

and for any $t > 0$,

$$\mathbb{P}(S > 2|a|_2\sqrt{t} + 2|a|_\infty t) \le e^{-t}\,.$$

References

[AL11] Pierre Alquier and Karim Lounici, *PAC-Bayesian bounds for sparse regression estimation with exponential weights*, Electron. J. Stat. **5** (2011), 127–145. MR2786484 (2012e:62240)

[Aud08] Jean-Yves Audibert, *Progressive mixture rules are deviation suboptimal*, Advances in Neural Information Processing Systems 20 (J.C. Platt, D. Koller, Y. Singer, and S. Roweis, eds.), MIT Press, Cambridge, MA, 2008, pp. 41–48.

[BRT09] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov, *Simultaneous analysis of Lasso and Dantzig selector*, Ann. Statist. **37** (2009), no. 4, 1705–1732. MR2533469

[BTW07] Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp, *Sparsity oracle inequalities for the Lasso*, Electron. J. Stat. **1** (2007), 169–194 (electronic). MR2312149 (2008h:62101)

[Cat99] O. Catoni, *Universal aggregation rules with exact bias bounds.*, Tech. report, Laboratoire de Probabilités et Modeles Aléatoires, Preprint 510., 1999.

[Cat04] Olivier Catoni, *Statistical learning theory and stochastic optimization*, Lecture Notes in Mathematics, vol. 1851, Springer-Verlag, Berlin, 2004, Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001. MR2163920 (2006d:62004)

[Coh66] Arthur Cohen, *All admissible linear estimates of the mean vector*, Ann. Math. Statist. **37** (1966), 458–463. MR0189164 (32 #6591)

[CT01] L. Cavalier and A. B. Tsybakov, *Penalized blockwise Stein's method, monotone oracles and sharp adaptive estimation*, Math. Methods Statist. **10** (2001), no. 3, 247–282, Meeting on Mathematical Statistics (Marseille, 2000). MR1867161 (2002i:62054)

[DRZ12] Dong Dai, Philippe Rigollet, and Tong Zhang, *Deviation optimal learning using greedy Q-aggregation*, Ann. Statist. **40** (2012), no. 3, 1878–1905.

[DS12] Arnak S. Dalalyan and Joseph Salmon, *Sharp oracle inequalities for aggregation of affine estimators*, Ann. Statist. **40** (2012), no. 4, 2327–2355. MR3059085

[DT07] Arnak S. Dalalyan and Alexandre B. Tsybakov, *Aggregation by exponential weighting and sharp oracle inequalities*, Learning theory, Lecture Notes in Comput. Sci., vol. 4539, Springer, Berlin, 2007, pp. 97–111. MR2397581

[DT08] A. Dalalyan and A.B. Tsybakov, *Aggregation by exponential weighting, sharp PAC-bayesian bounds and sparsity*, Machine Learning **72** (2008), no. 1, 39–61.

[FPRU10] Simon Foucart, Alain Pajor, Holger Rauhut, and Tino Ullrich, *The Gelfand widths of ℓ_p-balls for $0 < p \leq 1$*, J. Complexity **26** (2010), no. 6, 629–640. MR2735423 (2012b:41039)

[Gir08] Christophe Giraud, *Mixing least-squares estimators when the variance is unknown*, Bernoulli **14** (2008), no. 4, 1089–1107. MR2543587

[GN92] G. K. Golubev and M. Nussbaum, *Adaptive spline estimates in a nonparametric regression model*, Teor. Veroyatnost. i Primenen. **37** (1992), no. 3, 554–561. MR1214361

[Gru98] Marvin H. J. Gruber, *Improving efficiency by shrinkage*, Statistics: Textbooks and Monographs, vol. 156, Marcel Dekker Inc., New York, 1998, The James-Stein and ridge regression estimators. MR1608582 (99c:62196)

[JN00] Anatoli Juditsky and Arkadii Nemirovski, *Functional aggregation for nonparametric regression*, Ann. Statist. **28** (2000), no. 3, 681–712. MR1792783 (2001k:62059)

[Joh11] Iain M. Johnstone, *Gaussian estimation: Sequence and wavelet models*, Unpublished Manuscript., December 2011.

[LB06] Gilbert Leung and A.R. Barron, *Information theory and mixing least-squares regressions*, Information Theory, IEEE Transactions on **52** (2006), no. 8, 3396–3410.

[Lec07] Guillaume Lecué, *Optimal rates of aggregation in classification under low noise assumption*, Bernoulli **13** (2007), no. 4, 1000–1022. MR2364224 (2009c:62099)

[LM00] B. Laurent and P. Massart, *Adaptive estimation of a quadratic functional by model selection*, Ann. Statist. **28** (2000), no. 5, 1302–1338.

[LM09] Guillaume Lecué and Shahar Mendelson, *Aggregation via empirical risk minimization*, Probab. Theory Related Fields **145** (2009), no. 3-4, 591–613. MR2529440 (2010i:62114)

[LM12] ———, *General nonexact oracle inequalities for classes with a subexponential envelope*, Ann. Statist. **40** (2012), no. 2, 832–860. MR2933668

[Lou07] Karim Lounici, *Generalized mirror averaging and D-convex aggregation*, Math. Methods Statist. **16** (2007), no. 3, 246–259. MR2356820 (2009h:62052)

[LR14] Guillaume Lecué and Philippe Rigollet, *Optimal learning with q-aggregation*, Ann. Statist. **42** (2014), no. 1, 211–224.

[Nem00] Arkadi Nemirovski, *Topics in non-parametric statistics*, Lectures on probability theory and statistics (Saint-Flour, 1998), Lecture Notes in Math., vol. 1738, Springer, Berlin, 2000, pp. 85–277. MR1775640 (2001h:62074)

[Pin80] M. S. Pinsker, *Optimal filtration of square-integrable signals in Gaussian noise*, Probl. Inf. Transm. (Russian) **16** (1980), no. 2, 52–68. MR624591 (82j:93048)

[Pis81] G. Pisier, *Remarques sur un résultat non publié de B. Maurey*, Seminar on Functional Analysis, 1980–1981, École Polytech., Palaiseau, 1981, pp. Exp. No. V, 13. MR659306 (83h:46026)

[Rig12] Philippe Rigollet, *Kullback-Leibler aggregation and misspecified generalized linear models*, Ann. Statist. **40** (2012), no. 2, 639–665. MR2933661

[RT07] Ph. Rigollet and A. B. Tsybakov, *Linear and convex aggregation of density estimators*, Math. Methods Statist. **16** (2007), no. 3, 260–280. MR2356821 (2008m:62067)

[RT11] P. Rigollet and A. Tsybakov, *Exponential Screening and optimal rates of sparse estimation*, Ann. Statist. **39** (2011), no. 2, 731–771.

[RT12] ———, *Sparse estimation by exponential weighting*, Statistical Science **27** (2012), no. 4, 558–575.

[RWY11] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu, *Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls*, IEEE Trans. Inform. Theory **57** (2011), no. 10, 6976–6994. MR2882274 (2012k:62204)

[Ste56] Charles Stein, *Inadmissibility of the usual estimator for the mean of a multivariate normal distribution*, Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I (Berkeley and Los Angeles), University of California Press, 1956, pp. 197–206. MR0084922 (18,948c)

[Tsy03] A. B. Tsybakov, *Optimal rates of aggregation*, COLT, 2003, pp. 303–313.

[Tsy09] Alexandre B. Tsybakov, *Introduction to nonparametric estimation*, Springer Series in Statistics, Springer, New York, 2009, Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. MR2724359 (2011g:62006)

[WPGY11] Zhan Wang, Sandra Paterlini, Frank Gao, and Yuhong Yang, *Adaptive minimax estimation over sparse ℓ_q-hulls*, Arxiv:1108.1961 (2011).

[Yan99] Y. Yang, *Model selection for nonparametric regression*, Statistica Sinica **9** (1999), 475–500.

[Yan04] Yuhong Yang, *Aggregating regression procedures to improve performance*, Bernoulli **10** (2004), no. 1, 25–47. MR2044592 (2005b:62145)