

# Sharp optimal recovery in the Two Gaussian Mixture Model

Mohamed Ndaoud

MOHAMED.NDAOUD@ENSAE.FR

*Department of Statistics*

*CREST (UMR CNRS 9194), ENSAE*

*5, av. Henry Le Chatelier, 91764 Palaiseau, FRANCE*

## Abstract

In this paper, we study the non-asymptotic problem of exact recovery in the Two Gaussian mixture model when the centers are separated by some  $\Delta > 0$ . We present an optimal lower bound for the corresponding minimax Hamming risk improving on existing results. We also propose an optimal, efficient and adaptive procedure that is minimax rate optimal. Our procedure is based on a variant of Lloyd's iterations initialized by a spectral method. As a consequence of our study, we recover a sharp phase transition for the problem of exact recovery in the Gaussian mixture model. The latter phase transition happens around the critical threshold  $\Delta^*$  given by

$$\Delta^{*2} = \sigma^2 \log n \left( 1 + \sqrt{1 + 2 \frac{p}{n \log n}} \right).$$

## 1. Introduction

The problem of supervised or unsupervised clustering have gained huge interest in machine learning. In particular, many clustering algorithms are known to achieve good empirical results. A very useful model to study and compare these algorithms is the Gaussian mixture model. In this model, we assume that data are attributed to different centers and that we only have access to a noisy observation of data. For this specific model, authors considered either the problem of estimation of the centers Klusowski and Brinda (2016),Mixon et al. (2016) or of recovering the communities Lu and Zhou (2016),Giraud and Verzelen (2018),Royer (2017). This paper mainly focuses on the question related to community detection.

### 1.1 The Gaussian Mixture Model

We observe  $n$  independent random vectors  $Y_1, \dots, Y_n \in \mathbf{R}^p$ . We assume that there exists two unknown vectors  $\boldsymbol{\theta} \in \mathbf{R}^p$  and  $\eta \in \{-1, 1\}^n$ , such that for all  $i = 1, \dots, n$

$$Y_i = \boldsymbol{\theta} \eta_i + \sigma \xi_i, \tag{1}$$

where  $\sigma > 0$  and  $\xi_i$  is a standard Gaussian random vector. We denote by  $Y$  (resp.  $W$ ) the matrix with columns  $Y_1, \dots, Y_n$  (resp.  $W_1, \dots, W_n$ ). Model (1) is a special case of the general spiked model

$$Y = \boldsymbol{\theta} \boldsymbol{\eta}^\top + \sigma W.$$

We denote by  $\mathbf{P}_{(\boldsymbol{\theta}, \eta)}$  the distribution of  $Y$  in model (1) and by  $\mathbf{E}_{(\boldsymbol{\theta}, \eta)}$  the corresponding expectation. We assume that  $(\boldsymbol{\theta}, \eta)$  belongs to the following set

$$\Omega_\Delta = \{\boldsymbol{\theta} \in \mathbf{R}^p : \|\boldsymbol{\theta}\| \geq \Delta\} \times \{-1, 1\}^n,$$

where  $\Delta > 0$  is a given constant. The value  $\Delta$  characterizes the separation between the clusters and equivalently the strength of the signal.

In this paper, we study the problem of recovering the communities, that is, of estimating the vector  $\eta$ . As estimators of  $\eta$ , we consider any measurable functions  $\hat{\eta} = \hat{\eta}(Y_1, \dots, Y_n)$  of  $(Y_1, \dots, Y_n)$  taking values in  $\{-1, 1\}^n$ . We characterize the loss of a given  $\hat{\eta}$  as an estimator of  $\eta$  by the Hamming distance between  $\hat{\eta}$  and  $\eta$ , that is, by the number of positions at which  $\hat{\eta}$  and  $\eta$  differ:

$$|\hat{\eta} - \eta| \triangleq \sum_{j=1}^n |\hat{\eta}_j - \eta_j| = 2 \sum_{j=1}^n \mathbf{1}(\hat{\eta}_j \neq \eta_j).$$

Here,  $\hat{\eta}_j$  and  $\eta_j$  are the  $j$ th components of  $\hat{\eta}$  and  $\eta$ , respectively. Since  $\eta$  is only defined up to a sign change, it is more appropriate to consider the loss defined by

$$r(\hat{\eta}, \eta) := \min_{\nu \in \{-1, +1\}} |\hat{\eta} - \nu \eta|.$$

The expected loss of a given  $\hat{\eta}$  is defined as  $\mathbf{E}_{(\boldsymbol{\theta}, \eta)} r(\hat{\eta}, \eta)$ .

In the rest of the paper, we will always denote by  $\eta$  the vector to estimate, while  $\hat{\eta}$  will denote the corresponding estimator. We consider the following minimax risk

$$\Psi_\Delta := \inf_{\hat{\eta}} \sup_{(\boldsymbol{\theta}, \eta) \in \Omega_\Delta} \frac{1}{n} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} r(\hat{\eta}, \eta), \quad (2)$$

where  $\inf_{\hat{\eta}}$  denotes the infimum over all measurable estimators  $\hat{\eta}$  in  $\{-1, +1\}^n$ . A trivial lower bound on  $\Psi_\Delta$ , that we prove later, is given by

$$\Psi_\Delta \geq \frac{c}{1 + \Delta} e^{-\Delta^2/2} \quad (3)$$

for some  $c > 0$ . Inspecting the proof one may also notice that the corresponding oracle estimator  $\eta^*$  is given by

$$\eta_i^* = \text{sign} \left( Y_i^\top \boldsymbol{\theta} \right).$$

This oracle estimator assumes a prior knowledge of  $\boldsymbol{\theta}$ . It turns out that for  $p \geq n$ , there exists a regime where this lower bound is possibly not optimal, as pointed by Giraud and Verzelen (2018). The intuition being that for  $p$  larger than  $n$ ,  $\boldsymbol{\theta}$  is hard to estimate. To the best of our knowledge, there are no lower bounds for  $\Psi_\Delta$  that capture the estimation issue. This is one of the main questions addressed in the present paper.

**Notation.** In the rest of this paper we use the following notation. For given sequences  $a_n$  and  $b_n$ , we say that  $a_n = \mathcal{O}(b_n)$  (resp  $a_n = \Omega(b_n)$ ) when  $a_n \leq cb_n$  (resp  $a_n \geq cb_n$ ) for some absolute constant  $c > 0$ . We write  $a_n \asymp b_n$  when  $a_n = \mathcal{O}(b_n)$  and  $a_n = \Omega(b_n)$ . For  $x, y \in \mathbf{R}^p$ , we denote by  $x^\top y$  the Euclidean scalar product and  $\|x\|$  the corresponding norm of  $x$ . For  $x, y \in \mathbf{R}$ , we denote by  $x \vee y$  the maximum value between  $x$  and  $y$ . In particular  $x \vee 0$  will be denoted by  $x_+$ . To any matrix  $M \in \mathbf{R}^{n \times n}$ , we denote  $\|M\|_{op}$  its operator norm with respect to the  $L^2$ -norm and by  $\text{Tr}(M)$  its trace.  $\mathbf{I}_n$  denotes the identity matrix of dimension  $n$ . For  $z$  a standard Gaussian random variable, we denote by  $\Phi^c$  its complementary cumulative distribution function i.e.  $\forall t \in \mathbf{R}, \Phi^c(t) = \mathbf{P}(z > t)$ .

We assume that  $p, \sigma$  and  $\Delta$  depend on  $n$  and the asymptotic results correspond to the limit  $n \rightarrow \infty$ . All proofs are deferred to the Appendix.

## 1.2 Related literature

The present work can be confronted to two parallel lines of work.

### 1. Recovery in the sub-Gaussian mixture model:

Lu and Zhou (2016) are probably the first to present statistical guarantees for recovery in the sub-Gaussian mixture model using the well-known Lloyd’s algorithm Lloyd (1982). Their results require a better initialization than a random guess in addition to the condition

$$\Delta^2 = \Omega \left( \sigma^2 \left( 1 \vee \frac{p}{n} \right) \right),$$

in order to achieve *partial recovery* and

$$\Delta^2 = \Omega \left( \sigma^2 \log n \left( 1 \vee \frac{p}{n} \right) \right),$$

in order to achieve *exact recovery*. Both notions of partial and exact recovery are defined later. More recent work of Royer (2017) and Giraud and Verzelen (2018) have shown that previous conditions are not optimal in high dimension i.e.  $n = o(p)$ . In particular, in Giraud and Verzelen (2018), authors study an SDP relaxation of the Kmeans criterion that achieves *partial recovery* with a milder condition

$$\Delta^2 = \Omega \left( \sigma^2 \left( 1 \vee \sqrt{\frac{p}{n}} \right) \right), \quad (4)$$

and *exact recovery* given

$$\Delta^2 = \Omega \left( \sigma^2 \left( \log n \vee \sqrt{\log n \frac{p}{n}} \right) \right). \quad (5)$$

The previous conditions are the mildest in the literature, to the best of our knowledge, but no matching necessary conditions are known so far. In Giraud and Verzelen (2018), a good heuristic about optimality of their condition is discussed. In the supervised setting, where all labels are known similar conditions seem necessary to achieve either partial or exact recovery. It is still not clear whether optimal conditions in supervised mixture learning are also optimal in the unsupervised setting.

Another difference between previous works is computational. While, in Giraud and Verzelen (2018), an SDP relaxation is proposed, a faster algorithm through Lloyd’s is proposed in Lu and Zhou (2016). It remains not clear whether we can achieve partial (resp. exact) recovery under requirement (4) (resp. (5)) through faster methods compared to SDP relaxations, for instance through Lloyd’s iterations.

In Lu and Zhou (2016), authors suggest to initialize the Lloyd’s algorithm using a spectral method. It would be interesting to investigate whether Lloyd’s algorithm initialized by a spectral method, in the same spirit than Vempala and Wang (2004), can achieve optimal performance in the more general setting where  $p$  is allowed to be larger than  $n$ .

In this paper, we shed some light on these recent works. Specifically, we address the following questions that motivate the present work.

- Are conditions (4) and (5) necessary for both partial and exact recovery?
- Are optimal requirements similar in both supervised and unsupervised settings?
- Can we achieve similar results compared to Giraud and Verzelen (2018) using a faster algorithm?
- In case the answer to previous questions is positive, can we achieve the same results adaptively to all parameters?

## 2. Recovery in the Stochastic Block Model:

The Stochastic Block Model Holland et al. (1983) is probably the most popular framework for node clustering. This model with two communities can be seen as a particular case of model (1) when both the signal and the noise are symmetric matrices. A non symmetric variant of the SBM is the bipartite stochastic block model Feldman et al. (2015). Unlike the case of sub-Gaussian mixtures where most results are non-asymptotic, results of partial or exact recovery are mostly asymptotic for the SBM and its variants. This is due to the growing interest in sharp phase transitions in the SBM. In Abbe (2017), an open question is whether it is possible to characterize sharp phase transitions in other problems for instance in the Gaussian mixture model.

The first polynomial method achieving exact recovery in the SBM with two communities is due to Abbe et al. (2014). The algorithm splits the initial sample into two independent samples. A black-box algorithm is used on the first sample for partial recovery, then a local improvement is applied on the second sample. As stated in Abbe et al. (2014), it is not clear whether algorithms achieving partial recovery can be used to achieve exact recovery. It remains interesting to understand whether similar results can be achieved through direct algorithms ideally without the splitting step.

For the Bipartite SBM, sufficient computational conditions for exact recovery are presented in Feldman et al. (2015), Florescu and Perkins (2016). While the sharp phase transition for the problem of detection is fully answered in Florescu and Perkins (2016). It is still not clear whether these condition for exact recovery are optimal. More interestingly, the sufficient condition for exact recovery is different for  $p$  of the same order than  $n$  and for  $p$  larger than  $n^2$  for instance. This shows somehow a phase transition with respect to  $p$ , where for some critical dimension  $p^*$  the hardness of the problem changes.

We resume potential connections between our work and these recent developments in the following questions.

- Is it possible to characterize a sharp phase transition for exact recovery in model (1)?
- Are algorithms achieving partial recovery useful in order to achieve exact recovery in the sub-Gaussian mixture model?
- Are sufficient conditions for exact recovery in the Bipartite SBM optimal?
- Is there a critical dimension  $p^*$  that separates different regimes of hardness in the problem of exact recovery?

### 1.3 Main contribution

In this work, we provide a full answer to the question of exact recovery in the Gaussian mixture model. In particular, we give non-asymptotic lower bounds for the risk  $\Psi_\Delta$  and matching upper bounds through a variant of Lloyd's iterations initialized by a spectral method. To do so, we define a key quantity  $\mathbf{r}_n$  that turns out to be the right SNR of the problem. Define

$$\mathbf{r}_n = \frac{\Delta^2/\sigma^2}{\sqrt{\Delta^2/\sigma^2 + p/n}}.$$

This SNR is strictly larger than the expected one  $\Delta/\sigma$ . In particular, it states that hardness of the problem depends on the dimension  $p$ . Among other results we prove that for some  $c_1, c_2, C_1, C_2 > 0$ , we have

$$C_1 e^{-c_1 \mathbf{r}_n^2} \leq \Psi_\Delta \leq C_2 e^{-c_2 \mathbf{r}_n^2}.$$

Moreover, we give a sharp characterization of the previous constants.

Inspecting the proofs of lower bounds in Section 2, one may learn that, in a setting where no prior information on  $\boldsymbol{\theta}$  is given, the supervised learning estimator is optimal. Interestingly, supervised and unsupervised risks are almost equal, and the problem of community detection in the Gaussian mixture model is transparent to any prior information on labels as long as the centers are unknown.

As for the upper bound, we present and analyze a fully adaptive rate optimal computational procedure. In order to achieve optimal decay of the risk, it turns out that it is enough to consider  $\mathbf{H}(Y^\top Y)$  where for any squared matrix  $M$ ,  $\mathbf{H}(M) = M - \text{diag}(M)$  and  $\text{diag}(M)$  is the diagonal of  $M$ . We set  $\eta^0$  such that  $\eta^0 = \text{sign}(\hat{v})$  and  $\hat{v}$  is the eigenvector corresponding to the top eigenvalue of  $\mathbf{H}(Y^\top Y)$ . The risk of the latter estimator is studied in Section 3. In particular, we observe that  $\eta^0$  can achieve almost full recovery but is not rate optimal. The lack of rate optimality is mainly due to the fact that spectral methods do not benefit from the structure of binary vectors. As an improvement, we define, in Section 4, the iterative sequence of estimators  $(\hat{\eta}_k)_{k \geq 1}$  such that

$$\forall k \geq 0, \quad \hat{\eta}_{k+1} = \text{sign}(\mathbf{H}(Y^\top Y) \eta_k).$$

In comparison to Lu and Zhou (2016), we get better results, in particular for large  $p$ . In their approach, a spectral initialization on  $\boldsymbol{\theta}$  is considered and estimation of  $\boldsymbol{\theta}$  is handled at each iteration. The main difference compared to our procedure, lies in the fact that we get around estimation of  $\boldsymbol{\theta}$ . We only focus on the matrix  $\mathbf{H}(Y^\top Y)$ , that is almost blind to the direction of  $\boldsymbol{\theta}$ . In Giraud and Verzelen (2018), authors present an optimal but not sharp procedure. Our procedure is different in two ways. The first one being that it is not an SDP relaxation method and hence is faster. The second one, is that, through the operator  $\mathbf{H}$  we do not need to de-bias the Gram matrix, as the latter operator handles this task.

In section 5, we show the existence of a sharp phase transition for exact recovery in the Gaussian mixture model, around the threshold

$$\Delta_n^{*2} = \sigma^2 \log n \left( 1 + \sqrt{1 + 2 \frac{p}{n \log n}} \right).$$

In particular, this phase transition gives rise to two different regimes around a critical dimension  $p^* \asymp n \log(n)$ , showing that the hardness of exact recovery depends on whether  $p$  is larger or smaller than  $p^*$ .

## 2. Non-asymptotic fundamental limits in the Gaussian mixture model

This section is devoted to derive a sharp optimal lower bound for the risk  $\Psi_\Delta$ . As stated in the Introduction, a simple lower bound is given by (3). We give here a more precise statement.

**Proposition 1.** *For any  $\Delta > 0$ , we have*

$$\Psi_\Delta \geq \Phi^c(\Delta/\sigma).$$

Following similar arguments as in Ndaoud (2018), we decompose the minimax risk in two parts. Each part benefits either from information on  $\theta$  or  $\eta$ . According to Proposition 1, the hardness of recovering communities is mainly due to the lack of information concerning labels. It still benefits from the knowledge of  $\theta$ . In Giraud and Verzelen (2018), it becomes clear that for large  $p$ , the hardness of the problem results from the hardness of estimating  $\theta$ . Hence, and in order to capture this phenomenon, one may try to hide the information about the direction of  $\theta$  in order to make its estimation difficult.

More precisely, in order to lower bound the risk  $\Psi_\Delta$ , we place a prior on both  $\eta$  and  $\theta$ . Ideally, we would choose a Gaussian prior for  $\theta$  in order to make its estimation the hardest, but one should keep in mind that  $\theta$  is constrained by the set  $\Omega_\Delta$ . Based on similar arguments to derive lower bounds on constrained sets as in Butucea et al. (2018), we announce the next proposition. Let  $\pi = (\pi_\theta, \pi_\eta)$  be a product probability measure on  $\mathbf{R}^p \times \{-1, +1\}^n$  (a prior on  $(\theta, \eta)$ ). We denote by  $\mathbb{E}_\pi$  the expectation with respect to  $\pi$ .

**Proposition 2.** *Let  $\Delta > 0$  and  $\pi = (\pi_\theta, \pi_\eta)$  a product probability measure on  $\mathbf{R}^p \times \{-1, +1\}^n$ , then we have*

$$\Psi_\Delta \geq \frac{1}{2} \inf_{\hat{T} \in [-1, 1]^n} \frac{1}{n} \mathbb{E}_\pi \mathbf{E}_{(\theta, \eta)} |\hat{T} - \eta| - \pi_\theta(\|\theta\| < \Delta),$$

where  $\inf_{\hat{T} \in [-1, 1]^n}$  is the infimum over all estimators  $\hat{T}(Y) = (\hat{T}_1(Y), \dots, \hat{T}_n(Y))$  with values in  $[-1, 1]^n$ .

The previous proposition is useful to derive non-asymptotic lower bounds for constrained minimax risks. Still, for the corresponding lower bound to be optimal, we need the remainder term to be negligible. To do so, the prior on  $\theta$  must ensure that  $\|\theta\|$  is larger than  $\Delta$  with high probability. In the meantime this would make the problem of recovery easier. Hence it is clear that there exists some trade-off concerning the choice of  $\pi_\theta$ .

Let  $\pi^\alpha = (\pi_\theta^\alpha, \pi_\eta^\alpha)$  be a product prior on  $\mathbf{R}^p \times \{-1, +1\}^n$ , such that  $\pi_\theta^\alpha$  is random Gaussian vector with i.i.d entries of variance  $\alpha^2$ , and  $\pi_\eta^\alpha$  a random Rademacher vector with i.i.d entries of parameter  $1/2$ . For this specific choice of prior we get the following result.

**Theorem 3.** For any  $\alpha > 0$  and  $\pi^\alpha$  the product prior defined above, we have

$$\inf_{\hat{T} \in [-1,1]^n} \frac{1}{n} \mathbb{E}_{\pi^\alpha} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} |\hat{T} - \eta| \geq \frac{1}{n} \mathbb{E}_{\pi^\alpha} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} |\eta^{**} - \eta|,$$

where  $\eta^{**}$  is a supervised oracle estimator given by

$$\forall i = 1, \dots, n, \quad \eta_i^{**} = \text{sign} \left( Y_i^\top \left( \sum_{j \neq i} \eta_j Y_j \right) \right).$$

It is interesting to notice that the supervised oracle estimator  $\eta^{**}$  does not depend on  $\boldsymbol{\theta}$  explicitly but on its best estimator when other labels are known, under the Gaussian prior. The previous lower bound confirms the intuition that the supervised learning estimator is optimal in a minimax sense. For  $\sigma > 0$ , define  $G_\sigma$  such that

$$\forall t \in \mathbf{R}, \quad G_\sigma(t, \boldsymbol{\theta}) = \mathbf{P} \left( (\boldsymbol{\theta} + \sigma \xi_1)^\top \left( \boldsymbol{\theta} + \frac{\sigma}{n-1} \sum_{j=2}^n \xi_j \right) \leq \|\boldsymbol{\theta}\|^2 t \right), \quad (6)$$

where  $\xi_1, \dots, \xi_n$  are i.i.d standard Gaussian random vectors. Combining Proposition 2 and Theorem 3 and using the fact that all entries of the prior  $\pi^\alpha$  are i.i.d, we obtain the next Proposition.

**Proposition 4.** Let  $\Delta > 0$  and  $G_\sigma$  the function defined in (6). For any  $\alpha > 0$ , we have

$$\Psi_\Delta \geq \mathbb{E}_{\pi_\theta^\alpha} G_\sigma(0, \boldsymbol{\theta}) - \mathbf{P} \left( \sum_{j=1}^p \varepsilon_j^2 \leq \frac{\Delta^2}{\alpha^2} \right),$$

where  $\varepsilon_j$  are i.i.d standard Gaussian random variables.

We are now ready to state the main result of this section. As suggested in Giraud and Verzelen (2018), the main limit of the analysis in Lu and Zhou (2016) is partially due to the choice of the signal-to-noise ratio (SNR). We define here  $\mathbf{r}_n$ , a key quantity that plays the role of SNR in what follows. It is of the same order than the SNR presented in Giraud and Verzelen (2018), and is given by

$$\mathbf{r}_n = \frac{\Delta^2 / \sigma^2}{\sqrt{\Delta^2 / \sigma^2 + p/n}}. \quad (7)$$

**Theorem 5.** Let  $\Delta > 0$ . For  $n$  large enough, there exists a sequence  $\epsilon_n$  such that  $\epsilon_n = o(1)$  and

$$\Psi_\Delta \geq (1 - \epsilon_n) \Phi^c(\mathbf{r}_n(1 + \epsilon_n)).$$

It is worth saying that the result of Theorem 5 holds without any assumption on  $p$  and can be interpreted in a non-asymptotic sense by replacing  $\epsilon_n$  by some small  $c > 0$ . Moreover, and since  $\mathbf{r}_n \geq \Delta/\sigma$ , it is clear that it improves the lower bound in Proposition 1, at least for large values of  $n$ . This improvement is even more evident in the regime where  $\Delta^2/\sigma^2 = o(p/n)$ . This regime is mainly what we describe as the hard estimation regime.

### 3. On the spectral initialization

In the next two sections, we present a spectral estimator  $\eta^0$ , and analyze its non-asymptotic minimax risk. As it is the case, in SDP relaxations of the problem, the matrix of interest is the Gram matrix  $Y^\top Y$ . It is well known that the latter matrix suffers from a bias that grows with  $p$ . In Royer (2017) a de-biasing procedure is proposed using an estimator of the covariance on the noise. This step is important in order to make a given procedure adaptive to the noise level. Our approach is different but is still adaptive and consists in removing the diagonal entries of the Gram matrix. We give here some intuition about this procedure. Define the following linear operator  $\mathbf{H} : \mathbf{R}^{n \times n} \rightarrow \mathbf{R}^{n \times n}$ , such that

$$\forall M \in \mathbf{R}^{n \times n}, \quad \mathbf{H}(M) = M - \text{diag}(M),$$

where  $\text{diag}(M)$  is a diagonal matrix with the same diagonal as  $M$ . Going back to Theorem 3, we may observe that the oracle estimator  $\eta^{**}$  can be written as

$$\eta^{**} = \text{sign} \left( \mathbf{H} \left( Y^\top Y \right) \eta \right), \quad (8)$$

where the sign is applied entry-wise. This is the first time where the quantity  $\mathbf{H}(Y^\top Y)$  appears. Next, notice that the latter quantity can be decomposed as follows

$$\mathbf{H}(Y^\top Y) = \|\boldsymbol{\theta}\|^2 \eta \eta^\top + \mathbf{H}(W^\top W) + \mathbf{H}(W^\top \boldsymbol{\theta} \eta^\top + \eta \boldsymbol{\theta}^\top W) - \|\boldsymbol{\theta}\|^2 \mathbf{I}_n. \quad (9)$$

By isolating the signal term, the previous expression is similar to a spiked model, where the noise has a more complex structure. It turns out that the main driver of the noise is  $\mathbf{H}(W^\top W)$ . A technical lemma, that we state later, shows that our approach is probably an alternative to de-biasing the Gram matrix. More precisely, Lemma 18 gives that

$$\|\mathbf{H}(W^\top W)\|_{op} \leq 2\|W^\top W - \mathbf{E}(W^\top W)\|_{op}.$$

for any random matrix  $W$  with independent columns. Hence, the noise term can be controlled as if its covariance were known. Nevertheless, the previous operation may affect dramatically the signal since it also removes its diagonal entries. Luckily, the signal term is almost insensitive to this operation, since it is a rank-one matrix where the spike energy is spread all over the spike. For instance, we have

$$\|\mathbf{H}(\eta \eta^\top)\|_{op} = \left(1 - \frac{1}{n}\right) \|\eta \eta^\top\|_{op}.$$

Hence as  $n$  grows the signal does not get affected by removing the diagonal terms while we get rid of the bias in the noise. It is worth noticing that the previous trick succeeds thanks to the structure of  $\eta$  and can not be generalized to any spiked model.

Motivated by (9), the spectral estimator  $\eta^0$  is defined by

$$\eta^0 = \text{sign}(\hat{v}), \quad (10)$$

where  $\hat{v}$  is eigenvector corresponding to the top eigenvalue of  $\mathbf{H}(Y^\top Y)$ . The next result characterizes the non-asymptotic minimax risk of  $\eta^0$ .



**Theorem 6.** Let  $\Delta > 0$  and  $\eta^0$  the estimator given by (10). Under the condition  $\mathbf{r}_n \geq C$ , for some  $C > 0$ , we have

$$\sup_{(\boldsymbol{\theta}, \eta) \in \Omega_\Delta} \frac{1}{n} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} r(\eta^0, \eta) \leq \frac{C}{\mathbf{r}_n} + \frac{2}{n},$$

and

$$\sup_{(\boldsymbol{\theta}, \eta) \in \Omega_\Delta} \mathbf{P}_{(\boldsymbol{\theta}, \eta)} \left( \frac{1}{n} \left| \eta^\top \eta^0 \right| \leq 1 - \sqrt{\frac{\log n}{n}} - \frac{C}{\mathbf{r}_n} \right) \leq \epsilon_n \Phi^c(\mathbf{r}_n),$$

for some sequence  $\epsilon_n$  such that  $\epsilon_n = o(1)$ .

As we may expect the appropriate Hamming distance risk is decreasing with respect to  $\mathbf{r}_n$ . The residual term  $\frac{2}{n}$  is due to removing the diagonal and can be seen as the price to pay for adaptation. It is obvious that as  $\mathbf{r}_n$  gets larger than  $n$ , removing the diagonal terms is sub-optimal.

As  $\mathbf{r}_n \rightarrow \infty$ ,  $\eta^0$  achieves almost full recovery (cf. Definition 10). We show later that this condition is optimal but investigating its rate of convergence we find out that this method is sub-optimal in that sense. In particular,  $\eta^0$  can not achieve optimal exact recovery. To bring some evidence to the previous statement, we rely on asymptotic random matrix theory. In Benaych-Georges and Nadakuditi (2012), it is shown that, in the asymptotic where  $p/n \rightarrow c \in (0, 1]$  and when the noise is Gaussian, detection is possible only when  $\Delta^2 \geq \sqrt{c} \sigma^2$ . In that case, the correlation between  $\eta$  and its spectral approximation can not be larger than  $\sqrt{1 - \frac{c\sigma^2 + \Delta^2}{\Delta^2(1 + \Delta^2/\sigma^2)}}$ . When  $\mathbf{r}_n = \Omega(1)$ , we observe that  $\frac{c\sigma^2 + \Delta^2}{\Delta^2(1 + \Delta^2/\sigma^2)} \asymp \frac{1}{\mathbf{r}_n}$ . Hence, the decay in Theorem 6 is rate optimal, but certainly not sharp.

The condition  $\mathbf{r}_n = \Omega(1)$  is very natural, since it is necessary even for detection as shown in Banks et al. (2018). It is still obvious that spectral algorithms do not capture the particular structure of  $\eta$ . In the next section, we present an iterative procedure that is non-asymptotic minimax optimal and achieves optimal phase transitions.

#### 4. A rate optimal efficient algorithm

In this section, we present an algorithm that is minimax optimal, adaptive and faster than SDP relaxation. In the same spirit than Lu and Zhou (2016), we are tempted by using Lloyd's iterations. If properly initialized, Lloyd's algorithm may achieve the optimal rate under mild conditions after only a logarithmic number of steps. We present here a variant of Lloyd's iterations. Motivated by (8), and given an estimator  $\hat{\eta}^0$ , we define a sequence of estimators  $(\hat{\eta}^k)_{k \geq 0}$  such that

$$\forall k \geq 0, \quad \hat{\eta}^{k+1} = \text{sign} \left( \mathbf{H} \left( Y^\top Y \right) \hat{\eta}^k \right). \quad (11)$$

Notice that Lloyd's iterations correspond to the procedure (11), replacing  $\mathbf{H}(Y^\top Y)$  by  $Y^\top Y$ . If the initialization is good in some sense we describe below, then at each iteration  $\eta^k$  gets closer to  $\eta$  and achieves the minimax optimal rate after a logarithmic number of steps. The logarithmic number of steps is crucial computationally as many other iterative procedures.

**Theorem 7.** Let  $\Delta > 0$ ,  $\hat{\eta}^0$  an estimator satisfying

$$\frac{1}{n} \eta^\top \hat{\eta}^0 \geq 1 - \frac{C'}{\mathbf{r}_n}$$

for some  $C' > 0$  and  $(\hat{\eta}^k)_{k \geq 0}$  the corresponding iterative sequence (11). If  $\mathbf{r}_n \geq C$  for some  $C > 0$ , then after  $k = \lfloor 3 \log(n) \rfloor$  steps, we have

$$\sup_{(\boldsymbol{\theta}, \eta) \in \Omega_\Delta} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} \left( \frac{1}{n} |\hat{\eta}^k - \eta| \right) \leq (1 + \epsilon_n) \frac{2}{1 - \frac{C}{\sqrt{\mathbf{r}_n}}} \sup_{\|\boldsymbol{\theta}\| \geq \Delta} G_\sigma \left( \epsilon_n + \frac{C}{\sqrt{\mathbf{r}_n}}, \boldsymbol{\theta} \right) + \epsilon_n \Phi^c(\mathbf{r}_n),$$

for some sequence  $\epsilon_n$  such that  $\epsilon_n = o(1)$ .

As long as  $t$  is small, we recall that  $G(t, \boldsymbol{\theta})$  is close to  $G(0, \boldsymbol{\theta})$ . Theorem 7 can be interpreted as follows. Given a good initialization, the iterative procedure (11) achieves an error close to the supervised learning risk, within a logarithmic number of steps. Observing that, under the condition  $\mathbf{r}_n \geq C$  for some  $C > 0$ , the spectral estimator  $\eta^0$  is a good initializer, we state a general result showing that our variant of Lloyd's iterations initialized with a spectral estimator is minimax optimal.

**Theorem 8.** Let  $\Delta > 0$ . Let  $\eta^0$  be the spectral estimator defined in (10) and  $(\eta^k)_{k \geq 0}$  the corresponding iterative sequence (11). Assume that  $\mathbf{r}_n \geq C$  for some  $C > 0$ , then after  $k = \lfloor 3 \log(n) \rfloor$  steps we have

$$\sup_{(\boldsymbol{\theta}, \eta) \in \Omega_\Delta} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} r(\eta^k, \eta) \leq 2(1 + \epsilon_n) \frac{1}{1 - \frac{C}{\sqrt{\mathbf{r}_n}}} \Phi^c \left( \mathbf{r}_n \left( 1 - \epsilon_n - \frac{C}{\sqrt{\mathbf{r}_n}} \right) \right),$$

for some sequence  $\epsilon_n$  such that  $\epsilon_n = o(1)$ .

Notice that the upper bound in Theorem 8 is almost optimal, and gets closer to the optimal minimax rate as  $n, \mathbf{r}_n \rightarrow \infty$ . Hence, under mild conditions, we get a matching upper bound to the lower bound in Theorem 5. More interestingly, we figure out that a good initialization combined with smart iterations is almost equivalent to the supervised learning oracle. In fact the rate in Theorem 8 corresponds also to the rate of the supervised oracle estimator  $\eta^{**}$ . We conclude that unsupervised learning meets supervised learning in the Gaussian mixture model. The next Proposition, gives a full picture of the minimax risk  $\Psi_\Delta$ .

**Proposition 9.** Let  $\Delta > 0$ . For some  $c_1, c_2, C_1, C_2 > 0$  and  $n$  large enough, we have

$$C_1 e^{-c_1 \mathbf{r}_n^2} \leq \Psi_\Delta \leq C_2 e^{-c_2 \mathbf{r}_n^2}.$$

As a consequence, we give a complete answer to minimax optimality for the problem of recovery in model (1). Notice that the procedure we present here has a different rate of decay compared to the spectral procedure (10), that turns out to be non-asymptotically sub-optimal. This answers some questions addressed in the Introduction. We turn now to a more precise characterization of phase transitions.

## 5. Asymptotic analysis. Phase transitions

In this section, we conduct the asymptotic analysis of the problem of community recovery in the Gaussian mixture model. The results are derived as corollaries of the minimax bounds of previous sections. We will assume that  $n \rightarrow \infty$  and that parameters  $p, \sigma$  and  $\Delta$  depend on  $n$ . For the sake of readability we omit to write the index  $n$ .

The two asymptotic properties we study here are *exact recovery* and *almost full recovery*. We use this terminology following Butucea et al. (2018) but we define these properties in a different way.

**Definition 10.** Let  $(\Omega_{\Delta_n})_{n \geq 2}$  be a sequence of classes corresponding to  $(\Delta_n)_{n \geq 2}$ :

- We say that exact recovery is possible for  $(\Omega_{\Delta_n})_{n \geq 2}$  if there exists an estimator  $\hat{\eta}$  such that

$$\lim_{n \rightarrow \infty} \sup_{(\boldsymbol{\theta}, \eta) \in \Omega_{\Delta_n}} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} r(\hat{\eta}, \eta) = 0. \quad (12)$$

In this case, we say that  $\hat{\eta}$  achieves exact recovery.

- We say that almost full recovery is possible for  $(\Omega_{\Delta_n})_{n \geq 2}$  if there exists an estimator  $\hat{\eta}$  such that

$$\lim_{n \rightarrow \infty} \sup_{(\boldsymbol{\theta}, \eta) \in \Omega_{\Delta_n}} \frac{1}{n} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} r(\hat{\eta}, \eta) = 0. \quad (13)$$

In this case, we say that  $\hat{\eta}$  achieves almost full recovery.

It is of interest to characterize the sequence  $(\Delta_n)_{n \geq 1}$ , for which exact recovery and almost full recovery are possible. To describe the impossibility of exact or almost full recovery, we need the following definition.

**Definition 11.** Let  $(\Omega_{\Delta_n})_{n \geq 2}$  be a sequence of classes corresponding to  $(\Delta_n)_{n \geq 2}$ :

- We say that exact recovery is impossible for  $(\Omega_{\Delta_n})_{n \geq 2}$  if

$$\liminf_{n \rightarrow \infty} \inf_{\tilde{\eta}} \sup_{(\boldsymbol{\theta}, \eta) \in \Omega_{\Delta_n}} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} r(\tilde{\eta}, \eta) > 0, \quad (14)$$

- We say that almost full recovery is impossible for  $(\Omega_{\Delta_n})_{n \geq 2}$  if

$$\liminf_{n \rightarrow \infty} \inf_{\tilde{\eta}} \sup_{(\boldsymbol{\theta}, \eta) \in \Omega_{\Delta_n}} \frac{1}{n} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} r(\tilde{\eta}, \eta) > 0, \quad (15)$$

where  $\inf_{\tilde{\eta}}$  denotes the infimum over all estimators in  $\{-1, +1\}^n$ .

The following general characterization theorem is a straightforward corollary of the results of previous sections.

**Theorem 12.** (i) *Almost full recovery is possible for  $(\Omega_{\Delta_n})_{n \geq 2}$  if and only if*

$$\Phi^c(\mathbf{r}_n) \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (16)$$

*In this case, the estimator  $\eta^k$  defined in (10)-(11), with  $k = \lfloor 3 \log n \rfloor$ , achieves almost full recovery.*

(ii) *Exact recovery is impossible for  $(\Omega_{\Delta_n})_{n \geq 2}$  if*

$$\forall \epsilon > 0, \quad \liminf_{n \rightarrow \infty} n \Phi^c(\mathbf{r}_n(1 + \epsilon)) > 0 \quad \text{as } n \rightarrow \infty, \quad (17)$$

*and possible if for some  $\epsilon > 0$*

$$n \Phi^c(\mathbf{r}_n(1 - \epsilon)) \rightarrow 0 \quad \text{as } n \rightarrow \infty, \quad (18)$$

*In this case, the estimator  $\eta^k$  defined in (10)-(11), with  $k = \lfloor \log 3 \rfloor$ , achieves exact recovery.*

Although this theorem gives a complete solution to the problem, conditions (16), (17) and (18) are not quite explicit. Intuitively, we would like to get a “phase transition” values  $\Delta_n^*$  such that exact (or almost full) recovery is possible for  $\Delta_n$  greater than  $\Delta_n^*$  and is impossible for  $\Delta_n$  smaller than  $\Delta_n^*$ . Our aim now is to find such “phase transition” values. We first do it in the almost full recovery framework.

The next theorem is a consequence of Theorem 12. It describes a “phase transition” for  $\Delta_n$  in the problem of almost full recovery.

**Theorem 13.** (i) *If, for all  $n$  large enough,*

$$\Delta_n^2 \geq \sigma^2 A_n \frac{1 + \sqrt{1 + 4 \frac{p}{n A_n}}}{2}$$

*for an arbitrary sequence  $A_n \rightarrow \infty$ , as  $n \rightarrow \infty$ , then the estimator  $\eta^k$  defined in (10)-(11), with  $k = \lfloor \log 3 \rfloor$ , achieves almost full recovery.*

(ii) *Moreover, if there exists  $A > 0$  such that for all  $n$  large enough the reverse inequality holds:*

$$\Delta_n^2 \leq \sigma^2 A \frac{1 + \sqrt{1 + 4 \frac{p}{n A}}}{2} \quad (19)$$

*then almost full recovery is impossible.*

Theorem 13 shows that almost full recovery occurs if and only if

$$\sigma^2 \left(1 + \sqrt{p/n}\right) = o(\Delta_n^2). \quad (20)$$

We now turn to the problem of exact recovery. The “phase transition” for  $\Delta_n$  in the problem of exact recovery is described in the next theorem.

**Theorem 14.** (i) If

$$\Delta_n^2 \geq \sigma^2 A_n \frac{1 + \sqrt{1 + 4\frac{p}{nA_n}}}{2}$$

for all  $n$  large enough, where the sequence  $A_n$  is such that

$$A_n \geq 2 \log n (1 + \epsilon), \quad (21)$$

for some  $\epsilon > 0$ . Then the estimator  $\eta^k$  defined in (10)-(11), with  $k = \lfloor \log 3 \rfloor$ , achieves exact recovery.

(ii) If the complementary condition holds,

$$\Delta_n^2 \leq \sigma^2 A_n \frac{1 + \sqrt{1 + 4\frac{p}{nA_n}}}{2}$$

for all  $n$  large enough, where the sequence  $A_n$  is such that

$$A_n \leq 2 \log n (1 - \epsilon), \quad (22)$$

for any  $\epsilon > 0$ , then exact recovery is impossible.

Some remarks are in order here. First of all, Theorem 14 shows that the “phase transition” for exact recovery occurs at

$$\Delta_n^{*2} = \sigma^2 \log n \left( 1 + \sqrt{1 + 2\frac{p}{n \log n}} \right).$$

It is worth noticing that the previous sharp threshold for exact recovery holds for all values of  $p$ . In particular, there exists a critical dimension  $p^* \asymp n \log n$ . As long as  $p = o(p^*)$ , then  $\Delta_n^* = (1 + o(1))\sigma\sqrt{2 \log n}$ . In this case, the phase transition threshold for exact recovery, is the same as if  $\theta$  were known. While if  $p^* = o(p)$ , then  $\Delta_n^* = (1 + o(1))\sigma \left( 2\frac{p \log n}{n} \right)^{1/4}$ . This new condition includes the hardness of estimation, and  $p^*$  can be interpreted as a phase transition with respect to the dimension.

## 6. Discussion and open problems

The main objective of this paper, was to establish a phase transition for exact recovery in the Gaussian mixture model. While, all upper bounds remain valid in the case of sub-Gaussian noise, it would be interesting to generalize the methodology used to derive both lower and upper bounds in the case of multiple communities and general covariance structure of the noise. We also expect the procedure (10)-(11) to achieve optimal exact recovery in other problems, for instance in the Bipartite Stochastic Block Model.

We conclude this paper with an open question. Going back to the regime  $p^* = o(p)$ , we proved that for any  $\epsilon > 0$ , the condition

$$\Delta^2 \geq \sqrt{2}(1 - \epsilon)\sigma^2 \left( \frac{p}{p^*} \right)^{1/2}$$

is necessary to achieve exact recovery. This is a consequence of having a Gaussian prior on  $\theta$  which makes recovering its direction the hardest. We give here a heuristic arguing that this should hold independently on the choice of the prior as long as  $\theta$  is uniformly well-spread (i.e not sparse). Suppose that we put a Rademacher prior on  $\theta$  such that  $\theta = \frac{\Delta}{\sqrt{p}}\zeta$ , where  $\zeta$  is a random vector with i.i.d Rademacher entries of parameter 1/2. Following the same argument as in Proposition 1, it is clear that a necessary condition to get non-trivial correlation with  $\zeta$  is given by

$$\Delta^2 \geq c\sigma^2 \frac{p}{n},$$

for some  $c > 0$ . Observing that, in the hard estimation regime, we have

$$\left(\frac{p}{p^*}\right)^{1/2} = o\left(\frac{p}{n}\right),$$

it comes that, while exact recovery of  $\eta$  is possible, non-trivial correlation with  $\zeta$  is impossible. Hence, there is no hope from achieving exact recovery through partial recovery of  $\theta$  in the hard estimation regime.

**Conjecture 15.** *Let  $\Delta > 0$ . Assume that  $Y$  follows model (1).  $\eta$  is a random vector such that all entries are i.i.d Rademacher random variables of parameter 1/2, and  $\theta = \frac{\Delta}{\sqrt{p}}\zeta$  where  $\zeta$  is a random vector with i.i.d Rademacher entries of parameter 1/2. Assume that  $n \log n = o(p)$ . Prove or disprove that, for any  $\epsilon > 0$ ,*

$$\Delta^2 \geq (1 - \epsilon)\sigma^2 \sqrt{2 \frac{p \log n}{n}}$$

*is necessary to achieve exact recovery.*

In particular, a positive answer to the previous question will be very useful to derive optimal conditions for exact recovery in bipartite graph models among other problems.

## Acknowledgments

We would like to thank Christophe Giraud for stimulating discussions on clustering in Gaussian mixtures.

## References

- Emmanuel Abbe. Community detection and stochastic block models: recent developments. *arXiv preprint arXiv:1703.10146*, 2017.
- Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *arXiv preprint arXiv:1405.3267*, 2014.
- Jess Banks, Cristopher Moore, Roman Vershynin, Nicolas Verzelen, and Jiaming Xu. Information-theoretic bounds and phase transitions in clustering, sparse pca, and sub-matrix localization. *IEEE Transactions on Information Theory*, 2018.
- Florent Benaych-Georges and Raj Rao Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, 2012.
- Cristina Butucea, Mohamed Ndaoud, Natalia A Stepanova, and Alexandre B Tsybakov. Variable selection with hamming loss. *The Annals of Statistics*, 46(5):1837–1875, 2018.
- Vitaly Feldman, Will Perkins, and Santosh Vempala. Subsampled power iteration: a unified algorithm for block models and planted csp’s. In *Advances in Neural Information Processing Systems*, pages 2836–2844, 2015.
- Laura Florescu and Will Perkins. Spectral thresholds in the bipartite stochastic block model. In *Conference on Learning Theory*, pages 943–959, 2016.
- Christophe Giraud and Nicolas Verzelen. Partial recovery bounds for clustering with the relaxed  $k$  means. *arXiv preprint arXiv:1807.07547*, 2018.
- Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.
- Jason M Klusowski and WD Brinda. Statistical guarantees for estimating the centers of a two-component gaussian mixture by em. *arXiv preprint arXiv:1608.02280*, 2016.
- Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.
- Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.
- Yu Lu and Harrison H Zhou. Statistical and computational guarantees of lloyd’s algorithm and its variants. *arXiv preprint arXiv:1612.02099*, 2016.
- Dustin G Mixon, Soledad Villar, and Rachel Ward. Clustering subgaussian mixtures by semidefinite programming. *arXiv preprint arXiv:1602.06612*, 2016.
- Mohamed Ndaoud. Interplay of minimax estimation and minimax support recovery under sparsity. *arXiv preprint arXiv:1810.05478*, 2018.

Martin Royer. Adaptive clustering through semidefinite programming. In *Advances in Neural Information Processing Systems*, pages 1795–1803, 2017.

Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.



## Appendix

### 6.1 Main proofs

**Remark 16.** In all the proofs of lower bounds, we assume that half of  $(\eta_i)_{i \geq 1}$  are known. Without loss of generality  $\eta_1, \dots, \eta_{\lfloor n/2 \rfloor}$  are known. Observe that for any  $\tilde{\eta}$ , if  $\tilde{\eta}_i = \eta_i$  for  $i = 1, \dots, \lfloor n/2 \rfloor$ , then  $r(\tilde{\eta}, \eta) = |\tilde{\eta} - \eta|$ . Based on this observation we may assume that  $r(\tilde{\eta}, \eta) = |\tilde{\eta} - \eta|$  as long as we devide the minimax risk by a factor 2.

*Proof of Proposition 1.* Placing an independent Rademacher prior  $\pi$  on  $\eta$ , and fixing  $\boldsymbol{\theta}$ , it turns out that

$$\inf_{\tilde{\eta}} \sup_{(\boldsymbol{\theta}, \eta) \in \Omega_\Delta} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} |\hat{\eta} - \eta| \geq \sum_{j=1}^n \inf_{\bar{\eta}_j} \mathbf{E}_\pi |\bar{\eta}_j(Y_j) - \eta_j|, \quad (23)$$

where  $\bar{\eta}_j \in [-1, 1]$ . The last inequality holds because of independence of the priors. We define, for  $\epsilon \in \{-1, 1\}$ ,  $\tilde{f}_{\eta_i=\epsilon}^j$  the likelihood of the observation  $Y_j$  conditionally on  $\eta_i$ . Now Using Neyman-Pearson lemma, we get that

$$\eta_j^* = \begin{cases} 1 & \text{if } \tilde{f}_{\eta_i=1}(Y_j) \geq \tilde{f}_{\eta_i=-1}(Y_j), \\ -1 & \text{else.} \end{cases}$$

Hence and since the noise is Gaussian we get that

$$\eta_j^* = \text{sign} \left( \boldsymbol{\theta}^\top Y_j \right).$$

Replacing in (23), we get

$$\inf_{\bar{\eta}_j} \mathbf{E}_\pi |\bar{\eta}_j(Y_j) - \eta_j| = 2\Phi^c(\Delta/\sigma).$$

□

*Proof of Proposition 2.* Throughout the proof, we write for brevity  $A = \Omega_\Delta$ . Set  $\eta^A = \eta \mathbf{1}(\boldsymbol{\theta} \in A)$  and denote by  $\pi_A$  the probability measure  $\pi$  conditioned by the event  $\{\boldsymbol{\theta} \in A\}$ , that is, for any  $C \subseteq \{0, 1\}^d$ ,

$$\pi_A(C) = \frac{\pi(C \cap \{\boldsymbol{\theta} \in A\})}{\pi(\boldsymbol{\theta} \in A)}.$$

The measure  $\pi_A$  is supported on  $A$  and we have

$$\begin{aligned} \inf_{\hat{\eta}} \sup_{(\boldsymbol{\theta}, \eta) \in A} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} |\hat{\eta} - \eta| &\geq \inf_{\hat{\eta}} \mathbb{E}_{\pi_A} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} |\hat{\eta} - \eta| = \inf_{\hat{\eta}} \mathbb{E}_{\pi_A} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} |\hat{\eta} - \eta^A| \\ &\geq \sum_{j=1}^n \inf_{\hat{T}_j} \mathbb{E}_{\pi_A} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} |\hat{T}_j - \eta_j^A| \end{aligned}$$

where  $\inf_{\hat{T}_j}$  is the infimum over all estimators  $\hat{T}_j = \hat{T}_j(Y)$  with values in  $\mathbf{R}$ . According to Theorem 1.1 and Corollary 1.2 on page 228 in Lehmann and Casella (2006), there exists a Bayes estimator  $B_j^A = B_j^A(Y)$  such that

$$\inf_{\hat{T}_j} \mathbb{E}_{\pi_A} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} |\hat{T}_j - \eta_j^A| = \mathbb{E}_{\pi_A} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} |B_j^A - \eta_j^A|,$$

and this estimator is a conditional median of  $\eta_j^A$  given  $Y$ . Therefore,

$$\inf_{\hat{\eta}} \sup_{(\boldsymbol{\theta}, \eta) \in A} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} |\hat{\eta} - \eta| \geq \mathbb{E}_{\pi_A} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} \sum_{j=1}^n |B_j^A - \eta_j^A|. \quad (24)$$

Note that  $B_j^A \in [-1, 1]$  since  $\eta_j^A$  takes its values in  $[-1, 1]$ . Using this, we obtain

$$\begin{aligned} \inf_{\hat{T} \in [0, 1]^n} \mathbb{E}_{\pi} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} |\hat{T} - \eta| &\leq \mathbb{E}_{\pi} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} \sum_{j=1}^n |B_j^A - \eta_j| \\ &= \mathbb{E}_{\pi} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} \left( \sum_{j=1}^n |B_j^A - \eta_j| \mathbf{1}(\boldsymbol{\theta} \in A) \right) + \mathbb{E}_{\pi} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} \left( \sum_{j=1}^n |B_j^A - \eta_j| \mathbf{1}((\boldsymbol{\theta}, \eta) \in A^c) \right) \\ &\leq \mathbb{E}_{\pi_A} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} \sum_{j=1}^n |B_j^A - \eta_j^A| + \mathbb{E}_{\pi} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} \left( \sum_{j=1}^n |B_j^A - \eta_j| \mathbf{1}((\boldsymbol{\theta}, \eta) \in A^c) \right) \\ &\leq \mathbb{E}_{\pi_A} \mathbf{E}_{(\boldsymbol{\theta}, \eta)} \sum_{j=1}^n |B_j^A - \eta_j^A| + 2n\mathbf{P}((\boldsymbol{\theta}, \eta) \notin A). \end{aligned} \quad (25)$$

□

*Proof of Theorem 3.* Without loss of generality, we assume that  $\sigma = 1$ , replacing  $\Delta$  by  $\Delta/\sigma$ . We start by using the fact that

$$\mathbf{E}_{\pi} |\hat{\eta} - \eta| = \sum_{i=1}^n \mathbf{E}_{\pi_{-i}} \mathbf{E}_{\pi_i} (|\hat{\eta}_i - \eta_i| | (\eta_j)_{j \neq i}),$$

where  $\pi_i$  is the prior on  $\boldsymbol{\theta}$  and  $\eta_i$  while  $\pi_{-i}$  is the prior on  $(\eta_j)_{j \neq i}$ . We define, for  $\epsilon \in \{-1, 1\}$ ,  $\tilde{f}_{\eta_i = \epsilon}$  the likelihood of the observation  $Y$  knowing  $(\eta_j)_{j \neq i}$ . It is well known that the optimal decoder is given by

$$\eta_i^{**} = \begin{cases} 1 & \text{if } \tilde{f}_{\eta_i=1}(Y) \geq \tilde{f}_{\eta_i=-1}(Y), \\ -1 & \text{else.} \end{cases}$$

Using the independence of the rows of  $Y$  we have

$$\tilde{f}_{\eta_i = \epsilon}(Y) = \prod_{j=1}^p \frac{e^{-\frac{1}{2} L_j^{\top} \Sigma_{\epsilon}^{-1} L_j}}{|\Sigma_{\epsilon}|},$$

where  $L_j$  is the  $j$ -th row of  $Y$  and  $\Sigma_{\epsilon} = \mathbf{I}_n + \alpha^2 \eta_{\epsilon} \eta_{\epsilon}^{\top}$ .  $\eta_{\epsilon}$  is a binary vector such that  $\eta_i = \epsilon$  and the other components are known. It is easy to check that  $|\Sigma_{\epsilon}| = 1 + \alpha^2 n$ , hence it does not depend on  $\epsilon$ . A simple calculation, can also lead us to

$$\Sigma_{\epsilon}^{-1} = \mathbf{I}_n - \frac{\alpha^2}{1 + \alpha^2 n} \eta_{\epsilon} \eta_{\epsilon}^{\top}.$$

Hence

$$\begin{aligned}
\frac{\tilde{f}_{\eta_i=1}(Y)}{\tilde{f}_{\eta_i=-1}(Y)} &= \prod_{j=1}^p e^{-\frac{1}{2} L_j^\top (\Sigma_1^{-1} - \Sigma_{-1}^{-1}) L_j} \\
&= \prod_{j=1}^p e^{\frac{\alpha^2}{1+\alpha^2 n} L_{ji} \sum_{k \neq i} L_{jk} \eta_k} \\
&= e^{\frac{\alpha^2}{1+\alpha^2 n} \sum_{k \neq i} \eta_k \sum_{j=1}^p L_{jk} L_{ji}} = e^{\frac{\alpha^2}{1+\alpha^2 n} \langle Y_i, \sum_{k \neq i} \eta_k Y_k \rangle}.
\end{aligned}$$

It is now immediate that

$$\eta_i^{**} = \text{sign} \left( Y_i^\top \left( \sum_{k \neq i} \eta_k Y_k \right) \right).$$

□

*Proof of Proposition 4.* straightforward by combining Proposition 2 and Theorem 3. □

*Proof of Theorem 5.* Again, w.l.g assume that  $\sigma = 1$ . We prove the result following different cases.

- case  $\Delta \leq \frac{\log(n)^2}{\sqrt{n}}$ :

In this case we use Proposition 1.

Since  $0 \leq \frac{\Delta^2}{\sqrt{\Delta^2 + p/n}} \leq \Delta$ , we have  $|\Delta - \frac{\Delta^2}{\sqrt{\Delta^2 + p/n}}| \leq \frac{\log n^2}{\sqrt{n}}$ . Hence

$$\left| \Phi^c(\Delta) - \Phi^c \left( \frac{\Delta^2}{\sqrt{\Delta^2 + p/n}} \right) \right| \leq c \frac{\log n^2}{\sqrt{n}} \Phi^c \left( \frac{\Delta^2}{\sqrt{\Delta^2 + p/n}} \right),$$

for some  $c > 0$ . Hence we get the result with  $\epsilon_n = c \frac{\log n^2}{\sqrt{n}}$ .

- case  $\Delta \geq \sqrt{\frac{p \log(n)}{n}}$ :

In this case, we have  $\sqrt{1 + \frac{p}{n\Delta^2}} \frac{\Delta^2}{\sqrt{\Delta^2 + p/n}} = \Delta$ . It is easy to check that

$$\left| \sqrt{1 + \frac{p}{n\Delta^2}} - 1 \right| \leq \frac{1}{\log n}.$$

Hence

$$\Delta \leq \frac{\Delta^2}{\sqrt{\Delta^2 + p/n}} (1 + \epsilon_n),$$

for  $\epsilon_n = \frac{1}{\log(n)}$ . We conclude using Proposition 1.

- case  $\frac{\log n^2}{\sqrt{n}} < \Delta < \sqrt{\frac{p \log(n)}{n}}$ :

In what follows we use multiples sequences depending on  $n$  that we define at the end of the proof. Notice that  $p \geq \log n^3$  in this regime. We will use Proposition 4. Set  $\alpha^2$  such that

$$\alpha^2 = \frac{\Delta^2}{p(1 - \alpha_n)} \quad \text{and} \quad \nu_n = \sqrt{\frac{n\Delta^2}{p \log n^2}}.$$

It is easy to check that  $0 < \nu_n^2 \leq 1/\log n$ , Hence

$$\mathbf{P} \left( \sum_{j=1}^p \varepsilon_j^2 \leq \frac{\Delta^2}{\alpha^2} \right) = \mathbf{P} \left( \frac{1}{p} \sum_{j=1}^p (\varepsilon_j^2 - 1) \leq -\alpha \right) \leq e^{-c \frac{n}{\log n^2} \Delta^2},$$

for some  $c > 0$ . Hence, for any  $\epsilon_n \rightarrow 0$  we have

$$\mathbf{P} \left( \sum_{j=1}^p \varepsilon_j^2 \leq \frac{\Delta^2}{\alpha^2} \right) \leq e^{-c' \log n} \Phi^c \left( \Delta(1 + \epsilon_n) \right) \leq e^{-c' \log n} \Phi^c \left( \frac{\Delta^2}{\sqrt{\Delta^2 + p/n}} (1 + \epsilon_n) \right),$$

for some  $c' > 0$ . Since  $e^{-c' \log n} \rightarrow 0$ , then in order to conclude, we just need to prove that

$$\mathbb{E}_{\pi_{\theta}^{\alpha}} G_{\sigma}(0, \theta) \geq (1 - \epsilon_n) \Phi^c \left( \frac{\Delta^2}{\sqrt{\Delta^2 + p/n}} (1 + \epsilon_n) \right),$$

for some sequence  $\epsilon_n \rightarrow 0$ .

We recall that

$$\mathbb{E}_{\pi_{\theta}^{\alpha}} G_{\sigma}(0, \theta) = \mathbf{P} \left( \langle \theta + \xi_1, \theta + \frac{\xi_2}{\sqrt{n-1}} \rangle \leq 0 \right),$$

where  $\xi_1, \xi_2$  are two independent Gaussian random vector with i.i.d standard entries and  $\theta$  and independent Gaussian prior. Moreover, using independence, we have

$$\mathbf{P} \left( \langle \theta + \xi_1, \theta + \frac{\xi_2}{\sqrt{n-1}} \rangle \leq 0 \right) = \mathbf{P} \left( \varepsilon \sqrt{\|\theta\|^2 + \frac{\|\xi_2\|_2^2}{n-1} + \frac{2}{\sqrt{n-1}} \theta^{\top} \xi_2} \geq \|\theta\|^2 + \frac{1}{\sqrt{n-1}} \theta^{\top} \xi_2 \right),$$

where  $\varepsilon$  is a standard Gaussian random variable. Fix  $\theta$  and set the random event

$$\mathcal{A} = \left\{ \frac{\|\xi_2\|_2^2}{n-1} \geq \frac{p}{n-1} (1 - \zeta_n) \right\} \cap \left\{ |\theta^{\top} \xi_2| \leq \sqrt{n-1} \beta_n \|\theta\|^2 \right\}.$$

It is easy to check that

$$\mathbf{P}(\mathcal{A}^c) \leq e^{-c \log n^3 \zeta_n^2} + e^{-c \beta_n^2 n \|\theta\|^2}.$$

Hence conditioning on  $\theta$ , we have

$$\mathbf{P} \left( \langle \theta + \xi_1, \theta + \frac{\xi_2}{\sqrt{n-1}} \rangle \leq 0 \right) \geq \mathbf{E} \left[ \Phi^c \left( \frac{\|\theta\|^2 (1 + \beta_n)}{\sqrt{\|\theta\|^2 (1 - 2\beta_n) + \frac{p}{n-1} (1 - \zeta_n)}} \right) \mathbf{P}(\mathcal{A}) \right].$$

where the last expectation is over  $\boldsymbol{\theta}$ . Set now the random event  $\mathcal{B} = \{|\|\boldsymbol{\theta}\|^2 - \Delta^2| \leq \Delta^2 \gamma_n\}$ . Then

$$\mathbf{P}\left(\left\langle \boldsymbol{\theta} + \xi_1, \boldsymbol{\theta} + \frac{\xi_2}{\sqrt{n-1}} \right\rangle \leq 0\right) \geq \Phi^c(Z_n) \left(1 - e^{-c \log n^3 \zeta_n^2} - e^{-c \beta_n^2 (1-\gamma_n) \log^4 n}\right) \mathbf{P}(\mathcal{B}), \quad (26)$$

where  $Z_n := \frac{\Delta^2(1+\beta_n)(1+\gamma_n)}{\sqrt{\Delta^2(1-2\beta_n)(1-\gamma_n) + \frac{p}{n-1}(1-\zeta_n)}}$ . Now we may check that

$$\mathbf{P}(\mathcal{B}^c) = \mathbf{P}\left(\left|\sum_{j=1}^p \varepsilon_j^2 - \frac{\Delta^2}{\alpha^2}\right| \geq \frac{\Delta^2}{\alpha^2} \gamma_n\right).$$

Hence

$$\mathbf{P}(\mathcal{B}^c) \leq \mathbf{P}\left(\left|\sum_{j=1}^p \varepsilon_j^2 - p\right| \geq \frac{\Delta^2}{\alpha^2} \gamma_n - \left|p - \frac{\Delta^2}{\alpha^2}\right|\right).$$

Using the definition of  $\alpha^2$  we get

$$\mathbf{P}(\mathcal{B}^c) \leq \mathbf{P}\left(\left|\sum_{j=1}^p \varepsilon_j^2 - p\right| \geq p((1-\nu_n)\gamma_n - \nu_n)\right) \leq 2e^{-c \log n^3 \gamma_n^2}, \quad (27)$$

for some  $c > 0$  given that  $4\nu_n \leq \gamma_n \leq 1$ . Since  $\nu_n^2 \leq 1/\log n$ , by choosing  $\beta_n^2 = 1/\log n$ ,  $\gamma_n^2 = 16/\log n$  and  $\zeta_n^2 = 1/\log n$  the proof is complete combining (26) and (27).  $\square$

*Proof of Theorem 6.* We begin by writing that

$$Y^\top Y = \|\boldsymbol{\theta}\|^2 \frac{1}{n} \boldsymbol{\eta} \boldsymbol{\eta}^\top + Z_1,$$

where

$$Z_1 = \frac{1}{n} \boldsymbol{\eta} \boldsymbol{\theta}^\top W + \frac{1}{n} W^\top \boldsymbol{\theta} \boldsymbol{\eta}^\top + \frac{1}{n} W^\top W.$$

Next observe that

$$H(Y^\top Y) = \|\boldsymbol{\theta}\|^2 \frac{1}{n} \boldsymbol{\eta} \boldsymbol{\eta}^\top + Z_2,$$

where  $Z_2$  is given by

$$Z_2 = H(Z_1) - \|\boldsymbol{\theta}\|^2 \frac{1}{n} \mathbf{I}_n.$$

Based on the fact Lemma 18, we have

$$\|Z_2\|_\infty \leq 4 \left\| \frac{1}{n} \boldsymbol{\eta} \boldsymbol{\theta}^\top W \right\|_\infty + 2 \left\| \frac{1}{n} W W^\top - \mathbf{E}\left(\frac{1}{n} W W^\top\right) \right\|_\infty + \frac{\|\boldsymbol{\theta}\|^2}{n}. \quad (28)$$

Based on the Davis-Kahan  $\sin \theta$  theorem, we have that

$$\min_{\nu \in \{-1, 1\}} \left\| \hat{v} - \frac{1}{\sqrt{n}} \nu \boldsymbol{\eta} \right\|_2^2 \leq 4 \frac{\|Z_2\|_\infty}{\|\boldsymbol{\theta}\|^2}.$$

Moreover using Lemma 21, we get

$$r(\eta^0, \eta) \leq 8 \frac{\|Z_2\|_\infty}{\|\boldsymbol{\theta}\|^2}. \quad (29)$$

We can verify that  $\|\boldsymbol{\theta}\|^2 \geq c(1 + \sqrt{p/n})$ , and  $\sqrt{\|\boldsymbol{\theta}\|^2 + \frac{p}{n}} \geq C'(1 + \sqrt{p/n})$ . We conclude using Lemma 19 and Lemma 20 as long as (28), (29).  $\square$

*Proof of Theorem 7.* Define the random events  $\mathbf{B}$  and  $\mathbf{A}_i$  for  $i = 1, \dots, n$  such that

$$\mathbf{B} = \left\{ \frac{1}{n} \|\mathbf{H}(Z_2)\|_\infty \leq \frac{\|\boldsymbol{\theta}\|^2}{2} \left( \frac{C}{\sqrt{r}} + \frac{1}{n} \right) \right\},$$

and for all  $i = 1, \dots, n$

$$\mathbf{A}_i = \left\{ \frac{1}{n} \left\langle \mathbf{H}(Y^\top Y)_i, \eta \right\rangle \text{sign}(\eta_i) \geq C \frac{\|\boldsymbol{\theta}\|^2}{\sqrt{r}} \right\}.$$

where we have used same notation of the previous proof. Remember that

$$\mathbf{H}(Y^\top Y) = \|\boldsymbol{\theta}\|^2 \eta \eta^\top + \mathbf{H}(Z_2).$$

A simple calculation leads to

$$\left\langle \mathbf{H}(Y^\top Y)_i, \eta^k \right\rangle = \left\langle \mathbf{H}(Z_2)_i, \eta^k - \eta \right\rangle + \left\langle \mathbf{H}(Y^\top Y)_i, \eta \right\rangle - \|\boldsymbol{\theta}\|^2 \eta_i \left( n - \left\langle \eta^k, \eta \right\rangle \right).$$

Hence if  $\eta_i = -1$  and if  $\mathbf{A}_i$  is true then

$$\left\langle \frac{1}{n} \mathbf{H}(Y^\top Y)_i, \eta^k \right\rangle \leq \left\langle \frac{1}{n} \mathbf{H}(Z_2)_i, \eta^k - \eta \right\rangle - \|\boldsymbol{\theta}\|^2 \frac{C}{\sqrt{r}}.$$

Hence when  $\eta_i = -1$  we have

$$\mathbf{1}_{\left\{ \left\langle \frac{1}{n} \mathbf{H}(Y^\top Y)_i, \eta^k \right\rangle \geq 0 \right\}} \mathbf{1}_{\mathbf{A}_i} \leq \mathbf{1}_{\left\{ \left\langle \frac{1}{n} \mathbf{H}(Z_2)_i, \eta^k - \eta \right\rangle \geq \|\boldsymbol{\theta}\|^2 \frac{C}{\sqrt{r}} \right\}} \leq \frac{\left\langle \frac{1}{n} \mathbf{H}(Z_2)_i, \eta^k - \eta \right\rangle^2}{\|\boldsymbol{\theta}\|^4 \frac{C^2}{r}}.$$

similarly we get for  $\eta_i = 1$  that

$$\mathbf{1}_{\left\{ \left\langle \frac{1}{n} \mathbf{H}(Y^\top Y)_i, \eta^k \right\rangle \leq 0 \right\}} \mathbf{1}_{\mathbf{A}_i} \leq \frac{\left\langle \frac{1}{n} \mathbf{H}(Z_2)_i, \eta^k - \eta^* \right\rangle^2}{\|\boldsymbol{\theta}\|^4 \frac{C^2}{r}}.$$

It is clear that

$$\frac{1}{2} |\eta^{k+1} - \eta| \leq \sum_{\eta_i = -1} \mathbf{1}_{\left\{ \left\langle \frac{1}{n} \mathbf{H}(Y^\top Y)_i, \eta^k \right\rangle \geq 0 \right\}} + \sum_{\eta_i = 1} \mathbf{1}_{\left\{ \left\langle \frac{1}{n} \mathbf{H}(Y^\top Y)_i, \eta^k \right\rangle \leq 0 \right\}}.$$

Hence we get using the event  $\mathbf{A}_i$  that

$$\frac{1}{2n} |\eta^{k+1} - \eta| \leq \frac{\sum_{i=1}^n \left\langle \frac{1}{n} \mathbf{H}(Z_2)_i, \eta^k - \eta^* \right\rangle^2}{n \|\boldsymbol{\theta}\|^4 \frac{C^2}{r}} + \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\mathbf{A}_i^c}.$$

Using the observation that  $\|\eta^k - \eta^*\|_2^2 = 2|\eta^k - \eta^*|$  we get

$$\frac{1}{2n}|\eta^{k+1} - \eta|\mathbf{1}_B \leq \frac{C'}{2nr}|\eta^k - \eta^*|\mathbf{1}_B + \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{A_i^c}.$$

Since  $r > C'$ , we get that

$$\frac{1}{n}|\eta^k - \eta|\mathbf{1}_B \leq \left(\frac{C'}{r}\right)^k \frac{1}{n}|\eta^0 - \eta^*|\mathbf{1}_B + \frac{1}{1 - \frac{C'}{r}} \frac{2}{n} \sum_{i=1}^n \mathbf{1}_{A_i^c}.$$

In particular for  $k \geq 2 \log(n)$  and  $r \geq eC'$  we have

$$\left(\frac{C'}{r}\right)^k \frac{1}{n}|\eta^0 - \eta^*|\mathbf{1}_B \leq \frac{1}{n^2}.$$

Hence

$$\frac{1}{n}|\eta^k - \eta|\mathbf{1}_B \leq \frac{1}{n^2} + \frac{1}{1 - \frac{C'}{r}} \frac{2}{n} \sum_{i=1}^n \mathbf{1}_{A_i^c}.$$

Observe that if  $\frac{1}{1 - \frac{C'}{r}} \frac{2}{n} \sum_{i=1}^n \mathbf{1}_{A_i^c} = 0$  then  $\frac{1}{n}|\eta^k - \eta|\mathbf{1}_B = 0$ , if not then  $\frac{2}{n} \sum_{i=1}^n \mathbf{1}_{A_i^c} \geq \frac{2}{n}$ .

This leads to

$$\frac{1}{n}|\eta^k - \eta|\mathbf{1}_B \leq \left(\frac{1}{1 - \frac{C'}{r}} + o(1)\right) \frac{2}{n} \sum_{i=1}^n \mathbf{1}_{A_i^c}.$$

Finally we get

$$\frac{1}{n} \mathbf{E}(|\eta^k - \eta|) \leq \left(\frac{1}{1 - \frac{C'}{r}} + o(1)\right) \frac{2}{n} \sum_{i=1}^n \mathbf{P}(A_i^c) + \mathbf{P}(B^c).$$

The term  $\mathbf{P}(B^c)$  is upper bounded exactly as in the previous proof. For the other term observe that

$$\mathbf{P}(A_i^c) = \mathbf{P}\left(\|\boldsymbol{\theta}\|^2 \left(1 - \frac{C'}{r}\right) + \langle \xi_i, \boldsymbol{\theta} + \frac{1}{\sqrt{n}} \xi_{-i} \rangle + \langle \boldsymbol{\theta}, \frac{1}{\sqrt{n}} \xi_{-i} \rangle \leq 0\right) = G_\sigma\left(\frac{C'}{r}, \|\boldsymbol{\theta}\|^2\right).$$

□

## 6.2 Technical lemmas

**Lemma 17.** *Let  $A$  be a matrix in  $\mathbf{R}^{n \times n}$ , then*

$$\|H(A)\|_\infty \leq 2\|A\|_\infty.$$

*Proof.* From the linearity of  $H$ , we have that

$$\|H(A)\|_\infty \leq \|A\|_\infty + \|\text{diag}(A)\|_\infty.$$

Noticing that

$$\|\text{diag}(A)\|_\infty = \max_i |A_{ii}|,$$

we get that

$$\|\text{diag}(A)\|_\infty \leq \|A\|_\infty.$$

That concludes the proof.  $\square$

**Lemma 18.** *For any random matrix with independent columns, we have*

$$\|H(W^\top W)\|_\infty \leq 2\|W^\top W - \mathbf{E}(W^\top W)\|_\infty.$$

*Proof.* Since  $\mathbf{E}(W^\top W)$  is a diagonal matrix, it follows that

$$H(WW^\top) = H(WW^\top - \mathbf{E}(WW^\top)).$$

The result follows from Lemma 17.  $\square$

**Lemma 19.** *Let  $u \in \mathbf{S}^{p-1}$  and  $v \in \mathbf{S}^{n-1}$  then for some  $c, C > 0$*

$$\forall t \geq 2\sigma, \quad \mathbf{P}\left(\left\|\frac{1}{\sqrt{n}}Wvu^\top\right\|_{op} \geq t\right) \leq e^{-cnt/\sigma},$$

and

$$\mathbf{E}\left(\left\|\frac{1}{\sqrt{n}}Wvu^\top\right\|_{op}\right) \leq C\sigma^2.$$

*Proof.* We can easily check that

$$\left\|\frac{1}{\sqrt{n}}Wvu^\top\right\|_{op} \leq \frac{1}{\sqrt{n}}\|Wv\|_2.$$

Since  $\|v\|_2 = 1$ , we have that  $Wv$  is Gaussian with variance  $\sigma^2$ . We conclude using a tail inequality for quadratic forms of sub-Gaussian random variables using the fact that  $t \geq 2\sigma$ . The inequality in expectation is immediate by integration of the tail function.  $\square$

**Lemma 20.** *For some  $c, C, C' > 0$  we have*

$$\forall t \geq C\sigma^2 \left(1 \vee \sqrt{\frac{p}{n}}\right), \quad \mathbf{P}\left(\frac{1}{n}\|H(WW^\top)\|_{op} \geq t\right) \leq 2e^{-cnt/\sigma^2(1 \wedge \frac{tn}{p\sigma^2})},$$

and

$$\mathbf{E}\left(\frac{1}{n}\|H(WW^\top)\|_{op}\right) \leq C'\sigma^2(1 + \sqrt{p/n}).$$



*Proof.* Using Lemma 18, we get

$$\mathbf{P}\left(\frac{1}{n}\|H(WW^\top)\|_{op} \geq t\right) \leq \mathbf{P}\left(\frac{1}{n}\|WW^\top - \mathbf{E}(WW^\top)\|_{op} \geq t/2\right).$$

Now based on Lemma 1 in Royer (2017), we get moreover that

$$\mathbf{P}\left(\frac{1}{n}\|H(WW^\top)\|_{op} \geq t\sigma^2\right) \leq 9^n 2e^{-cnt(1 \wedge tn/p)},$$

for some  $c > 0$ . For  $t \geq C(1 \vee \sqrt{p/n})\sigma^2$  with  $C$  large enough, we get  $ct(1 \wedge tn/p\sigma^2) \geq 4\sigma^2 \log 9$ , hence

$$\mathbf{P}\left(\frac{1}{n}\|H(WW^\top)\|_{op} \geq t\right) \leq e^{-c'nt/\sigma^2(1 \wedge tn/p\sigma^2)},$$

for some  $c' > 0$ . The result in expectation is immediate by integration.  $\square$

**Lemma 21.** *For any  $x \in \{-1, 1\}^n$  and  $y \in \mathbf{R}^n$  such that  $\|y\|_2 = 1$ , we have*

$$\frac{1}{n}|x - \text{sign}(y)| \leq 2\left\|\frac{x}{\sqrt{n}} - y\right\|_2^2.$$

*Proof.* Observe that if  $x \in \{-1, 1\}$ , then

$$|x - \text{sign}(y)| = 2\mathbf{1}(x \neq \text{sign}(y)) \leq 2n\left|\frac{x}{\sqrt{n}} - y\right|^2$$

$\square$