Centre de recherche en économie et statistique

Série des Documents de Travail

n° 2015-06

Optimal bounds for aggregation of affine estimators

P.Bellec[1]

[1] CREST, ENSAE, UMR CNRS 9194. E-mail : pierre.bellec@ensae.fr

Optimal bounds for aggregation of affine estimators

Pierre C. Bellec [*†‡]

June 30, 2015

Abstract

This paper deals with aggregation of estimators in the context fixed design regression, with heteroscedastic and subgaussian noise. We derive sharp oracle inequalities in deviation for model selection type aggregation of affine estimators when the noise is subgaussian. Explicit numerical constants are given for Gaussian noise and the procedure is robust to variance misspecification. Then we present a new concentration result that is sharper than the Hanson-Wright inequality under the Bernstein condition on the noise. This allows us to improve the sharp oracle inequality obtained in the subgaussian case. Finally, we show that up to numerical constants, the optimal sparsity oracle inequality previously obtained for Gaussian noise holds in the subgaussian case. The exact knowledge of the variance of the noise is not needed to construct the estimator that satisfies the sparsity oracle inequality.

1 Introduction

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

$$Y_i = f_i + \xi_i, \qquad i = 1, ..., n, \tag{1.1}$$

where the noise random variables $\xi_1, ..., \xi_n$ are zero-mean, subgaussian random variables. We measure the quality of estimation of the unknown vector $\mathbf{f}$ with the squared euclidean norm in $\mathbf{R}^n$:

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

for any estimator $\hat{f}$ of $\mathbf{f}$. When the noise random variables are normal, this is the Gaussian sequence model, which has been extensively studied [30].

*CREST-ENSAE, UMR CNRS 9194, 3 avenue Pierre Larousse, 92245 Malakoff Cedex, France

†CMAP, Ecole Polytechnique, 91120 Palaiseau, France

‡This work was supported by the grant Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047).

Two related statistical problems are tackled in the present paper: sparsity oracle inequalities and aggregation of affine estimators. Given a design matrix $\mathbb{X}$ with p columns, an estimator $\hat{\mu}$ of $\mathbf{f}$ is said to achieve a sparsity oracle inequality if it satisfies

$$\|\hat{\mu} - \mathbf{f}\|_2^2 \leq \min_{\theta \in \mathbf{R}^p} \left(C \|\mathbb{X}\theta - \mathbf{f}\|_2^2 + \Delta(|\theta|_0) \right), \tag{1.2}$$

with high probability or in expectation. In (1.2), $C \geq 1$, $|\theta|_0$ is the number of non-zero coefficients of θ and Δ is an increasing function, which may also depend on problem parameters such as the variance of the noise or the design matrix $\mathbb{X}$. See (1.3) below for a typical example of such function $\Delta(\cdot)$. Results similar to (1.2) are of mainstream interest in theoretical statistics, in particular when the number of covariates p exceed the number of observations. First approach to get such results can be found in [21], and in an expanded form in [8, 9], under Gaussian noise with a leading constant $C > 1$. The drawback of having $C > 1$ cannot be repaired for these penalized model selection procedures [23, Section 6.4.2 and Proposition 6.1]. More recently, aggregation methods based on exponential weights [36, 35, 41] and then Q-aggregation [14] were shown to achieve *sharp* oracle inequalities similar to (1.2). Here, *sharp* means that the oracle inequality has leading constant $C = 1$. These sharp oracle inequalities were proved for Gaussian noise with known variance. In Section 4.2, we propose a new aggregation method that satisfies the optimal sharp oracle inequality, under Subgaussian noise (the Gaussianity assumption is relaxed), and only an upper bound on the variance is needed to construct the estimator whereas previous methods require the exact knowledge of the variance.

The second problem tackled in this paper is the aggregation of affine estimators. Several estimators have been proposed to recover the unknown vector $\mathbf{f}$ from the observations: the Ordinary Least Squares, the Ridge regressors, 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 strengths and weaknesses in different scenarios, so it is important be able to mimic the best among a given family of estimators, without any assumption on the unknown regression vector $\mathbf{f}$. The problem of mimicking the best estimator in a given finite set is the problem of model-selection type aggregation, which was introduced in [34, 42]. More precisely, let $\hat{\mu}_1, ..., \hat{\mu}_M$ be M estimators of $\mathbf{f}$ based on the data $Y_1, ..., Y_n$. The goal is to construct a new estimator or aggregate $\hat{f}$ with the same data $Y_1, ..., Y_n$, which satisfies with probability greater than $1 - \epsilon$ the sharp oracle inequality

$$\left\| \hat{f} - \mathbf{f} \right\|_2^2 \leq \min_{j=1,...,M} \|\hat{\mu}_j - \mathbf{f}\|_2^2 + \delta_{n,M}(\epsilon),$$

where $\delta_{n,M}(\cdot)$ is a function of ϵ that should be small.

A first approach to mimic the best estimator in a given family is to use independence by assuming that the estimators $\hat{\mu}_1, ..., \hat{\mu}_n$ are independent of the observations $Y_1, ..., Y_n$ used for the aggregation step. For example, assume that two independent samples $(Y_1, ..., Y_n)$

and $(Y'_1, ..., Y'_n)$ are available, with Y_i and Y'_i independent and identically distributed for all $i = 1, ..., n$. Then one can use the sample $Y_1, ..., Y_n$ to construct the estimators $\hat{\mu}_1, ..., \hat{\mu}_M$ and use the independent sample $Y'_1, ..., Y'_n$ to aggregate them. For the aggregation step, conditionally on $Y_1, ..., Y_M$, the estimators $\hat{\mu}_1, ..., \hat{\mu}_M$ can be considered deterministic, thanks to independence. It is possible to obtain such independent samples when the noise is Gaussian and the variance is known, with *sample cloning* [41, Lemma 2.1], at the cost of a factor 2 in the variance of the observations. However, this technique is specific to the Gaussian case and cannot be used when the noise is only assumed to be subgaussian as in the present paper.

Among the procedures available to estimate $\mathbf{f}$, several are linear in the observations $Y_1, ..., Y_n$. It is the case for example of the Least Squares and the Ridge regressors, whereas the shrinkage estimators and the Stein estimator are non-linear functions of the observations. A description of the estimators that are linear or affine in the observations is given in [15, Section 1.2], [1] and references therein. This linear behavior of the estimators $\hat{\mu}_1, ..., \hat{\mu}_M$ makes it possible to explicitly treat the dependence between the estimators $\hat{\mu}_1, ..., \hat{\mu}_M$ and the data $Y_1, ..., Y_n$ used to aggregate them. Leung and Barron [32] studied the problem of aggregation of projection estimators, and derived sharp oracle inequalities in expectation with a procedure based on exponential weights. Then, Dalalyan and Salmon [15] and Dai et al. [14] gave insights on how to construct an aggregate to mimic the best candidate among a set or affine estimators. Here we also consider affine estimators. Let $\mathbf{y} = (Y_1, ..., Y_n)^T$ be the vector of observations. An affine estimator is of the form $\hat{\mu}_j = A_j \mathbf{y} + b_j$ for a deterministic matrix A_j of size $n \times n$ and a deterministic vector $b_j \in \mathbf{R}^n$.

We consider in Section 3 that the variances of the noise random variables $\xi_1, ..., \xi_n$ are known and in Section 4.2 that an upper bound on the subgaussian norm of the noise vector is known. We refer the reader to [24] and the survey [25] for the problem of estimating the unknown vector $\mathbf{f}$ when the variance of the noise is unknown, which is outside of the scope of the present paper.

As in the papers [15, 14], we consider the problem of aggregation of M affine estimators with a prior probability distribution $\pi_1, ..., \pi_M$ on the finite set of indices $\{1, ..., M\}$. Prior weights is a common ingredient in deriving sharp oracle inequalities for model-selection type aggregation [16, 13, 31, 5]. An example of such an oracle inequality is (1.4) below. The use of sparsity-inducing prior weights is crucial to prove sparsity oracle inequalities via sparsity pattern aggregation [36, 35, 14, 41]. When the noise is Gaussian with variance σ^2, the following sparsity oracle inequality was shown in [14] for an estimator $\hat{\mu}$ and a design matrix $\mathbb{X}$ with p columns: with probability greater than $1 - 2\exp(-x)$,

$$\|\hat{\mu} - \mathbf{f}\|_2^2 \leq \min_{\theta \in \mathbf{R}^p} \left(\|\mathbb{X}\theta - \mathbf{f}\|_2^2 + c\,\sigma^2 |\theta|_0 \log\left(\frac{ep}{1 \vee |\theta|_0} \right) \right) + c'\sigma^2 x, \tag{1.3}$$

where $c, c' > 0$ are absolute constants and $|\theta|_0$ denotes the number of non-zero coefficients of θ. A similar result in expectation was shown in [36, 41], also with the assumption that the noise random variables are normal with known variance. In Section 4.2, we propose an

estimator that achieves a similar sparsity oracle inequality in deviation, but we only assume that the noise vector is subgaussian. It extends the previous results [36, 35, 14, 41] to the subgaussian setting and the statistician does not need to know the exact knowledge of the variance to construct the estimator.

The papers [15, 14] derived different procedures that satisfy sharp oracle inequalities for the problem of aggregation of affine estimators when the noise random variable are Gaussian. Dalalyan and Salmon [15] proposed an estimator $\hat{\mu}^{EW}$ based on exponential weights, for which a sharp oracle inequality holds in expectation:

$$\mathbb{E}\left\|\mathbf{f}-\hat{\mu}^{EW}\right\|_2^2 \leq \min_{j=1,\dots,M}\left(\mathbb{E}\left\|\hat{\mu}_j-\mathbf{f}\right\|_2^2+\beta\log\frac{1}{\pi_j}\right), \tag{1.4}$$

where β is a constant proportional to the largest variance of the noise random variables. This oracle inequality in expectation holds for $\hat{\mu}^{EW}$ under a commutativity assumption on the matrices A_j, which is enough to apply this oracle inequality to orthogonal projectors on a set of coordinates. In the case where the matrices A_j are not symmetric, [15] 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 $\hat{\mu}^{EW}$ achieves this inequality in expectation, it was shown in [13] that this procedure cannot achieve a similar result in deviation, with an unavoidable error term of order $\sqrt{n}$. In Dai et al. [14], a sharp oracle inequality in deviation is derived for an estimator $\hat{\mu}^{Q}$ based on Q-aggregation [13]. Namely, the estimator $\hat{\mu}^{Q}$ satisfies with probability greater than $1-\delta$:

$$\left\|\mathbf{f}-\hat{\mu}^{Q}\right\|_2^2 \leq \min_{j=1,\dots,M}\left(\left\|\hat{\mu}_j-\mathbf{f}\right\|_2^2+4\sigma^2\mathrm{Tr}(A_j)+\beta\log\frac{1}{\pi_j}\right)+\beta\log\frac{1}{\delta}, \tag{1.5}$$

where β is a constant and the noise random variables are i.i.d. with variance σ^2. This bound shows that it is possible to achieve oracle inequalities in deviation in the context of aggregation of affine estimators. However the extra term $4\sigma^2\mathrm{Tr}(A_j)$ may be large in common situation where the trace of some matrices A_j is large. For example, if one aggregates the estimators $\hat{\mu}_1=\lambda_1\mathbf{y},\dots,\hat{\mu}_M=\lambda_M\mathbf{y}$, for some positive real numbers $\lambda_1,\dots,\lambda_M$ with the uniform prior $\pi_j=1/M$ for all $j=1,\dots,M$, then the remainder term $4\sigma^2\mathrm{Tr}(A_j)$ in the above oracle inequality is of order $\sigma^2 n\lambda_j$ for each $j=1,\dots,M$, which is large relatively to the optimal rate $\sigma^2\log M$. This term $4\sigma^2\mathrm{Tr}(A_j)$ makes the previous oracle inequality suitable only for scenarios where the matrices A_j have small trace.

The contributions of the present paper are the following:

- We propose in Theorem 3.1 an estimator that satisfies a sharp oracle inequality in deviation. The remainder term only involves $\log(\pi_j^{-1})$, as opposed to (1.5) where the remainder term has an extra term proportional to $\sigma^2\mathrm{Tr}(A_j)$. Thus our estimator is suitable for situations involving matrices A_j with large trace, and it recovers the optimal rate proportional to $\log M$ when the uniform prior is used. The assumptions

on the matrices $A_1, ..., A_M$ are relaxed. In particular, they can be non-symmetric and have negative eigenvalues.

- In Theorem 3.2, we prove that this procedure is robust to variance misspecification.
- We show in Theorem 3.3 that the result of Theorem 3.1 can be extended to subgaussian noise. In earlier results [15, 14], only Gaussian noise was considered. The noise distributions under which Theorem 3.3 holds are given in Assumptions 3.1, 3.2 and 3.3.
- In order to prove Theorem 3.3 under Assumption 3.3, we derive in Theorem 3.4 a new concentration result for quadratic forms of independent random variables. It is sharper than the Hanson-Wright inequality under Assumption 3.3.
- Using sparsity pattern aggregation, we derive a sparsity oracle inequality in deviation when the noise vector is subgaussian, without assuming independence of the noise components. Theorem 4.1 recovers up to absolute constants the sparsity oracle inequality obtained when the noise is Gaussian [36, 35, 14].

The paper is organized as follows. In Section 2 we define the notation used throughout the paper. Section 3 defines an estimator and shows that it achieves sharp oracle inequalities in deviation for aggregation of affine estimators under three different assumptions on the noise. In Section 4, we derive sparsity oracle inequalities. The concentration inequalities used in the paper are given in Appendix A and the proofs are given in Appendix B.

2 Notation

We study an aggregation problem for the regression model with fixed design and heteroscedastic subgaussian noise. A random variable X is said to be subgaussian if and only if the quantity

$$\|X\|_{\psi_2} = \sup_{p \geq 1} p^{-1/2} \left(\mathbb{E}|X|\right)^{1/p}$$

is finite. Several other definitions are used in the literature, see [43, Section 5.2.3] for a review of their equivalence.

Let $(f_1, ..., f_n)^T \in \mathbf{R}^n$ be an unknown regression vector. We observe n random variables (1.1) where $\xi_1, ..., \xi_n$ are subgaussian random variables, with $\mathbb{E}[\xi_i] = 0$ and $\mathbb{E}[\xi_i^2] = \sigma_i^2$. The model is heteroscedastic, which means that the random variables $\xi_1, ..., \xi_n$ may have different variances. It can be rewritten in the vector form $\mathbf{y} = \mathbf{f} + \boldsymbol{\xi}$ where $\mathbf{y} = (Y_1, ..., Y_n)^T$, $\mathbf{f} = (f_1, ..., f_n)^T$ and $\boldsymbol{\xi} = (\xi_1, ..., \xi_n)^T$.

For any estimator $\hat{f}_n$ of $\mathbf{f}$, we measure the quality of estimation of $\mathbf{f}$ with the loss $\|\mathbf{f} - \hat{f}_n\|_2^2$ where $\|\cdot\|_2$ denotes the Euclidean norm in $\mathbf{R}^n$. Let $M \geq 2$. As in [15, 14], we consider M affine estimators of the form

$$\hat{\mu}_j = A_j \mathbf{y} + b_j, \qquad j = 1, ..., M.$$

The matrices $A_1, ..., A_M$ and the vectors $b_1, ..., b_M \in \mathbf{R}^n$ are deterministic. Define the simplex in $\mathbf{R}^M$:

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

and for any $\theta \in \Lambda^M$, let $\hat{\mu}_\theta = \sum_{j=1}^M \theta_j \hat{\mu}_j$. Let $e_1, ..., e_M$ be the vectors of the canonical basis in $\mathbf{R}^M$. Then $\hat{\mu}_j = \hat{\mu}_{e_j}$ for all $j = 1, ..., M$.

Finally, for any $n \times n$ real matrix $A = (a_{i,j})_{i,j=1,...,n}$, define the operator norm of A, the Hilbert-Schmidt (or Frobenius) norm of A and the nuclear norm of A respectively by:

$$|||A|||_2 = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}, \quad \|A\|_{\mathrm{HS}} = \sqrt{\sum_{i,j=1,...,n} a_{i,j}^2}, \quad \|A\|_1 = \mathrm{Tr}\left(\sqrt{A^T A}\right).$$

3 Model-selection type oracle inequalities

3.1 A penalized procedure over the simplex

For any $\theta \in \Lambda^M$ define

$$\hat{H}_n(\theta) = \|\hat{\mu}_\theta\|_2^2 - 2\mathbf{y}^T \hat{\mu}_\theta + 2\sum_{j=1}^M \theta_j \mathrm{Tr}(D_\sigma A_j D_\sigma) \tag{3.1}$$

$$+ \tfrac{1}{2}\, \widehat{\mathrm{pen}}(\theta) + \beta \sum_{j=1}^M \theta_j \log \frac{1}{\pi_j},$$

where $\beta > 0$ is a constant, $D_\sigma = \mathrm{diag}(\sigma_1, ..., \sigma_n)$ and

$$\widehat{\mathrm{pen}}(\theta) = \sum_{j=1}^M \theta_j \, \|\hat{\mu}_\theta - \hat{\mu}_j\|_2^2 . \tag{3.2}$$

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

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \Lambda^M} \hat{H}_n(\theta). \tag{3.3}$$

When θ is fixed and deterministic, the term

$$\|\hat{\mu}_\theta\|_2^2 - 2\mathbf{y}^T \hat{\mu}_\theta + 2\sum_{j=1}^M \theta_j \mathrm{Tr}(D_\sigma A_j D_\sigma) \tag{3.4}$$

in the definition of $\hat{H}_n$ is an unbiased estimate of the quantity

$$\|\hat{\mu}_\theta\|_2^2 - 2\mathbf{f}^T \hat{\mu}_\theta = \|\hat{\mu}_\theta - \mathbf{f}\|_2^2 - \|\mathbf{f}\|_2^2 , \tag{3.5}$$

which is the quantity of interest $\|\hat{\mu}_\theta - \mathbf{f}\|_2^2$ up to the additive constant $\|\mathbf{f}\|_2^2$. The term involving the trace of the matrices $D_\sigma A_j D_\sigma$ comes from the quadratic term in $\boldsymbol{\xi}$:

$$\sum_{j=1}^{M} \theta_j \mathrm{Tr}(D_\sigma A_j D_\sigma) = \mathbb{E}[\sum_{j=1}^{M} \theta_j \boldsymbol{\xi}^T A_j \boldsymbol{\xi}] = \mathbb{E}[\boldsymbol{\xi}^T \hat{\mu}_\theta].$$

The estimators from [32, 15, 14] are all obtained with an unbiased estimate of the quantity (3.5), so the term (3.4) comes as no surprise in the definition of $\hat{H}_n$.

The penalty (3.2) 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 [13, 31, 5]. Since the estimators $\hat{\mu}_1, ..., \hat{\mu}_M$ depend on the data, the penalty (3.2) is data-driven, which is not the case when the estimators to aggregate are deterministic vectors as in [13]. In order to give some geometric insights on the penalty (3.2), let $c \in \mathbf{R}^n$ satisfies the M linear equations $2c^T\hat{\mu}_j = \|\hat{\mu}_j\|_2^2$ and assume only in the rest of this paragraph that c is well defined, even though this assumption cannot be fulfilled for $M > n$. Then

$$\widehat{\mathrm{pen}}(\theta) = \sum_{j=1}^{M} \theta_j \|\hat{\mu}_j\|_2^2 - \|\hat{\mu}_\theta\|_2^2 = 2c^T\hat{\mu}_\theta - \|\hat{\mu}_\theta\|_2^2 = \|c\|_2^2 - \|\hat{\mu}_\theta - c\|_2^2 \,. \tag{3.6}$$

Assume also only in this paragraph that the function $\theta \to \hat{\mu}_\theta$ is bijective from the simplex Λ^M to the convex hull of $\{\hat{\mu}_1, ..., \hat{\mu}_M\}$. Then we can write $\widehat{\mathrm{pen}}(\theta) = g(\hat{\mu}_\theta)$ for some function g defined on the convex hull of $\{\hat{\mu}_1, ..., \hat{\mu}_M\}$. Equation (3.6) shows that the level sets of the function g are euclidean balls centered at c. The function g is non-negative, it is minimal at the extreme points $\hat{\mu}_1, ..., \hat{\mu}_M$ since $g(\hat{\mu}_j) = 0$ for all $j = 1, ..., M$ and g is maximal at the projection of c on the convex hull of $\{\hat{\mu}_1, ..., \hat{\mu}_M\}$. Intuitively, the penalty (3.2) pushes θ away from the center of the simplex towards the vertices. Thus, the level sets of the function $\theta \to \widehat{\mathrm{pen}}(\theta)$ in $\mathbf{R}^M$ are ellipsoids centered at θ_c, where θ_c is the unique point in $\mathbf{R}^M$ such that $\hat{\mu}_{\theta_c} = c$. If $M > n$ or if the vector c is not well defined, the level sets of $\widehat{\mathrm{pen}}(\cdot)$ are more intricate and cannot be described as simply.

Finally, the term

$$\beta \sum_{j=1}^{M} \theta_j \log \frac{1}{\pi_j} \tag{3.7}$$

allows to weight the candidates $\hat{\mu}_1, ..., \hat{\mu}_M$ with the prior probability distribution $(\pi_j)_{j=1,...,M}$ based on some prior knowledge about the estimators $\hat{\mu}_1, ..., \hat{\mu}_M$. In many practical cases, no prior knowledge is available and uniform weights ($\pi_j = 1/M$ for all $j = 1, ..., M$) will be used, in that case the term (3.7) is constant which means that β does not appear in the definition of the procedure. Note that the prior probability distribution $(\pi_j)_{j=1,...,M}$ is deterministic and cannot depend on the data $Y_1, ..., Y_n$. For example, if the estimators are projection estimators, one can set prior weights that decrease with the rank of the projections [35], we use this strategy in Section 4.2. The same term is used in [31] whereas

[14] uses the Kullback-Leibler divergence of θ from π. It is shown in [13] that for aggregation of deterministic vectors, one may use a quantity of the form $\beta \sum_{j=1}^M \theta_j \log(\rho(\theta_j)/\pi_j)$ where $\rho(\cdot)$ satisfies $\rho(t) \geq t$ and $t \to t\log(\rho(t))$ is convex. This suggests that we could use the Kullback-Leibler divergence of θ from π instead of (3.7), but in their current form, our proofs only hold with the "linear entropy" (3.7).

Finally, notice that the function $\hat{H}_n$ is convex, as it has the form $\hat{H}_n(\theta) = \frac{1}{2}\|\hat{\mu}_\theta\|_2^2 + \mathrm{lin}(\theta)$ where $\mathrm{lin}(\cdot)$ is a linear function. This can be seen using (B.2) with $g = 0$. Thus minimizing $\hat{H}_n$ over the simplex is a quadratic program for which efficient algorithms are available. The convexity of $\hat{H}_n$ also proves that $\hat{\theta}$ is well defined, although it may not be unique (for example if all $\hat{\mu}_j$ are the same then $\hat{H}_n$ is constant on the simplex).

Under homoscedastic Gaussian noise, the estimator (3.3) becomes (3.8) below and satisfies the following oracle inequality. Theorem 3.1 is proved in Appendix B.2.

Theorem 3.1. *Let $M \geq 2$. For $j = 1, ..., M$, consider the estimator $\hat{\mu}_j = A_j\mathbf{y} + b_j$ and assume that $|||A_j|||_2 \leq 1$. Let $(\pi_1, ..., \pi_M)^T \in \Lambda^M$. Assume that the noise random variables $\xi_1, ...\xi_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$. Let*

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \Lambda^M} \|\hat{\mu}_\theta - \mathbf{y}\|_2^2 + 2\sigma^2 \sum_{j=1}^M \theta_j \mathrm{Tr}(A_j) + \tfrac{1}{2}\,\widehat{\mathrm{pen}}(\theta) + 34\sigma^2 \sum_{j=1}^M \theta_j \log\frac{1}{\pi_j}. \tag{3.8}$$

Then for all $x > 0$, with probability greater than $1 - 2\exp(-x)$,

$$\|\hat{\mu}_{\hat{\theta}} - \mathbf{f}\|_2^2 \leq \min_{j=1,...,M} \left(\|\hat{\mu}_j - \mathbf{f}\|_2^2 + 68\sigma^2 \log\frac{1}{\pi_j} \right) + 34\sigma^2 x. \tag{3.9}$$

The sharp oracle inequality in deviation given in [14] presents an additive term proportional to $\sigma^2\mathrm{Tr}(A_j)$, as in (1.5). 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 $\sigma^2\mathrm{Tr}(A_j)$ are not meaningful for the problem of aggregation of affine estimators, and Theorem 3.1 improves upon the earlier result of [14]. When the uniform prior is used, i.e., $\pi_j = 1/M$ for all $j = 1, ..., M$, the sharp oracle inequality (3.17) matches the lower bound from [36, Proof of Theorem 5.3 with $S = 1$] showing that 3.17 is optimal in a minimax sense.

We relax all assumptions on the matrices $A_1, ..., A_M$, for instance they may be non-symmetric and have negative eigenvalues. Earlier works studied projection matrices [32], assumed some commutativity property of the matrices [15] or their symmetry and positive semi-definiteness [14]. Although it is shown in [12] that all admissible linear estimators are symmetric with non-negative eigenvalues, some linear estimators used in practice are not symmetric. For example, the last example of [15, Section 1.2] ("moving averages"), exhibits linear estimators that need not be symmetric: if two neighbors of the graph i, j have a different number of neighbours, then $a_{ij} \neq a_{ji}$. Our result also shows that the restrictions

on the matrices $A_1, ..., A_M$ present in [32, 15, 14] were not intrinsic to the problem of aggregation of affine estimators.

For clarity we assume in Theorem 3.1 that the operator norm of each matrix is bounded by 1. This assumption is not restrictive since all linear estimators of the form $A\mathbf{y}$ satisfy $|||A|||_2 \leq 1$ [12]. This assumption can be relaxed, as seen in Theorem 3.3 below.

The next section provides a similar result when the variance is not known, and Section 3.3 generalizes Theorem 3.1 to non-Gaussian noise distributions.

3.2 Robustness to variance misspecification

In order to construct the estimator (3.8), the knowledge of the variance of the noise is needed. However, the following proposition shows that the procedure (3.8) is robust to variance misspecification, i.e., the result 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 3.2 (Aggregation under variance misspecification)**.** *Let $M \geq 2$ and $\delta > 0$. For $j = 1, ..., M$, consider the estimator $\hat{\mu}_j = A_j\mathbf{y} + b_j$. Let $(\pi_1, ..., \pi_M)^T \in \Lambda^M$. Assume that the noise random variables $\xi_1, ...\xi_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$. Let $\hat{\sigma}^2$ be an estimator possibly constructed with the observation $\mathbf{y}$, and assume that*

$$\forall j = 1, ..., M,\ A_j = A_j^T = A_j^2, \qquad \delta := \mathbb{P}\left(|\sigma^2 - \hat{\sigma}^2| > \frac{1}{8}\sigma^2\right) < 1. \tag{3.10}$$

Let $\hat{\beta} = (448/7)\hat{\sigma}^2$ and let $\hat{\theta} = \operatorname{argmin}_{\theta \in \Lambda^M} W_n(\theta)$ where

$$W_n(\theta) := \|\hat{\mu}_\theta\|_2^2 - 2\mathbf{y}^T\hat{\mu}_\theta + 2\hat{\sigma}^2 \sum_{j=1}^M \theta_j \mathrm{Tr}(A_j) + \tfrac{1}{2}\,\widehat{\mathrm{pen}}(\theta) + \hat{\beta}\sum_{j=1}^M \theta_j \log\frac{1}{\pi_j}. \tag{3.11}$$

Then for all $x > 0$, with probability greater than $1 - \delta - 2\exp(-x)$,

$$\|\hat{\mu}_{\hat{\theta}} - \mathbf{f}\|_2^2 \leq \min_{j=1,...,M}\left(\|\hat{\mu}_j - \mathbf{f}\|_2^2 + 144\sigma^2 \log\frac{1}{\pi_j}\right) + 56\sigma^2 x. \tag{3.12}$$

The proof of Theorem 3.2 is given in Appendix B.3. In the condition (3.10), the matrices $A_1, ..., A_M$ are assumed to be orthogonal projectors, so Theorem 3.2 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. But broader families of matrices can be aggregated as well. By looking at the proof (in particular (B.13)), one can see that condition (3.10) can be replaced by

$$\gamma \leq \min_{j,k=1,...,M:\ \mathrm{Tr}(A_j - A_k) \neq 0} \frac{\|A_j - A_k\|_{\mathrm{HS}}^2}{|\mathrm{Tr}(A_j - A_k)|}, \qquad \delta := \mathbb{P}\left(|\sigma^2 - \hat{\sigma}^2| > \frac{\gamma}{8}\sigma^2\right) < 1,$$

for some $\gamma > 0$. Then by setting $\hat{\beta} = 448\hat{\sigma}^2/(7\gamma)$ in (3.11), the estimator $\hat{\mu}_{\hat{\theta}}$ satisfies a sharp oracle inequality similar to (3.12). Lemma B.1 ensures that the inequality on the left of the previous display holds with $\gamma = 1$.

The papers [24, 25, 3] aim at performing aggregation of Least Squares estimators when σ^2 is unknown, but unlike Theorem 3.2 the sharp oracle inequalities obtained have a leading constant greater than 1.

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

Example 3.1 (An independent estimator $\hat{\sigma}^2$ is accessible). In [15, Section 3.1], two contexts are given where an unbiased estimator of the covariance matrix, independent from $\mathbf{y}$, is available. First, 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. Second, it is shown that one can construct an estimator $\hat{\sigma}^2$ if the noise comes from the device: one can use a known signal in order to evaluate the noise.

Example 3.2 (Difference based esimators). In nonparametric regression where the non-random design points are equispaced 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 [26, 20], and is efficient when the design points lie in a multidimensional space [33]. For images, where the underlying space is 2-dimensional, efficient methods which require no multiplication are available [29].

Example 3.3 (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 [40] allows a joint estimation of the regression coefficients and the noise level σ^2. The estimator $\hat{\sigma}^2$ of the scaled LASSO converges in probability to the true noise level σ^2 [40, Theorem 1], and $\hat{\sigma}^2/\sigma^2$ is asymptotically normal [40, Equation (19)]. Second, [6] proposes to estimate σ^2 with with a recursive procedure that uses LASSO residuals, and non-asymptotic guarantees are proved [6, Supplementary material]. Third, [7] provides non-asymptotic bounds on the estimation of σ^2 by the residuals of the Square-Root LASSO [7, Theorem 2] and these bounds imply consistency. In Theorem 3.2, 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 [6, 40].

To our knowledge, Theorem 3.2 is the first aggregation result, with leading constant 1, that is robust to variance misspecification.

3.3 Robustness to non-Gaussian noise distributions

We state here the three different assumptions under which Theorem 3.1 can be generalized. The value of β given below is used in the construction of the estimator $\hat{\theta}$ defined in (3.3). The value of β depends on the assumption on the noise.

The constant $L > 0$ is independent of the noise and its role will be specified in Theorem 3.3. It can be chosen equal to $\sup_{j,k=1,\dots,M} |||A_j - A_k|||_2/2$ so its value is always

known by the practitioner. For example, for projection estimators $L = 1$ is a suitable choice.

Assumption 3.1 (Gaussian noise)**.** *Assume that the noise components $\xi_1, ...\xi_n$ are normal, independent, zero-mean, and ξ_i has variance σ_i^2. In this case, let*

$$\beta = (12 + 16L + 6L^2) \left(\max_{i=1,...,n} \sigma_i^2 \right). \tag{3.13}$$

Assumption 3.2 (Subgaussian noise)**.** *Let $K > 0$ and assume that the noise components $\xi_1, ..., \xi_n$ are independent, zero-mean, $\|\xi_i\|_{\psi_2} \leq K$ and ξ_i has variance σ_i^2. Here, let*

$$\beta = K^2 \left(c_{w_1}(2 + L)2L + 2c_h^2(1 + L)^2 + \tfrac{1}{2}c_{w_2}^2 \max_{i=1,...,n} \frac{\|\xi_i\|_{\psi_2}^2}{\sigma_i^2}(2 + L)^2 \right), \tag{3.14}$$

where c_{w_1}, c_{w_2} and c_h are the absolute constants given in Propositions A.2 and A.3.

Assumption 3.3 (Bernstein condition on $\xi_1^2, ..., \xi_n^2$)**.** *Let $K > 0$ and assume that the noise components $\xi_1, ..., \xi_n$ are independent and satisfy*

$$\forall p \geq 1, \qquad \mathbb{E}|\xi_i|^{2p} \leq \tfrac{1}{2}\, p!\, \sigma_i^2\, K^{2(p-1)}. \tag{3.15}$$

Here, let

$$\beta = 392 + 1408L + 608L^2. \tag{3.16}$$

Assumption 3.3 is the natural assumption to derive a Bernstein concentration inequality for the sum of random variables $\xi_1^2 + ... + \xi_n^2$. Although Assumption 3.3 is less common than Assumptions 3.1 and 3.2, its interest resides in the concentration inequality given in Theorem 3.4, which is sharper than the Hanson-Wright inequality. Under this assumption, it is possible to remove the expression $\max_{i=1,...,n} \|\xi_i\|_{\psi_2}/\sigma_i$ from the value of β.

Theorem 3.3. *Let $L > 0$ be a positive real number and $M \geq 2$. For $j = 1, ..., M$, consider the estimators $\hat{\mu}_j = A_j\mathbf{y} + b_j$ with $b_j \in \mathbf{R}^n$ and A_j a real matrix of size $n \times n$. Assume that the matrices $A_1, ..., A_M$ satisfy $|||A_j - A_k|||_2 \leq 2L$ for any j, k.*

Assume one of the Assumptions 3.1, 3.2 or 3.3 on the noise $\boldsymbol{\xi} = (\xi_1, ..., \xi_n)$ and set the value of β accordingly. Once β is set, let $\hat{\theta}$ be defined in (3.3). Then for all $x > 0$, with probability greater than $1 - 2\exp(-x)$,

$$\|\hat{\mu}_{\hat{\theta}} - \mathbf{f}\|_2^2 \leq \min_{j=1,...,M} \left(\|\hat{\mu}_j - \mathbf{f}\|_2^2 + 2\beta \log \frac{1}{\pi_j} \right) + \beta x. \tag{3.17}$$

In most practical cases where uniform prior weights are used, the value of β is not needed to construct the procedure (3.3) since the term (3.7) is constant. The proof of Theorem 3.3 is given in Appendix B.2.

Theorem 3.3 is a generalization of Theorem 3.1. Hence it also improves upon earlier results: all assumptions on the matrices are relaxed and the remainder term of the oracle

inequality is independent of the trace of the matrices $A_1, ..., A_M$, as opposed to the bound (1.5) proved in [14].

One of the contribution of the present paper is to provide a sharp oracle inequality such as (3.17) under Subgaussian noise. To our knowledge, (3.17) is the first result on sharp oracle inequality in deviation for model selection type aggregation obtained without assuming that the noise is Gaussian.

Under Assumption 3.2 (Subgaussian noise), our analysis leads to a remainder term that can be large for random variables that have pathologically small variance relatively to their subgaussian norm: β defined in (3.14) is proportional to $\max_{i=1,...,n} \|\xi_i\|_{\psi_2}/\sigma_i$. Under Assumption 3.3 which is slightly stronger and prevents the variance from being pathologically small, this issue can be fixed. We will come back to Assumption 3.3 in Section 3.5 below.

The constant β in the oracle inequality is of the order $K^2(1 \vee L^2)$, where K^2 is the supremum of the variances or the supremum of the squared subgaussian norms, and $2L$ upper bounds the operator norms of all $A_j - A_k$ for $j, k = 1, ..., M$. In most practical cases, L will be smaller than 1 since all admissible estimators of the form $A_j\mathbf{y}$ satisfy $|||A_j|||_2 \leq 1$ [12], thus the fact that β is proportional to $1 \vee L^2$ is not an issue. Interestingly, the operator norm of the matrices $A_1, ..., A_M$ does not appear in the sharp oracle inequality in expectation given in [15], while it plays a crucial role here. On the other hand, the factor K^2 may be more problematic, especially for heteroscedastic noise: β is proportional to the largest variance (resp. the largest subgaussian norm) even if most of the noise random variables have small variance (resp. small subgaussian norm).

3.4 Outline of the proof of Theorem 3.3

The following lemma shows that we can derive a sharp oracle inequality for the estimator $\hat{\mu}_{\hat{\theta}}$ by controlling the concentration of terms of the form $\boldsymbol{\xi}^T Q \boldsymbol{\xi}$ and $\boldsymbol{\xi}^T v$, where Q is a $n \times n$ deterministic matrix and v is a deterministic vector in $\mathbf{R}^n$. The following lemma proved in Appendix B.1.

Lemma 3.1. *Let $\hat{\theta}$ be defined in (3.3). Then almost surely,*

$$\|\hat{\mu}_{\hat{\theta}} - \mathbf{f}\|_2^2 \leq \min_{J^*=1,...,M} \left(\|\hat{\mu}_{J^*} - \mathbf{f}\|_2^2 + 2\beta \log \frac{1}{\pi_{J^*}} \right) + \max_{j,k=1,...,M} \zeta_{j,k}$$

where

$$\begin{aligned} \zeta_{j,k} = \; & \boldsymbol{\xi}^T Q_{j,k} \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q_{j,k} \boldsymbol{\xi}] \\ & + \boldsymbol{\xi}^T v_{j,k} \\ & - \beta \log \frac{1}{\pi_k \pi_j} - \tfrac{1}{2} \|(A_k - A_j) D_\sigma\|_{\mathrm{HS}}^2 - \tfrac{1}{2} \|(A_k - A_j)\mathbf{f} + b_k - b_j\|_2^2 , \end{aligned} \tag{3.18}$$

and the matrix D_σ, the matrices $Q_{j,k}$ and the vectors $v_{j,k}$ are defined by

$$\begin{aligned} D_\sigma &:= \mathrm{diag}(\sigma_1, ..., \sigma_n), \\ Q_{j,k} &:= 2(A_k - A_j) - \tfrac{1}{2}(A_k - A_j)^T(A_k - A_j), \qquad (3.19) \\ v_{j,k} &:= 2\left(I_{n\times n} - \tfrac{1}{2}(A_k - A_j)^T\right)((A_k - A_j)\mathbf{f} + b_k - b_j)\,. \qquad (3.20) \end{aligned}$$

In Appendix B.2, we prove Theorem 3.3 by applying Lemma 3.1 and controlling the concentration of terms of the form $\boldsymbol{\xi}^T Q_{j,k}\boldsymbol{\xi}$ and $\boldsymbol{\xi}^T v_{j,k}$ under the different Assumptions 3.1, 3.2 and 3.3.

A sketch of the proof of Theorem 3.3 under Assumption 3.1 (Gaussian Noise) goes as follows. The quantity $W_{j,k}^{\mathrm{linear}} := \frac{1}{2}\|(A_k - A_j)\mathbf{f} + b_k - b_j\|_2^2$ in (3.18) is of the order of the variance of $\boldsymbol{\xi}^T v_{j,k}$. Using (A.1) applied to $v = v_{j,k}$, it is shown that for all $t > 0$, with probability greater than $1 - \exp(-t)$,

$$\boldsymbol{\xi}^T v_{j,k} - W_{j,k}^{\mathrm{linear}} \leq \gamma\beta t,$$

where $\gamma \in (0, 1)$ and β is the constant given in (3.13). Similarly, the quantity $W_{j,k}^{\mathrm{quad}} := \frac{1}{2}\|(A_k - A_j)D_\sigma\|_{\mathrm{HS}}^2$ in (3.18) is of the order of the variance of $\boldsymbol{\xi}^T Q_{j,k}\boldsymbol{\xi}$. Using the concentration inequality (A.2) applied to $Q_{j,k}$, we prove that with probability greater than $1 - \exp(-t)$,

$$\boldsymbol{\xi}^T Q_{j,k}\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q_{j,k}\boldsymbol{\xi}] - W_{j,k}^{\mathrm{quad}} \leq (1 - \gamma)\beta t.$$

For fixed j and k, these concentration inequalities and the union bound lead to

$$\forall t > 0, \qquad \mathbb{P}\left(\zeta_{j,k} + \beta \log\frac{1}{\pi_j\pi_k} > \beta t\right) \leq 2\exp(-t).$$

Finally, the non-random term $-\beta\log\frac{1}{\pi_k\pi_j}$ is used to perform the union bound on $j, k = 1, ..., M$, such that for all $x > 0$,

$$\begin{aligned} \mathbb{P}\left(\max_{j,k=1,...,M}\zeta_{j,k} > \beta x\right) &\leq \sum_{j,k=1,...,M}\mathbb{P}\left(\zeta_{j,k} + \beta\log\frac{1}{\pi_j\pi_k} > \beta(x + \log\frac{1}{\pi_j\pi_k})\right) \\ &\leq \sum_{j,k=1,...,M}\pi_j\pi_k 2\exp(-x) = 2\exp(-x). \end{aligned}$$

The proof is similar under the two other assumptions 3.2 and 3.3, but different concentration inequalities are used. The proof of Lemma 3.1 can be found in Appendix B.1 and the proof of Theorem 3.3 is given in Appendix B.2.

3.5 Assumption 3.3: examples and concentration inequality

The goal of this section is to present the motivation behind Assumption 3.3 and to present the concentration inequality of Theorem 3.4. This concentration inequality is of independent interest as it provides sharper bounds than the Hanson-Wright inequality.

This assumption is sufficient to remove the quantity $\max_{i=1,...,n} \|\xi_i\|_{\psi_2}/\sigma_i$ from the expression (3.14) of β in the sharp oracle inequality of Theorem 3.3. It was the weakest assumption we could find that allowed us to remove the quantity $\max_{i=1,...,n} \|\xi_i\|_{\psi_2}/\sigma_i$.

Example 3.4. Centered variables almost surely bounded by K and zero-mean Gaussian random variables with variance smaller than K^2 satisfy (3.15).

Example 3.5 (Log-concave random variables). In [38], the authors consider a slightly stronger condition [38, Definition 1.1]. They consider random variables Z satisfying for any integer $p \geq 1$ and some constant K:

$$\mathbb{E}[|Z|^p] \leq \; p\, K\, \mathbb{E}[|Z|^{p-1}], \tag{3.21}$$

and they showed in [38, Section 7] that any distribution that is log-concave satisfies (3.21). Thus, if X^2 is log-concave then our assumption (3.15) holds. See [2, Section 6] for a comprehensive list of the common log-concave distributions.

The next theorem provides a concentration inequality for quadratic forms of independent random variables satisfying the moment assumption (3.15). It is sharper than the Hanson-Wright inequality given in Proposition A.3.

Theorem 3.4. *Assume that the noise random variable $\boldsymbol{\xi} = (\xi_1, ..., \xi_n)^T$ satisfies Assumption 3.3 for some $K > 0$. Let A be any $n \times n$ real matrix. Then for all $t > 0$,*

$$\mathbb{P}\left(\boldsymbol{\xi}^T A \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T A \boldsymbol{\xi}] > t\right) \leq \exp\left(-\min\left(\frac{t^2}{192K^2\,\|AD_\sigma\|_{\mathrm{HS}}^2}, \frac{t}{256K^2|||A|||_2}\right)\right), \tag{3.22}$$

where $D_\sigma = \mathrm{diag}(\sigma_1, ..., \sigma_n)$. Furthermore, for any $x > 0$, with probability greater than $1 - \exp(-x)$,

$$\boldsymbol{\xi}^T A \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T A \boldsymbol{\xi}] \leq 256K^2|||A|||_2 x + 8\sqrt{3}K\,\|AD_\sigma\|_{\mathrm{HS}}\sqrt{x}. \tag{3.23}$$

The proof of Theorem 3.4 is given in the supplementary material. A key ingredient to prove this concentration result is a decoupling inequality [19, 18]. A simple decoupling inequality for quadratic forms can be found in [44] or [22, Theorem 8.11], and we use this result in order to prove Theorem 3.4.

When t is small, the right hand side of (3.22) becomes

$$\exp\left(-\frac{t^2}{192K^2\,\|AD_\sigma\|_{\mathrm{HS}}^2}\right),$$

whereas the right hand side of the Hanson-Wright inequality (A.4) yields

$$\exp\left(-c\frac{t^2}{K^4\left\|A\right\|_{\mathrm{HS}}^2}\right),$$

for some absolute constant $c>0$. The element of the diagonal matrix D_σ are bounded by K, so Theorem 3.4 gives a sharper bound than the Hanson-Wright inequality in this regime. Under the moment assumption (3.15), we were able to remove the factor $\max_{i=1,\dots,n}(\|\xi_i\|_{\psi_2}/\sigma_i)$ using the concentration inequality from Theorem 3.4.

In particular, the sharp oracle inequality (3.17) with β given in (3.16) holds for all the noise distributions described in Examples 3.4 and 3.5.

3.6 Simulations

In this section we implement the procedure (3.3) on synthetic data and compare its performance with that of exponential weights estimator. The uniform prior is used.

Experiment 3.1 (Well specified least squares). Let $n=100$, $\mathbf{f}=2\sigma e_1$, $M=n$ and for all $j=1,\dots,M$, $\hat{\mu}_j=(e_j^T\mathbf{y})e_j$, i.e., $\hat{\mu}_j$ is the projection on the j-th coordinate.

Experiment 3.2 (Small misspecifications). Let $n=100$, $\mathbf{f}=2\sigma\sum_{i=1}^{10}e_i$. The set $\{\hat{\mu}_1,\dots,\hat{\mu}_M\}$ is $\cup_{r=0,2,4}\cup_{i=1}^{n-r}\{\sum_{k=i}^{i+r}(e_k^T\mathbf{y})e_k\}$ and $M=3n-6$.

Experiment 3.3 (Medium misspecifications). Let $n=100$, $\mathbf{f}=2\sigma\sum_{i=1}^{20}e_i$, M and $\hat{\mu}_1,\dots,\hat{\mu}_M$ are the same as in Experiment 3.2.

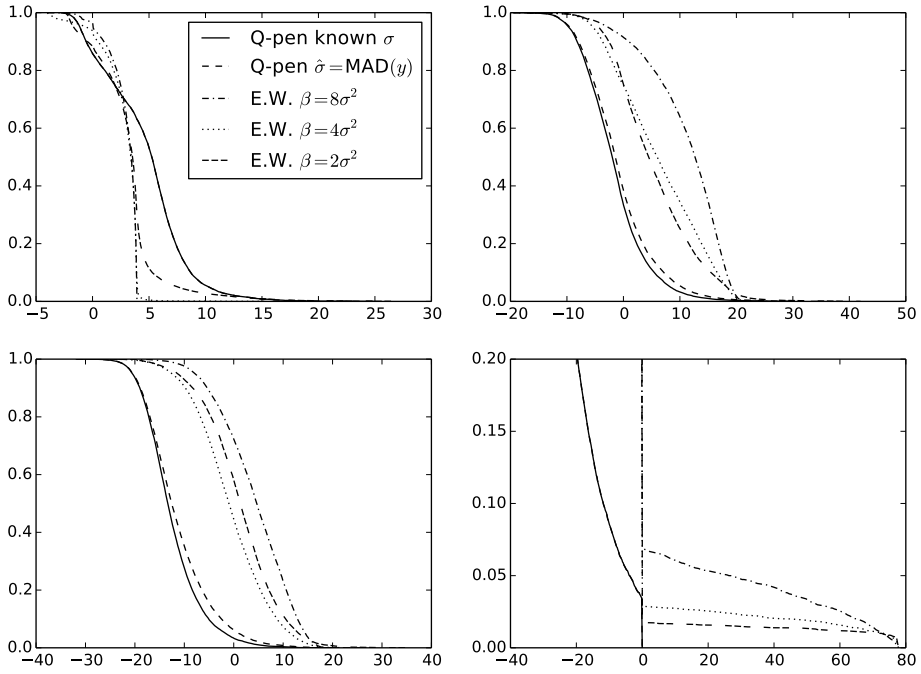
Experiment 3.4 (High misspecifiations). Let $n=100$, $\mathbf{f}=0$ and $M=8\sqrt{n}$. Let $\hat{\mu}_1=\sigma\sqrt{n}e_1$ and $\hat{\mu}_j=\sigma(\sqrt{n}+1)e_2$ for all $j\geq 2$. This example is inspired by [13, Section 2.1].

Figure 1 shows the performance of Exponential weights and the procedure of the present paper, denoted by Q-PEN in the figure. Exponential weights are computed with temperature parameters $2\sigma^2, 4\sigma^2$ and $8\sigma^2$ (the temperature $4\sigma^2$ is recommended in [32, 17] and the temperature $8\sigma^2$ is recommended in [15]). The procedure (3.3) is implemented with the true value of σ and with a data-driven estimate $\hat{\sigma}^2=\textsc{mad}(\mathbf{y})$ where MAD is the normalized Mean Absolute Deviation which provides a rough upper bound of the variance. The procedure (3.3) is implemented using the default quadratic program solver of CVXOPT. We perform $S=10000$ replications. For different estimators $\hat{\mu}$, Figure 1 shows the empirical tail of the excess risk given by

$$t\to\frac{1}{S}\sum_{k=1}^{S}\mathbf{1}_k\left(\left\{\|\hat{\mu}-\mathbf{f}\|^2-\min_{j=1,\dots,M}\|\hat{\mu}_j-\mathbf{f}\|^2>\sigma^2 t\right\}\right),$$

where $\mathbf{1}_k(E)=1$ if the event E holds at the k-th replication, and 0 otherwise. Note that since the estimators are valued in the convex hull of $\{\hat{\mu}_1,\dots,\hat{\mu}_M\}$, the excess risk may be smaller than 0. This explains the negative values in Figure 1. Exponential weights outperform the

Figure 1: Empirical tail of the excess risk for Exponential Weights and Q-PEN in Experiments 3.1 (top-left), 3.2 (top-right), 3.3 (bottom-left) and 3.4 (bottom-right). For Experiments 3.1 and 3.4, the two Q-PEN curves are the same.



procedure of the present paper when the model is well specified (Experiment 3.1). When the model presents small misspecifications (Experiments 3.2 and 3.3), the procedure of the present paper outperforms exponential weights. In the case of large misspecifications (Experiment 3.4), exponential weights are clearly suboptimal and present a large risk on an event of small probability, confirming the theoretical results of [13, Section 2.1]. In these experiments, a data-driven estimate for the variance in the procedure Q-PEN performs similarly to the procedure with known variance.

4 Sparsity oracle inequality

The goal of this section is to prove sparsity oracle inequalities. We are given p deterministic vectors in $\mathbf{R}^n$ that are the columns of a $n \times p$ real matrix $\mathbb{X}$, and the goal is to find an estimator $\hat{\theta} \in \mathbf{R}^p$ such that the quantity $\|\mathbb{X}\hat{\theta} - \mathbf{f}\|_2^2$ is close to $\|\mathbb{X}\theta^* - \mathbf{f}\|_2^2$ for some sparse $\theta^* \in \mathbf{R}^p$ for which $\mathbb{X}\theta^*$ is a good approximation of the unknown regression vector $\mathbf{f}$.

4.1 Adaptation to the variance

First, we derive a sharp oracle inequality under one of the assumptions from Section 3.3 for homoscedastic noise, assuming that a sparsity parameter $k \geq 1$ is known. This parameter k is an upper bound on the sparsity of some vector θ^* such that $\mathbb{X}\theta^*$ is a good approximation of $\mathbf{f}$.

Consider the family of estimators $\{\hat{\mu}_1, ..., \hat{\mu}_M\}$ where for each $j = 1, ..., M$, $\hat{\mu}_j = A_j\mathbf{y}$ and A_j is the projection matrix on a linear span of k linearly independent columns of $\mathbb{X}$. In particular, $M \leq \binom{p}{k}$. The estimator $\hat{\mu}_j$ is also the least squares estimator on the subspace V_j of dimension k generated by these k columns of $\mathbb{X}$.

Now consider the estimator $\hat{\theta}_M^{(k)} \in \mathbf{R}^M$ defined by

$$\hat{\theta}_M^{(k)} = \operatorname*{argmin}_{\theta \in \Lambda^M} \left(\|\hat{\mu}_\theta\|_2^2 - 2\mathbf{y}^T\hat{\mu}_\theta + \frac{1}{2}\,\widehat{\mathrm{pen}}(\theta) \right),$$

where $\widehat{\mathrm{pen}}(\cdot)$ is the penalty (3.2). It is exactly the procedure (3.3) from Theorem 3.3 with the uniform prior ($\pi_j = 1/M$ for all $j = 1, ..., M$) since the noise random variables $\xi_1, ..., \xi_n$ have the same variance and the projection matrices $A_1, ..., A_M$ have the same trace equal to k. This procedure is fully adaptive with respect to the unknown variance of the noise. The following result is a direct consequence of Theorem 3.3.

Corollary 4.1. *Let $k \geq 1$. Assume that the variance of the noise components are the same: $\mathbb{E}[\xi_i^2]$ are equal for all $i = 1, ..., n$. Assume one of the Assumptions 3.1, 3.2 or 3.3 on the noise $\boldsymbol{\xi} = (\xi_1, ..., \xi_n)$ and set the value of β accordingly with $L = 1$. Let $\hat{\theta}_M^{(k)}$ be the estimator defined above. Then for all $x > 0$, with probability greater than $1 - 3\exp(-x)$,*

$$\left\| \hat{\mu}_{\hat{\theta}_M^{(k)}} - \mathbf{f} \right\|_2^2 \leq \min_{\theta \in \mathbf{R}^p,\, |\theta|_0 \leq k} \left(\|\mathbb{X}\theta - \mathbf{f}\|_2^2 \right) + c\,\beta \left(k \log\left(\frac{ep}{k} \right) + x \right),$$

for some absolute constant $c > 0$.

Thus, we see that the knowledge of the variance is not needed to get a sharp oracle inequality when the parameter k (an upper bound on the sparsity) is known.

4.2 Adaptation to sparsity

Here, we propose a new aggregation method that is adaptive to the sparsity and only requires an estimator $\hat{K}^2$ that upper bounds the subgaussian norm of the noise with high probability. We will make the following assumption on the noise.

Assumption 4.1 (Subgaussian noise)**.** *Let $K > 0$ and assume that the random vector $\boldsymbol{\xi}$ satisfies:*

$$\forall \alpha \in \mathbf{R}^n, \qquad \mathbb{E}\exp(\alpha^T \boldsymbol{\xi}) \leq \exp\left(\frac{\|\alpha\|_2^2 K^2}{2}\right).$$

Contrary to the previous section, the components of $\boldsymbol{\xi}$ are not assumed to be independent. The same assumption is made in [13] for aggregation of deterministic vectors. Under this assumption, the authors of [28] proved the concentration inequality (A.8) and we use this concentration result to prove the following oracle inequality for aggregation of Least Squares estimators. Given an estimator $\hat{K}^2$, define for any $\theta \in \Lambda^M$

$$\hat{V}_n(\theta) = \|\hat{\mu}_\theta\|_2^2 - 2\mathbf{y}^T\hat{\mu}_\theta + \tfrac{1}{2}\,\widehat{\mathrm{pen}}(\theta) + 32\hat{K}^2\sum_{j=1}^M \theta_j \log\frac{1}{\pi_j},$$

where $\widehat{\mathrm{pen}}(\cdot)$ is the penalty (3.2). We consider the estimator $\hat{\mu}_{\hat{\theta}}$ of $\mathbf{f}$ where

$$\hat{\theta} \in \operatorname*{argmin}_{\theta\in\Lambda^M} \hat{V}_n(\theta). \tag{4.1}$$

The function $\hat{V}_n$ is equal to the sum of $\hat{H}_n$ (3.1) and some linear function of θ. Thus $\hat{V}_n$ is also convex and minimizing $\hat{V}_n$ over the simplex is a quadratic program.

Proposition 4.1. *Let $K > 0$ be the smallest positive number such that the random vector $\boldsymbol{\xi}$ satisfies Assumption 4.1. For all $j = 1, ..., M$, let $b_j \in \mathbf{R}^n$ and let A_j be a square matrix of size n that satisfies $A_j = A_j^T = A_j^2$. Let $(\pi_1, ..., \pi_M) \in \Lambda^M$ such that for all $j = 1, ..., M$, $\mathrm{Tr}(A_j) \leq \log(\pi_j^{-1})$. Let $\hat{K} > 0$ be a given estimator and let $\hat{\theta}$ be defined in (4.1). Let $\delta := \mathbb{P}(\hat{K}^2 < K^2)$. Then for all $x > 0$, with probability greater than $1 - \delta - 2\exp(-x)$,*

$$\|\hat{\mu}_{\hat{\theta}} - \mathbf{f}\|_2^2 \leq \min_{j=1,...,M}\left(\|\hat{\mu}_j - \mathbf{f}\|_2^2 + 64\hat{K}^2\log\frac{1}{\pi_j}\right) + 28K^2x. \tag{4.2}$$

Proposition 4.1 is proved in Appendix B.4. Compared to (3.17), this oracle inequality holds for orthogonal projectors under the constraint $\mathrm{Tr}(A_j) \leq \log(\pi_j^{-1})$ for all $j = 1, ..., M$. However, this oracle inequality presents some advantages. First, it holds under Assumption 4.1 which is weaker that the noise assumptions of Section 3 since the noise coordinates do not need to be independent. Second, the quantity $\max_{i=1,...,n} \|\xi_i\|_{\psi_2}/\sigma_i$ appearing in (3.14) is not present here, which is possible thanks to the constraint $\mathrm{Tr}(A_j) \leq \log(\pi_j^{-1})$. Finally, one does not need to know the variance of the noise in order to compute the proposed estimator; its construction only relies on $\hat{K}$ which can be any estimate that *upper bounds* the subgaussian norm of the random vector $\boldsymbol{\xi}$. For instance, assume that $\boldsymbol{\xi}$ is zero-mean Gaussian with covariance matrix $\sigma^2 I_{n\times n}$, and assume that an estimator $\hat{\sigma}^2$ of σ^2 is accessible, and that this estimator has bounded bias. Let $\gamma > 1$ and $\epsilon = \mathbb{P}(\hat{\sigma}^2 < \sigma^2/\gamma)$. The quantity ϵ is likely to be small if $\hat{\sigma}^2$ has a bounded bias and γ is large enough. Then one can use the upper bound $\hat{K}^2 = \gamma\hat{\sigma}^2$ in Proposition 4.1, which yields that with probability greater than $1 - 3\epsilon$,

$$\|\hat{\mu}_{\hat{\theta}} - \mathbf{f}\|_2^2 \leq \min_{j=1,...,M} \left(\|\hat{\mu}_j - \mathbf{f}\|_2^2 + 64\gamma\hat{\sigma}^2 \log\frac{1}{\pi_j} \right) + 28\sigma^2 \log(1/\epsilon).$$

Thus, γ is used to perform a trade-off between the probability estimate and the remainder term of the oracle inequality. By using an upper bound for $\hat{K}^2$ in Proposition 4.1 the oracle inequality holds with slightly worse constants but with high probability. Examples of estimators $\hat{\sigma}^2$ are given in Section 3.2.

We now use the oracle inequality (4.2) to perform sparsity pattern aggregation [36, 35, 14, 41]. For each subset $J \subset \{1, ..., p\}$, let $\hat{\mu}_J^{LS}$ be the Least Squares estimator on the linear span of the columns of $\mathbb{X}$ whose indices are in J. This estimator satisfies the oracle inequality (B.14) with $d \leq |J|$, where $|J|$ denotes the cardinal of J and d is the dimension of the linear span of the columns whose indices are in J. We aggregate these 2^p Least Squares estimators using the method (4.1) and the prior distribution given by $\pi_J \propto e^{-|J|}\binom{p}{|J|}^{-1}$. As sparsity pattern aggregation is not central in the present paper, we keep this presentation short and refer the reader to [36, 35, 14, 41] for more details on sparsity pattern aggregation and the construction of Least Squares estimators.

Given a subset $J \subset \{1, ..., p\}$, the Least Squares estimator $\hat{\mu}_J^{LS}$ is of the form $\hat{\mu}_J^{LS} = A_J\mathbf{y}$ for some projection matrix A_J and the inequality $\mathrm{Tr}(A_J) \leq |J| \leq \log(\pi_J^{-1})$ holds [35, Section 5.2.1, the normalizing constant is greater than 1]. Define $\hat{\theta}^{SPA}$ such that

$$\mathbb{X}\hat{\theta}^{SPA} = \hat{\mu}_{\hat{\theta}}, \tag{4.3}$$

where $\hat{\theta}$ is the estimator from (4.1) and $\hat{\mu}_{\hat{\theta}}$ is obtained by aggregating the $M = 2^p$ estimators $(\hat{\mu}_J^{LS})_{J\subset\{1,...,p\}}$. Then the following sparsity oracle inequality holds, where $|\theta|_0$ is the number of non-zero coefficients of θ.

Theorem 4.1. *Let $\mathbb{X}$ be a deterministic design matrix with p columns Let $K > 0$ be the smallest positive real number such that the noise random $\boldsymbol{\xi}$ satisfies Assumption 4.1. Let $\hat{K}$*

be a given estimator and let $\delta \coloneqq \mathbb{P}(\hat{K}^2 < K^2)$. *Then, the sparsity pattern aggregate* $\hat{\theta}^{SPA}$ *defined in (4.3) satisfies with probability greater than* $1 - \delta - 3\exp(-x)$,

$$\left\|\mathbb{X}\hat{\theta}^{SPA} - \mathbf{f}\right\|_2^2 \le \inf_{\theta \in \mathbf{R}^p} \left[\|\mathbb{X}\theta - \mathbf{f}\|_2^2 + 31K^2 x \right.$$
$$\left. + (64\hat{K}^2 + 4K^2)\left(\frac{1}{2} + 2|\theta|_0 \log\left(\frac{ep}{1 \vee |\theta|_0}\right)\right)\right]. \quad (4.4)$$

Theorem 4.1 is proved in Appendix B.4. It improves upon the previous results on sparsity pattern aggregation [14, 36, 35, 41] in several aspects.

First, the noise $\boldsymbol{\xi}$ is only assumed to be subgaussian and its components need not be independent, whereas previous results only hold under Gaussianity and independence of the noise components. Theorem 4.1 shows that the optimal bounds previously known for Gaussian noise [14, 36, 35, 41] are of the same form when the noise is only assumed to be subgaussian.

Second, to construct the aggregates in [14, 36, 35, 41] one needs the exact knowledge of the covariance matrix of the noise. In Theorem 4.1, only an upper bound of the subgaussian norm of the noise is needed to construct the estimator.

Third, we do not split the data in order to perform sparsity pattern aggregation, as opposed to the "sample cloning" approach [41, Lemma 2.1]. Sample cloning is possible only for Gaussian noise when the variance is known; it cannot be used here as $\boldsymbol{\xi}$ can be any subgaussian vector.

The estimator of Theorem 4.1 achieves the minimax rate for any intersection of ℓ_0 and ℓ_q balls, where $q \in (0, 2)$. This can be shown by applying the arguments of [14, 41] and bounding the right hand side of (4.4). Indeed, although [14, 41] consider only normal random variables, the argument does not depend on the noise distribution.

The result above holds without any assumption on the design matrix $\mathbb{X}$, as opposed to the LASSO or the Dantzig estimators which need assumptions on the design matrix $\mathbb{X}$ to achieve sparsity oracle inequalities similar but weaker than (4.4).

The interest of the LASSO and the Dantzig estimators is that they can be computed efficiently for large p. The sparsity pattern aggregate based on exponential weights can also be computed efficiently usind MCMC methods [36]. The estimator $\hat{\theta}^{SPA}$ proposed here suffers the same drawback as [8] or the sparsity pattern aggregate performed with Q-aggregation [14]: it is not known whether these estimators can be computed in polynomial time, which makes them useful only for relatively small p.

A Concentration inequalities

In this appendix, we gather the concentration inequalities used to prove Theorem 3.3.

A.1 Gaussian concentration

Let X be a zero-mean Gaussian random variable with variance σ^2. A standard bound on the Gaussian tail is

$$\forall x > 0, \qquad \mathbb{P}\left(X > \sigma\sqrt{2x}\right) \leq \exp(-x).$$

Let $v \in \mathbf{R}^n$ and let $\xi_1, ..., \xi_n$ be zero-mean independent Gaussian random variables with $\mathbb{E}[\xi_i^2] = \sigma_i^2$ for all i. Then $v^T\boldsymbol{\xi}$ is Gaussian with variance $\|D_\sigma v\|_2^2$ and thus

$$\forall x > 0, \qquad \mathbb{P}\left(v^T\boldsymbol{\xi} > \|D_\sigma v\|_2 \sqrt{2x}\right) \leq \exp(-x). \tag{A.1}$$

Proposition A.1 (Gaussian chaos of order 2)**.** *Let $\xi_1, ..., \xi_n$ be independent zero-mean normal random variables with for all $i = 1, ..., n$, $\mathbb{E}[\xi_i^2] = \sigma_i^2$. Let A be any $n \times n$ real matrix. Then for any $x > 0$,*

$$\mathbb{P}\left(\boldsymbol{\xi}^T A\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T A\boldsymbol{\xi}] > 2\|D_\sigma A D_\sigma\|_{\mathrm{HS}}\sqrt{x} + 2|||D_\sigma A D_\sigma|||_2 x\right) \leq \exp(-x). \tag{A.2}$$

A proof of this concentration result is given in [10, Example 2.12] for diagonal-free matrices. A slightly modified version is available in [39, Theorem 2.2] for positive semi-definite matrices. It can be easily extended to the general case via the following argument.

Proof of Proposition A.1. First, notice that if the result holds for standard normal random variables with variance 1, then by considering the random variables $\xi_i' = \xi_i/\sigma_i$ and the matrix $M = D_\sigma A D_\sigma$, the result also holds when $\xi_1, ..., \xi_n$ have variances different than 1. Thus in the following we assume without loss of generality that $\sigma_i = 1$ for all $i = 1, ..., n$.

Second, if the result holds for all symmetric matrices A, then for a non-symmetric matrix A one can consider $B = \frac{A+A^T}{2}$ which is symmetric. Then $\boldsymbol{\xi}^T B\boldsymbol{\xi} = \boldsymbol{\xi}^T A\boldsymbol{\xi}$ and by the triangle inequality,

$$|||B|||_2 \leq \frac{|||A|||_2 + \left|\left|\left|A^T\right|\right|\right|_2}{2} = |||A|||_2, \qquad \|B\|_{\mathrm{HS}} \leq \frac{\|A\|_{\mathrm{HS}} + \left\|A^T\right\|_{\mathrm{HS}}}{2} = \|A\|_{\mathrm{HS}}\,.$$

Thus if the concentration inequality (A.2) holds for the symmetric matrix B, it will also hold for the non-symmetric matrix A. Without loss of generality, we can consider only symmetric matrices.

Let $\xi_1, ..., \xi_n$ be standard normal random variables and let A be a symmetric matrix. There exists an invertible square matrix U with $U^T = U^{-1}$ such that $A = U^T\Lambda U$ for some diagonal matrix $\Lambda = \mathrm{diag}(\mu_1, ..., \mu_n)$. By rotational invariance of the normal distribution, if $(X_1, ..., X_n)^T = U\boldsymbol{\xi}$ then $X_1, ..., X_n$ are i.i.d. standard normal random variables. As $\mathbb{E}[\boldsymbol{\xi}^T A\boldsymbol{\xi}] = \mathrm{Tr}A = \sum_{i=1}^n \mu_i$,

$$\boldsymbol{\xi}^T A\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T A\boldsymbol{\xi}] = \sum_{i=1}^n \mu_i(X_i^2 - 1).$$

The rest of the proof can be treated exactly as in the proof of [10, Example 2.12], using the bound

$$\forall \lambda \in (0, 1/2), \quad \log \mathbb{E} \exp(\lambda(\xi_i^2 - 1) \leq \frac{\lambda^2}{1 - 2\lambda},$$

without assuming that A is diagonal-free. □

A.2 Subgaussian concentration

Again, we present tools to control terms of the form $\boldsymbol{\xi}^T Q \boldsymbol{\xi}$ and $v^T \boldsymbol{\xi}$ that appear in Lemma 3.1. Proposition A.2 below provides a concentration result for the latter.

Proposition A.2 (Hoeffding-type inequality [43, Section 5.2.3])**.** *There exists an absolute constant $C_H > 0$ such that the following holds. Let $n \geq 1$ and $\xi_1, ..., \xi_n$ be independent zero-mean subgaussian random variables with $\max_{i=1,...,n} \|\xi_i\|_{\psi_2} \leq K$ for some real number $K > 0$. Let $v \in \mathbf{R}^n$.*

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

$$\boldsymbol{\xi}^T v \leq C_H K \|v\|_2 \sqrt{x} \tag{A.3}$$

where $\boldsymbol{\xi} = (\xi_1, ..., \xi_n)^T$.

The concentration result for a quadratic form of independent zero-mean subgaussian random variables given in Proposition A.3 below is known as the Hanson-Wright inequality. First versions of this inequality can be found in Hanson and Wright [27] and Wright [45], although with a weaker statement than Proposition A.3 below since these results involve $\| |(|a_{ij}|)| \|_2$ instead of $\|A\|_2$. Recent proofs of this concentration inequality with $\|A\|_2$ instead of $\| |(|a_{ij}|)| \|_2$ can be found in Rudelson and Vershynin [37] or Barthe and Milman [4, Theorem A.5].

Proposition A.3 (Hanson-Wright inequality [37])**.** *There exist absolute constants $c_{w_1}, c_{w_2}, c > 0$ such that the following holds. Let $n \geq 1$ and $\xi_1, ..., \xi_n$ be independent zero-mean subgaussian random variables with $\max_{i=1,...,n} \|\xi_i\|_{\psi_2} \leq K$ for some real number $K > 0$. Let A be any $n \times n$ real matrix. Then for all $t > 0$,*

$$\mathbb{P}\left(\boldsymbol{\xi}^T A \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T A \boldsymbol{\xi}] > t\right) \leq \exp\left(-c \min\left(\frac{t^2}{K^4 \|A\|_{\mathrm{HS}}^2}, \frac{t}{K^2 \|A\|_2}\right)\right) \tag{A.4}$$

where $\boldsymbol{\xi} = (\xi_1, ..., \xi_n)^T$. Furthermore, for any $x > 0$, with probability greater than $1 - \exp(-x)$,

$$\boldsymbol{\xi}^T A \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T A \boldsymbol{\xi}] \leq c_{w_1} K^2 \|A\|_2 x + c_{w_2} K^2 \|A\|_{\mathrm{HS}} \sqrt{x}. \tag{A.5}$$

A.3 Concentration under Assumption 3.3

Proposition A.4. *Let $v \in \mathbf{R}^n$ and $K > 0$. Let $\xi_1, ..., \xi_n$ be n independent random variables satisfying the moment assumption (3.15). The following Hoeffding-type inequality holds:*

$$\mathbb{P}\left(v^T \boldsymbol{\xi} > 2K \|v\|_2 \sqrt{x}\right) \leq \exp(-x). \tag{A.6}$$

Proposition A.4 is proved in the supplementary material.

A.4 Concentration of subgaussian vectors

A direct consequence of Assumption 4.1 on the random vector $\boldsymbol{\xi}$ is the following Hoeffding-type concentration inequality:

$$\mathbb{P}\left(\alpha^T \boldsymbol{\xi} > K \|\alpha\|_2 \sqrt{2x}\right) \leq \exp(-x). \tag{A.7}$$

Under Assumption 4.1, the following concentration inequality was proven in [28].

Proposition A.5 (One sided concentration [28])**.** *Let $\boldsymbol{\xi}$ be a random vector in $\mathbf{R}^n$ satisfying Assumption 4.1 for some $K > 0$. Let A be a real $n \times n$ positive semi-definite symmetric matrix. Then for all $x > 0$, with probability greater than $1 - \exp(-x)$,*

$$\boldsymbol{\xi}^T A \boldsymbol{\xi} \leq K^2 \left(\mathrm{Tr} A + 2 \|A\|_{\mathrm{HS}} \sqrt{x} + 2 |||A|||_2 x\right). \tag{A.8}$$

This result is remarkable as it holds with the same constants as in the Gaussian case (A.2), under the weak Assumption 4.1. Similar results are obtained in a different form in [39]. Unlike the previous concentration results given in Appendix A used in Section 3, the above inequality is only one-sided, and it is not known if the above result holds as a two-sided inequality or without the positive semi-definiteness of A. Another difference with the concentration inequalities of Appendix A is that the term $\mathrm{Tr} A$ in (A.8) is an upper bound on the expectation of $\boldsymbol{\xi}^T A \boldsymbol{\xi}$ up to constants. It is not known whether this concentration inequality holds with the constant term $K^2 \mathrm{Tr} A$ replaced by $\mathbb{E}[\boldsymbol{\xi}^T A \boldsymbol{\xi}]$. The following corollary extends Proposition A.5 to general matrices.

Corollary A.1 (Corollary of Proposition A.5 for any real matrix A)**.** *Under Assumption 4.1 and for any real matrix A, with probability greater than $1 - \exp(-x)$, the following holds:*

$$\boldsymbol{\xi}^T A \boldsymbol{\xi} \leq K^2 \left(\|A\|_1 + 2 \|A\|_{\mathrm{HS}} \sqrt{x} + 2 |||A|||_2 x\right). \tag{A.9}$$

Proof. To see this, let $A_s := \frac{1}{2}(A + A^T)$ and consider $|A_s| := \sqrt{A_s^2}$, the unique square root of the positive semi-definite matrix A_s^2. By definition of $|A_s|$ and the triangle inequality,

$$\begin{aligned} &\boldsymbol{\xi}^T A \boldsymbol{\xi} = \boldsymbol{\xi}^T A_s \boldsymbol{\xi} \leq \boldsymbol{\xi}^T |A_s| \boldsymbol{\xi}, \qquad && \mathrm{Tr}(|A_s|) = \|A_s\|_1 \leq \|A\|_1, \\ &|||A_s|||_2 = |||A_s|||_2 \leq |||A|||_2, \qquad && \||A_s|\|_{\mathrm{HS}} = \|A_s\|_{\mathrm{HS}} \leq \|A\|_{\mathrm{HS}}. \end{aligned}$$

Thus applying (A.8) to the matrix $|A_s|$ proves (A.9). □

B Proofs

B.1 Proof of Lemma 3.1

We start with some preliminary remarks. If $R_n(\theta) \coloneqq \|\hat{\mu}_{\hat{\theta}}\|_2^2 - 2\mathbf{y}^T\hat{\mu}_{\hat{\theta}}$, $R_n(\cdot)$ is differentiable and the following identify holds for any $j = 1, ..., M$ and $\theta \in \Lambda^M$:

$$\nabla R_n(\theta)^T(e_j - \theta) = \|\hat{\mu}_j - \mathbf{f}\|_2^2 - \|\hat{\mu}_\theta - \mathbf{f}\|_2^2 - 2\boldsymbol{\xi}^T(\hat{\mu}_j - \hat{\mu}_\theta) - \|\hat{\mu}_\theta - \hat{\mu}_j\|_2^2 . \tag{B.1}$$

The penalty (3.2) satisfies for any $g \in \mathbf{R}^n$ and any $\theta \in \Lambda^M$:

$$\sum_{j=1}^M \theta_j \|\hat{\mu}_j - g\|_2^2 = \|\hat{\mu}_\theta - g\|_2^2 + \widehat{\mathrm{pen}}(\theta). \tag{B.2}$$

This can be shown by using simple properties of the Euclidean norm, or by noting that the equality above is a bias-variance decomposition. The function $\widehat{\mathrm{pen}}(\cdot)$ is differentiable and for any $j = 1, ..., M$, and $\theta \in \Lambda^M$, one can check that

$$\begin{aligned} \tfrac{1}{2}\nabla\, \widehat{\mathrm{pen}}(\theta)^T(e_j - \theta) &= \tfrac{1}{2}\|\hat{\mu}_\theta - \hat{\mu}_j\|_2^2 - \tfrac{1}{2}\,\widehat{\mathrm{pen}}(\theta), \\ &= \|\hat{\mu}_\theta - \hat{\mu}_j\|_2^2 - \tfrac{1}{2}\sum_{k=1}^M \theta_k \|\hat{\mu}_j - \hat{\mu}_k\|_2^2 , \end{aligned} \tag{B.3}$$

where we used (B.2) with $g = \hat{\mu}_j$ for the last equality.

Proof of Lemma 3.1. Let $J^* = 1, ..., M$ be a deterministic integer. Since $\hat{\theta}$ minimizes $\hat{H}_n$ over the simplex and $\hat{H}_n$ is convex and differentiable, a simple consequence of the KKT conditions [11, 4.2.3, equation (4.21)] yields:

$$\nabla \hat{H}_n(\hat{\theta})^T(e_{J^*} - \hat{\theta}) \geq 0. \tag{B.4}$$

Let $W \coloneqq \nabla \hat{H}_n(\hat{\theta})^T(e_{J^*} - \hat{\theta})$. By simple algebraic calculations using (B.1) and (B.3), we have

$$\begin{aligned} W = \|\hat{\mu}_{J^*} - \mathbf{f}\|_2^2 - \|\hat{\mu}_{\hat{\theta}} - \mathbf{f}\|_2^2 - 2\boldsymbol{\xi}^T(\hat{\mu}_{J^*} - \hat{\mu}_{\hat{\theta}}) \\ + 2\mathrm{Tr}(D_\sigma A_{J^*} D_\sigma) - \sum_{k=1}^M \hat{\theta}_k \mathrm{Tr}(D_\sigma A_k D_\sigma) \\ - \tfrac{1}{2}\sum_{k=1}^M \hat{\theta}_k \|\hat{\mu}_k - \hat{\mu}_{J^*}\|_2^2 + \beta \log\frac{1}{\pi_{J^*}} - \beta \sum_{k=1}^M \hat{\theta}_k \log\frac{1}{\pi_k}. \end{aligned}$$

Since for all $j = 1, ..., M$, $\mathrm{Tr}(D_\sigma A_j D_\sigma) = \mathbb{E}[\boldsymbol{\xi}^T A_j \boldsymbol{\xi}]$, (B.4) can be rewritten

$$\|\hat{\mu}_{\hat{\theta}} - \mathbf{f}\|_2^2 \leq \|\hat{\mu}_{J^*} - \mathbf{f}\|_2^2 + 2\beta \log\frac{1}{\pi_{J^*}} + Z(J^*, \hat{\theta}), \tag{B.5}$$

where for all $J^* = 1, ..., M$ and $\theta \in \Lambda^M$,

$$\begin{aligned} Z(J^*, \theta) \coloneqq \quad & 2\boldsymbol{\xi}^T(\hat{\mu}_\theta - \hat{\mu}_{J^*}) - 2\sum_{k=1}^{M} \theta_k \mathbb{E}[\boldsymbol{\xi}^T(A_k - A_{J^*})\boldsymbol{\xi}] \\ & - \tfrac{1}{2}\sum_{k=1}^{M} \theta_k \|\hat{\mu}_{J^*} - \hat{\mu}_k\|_2^2 - \beta \sum_{k=1}^{M} \theta_k \log\frac{1}{\pi_k} - \beta \log\frac{1}{\pi_{J^*}}. \end{aligned}$$

The quantity $Z(J^*, \theta)$ is affine in its second argument $\theta \in \Lambda^M$ thus it is maximized at a vertex of Λ^M, and the following upper bounds hold:

$$Z(J^*, \hat{\theta}) \leq \max_{\theta \in \Lambda^M} Z(J^*, \theta) = \max_{k=1,...,M} Z(J^*, e_k) \leq \max_{j,k=1,...,M} Z(j, e_k). \tag{B.6}$$

Let $\zeta_{j,k} \coloneqq Z(j, e_k)$ for all $j, k = 1, ..., M$. From (B.5) and (B.6),

$$\|\hat{\mu}_{\hat{\theta}} - \mathbf{f}\|_2^2 \leq \|\hat{\mu}_{J^*} - \mathbf{f}\|_2^2 + 2\beta \log\frac{1}{\pi_{J^*}} + \max_{j,k=1,...,M} \zeta_{j,k},$$

where

$$\zeta_{j,k} = 2\boldsymbol{\xi}^T(\hat{\mu}_k - \hat{\mu}_j) - 2\mathbb{E}[\boldsymbol{\xi}^T(A_k - A_j)\boldsymbol{\xi}] - \tfrac{1}{2}\|\hat{\mu}_k - \hat{\mu}_j\|_2^2 - \beta \log\frac{1}{\pi_k \pi_j}.$$

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

$$\|\hat{\mu}_k - \hat{\mu}_j\|_2^2 = \|B_{jk}\boldsymbol{\xi}\|_2^2 + \|B_{jk}\mathbf{f} + b_k - b_j\|_2^2 + 2\boldsymbol{\xi}^T B_{jk}^T(B_{jk}\mathbf{f} + b_k - b_j). \tag{B.7}$$

After some algebra, we get

$$\begin{aligned} \zeta_{j,k} = \ & \boldsymbol{\xi}^T Q_{j,k}\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q_{j,k}\boldsymbol{\xi}] + \boldsymbol{\xi}^T v_{j,k} \\ & - \beta \log\frac{1}{\pi_k \pi_j} - \tfrac{1}{2}\|B_{jk}D_\sigma\|_{\mathrm{HS}}^2 - \tfrac{1}{2}\|B_{jk}\mathbf{f} + b_k - b_j\|_2^2 \end{aligned}$$

where we used the equality $\|B_{jk}D_\sigma\|_{\mathrm{HS}}^2 = \mathbb{E}[\|B_{jk}\boldsymbol{\xi}\|_2^2]$ and where $Q_{j,k}$ and $v_{j,k}$ are defined in (3.19) and (3.20), respectively. □

B.2 Proof of Theorem 3.3

Theorem 3.1 is a direct consequence of Theorem 3.3 for homoscedastic Gaussian noise, with $L = 1$.

In order to prove Theorem 3.3, we need the following notation. We refer the reader to Appendix B.3 for a proof with similar arguments and less notational complexity. Let $K, C_{W_1}, C_{W_2}, C_H > 0$ and a diagonal matrix $\bar{D}$ be parameters that are specified below

for each assumption. For any $v \in \mathbf{R}^n$ and any real matrix Q, consider the following concentration inequalities: $\forall x > 0$,

$$\mathbb{P}\left(v^T \boldsymbol{\xi} > C_H K \|v\|_2^2\right) \leq \exp(-x), \tag{B.8}$$

$$\mathbb{P}\left(\boldsymbol{\xi}^T Q \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q \boldsymbol{\xi}] > C_{W_2} K \left\|Q\bar{D}\right\|_{\mathrm{HS}} \sqrt{x} + C_{W_1} K^2 |||Q|||_2 x\right) \leq \exp(-x). \tag{B.9}$$

Let $(\bar{d}_i)_{i=1,\dots,}$ be the diagonal elements of the matrix $\bar{D}$ and let

$$\beta = K^2 \left(C_{W_1}(2+L)2L + 2C_H^2(1+L)^2 + \tfrac{1}{2} C_{W_2}^2 \max_{i=1,\dots,n} \frac{\bar{d}_i^2}{\sigma_i^2}(2+L)^2 \right). \tag{B.10}$$

The above concentration inequalities are satisfied under the three assumptions on the noise, with different constants:

- Under Assumption 3.1, set $K = \max_{i=1,\dots,n} \sigma_i$, $\bar{D} = D_\sigma$ and $C_H = \sqrt{2}$, $C_{W_1} = 2$, $C_{W_2} = 2$. With this choice of constants, the value of β (B.10) is equal to the value (3.13), (B.8) becomes exactly (A.1) and (B.9) is a consequence of (A.2) applied to the matrix Q and the random vector $\boldsymbol{\xi}$.

- Under Assumption 3.2, K is given in the assumption, set

  $$\bar{D} = \mathrm{diag}(\|\xi_1\|_{\psi_2}, \dots, \|\xi_n\|_{\psi_2}),$$

  $C_H = c_h$, $C_{W_1} = c_{w_1}$ and $C_{W_2} = c_{w_2}$ where c_h, c_{w_1} and c_{w_2} are the numerical constants from Propositions A.2 and A.3. With this choice of constants, the value of β (B.10) is equal to the value (3.14), (B.8) becomes exactly (A.3) and (B.9) is a direct consequence of (A.5) applied to the random vector $(\frac{\xi_1}{\|\xi_1\|_{\psi_2}}, \dots, \frac{\xi_n}{\|\xi_n\|_{\psi_2}})$ and the matrix $\bar{D}Q\bar{D}$.

- Under Assumption 3.3, K is given in the assumption, set $\bar{D} = D_\sigma$ and $C_H = 2$, $C_{W_1} = 256$, $C_{W_2} = 8\sqrt{3}$. With this choice of constants, the value of β (B.10) is equal to the value (3.16), (B.8) becomes exactly (A.6) and (B.9) becomes exactly (3.23) applied to the random vector $\boldsymbol{\xi}$ and the matrix Q.

Proof of Theorem 3.3. Let $x > 0$. The concentration inequalities (B.8) and (B.9) always hold, with different constants depending on the assumption on the noise as explained above.

Using Lemma 3.1, it is enough to upper bound $\max_{j,k=1,\dots,M} \zeta_{j,k}$ where $\zeta_{j,k}$ is defined in (3.18). Let $j, k = 1, \dots, M$ be fixed, and let $B_{j,k} = A_k - A_j$. We apply the concentration inequality (B.9) to the matrix $Q_{j,k}$ (3.19) and the concentration inequality (B.8) to the vector $v_{j,k}$ (3.20). With the union bound, on the event where both concentration inequalities hold we get that with probability greater than $1 - 2\exp(-x)$,

$$\begin{aligned}\zeta_{j,k} \leq & C_{W_1} K^2 |||Q_{j,k}|||_2 x + C_{W_2} K \left\|Q_{j,k}\bar{D}\right\|_{\mathrm{HS}} \sqrt{x} + C_H K \|v_{j,k}\|_2 \sqrt{x} \\ & - \beta \log \frac{1}{\pi_k \pi_j} - \tfrac{1}{2} \|B_{j,k} D_\sigma\|_{\mathrm{HS}}^2 - \tfrac{1}{2} \|B_{j,k}\mathbf{f} + b_k - b_j\|_2^2 .\end{aligned} \tag{B.11}$$

Using properties of the operator norm and the Hilbert-Schmidt norm with (3.20), (3.19):

$$\begin{aligned}
\|v_{j,k}\|_2 &\leq 2\left(1+\tfrac{1}{2}|||B_{j,k}|||_2\right)\|B_{j,k}\mathbf{f}+b_k-b_j\|_2\,,\\
|||Q_{j,k}|||_2 &\leq \left(2+\tfrac{1}{2}|||B_{j,k}|||_2\right)|||B_{j,k}|||_2,\\
\left\|Q_{j,k}\bar{D}\right\|_{\mathrm{HS}} &\leq 2\left\|B_{j,k}\bar{D}\right\|_{\mathrm{HS}}+\tfrac{1}{2}\left\|B_{j,k}^T B_{j,k}\bar{D}\right\|_{\mathrm{HS}}\,,\\
&\leq \left(2+\tfrac{1}{2}|||B_{j,k}|||_2\right)\left\|B_{j,k}\bar{D}\right\|_{\mathrm{HS}}
\end{aligned}$$

where we used in the last display that for any square matrices M, C, $\|MC\|_{\mathrm{HS}} \leq |||M|||_2\,\|C\|_{\mathrm{HS}}$. We plug these inequalities in (B.11):

$$\begin{aligned}
\zeta_{j,k} \leq &\left(2+\tfrac{1}{2}|||B_{j,k}|||_2\right)\left(C_{W_1}K^2|||B_{j,k}|||_2 x + C_{W_2}K\left\|B_{j,k}\bar{D}\right\|_{\mathrm{HS}}\sqrt{x}\right)\\
&+2C_H K\left(1+\tfrac{1}{2}|||B_{j,k}|||_2\right)\|B_{j,k}\mathbf{f}+b_k-b_j\|_2\sqrt{x}\\
&-\beta\log\frac{1}{\pi_k\pi_j}-\tfrac{1}{2}\|B_{j,k}D_\sigma\|_{\mathrm{HS}}^2-\tfrac{1}{2}\|B_{j,k}\mathbf{f}+b_k-b_j\|_2^2\,.
\end{aligned}$$

We apply the inequality $st \leq \frac{s^2+t^2}{2}$ twice, first with

$$s = C_{W_2}K\left(2+\tfrac{1}{2}|||B_{j,k}|||_2\right)\frac{\left\|B_{j,k}\bar{D}\right\|_{\mathrm{HS}}}{\|B_{j,k}D_\sigma\|_{\mathrm{HS}}}\sqrt{x}$$

and $t = \|B_{j,k}D_\sigma\|_{\mathrm{HS}}$, second with $s = 2C_HK\left(1+\frac{1}{2}|||B_{j,k}|||_2\right)\sqrt{x}$ and $t = \|B_{j,k}\mathbf{f}+b_k-b_j\|_2$. In both cases, the term $\frac{t^2}{2}$ cancels and we obtain

$$\begin{aligned}
\zeta_{j,k} \leq &K^2 x\Big(C_{W_1}\left(2+\tfrac{1}{2}|||B_{j,k}|||_2\right)|||B_{j,k}|||_2+\tfrac{1}{2}C_{W_2}^2\frac{\left\|B_{j,k}\bar{D}\right\|_{\mathrm{HS}}^2}{\|B_{j,k}D_\sigma\|_{\mathrm{HS}}^2}\left(2+\tfrac{1}{2}|||B_{j,k}|||_2\right)^2\Big)\\
&+2C_H^2K^2\left(1+\tfrac{1}{2}|||B_{j,k}|||_2\right)^2 x-\beta\log\frac{1}{\pi_k\pi_j}.
\end{aligned}$$

Let $(b_{i,l})_{i,l=1,\dots,n}$ be the elements of the matrix $B_{j,k} = A_k - A_j$, and $(\bar{d}_i)_{i=1,\dots,}$ be the diagonal elements of the matrix $\bar{D}$. Since

$$\frac{\left\|B_{j,k}\bar{D}\right\|_{\mathrm{HS}}^2}{\|B_{j,k}D_\sigma\|_{\mathrm{HS}}^2} = \frac{\sum_{i,l}\bar{d}_i^2 b_{i,l}^2}{\sum_{i,l}\sigma_i^2 b_{i,l}^2} \leq \max_{i=1,\dots,n}\frac{\bar{d}_i^2}{\sigma_i^2},$$

we obtain $\zeta_{j,k} \leq \beta x - \beta\log\frac{1}{\pi_k\pi_j}$ where β is given in (B.10).

For any $t > 0$, let $x = t + \log\frac{1}{\pi_k\pi_j}$. The inequality $\zeta_{j,k} \leq \beta t$ holds with probability greater than $1 - 2\pi_j\pi_k\exp(-t)$. Using the union bound on $j,k = 1,\dots,M$, we have $\max_{j,k=1,\dots,M}\zeta_{j,k} \leq \beta t$ with probability greater than $1 - \sum_{j,k=1,\dots,M}2\pi_j\pi_k\exp(-t) = 1 - 2\exp(-t)$. □

B.3 Proof of Theorem 3.2

The following inequality will be useful.

Lemma B.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*

$$|\mathrm{Tr}(A - B)| \leq \|A - B\|_{\mathrm{HS}}^2 . \tag{B.12}$$

Proof. Without loss of generality, assume that $\mathrm{Tr}A \geq \mathrm{Tr}B$. As $\|A - B\|_{\mathrm{HS}}^2 = \|A\|_{\mathrm{HS}}^2 + \|B\|_{\mathrm{HS}}^2 - 2\mathrm{Tr}(AB)$ and $\|A\|_{\mathrm{HS}}^2 = \mathrm{Tr}A$, (B.12) is equivalent to $2\mathrm{Tr}(AB) \leq 2\mathrm{Tr}(B)$. Notice that for projection matrices, $\mathrm{Tr}(AB) = \|AB\|_{\mathrm{HS}}^2 \leq |||A|||_2^2 \|B\|_{\mathrm{HS}}^2 \leq \|B\|_{\mathrm{HS}}^2 = \mathrm{Tr}(B)$ and the proof is complete. □

Proof of Theorem 3.2. With the previous notation, $D_\sigma = \sigma I_{n\times n}$, $A_j^T = A_j = A_j^2$. We perform the same calculations as in the proof of Lemma 3.1 and obtain that almost surely

$$\|\hat{\mu}_{\hat{\theta}} - \mathbf{f}\|_2^2 \leq \min_{J^*=1,\dots,M} \left(\|\hat{\mu}_{J^*} - \mathbf{f}\|_2^2 + 2\hat{\beta} \log \frac{1}{\pi_{J^*}} \right) + \max_{j,k=1,\dots,M} \zeta_{j,k},$$

where

$$\begin{aligned} \zeta_{j,k} = \; & \boldsymbol{\xi}^T Q_{j,k} \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q_{j,k} \boldsymbol{\xi}] + \boldsymbol{\xi}^T v_{j,k} - \tfrac{1}{2} \|(A_k - A_j)\mathbf{f} + b_k - b_j\|_2^2 \\ & + 2(\hat{\sigma}^2 - \sigma^2)\mathrm{Tr}(A_j - A_k) - \frac{\sigma^2}{2} \|A_k - A_j\|_{\mathrm{HS}}^2 - \hat{\beta} \log \frac{1}{\pi_k \pi_j}, \end{aligned}$$

the matrices $Q_{j,k}$ and the vectors $v_{j,k}$ are defined in (3.19) and (3.20). Let $j, k \in \{1, ..., M\}$. The assumption on $\hat{\sigma}^2$ and (B.12) yield that on an event Ω_0 of probability greater than $1 - \delta$,

$$2|(\hat{\sigma}^2 - \sigma^2)\mathrm{Tr}(A_j - A_k)| \leq \frac{\sigma^2}{4} \|A_j - A_k\|_{\mathrm{HS}}^2 . \tag{B.13}$$

Let $\beta^* = 56\sigma^2$. On the event Ω_0, $\hat{\beta} \geq \beta^*$ thus

$$\begin{aligned} \zeta_{j,k} \leq \; & \boldsymbol{\xi}^T Q_{j,k} \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q_{j,k} \boldsymbol{\xi}] + \boldsymbol{\xi}^T v_{j,k} - \tfrac{1}{2} \|(A_k - A_j)\mathbf{f} + b_k - b_j\|_2^2 \\ & - \frac{\sigma^2}{4} \|A_k - A_j\|_{\mathrm{HS}}^2 - \beta^* \log \frac{1}{\pi_k \pi_j}. \end{aligned}$$

Note that $|||Q_{j,k}|||_2 \leq 6$, $\|Q_{j,k}\|_{\mathrm{HS}} \leq 3\|A_k - A_j\|_{\mathrm{HS}}$ and $\|v_{j,k}\|_2 \leq 4\|(A_k - A_j)\mathbf{f} + b_k - b_j\|_2$. We apply (A.2) to the matrix $Q_{j,k}$ and (A.1) to the vector $v_{j,k}$. For all $x > 0$, it yields that on an event $\Omega_{j,k}(x)$ of probability greater than $1 - 2\exp(-x)$,

$$\begin{aligned} \boldsymbol{\xi}^T Q_{j,k} \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q_{j,k} \boldsymbol{\xi}] &\leq \sigma^2 (12x + 6\|A_k - A_j\|_{\mathrm{HS}} \sqrt{x}), \\ &\leq 48\sigma^2 x + \frac{1}{4} \|A_k - A_j\|_{\mathrm{HS}} , \\ \text{and} \quad \boldsymbol{\xi}^T v_{j,k} &\leq \sigma\sqrt{x} 4 \|(A_k - A_j)\mathbf{f} + b_k - b_j\|_2 , \\ &\leq 8\sigma^2 x + \frac{1}{2} \|(A_k - A_j)\mathbf{f} + b_k - b_j\|_2^2 . \end{aligned}$$

On the event $\Omega_0 \cap \Omega_{j,k}(x + \log(1/(\pi_k \pi_j))$, we have $\zeta_{j,k} \leq \beta^* x$. Using the union bound, the probability of the event $\Omega_0 \cap (\cap_{j,k=1,\dots,M} \Omega_{j,k}(x + \log(1/(\pi_k \pi_j))))$ is at least

$$1 - \delta - \sum_{j,k=1,\dots,M} 2\exp(-x)\pi_j \pi_k = 1 - \delta - 2\exp(-x).$$

Finally, on this event $\max_{j,k=1,\dots,M} \zeta_{j,k} \leq \beta^* x$ and $2\hat{\beta} \leq 18\beta^*/7 = 144\sigma^2$ which completes the proof. □

B.4 Sparsity oracle inequality

In [28], the authors prove the following oracle inequality for the Least Squares estimator $\hat{\mu}_V^{LS}$ on a d-dimensional linear subspace V of $\mathbf{R}^n$. The Least Squares estimator $\hat{\mu}_V^{LS}$ is defined as the orthogonal projection of $\mathbf{y}$ on the linear subspace V.

Lemma B.2 ([28]). *Under Assumption 4.1, with probability greater than* $1 - \exp(-x)$*:*

$$\begin{aligned} \left\| \hat{\mu}_V^{LS} - \mathbf{f} \right\|_2^2 &\leq \min_{\mu \in V} \|\mu - \mathbf{f}\|_2^2 + K^2(d + 2\sqrt{dx} + 2x), \\ &\leq \min_{\mu \in V} \|\mu - \mathbf{f}\|_2^2 + K^2(2d + 3x). \end{aligned} \tag{B.14}$$

We now use this result to prove Proposition 4.1.

Proof of Proposition 4.1. Let $\hat{\beta} = 32\hat{K}^2$. Let $J^* = 1, ..., M$ be a deterministic integer. Since $\hat{\theta}$ minimizes $\hat{V}_n$ over the simplex and $\hat{V}_n$ is convex and differentiable, a simple consequence of the KKT conditions [11, 4.2.3, equation (4.21)] yields:

$$\nabla \hat{V}_n(\hat{\theta})^T (e_{J^*} - \hat{\theta}) \geq 0. \tag{B.15}$$

Let $W \coloneqq \nabla \hat{V}_n(\hat{\theta})^T (e_{J^*} - \hat{\theta})$. Using (B.1), (B.3) and some algebra, we obtain

$$\begin{aligned} W = &\|\hat{\mu}_{J^*} - \mathbf{f}\|_2^2 - \|\hat{\mu}_{\hat{\theta}} - \mathbf{f}\|_2^2 - 2\boldsymbol{\xi}^T(\hat{\mu}_{J^*} - \hat{\mu}_{\hat{\theta}}) \\ &- \tfrac{1}{2} \sum_{k=1}^M \hat{\theta}_k \|\hat{\mu}_k - \hat{\mu}_{J^*}\|_2^2 + \hat{\beta} \log \frac{1}{\pi_{J^*}} - \hat{\beta} \sum_{k=1}^M \hat{\theta}_k \log \frac{1}{\pi_k}. \end{aligned}$$

Inequality (B.15) can be rewritten as

$$\|\hat{\mu}_{\hat{\theta}} - \mathbf{f}\|_2^2 \leq \|\hat{\mu}_{J^*} - \mathbf{f}\|_2^2 + 2\hat{\beta} \log \frac{1}{\pi_{J^*}} + z(J^*, \hat{\theta})$$

where

$$z(J^*, \hat{\theta}) \coloneqq 2\boldsymbol{\xi}^T(\hat{\mu}_{\hat{\theta}} - \hat{\mu}_{J^*}) - \tfrac{1}{2} \sum_{k=1}^M \hat{\theta}_k \|\hat{\mu}_k - \hat{\mu}_{J^*}\|_2^2 - \hat{\beta} \log \frac{1}{\pi_{J^*}} - \hat{\beta} \sum_{k=1}^M \hat{\theta}_k \log \frac{1}{\pi_k}.$$

The function $z(J^*, \cdot)$ is affine in its second argument. Thus it is maximized at a vertex of Λ^M, and

$$z(J^*, \hat{\theta}) \leq \max_{\theta \in \Lambda^M} z(J^*, \theta) = \max_{k=1,\dots,M} z(J^*, e_k) \leq \max_{j,k=1,\dots,M} z(j, e_k).$$

As it holds for all deterministic $J^* = 1, ..., M$, we proved that

$$\|\hat{\mu}_{\hat{\theta}} - \mathbf{f}\|_2^2 \leq \min_{J^*=1,\dots,M} \left(\|\hat{\mu}_{J^*} - \mathbf{f}\|_2^2 + 2\hat{\beta} \log \frac{1}{\pi_{J^*}} \right) + \max_{j,k=1,\dots,M} \zeta_{jk},$$

where

$$\zeta_{jk} \coloneqq z(j, e_k) = \quad 2\boldsymbol{\xi}^T(\hat{\mu}_k - \hat{\mu}_j) - \hat{\beta} \log \frac{1}{\pi_j \pi_k} - \tfrac{1}{2} \|\hat{\mu}_k - \hat{\mu}_j\|_2^2 .$$

Let $B_{jk} = A_k - A_j$, and note that $\|\!|B_{jk}|\!\|_2 \leq 2$ because A_k and A_j are orthogonal projectors. Using $\hat{\mu}_k - \hat{\mu}_j = B_{jk}\boldsymbol{\xi} + (B_{jk}\mathbf{f} + b_k - b_j)$ and (B.7), we get

$$\zeta_{jk} = \quad 2\boldsymbol{\xi}^T(A_k - A_j)\boldsymbol{\xi} + \boldsymbol{\xi}^T \alpha_{jk} - \tfrac{1}{2} \|B_{jk}\mathbf{f} + b_k - b_j\|_2^2 - \tfrac{1}{2} \|B_{jk}\boldsymbol{\xi}\|_2^2 - \hat{\beta} \log \frac{1}{\pi_j \pi_k},$$

where $\alpha_{jk} \coloneqq 2(I_{n\times n} - \frac{1}{2}B_{jk}^T)(B_{jk}\mathbf{f} + b_k - b_j)$. The vector α_{jk} satisfies

$$\|\alpha_{jk}\|_2 \leq 2(1 + \tfrac{1}{2}\|\!|B_{jk}|\!\|_2) \|B_{jk}\mathbf{f} + b_k - b_j\|_2 \leq 4 \|B_{jk}\mathbf{f} + b_k - b_j\|_2 .$$

We also have $-\|B_{jk}\boldsymbol{\xi}\|_2^2 \leq 0$ almost surely.

Let $x > 0$. We now apply the concentration inequality (A.9) to the matrix $2B_{jk}$ and the Hoeffding-type inequality (A.7) to the vector α_{jk}. Using the union bound, the following holds with probability greater than $1 - 2\exp(-x)$:

$$\begin{aligned} \zeta_{jk} \leq \quad & K^2 \left(2 \|B_{jk}\|_1 + 4\|\!|B_{jk}|\!\|_2 x + 4 \|B_{jk}\|_{\mathrm{HS}} \sqrt{x} \right) \\ & + 2K(1 + \tfrac{1}{2}\|\!|B_{jk}|\!\|_2) \|B_{jk}\mathbf{f} + b_k - b_j\|_2 \sqrt{2x} - \tfrac{1}{2} \|B_{jk}\mathbf{f} + b_k - b_j\|_2^2 \\ & - \hat{\beta} \log \frac{1}{\pi_j \pi_k}. \end{aligned}$$

We upper bound the first line of the RHS of the previous display. By the triangle inequality, and the assumption $\mathrm{Tr}(A_j) \leq \log(\pi_j^{-1})$, we have $\|B_{jk}\|_1 \leq \mathrm{Tr}(A_j + A_k) \leq \log((\pi_j \pi_k)^{-1})$. Using simple inequalities,

$$\|B_{jk}\|_{\mathrm{HS}} \sqrt{x} \leq (\|A_j\|_{\mathrm{HS}} + \|A_k\|_{\mathrm{HS}})\sqrt{x} \leq (\|A_j\|_{\mathrm{HS}}^2 + \|A_k\|_{\mathrm{HS}}^2 + 2x)/2 \leq \frac{1}{2} \log \frac{1}{\pi_j \pi_k} + x.$$

Thus, $2 \|B_{jk}\|_1 + 4\|\!|B_{jk}|\!\|_2 x + 4 \|B_{jk}\|_{\mathrm{HS}} \sqrt{x} \leq K^2(12x + 4 \log \frac{1}{\pi_j \pi_k})$.

Now we upper bound the second line. We apply the inequality $st \leq \frac{s^2+t^2}{2}$ with $t = \|B_{jk}\mathbf{f} + b_k - b_j\|_2$ and $s = 2K(1 + \frac{1}{2}|||B_{jk}|||_2)\sqrt{2x}$:

$$\begin{aligned} &2K(1 + \tfrac{1}{2}|||B_{jk}|||_2) \|B_{jk}\mathbf{f} + b_k - b_j\|_2 \sqrt{2x} - \tfrac{1}{2} \|B_{jk}\mathbf{f} + b_k - b_j\|_2^2 \\ &\qquad = st - \frac{t^2}{2} \leq \frac{s^2}{2} = 4K^2(1 + \tfrac{1}{2}|||B_{jk}|||_2)^2 x \leq 16K^2 x. \end{aligned}$$

For any $x' > 0$, let $x_{jk} = x' + \frac{1}{\pi_j \pi_k}$. By setting $x = x_{jk}$, the above displays yield the following bound on ζ_{jk}, with probability greater than $1 - 2\pi_j\pi_k \exp(-x')$:

$$\zeta_{jk} \leq 28K^2 x_{jk} - (\hat{\beta} - 4K^2) \log \frac{1}{\pi_j \pi_k} = 28K^2 x' - (\hat{\beta} - 32K^2) \log \frac{1}{\pi_j \pi_k}.$$

Using a union bound, we obtain that on an event of probability greater than $1 - \delta - 2\sum_{j=1}^M \sum_{k=1}^M \pi_j \pi_k \exp(-x') = 1 - \delta - 2\exp(-x')$, we have $\hat{\beta} \geq 32K^2$ and

$$\max_{j,k=1,\dots,M} \zeta_{jk} \leq 28K^2 x'.$$

□

Proof of Theorem 4.1. Let $\bar{\theta} \in \mathbf{R}^p$ be a minimizer of the right hand side of (4.4) and let $\bar{J} \subset \{1, ..., p\}$ be the support of $\bar{\theta}$, hence $|\bar{\theta}|_0 = |\bar{J}|$. Since the RHS of (4.4) is random, $\bar{\theta}$ and its support are also random.

Let $t > 0$. For each support $J \subset \{1, ..., p\}$, the oracle inequality (B.14) applied to $x = t + \log(\pi_J^{-1})$ yields that with probability greater than $1 - \pi_J \exp(-t)$,

$$\left\|\hat{\mu}_J^{LS} - \mathbf{f}\right\|_2^2 \leq \left\|\mathbb{X}\bar{\theta} - \mathbf{f}\right\|_2^2 + K^2 \left(2|\bar{\theta}|_0 + 3\log\left(\frac{1}{\pi_J}\right) + 3t\right). \tag{B.16}$$

With the union bound, (B.16) holds simultaneously for all $J \subset \{1, ..., p\}$ with probability greater than $1 - \exp(-t) = 1 - \sum_{J \subset \{1,\dots,p\}} \pi_J \exp(-t)$.

We apply the oracle inequality of Proposition 4.1 and the oracle inequality (B.16) to $\hat{\mu}_{\bar{J}}^{LS}$. With the union bound, we have with probability greater than $1 - \delta - 3\exp(-t)$:

$$\begin{aligned} \left\|\mathbb{X}\hat{\theta}^{SPA} - \mathbf{f}\right\|_2^2 &\leq \left\|\hat{\mu}_{\bar{J}}^{LS} - \mathbf{f}\right\|_2^2 + 64\hat{K}^2 \log \frac{1}{\pi_{\bar{J}}} + 28K^2 t, \\ \left\|\hat{\mu}_{\bar{J}}^{LS} - \mathbf{f}\right\|_2^2 &\leq \left\|\mathbb{X}\bar{\theta} - \mathbf{f}\right\|_2^2 + K^2 \left(2|\bar{\theta}|_0 + 3\log\left(\frac{1}{\pi_{\bar{J}}}\right) + 3t\right), \end{aligned}$$

where $A_{\bar{J}}$ is the projection matrix such that $\hat{\mu}_{\bar{J}}^{LS} = A_{\bar{J}}\mathbf{y}$. The following bound can be found in with the following bound from [35, Section 5.2.1]:

$$\log \frac{1}{\pi_{\bar{J}}} \leq 2|\bar{\theta}|_0 \log\left(\frac{ep}{1 \vee |\bar{\theta}|_0}\right) + \frac{1}{2}.$$

Summing the two oracle inequalities above and applying the upper bound on $\log \frac{1}{\pi_{\bar{J}}}$ completes the proof. □

Acknowledgement

We would like to thank Alexandre Tsybakov for valuable comments on previous versions of this manuscript.

References

[1] Sylvain Arlot and Francis R Bach. Data-driven calibration of linear estimators with minimal penalties. In *Advances in Neural Information Processing Systems*, pages 46–54, 2009.

[2] Mark Bagnoli and Ted Bergstrom. Log-concave probability and its applications. *Economic theory*, 26(2):445–469, 2005.

[3] Yannick Baraud, Christophe Giraud, and Sylvie Huet. Estimator selection in the gaussian setting. 50(3):1092–1119, 2014.

[4] Franck Barthe and Emanuel Milman. Transference principles for log-sobolev and spectral-gap with applications to conservative spin systems. *Communications in Mathematical Physics*, 323(2):575–625, 2013. ISSN 0010-3616. doi: 10.1007/s00220-013-1782-2. URL http://dx.doi.org/10.1007/s00220-013-1782-2.

[5] Pierre C. Bellec. Optimal exponential bounds for aggregation of density estimators. *arXiv preprint arXiv:1405.3907*, 2014.

[6] Alexandre Belloni and Victor Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 05 2013. doi: 10.3150/11-BEJ410. URL http://dx.doi.org/10.3150/11-BEJ410.

[7] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Pivotal estimation via square-root lasso in nonparametric regression. *Ann. Statist.*, 42(2):757–788, 04 2014. doi: 10.1214/14-AOS1204. URL http://dx.doi.org/10.1214/14-AOS1204.

[8] Lucien Birgé and Pascal Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268, 2001.

[9] Lucien Birgé and Pascal Massart. Minimal penalties for gaussian model selection. *Probability Theory and Related Fields*, 138(1-2):33–73, 2007. ISSN 0178-8051. doi: 10.1007/s00440-006-0011-8. URL http://dx.doi.org/10.1007/s00440-006-0011-8.

[10] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[11] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2009.

[12] Arthur Cohen. All admissible linear estimates of the mean vector. *The Annals of Mathematical Statistics*, pages 458–463, 1966.

[13] D. Dai, P. Rigollet, and T. Zhang. Deviation optimal learning using greedy Q-aggregation. *The Annals of Statistics*, 40(3):1878–1905, 2012.

[14] D. Dai, P. Rigollet, Xia L., and Zhang T. Aggregation of affine estimators. *Electon. J. Stat.*, 8:302–327, 2014.

[15] Arnak S. Dalalyan and Joseph Salmon. Sharp oracle inequalities for aggregation of affine estimators. *The Annals of Statistics*, 40(4):2327–2355, 2012.

[16] Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Learning Theory*, volume 4539 of *Lecture Notes in Computer Science*, pages 97–111. Springer Berlin Heidelberg, 2007. ISBN 978-3-540-72925-9. doi: 10.1007/978-3-540-72927-3_9. URL http://dx.doi.org/10.1007/978-3-540-72927-3_9.

[17] Arnak S Dalalyan and Alexandre B Tsybakov. Sparse regression learning by aggregation and langevin monte-carlo. *Journal of Computer and System Sciences*, 78(5):1423–1443, 2012.

[18] Victor De la Pena and Evarist Giné. *Decoupling: from dependence to independence*. Springer, New York, 1999.

[19] Victor H. de la Pena and S. J. Montgomery-Smith. Decoupling inequalities for the tail probabilities of multivariate u-statistics. *The Annals of Probability*, 23(2):806–816, 04 1995. doi: 10.1214/aop/1176988291. URL http://dx.doi.org/10.1214/aop/1176988291.

[20] Holger Dette, Axel Munk, and Thorsten Wagner. Estimating the variance in nonparametric regression—what is a reasonable choice? *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(4):751–764, 1998.

[21] David L Donoho, Iain M Johnstone, Jeffrey C Hoch, and Alan S Stern. Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 41–81, 1992.

[22] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*. Springer, 2013.

[23] Sébastien Gerchinovitz. *Prediction of individual sequences and prediction in the statistical framework: some links around sparse regression and aggregation techniques*. PhD thesis, Université Paris Sud-Paris XI, 2011. URL https://tel.archives-ouvertes.fr/tel-00653550.

[24] Christophe Giraud. Mixing least-squares estimators when the variance is unknown. *Bernoulli*, 14(4):1089–1107, 2008.

[25] Christophe Giraud, Sylvie Huet, and Nicolas Verzelen. High-dimensional regression with unknown variance. *Statistical Science*, 27(4):500–518, 11 2012. doi: 10.1214/12-STS398. URL http://dx.doi.org/10.1214/12-STS398.

[26] Peter Hall, JW Kay, and DM Titterinton. Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, 77(3):521–528, 1990.

[27] David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3): 1079–1083, 1971.

[28] Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.*, 17:no. 52, 1–6, 2012. ISSN 1083-589X. doi: 10.1214/ECP.v17-2079. URL http://ecp.ejpecp.org/article/view/2079.

[29] John Immerkaer. Fast noise variance estimation. *Computer vision and image understanding*, 64(2):300–302, 1996.

[30] I. M. Johnstone. Function estimation and gaussian sequence models. *Unpublished manuscript*, 2(5.3):2, 2002.

[31] Guillaume Lecué and Philippe Rigollet. Optimal learning with Q-aggregation. *Ann. Statist.*, 42(1):211–224, 2014.

[32] Gilbert Leung and Andrew R. Barron. Information theory and mixing least-squares regressions. *Information Theory, IEEE Transactions on*, 52(8):3396–3410, 2006.

[33] Axel Munk, Nicolai Bissantz, Thorsten Wagner, and Gudrun Freitag. On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):19–41, 2005.

[34] Arkadi Nemirovski. Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1998)*, volume 1738 of *Lecture Notes in Mathematics*. Springer, Berlin, 2000.

[35] P. Rigollet and A. Tsybakov. Sparse estimation by exponential weighting. *Statistical Science*, 27(4):558–575, 2012.

[36] Philippe Rigollet and Alexandre Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011.

[37] Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electron. Commun. Probab.*, 18:no. 82, 1–9, 2013. ISSN 1083-589X. doi: 10.1214/ECP.v18-2865. URL http://ecp.ejpecp.org/article/view/2865.

[38] Warren Schudy and Maxim Sviridenko. Concentration and moment inequalities for polynomials of independent random variables. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 437–446. SIAM, 2012.

[39] Vladimir Spokoiny and Mayya Zhilova. Sharp deviation bounds for quadratic forms. *Mathematical Methods of Statistics*, 22(2):100–113, 2013.

[40] Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 2012.

[41] A.B. Tsybakov. Aggregation and minimax optimality in high-dimensional estimation. In *Proceedings of the International Congress of Mathematicians*, Seoul, 2014. To appear.

[42] Alexandre B. Tsybakov. Optimal rates of aggregation. In *Learning Theory and Kernel Machines*, pages 303–313. Springer, 2003.

[43] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[44] Roman Vershynin. A simple decoupling inequality in probability theory. 2011. URL http://www-personal.umich.edu/~romanv/papers/decoupling-simple.pdf.

[45] Farrol Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables whose distributions are not necessarily symmetric. *The Annals of Probability*, 1(6):1068–1070, 1973.

Supplement: additional proofs

C Bounds on moment generating functions

The condition (3.15) leads to the following bounds on the moment generating functions of X and X^2, which are crucial to prove Theorem 3.4.

Proposition C.1. *Let $K > 0$ and let ξ_i be a random variable satisfying (3.15) with $\sigma_i^2 = \mathbb{E}[\xi_i^2]$. Then for all $s \in \mathbf{R}$:*

$$\mathbb{E}\exp(s\xi_i) \leq \exp(s^2K^2). \tag{C.1}$$

Furthermore, if $0 \leq 2sK^2 \leq 1$, then

$$\mathbb{E}\exp(s\xi_i^2 - s\sigma_i^2) \leq \exp(s^2\sigma_i^2K^2), \tag{C.2}$$

$$\mathbb{E}\exp(s\xi_i^2) \leq \exp\left(\frac{3}{2}s\sigma_i^2\right). \tag{C.3}$$

Inequality (C.1) shows that a random variable X satisfying the moment assumption (3.15) is subgaussian and its ψ_2 norm is bounded by K up to a multiplicative absolute constant. The proof of Proposition C.1 is based on Taylor expansions and some algebra.

Proof of Proposition C.1. To simplify the notation, let $X = \xi_i$ and $\sigma = \sigma_i$. We first prove (C.2). We apply the assumption on the even moments of X:

$$\begin{aligned}\mathbb{E}\exp(sX^2) &= 1 + s\sigma^2 + \sum_{p\geq 2}\frac{s^p\mathbb{E}X^{2p}}{p!},\\ &\leq 1 + s\sigma^2 + \frac{\sigma^2 s}{2}\sum_{k=1}^{\infty}(sK^2)^k = 1 + s\sigma^2 + \frac{\sigma^2K^2s^2}{2(1-sK^2)},\end{aligned}$$

and using the inequality $0 < 2sK^2 \leq 1$, we obtain:

$$\mathbb{E}\exp(sX^2) \leq 1 + s\sigma^2 + \sigma^2s^2K^2 \leq \exp(s\sigma^2 + s^2\sigma^2K^2),$$

which completes the proof of (C.2). Inequality (C.3) is a direct consequence of (C.2) after applying again the inequality $2sK^2 \leq 1$.

We now prove (C.1). Using the Cauchy-Schwarz inequality and the assumption on the moments for $p = 2$, we get $\sigma^4 \leq \mathbb{E}[\xi^4] \leq \sigma^2K^2$, so $\sigma \leq K$. Let $p \geq 1$. For the even terms of the expansion of $\mathbb{E}\exp(sX)$, we get:

$$\frac{s^{2p}\mathbb{E}X^{2p}}{(2p)!} \leq \tfrac{1}{2}(sK)^{2p}\frac{p!}{(2p)!} \leq \tfrac{1}{2}\frac{(sK)^{2p}}{p!},$$

where for the last inequality we used $(p!)^2 \leq (2p)!$. For the odd terms, by using the Jensen inequality for $p \geq 1$:

$$\frac{s^{2p+1}\mathbb{E}X^{2p+1}}{(2p+1)!} \leq \frac{s^{2p+1}(\mathbb{E}X^{2p+2})^{\frac{2p+1}{2p+2}}}{(2p+1)!} \leq |sK|^{2p+1}\frac{\left(\frac{(p+1)!}{2}\right)^{\frac{2p+1}{2p+2}}}{(2p+1)!},$$
$$\leq \tfrac{1}{2}|sK|^{2p+1}\frac{(p+1)!}{(2p+1)!}.$$

If $|sK| > 1$, we use the inequality $(p+1)!^2 \leq (2p+1)!$ to obtain

$$\frac{s^{2p+1}\mathbb{E}X^{2p+1}}{(2p+1)!} \leq \frac{|sK|^{2(p+1)}}{2((p+1)!)},$$

and by combining the inequality for the even and the odd terms:

$$\begin{aligned}\mathbb{E}\exp(sX) &= 1 + \sum_{p\geq 1}\frac{s^{2p}\mathbb{E}X^{2p}}{(2p)!} + \frac{s^{2p+1}\mathbb{E}X^{2p+1}}{(2p+1)!},\\ &\leq 1 + \tfrac{1}{2}\sum_{p\geq 1}\frac{(sK)^{2p}}{p!} + \frac{|sK|^{2(p+1)}}{(p+1)!},\\ &\leq 1 + \sum_{p\geq 1}\frac{(sK)^{2p}}{p!} = \exp(s^2K^2).\end{aligned}$$

If $|sK| \leq 1$, we use the inequality $(p+1)!p! \leq (2p+1)!$ to obtain

$$\frac{s^{2p+1}\mathbb{E}X^{2p+1}}{(2p+1)!} \leq \frac{(sK)^{2p}}{2(p!)},$$

and by combining the inequality for the even and the odd terms:

$$\begin{aligned}\mathbb{E}\exp(sX) &= 1 + \sum_{p\geq 1}\frac{s^{2p}\mathbb{E}X^{2p}}{(2p)!} + \frac{s^{2p+1}\mathbb{E}X^{2p+1}}{(2p+1)!},\\ &\leq 1 + \tfrac{1}{2}\sum_{p\geq 1}\frac{(sK)^{2p}}{p!} + \frac{(sK)^{2p}}{p!} = 1 + \sum_{p\geq 1}\frac{(sK)^{2p}}{p!} = \exp(s^2K^2).\end{aligned}$$

□

D Proof of Proposition A.4

Proof. It is a direct application of (C.1) combined with the Chernoff bound. Let $s = 2K\|v\|_2\sqrt{x}$. The Chernoff bound and the independence of $\xi_1, ..., \xi_n$ yield, for all $\lambda > 0$,

$$\mathbb{P}\left(v^T\boldsymbol{\xi} > s\right) \leq \exp(-\lambda s)\prod_{i=1}^{n}\mathbb{E}[\exp(\lambda v_i\xi_i)] \leq \exp(-\lambda s + \lambda^2K^2\|v\|_2^2).$$

Now set $\lambda = \sqrt{x}/(K\|v\|_2)$ such that the RHS of the previous display is minimized, and the proof is complete. □

E Proof of Theorem 3.4

The goal of this section is to prove Theorem 3.4. We start with preliminary calculations that will be useful in the proof. Let A be any $n \times n$ real matrix. Let $\lambda > 0$ satisfy

$$128|||A|||_2 K^2 \lambda \leq 1, \tag{E.1}$$

and define

$$\eta = 32K^2\lambda^2. \tag{E.2}$$

The inequality (E.1) can be rewritten in terms of η:

$$512K^2|||A|||_2^2 \eta \leq 1. \tag{E.3}$$

Let A_0 be the matrix A with the diagonal entries set to 0. Then, using the triangle inequality with $A_0 = A - \mathrm{diag}(a_{11}, ..., a_{nn})$ and $|a_{ii}| \leq |||A|||_2$ for all $i = 1, ..., n$, we obtain

$$|||A_0|||_2 \leq 2|||A|||_2. \tag{E.4}$$

Let $B = A_0^T A_0 = (b_{ij})_{i,j=1,...,n}$ and let B_0 be the matrix B with the diagonal entries set to 0. Then

$$\forall i = 1, ..., n, \qquad 0 \leq b_{ii} = \sum_{j \neq i} a_{ji}^2 \leq |||A|||_2^2. \tag{E.5}$$

By using the decomposition $B_0 = B - \mathrm{diag}(b_{11}, ..., b_{nn})$ and the inequality $\|v + v'\|_2^2 \leq 2\|v\|_2^2 + 2\|v'\|_2^2$, (E.5) and (E.4), we have:

$$\begin{aligned}
\|B_0\boldsymbol{\xi}\|_2^2 &\leq 2\|B\boldsymbol{\xi}\|_2^2 + 2\sum_{i=1}^n b_{ii}^2\xi_i^2, \\
&\leq 2|||A_0|||_2^2 \|A_0\boldsymbol{\xi}\|_2^2 + 2|||A|||_2^2 \sum_{i=1}^n b_{ii}\xi_i^2, \\
&\leq 8|||A|||_2^2 \|A_0\boldsymbol{\xi}\|_2^2 + 2|||A|||_2^2 \sum_{i=1}^n b_{ii}\xi_i^2.
\end{aligned}$$

Combining the previous display with (E.3), we obtain for any $K > 0$:

$$\begin{aligned}
16K^2\eta^2 \|B_0\boldsymbol{\xi}\|_2^2 &\leq (512K^2|||A|||_2^2\eta)\left(\frac{\eta}{4}\|A_0\boldsymbol{\xi}\|_2^2 + \frac{\eta}{16}\sum_{i=1}^n b_{ii}\xi_i^2\right), \\
&\leq \frac{\eta}{4}\|A_0\boldsymbol{\xi}\|_2^2 + \frac{\eta}{16}\sum_{i=1}^n b_{ii}\xi_i^2.
\end{aligned} \tag{E.6}$$

Proof of Theorem 3.4. Throughout the proof, let $\lambda > 0$ satisfy (E.1). The value of λ will be specified later.

First we treat the diagonal terms by bounding the moment generating function of

$$S_{\text{diag}} \coloneqq \sum_{i=1}^{n} a_{ii}\xi_i^2 - \sum_{i=1}^{n} a_{ii}\sigma_i^2.$$

Using the independence of $\xi_1, ..., \xi_n$ and (C.2) with $s = a_{ii}\lambda$ with each $i = 1, ..., n$:

$$\mathbb{E}\exp(\lambda S_{\text{diag}}) \leq \exp\left(\lambda^2 \sum_{i=1}^{n} a_{ii}^2\sigma_i^2 K^2\right), \tag{E.7}$$

provided that for all $i = 1, ..., n$, $2|a_{ii}|\lambda K^2 \leq 1$ which is satisfied as (E.1) holds and $|a_{ii}| \leq |||A|||_2$.

Now we bound the moment generating function of the off-diagonal terms. Let

$$S_{\text{off-diag}} \coloneqq \sum_{i,j=1,...,n:i\neq j} a_{ij}\xi_i\xi_j.$$

Let the random vector $\boldsymbol{\xi}' = (\xi_1', ..., \xi_n')^T$ be independent of $\boldsymbol{\xi}$ with the same distribution as $\boldsymbol{\xi}$. We apply the decoupling inequality [44] (see also [22, Theorem 8.11]) to the convex function $s \to \exp(\lambda s)$:

$$\mathbb{E}\exp(\lambda S_{\text{off-diag}}) \leq \mathbb{E}\exp\left(4\lambda \sum_{i,j=1,...,n:i\neq j} a_{ij}\xi_i'\xi_j\right).$$

Conditionally on $\xi_1, ..., \xi_n$, for each $i = 1, ..., n$, we use the independence of $\xi_1', ..., \xi_n'$ and (C.1) applied to ξ_i' with $s = 4\sum_{j=1,...,n:i\neq j} a_{ij}\xi_j$:

$$\begin{aligned}\mathbb{E}\exp\left(4\lambda \sum_{i\neq j} a_{ij}\xi_i'\xi_j\right) &\leq \mathbb{E}\exp\left(16K^2\lambda^2 \sum_{i=1,...,n}\left(\sum_{j=1,...,n:i\neq j} a_{ij}\xi_j\right)^2\right), \\ &= \mathbb{E}\exp\left(16K^2\lambda^2 \|A_0\boldsymbol{\xi}\|_2^2\right) = \mathbb{E}\exp\left(\frac{\eta}{2}\|A_0\boldsymbol{\xi}\|_2^2\right),\end{aligned}$$

where η is defined in (E.2) and A_0 is the matrix A with the diagonal entries set to 0. Let $B = A_0^T A_0 = (b_{ij})_{i,j=1,...,n}$. Then $\|A_0\boldsymbol{\xi}\|_2^2 = \sum_{i=1}^n b_{ii}\xi_i^2 + \sum_{i\neq j} b_{ij}\xi_i\xi_j$.

We use the Cauchy-Schwarz inequality to separate the diagonal terms from the off-diagonal ones:

$$\left(\mathbb{E}\exp(\frac{\eta}{2}\|A_0\boldsymbol{\xi}\|_2^2)\right)^2 \leq \mathbb{E}\exp\left(\eta\sum_{i=1}^{n} b_{ii}\xi_i^2\right)\mathbb{E}\exp\left(\eta\sum_{i\neq j} b_{ij}\xi_i\xi_j\right). \tag{E.8}$$

For the off-diagonal terms of (E.8), using the decoupling inequality [44] (see also [22, Theorem 8.11]) we have:

$$\mathbb{E}\exp\left(\eta\sum_{i\neq j}b_{ij}\xi_i\xi_j\right)\leq\mathbb{E}\exp\left(4\eta\sum_{i\neq j}b_{ij}\xi'_i\xi_j\right).$$

Again, conditionally on $\xi_1,...,\xi_n$, for each $j=1,...,n$, we use (C.1) applied to ξ'_i and the independence of $\xi'_1,...,\xi'_n$:

$$\begin{aligned}\mathbb{E}\exp\left(4\eta\sum_{i\neq j}b_{ij}\xi'_i\xi_j\right)&\leq\mathbb{E}\exp\left(16K^2\eta^2\sum_{i=1}^n\left(\sum_{j=1,...,n:\ i\neq j}b_{ij}\xi_j\right)^2\right),\\&=\mathbb{E}\exp\left(16K^2\eta^2\left\|B_0\boldsymbol{\xi}\right\|_2^2\right),\\&\leq\mathbb{E}\exp\left(\frac{\eta}{4}\left\|A_0\boldsymbol{\xi}\right\|_2^2+\frac{\eta}{16}\sum_{i=1}^n b_{ii}\xi_i^2\right),\end{aligned}$$

where we used the preliminary calculation (E.6) for the last display. Finally, the Cauchy-Schwarz inequality yields

$$\mathbb{E}\exp\left(4\eta\sum_{i\neq j}b_{ij}\xi_i\xi'_j\right)\leq\sqrt{\mathbb{E}\exp\left(\frac{\eta}{2}\left\|A_0\boldsymbol{\xi}\right\|_2^2\right)}\sqrt{\mathbb{E}\exp\left(\frac{\eta}{8}\sum_{i=1}^n b_{ii}\xi_i^2\right)}.$$

We plug this upper bound back into (E.8). After rearranging, we find

$$\left(\mathbb{E}\exp(\frac{\eta}{2}\left\|A_0\boldsymbol{\xi}\right\|_2^2)\right)^{3/2}\leq\mathbb{E}\exp\left(\eta\sum_{i=1}^n b_{ii}\xi_i^2\right)\sqrt{\mathbb{E}\exp\left(\frac{\eta}{8}\sum_{i=1}^n b_{ii}\xi_i^2\right)}.$$

As $b_{ii}\geq 0$, this implies:

$$\mathbb{E}\exp(\frac{\eta}{2}\left\|A_0\boldsymbol{\xi}\right\|_2^2)\leq\mathbb{E}\exp\left(\eta\sum_{i=1}^n b_{ii}\xi_i^2\right).$$

For each $i=1,...,n$, we apply (C.3) to the variable ξ_i with $s=b_{ii}\eta\geq 0$. Using the independence of $\xi_1^2,...,\xi_n^2$, we obtain:

$$\begin{aligned}\mathbb{E}\exp\left(\eta\sum_{i=1}^n b_{ii}\xi_i^2\right)&=\prod_{i=1}^n\mathbb{E}\exp(\eta b_{ii}\xi_i^2),\\&\leq\exp\left(\frac{3}{2}\eta\sum_{i=1}^n b_{ii}\sigma_i^2\right)=\exp\left(\frac{3}{2}\eta\left\|A_0D_\sigma\right\|_{\mathrm{HS}}^2\right).\end{aligned}$$

provided that for all $i = 1, ..., n$, $2K^2 b_{ii}\eta \leq 1$ which is satisfied thanks to (E.1) and (E.5).

We remove η from the above displays using its definition (E.2):

$$\mathbb{E}\exp(\lambda S_{\text{off}-\text{diag}}) \leq \exp\left(48\lambda^2 K^2 \|A_0 D_\sigma\|_{\text{HS}}^2\right), \tag{E.9}$$

where A_0 is the matrix A with the diagonal entries set to 0.

Now we combine the bound on the moment generating function of S_{diag} and $S_{\text{off}-\text{diag}}$, given respectively in (E.7) and (E.9). Using the Chernoff bound and the Cauchy-Schwarz inequality: we have that for all λ satisfying (E.1),

$$\begin{aligned}
\mathbb{P}\left(S_{\text{diag}} + S_{\text{off}-\text{diag}} > t\right) &\leq \exp(-\lambda t)\mathbb{E}[\exp(\lambda S_{\text{diag}})\exp(\lambda S_{\text{off}-\text{diag}})], \\
&\leq \exp\left(-\lambda t\right)\sqrt{\mathbb{E}[\exp(2\lambda S_{\text{diag}})]}\sqrt{\mathbb{E}[\exp(2\lambda S_{\text{off}-\text{diag}})]}, \\
&\leq \exp\left(-\lambda t + \lambda^2 K^2\left(\sum_{i=1}^n \sigma_i^2 a_{ii}^2 + 48\|A_0 D_\sigma\|_{\text{HS}}^2\right)\right), \\
&\leq \exp\left(-\lambda t + 48\lambda^2 K^2 \|AD_\sigma\|_{\text{HS}}^2\right), \qquad \text{(E.10)}
\end{aligned}$$

where for the last display we used the equality

$$\|AD_\sigma\|_{\text{HS}}^2 = \sum_{i,j=1,\ldots,n} a_{ij}^2\sigma_i^2 = \|A_0 D_\sigma\|_{\text{HS}} + \sum_{i=1}^n a_{ii}^2\sigma_i^2.$$

It now remains to choose the parameter λ. The unconstrained minimum of (E.10) is attained at $\bar{\lambda} = t/(96K^2\|AD_\sigma\|_{\text{HS}}^2)$. If $\bar{\lambda}$ satisfies the constraint (E.1), then

$$\mathbb{P}\left(S_{\text{diag}} + S_{\text{off}-\text{diag}} > t\right) \leq \exp\left(\frac{-t^2}{192K^2\|AD_\sigma\|_{\text{HS}}^2}\right).$$

On the other hand, if $\bar{\lambda}$ does not satisfy (E.1), then the constraint (E.1) is binding and the minimum of (E.10) is attained at $\lambda_{\text{b}} = 1/(128|||A|||_2 K^2) < \bar{\lambda}$. In this case,

$$-t\lambda_{\text{b}} + \lambda_{\text{b}}^2 48K^2\|AD_\sigma\|_{\text{HS}}^2 \leq -t\lambda_{\text{b}} + \lambda_{\text{b}}\bar{\lambda}48K^2\|AD_\sigma\|_{\text{HS}}^2 = -t\lambda_{\text{b}} + \frac{t}{2}\lambda_{\text{b}} = -\frac{t}{256K^2|||A|||_2}.$$

Combining the two regimes, we obtain

$$\mathbb{P}\left(S_{\text{diag}} + S_{\text{off}-\text{diag}} > t\right) \leq \exp\left(-\min\left(\frac{t^2}{192K^2\|AD_\sigma\|_{\text{HS}}^2}, \frac{t}{256K^2|||A|||_2}\right)\right).$$

The proof of (3.22) is complete.

Now we prove (3.23). The function

$$t \to x(t) = \min\left(\frac{t^2}{192K^2\|AD_\sigma\|_{\text{HS}}^2}, \frac{t}{256K^2|||A|||_2}\right)$$

is increasing and bijective from the set of positive real numbers to itself. Furthermore, for all $t > 0$,

$$t \leq 8\sqrt{3}K \|AD_\sigma\|_{\mathrm{HS}} \sqrt{x(t)} + 256K^2 |||A|||_2 x(t),$$

so the variable change $x = x(t)$ completes the proof of (3.23). □