Computation-information gap in high-dimensional clustering

Bertrand Even, Christophe Giraud, Nicolas Verzelen

Computation-information gap in high-dimensional clustering

Bertrand Even[1], Christophe Giraud[1], and Nicolas Verzelen[2]

[1] *Université Paris-Saclay, Laboratoire de mathématiques d'Orsay, Orsay, France*
[2] *INRAE, MISTEA, Univ. Montpellier, Montpellier, France*

February 29, 2024

Abstract

We investigate the existence of a fundamental computation-information gap for the problem of clustering a mixture of isotropic Gaussian in the high-dimensional regime, where the ambient dimension p is larger than the number n of points. The existence of a computation-information gap in a specific Bayesian high-dimensional asymptotic regime has been conjectured by [1] based on the replica heuristic from statistical physics. We provide evidence of the existence of such a gap generically in the high-dimensional regime $p \geq n$, by (i) proving a non-asymptotic low-degree polynomials computational barrier for clustering in high-dimension, matching the performance of the best known polynomial time algorithms, and by (ii) establishing that the information barrier for clustering is smaller than the computational barrier, when the number K of clusters is large enough. These results are in contrast with the (moderately) low-dimensional regime $n \geq \text{poly}(p, K)$, where there is no computation-information gap for clustering a mixture of isotropic Gaussian. In order to prove our low-degree computational barrier, we develop sophisticated combinatorial arguments to upper-bound the mixed moments of the signal under a Bernoulli Bayesian model.

1 Introduction

We investigate the problem of clustering a mixture of isotropic Gaussian in a high-dimensional set-up. The problem of clustering a mixture of Gaussian is a classical problem, which has lead to a large literature both in statistics and in machine learning [1–14].

Set-up. We observe a set of n points $Y_1, \dots, Y_n \in \mathbb{R}^p$, which have been generated as follows. For some unknown vectors $\mu_1, \dots, \mu_K \in \mathbb{R}^p$, some unknown $\sigma > 0$, and an unknown partition $G^* = \{G_1^*, \dots, G_K^*\}$ of $\{1, \dots, n\}$, the points $Y_1, \dots, Y_n$ are sampled independently with distribution

$$Y_i \sim \mathcal{N}(\mu_k, \sigma^2 I_p), \quad \text{for } i \in G_k^*.$$

We focus in this paper on the high-dimensional setting $p \geq n$, with balanced clusters

$$\frac{\max_k |G_k^*|}{\min_k |G_k^*|} \leq \alpha, \quad \text{for some } \alpha \geq 1. \tag{1}$$

Curse of dimensionality. For $K = 2$ clusters, in low dimension $p \ll n$, it is well known that the probability of misclassifying a new data point given the label of all the others decays like $\exp(-c\Delta^2)$ with the separation

$$\Delta^2 = \min_{l \neq r \in [1,K]} \frac{\|\mu_r - \mu_l\|^2}{2\sigma^2} \ . \tag{2}$$

In the high-dimensional regime $p \gg n$, the variance of the estimation of the high-dimensional means μ_k leads to the slower rate $\exp\left(-c' \frac{n}{p} \Delta^4\right)$ when $\Delta^2 \leq p/n$, see [7].

This curse of dimensionality for the classification problem has some repercussion on the clustering problem. When $K = 2$, and $p \geq n$ are large, [15] proved that a separation

$$\Delta^2 > 2\sqrt{\frac{2p\log(n)}{n}}$$

is necessary in order to perfectly recover the clusters, and also sufficient to recover them in polynomial time. For larger K, [7] proved that an SDP relaxation of Kmeans [16] provides a non-trivial clustering for $\Delta^2 \gtrsim \sqrt{pK^2/n}$, when $p \geq n$, where $\gtrsim$ hides a multiplicative constant depending only on α. Perfect clustering is also possible with single-linkage hierarchical clustering when $\Delta^2 \gtrsim \sqrt{p\log(n)} + \log(n)$ –see Appendix F for details on hierarchical clustering–, so, when $p \geq n$, non-trivial clustering is possible in polynomial time for

$$\Delta^2 \gtrsim \sqrt{\frac{pK^2}{n}} \wedge \sqrt{p\log(n)}. \tag{3}$$

Some non-rigorous arguments from statistical physics suggest that this minimal separation for non-trivial clustering in polynomial time may be optimal, up to a possible $\sqrt{\log(n)}$ factor for the second term. Indeed, building on the replica heuristic from statistical physics, [1] conjectures that, when the means μ_k are drawn i.i.d. with Gaussian $\mathcal{N}(0, p^{-1}\bar{\Delta}^2 I_p)$ distribution in $\mathbb{R}^p$, in the asymptotic regime where n, p go to infinity with $p/n \to \gamma \in [(K/2-2)^{-2}, +\infty)$, non-trivial clustering is possible in polynomial time only for $\bar{\Delta}^2 > \sqrt{\gamma K^2}$, while it is possible without computational constraints for $\bar{\Delta}^2 > 2\sqrt{\gamma K \log(K)}$, see also [17] for the problem of cluster detection.

These results are in contrast with the moderately low-dimensional setting, where it follows from [13] that for $n \geq \text{poly}(p, K)$, non-trivial clustering is possible in polynomial time when $\Delta^2 \gtrsim (\log(K))^{1+c}$, with $c > 0$, almost matching the information minimal separation $\Delta^2 \gtrsim \log(K)$ from [6, 10, 12], up to a small power of $\log(K)$. This set of results leaves open two fundamental questions:

1. Can we design a polynomial-time algorithm achieving non-trivial clustering for a separation smaller than (3) in the high-dimensional setting $p \geq n$?
2. What is the minimal separation Δ^2 necessary for non-trivial clustering in high-dimension, and is there a computation-information gap as conjectured in [1]?

Our contribution. We provide an answer to these two fundamental questions.

1. Our first contribution is to prove a non-asymptotic low-degree polynomial lower bound suggesting that the separation (3) is minimal, up to a possible polylog(n) factor, for clustering in polynomial time in the high-dimensional setting $p \geq n$.
2. Our second contribution is to prove that the information barrier for non-trivial clustering is

$$\Delta^2 \gtrsim \log(K) \vee \sqrt{\frac{pK\log(K)}{n}}, \tag{4}$$

with the exact Kmeans algorithm beeing (without surprise) information rate-optimal.

These two results provide evidence for the existence of a computation-information gap for the problem of clustering a mixture of isotropic Gaussians in high-dimension $p \geq n$, when the number K of clusters is larger than some constant K_0; confirming and generalizing the gap conjectured in [1].

$\Delta^2 \lesssim \sqrt{\frac{pK\log(K)}{n}}$	$\sqrt{\frac{pK\log(K)}{n}} \lesssim \Delta^2 \widetilde{\ll} \sqrt{\frac{pK^2}{n}} \wedge \sqrt{p}$	$\Delta^2 \gtrsim \sqrt{\frac{pK^2}{n}} \wedge \sqrt{p\log(n)}$
Impossible	Hard	Easy

Clustering hardness in high-dimension $p \geq n$. Here, $\widetilde{\ll}$ hides polylog(n) factors.

The main difficulty of the proof of the low-degree computational barrier is to bound mixed moments of the high-dimensional signal, drawn under a Bernoulli Bayesian model defined in Section 2. To derive these pivotal bounds, we develop sophisticated combinatorial arguments.

Literature review. The problem of clustering in high-dimension has been investigated in [4, 7, 8]. The latter provide some state-of-the-art controls on the (partial or perfect) recovery of the clusters in polynomial time, based on an SDP relaxation of Kmeans [16]. [15] considers the problem of perfect recovery when there are $K = 2$ clusters, identifying a sharp threshold for information-possible perfect recovery, and proving that perfect clustering is possible in polynomial time above this threshold with a simple Lloyd algorithm. In particular, there is no computation-information gap for a mixture of $K = 2$ isotropic Gaussian, whatever the ambient dimension p. In a Bayesian setting with a Gaussian prior on the μ_k, [1] conjectures a computation-information gap for clustering in the asymptotic limit where $p/n \to \gamma \in [(K/2-2)^{-2}, +\infty)$. Similarly, [17] proves that the information threshold for detecting the existence of clusters is smaller than the spectral detection threshold, when K is large enough. Interestingly, this information barrier for cluster detection in this Bayesian setting matches, up to a possible constant, the information barrier (4) for clustering, so that there is no test-estimation gap at the information level. On a different perspective, the estimation of the parameters of a Gaussian mixture distribution in high-dimension has been addressed in [18].

Contrary to our high-dimensional setting $p \geq n$, there is no computation-information gap for learning mixture of isotropic Gaussian [6, 10, 12, 13] in a moderately low-dimensional setting $n \geq \text{poly}(p, K)$. Some computation-information gaps have yet been shown in moderately low-dimension for learning mixture of non-isotropic Gaussian with unknown covariance. In such a setting, [19] and [14] establish some lower-bounds for the running time of any Statistical-Query algorithm (SQ-algorithm), enforcing a computation-information gap between SQ-algorithms and information optimal algorithms. We emphasize that in our high-dimensional setting, contrary to the moderately low-dimensional case, the computation-information gap is not induced by some non-isotropic effects. Indeed, when $p \geq n$, the computation-information gap shows up for isotropic Gaussian mixture. In addition, the performances of polynomial time estimators are similar for isotropic Gaussian and anisotropic subGaussian mixtures [7]. We refer to Section 4 for (i) a detailed discussion on the differences between the high and moderately low-dimensional settings, and (ii) a discussion highlighting that optimal clustering rates cannot be simply derived from estimation rates.

The low-degree polynomial model of computation requires the output of the algorithm to be computed by a low-degree polynomial of the entries of the input data. Many state-of-the-art algorithms, including spectral methods and approximate message passing algorithms, can be approximated by low-degree polynomials, and the class of low-degree polynomials is as powerful as the best known polynomial-time algorithms for many canonical problems, including planted clique [20], community detection [21], sparse PCA [22], and tensor PCA [23]. A low-degree polynomial lower bound is then a compelling evidence for computational hardness of a learning problem. Low-degree lower bounds have been first introduced in [20] –see also [24]–, and then extended in many settings. [25] and [26] provide some generic techniques for proving low-degree lower bounds in a wide range of situations. For example, building on these results, [27] provides evidence for the computation-information gap conjectured in [28] for clustering in the Stochastic Block Model.

In our setting, the random partition and the high number of dimensions cause a high dependence between the signal vectors. Thus, we have to appeal to delicate combinatorial arguments in order to upper-bound the mixed moments of the signal. We discuss this more precisely in the sketch of the proof of Theorem 1, in Section 2, and in the proof of the theorem in Appendix A.

Outline and notation. We state our computational lower bound in Section 2, we analyse the information-barrier for partial and perfect recovery in Section 3, and we discuss these results and their connections with the literature in Section 4. All the proofs are deferred to the appendices, though a sketch of the proof of the computational lower bound is provided in Section 2.

Throughout this documents, we use $\|\cdot\|_q$ for the L_q norm of a vector or of the entries of a matrix. For $q = 2$, we simply write $\|\cdot\|$ for the Euclidean norm of a vector, and $\|\cdot\|_F$ for the Frobenius norm of a matrix. The notations $\|\cdot\|_{op}$ and $\|\cdot\|_*$ respectively stand for the operator norm and the nuclear norm of a matrix.

We denote by $\mathcal{P}_\alpha$ the set of partitions of $[1, n]$ fulfilling (1). For a partition $G = \{G_1, \dots, G_K\}$, we define k_i^G as the integer such that $Y_i \in G_{k_i}$, and the partnership matrix $M_{ij}^G = \mathbf{1}_{k_i^G = k_j^G}$. For $G = G^*$, we simply write $k_i^* = k_i^{G^*}$ and $M^* = M^{G^*}$. We also define the proportion of misclassified points as

$$err(G, G^*) = \frac{1}{2n} \min_{\pi \in \mathcal{S}_K} \sum_{k=1}^{K} |G_k^* \Delta G_{\pi(k)}| \ , \tag{5}$$

where $A \Delta B$ stands the symmetric difference between the sets A and B, and $\mathcal{S}_K$ denotes the set of all permutations of $[1, K]$.

2 Low degree polynomial lower bound

Low degree polynomials are not well suited for directly outputting a partition $\widehat{G}$, which is combinatorial in nature. Instead, we focus on the problem of estimating the partnership matrix $M_{ij}^* = \mathbf{1}_{k_i^* = k_j^*}$ with low-degree polynomials. It turns out that estimating M^* and (partially) recovering the partition G^* are closely related. On the one hand, for any partition G, we have

$$\begin{aligned}\frac{1}{n(n-1)} \|M^G - M^*\|_F^2 &\leq \frac{1}{n(n-1)} \min_{\pi \in \mathcal{S}_K} \sum_{i \neq j = 1}^{n} \left(\mathbf{1}_{\pi(k_i^G) \neq k_i^*} \vee \mathbf{1}_{\pi(k_j^G) \neq k_j^*} \right) \\ &\leq \frac{2}{n} \min_{\pi \in \mathcal{S}_K} \sum_{i=1}^{n} \mathbf{1}_{\pi(k_i^G) \neq k_i^*} \ \leq 2 \ err(G, G^*),\end{aligned}$$

so, if it is possible to cluster (in polynomial time) with error $err(\hat{G}, G^*) \leq \rho$, then we can estimate M^* (in polynomial time) with error $n^{-2}\|M^{\hat{G}} - M^*\|_F^2 \leq 2\rho$. On the other hand, for any estimator $\widehat{M}$, we have

$$\sum_{i,j=1}^{n} \mathbf{1}_{|\widehat{M}_{ij} - M_{ij}^*| \geq 1/2} \leq 4\|\widehat{M} - M^*\|_F^2.$$

So, when $\|\widehat{M} - M^*\|_F^2 < 1/4$, pairing together points i, j fulfilling $\widehat{M}_{i,j} \geq 1/2$ provides a valid partition, equal to G^*.

To provide evidence of a computational barrier for the clustering problem, we build on this connection by proving a low-degree polynomial lower bound for the estimation of M^*. In this section, we consider the following generative prior for the partition G^* and the means $\mu_1, \dots, \mu_K$.

Definition 1. *We draw $k_1, \dots, k_n$ i.i.d. uniformly on $[1, K]$. For a given $\bar{\Delta} > 0$, independently from $(k_i)_{i\in[1,n]}$, we draw $\mu_1, \dots, \mu_K \in \mathbb{R}^p$ i.i.d. uniformly distributed on the hypercube $\mathcal{E} = \{+\varepsilon, -\varepsilon\}^p$, with $\varepsilon^2 = \frac{1}{p}\bar{\Delta}^2\sigma^2$. Then, conditionally on $(k_i)_{i=1,\dots,n}$ and $(\mu_k)_{k=1,\dots,K}$, the Y_i are independent with $\mathcal{N}(\mu_{k_i}, \sigma^2 I_p)$ distribution. The partition G^* is obtained from the k_i's with the canonical partitioning $G_k^* = \{i \in [1, n],\ k_i = k\}$, and $M_{ij}^* = \mathbf{1}_{k_i=k_j}$.*

We observe that in this model, for any $k \neq \ell$ the normalized square distance $\|\mu_k - \mu_\ell\|^2/(2\sigma^2)$ is equal to $\bar{\Delta}^2\left(1 + O_{\mathbb{P}}(p^{-1/2})\right)$.

Let $\mathbb{R}_D[Y]$ be the set of all polynomials in the observations $(Y_{ij})_{i,j\in[1,n]\times[1,p]}$ of degree at most D. We consider the degree-D minimum mean squared error defined similarly as in [26] by

$$MMSE_{\leq D} := \inf_{f_{ij}\in\mathbb{R}_D[Y]} \ \frac{1}{n(n-1)} \sum_{i\neq j=1}^{n} \mathbb{E}\left[(f_{ij}(Y) - M_{ij}^*)^2\right] \ . \tag{6}$$

We observe that the trivial estimator $\widehat{M}_{ii} = 1$ and $\widehat{M}_{ij} = 1/K$ for $i \neq j$ has a mean square error $\frac{1}{n(n-1)}\,\mathbb{E}\left[\|\widehat{M} - M^*\|_F^2\right] = \frac{1}{K} - \frac{1}{K^2}$. Our main result is the next theorem, which identifies a regime where low-degree polynomials cannot perform significantly better than the trivial estimator in the high-dimensional setting $p \geq n$. We refer to Appendix A for a proof of this theorem.

Theorem 1. *Let $D \in \mathbb{N}$. If $p \geq n$ and $\zeta_n := \frac{\bar{\Delta}^4 D^8 (1+D)^4}{p} \max\left(\frac{n}{K^2}, 1\right) < 1$, then under the prior of Definition 1, we have*

$$MMSE_{\leq D} \geq \frac{1}{K} - \frac{1}{K^2}\left(1 + \frac{\zeta_n}{(1-\sqrt{\zeta_n})^3}\right). \tag{7}$$

In particular, if $\bar{\Delta}^2 \ll D^{-6}\left(\sqrt{\frac{pK^2}{n}} \wedge \sqrt{p}\right)$, then $MMSE_{\leq D} = \frac{1}{K} - \frac{1+o(1)}{K^2}$.

Remark: The same result holds (with a different power of D), when, in Definition 1, the prior on the μ_k is i.i.d. $\mathcal{N}(0, \varepsilon I_p)$, instead of i.i.d. uniform on $\mathcal{E} = \{+\varepsilon, -\varepsilon\}^p$.

The second part of Theorem 1 ensures that, when $p \geq n$, low-degree polynomials with degree $D \leq (\log(n))^{1+\eta}$ do not perform better than the trivial estimator when

$$\bar{\Delta}^2 \ll (\log n)^{-6(1+\eta)}\left(\sqrt{\frac{pK^2}{n}} \wedge \sqrt{p}\right).$$

Since lower-bounds for low-degree polynomials with degree $D \leq (\log(n))^{1+\eta}$ are considered as evidence of the computational hardness of the problem, Theorem 1 suggests computational hardness of estimating M^* when $\bar{\Delta}^2 \ll (\log n)^{-6(1+\eta)}\left(\sqrt{\frac{pK^2}{n}} \wedge \sqrt{p}\right)$ and $p \geq n$. Since, as made explicit above, estimation of M^* is possible in polynomial time when clustering is possible in polynomial time, this provides compelling evidence for the computational hardness of the clustering problem in this regime. Conversely, we explained in the introduction, that non-trivial clustering is possible in polynomial-time under the almost matching condition $\Delta^2 \gtrsim \sqrt{\frac{pK^2}{n}} \wedge \sqrt{p\log(n)}$.

Sketch of the proof By linearity of the loss function, we only consider, without loss of generality, the problem of estimating $x = M_{12}^*$ when $\sigma^2 = 1$. We need to prove that

$$\inf_{f\in\mathbb{R}_D[Y]} \ \mathbb{E}\left[(f(Y) - x)^2\right] \geq \frac{1}{K} - \frac{1}{K^2}\left(1 + \frac{\zeta_n}{(1-\sqrt{\zeta_n})^3}\right).$$

Since $\mathbb{E}[x^2] = 1/K$, the problem boils down –see [26]– to proving that, the so-called low degree correlation $corr_{\leq D}$ satisfies the following

$$corr_{\leq D} := \sup_{\substack{f \in \mathbb{R}_D[Y] \\ \mathbb{E}[f^2(Y)]=1}} \mathbb{E}(f(Y)x) \leq \frac{1}{K}\sqrt{1 + \frac{\zeta_n}{(1-\sqrt{\zeta_n})^3}} \ .$$

Interestingly, we can rewrite the observed matrix Y as a Gaussian additive model $Y = X + E$, where $E \in \mathbb{R}^{n\times p}$ is made of independent standard normal entries, and $X = A\underline{\mu}$ where the matrix $A \in \{0,1\}^{n\times K}$ contains exactly one non-zero entry on each row and its position is sampled uniformly at random, and where the matrix $\epsilon^{-1}\underline{\mu} \in \mathbb{R}^{p\times K}$ is made of independent Rademacher random variables.

This allows us to apply the general results of [26], which bound the low degree correlation in terms of a sum of cumulants

$$corr^2_{\leq D} \leq \sum_{\substack{\alpha \in \mathbb{N}^{n\times p} \\ |\alpha| \leq D}} \frac{\kappa_\alpha^2}{\alpha!} \ , \tag{8}$$

where the α's run over all integer valued matrices whose sum is at most D, and where κ_α is the cumulants of the random variables $(x, \underbrace{X_{1,1}, \dots, X_{1,1}}_{\alpha_{1,1}}, \dots, \underbrace{X_{i,j}, \dots, X_{i,j}}_{\alpha_{i,j}}, \dots)$. The bound (8) turned out to be instrumental for establishing low degree polynomials lower bounds for submatrix estimation [26], and for Stochastic Block model (SBM) estimation [27]. In these two works, the authors follow a two-steps approach: first, they prune the sum in (8) by characterizing all the cumulants that are equal to zero. Second, they bound the cumulants as a polynomial sum of mixed moments.

In comparison to the above works, we use the same general strategy, but the structure of the signal matrix X is more involved. Indeed, in submatrix problem, the matrix X only contains a single non-zero blocks whereas, for SBM, X is, up to a permutation, a block-diagonal matrix. Here, we need to leverage on the fact that the rectangular matrix $X = A\underline{\mu}$ jointly involves a random partition matrix A and an high-dimensional random matrix $\underline{\mu}$. As a consequence, we need to rely on more subtle arguments both for the pruning step, that is for characterizing null cumulants, and for bounding mixed moments with respect to the entries of X.

For that purpose, we represent $\alpha \in \mathbb{N}^{n\times p}$ as a bi-partite multigraph $\mathcal{G}_\alpha$ between the set $[n]$ of points, and the set $[p]$ of variables and we write $\mathcal{G}_\alpha^-$ for its restriction to non-isolated nodes. In Lemma 5, we first establish that the cumulant κ_α is null unless the graph $\mathcal{G}_\alpha^-$ satisfies the three following properties: (i) $\mathcal{G}_\alpha^-$ is connected, (ii) Both the first and the second points belong to the connected component, and (iii) Each variable in $\mathcal{G}_\alpha^-$ is connected to at least two distinct nodes. Indeed, if at least one of these properties is not satisfied, it is possible to partition $(x, \dots, \underbrace{X_{i,j}, \dots, X_{i,j}}_{\alpha_{i,j}}, \dots)$ into two set of independent random variable, which implies the nullity of the cumulant.

Now that we have pruned the sum in (8) by restricting ourselves to such matrices α, we need to control the non-zero cumulants κ_α. Since cumulants express as linear combination of moments, we bound mixed moments of the form $\mathbb{E}[X^\gamma] = \mathbb{E}[\prod_{i=1}^n \prod_{j=1}^p X_{i,j}^{\gamma_{ij}}]$ and $\mathbb{E}[xX^\gamma]$, for matrices $\gamma \in \mathbb{N}^{n\times p}$. We establish in Lemma 6 that

$$\mathbb{E}[X^\gamma] \leq \varepsilon^{|\gamma|} \min\left(1, |\gamma|^{|\gamma|}\left(\frac{1}{K}\right)^{l_\gamma - \frac{|\gamma|}{2} - CC_\gamma}\right) \ , \tag{9}$$

where $|\gamma| = \sum_{ij}\gamma_{ij}$ is the number of edges of $\mathcal{G}_\gamma^-$, CC_γ is the number of connected components of $\mathcal{G}_\gamma^-$, and l_γ is the number of nodes.

For establishing (9), we first rely on the fact that the entries of $\underline{\mu}$ are independent and follow a symmetric distributions. Since $X = A\underline{\mu}$, where we recall that $A_{ik} = \mathbf{1}_{k_i=k}$ encodes the partition of the n points, we have

$$X^\gamma = \prod_{k=1}^{K} \prod_{j=1}^{p} \underline{\mu}_{kj}^{\sum_{i=1}^n A_{i,k}\gamma_{ij}} \ .$$

The conditional expectation of X^γ given A is therefore non-zero (and is equal to $\epsilon^{|\gamma|}$), if and only if, $\sum_{i=1}^n A_{i,k}\gamma_{ij}$ is even for all (k,j). We call the latter a (A,γ) parity property. Since $\epsilon^{-1}\underline{\mu}_{kj}$ is a Rademacher random variable, it follows from the above that

$$\mathbb{E}[X^\gamma] = \epsilon^\gamma \, \mathbb{P}[(A,\gamma) \text{ satisfies the parity property}] \ .$$

Next, we characterize in Lemma 8 the partition matrices A (or equivalently the partition G^*) that satisfy the parity property. In particular, we show that the partition induced by G^* on the set of non-zero rows of γ only contains a small number of groups. More precisely, we bound this number of groups in terms of $|\gamma|$, the number of non-zero rows of γ, the number of non-zero columns of γ, and CC_γ, the number of connected components of $\mathcal{G}_\gamma^-$. In turn, this condition on the number of groups enforces that $\mathbb{P}[(A,\gamma)$ satisfies the parity property$]$ is small. This combinatorial argument for establishing Lemma 8 is the main technical result in our proof.

Finally, we build upon the mixed moment bounds (9) to control the cumulants κ_α. Coming back to (8), this allows us to conclude.

3 Information barrier

3.1 Clustering below the computational barrier with exact Kmeans

For a mixture of isotropic Gaussian, the partition $\hat{G}$ maximizing the likelihood is the exact K-means partitioning, which minimizes the criterion

$$\hat{G} \in \operatorname*{argmin}_{G \in \mathcal{G}_K} \mathrm{Crit(G)} \ , \quad \text{where} \quad \mathrm{Crit(G)} = \sum_{k=1}^{K} \sum_{a \in G_k} \left\| Y_a - \frac{1}{|G_k|} \sum_{b \in G_k} Y_b \right\|^2 , \tag{10}$$

with $\mathcal{G}_K$ the set of partitions of $[1,n]$ in K groups. Minimizing $\mathrm{Crit}(G)$ is NP-hard in general, and even hard to approximate [29].

Next theorem proves that exact K-means succeeds to produce non-trivial clustering for a separation smaller than the computational barrier (3) for $p \geq n$, and for K larger than some constant K_0.

Theorem 2. *Assume that G^* belongs to the set of balanced partitions $\mathcal{P}_\alpha$. Then, there exist some constants c, c', c'' depending only on α, such that the following holds. If*

$$\Delta^2 \geq c \left(\log(K) \vee \sqrt{\frac{pK\log(K)}{n}} \right) \ , \tag{11}$$

then, we have with probability at least $1 - c'/n^2$

$$err(\hat{G}, G^*) \leq e^{-c''s^2}, \quad \textit{where} \quad s^2 = \Delta^2 \wedge \frac{n\Delta^4}{pK}. \tag{12}$$

This result follows from the more precise Theorem 5 stated and proved in Appendix B. The rate $e^{-c''s^2}$ for the proportion of misclassified points, matches the optimal probability of wrongly

classifying a data point given the label of all the others [7]. The term Δ^2 in s^2 corresponds to the rate in low-dimension, while the term $\frac{n\Delta^4}{pK}$ is induced by the minimal error $\sigma\sqrt{pK/n}$ for estimating the means μ_k in dimension p with n/K observations. We underline yet in Section 4, that the minimal separation (11) for clustering cannot be readily derived from the minimal estimation rate for the means.

In the high-dimensional setting $p \geq n$, Theorem 2 ensures that for

$$\Delta^2 \gtrsim \sqrt{\frac{pK\log(K)}{n}}$$

the exponential exponent $c''s^2$ is larger than $(1+\eta)\log(K)$, so the proportion of misclustered points by exact Kmeans is smaller than $1/K^{1+\eta}$. Exact Kmeans then performs a non-trivial clustering in this regime, breaking the computational barrier (3) established in the previous section, when K is larger than some constant K_0.

3.2 Information lower bound

For $\bar{\Delta} > 0$, let $\Theta_{\bar{\Delta}}$ denote the set of K-tuples $\mu_1, \ldots, \mu_K \in (\mathbb{R}^p)^K$ that satisfy $\Delta \geq \bar{\Delta}$, with Δ defined in (2). Given $\mu_1, \ldots, \mu_K \in (\mathbb{R}^p)^K$ and G a partition of $[1, n]$, we denote by $\mathbb{P}_{\mu,G}$ the probability distribution of the random variables $(Y_1, \ldots, Y_n) \in (\mathbb{R}^p)^n$ generated as follows: $Y_1, \ldots, Y_n$ are independent, and $Y_i \sim \mathcal{N}(\mu_k, \sigma^2 I_p)$ when $i \in G_k$. The next result, proved in Appendix C, provides a minimax lower bound on the partial recovery of a partition.

Theorem 3. *There exist c,c', C and K_0 numerical constants such that the following holds. Assume that $p \geq c\log(K)$, $K \geq K_0$, $n \geq 2K$, and $\alpha \geq \frac{3}{2}$. Then, for any estimator $\hat{G}$, we have*

$$\sup_{G\in\mathcal{P}_\alpha} \sup_{\mu\in\Theta_{\bar{\Delta}}} \mathbb{E}_{\mu,G}[err(\hat{G}, G)] \geq C\,, \quad \textit{when} \quad \bar{\Delta}^2 \leq c'\left(\log(K) \vee \sqrt{\frac{pK\log(K)}{n}}\right).$$

For exact Kmeans, according to (12) from Theorem 3, non-trivial clustering is achieved for balanced clusters when

$$s^2 = \Delta^2 \wedge \frac{n\Delta^4}{pK} \gtrsim \log(K),$$

or equivalently $\Delta^2 \gtrsim \log(K) \vee \sqrt{\frac{pK\log(K)}{n}}$. Exact Kmeans is then information optimal for non-trivial clustering, and the information threshold for non-trivial clustering is

$$\Delta^2 \gtrsim \log(K) \vee \sqrt{\frac{pK\log(K)}{n}}.$$

We observe that this information barrier for clustering matches, up to a possible constant, the information barrier $\bar{\Delta}^2 \geq 2\sqrt{\gamma K\log(K)} + 2\log(K)$ established in [17] for detecting clusters in a Bayesian setting with a Gaussian prior $\mathcal{N}(0, p^{-1}\bar{\Delta}^2 I_p)$ on the μ_k. Hence, there is no (significant) test-estimation gap at the information level. We refer to Section 4 for a comparison of the information rates for clustering and estimation.

We complement this result with a lower bound for perfect recovery of the planted partition, proved in Appendix D.

Theorem 4. *There exist numerical constants c, C and n_0 such that the following holds. Assume that $n \geq 9K/2$, $\alpha \geq \frac{3}{2}$ and $n \geq n_0$. Then, for any estimator $\hat{G}$,*

$$\sup_{G\in\mathcal{P}_\alpha} \sup_{\mu\in\Theta_{\bar{\Delta}}} \mathbb{P}_{\mu,G}\left[\hat{G} \neq G\right] > C\,, \quad \textit{when} \quad \bar{\Delta}^2 \leq c\left(\log(n) \vee \sqrt{\frac{pK\log(n)}{n}}\right).$$

For $K \leq \log(n)$, we recover the optimal separation from [15] ($K = 2$) and [30], up to a multiplicative constant. Perfect recovery corresponds to a proportion of misclustered points $err(\hat{G}, G^*)$ smaller than $1/n$. For exact Kmeans, according to (12), perfect recovery is achieved for

$$s^2 = \Delta^2 \wedge \frac{n\Delta^4}{pK} \gtrsim \log(n),$$

or equivalently $\Delta^2 \gtrsim \log(n) \vee \sqrt{\frac{pK\log(n)}{n}}$. Exact Kmeans is then also optimal for exact recovery, and the information threshold for perfect clustering is then

$$\Delta^2 \gtrsim \log(n) \vee \sqrt{\frac{pK\log(n)}{n}}. \tag{13}$$

When $K \lesssim \log(n)$, [7] shows that an SDP relaxation of Kmeans [16] also succeeds to perfectly recover the clusters when (13) is met, so there is no separation in this regime –see also [30]. Yet, in the high-dimensional setting $p \geq n$, we observe that the threshold (13) is smaller than the computational barrier (3) when $K \gtrsim \log(n)$, so there is also a computation-information gap for perfect recovery in this regime, thereby confirming the conjecture of [30].

4 Discussion

4.1 Comparison to the moderately low-dimensional setting

No non-isotropic effect. We emphasize that compared to the moderately low-dimensional setting $n \geq \text{poly}(p, K)$, the computational hardness of clustering in the high-dimensional regime $p \geq n$ is not driven by any non-isotropic effect. Indeed, contrary to the low-dimensional setting where there is no computation-information gap for learning mixture of isotropic Gaussian [6, 13], we prove the computation-information gap for a mixture of Gaussians with covariances known to be all equal to the identity. Furthermore, there is no difference between the Gaussian and the sub-Gaussian setting, in the sense that for a mixture of possibly anisotropic sub-Gaussian distribution, clustering is also possible in polynomial time above the computational barrier (3) established for Gaussian mixture, for example with an SDP relaxation of Kmeans [7, 16] for $K \leq \sqrt{n}$, or with single linkage hierarchical clustering [31] for $K > \sqrt{n}$.

Comparison to moderately low dimension. Contrary to the high-dimensional setting, where the computational hardness seems tightly related to the BBP transition for the largest eigenvalue of the Gram matrix of the observations [32, 33], the computational hardness in moderately low-dimensional settings $n \geq \text{poly}(p, K)$ is completely driven by the unknown non-isotropy of the components of the mixture. A first example of non-isotropic mixture giving rise to a computation-information gap is the so-called example of "parallel pancakes" [19]. In this example, the K unknown centers of the Gaussian distribution are aligned along an unknown direction v, and the unknown covariances are all equal to the identity, except in the direction v, where they are very thin. The key feature of this construction, is that the $2K - 1$ first moments of the mixture distribution match those of a standard Gaussian, so that it is impossible to figure out the direction v from the $2K - 1$ first moments. As a consequence, [19] proves a lower-bound for the running time of any Statistical-Query algorithm (SQ-algorithm), enforcing a computation-information gap between SQ-algorithms and information-optimal algorithms in this moderately low-dimensional setting.

This approach has been extended by [14], who has adapted this construction for centers with separation $\Delta^2 \geq k^\eta$ much larger than the information-minimal separation $\Delta^2 \gtrsim \log(K)$ [6] in moderately low dimension. For such a large separation, the centers are drawn according to a

standard Gaussian on a (unknown) random subspace of dimension $d \approx \Delta^2$. [14] then proves again a lower-bound for the running time of any Statistical-Query algorithm (SQ-algorithm), enforcing a computation-information gap between SQ-algorithms and information optimal algorithms in this (moderately low-dimensional) setting.

4.2 Comparing estimation and clustering rates

Assume with no loss of generality that $\sigma^2 = 1$. Theorems 2-3 show that the information-minimal separation for clustering is $\Delta^2 \gtrsim \log(K) \vee \sqrt{Kp\log(K)/n}$. When the separation Δ^2 is larger than $\log(K)$, it is known that the information-minimal rate for estimating the means μ_k is at least $\sqrt{Kp/n}$. This rate stems from the fact that we estimate p-dimensional vectors with about n/K observations for each of them. In the moderately low-dimensional regime $n \geq pK^3$, estimation at this rate (up to possible log factors) is actually information-possible [10]. We then underline that for $\log(K) \leq Kp/n$, at the information-minimal separation for clustering $\Delta^2 \asymp \sqrt{Kp\log(K)/n} \geq \log(K)$, we cannot estimate the means μ_k better than with a precision $\sqrt{Kp/n} \asymp \Delta^2$, up to log factors. This precision is much larger than the minimum distance $\sqrt{2}\Delta$ between the means. In particular, a natural *estimate-then-cluster* strategy that would consist in (i) estimating the means with precision at least Δ, and then (ii) apply Linear Discriminant Analysis with the estimated means $\hat{\mu}_k$, would require a separation at least $\Delta^2 \gtrsim \log(K) \vee (Kp/n)$ (up to possible log factors), which is much larger than the information-minimal separation (11) for clustering. The message is then that optimal rates for the estimation problem do not directly provide useful information for the clustering problem.

4.3 Computation-information gap for partnership matrix estimation

While our primary interest is on clustering, we point out below that our results provide evidence for the existence of a computation-information gap for the estimation of the partnership matrix M^* in Frobenius norm, in the high-dimensional regime $p \geq n$.

As discussed in Section 2, starting from a partition $\hat{G}$, we can estimate the partnership matrix $M^*_{ij} = \mathbf{1}_{k^*_i = k^*_j}$ with $M^{\hat{G}} = \mathbf{1}_{k^{\hat{G}}_i = k^{\hat{G}}_j}$. The mean squared error $\frac{1}{n(n-1)}\|M^{\hat{G}} - M^*\|_F^2$ is then upper bounded by twice the clustering error $err(\hat{G}, G^*)$. Relying on this connection, we prove below that, when the dimension is high $p \geq n$, and when there is a large number of points $n \gtrsim K^2\log(n)$, the exact Kmeans provides an estimation of M^* below the computational barrier in the generative model of Definition 1.

Corollary 1. *Let us consider the generative model of Definition 1. Assume that $n \geq cK^2\log(n)$, with $c > 0$ a numerical constant, and $p \geq n$. There exists c' and C, two numerical constants, such that the partnership matrix estimation induced by the exact Kmeans partition $\hat{G}$ fulfills*

$$\frac{1}{n(n-1)}\,\mathbb{E}\left[\|M^{\hat{G}} - M^*\|_F^2\right] \leq \frac{C}{n^2}\,, \quad \textit{when} \quad \Delta^2 \geq c'\sqrt{\frac{pK\log(n)}{n}}\,.$$

We refer to Section E for a proof of this corollary of Theorem 5 stated in Appendix B, which slightly generalizes Theorem 2. Thus, when $\sqrt{\frac{pK}{n}\log(n)} \lesssim \Delta^2 \ll \sqrt{\frac{pK^2}{n}}$, the error obtained with the exact K-means estimator decays much faster than the trivial error $\frac{1}{K}$ obtained with the trivial estimator ρ, and with the best polynomial of degree at most $D(n) = \log(n)^{1+\eta}$, when $\Delta^2 \ll (\log n)^{-6(1+\eta)}\sqrt{pK^2/n}$.

4.4 Limitations

Inspired from [26], our analysis for establishing a low-degree polynomial lower bound has the nice feature to provide a rigorous and non-asymptotic computational lower-bound, but it has the drawback to be a bit rough, and a spurious polylog(n) factor shows up in the computational lower bound. Removing this polylog(n) factor in the proof would probably require a different strategy for bounding the correlation $corr_D^2$, by precisely keeping track of all terms. The complexity of such an analysis would go well beyond the complexity of our proof of Theorem 1. Given the already high complexity of our proof, we do not intend to pursue in this direction.

Another drawback of our analysis is that it is limited to the dimension range $p \geq n$, while a computation-information gap may also exist for smaller values of the ambient dimension p. Indeed, when the means μ_k are drawn i.i.d. with Gaussian $\mathcal{N}(0, p^{-1}\bar{\Delta}^2 I_p)$ distribution in $\mathbb{R}^p$, and in the asymptotic regime where n, p go to infinity with $p/n \to \gamma \in [(K/2-2)^{-2}, +\infty)$, [1] conjectures that non-trivial clustering is possible in polynomial time only for $\bar{\Delta}^2 > \sqrt{\gamma K^2}$, while it is possible without computational constraints for $\bar{\Delta}^2 \gtrsim \sqrt{\gamma K \log(K)} \vee \log(K)$, which is smaller for $K \gtrsim 1 \vee (\gamma^{-1/2} \log K)$. Similarly, for the problem of detecting the existence of clusters in the same specific setting, [17] shows that spectral detection is not possible at the information threshold for $\gamma \gtrsim (\log(K)/K)^2$. Indeed, from BBQ transition, the largest eigenvalue of the Gram matrix of the data points singles out of the bulk of the spectrum only for $\bar{\Delta}^2 \geq \sqrt{\gamma K^2}$, while detection is information possible for $\bar{\Delta}^2 \geq 2\sqrt{\gamma K \log(K)} + 2\log(K)$. These two results suggest the existence of a computation-information gap not only for $p \geq n$, as considered in this paper, but more generally for $p \gtrsim n(\log(K)/K)^2$ and $K \geq K_0$. In Appendix A.6, we adapt the proof of Theorem 1 in order to provide a computational lower-bound when $p \leq n$. We believe that our computational barrier is not tight in this regime, yet it already provides evidence for the existence of a computation-information gap when

$$\frac{n}{K} \vee K \,\widetilde{\ll}\, p \leq n\,,$$

where $\widetilde{\ll}$ hides polylog(n) factors. Proving a more tight computational barrier for the range $n(\log(K)/K)^2 \lesssim p \leq n$ is left for future investigation.

References

1. Lesieur, T. *et al. Phase transitions and optimal algorithms in high-dimensional Gaussian mixture clustering* in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (2016), 601–608.
2. Dasgupta, S. *Learning mixtures of Gaussians* in *40th Annual Symposium on Foundations of Computer Science (Cat. No.99CB37039)* (1999), 634–644.
3. Vempala, S. & Wang, G. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences* **68.** Special Issue on FOCS 2002, 841–860 (2004).
4. Lu, Y. & Zhou, H. H. Statistical and Computational Guarantees of Lloyd's Algorithm and its Variants. *ArXiv e-prints.* arXiv: `1612.02099 [math.ST]` (Dec. 2016).
5. Diakonikolas, I., Kane, D. M. & Stewart, A. *List-decodable robust mean estimation and learning mixtures of spherical gaussians* in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing* (Association for Computing Machinery, 2018), 1047–1060.
6. Regev, O. & Vijayaraghavan, A. *On Learning Mixtures of Well-Separated Gaussians* in *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)* (Oct. 2017), 85–96.
7. Giraud, C. & Verzelen, N. Partial recovery bounds for clustering with the relaxed K-means. *Mathematical Statistics and Learning* **1,** 317–374 (2019).
8. Fei, Y. & Chen, Y. *Hidden Integrality of SDP Relaxations for Sub-Gaussian Mixture Models* in *Proceedings of the 31st Conference On Learning Theory* **75** (PMLR, 2018), 1931–1965.
9. Chen, X. & Yang, Y. Hanson–Wright inequality in Hilbert spaces with application to K-means clustering for non-Euclidean data. *Bernoulli* **27,** 586–614 (2021).
10. Kwon, J. & Caramanis, C. *The EM Algorithm gives Sample-Optimality for Learning Mixtures of Well-Separated Gaussians* in *Proceedings of Thirty Third Conference on Learning Theory* (eds Abernethy, J. & Agarwal, S.) **125** (PMLR, Sept. 2020), 2425–2487.
11. Segol, N. & Nadler, B. Improved convergence guarantees for learning Gaussian mixture models by EM and gradient EM. *Electronic Journal of Statistics* **15,** 4510–4544 (2021).
12. Romanov, E., Bendory, T. & Ordentlich, O. On the Role of Channel Capacity in Learning Gaussian Mixture Models. *Proceedings of Machine Learning Research vol 178:1–50* (2022).
13. Liu, A. & Li, J. *Clustering mixtures with almost optimal separation in polynomial time* in *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing* (Association for Computing Machinery, Rome, Italy, 2022), 1248–1261.
14. Diakonikolas, I., Kane, D. M., Pittas, T. & Zarifis, N. *SQ Lower Bounds for Learning Mixtures of Separated and Bounded Covariance Gaussians* in *Proceedings of Thirty Sixth Conference on Learning Theory* (eds Neu, G. & Rosasco, L.) **195** (PMLR, Dec. 2023), 2319–2349.
15. Ndaoud, M. Sharp optimal recovery in the two component Gaussian mixture model. *The Annals of Statistics* **50,** 2096–2126 (2022).
16. Peng, J. & Wei, Y. Approximating K-means-type Clustering via Semidefinite Programming. *SIAM J. on Optimization* **18,** 186–205 (Feb. 2007).
17. Banks, J., Moore, C., Vershynin, R., Verzelen, N. & Xu, J. Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization. *IEEE Transactions on Information Theory* **64,** 4872–4894 (2018).
18. Doss, N., Wu, Y., Yang, P. & Zhou, H. H. Optimal estimation of high-dimensional Gaussian location mixtures. *The Annals of Statistics* **51,** 62–95 (2023).
19. Diakonikolas, I., Kane, D. M. & Stewart, A. *Statistical Query Lower Bounds for Robust Estimation of High-Dimensional Gaussians and Gaussian Mixtures* in *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)* (2017), 73–84.

20. Barak, B. *et al.* A Nearly Tight Sum-of-Squares Lower Bound for the Planted Clique Problem. *SIAM Journal on Computing* **48,** 687–735. eprint: `https://doi.org/10.1137/17M1138236` (2019).

21. Hopkins, S. B. & Steurer, D. *Efficient Bayesian Estimation from Few Samples: Community Detection and Related Problems* in *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)* (2017), 379–390.

22. Ding, Y., Kunisky, D., Wein, A. S. & Bandeira, A. S. Subexponential-time algorithms for sparse PCA. *Foundations of Computational Mathematics,* 1–50 (2023).

23. Hopkins, S. B. *et al. The Power of Sum-of-Squares for Detecting Hidden Structures* in *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE Computer Society, Los Alamitos, CA, USA, Oct. 2017), 720–731.

24. Hopkins, S. *Statistical inference and the sum of squares method* PhD thesis (Cornell University, 2018).

25. Kunisky, D., Wein, A. S. & Bandeira, A. S. *Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio* in *ISAAC Congress (International Society for Analysis, its Applications and Computation)* (2019), 1–50.

26. Schramm, T. & Wein, A. S. Computational barriers to estimation from low-degree polynomials. *The Annals of Statistics* **50,** 1833–1858 (2022).

27. Luo, Y. & Gao, C. *Computational Lower Bounds for Graphon Estimation via Low-degree Polynomials* 2023. arXiv: `2308.15728 [math.ST]`.

28. Decelle, A., Krzakala, F., Moore, C. & Zdeborová, L. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E* **84,** 066106 (6 Dec. 2011).

29. Awasthi, P., Charikar, M., Krishnaswamy, R. & Sinop, A. K. *The Hardness of Approximation of Euclidean k-Means* in *31st International Symposium on Computational Geometry (SoCG 2015)* **34** (Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2015), 754–767.

30. Chen, X. & Yang, Y. Cutoff for exact recovery of gaussian mixture models. *IEEE Transactions on Information Theory* **67,** 4223–4238 (2021).

31. Giraud, C. *Introduction to high-dimensional statistics* (CRC Press, Boca Raton, FL, 2021).

32. Baik, J., Arous, G. B. & Péché, S. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability* **33,** 1643–1697 (2005).

33. Debashis, P. ASYMPTOTICS OF SAMPLE EIGENSTRUCTURE FOR A LARGE DIMENSIONAL SPIKED COVARIANCE MODEL. *Statistica Sinica* **17,** 1617–1642 (2007).

34. Novak, J. Three lectures on free probability. *Random matrix theory, interacting particle systems, and integrable systems* **65,** 13 (2014).

A Proof of Theorem 1

With no loss of generality, we assume in all the proof that $\sigma^2 = 1$. Let $D \in \mathbb{N}$. We recall the assumption that $p \geq n$, and

$$\zeta_n = \frac{\bar{\Delta}^4 D^8 (1+D)^4}{p} \max\left(\frac{n}{K^2}, 1\right) < 1.$$

Since the minimization problem defining $MMSE_{\leq D}$ in Equation (6) is separable, and since the random variables M^*_{ij} are exchangeable, the $MMSE_{\leq D}$ can be reduced to

$$\begin{aligned} MMSE_{\leq D} &= \frac{1}{n(n-1)} \sum_{i \neq j = 1}^{n} \inf_{f_{ij} \in \mathbb{R}_D(Y)} \mathbb{E}\left[\left(f_{ij}(Y) - M^*_{ij}\right)^2\right] \\ &= \inf_{f \in \mathbb{R}_D(Y)} \mathbb{E}\left[\left(f(Y) - M^*_{12}\right)^2\right] \ . \end{aligned}$$

In the remaining of the proof, we write $x = M^*_{12} = \mathbf{1}_{k_1 = k_2}$. Then, our goal is to upper-bound

$$MMSE_{\leq D} = \inf_{f \in \mathbb{R}_D(Y)} \mathbb{E}\left[\left(f(Y) - x\right)^2\right] \ .$$

As noticed by [26], the $MMSE_{\leq D}$ can be further decomposed as

$$MMSE_{\leq D} = \mathbb{E}[x^2] - corr^2_{\leq D} = \frac{1}{K} - corr^2_{\leq D} \ ,$$

where $corr^2_{\leq D}$ is the degree-D maximum correlation

$$corr_{\leq D} := \sup_{\substack{f \in \mathbb{R}_D[Y] \\ \mathbb{E}[f^2(Y)]=1}} \mathbb{E}(f(Y)x) = \sup_{\substack{f \in \mathbb{R}_D[Y] \\ \mathbb{E}(f^2(Y)) \neq 0}} \frac{\mathbb{E}[f(Y)x]}{\sqrt{\mathbb{E}(f^2(Y))}} \ . \tag{14}$$

Hence, in order to prove Theorem 1, it is enough to prove that

$$corr^2_D \leq \frac{1}{K^2}\left(1 + \frac{\zeta_n}{(1-\sqrt{\zeta_n})^3}\right) \ .$$

The model of Definition 1 is a particular instance of the Additive Gaussian Noise Model considered in [26]. Hence, we can use Theorem 2.2 from [26] that we recall here. For a matrix $\alpha \in \mathbb{N}^{n \times p}$, we define $|\alpha| = \sum_{i=1}^n \sum_{j=1}^p \alpha_{ij}$ and $\alpha! = \prod_{i=1}^n \prod_{j=1}^p \alpha_{ij}!$. Given another matrix $\beta \in \mathbb{N}^{n \times p}$, we write $\binom{\alpha}{\beta} = \prod_{i=1}^n \prod_{j=1}^p \binom{\alpha_{ij}}{\beta_{ij}}$. Finally, given a matrix $Q \in \mathbb{R}^{n \times p}$, we write $Q^\alpha = \prod_{i=1}^n \prod_{j=1}^p Q_{ij}^{\alpha_{ij}}$. We define $X \in \mathbb{R}^{n \times p}$ the signal matrix whose i-th row is the vector μ_{k_i}, so that, conditionally on X, the Y_{ij} are independent with $\mathcal{N}(X_{ij}, 1)$ distribution. Throughout this proof, we write $X_i = \mu_{k_i}$.

Proposition 1. *[26] The degree D maximum correlation satisfies the upper-bound*

$$corr^2_{\leq D} \leq \sum_{\substack{\alpha \in \mathbb{N}^{n \times p} \\ |\alpha| \leq D}} \frac{\kappa_\alpha^2}{\alpha!} \ , \tag{15}$$

where κ_α for $\alpha \in \mathbb{N}^{n \times p}$ is defined recursively by

$$\kappa_\alpha = \mathbb{E}[x X^\alpha] - \sum_{\beta \lneq \alpha} \mathbb{E}[X^{\alpha - \beta}] \binom{\alpha}{\beta} \kappa_\beta \ . \tag{16}$$

[26] observe that, for $\alpha \in \mathbb{N}^{n\times p}$, the quantity κ_α corresponds to the cumulant $\kappa(x, X_{a_1}, \dots, X_{a_m})$, where $\{a_1, \dots, a_m\}$ is the multiset that contains α_{ij} copies of (i,j), for $i,j \in [1,n]\times[1,p]$. We refer e.g. to the lecture notes [34] for more details on cumulants. In the remainder of the proof, we first characterize the matrices α for which $\kappa_\alpha \neq 0$, and we provide an upper bound on the corresponding cumulants.

Let us first provide sufficient conditions on α, so that the corresponding cumulant κ_α is zero. For that purpose, it is convenient to represent $\alpha \in \mathbb{N}^{n\times p}$ as a bipartite multi-graph. More precisely, we define $\mathcal{G}_\alpha$ as the bipartite multi-graph on two disjoint sets of nodes $U = \{u_i, i \in [1,n]\}$ and $V = \{v_j, j \in [1,p]\}$, with α_{ij} edges between u_i and v_j, for $i,j \in [1,n]\times[1,p]$. For a given $\alpha \in \mathbb{N}^{n\times p}$, the ℓ^1-norm $|\alpha| = \sum_{i=1}^n \sum_{j=1}^p \alpha_{ij}$ corresponds to the number of multi-edges of $\mathcal{G}_\alpha$. Besides, we also denote by $\mathcal{G}_\alpha^-$ the graph $\mathcal{G}_\alpha$ from which we have removed isolated nodes, and we write l_α for the number of nodes of $\mathcal{G}_\alpha^-$. We define m_α (resp. r_α) as the number of nodes of $\mathcal{G}_\alpha^- \cap U$ (resp. $\mathcal{G}_\alpha^- \cap V$), so that $l_\alpha = m_\alpha + r_\alpha$. The following lemma is proved in Section A.2.

Lemma 1. *Let $\alpha \in \mathbb{N}^{n\times p}$ be non-zero such that $\kappa_\alpha \neq 0$. Then, $\mathcal{G}_\alpha^-$ is connected and contains both u_1 and u_2. Moreover, all the nodes v_j of $\mathcal{G}_\alpha^-$ are connected to at least two distinct nodes of $\mathcal{G}_\alpha^-$.*

This lemma is proved in Section A.1. The proof mostly relies on the property that a cumulant of two sets of independents random variable is zero. As a corollary, we deduce the following properties for matrices α, such that κ_α is non-zero.

Lemma 2. *Let $\alpha \in \mathbb{N}^{n\times p}$ be non-zero. If $\kappa_\alpha \neq 0$, then $m_\alpha \geq 2$, $|\alpha| \geq 2r_\alpha$ and $|\alpha| \geq r_\alpha + m_\alpha - 1$.*

Proof of Lemma 2. The first property ($m_\alpha \geq 2$) holds because u_1 and u_2 are spanned by $\mathcal{G}_\alpha^-$. The last property ($|\alpha| \geq r_\alpha + m_\alpha - 1$) holds because the multigraph $\mathcal{G}_\alpha^-$ is connected. Finally, we know that each node of $\mathcal{G}_\alpha^-$ that is also in V has degree at least 2. Since the graph $\mathcal{G}_\alpha^-$ is bipartite, all these edges are distinct and we deduce that $|\alpha| \geq 2r_\alpha$. □

In order to upper-bound $corr^2_{\leq D}$, we need to upper-bound the cumulants κ_α for all $\alpha \in \mathbb{N}^{n\times p}$ that satisfy the conditions of Lemma 1.

Lemma 3. *Let $\alpha \in \mathbb{N}^{n\times p}$ be such that $\kappa_\alpha \neq 0$. We have*

$$|\kappa_\alpha| \leq \varepsilon^{|\alpha|} (1+|\alpha|)^{|\alpha|} \min\left(\frac{1}{K}, |\alpha|^{|\alpha|} \left(\frac{1}{K}\right)^{l_\alpha - \frac{|\alpha|}{2} - 1}\right) . \tag{17}$$

Now we gather the three above lemmas to control the sum $\sum_{\alpha:\ |\alpha|\leq D} \kappa_\alpha^2/\alpha!$ in (15). We reorganize the sum over α by organizing it according to the values of m_α, r_α, and $|\alpha|$, which respectively correspond to the number of u-nodes, v-nodes and edges in $\mathcal{G}_\alpha^-$.

Lemma 4. *Given $m \geq 2$, $r \geq 1$, $d \geq \max(r+m-1, 2r)$, there exists at most $p^r n^{m-2} d^{2d}$ matrices $\alpha \in \mathbb{N}^{n\times p}$ such $\kappa_\alpha \neq 0$, $m_\alpha = m$, $r_\alpha = r$ and $|\alpha| = d$.*

Proof of Lemma 4. Since $\kappa_\alpha \neq 0$, Lemma 1 ensures that both u_1 and u_2 are nodes of $\mathcal{G}_\alpha^-$. Hence, there are less than $n^{m-2}p^r$ possibilities for choosing the remaining nodes. By assumption $d \geq \max(m,r)$. For each edge, there are at most $mr \leq d^2$ possibilities. Since $\mathcal{G}_\alpha^-$ has d edges, we have less than d^{2d} possibilities for choosing these edges. Since $\mathcal{G}_\alpha^-$ is one to one with α, we conclude that there are less than $p^r n^{m-2} d^{2d}$ matrices α satisfying the given constraints. □

Combining the bounds on the cumulants of Lemma 3 and Lemma 4, we are in position to control $corr^2_D$. We slice the sum of the cumulants according to r_α, m_α and d_α, and we use below the

notation $\mathcal{D}_d = \{(r,m) \in [1,d] \times [2,d] : \max(m+r-1, 2r) \leq d\}$. We recall that X_{ij} takes value in $\{-\varepsilon, +\varepsilon\}$, with $\varepsilon = \bar{\Delta}/\sqrt{p}$, so that

$$\begin{aligned}
corr_D^2 \leq & \sum_{\substack{\alpha \in \mathbb{N}^{n\times p} \\ |\alpha| \leq D}} \kappa_\alpha^2 \\
\leq & \kappa_0^2 + \sum_{d=1}^{D} \sum_{(r,m)\in\mathcal{D}_d} p^r n^{m-2} d^{2d} \varepsilon^{2d} (1+d)^{2d} \min\left(\frac{1}{K^2}, d^{2d}\left(\frac{1}{K}\right)^{2m+2r-d-2}\right) \\
\leq & \frac{1}{K^2} + \sum_{d=1}^{D} \sum_{(r,m)\in\mathcal{D}_d} p^r n^{m-2} (\varepsilon^2 D^4 (1+D)^2)^d \min\left(\frac{1}{K^2}, \left(\frac{1}{K}\right)^{2m+2r-d-2}\right). \qquad (18)
\end{aligned}$$

Since $\zeta_n = \frac{\bar{\Delta}^4 D^8 (1+D)^4}{p} \max\left(\frac{n}{K^2}, 1\right)$, with $\bar{\Delta}^2 = p\epsilon^2$, we get

$$\begin{aligned}
& corr_D^2 - \frac{1}{K^2} \\
\leq & \sum_{d=1}^{D} \sum_{(r,m)\in\mathcal{D}_d} \zeta_n^{d/2} p^{-(d/2-r)} n^{m-2} \left(\frac{1}{\max(1, n/K^2)}\right)^{d/2} \min\left(\frac{1}{K^2}, \left(\frac{1}{K}\right)^{2m+2r-d-2}\right) \\
\leq & \frac{1}{K^2} \sum_{d=1}^{D} \sum_{(r,m)\in\mathcal{D}_d} \zeta_n^{d/2} n^{r+m-d/2-2} \left(\frac{1}{\max(1, n/K^2)}\right)^{d/2} \min\left(1, \left(\frac{1}{K}\right)^{2m+2r-d-4}\right), \qquad (19)
\end{aligned}$$

where we used in the last line that $n \leq p$ and $r \leq d/2$. Let us check that each term in the sum is upper-bounded by $\zeta_n^{d/2}$, by considering apart the cases $d/2 \geq m+r-2$ and $d/2 < m+r-2$.

When $d/2 \geq m+r-2$, the exponent of n is non-positive, so that

$$n^{r+m-d/2-2} \left(\frac{1}{\max(1, n/K^2)}\right)^{d/2} \min\left(1, \left(\frac{1}{K}\right)^{2m+2r-d-4}\right) \leq 1 \ .$$

When $d/2 < m+r-2$, we can upper bound the minimum by $K^{-(2m+2r-d-4)}$, so that

$$\begin{aligned}
& n^{r+m-d/2-2} \left(\frac{1}{\max(1, n/K^2)}\right)^{d/2} \min\left(1, \left(\frac{1}{K}\right)^{2m+2r-d-4}\right) \\
& \qquad \leq \left(\frac{n}{K^2}\right)^{m+r-d/2-2} \left(\frac{1}{\max(1, n/K^2)}\right)^{d/2} .
\end{aligned}$$

If $n \leq K^2$, the latter expression is smaller or equal to one. If $n \geq K^2$, this last expression, is equal to $(n/K^2)^{m+r-d-2}$ and is also smaller or equal to one since $d \geq m+r-1$ for $(r,m) \in \mathcal{D}_d$. Back to (19), and relying on the assumption $\zeta_n < 1$, we conclude that

$$\begin{aligned}
corr_D^2 & \leq \frac{1}{K^2} + \frac{1}{K^2} \sum_{d=1}^{D} \sum_{(r,m)\in\mathcal{D}_d} \zeta_n^{d/2} \\
& \leq \frac{1}{K^2} \left[1 + \sum_{d=2}^{D} \frac{d(d-1)}{2} \zeta_n^{d/2}\right] \\
& \leq \frac{1}{K^2} \left[1 + \frac{\zeta_n}{(1-\sqrt{\zeta_n})^3}\right] \quad . \qquad (20)
\end{aligned}$$

A.1 Proof of Lemma 1

In order to prove Lemma 1, we will use a classical property of cumulants, that we recall here.

Lemma 5. *[e.g. [34]] Let $X_1, \dots, X_r$ be random variables on the same space Ω. If there exists a partition A, B of $[1, r]$ such that $(X_i)_{i\in A}$ is independant from $(X_i)_{i\in B}$, then the cumulant $\kappa(X_1, \dots, X_r)$ is zero.*

We prove below that $\kappa_\alpha = 0$ if one of the three following properties is satisfied:

(i) u_1 or u_2 are not spanned by $\mathcal{G}_\alpha^-$,

(ii) a node of $\mathcal{G}_\alpha^-$ which also belong to V is connected to at most one node in U,

(iii) $\mathcal{G}_\alpha^-$ is not connected.

We denote by $U_\alpha \subset U$ (resp. $V_\alpha \subset V$) the set of nodes of $\mathcal{G}_\alpha^-$ that also belong to U (resp. V). We denote by E_α the set of edges of $\mathcal{G}_\alpha^-$.

Let us first show that (i) is a sufficient condition for $\kappa_\alpha = 0$. By symmetry, we suppose that u_1 is not spanned by $\mathcal{G}_\alpha^-$. Then, k_1 (the group corresponding to u_1) is independent from the family of random variables $(X_{ij})_{u_i,v_j\in E_\alpha}$. Hence, the random variable $x = \mathbf{1}_{k_1=k_2}$ is also independent from $(X_{ij})_{u_i,v_j\in E_\alpha}$. Together with Lemma 5, this implies the nullity of the cumulant κ_α.

Then, let us show that (ii) is also a sufficient condition for $\kappa_\alpha = 0$. We suppose that there exists $j_0 \in [1, p]$ such that v_{j_0} is connected with only one node $u_{i_0} \in U$. Conditionally on $(k_i)_{i\in[1,K]}$ and on $(\mu_{k,j})_{k,j\in[1,k]\times([1,p]\setminus\{j_0\})}$, the variable X_{i_0,j_0} is uniformly distributed on $\{-\varepsilon, \varepsilon\}$. This implies the independence of X_{i_0,j_0} with $((X_{ij})_{u_i,v_j\in E_\alpha\setminus(i_0,j_0)}, x)$. Lemma 5 then leads to the nullity of the cumulant κ_α.

Finally, let us show that (iii) is a sufficient condition for $\kappa_\alpha = 0$. Let $\alpha \in \mathbb{N}^{n\times p}$ such that $\mathcal{G}_\alpha^-$ has at least two connected components. Let us denote C_1 and C_2 two of these connected components. At least one of them does not contain both u_1 and u_2. We suppose by symmetry that C_1 does not contain both these nodes (we suppose that it does not contain u_2 for example). We denote $E_1 = E_\alpha \cap ((U \cap C_1) \times (V \cap C_1))$ which corresponds to the edges of $\mathcal{G}_\alpha^-$ which connect points from $U_\alpha \cap C_1$ to points from $V_\alpha \cap C_1$. We will show that the families of random variables $(X_{ij})_{u_i,v_j\in E_1}$ and $(x, (X_{ij})_{u_i,v_j\in E_\alpha\setminus E_1})$ are independent, which will lead to the nullity of κ_α, using Lemma 5.

For sake of clarity, we begin by dealing with the simple case where C_1 also does not contain u_1. So, the intersection of C_1 with $\{u_1, u_2\}$ is empty. For $u_i, v_j \in E_\alpha$, $X_{ij} = \mu_{k_i,j}$. By definition of our model, the family $\left((k_i)_{u_i\in C_1\cap U}, (\mu_{k,j})_{k\in[1,K],v_j\in C_1\cap V}\right)$ is independent from the family $\left((k_i)_{u_i\in U\setminus C_1}, (\mu_{k,j})_{k\in[1,K],v_j\in V\setminus C_1}\right)$. On the one hand, the random variables $(X_{ij})_{u_i,v_j\in E_1}$ are measurable with respect to $(k_i)_{u_i\in C_1\cap U}$ and $(\mu_{k,j})_{k\in[1,K],v_j\in C_1\cap V}$. On the other hand, the random variables $(x = \mathbf{1}_{k_1=k_2}, (X_{ij})_{u_i,v_j\in E_\alpha\setminus E_1})$ are measurable with respect to $(k_i)_{u_i\in U\setminus C_1}$ and $(\mu_{k,j})_{k\in[1,K],v_j\in V\setminus C_1}$. This leads to the independence of $(X_{ij})_{u_i,v_j\in E_1}$ with $(x, (X_{ij})_{u_i,v_j\in E_\alpha\setminus E_1})$. Lemma 5 implies the nullity of κ_α.

Now, let us deal with the more complex case where C_1 contains u_1. Again, the random variables x and $(X_{ij})_{u_i,v_j\in E_\alpha\setminus E_1}$ are measurable with respect to $(k_i)_{u_i\in U\setminus C_1}$ and $(\mu_{k,j})_{k\in[1,K],v_j\in V\setminus C_1}$. The difference with the previous case lies in the fact that, since $u_1 \in C_1$, we lose the independence of $(k_i)_{i\in C_1}$ with $(x = \mathbf{1}_{k_1=k_2}, (X_{ij})_{u_i,v_j\in E_\alpha\setminus E_1})$. Instead, we will show the independence of the partition induced by the k_i's on $C_1 \cap U$ with $(x = \mathbf{1}_{k_1=k_2}, (X_{ij})_{u_i,v_j\in E_\alpha\setminus E_1})$. We denote $\hat{G}$ this partition. Two nodes u_i and $u_{i'}$ of $C_1 \cap U$ are in the same group of $\hat{G}$ if and only if $k_i = k_{i'}$.

Let $(u_i, v_j) \in E_1$. We denote $A \in \hat{G}$ the group of $\hat{G}$ containing u_i. Then, $X_{ij} = \frac{1}{|A|}\sum_{i'\in A} \mu_{k_{i'},j}$. So, the family $(X_{ij})_{u_i,v_j\in E_1}$ is entirely defined by $\hat{G}$ and by the centers of this partition for coordinates $j \in [1, p] \cap C_1$. For $A \in \hat{G}$, we denote $\mu_{A,j} = \frac{1}{|A|}\sum_{u_i\in A} \mu_{k_i,j}$, which is the j-th coordinate of the center of the group A.

The family $(x, (X_{ij})_{u_i,v_j \in E_\alpha \setminus E_1})$ is measurable with respect to the family

$$\mathcal{X}_1 := \left((k_i)_{u_i \in U \setminus C_1}, (\mu_{k,j})_{k,u_j \in [1,K] \times (V \setminus C_1)}, \mathbf{1}_{k_1 = k_2}\right) \ .$$

Since the intersection of $U \cap C_1$ with $(U \setminus C_1) \cup \{u_1, u_2\}$ contains only u_1, it is clear that the family $\left(\mathbf{1}_{k_i = k_{i'}}\right)_{(u_i,u_{i'}) \in (U \cap C_1)^2}$ is independent from $\mathcal{X}_1$. Then, $\hat{G}$ is independent from $\mathcal{X}_1$.

We now condition to the k_i's for $i \in [1,n]$, and to $\mathcal{X}_1$. Then, $\hat{G}$ is fixed. Since the $\mu_{k,j}$'s are drawn independently, we deduce that, with our conditioning, the $\mu_{k,j}$'s, for $j \in V \cap C_1$, are still drawn independently and uniformly on $\{-\varepsilon, \varepsilon\}$. For $A \in \hat{G}$, there exists $k_A \in [1,K]$ which satisfies; for all $j \in V \cap C_1$ $\mu_{A,j} = \mu_{k_A,j}$. The application $A \to k_A$ being an injection, we deduce that the $\mu_{k_A,j}$'s, for $A \in \hat{G}$ and $j \in V \cap C_1$ are independent and uniformly drawn from $\{-\varepsilon, +\varepsilon\}$.

Let us summarize this; the partition $\hat{G}$ is independent from $\mathcal{X}_1$, and conditionally on $\hat{G}$ and $\mathcal{X}_1$, the $\mu_{A,j}$'s, for $A \in \hat{G}$ and $v_j \in V \cap C_1$, are independently and uniformly drawn from $\{-\varepsilon, +\varepsilon\}$. Together with the fact that the family $(X_{ij})_{u_i,v_j \in E_1}$ is measurable with respect to $\hat{G}$ and the $\mu_{A,j}$'s, for $A \in \hat{G}$ and $v_j \in V \cap C_1$, this leads to the independence of $(X_{ij})_{u_i,v_j \in E_1}$ with $\mathcal{X}_1$.

Since $(x, (X_{ij})_{u_i,v_j \in E_\alpha \setminus E_1})$ is measurable with respect to $\mathcal{X}_1$, we deduce the independence of the families $(X_{ij})_{u_i,v_j \in E_1}$ and $(x, (X_{ij})_{u_i,v_j \in E_\alpha \setminus E_1})$. Hence, Lemma 5 leads to the nullity of the cumulant κ_α. This concludes our proof.

A.2 Proof of Lemma 3

First, we provide an upper-bound on the moments of the form $\mathbb{E}[X^\gamma]$ and $\mathbb{E}[xX^\gamma]$.

Lemma 6. *Let $\gamma \in \mathbb{N}^{n \times p}$. We denote by CC_γ the number of connected components of $\mathcal{G}_\gamma$ and by l_γ the number of nodes of $U \cup V$ spanned by $\mathcal{G}_\gamma$. Then, we both have*

$$\mathbb{E}[X^\gamma] \leq \varepsilon^{|\gamma|} \min\left(1, |\gamma|^{|\gamma|} \left(\frac{1}{K}\right)^{l_\gamma - \frac{|\gamma|}{2} - CC_\gamma}\right) \ ,$$

and

$$\mathbb{E}[xX^\gamma] \leq \varepsilon^{|\gamma|} \min\left(\frac{1}{K}, |\gamma|^{|\gamma|} \left(\frac{1}{K}\right)^{l_\gamma - \frac{|\gamma|}{2} - CC_\gamma}\right) \ .$$

This lemma, which is the core of our arguments, is shown in the next subsection. Here, we deduce the upper-bound (17) on the cumulant from this lemma.

We proceed by induction on $\alpha \in \mathbb{N}^{n \times p}$. For the initialization , we have $\kappa_0 = \mathbb{E}[x] = \frac{1}{K}$. Then, we take $\alpha \in \mathbb{N}^{n \times p}$, and we suppose that, for all $\beta \lneq \alpha$, we have

$$|\kappa_\beta| \leq \varepsilon^{|\beta|} (1 + |\beta|)^{|\beta|} \min\left(\frac{1}{K}, |\beta|^{|\beta|} \left(\frac{1}{K}\right)^{l_\beta - \frac{|\beta|}{2} - 1}\right) \ .$$

From Lemma 1, we can suppose that $\mathcal{G}_\alpha$ only has one connected component, otherwise $\kappa_\alpha = 0$. We recall that, given $\gamma \in \mathbb{N}^{n \times p}$, l_γ and CC_γ respectively stand for the number of nodes spanned by $\mathcal{G}_\gamma$, and the number of connected components of $\mathcal{G}_\gamma$. The Definition (16) of κ_α, Lemma 6, and

the induction hypothesis imply that

$$
\begin{aligned}
|\kappa_\alpha| \leq & \mathbb{E}[xX^\alpha] + \sum_{0<\beta\lneq\alpha} \binom{\alpha}{\beta} \mathbb{E}[X^{\alpha-\beta}]|\kappa_\beta| + |\kappa_0|\,\mathbb{E}[X^\alpha] \\
\leq & \varepsilon^{|\alpha|}|\alpha|^\alpha \left(\frac{1}{K}\right)^{l_\alpha - \frac{|\alpha|}{2}-1} \\
& + \sum_{0<\beta\lneq\alpha} \binom{\alpha}{\beta} |\varepsilon|^{|\beta|+|\alpha-\beta|}(1+|\beta|)^{|\beta|}|\beta|^{|\beta|}|\alpha-\beta|^{|\alpha-\beta|} \left(\frac{1}{K}\right)^{l_\beta + l_{\alpha-\beta} - \frac{|\beta|}{2} - \frac{|\alpha-\beta|}{2} - 1 - CC_{\alpha-\beta}} \\
& + \frac{1}{K}\varepsilon^{|\alpha|}|\alpha|^\alpha \left(\frac{1}{K}\right)^{l_\alpha - \frac{|\alpha|}{2}-1} \\
\leq & 2\varepsilon^{|\alpha|}|\alpha|^\alpha \left(\frac{1}{K}\right)^{l_\alpha - \frac{|\alpha|}{2}-1} \\
& + \varepsilon^{|\alpha|}|\alpha|^\alpha \left(\frac{1}{K}\right)^{-\frac{|\alpha|}{2}} \sum_{0<\beta\lneq\alpha} \binom{\alpha}{\beta} (1+|\beta|)^{|\beta|} \left(\frac{1}{K}\right)^{l_\beta + l_{\alpha-\beta} - 1 - CC_{\alpha-\beta}} \quad . \qquad (21)
\end{aligned}
$$

Claim: For any $0 < \beta \lneq \alpha$, we have $l_\beta + l_{\alpha-\beta} - C_{\alpha-\beta} \geq l_\alpha$.

We first show this claim. We have supposed that $\mathcal{G}_\alpha$ only has one connected component. We denote $C_1, \ldots, C_{CC_{\alpha-\beta}} \subset U \cup V$ the connected components of $\mathcal{G}_{\alpha-\beta}$. For all $s \in [1, CC_{\alpha-\beta}]$, there exists $x \in C_s$ which is spanned by $\mathcal{G}_\beta$. Indeed, otherwise, since the set of edges of $\mathcal{G}_\alpha$ is the union of the edges of $\mathcal{G}_\beta$ and $\mathcal{G}_{\alpha-\beta}$, C_s would also be a connected component of $\mathcal{G}_\alpha$ which does not span the nodes spanned by $\mathcal{G}_\beta$. This contradicts the connectivity of $\mathcal{G}_\alpha$. So, there exist at least $CC_{\alpha-\beta}$ distinct points of $U \cup V$ which are spanned both by $\mathcal{G}_{\alpha-\beta}$ and $\mathcal{G}_\beta$. Since the nodes spanned by $\mathcal{G}_\alpha$ are spanned by $\mathcal{G}_{\alpha-\beta}$ or $\mathcal{G}_\beta$, this leads to $l_\beta + l_{\alpha-\beta} - CC_{\alpha-\beta} \geq l_\alpha$.

Now that we proved the claim, we plug it in (21). This leads to

$$
\begin{aligned}
|\kappa_\alpha| \leq & \varepsilon^{|\alpha|}|\alpha|^{|\alpha|} \left(\frac{1}{K}\right)^{l_\alpha - \frac{|\alpha|}{2}-1} \left(2 + \sum_{0<\beta\lneq\alpha} \binom{\alpha}{\beta} (1+|\beta|)^{|\beta|}\right) \\
\leq & \varepsilon^{|\alpha|}|\alpha|^{|\alpha|} \left(\frac{1}{K}\right)^{l_\alpha - \frac{|\alpha|}{2}-1} \left(2 + \sum_{0<\beta\lneq\alpha} \binom{\alpha}{\beta} |\alpha|^{|\beta|}\right) \\
\leq & \varepsilon^{|\alpha|}|\alpha|^{|\alpha|} \left(\frac{1}{K}\right)^{l_\alpha - \frac{|\alpha|}{2}-1} \left(2 + \sum_{w=1}^{|\alpha|-1} \binom{|\alpha|}{w} |\alpha|^w\right) \\
\leq & \varepsilon^{|\alpha|}|\alpha|^{|\alpha|}(1+|\alpha|)^{|\alpha|} \left(\frac{1}{K}\right)^{l_\alpha - \frac{|\alpha|}{2}-1} \quad .
\end{aligned}
$$

This inequality proves the first part of the sought upper-bound of Lemma 3.

It remains to prove that $|\kappa_\alpha| \leq \varepsilon^{|\alpha|}(1+|\alpha|)^{|\alpha|}\frac{1}{K}$. For all $\beta \lneq \alpha$, we know from Lemma 6 that

$\mathbb{E}[X^\beta] \leq \varepsilon^{|\beta|}$ and $\mathbb{E}[xX^\alpha] \leq \frac{1}{K}\varepsilon^{|\alpha|}$. Together with the induction hypothesis, this leads to

$$\begin{aligned}
|\kappa_\alpha| \leq & \varepsilon^{|\alpha|}\frac{1}{K} + \sum_{\beta \lneq \alpha} \binom{\alpha}{\beta}(1+|\beta|)^{|\beta|}\frac{1}{K}\varepsilon^{|\beta|}\varepsilon^{|\alpha-\beta|} \\
\leq & \varepsilon^{|\alpha|}\frac{1}{K}\left(1 + \sum_{\beta \lneq \alpha}\binom{\alpha}{\beta}|\alpha|^{|\beta|}\right) \\
\leq & \varepsilon^{|\alpha|}(1+|\alpha|)^{|\alpha|}\frac{1}{K} \ .
\end{aligned}$$

This concludes the induction; for all $\alpha \in \mathbb{N}^{n\times p}$, we have

$$|\kappa_\alpha| \leq \varepsilon^{|\alpha|}(1+|\alpha|)^{|\alpha|}\min\left(\frac{1}{K}, |\alpha|^{|\alpha|}\left(\frac{1}{K}\right)^{l_\alpha - \frac{|\alpha|}{2}-1}\right) \ .$$

A.3 Proof of Lemma 6

Let $\gamma \in \mathbb{N}^{n\times p}$ such that $\mathcal{G}_\gamma$ has CC_γ connected components and spans l_γ nodes. Let us first upper bound $\mathbb{E}[X^\gamma]$. We denote m_γ the number of nodes of U spanned by $\mathcal{G}_\gamma$ and r_γ the number of nodes of V spanned by $\mathcal{G}_\gamma$. We suppose by symmetry that γ is supported on $[1, m_\gamma] \times [1, r_\gamma]$. We denote $\hat{G}^\gamma$ the partition induced on $[1, m_\gamma]$ by $k_1, \ldots, k_{m_\gamma}$.

Lemma 7. *If, for all $j \in [1, r_\gamma]$ and for all groups $A \in \hat{G}^\gamma \subset [1, m_\gamma]$, we have $\sum_{i\in A}\gamma_{ij} \equiv 0$ [2], then, $\mathbb{E}[X^\gamma|\hat{G}^\gamma] = \varepsilon^{|\gamma|}$. Otherwise, we have $\mathbb{E}[X^\gamma|\hat{G}^\gamma] = 0$.*

This lemma, proved in Section A.4, implies

$$\mathbb{E}[X^\gamma] = \varepsilon^{|\gamma|}\,\mathbb{P}\left[\forall j \in [1, r_\gamma],\ \forall A \in \hat{G}^\gamma,\ \sum_{i\in A}\gamma_{ij} \equiv 0\ [2]\right] \ . \tag{22}$$

Let G be a partition of $[1, m_\gamma]$ with $|G| \leq K$, and let us upper-bound $\mathbb{P}[\hat{G}^\gamma = G]$. We write $G = \{G_1, \ldots, G_{|G|}\}$. We take, for $k \in [1, |G|]$, $i_k \in G_k$. Then, $\hat{G}^\gamma = G$ implies that, for all $k \in [1, |G|]$, for all $i \in G_k$, the equality $k_i = k_{i_k}$ holds. And,

$$\begin{aligned}
\mathbb{P}\left[\forall k \in [1, |G|],\ \forall i \in G_k \setminus \{i_k\},\ k_i = k_{i_k}\right] &= \prod_{k\in[1,|G|]} \mathbb{P}\left[\forall i \in G_k \setminus \{i_k\},\ k_i = k_{i_k}\right] \\
&= \prod_{k\in[1,|G|]}\left(\frac{1}{K}\right)^{|G_k|-1} \\
&= \left(\frac{1}{K}\right)^{m_\gamma - |G|} \ .
\end{aligned}$$

This equality leads to

$$\mathbb{P}[\hat{G}^\gamma = G] \leq \left(\frac{1}{K}\right)^{m_\gamma - |G|} \ .$$

The next lemma upper-bounds the number of groups of a partition G of $[1, m_\gamma]$ such that, for all $j \in [1, r_\gamma]$ and all groups $A \in G$, we have $\sum_{i\in A}\gamma_{ij} \equiv 0$ [2]. Its proof, given in Section A.5, relies on delicate combinatorial arguments.

Lemma 8. *Let G be a partition of $[1, m_\gamma]$ satisfying, for all $j \in [1, r_\gamma]$, all groups $A \in G$, the equality $\sum_{i\in A}\gamma_{ij} \equiv 0$ [2]. Then, the number of groups satisfies the following inequality*

$$|G| \leq \frac{|\gamma|}{2} - r_\gamma + CC_\gamma \ .$$

Applying this lemma together with the fact that there are at most $m_\gamma^{m_\gamma}$ partitions of $[1, m_\gamma]$ leads to

$$\mathbb{P}\left[\forall j \in [1, r_\gamma],\ \forall A \in \hat{G}^\gamma,\ \sum_{i \in A} \gamma_{ij} \equiv 0\ [2]\right] \leq m_\gamma^{m_\gamma} \left(\frac{1}{K}\right)^{m_\gamma + r_\gamma - CC_\gamma - \frac{|\gamma|}{2}} .$$

We plug this inequality in (22) and get, using $m_\gamma \leq |\gamma|$,

$$\mathbb{E}[X^\gamma] \leq \varepsilon^{|\gamma|} |\gamma|^{|\gamma|} \left(\frac{1}{K}\right)^{m_\gamma + r_\gamma - CC_\gamma - \frac{|\gamma|}{2}} .$$

Moreover, since $\mathbb{P}\left[\forall j \in [1, r_\gamma],\ \forall A \in \hat{G}^\gamma,\ \sum_{i \in A} \gamma_{ij} \equiv 0\ [2]\right] \leq 1$, we get

$$\mathbb{E}[X^\gamma] \leq \varepsilon^{|\gamma|} \min\left(1,\ |\gamma|^{|\gamma|} \left(\frac{1}{K}\right)^{m_\gamma + r_\gamma - CC_\gamma - \frac{|\gamma|}{2}}\right) .$$

Now, let us upper bound $\mathbb{E}[xX^\gamma]$. Since the random variables x and $|X^\gamma|/\epsilon^{|\gamma|}$ belong to $[0, 1]$ almost surely, we obtain

$$\mathbb{E}[xX^\gamma] \leq \min(\epsilon^{|\gamma|}\mathbb{E}[x], \mathbb{E}[X^\gamma]) . \tag{23}$$

Then, since $\mathbb{E}[x] = 1/K$, we can deduce the desired bound from the previous case.

A.4 Proof of Lemma 7

We suppose first that, for all $j \in [1, r_\gamma]$ and for all groups $A \in \hat{G}^\gamma$, we have $\sum_{i \in A} \gamma_{ij} \equiv 0\ [2]$. Consider a specific (i, j). If $k_i = k$, we have $X_{ij} = \mu_{k,j}$. This implies that

$$X^\gamma = \prod_{k,j \in [1,K] \times [1, r_\gamma]} \mu_{k,j}^{\sum_{i, k_i = k} \gamma_{i,j}} .$$

Moreover, by hypothesis, conditionally on $\hat{G}^\gamma$ for $k \in [1, K]$ and $j \in [1, r_\gamma]$, $\sum_{i, k_i = k} \gamma_{i,j} \equiv 0\ [2]$ and $|\mu_{k,j}| = \varepsilon$. Hence, we have

$$X^\gamma = |\varepsilon|^{\sum \gamma_{ij}} = |\varepsilon|^{|\gamma|} .$$

This leads to the sought equality

$$\mathbb{E}[X^\gamma | \hat{G}^\gamma] = |\varepsilon|^{\sum \gamma_{ij}} = |\varepsilon|^{|\gamma|} .$$

Now, let us suppose that there exists $j \in [1, r_\gamma]$, a group $A \in \hat{G}^\gamma$, such that $\sum_{i \in A} \gamma_{ij} \equiv 1\ [2]$. We have as before

$$X^\gamma = \prod_{k,j \in [1,K] \times [1, r_\gamma]} \mu_{k,j}^{\sum_{i, k_i = k} \gamma_{i,j}} .$$

By independence of the $\mu_{k,j}$'s, for $k \in [1, K]$ and $j \in [1, p]$, both between themselves and with the k_i's, for $i \in [1, n]$, we deduce that, conditionally on the k_i's, for such k_i's that induce $\hat{G}^\gamma$, that

$$\mathbb{E}[X^\gamma | (k_i)_{i \in [1,n]}] = \prod_{k,j \in [1,K] \times [1, r_\gamma]} \mathbb{E}[\mu_{k,j}^{\sum_{i, k_i = k} \gamma_{i,j}} | (k_i)_{i \in [1,n]}] .$$

Let us denote $k' \in [1, K]$ and $j' \in [1, r_\gamma]$ that satisfies; $\sum_{i,k_i=k'} \gamma_{i,j'} \equiv 1 \ [2]$. This, together with the fact that the probability distribution of $\mu_{k',j'}$ is symmetric, leads to $\mathbb{E}\left[\mu_{k',j'}^{\sum_{i,k_i=k'} \gamma_{i,j'}} | (k_i)_{i\in[1,n]}\right] = 0$. This implies $\mathbb{E}\left[X^\gamma | (k_i)_{i\in[1,n]}\right] = 0$. Thus, we get the sought equality

$$\mathbb{E}[X^\gamma | \hat{G}^\gamma] = 0 \ .$$

A.5 Proof of Lemma 8

In this proof, given a partition G of a subset of $[1, n]$, and given $\gamma \in \mathbb{N}^{n\times p}$, we say γ is even with respect to G if the following holds: $\sum_{i\in A} \gamma_{ij} \equiv 0 \ [2]$ for all $A \in G$ and $j \in [1, p]$. We recall that m_γ, r_γ, and CC_γ respectively stand for the number of nodes in $U \cap \mathcal{G}_\gamma^-$, the number of nodes in $V \cap \mathcal{G}_\gamma^-$, and the number of connected components of the graph $\mathcal{G}_\gamma^-$. We prove in this section the following claim, which rephrases Lemma 8. For $\gamma \in \mathbb{N}^{n\times p}$ and a partition G of $\{i \in [1, n], \ \exists j \in [1, p], \ \gamma_{ij} > 0\}$, if G is even with respect to γ, then

$$|G| \leq \frac{|\gamma|}{2} - r_\gamma + CC_\gamma \ . \tag{24}$$

For $\beta \leq \gamma$ a restriction of γ to one of the connected component of $\mathcal{G}_\gamma^-$, and for $j \in [1, p]$, the vector $\beta_{:j}$ is either null or equal to $\gamma_{:j}$. So, if each restriction of γ to a connected component of $\mathcal{G}_\gamma^-$ is even with respect to a partition G, then γ is also even with respect to G. Hence, it is sufficient to prove this lemma when $\mathcal{G}_\gamma^-$ is connected. We proceed with by induction on $r_\gamma > 0$.

Initialization. If $r_\gamma = 1$; we suppose by symmetry that γ is supported on $[1, m_\gamma] \times \{1\}$. Again, we proceed by induction. If $|\gamma| = 1$, then there exists no partition that satisfy the conditions of Lemma 8 and so the proposition is true. We suppose that $|\gamma| > 1$ and that the proposition is true for all $\beta \lneq \gamma$. We distinguish two cases.

First, we consider the case where, for all $i \in [1, m_\gamma]$, $\gamma_{i1} \equiv 1 \ [2]$. Let G a partition of $[1, m_\gamma]$ for which γ is even. Then, each group of G must have an even number of elements. So, each group of G is of cardinality at least 2. Hence, there are at most $\frac{m_\gamma}{2}$ groups in the partition G. Moreover, $|\gamma| \geq m_\gamma$. Thus, $|G| \leq \frac{|\gamma|}{2}$.

Now, let us consider the case where there exists $i_0 \in [1, m_\gamma]$ such that $\gamma_{i_0 1} \equiv 0 \ [2]$. We define γ' by $\gamma'_{ij} = \gamma_{ij}\mathbf{1}_{i\neq i_0}$. Let G a partition of $[1, m_\gamma]$ for which γ is even. We define G' the partition induced on $[1, m_\gamma] \setminus \{i_0\}$. Then $|G| \leq |G'| + 1$. The fact that $\gamma_{i_0 1} \equiv 0 \ [2]$ implies that γ is also even with respect to G'. Since $\gamma \equiv \gamma' \ [2]$, we deduce that γ' is even with respect to G'. Applying the induction hypothesis on G' leads to $|G'| \leq \frac{|\gamma'|}{2}$. Thus, $|G| \leq \frac{|\gamma'|}{2} + 1$. Since $\gamma_{i_0 1} \geq 2$, we obtain $|G| \leq \frac{|\gamma|}{2}$, which concludes the proof of the initialization.

Induction step. Now, we suppose that the following holds. For all $r' < r_\gamma$, all β satisfying $r_\beta = r'$ and $\mathcal{G}_\beta^-$ connected, for all partition G of $\{i \in [1, n], \exists j \in [1, p], \ \beta_{ij} > 0\}$ for which β is even, we have the inequality $|G| \leq \frac{|\beta|}{2} - r_\beta + 1$. Let us prove that, for G a partition of $\{i \in [1, n], \exists j \in [1, p], \ \gamma_{ij} > 0\}$, if γ is even with respect to G, the inequality $|G| \leq \frac{|\gamma|}{2} - r_\gamma + 1$ also holds.

We suppose by symmetry that γ is supported on $[1, m_\gamma] \times [1, r_\gamma]$. We call $\mathcal{V}_\gamma$ the graph on $[1, r_\gamma]$ where we connect nodes j and j', if they have a common neighbour in $\mathcal{G}_\gamma^-$, see Figure 1. More formally, for all $j, j' \in [1, p]$, there exists an edge in $\mathcal{V}_\gamma$ between j and j', if and only if, there exists $i \in [1, n]$ such that $\gamma_{ij}\gamma_{ij'} > 0$.

Example.
Let us illustrate the construction of the graphs $\mathcal{V}_\gamma$ and $\mathcal{V}_{\gamma'}$ on an example. Let γ be the matrix below, which is even with respect to the partition $\{1,2,3,4\},\{5\}$

$$\gamma = \begin{pmatrix} 2 & 1 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 & 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 & 2 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 \end{pmatrix}.$$

Figure 1 represents respectively $\mathcal{G}_\gamma^-$ and the corresponding graph $\mathcal{V}_\gamma$. We also represent the set $L := \{i \in [1, m_\gamma],\ \exists i' \in U_{\gamma'},\ \exists A \in G,\ \{i, i'\} \subset A\}$. We do not represent here multi-edges.

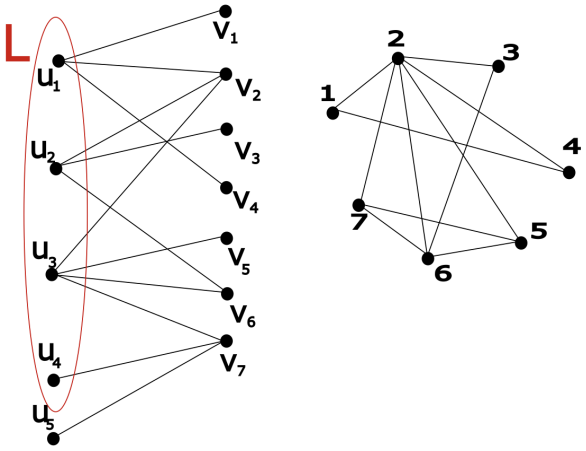

Figure 1: The graph $\mathcal{G}_\gamma^-$ (on the left), and the corresponding graph $\mathcal{V}_\gamma$ (on the right).

For $j, j' \in [1, p]$, the nodes j and j' are in the same connected component for $\mathcal{V}_\gamma$, if and only if, v_j and $v_{j'}$ are in the same connected component for $\mathcal{G}_\gamma$. Hence, the graph $\mathcal{V}_\gamma$ is connected. As a consequence, there exists a spanning tree of this connected component. By symmetry, we can suppose that the node r_γ is a leaf of this tree. This implies that the graph induced by $\mathcal{V}_\gamma$ on $[1, r_\gamma - 1]$ is also connected. We define γ' by $\gamma'_{ij} = \gamma_{ij}\mathbf{1}_{j \neq r_\gamma}$. We write $\mathcal{V}_{\gamma'}$ for the graph induced by $\mathcal{V}_\gamma$ on $[1, r_\gamma - 1]$. The graph $\mathcal{V}_{\gamma'}$ is obtained by removing the node r_γ and the edges connected to it, as represented in Figure 2. This implies that $\mathcal{V}_{\gamma'}$, and therefore also $\mathcal{G}_{\gamma'}^-$ are connected graphs. We denote $U_{\gamma'} = \{i \in [1, n], \exists j \in [1, p], \gamma'_{ij} > 0\}$.

Example (continued). In our example, we have $r_\gamma = 7$, and the node 7 is a leaf of a spanning tree of $\mathcal{V}_\gamma$. After setting to zero the column $j = 7$, we obtain

$$\gamma' = \begin{pmatrix} 2 & 1 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 & 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 & 2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The corresponding graphs $\mathcal{G}^-_{\gamma'}$ and $\mathcal{V}^-_{\gamma'}$, built by removing v_7 and the edges connecting it, are represented in Figure 2.

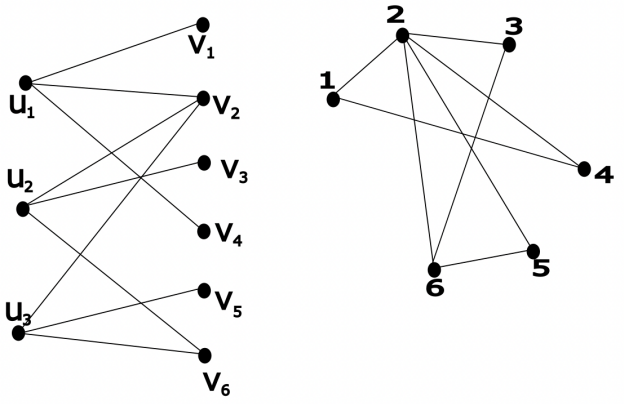

Figure 2: The graph $\mathcal{G}_{\gamma'}$ (on the left), and the corresponding graph $\mathcal{V}_{\gamma'}$ (on the right).

Let G a partition of $[1, m_\gamma]$ such that γ is even with respect to G. The r_γ-th column of γ' being null, we have that, for all $A \in G$, $\sum_{i \in A} \gamma'_{ir_\gamma} = 0$.

For $j \neq r_\gamma$, the j-th column of γ is equal to that of γ'. This implies, since γ is even with respect to G, that $\sum_{i \in A} \gamma'_{ij} \equiv 0 \ [2]$ for any A in G.

Thus, for all groups $A \in G$, all $j \in [1, r_\gamma]$, we have $\sum_{i \in A} \gamma'_{ij} \equiv 0 \ [2]$. This means that γ' is even with respect to the partition G. Now, we distinguish in the set U_γ different type of nodes. First, we have the set $U_{\gamma'} = \{i \in [1, n], \exists j \in [1, r_\gamma - 1], \gamma'_{ij} > 0\}$ of nodes of U spanned by $\mathcal{G}^-_{\gamma'}$. Then, we define $L = \{i \in [1, m_\gamma], \ \exists i' \in U_{\gamma'}, \ \exists A \in G, \ \{i, i'\} \subset A\}$, the set of nodes of U which are linked to $U_{\gamma'}$ through the partition G. Note that $U_{\gamma'} \subset L$. Finally, we consider the remaining nodes $[1, m_\gamma] \setminus L$.

We respectively define G' the restriction of G to $U_{\gamma'}$ and G'' the restriction of G to $[1, m_\gamma] \setminus L$. By definition of L, we have $|G| = |G'| + |G''|$.

Let us first use the induction hypothesis on $|G'|$. To do so, let us prove that γ' is even with respect to G'. We know that γ' is even with respect to G. Moreover, every group A' of G' is the restriction of a group A of G to $U_{\gamma'}$. This implies that, for all $j \in [1, r_\gamma - 1]$, the equality $\sum_{i \in A'} \gamma'_{ij} = \sum_{i \in A} \gamma'_{ij}$ holds. Hence, γ' is even with respect to G'. This allows us to apply the induction hypothesis and leads to $|G'| \leq \frac{|\gamma'|}{2} - r_\gamma + 2$.

Now, we upper bound $|G''|$. For all group A of G'', it is clear by definition of L that A is also a group of G which contains only elements from $[1, m_\gamma] \setminus L$. Hence, $\sum_{i \in A} \gamma_{ir_\gamma} \equiv 0 \ [2]$. Thus, we know from the initialization step of the induction that $|G''| \leq \frac{|\gamma''|}{2}$, where we define γ'' as the restriction of γ to $(\cup_{A \in G''} A) \times \{r_\gamma\}$. This leads to $|G| = |G'| + |G''| \leq \frac{|\gamma''|}{2} + \frac{|\gamma'|}{2} - r_\gamma + 2$. It remains to prove that $|\gamma'| + |\gamma''| \leq |\gamma| - 2$. It is clear that $|\gamma| = |\gamma'| + |\gamma''| + \sum_{i \in L} \gamma_{ir_\gamma}$. We

distinguish three cases.

First case: If $|\{i \in [1, m_\gamma],\ \gamma_{ir_\gamma} \equiv 1\ [2]\} \cap U_{\gamma'}| = 1$. We call i_0 the only point in this set. We denote A the group of G such that $i_0 \in A$. By hypothesis, $\sum_{i\in A} \gamma_{ir_\gamma} \equiv 0\ [2]$. Hence, there exists $i \neq i_0 \in A$ which satisfies $\gamma_{ir_\gamma} \equiv 1\ [2]$. Thus, since $A \subset L$, we get $\sum_{i\in L} \gamma_{ir_\gamma} \geq 2$. This leads to $|\gamma| \geq |\gamma'| + |\gamma''| + 2$.

Second case: If $|\{i \in [1, m_\gamma],\ \gamma_{ir_\gamma} \equiv 1\ [2]\} \cap U_{\gamma'}| = 0$. Since $\mathcal{G}_\gamma$ is connected, there exists $i \in U_{\gamma'}$ such that $\gamma_{ir_\gamma} > 0$. Since $|\{i \in [1, m_\gamma],\ \gamma_{ir_\gamma} \equiv 1\ [2]\} \cap U_{\gamma'}| = 0$, we have $\gamma_{ir_\gamma} \geq 2$. The fact that $U_{\gamma'} \subset L$ leads to $\sum_{i\in L} \gamma_{ir_\gamma} \geq 2$. Hence, $|\gamma| \geq |\gamma'| + |\gamma''| + 2$.

Third case: If $|\{i \in [1, m_\gamma],\ \gamma_{ir_\gamma} \equiv 1\ [2]\} \cap U_{\gamma'}| \geq 2$. In this case, $\sum_{i\in U_{\gamma'}} \gamma_{ir_\gamma} \geq 2$. This leads to $|\gamma| \geq |\gamma'| + |\gamma''| + 2$.

This concludes our induction and leads to the sought inequality $|G| \leq \frac{|\gamma|}{2} - r_\gamma + 1$ for any γ such that $\mathcal{G}_\gamma^-$ is connected and for any partition G of $\{i \in [1, n],\ \exists j \in [1, p],\ \gamma_{ij} > 0\}$ that is even with respect to γ. In the general case where $\mathcal{G}_\alpha$ has CC_γ connected components, we readily get $|G| \leq \frac{|\gamma|}{2} - r_\gamma + CC_\gamma$.

A.6 A computational barrier for $p \leq n$

In this section, we adapt the proof of Theorem 1 to provide a computational lower bound when $p \leq n$. This lower-bound provides evidence for the existence of a computation-information gap for $\frac{n}{K} \vee K \widetilde{\ll} p \leq n$, where $\widetilde{\ll}$ hides some polylog(n) factors. We believe yet, that our computational lower-bound is not tight in this regime.

Proposition 2. *Let $D \in \mathbb{N}$. If $p \leq n$ and $\overline{\zeta}_n := \frac{\bar{\Delta}^4 D^8 (1+D)^4 n}{p^2} \max\left(\frac{n}{K^2}, 1\right) < 1$, then under the prior of Definition 1, we have*

$$MMSE_{\leq D} \geq \frac{1}{K} - \frac{1}{K^2}\left(1 + \frac{\overline{\zeta}_n}{(1 - \sqrt{\overline{\zeta}_n})^3}\right) .$$

In particular, if $\bar{\Delta}^2 \ll D^{-6}\left(\frac{pK}{n} \wedge \frac{p}{\sqrt{n}}\right)$, then $MMSE_{\leq D} = \frac{1}{K} - \frac{1+o(1)}{K^2}$.

The second statement of Proposition 2 states that low-degree polynomials with degree $D \leq (\log(n))^{1+\eta}$ do not perform better than the trivial estimator when

$$\bar{\Delta}^2 \widetilde{\ll} \left(\frac{pK}{n} \wedge \frac{p}{\sqrt{n}}\right),$$

where $\widetilde{\ll}$ hides polylog(n) factors. Since lower-bounds for low-degree polynomials with degree $D \leq (\log(n))^{1+\eta}$ are considered as evidence of the computational hardness of the problem, this suggests computational hardness of estimating M^* when $\bar{\Delta}^2 \widetilde{\ll} \left(\frac{pK}{n} \wedge \frac{p}{\sqrt{n}}\right)$ and $p \leq n$. Since, as made explicit in Section 2, estimation of M^* is possible in polynomial time when clustering is possible in polynomial time, this provides compelling evidence for the computational hardness of the clustering problem in this regime.

Comparing the computational lower bound $\bar{\Delta}^2 \widetilde{\gg} \left(\frac{pK}{n} \wedge \frac{p}{\sqrt{n}}\right)$ for $p \leq n$, to the information barrier

$$\Delta^2 \gtrsim \log(K) \vee \sqrt{\frac{pK \log(K)}{n}},$$

we observe that there is a computation-information gap when

$$1 \vee \sqrt{\frac{pK}{n}} \widetilde{\ll} \left(\frac{pK}{n} \wedge \frac{p}{\sqrt{n}} \right),$$

which happens when

$$\frac{n}{K} \vee K \widetilde{\ll} p \leq n\,.$$

Proof of Proposition 2. We argue exactly as in the proof of Theorem 1 by upper bounding $corr_D^2$. In the proof of Theorem 1, we only used the assumption that $n \leq p$ in (19). Hence, we start from (18) by plugging the definition of $\overline{\zeta}_n = \bar{\Delta}^4 D^4 (1+D)^2 \frac{n}{p^2} \max\left(\frac{n}{K^2}, 1\right)$. This leads us to

$$\begin{aligned}
&corr_D^2 - \frac{1}{K^2} \\
\leq &\frac{1}{K^2} \sum_{d=1}^{D} \sum_{(r,m)\in\mathcal{D}_d} \overline{\zeta}_n^{d/2} p^r n^{m-2-d/2} \left(\frac{1}{\max(1, n/K^2)} \right)^{d/2} \min\left(1, \left(\frac{1}{K}\right)^{2m+2r-d-4}\right) \\
\leq &\frac{1}{K^2} \sum_{d=1}^{D} \sum_{(r,m)\in\mathcal{D}_d} \overline{\zeta}_n^{d/2} n^{m+r-2-d/2} \left(\frac{1}{\max(1, n/K^2)} \right)^{d/2} \min\left(1, \left(\frac{1}{K}\right)^{2m+2r-d-4}\right), \qquad (25)
\end{aligned}$$

where we used in the last line that $n \geq p$. Note that this upper bound is exactly the same as in (19) except that ζ_n has been replaced by $\overline{\zeta}_n$. Hence, arguing as in the proof of Theorem 1, we arrive at the similar conclusion

$$corr_D^2 \leq \frac{1}{K^2} \left[1 + \frac{\overline{\zeta}_n}{(1 - \sqrt{\overline{\zeta}_n})^3} \right].$$

□

B Proof of Theorem 2

Without loss of generality, we assume throughout this proof that $\sigma = 1$. We will use the following notation. For $i \in [1, n]$, we decompose $Y_i = \mathbb{E}(Y_i) + E_i = \mu_k + E_i$, if $i \in G_k^*$. Then, $(E_i)_{i\in[1,n]}$ are independent vectors, with distribution $\mathcal{N}(0, I_p)$. We denote:

- $Y \in \mathbb{R}^{n\times p}$ whose i-th row is the vector Y_i,
- $A \in \mathbb{R}^{n\times K}$ the membership matrix defined by $A_{ik} = \mathbf{1}_{i\in G_k^*}$,
- $E \in \mathbb{R}^{n\times p}$ the noise matrix whose i-th row is the Gaussian vector E_i,
- $\mu \in \mathbb{R}^{K\times p}$ whose k-th row is μ_k.

We then have the relation

$$Y = A\mu + E.$$

Let us also denote $m = \min_{k\in[1,K]}$ the minimal size of the clusters, and $m^+ = \max_{k\in[1,K]}$ the maximal size. The hypothesis $G^* \in \mathcal{P}_\alpha$ is equivalent to $\frac{m^+}{m} \leq \alpha$. We define the signal-to-noise ratio $\tilde{s}^2 = \Delta^2 \wedge \frac{\Delta^4 m}{p}$. We note that $\frac{1}{\alpha} s^2 \leq \tilde{s}^2 \leq s^2$, where s^2 is the signal-to-noise ratio defined in (12).

We prove in this section a more general theorem, which induces directly Theorem 2.

Theorem 5. *There exist positive numerical constants c, c′, c″ such that the following holds. If*

$$\tilde{s}^2 \geq c(\log(n/m) \vee \frac{m^+}{m}) \ , \tag{26}$$

then, we have $err(\hat{G}, G^*) \leq e^{-c''\tilde{s}^2}$ *with probability at least* $1 - c'/n^2$.

First, let us formulate the K-means criterion in an alternative way. Given $G = \{G_1, ..., G_K\}$ a partition, let us define the normalized-partnership matrix $B(G)$ by:

$$B_{ij} = \sum_{k \leq K} \mathbf{1}_{i,j \in G_k} \frac{1}{|G_k|} \ .$$

The application $G \to B(G)$ is a bijection on

$$\mathcal{B} = \{B \in S_n(\mathbb{R})^+ : \ B_{ij} \geq 0, \mathrm{Tr}(B) = K, B1 = 1, B^2 = B\} \ .$$

We refer to [16] and to Chapter 12.4 of [31] for this last statement. It implies the following proposition.

Proposition 3 ([16])**.** *Finding* $\widehat{G} \in \mathrm{argmin}_G \mathrm{Crit}(G)$ *is equivalent to finding*

$$\widehat{B} \in \underset{B \in \mathcal{B}}{\mathrm{argmax}} \langle YY^T, B \rangle \ .$$

We will denote in the following $\hat{B}$ a minimiser of $\langle YY^T, B\rangle$ over the set $\mathcal{B}$. We note that the convex relaxation of this problem introduced in [16] and studied in [7] is the minimiser of $\langle YY^T, B\rangle$ over the convex set

$$\mathcal{C} = \left\{B \in S_n(\mathbb{R})^+ : \ B_{ij} \geq 0, \ \mathrm{Tr}(B) = K, \ B1 = 1\right\} \ .$$

Since we have $\mathcal{B} \subset \mathcal{C}$, all the bounds obtained in [7] for the matrices in $\mathcal{C}$ are *de facto* valid for $\mathcal{B}$.

The proof of Theorem 5 follows the same main steps as the analysis of relaxed K-means in [7]. The main difference is in the delicate proof of Lemma 11 below. First, we use that, for any partition G, the proportion $err(G, G^*)$ of misclustered points is controlled as a function of the l_1 norm $\|B^* - B^*B\|_1$, where B is the normalized-partnership matrix associated to G. More precisely, we show, in Section B.2, the following Lemma.

Lemma 9. *Consider two partitions* G^* *and* G*. Write* B^* *and* B *for the corresponding normalized-partnership matrices. For some numerical constant* $c > 0$*, we have*

$$err(G, G^*) \leq c \left(\frac{m^+}{m}\right) \frac{\|B^* - B^*B\|_1}{n} \ .$$

As a consequence, we only have to control the l_1 error $\|B^* - B^*\hat{B}\|_1$. Again, as in [7], we start from the optimality condition that defines $\hat{B}$, that is

$$\langle YY^T, \hat{B} - B^* \rangle \geq 0 \ .$$

By definition, we have $Y = A\mu + E$. Plugging this expression in the above inequality leads to

$$\langle A\mu(A\mu)^T, B^* - \hat{B} \rangle \leq \langle A\mu E^T + E(A\mu)^T, \hat{B} - B^* \rangle + \langle EE^T, \hat{B} - B^* \rangle \ . \tag{27}$$

We call $\langle A\mu(A\mu)^T, \hat{B} - B^* \rangle$ the signal term, $\langle EE^T, \hat{B} - B^* \rangle$ the quadratic noise term, and $\langle A\mu E^T + E(A\mu)^T, \hat{B} - B^* \rangle$ the cross term. The three following lemmas control each of these three terms. For any $B \in \mathcal{B}$, we denote $\delta_B = \|B^* - B^*B\|_1$.

Lemma 10. *For all $B \in \mathcal{B}$, we have*

$$\langle A\mu(A\mu)^T, B^* - B\rangle = \langle S, B^* - B\rangle \geq \frac{1}{2}\Delta^2\delta_B \ ,$$

where $S_{ab} = -\frac{1}{2}\|\mu_a - \mu_b\|^2$ for a and b in $[1, n]$.

Lemma 11. *There exists numerical constants c_1 and c_2 such that the following holds with probability at least $1 - \frac{c_1}{n^2}$. Simultaneously for all $B \in \mathcal{B}$, we have:*

$$\begin{aligned}\langle EE^T, B - B^*\rangle \leq & c_2\delta_B\left(\log(n/m) \vee \frac{m^+}{m} + \sqrt{\frac{p}{m}\left(\log(n/m) \vee \frac{m^+}{m}\right)}\right) \\ & + c_2\delta_B\left(\sqrt{\frac{p}{m}\log(n\frac{K^3}{\delta})} + \log(n\frac{K^3}{\delta})\right) \ .\end{aligned}$$

Lemma 12. *There exist constants c_3 and c_4 such that the following holds with probability at least $1 - \frac{c_3}{n^2}$. Simultaneously for all $B \in \mathcal{P}$, we have*

$$\langle A\mu E^T + E(A\mu)^T, B - B^*\rangle \leq c_4\sqrt{\langle S, B^* - B\rangle}\sqrt{\delta_B\log(\frac{nK^3}{\delta})} \ ,$$

where $S_{ab} = -\frac{1}{2}\|\mu_a - \mu_b\|^2$.

Lemmas 10 and 12, taken from [7], are in fact valid for the larger class of matrices $B \in \mathcal{C}$. The difference with [7] lies in the quadratic noise term controlled by Lemma 11. For this quadratic term, we can get an upper-bound over the class $\mathcal{B}$, that is significantly smaller than over the class $\mathcal{C}$. Indeed, here, we can leverage the fact that the class $\mathcal{B}$ of matrices is finite. Instead of having an upper-bound of the order of n/m, we get a smaller upper-bound of the order of $\log(n/m) \vee \frac{m^+}{m}$. This is the main reason for which the Condition (26) for exact Kmeans is less stringent than the condition $\tilde{s} \geq n/m$ for the SDP relaxation of Kmeans. We refer to Section B.1 for a proof of Lemma 11, which is the main hurdle for proving Theorem 5. We note that when the constant c of (26) is large enough, the term $c_2\delta_B\left(\log(n/m) \vee \frac{m^+}{m} + \sqrt{\frac{p}{m}\left(\log(n/m) \vee \frac{m^+}{m}\right)}\right)$ is bounded by $\frac{\Delta^2\delta_B}{8}$. That allows us to neglect it in (27) up to a multiplicative constant.

Combining (27) with Lemmas 10–12, we deduce that, for some constants c' and c'' the following holds with probability at least $1 - \frac{c'}{n^2}$

$$\langle S, B^* - \hat{B}\rangle \leq c''\left[\sqrt{\langle S, B^* - \hat{B}\rangle}\sqrt{\delta_{\hat{B}}\log(\frac{nK^3}{\delta_{\hat{B}}})} + \delta_{\hat{B}}\left(\sqrt{\frac{p}{m}\log(n\frac{K^3}{\delta_{\hat{B}}})} + \log(n\frac{K^3}{\delta_{\hat{B}}})\right)\right] \ .$$

Below, we write for convenience $a \lesssim b$ for $a \leq c\, b$, with c a positive numerical constant that may vary from line to line. The above bound implies

$$\langle S, B^* - \hat{B}\rangle \lesssim \delta_{\hat{B}}\log(\frac{nK^3}{\delta_{\hat{B}}}) \vee \delta_{\hat{B}}\sqrt{\frac{p}{m}\log(\frac{nK^3}{\delta_{\hat{B}}})} \ .$$

Moreover, by Lemma 10, we have $\frac{1}{2}\Delta^2\delta_{\hat{B}} \leq \langle S, B^* - \hat{B}\rangle$. This leads us to

$$\Delta^2\delta_{\hat{B}} \lesssim \delta_{\hat{B}}\log(\frac{nK^3}{\delta_{\hat{B}}}) \vee \delta_{\hat{B}}\sqrt{\frac{p}{m}\log(\frac{nK^3}{\delta_{\hat{B}}})} \ ,$$

which, in turn, implies that

$$\log(\frac{nK^3}{\delta_{\hat{B}}}) \gtrsim \Delta^2 \wedge \frac{\Delta^4 m}{p} = \tilde{s}^2 \ .$$

Thus, for some numerical constant c_0, we have

$$\delta_{\hat{B}} \leq nK^3 e^{-c_0 \tilde{s}^2} \ .$$

Coming back to Lemma 9, we conclude that the proportion of misclassified points satisfies $err(\hat{G}, G^*) \leq (\frac{m^+}{m})K^3 e^{-c_0\tilde{s}^2} \leq (\frac{n}{m})^4 e^{-c_0\tilde{s}^2}$. Therefore, provided that the constant c in Condition (26) is large enough, there exists a constant c'' such that

$$\mathrm{err}(\hat{G}, G^*) \leq e^{-c''\tilde{s}^2} \ .$$

This concludes the proof.

B.1 The quadratic noise term (Proof of Lemma 11)

In this section, we will upper-bound uniformly over all $\mathcal{B}$ the noise term $\langle EE^T, \hat{B} - B^* \rangle$.

Let us decompose, for $B \in \mathcal{B}$, the term $B - B^*$. We get: $B - B^* = B^*(B - B^*) + (B - B^*)B^* + (I_n - B^*)(B - B^*)(I_n - B^*) - B^*(B - B^*)B^*$. We also remark that, since for all $B \in \mathcal{B}$, $\mathrm{Tr}(B) = K$, we have $\langle I_n, B - B^* \rangle = 0$. We thus get

$$\begin{aligned}\langle EE^T, B - B^* \rangle =& 2\langle EE^T - pI_n, B^*(B - B^*) \rangle \\ &+ \langle EE^T - pI_n, (I_n - B^*)(B - B^*)(I_n - B^*) \rangle \\ &- \langle EE^T - pI_n, B^*(B - B^*)B^* \rangle.\end{aligned}$$

We will control each of these terms.

First, let us deal with the term in $(I_n - B^*)(B - B^*)(I_n - B^*)$. We show, using Hanson-Wright lemma and union bounds over the sets

$$\mathcal{B}_j := \{B \in \mathcal{B}, \delta_B \in [j-1, j]\}, \ \ j \leq 2n \ ,$$

the following lemma (see section B.1.1 for the proof of this lemma).

Lemma 13. *There exists c_5, c_6 two positive numerical constants such that the following holds with probability at least $1 - \frac{c_5}{n^2}$. Simultaneously for all $B \in \mathcal{B}$, we have*

$$\langle EE^T - pI_n, (I - B^*)(B - B^*)(I - B^*) \rangle \leq c_6 \delta_B \left(\log(n/m) \vee \frac{m^+}{m} + \sqrt{\frac{p}{m}\left(\log(n/m) \vee \frac{m^+}{m} \right)} \right).$$

Now we will upper bound the other quadratic terms (the ones corresponding to $B^*(B - B^*)$ and to $B^*(B - B^*)B^*$). To do so, we will distinguish two cases:

1. We will use the bound obtained in [7] for the matrices B such that $\delta_B \leq m$;
2. We will then show a similar result than Lemma 13 for the matrices B that fulfill $\delta_B \geq m$.

For the first point, we will use Lemma 7 of [7] which states:

Lemma 14. *There exist positive numerical constants c_7 and c_8 such that the following holds with probability at least $1 - \frac{c_7}{n^2}$. Simultaneously for all $B \in \mathcal{B}$, we have*

$$\langle EE^T - pI_n, B^*(B - B^*) \rangle \leq \frac{c_8 \delta_B}{\sqrt{m}} \left(\sqrt{p \log(n \frac{K^3}{\delta_B})} + \sqrt{\delta_B + 1} \log(n \frac{K^3}{\delta_B}) \right) \ .$$

The same result holds for the term in $B^*(B - B^*)B^*$.

This lemma implies that, with probability at least $1 - \frac{c_7}{n^2}$, for all matrices B fulfilling $\delta_B \leq m$,

$$\langle EE^T - pI_n, 2B^*(B - B^*) - B^*(B - B^*)B^* \rangle \leq 6c_8\delta \left(\sqrt{\frac{p}{m} \log(n\frac{K^3}{\delta})} + \log(n\frac{K^3}{\delta}) \right).$$

For the second point, we will do as for Lemma 13 and use Hanson-Wright Lemma and union bounds over the sets $\mathcal{B}_j$ to get the following Lemma (see Section B.1.4 for a proof of this lemma).

Lemma 15. *There exists c_9 and c_{10} two positive numerical constants such that the following holds. With probability higher than $1 - \frac{c_9}{n^2}$, we have simultaneously for all $B \in \mathcal{B}$ such that $\|B^* - B^*B\|_1 \geq m$,*

$$\langle EE^T - pI_n, B^*(B - B^*) \rangle \leq c_{10}\delta_B \left(\log(n/m) \vee \frac{m^+}{m} + \sqrt{\frac{p}{m} \left(\log(n/m) \vee \frac{m^+}{m} \right)} \right).$$

The same result holds for the term in $-B^*(B - B^*)B^*$.

Combining the results from Lemma 13-15, we get the bound of Lemma 11.

B.1.1 Proof of Lemma 13

Let us use Hanson-Wright Lemma, that we recall here (see e.g. [31])

Lemma 16. (Hanson-Wright inequality) *Let $\varepsilon \sim \mathcal{N}(0, I_d)$ for some d. Let S be a real symmetric $p \times p$ matrix. Then, for all $x > 0$,*

$$\mathbb{P}\left[\varepsilon^T S\varepsilon - \mathrm{Tr}(S) > \sqrt{8\|S\|_F^2 x} \vee (8\|S\|_{op} x) \right] \leq e^{-x}.$$

That inequality implies that, for a fixed matrix $B \in \mathcal{B}$, with probability at least $1 - e^{-x}$,

$$\begin{aligned}\langle EE^T - pI_n, (I - B^*)(B - B^*)(I - B^*) \rangle \leq & \sqrt{8px\|(I - B^*)(B - B^*)(I - B^*)\|_F^2} \\ & \vee (8\|(I - B^*)(B - B^*)(I - B^*)\|_{op} x) \ .\end{aligned}$$

Our task will be to bound both $\|(I - B^*)(B - B^*)(I - B^*)\|_{op}$ and $\|(I - B^*)(B - B^*)(I - B^*)\|_F$ with respect to δ_B. We will then do an union bound on all $B \in \mathcal{B}_j := \{B \in \mathcal{B}, \delta_B \in [j-1, j]\}$, and finally a union bound on all j in $[1, 2n]$. From Lemma 1 of [7], we have $\delta_B = 2\sum_{l \neq k} |B_{G_k^* G_l^*}|_1$. For all $i \in G_k^*$, $\sum_{l \neq k} |B_{iG_l^*}| \leq 1$. This implies $\delta_B \leq 2n$. Hence, it is sufficient to consider $j \leq 2n$. We have, for $j \in [1, 2n]$ and $B \in \mathcal{B}_j$,

1. $\|(I - B^*)(B - B^*)(I - B^*)\|_{op} \leq \|(I - B^*)(B - B^*)(I - B^*)\|_*$ where $\|\cdot\|_*$ denotes the nuclear norm. And, Lemma 1 of [7] shows that $\|(I - B^*)(B - B^*)(I - B^*)\|_* \lesssim \frac{1}{m}\delta_B$. Hence: $\|(I - B^*)(B - B^*)(I - B^*)\|_{op} \lesssim \frac{1}{m}\delta_B$;
2. Since $(I - B^*)(B - B^*)(I - B^*)$ is a difference of product of projections, all its eigen-values are bounded by 2 in absolute value. Hence: $\|(I - B^*)(B - B^*)(I - B^*)\|_{op} \lesssim \frac{1}{j}\delta_B$;
3. $\|(I - B^*)(B - B^*)(I - B^*)\|_F \lesssim \|(I - B^*)(B - B^*)(I - B^*)\|_*$ and therefore, we also have $\|(I - B^*)(B - B^*)(I - B^*)\|_F \lesssim \frac{1}{m}\delta_B$;
4. Since all the eigen-values of $(I - B^*)(B - B^*)(I - B^*)$ are bounded by 2 in absolute value, we have $\|(I - B^*)(B - B^*)(I - B^*)\|_F \lesssim \sqrt{\|(I - B^*)(B - B^*)(I - B^*)\|_*}$. Thus, $\|(I - B^*)(B - B^*)(I - B^*)\|_F \lesssim \frac{1}{\sqrt{mj}}\delta_B$.

These inequalities on the norms imply that, for $B \in \mathcal{B}_j$ and $x > 0$, with probability at least $1 - e^{-x}$,

$$\langle EE^T - I_n, (I - B^*)(B - B^*)(I - B^*)\rangle \lesssim \delta_B \left(\sqrt{\frac{p}{m} \frac{x}{m \vee j}} + \frac{x}{m \vee j} \right).$$

Doing an union bound on $\mathcal{B}_j$, it appears that for $x > 0$, this inequality remains true with probability at least $1 - |\mathcal{B}_j| e^{-x}$, simultaneously on all $B \in \mathcal{B}_j$. This implies that for a positive constant c_1, with probability at least $1 - \frac{c_1}{2n^3}$, simultaneously on all $B \in \mathcal{B}_j$:

$$\langle EE^T - I_n, (I - B^*)(B - B^*)(I - B^*)\rangle \lesssim \delta_B \left(\sqrt{\frac{p}{m} \frac{\log(|\mathcal{B}_j| \vee n)}{j \vee m}} \vee \frac{\log(|\mathcal{B}_j| \vee n)}{j \vee m} \right) .$$

The next lemma, proved in Section B.1.2, bounds the cardinality of $|\mathcal{B}_j|$.

Lemma 17. *For any $j \in [1, 2n]$, $|\mathcal{B}_j| \leq \binom{n}{j \wedge n} K^{3j} 2^{j \frac{m^+}{m}}$.*

Hence, with probability at least $1 - \frac{c_1}{2n^3}$, simultaneously on all $B \in \mathcal{B}_j$,

$$\begin{aligned}\langle EE^T - I_n, (I - B^*)(B - B^*)(I - B^*)\rangle \lesssim & \delta_B \Bigg(\sqrt{\frac{p}{m} \frac{\log(n) + \log(\binom{n}{j \wedge n}) + j \log(K) + j \frac{m^+}{m}}{j \vee m}} \\ & + \frac{\log(n) + \log(\binom{n}{j \wedge n}) + j \log(K) + j \frac{m^+}{m}}{j \vee m} \Bigg) .\end{aligned}$$

And,

- We have $\log \binom{n}{j \wedge n} \lesssim j \log(\frac{n}{j \wedge n})$. The function $x \to x \log(\frac{n}{x})$ being increasing on $]0, \frac{n}{e}]$, we get that, when $j \leq m$ and $m \leq \frac{n}{e}$, $j \log(\frac{n}{j \wedge n}) \lesssim m \ln(\frac{n}{m})$. When $j \geq m$, we have $j \log(\frac{n}{j \wedge n}) \leq j \log(\frac{n}{m})$. Moreover, if $m \geq \frac{n}{e}$ and $j \leq n$, $j \log(\frac{n}{j \wedge n}) \leq \frac{n}{e} \log(e) \lesssim m \log(\frac{n}{m})$. If $m \geq \frac{n}{e}$ and $j \geq n$, then $j \log(\frac{n}{j \wedge n}) = 0 \leq m \log(n/m)$. Hence, we get that $\log \binom{n}{j \wedge n} \lesssim (m \vee j) \log(\frac{n}{m})$;
- $j \log(K) \leq j \log(\frac{n}{m})$ since $K \leq \frac{n}{m}$;
- The fact that the function $x \to x \log(\frac{n}{x})$ is increasing on $]0, \frac{n}{e}]$ also implies that, if $m \leq \frac{n}{e}$, $\log(n) \lesssim m \log(\frac{n}{m})$. If $m \geq \frac{n}{e}$, we have $\log(n) \lesssim n \lesssim \frac{n}{e} \lesssim m \log(n/m)$.

Thus, with the same probability $1 - \frac{c_1}{2n^3}$, simultaneously on all $B \in \mathcal{B}_j$,

$$\langle EE^T - I_n, (I - B^*)(B - B^*)(I - B^*)\rangle \lesssim \delta_B \left(\log(n/m) \vee \frac{m^+}{m} + \sqrt{\frac{p}{m} \left(\log(n/m) \vee \frac{m^+}{m} \right)} \right).$$

An union bound on $j \in [1, 2n]$ concludes the proof of the lemma.

B.1.2 Proof of Lemma 17

In this section, we consider the partitions G such that $B(G) \in \mathcal{B}_j$. For $B \in \mathcal{B}_j$, denote $B_{G_k^* G_l^*}$ the restriction of B where we keep the rows belonging to G_k^* and the columns belonging to G_l^*. From Lemma 1 of [7], we get that, $\delta_B = 2 \sum_{l \neq k} |B_{G_k^* G_l^*}|$, for $B \in \mathcal{C}$. This equality tells us that if a point $i \in G_r^*$, for some $r \in [1, K]$, is linked in G to a majority of points not belonging to the same G_r^*, then the contribution of $2 \sum_{k : k \neq r} |B_{i G_r^*}|$ in δ_B is at least of one. Indeed, denoting by l the index such that $i \in G_l$, we have

$$\begin{aligned} 2 \sum_{k : k \neq r} |B_{i G_r^*}| &= 2 \sum_{j \in G_l \setminus G_r^*} \frac{1}{|G_l|} \\ &= 2 \frac{|G_l \setminus G_r^*|}{|G_l|} \geq 1 . \end{aligned}$$

In order to bound $|\mathcal{B}_j|$, we will find, for a partition G whose normalized-partnership matrix is in $\mathcal{B}_j$, a labelling $G_1, ..., G_K$ for which we can upper-bound the possibilities for choosing points that are in some $G_r^* \cap G_l$ for $l \neq r$.

The equality that fulfills δ_B will help us bound the possibilities of choosing the points that are in some $G_r^* \cap G_l$, with $l \neq r$ and $\frac{|G_l \cap G_r^*|}{|G_l|} \leq \frac{1}{2}$. On the other hand, the following lemma allows us to control the number of cotuple (l, r), with $l \neq r$, satisfying $\frac{|G_l \cap G_r^*|}{|G_l|} > \frac{1}{2}$, by stating that for any of these cotuple, we can consider that all the points belonging to G_l^* add a contribution of at least one to δ_B.

Lemma 18. *Let $G = \{G_1, ..., G_K\}$ be a partition of $[1, n]$. There exists $\phi : [1, K] \to [1, K]$ a bijection such that the following holds. For all $l \neq r$ such that $\frac{|G_{\phi(l)} \cap G_r^*|}{|G_{\phi(l)}|} > \frac{1}{2}$, we have $\max_{l'} \frac{|G_l^* \cap G_{\phi(l')}|}{|G_{\phi(l')}|} \leq \frac{1}{2}$.*

This lemma is proved in Section B.1.3. Let G such that $B(G) \in \mathcal{B}_j$. Without loss of generality, we can suppose that the bijection ϕ considered in Lemma 18 is the identity. Fix any $l \neq r$ in $[1, K]$. We consider two cases

1. $\frac{|G_l^* \cap G_r|}{|G_r|} \leq \frac{1}{2}$. For any $i \in G_l^* \cap G_r$, we have
$$\sum_{k:k\neq l} |B_{iG_k^*}| = \sum_{t \in [1,n] \setminus G_l^*} B_{it} \geq \frac{1}{|G_r|} |G_r \setminus G_l^*| \geq \frac{1}{2} .$$

2. $\frac{|G_l^* \cap G_r|}{|G_r|} > \frac{1}{2}$. Consider any $i \in G_r^*$. Then, denoting l' the index such that $i \in G_{l'}$, we have from Lemma 18 that $\frac{|G_r^* \cap G_{l'}|}{|G_{l'}|} \leq \frac{1}{2}$. Arguing as above, this implies $\sum_{k\neq r} |B_{iG_k^*}| \geq \frac{1}{2}$. Hence, $\sum_{k:k\neq r} |B_{G_r^* G_k^*}| \geq \frac{m}{2}$.

Since $B \in \mathcal{B}_j$, we deduce from the above discussion that (i) there are at most $j \wedge n$ points that belong to $G_l^* \cap G_r$ for some $l \neq r$ such that $\frac{|G_l^* \cap G_r|}{|G_r|} \leq \frac{1}{2}$ and (ii) that $|\{(l, r), l \neq r, \frac{|G_l^* \cap G_r|}{|G_r|} > \frac{1}{2}\}| \leq \frac{j}{m}$.

Hence, $|\mathcal{B}_j|$ is upper-bounded by the number of partitions satisfying these two conditions. Such partitions are fully defined by the points i that are in some $G_l^* \cap G_r$, for $l \neq r$. So, it is sufficient to count the possibilities of choosing these points:

- We choose $j \wedge m$ points, and for each of these points we decide a group G_r where it is sent (r is not necessarily different than the index of the group of G^* that contains i); we have $\binom{n}{j\wedge n} K^{j\wedge n}$ such possibilities,
- We choose $\frac{j}{m}$ couples (l, r) and, for each of these couples, we choose a subset of G_l^* that will be a subset of G_r. That gives less than $2^{j\frac{m^+}{m}} K^{2\frac{j}{m}}$ possibilities.

Hence, we conclude that $|\mathcal{B}_j| \leq \binom{n}{j\wedge n} K^{3j} 2^{j\frac{m^+}{m}}$.

B.1.3 Proof of Lemma 18

Let $G = \{G_1, ..., G_K\}$ a partition of $[1, n]$. For $l \in [1, K]$ such that there exists $l' \in [1, K]$ which satisfies $\frac{|G_l^* \cap G_{l'}|}{|G_{l'}|} > \frac{1}{2}$, we define $\phi(l) = l'$. The mapping ϕ is an injection since for all $l' \in [1, K]$, there is at most one index l such that $\frac{|G_l^* \cap G_{l'}|}{|G_{l'}|} > \frac{1}{2}$. Then, we can expand ϕ to $[1, K]$ in order to have a permutation of $[1, K]$.

The permutation ϕ verifies the following assertion. For $l \in [1, K]$, if there exists l' such that $\frac{|G_l^* \cap G_{\phi(l')}|}{|G_{\phi(l')}|} > \frac{1}{2}$, then, we also have $\frac{|G_l^* \cap G_{\phi(l)}|}{|G_{\phi(l)}|} > \frac{1}{2}$. So, for $l \neq r$ such that $\frac{|G_{\phi(l)} \cap G_r^*|}{|G_{\phi(l)}|} > \frac{1}{2}$, there exists no l' such that $\frac{|G_l^* \cap G_{\phi(l')}|}{|G_{\phi(l')}|} > \frac{1}{2}$. Indeed, otherwise, we would also have $\frac{|G_l^* \cap G_{\phi(l)}|}{|G_{\phi(l)}|} > \frac{1}{2}$. That would contradict the hypothesis $\frac{|G_r^* \cap G_{\phi(l)}|}{|G_{\phi(l)}|} > \frac{1}{2}$.

This concludes the proof of the lemma.

B.1.4 Proof of Lemma 15

For any B in $\mathcal{B}$, Lemma 16 implies that, for $L \geq 0$, with probability at least $1 - e^{-L}$,

$$\begin{aligned}\langle EE^T - I_n, B^*(B - B^*)\rangle \leq & \sqrt{8pL\|B^*(B - B^*)\|_F^2} \\ & \vee (8\|B^*(B - B^*)\|_{op} L)\end{aligned}$$

Following the same arguments as in Section B.1.1, it is sufficient to show that, for $j \geq m$ and $B \in \mathcal{B}_j$, $\|B^*(B - B^*)\|_F \lesssim \frac{1}{\sqrt{mj}}\delta_B$ and $\|B^*(B - B^*)\|_{op} \lesssim \frac{1}{j}\delta_B$.

1. $B^*(B - B^*)$ is a product of a projection and a difference of projections. That implies that all its eigen-values are bounded by 2 in absolute value. Hence: $\|B^*(B - B^*)\|_{op} \lesssim \frac{1}{j}\delta_B$.

2. All the coefficients of B^* are non-negative and bounded by $\frac{1}{m}$. Since the coefficients of B^*B and of B^*B^* are convex combinations of coefficients of B^*, it comes forward that all its coefficients are also bounded by $\frac{1}{m}$ in absolute value. That implies that the coefficients of $B^*(B - B^*)$ are bounded by $\frac{2}{m}$ in absolute value. Hence: $\|B^*(B - B^*)\|_F \lesssim \frac{1}{\sqrt{m}}\sqrt{\|B^*(B - B^*)\|_1} \lesssim \frac{1}{\sqrt{jm}}\delta_B$.

That concludes the proof for the term in $B^*(B - B^*)$.

For the term in $B^*(B - B^*)B^*$,we use the same arguments and add the fact that $\|B^*(B - B^*)B^*\|_1 = \|B^*(B - B^*)\|_1$ (see Lemma 1 of [7]).

B.2 Proof of Lemma 9

In this section, we bound the proportion of misclassified points

$$err(G, G^*) = \frac{1}{2n} \min_{\pi \in \mathcal{S}_K} \sum_{k=1}^{K} |G_k^* \Delta G_{\pi(k)}| \ ,$$

with respect to $\|B^* - B^*B\|_1$, for a given partition G and its normalized-partnership matrix B.

We consider G a partition. By Lemma 18, there exists a bijection $\phi : [1, K] \to [1, K]$ such that the following holds. For all $l \neq r$, if $\frac{|G_{\phi(l)} \cap G_r^*|}{|G_{\phi(l)}|} > \frac{1}{2}$, then for all l', $\frac{|G_{\phi(l')} \cap G_l^*|}{|G_{\phi(l')}|} \leq \frac{1}{2}$. Without loss of generality, we suppose that this bijection is the identity.

Then, by definition,

$$err(G, G^*) \leq \frac{1}{2n} \sum_{k=1}^{K} |G_k^* \Delta G_k| \ .$$

So, we have

$$err(G, G^*) \leq \frac{1}{n} \sum_{l \neq r} \sum_{i \in [1,n]} \mathbf{1}_{i \in G_r^* \cap G_l} \ .$$

Let $l \neq r$. If $\frac{|G_r^* \cap G_l|}{|G_l|} \leq \frac{1}{2}$, each $i \in G_r^* \cap G_l$ adds a contribution of at least 1 in $\|B^* - B^*B\|_1$. Moreover, if $\frac{|G_r^* \cap G_l|}{|G_l|} > \frac{1}{2}$, each point of G_l^* adds a contribution of at least 1 in $\|B^* - B^*B\|_1$. So, we can match any set $G_r^* \cap G_l$ such that $\frac{|G_r^* \cap G_l|}{|G_l|} > \frac{1}{2}$, to a group G_l^* which contains points all adding a contribution of at least 1 to $\|B^* - B^*B\|_1$. Since $|G_l^*| \geq m$ and $|G_r^* \cap G_l| \leq m^+$, $|G_l^*| \geq \frac{m}{m^+}|(G_r^* \cap G_l|$. Let us denote A_0 the set of points that are in some $G_r^* \cap G_l$, with $l \neq r$, satisfying $\frac{|G_r^* \cap G_l|}{|G_l|} > \frac{1}{2}$. We also denote A_1 the set of points that are in some $G_r^* \cap G_l$, satisfying $\frac{|G_r^* \cap G_l|}{|G_l|} \leq \frac{1}{2}$. We then have

$$\begin{aligned} |A_0| &= \sum_{l \neq r : \frac{|G_r^* \cap G_l|}{|G_l|} > \frac{1}{2}} |G_r^* \cap G_l| \\ &\leq \frac{m^+}{m} \sum_{l \neq r : \frac{|G_r^* \cap G_l|}{|G_l|} > \frac{1}{2}} |G_l^*| \ . \end{aligned}$$

Given $l \neq r$ such that $\frac{|G_r^* \cap G_l|}{|G_l|} > \frac{1}{2}$, the following assertion is satisfied: For all $l' \in [1, K]$, $\frac{|G_l^* \cap G_{l'}|}{|G_{l'}|} \leq \frac{1}{2}$. Hence, $|G_l^*| = \sum_{l' : \frac{|G_l^* \cap G_{l'}|}{|G_{l'}|} \leq \frac{1}{2}} |G_l^* \cap G_{l'}|$. Moreover, for all $l \in [1, K]$, there exists at most one index r such that $\frac{|G_r^* \cap G_l|}{|G_l|} > \frac{1}{2}$. Hence,

$$\begin{aligned} |A_0| \leq & \frac{m^+}{m} \sum_{l', l : \frac{|G_l^* \cap G_{l'}|}{|G_{l'}|} \leq \frac{1}{2}} |G_l^* \cap G_{l'}| \\ \leq & \frac{m^+}{m} \|B^* - B^*B\|_1 \ , \end{aligned}$$

where the last inequality comes from the fact that, for $l, l' \in [1, K]$, if $\frac{|G_l^* \cap G_{l'}|}{|G_{l'}|} \leq \frac{1}{2}$, each point of $|G_l^* \cap G_{l'}|$ adds a contribution of at least 1 to $\|B^* - B^*B\|$.

Since all $i \in A_1$ adds a contribution of at least 1 to $\|B^* - B^*B\|_1$, we have $|A_1| \leq \|B^* - B^*B\|_1$. This leads to $\sum_{l \neq r} \sum_{i \in [1,n]} \mathbf{1}_{i \in G_r^* \cap G_l} = |A_0| + |A_1| \leq \|B^* - B^*B\|_1 (1 + \frac{m^+}{m}) \leq 2\|B^* - B^*B\|_1 \frac{m^+}{m}$. Therefore, the proportion of misclustered points is upper-bounded by

$$err(G, G^*) \leq 2\frac{m^+}{m} \|B^* - B^*B\|_1 \ .$$

This concludes the proof of Lemma 9.

B.3 The signal term (Proof of Lemma 10)

The term $\langle A\mu(A\mu)^T, B^* - B \rangle$ is already dealt in [7]. We re-derive Lemma 10 for the sake of completeness. We denote $S \in \mathbb{R}^{n \times n}$ the matrix defined by $S_{ab} = -0.5\|\mu_k - \mu_l\|^2$, if $a \in G_k^*$ and $b \in G_l^*$. For $B \in \mathcal{C}$, we get

$$\begin{aligned} \langle A\mu(A\mu)^T, B^* - B \rangle &= \sum_{a,b \in [1,n]} \langle \mu_a, \mu_b \rangle (B_{ab}^* - B_{ab}) \\ &= \sum_{a,b \in [1,n]} \left(-0.5\|\mu_a - \mu_b\|^2 + 0.5\|\mu_a\|^2 + 0.5\|\mu_b\|^2 \right) (B_{ab}^* - B_{ab}) \\ &= \sum_{a,b \in [1,n]} -0.5\|\mu_a - \mu_b\|^2 (B_{ab}^* - B_{ab}) \\ &= \langle S, B^* - B \rangle \ , \end{aligned}$$

where the third equality comes for the fact that, for $a \in [1,n]$,

$$\sum_{b\in[1,n]} \|\mu_a\|^2 (B_{ab} - B^*_{ab}) = 0 \ .$$

The term S_{ab} being null when a and b are in the same group, we have

$$\begin{aligned}\langle S, B - B^*\rangle =& 0.5 \sum_{i\neq k} \sum_{a\in G_k^*} \sum_{b\in G_i^*} \|\mu_k - \mu_i\|^2 B_{ab} \\ \geq& \Delta^2 \sum_{i\neq k} \sum_{a\in G_k^*} \sum_{b\in G_i^*} B_{ab} \\ \geq& \Delta^2 \frac{1}{2} \|B^* - B^* B\|_1 \ ,\end{aligned}$$

where the last inequality comes from Lemma 1 of [7], which states that for any matrix B belonging to the larger class of matrix $\mathcal{C}$, we have the equality $\|B^* - B^* B\|_1 = 2\sum_{i\neq k}\sum_{a\in G_k^*}\sum_{b\in G_i^*} B_{ab}$.

This concludes the proof of Lemma 10.

C Proof of Theorem 3

In this section, we prove Theorem 3. We suppose that $p \geq c\log(K)$, for c a numerical constant that we will choose large enough later. Without loss of generality, we suppose throughout this proof that $\sigma = 1$. We suppose $n \geq 2K$, $K \geq K_0$, for K_0 a numerical constant that we will choose large enough, and $\alpha \geq \frac{3}{2}$.

Given ρ a probability distribution on $(\mathbb{R}^p)^K$ and a partition G of $[1,n]$ in K groups, we define the probability distribution on $(\mathbb{R}^p)^n$ by

$$\mathbb{P}_{\rho,G}(B) = \int \mathbb{P}_{\mu,G}(B) d\rho(\mu) \ .$$

To prove Theorem 3, we will use three lemmas. The first one, proved in Section C.3, is a consequence of Fano's lemma that we recall in Section C.3.

Lemma 19. *Let ρ be a probability measure on $\mathbb{R}^{p\times K}$. For any finite set $A \subset \mathcal{P}_\alpha$, any partition $G^{(0)} \in \mathcal{P}_\alpha$, we have the following inequality*

$$\inf_{\hat{G}} \frac{2}{|A|} \sum_{G\in A} \mathbb{E}_{\rho,G}[err(G,\hat{G})] \geq \min_{G\neq G'\in\mathcal{P}_\alpha} err(G,G') \left(1 - \frac{1 + \frac{1}{|A|}\sum_{G\in A} KL\left(\mathbb{P}_{\rho,G}, \mathbb{P}_{\rho,G^{(0)}}\right)}{\log|A|}\right).$$

The second lemma is a reduction lemma, which plays the same role as Lemma 26 in the proof of Theorem 4. We prove it in Section C.5.

Lemma 20. *Suppose that there exists a probability distribution ρ on $\mathbb{R}^{p\times K}$ and a numerical constant $C > 0$ satisfying*

$$\inf_{\hat{G}} \sup_{G\in\mathcal{P}_\alpha} \mathbb{E}_{\rho,G}\left[err(\hat{G},G)\right] - \rho(\mathbb{R}^{p\times K} \setminus \Theta_{\bar{\Delta}}) \geq C \ .$$

Then, we have

$$\inf_{\hat{G}} \sup_{\mu\in\Theta_{\bar{\Delta}}} \sup_{G\in\mathcal{P}_\alpha} \mathbb{E}_{\mu,G}\left[err(\hat{G},G)\right] \geq C \ .$$

Finally, the third lemma helps us choose the set A of partitions to whom we will apply Lemma 19. We prove it in Section C.6. We define $\overline{G}$ the partition of $[1,n]$ defined by; $i \in \overline{G}_k$ if and only if $i \equiv k \ [K]$. It is clear that $\overline{G} \in \mathcal{P}_{\frac{3}{2}}$.

Lemma 21. *We suppose that the constant K_0 such that $K \geq K_0$ is large enough and that $n \geq 2K$. There exists $S \subset \mathcal{P}_{\frac{3}{2}}$ which satisfies:*

- *There exists a numerical constant $c' > 0$ such that $\log|S| \geq c'' n \log(K)$,*
- *There exists a numerical constant $a > 0$ such that, for $G \neq G' \in S$, $err(G, G') \geq a$,*
- *For all $k \in [1, K]$, for all $G \in S$, $|G_k| = |\overline{G}_k|$.*

We distinguish two cases. In the first one, we suppose that there exists a numerical constant c_1, that we will choose small enough, such that $\bar{\Delta}^2 \leq c_1 \log(K)$. In the second one, we will suppose that $c_1 \log(K) \leq \bar{\Delta}^2 \leq c_2 \sqrt{\frac{p}{n} K \log(K)}$, for c_2 a numerical constant that we will also choose small enough.

C.1 Case $\bar{\Delta}^2 \leq c_1 \log(K)$

In this section, we suppose that $\bar{\Delta}^2 \leq c_1 \log(K)$. We suppose that the constant c such that $p \geq c \log(K)$ is larger than $4/\log(2)$, for having $\frac{p}{4}\log(2) \geq \log(K)$. Our choice of the centers $\mu_1, \ldots, \mu_K$ relies on the following lemma, proved in Section C.7.

Lemma 22. *If $\frac{p}{4}\log(2) \geq \log(K)$, there exists $\mu_1, \ldots, \mu_K$ in $\mathbb{R}^p$ such that $\frac{1}{2}\min_{l \neq r} \|\mu_l - \mu_r\|^2 \geq \bar{\Delta}^2$ and $\frac{1}{2}\max_{l \neq r} \|\mu_l - \mu_r\|^2 \leq 4\bar{\Delta}^2$.*

We consider $\mu = (\mu_1, \ldots, \mu_K)$ in $(\mathbb{R}^p)^K$ given by this lemma. In particular, we have $\mu \in \Theta_{\bar{\Delta}}$.

We recall that $\overline{G}$ is the partition of $[1, n]$ defined by; $i \in \overline{G}_k$ if and only if $i \equiv k \ [K]$. Given G taken in the set S defined in Lemma 21, let us compute $KL(\mathbb{P}_{\mu,G}, \mathbb{P}_{\mu,\overline{G}})$. We denote $\mathbb{P}_{\mu,G}(Y_i)$ the marginal law of Y_i under the joint law $\mathbb{P}_{\mu,G}$. By independence of all the Y_i, we have that

$$KL(\mathbb{P}_{\mu,G}, \mathbb{P}_{\mu,\overline{G}}) = \sum_{i=1}^{n} KL(\mathbb{P}_{\mu,G}(Y_i), \mathbb{P}_{\mu,\overline{G}}(Y_i)) \ .$$

Given $i \in [1, n]$, with k and l such that $i \in G_k \cap \overline{G}_l$, we have $KL(\mathbb{P}_{\mu,G}(Y_i), \mathbb{P}_{\mu,\overline{G}}(Y_i)) = \frac{\|\mu_k - \mu_l\|^2}{2}$. A fortiori, using Lemma 22,

$$KL(\mathbb{P}_{\mu,G}, \mathbb{P}_{\mu,\overline{G}}) \leq 4n\bar{\Delta}^2 \leq 4c_1 n \log(K) \ .$$

Applying this, together with $\log|S| \geq c'' n \log(K)$ and Lemma 19 with $\rho = \delta_{(\mu_1,\ldots,\mu_k)}$ leads to

$$\inf_{\hat{G}} \frac{2}{|S|} \sum_{G \in S} \mathbb{E}_{\mu,G}\left[err(\hat{G}, G)\right] \geq a\left(1 - \frac{1 + 4c_1^2 n \log(K)}{\log|S|}\right) \geq a\left(1 - \frac{1 + 4c_1^2 n \log(K)}{c'' n \log(K)}\right) \ .$$

The quantity $a\left(1 - \frac{1+4c_1^2 n \log(K)}{c'' n \log(K)}\right)$ being larger than $\frac{a}{2}$, supposing c_1 is small enough and K_0 large enough, there exists a constant C such that

$$\inf_{\hat{G}} \frac{2}{|S|} \sum_{G \in S} \mathbb{E}_{\mu,G}\left[err(\hat{G}, G)\right] \geq C \ .$$

A fortiori, since $\mu \in \Theta_{\bar{\Delta}}$ and $S \subset \mathcal{P}_\alpha$,

$$\inf_{\hat{G}} \sup_{G \in \mathcal{P}_\alpha} \sup_{\mu \in \Theta_{\bar{\Delta}}} \mathbb{E}_{\mu,G}\left[err(\hat{G}, G)\right] \geq C \ .$$

This concludes the proof of the theorem in this case.

C.2 Case $c_1 \log(K) \leq \bar{\Delta}^2 \leq c_2 \sqrt{\frac{p}{n} K \log(K)}$

In this section, we suppose that $c_1 \log(K) \leq \bar{\Delta}^2 \leq c_2 \sqrt{\frac{p}{n} K \log(K)}$, with c_2 a numerical constant that we will choose small enough. We still suppose that $K \geq K_0$.

Let us define ρ the uniform distribution on the hypercube $\mathcal{E} = \{-\varepsilon, +\varepsilon\}^{pK}$, where $\varepsilon = \sqrt{\frac{2}{p}}\bar{\Delta}$. We show in Section C.4 the following lemma, which controls $KL(\mathbb{P}_{\rho,G}, \mathbb{P}_{\rho,\overline{G}})$, for $G \in S$.

Lemma 23. *If c_2 is small enough with respect to c_1 and $c_1 \log(K) \leq \bar{\Delta}^2 \leq c_2 \sqrt{\frac{p}{n} K \log(K)}$, there exists a numerical constant $c > 0$ such that, for all $G \in S$, $KL(\mathbb{P}_{\rho,G}, \mathbb{P}_{\rho,\overline{G}}) \leq c c_2^2 n \log(K)$.*

Combining this lemma with Lemma 19 implies that

$$\inf_{\hat{G}} \frac{2}{|S|} \sum_{G \in S} \mathbb{E}_{\rho,G}\left[err(\hat{G}, G)\right] \geq \min_{G \neq G' \in S} err(G, G') \left(1 - \frac{1 + c c_2^2 n \log(K)}{c'' n \log(K)}\right) \ .$$

By definition of S in Lemma 21, given $G \neq G' \in S$, we have $err(G, G') \geq a$. Hence,

$$\inf_{\hat{G}} \frac{2}{|S|} \sum_{G \in S} \mathbb{E}_{\rho,G}\left[err(\hat{G}, G)\right] \geq a \left(1 - \frac{1 + c c_2^2 n \log(K)}{c'' n \log(K)}\right) \ .$$

This last quantity is larger than $\frac{a}{2}$ provided c_2 is small enough and K_0 large enough. Hence, since $S \subset \mathcal{P}_\alpha$, this implies

$$\inf_{\hat{G}} \sup_{G \in \mathcal{P}_\alpha} \mathbb{E}_{\rho,G}\left[err(\hat{G}, G)\right] \geq \frac{a}{2} \ .$$

In order to apply Lemma 20, it remains to upper-bound the quantity $\rho(\mathbb{R}^{p \times K} \setminus \theta_{\bar{\Delta}})$. The next lemma, proved in Section C.8, provides such an upper-bound.

Lemma 24. *For all $K \geq 2$, $p > 1$ and $\bar{\Delta} > 0$, the probability distribution ρ satisifes*

$$\rho(\mathbb{R}^{p \times K} \setminus \Theta_{\bar{\Delta}}) \leq \frac{K(K-1)}{2} e^{-p/8} \ .$$

In particular, if the constant c such that $p \geq c \log(K)$ is large enough, the quantity $\frac{K(K-1)}{2} e^{-p/8}$ is smaller than $a/4$. This, together with Lemma 20, leads to the existence of a numerical constant $C > 0$ such that

$$\inf_{\hat{G}} \sup_{\mu \in \Theta_{\bar{\Delta}}} \sup_{G \in \mathcal{P}_\alpha} \mathbb{E}_{\mu,G}\left[err(\hat{G}, G)\right] > C \ .$$

This concludes the proof of the Theorem 3.

C.3 Proof of Lemma 19

Lemma 19 is a simple derivation from Fano's Lemma that we recall here (see e.g [31]).

Lemma 25 (Fano's Lemma)**.** *Let $(\mathbb{P}_j)_{j \in [1,M]}$ be a set of probability distributions on some set $\mathcal{Y}$. For any probability distribution $\mathbb{Q}$ such that for all $j \in [1, M]$, $\mathbb{P}_j << \mathbb{Q}$,*

$$\inf_{\hat{J}: \mathcal{Y} \to [1,M]} \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}_j\left(\hat{J}(Y) \neq j\right) \geq 1 - \frac{1 + \frac{1}{M} \sum_{j=1}^{M} KL(\mathbb{P}_j, \mathbb{Q})}{\log(M)} \ ,$$

where we recall that $KL(\mathbb{P}, \mathbb{Q}) = \int \log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{P}$ stands for the Kullback-Leibler divergence between $\mathbb{P}$ and $\mathbb{Q}$.

Let A be a subset of $\mathcal{P}_\alpha$. Denote $G^{(1)}, \dots, G^{(|A|)}$ the elements of A. Given any estimator $\hat{G}$, we denote $\hat{j}$ an index that minimises $err(\hat{G}, G^{(j)})$.

For any $j \in [1, |A|]$, using the definition of $\hat{j}$ together with the fact that the function err satisfies the triangular inequality (for the second inequality), we have

$$\begin{aligned}\min_{i \neq i'} err(G^{(i)}, G^{(i')}) \mathbf{1}_{\hat{j} \neq j} &\leq err(G^{(\hat{j})}, G^{(j)}) \\ &\leq err(\hat{G}, G^{(\hat{j})}) + err(\hat{G}, G^{(j)}) \\ &\leq 2 err(\hat{G}, G^{(j)}) \ .\end{aligned}$$

Applying the expectation to this last inequality, we have, for $j \in [1, |A|]$,

$$2\, \mathbb{E}_{\rho, G^{(j)}} \left[err(\hat{G}, G^{(j)}) \right] \geq \min_{i \neq i'} err(G^{(i)}, G^{(i')})\, \mathbb{P}_{\rho, G^{(j)}} \left[\hat{j} \neq j \right] \ .$$

Summing over $G \in A$ leads to

$$\frac{2}{|A|} \sum_{G \in A} \mathbb{E}_{\rho, G} \left[err(\hat{G}, G) \right] \geq \min_{G, G' \in A} err(G, G') \min_{\hat{G}} \frac{1}{|A|} \sum_{G^{(j)} \in A} \mathbb{P}_{\rho, G^{(j)}} \left[\hat{j} \neq j \right].$$

Applying Fano's Lemma 25, we get the sought inequality

$$\inf_{\hat{G}} \frac{2}{|A|} \sum_{G \in A} \mathbb{E}_{\rho, G} \left[err(\hat{G}, G) \right] \geq \min_{G, G' \in A} err(G, G') \left(1 - \frac{1 + \frac{1}{|A|} \sum_{G \in A} KL(\mathbb{P}_{\rho, G}, \mathbb{P}_{\rho, G^{(0)}})}{\log(|A|)} \right).$$

This concludes the proof of the Lemma 19 .

C.4 Proof of Lemma 23

For $G \in S$, let us compute $KL(\mathbb{P}_{\rho, G}, \mathbb{P}_{\rho, \overline{G}})$. We recall that $i \in \overline{G}_k$ if and only if $i \equiv k \ [K]$. From Lemma 21, we have that, for $k \in [1, K]$, $|G_k| = |\overline{G}_k|$. We define m_k this quantity. Given $k, l \in [1, K]$, we write $m_{kl} = |\overline{G}_k \cap G_l|$.

We recall the definition $KL(\mathbb{P}_{\rho, G}, \mathbb{P}_{\rho, \overline{G}}) = \int \log \left(\frac{d\,\mathbb{P}_{\rho, G}}{d\,\mathbb{P}_{\rho, \overline{G}}} \right) d\,\mathbb{P}_{\rho, G}$. We write $\mathbb{P}_{0, G}$, or equivalentely $\mathbb{P}_{0, \overline{G}}$, the distribution of $(Y_i)_{i \in [1, n]}$ when $\mu_1 = \dots, \mu_K = 0$ almost surely. Under this distribution, the Y_i's are drawn independently according $\mathbb{N}(0, I_p)$. First, we will compute the quantity

$$\frac{d\,\mathbb{P}_{\rho, G}}{d\,\mathbb{P}_{\rho, \overline{G}}} = \frac{\frac{d\,\mathbb{P}_{\rho, G}}{d\,\mathbb{P}_{0, G}}}{\frac{d\,\mathbb{P}_{\rho, \overline{G}}}{d\,\mathbb{P}_{0, G}}} = \frac{\frac{d\,\mathbb{P}_{\rho, G}}{d\,\mathbb{P}_{0, G}}}{\frac{d\,\mathbb{P}_{\rho, \overline{G}}}{d\,\mathbb{P}_{0, \overline{G}}}} \ , \tag{28}$$

where the second equality comes from the fact that $\mathbb{P}_{0, G} = \mathbb{P}_{0, \overline{G}}$. Given a probability distribution $\mathbb{P}$ on some Euclidean space, which is absolutely continuous with respect to the Lebesgue measure, we write $d\,\mathbb{P}$ for the density of this distribution with respect to the Lebesgue measure. For the numerator in (28), we have

$$\begin{aligned}\frac{d\,\mathbb{P}_{\rho, G}}{d\,\mathbb{P}_{0, G}}(Y) =& \frac{\mathbb{E}_\rho \left[d\,\mathbb{P}_{\mu, G}(Y) \right]}{\mathbb{E}_0 \left[d\,\mathbb{P}_{\mu, G}(Y) \right]} \\ =& \frac{\mathbb{E}_\rho \left[\prod_{k \in [1, K]} \prod_{i \in G_k} \exp\left(-\frac{1}{2} \|Y_i - \mu_k\|^2 \right) \right]}{\mathbb{E}_0 \left[\prod_{k \in [1, K]} \prod_{i \in G_k} \exp\left(-\frac{1}{2} \|Y_i - \mu_k\|^2 \right) \right]} \\ =& \frac{\mathbb{E}_\rho \left[\prod_{d \in [1, p]} \prod_{k \in [1, K]} \prod_{i \in G_k} \exp\left(-\frac{1}{2} \left(Y_{i, d} - \mu_{k, d} \right)^2 \right) \right]}{\prod_{d \in [1, p]} \prod_{k \in [1, K]} \prod_{i \in G_k} \exp\left(-\frac{1}{2} Y_{i, d}^2 \right)} \ .\end{aligned}$$

Using the independence of the $\mu_{k,d}$'s under the law ρ, we get that

$$\begin{aligned}\frac{d\,\mathbb{P}_{\rho,G}}{d\,\mathbb{P}_{0,G}}(Y) &= \prod_{d\in[1,p]}\prod_{k\in[1,K]}\frac{\mathbb{E}_\rho\left[\prod_{i\in G_k}\exp\left(-\frac{1}{2}(Y_{i,d}-\mu_{k,d})^2\right)\right]}{\prod_{i\in G_k}\exp\left(-\frac{1}{2}Y_{i,d}^2\right)}\\ &= \prod_{d\in[1,p]}\prod_{k\in[1,2]}\mathbb{E}_\rho\left[\prod_{i\in G_k}\exp\left(-\frac{1}{2}\left((Y_{i,d}-\mu_{k,d})^2-(Y_{i,d})^2\right)\right)\right]\\ &= \prod_{d\in[1,p]}\prod_{k\in[1,K]}\mathbb{E}_\rho\left[\prod_{i\in G_k}\exp\left(Y_{i,d}\mu_{k,d}-\frac{\varepsilon^2}{2}\right)\right]\\ &= \prod_{d\in[1,p]}\prod_{k\in[1,K]}e^{\frac{-m_k\varepsilon^2}{2}}\cosh\left(\sum_{i\in G_k}\varepsilon Y_{i,d}\right)\ .\end{aligned}$$

Similarly, we have

$$\frac{d\,\mathbb{P}_{\rho,\overline{G}}}{d\,\mathbb{P}_{0,\overline{G}}}(Y) = \prod_{d\in[1,p]}\prod_{k\in[1,K]}e^{\frac{-m_k\varepsilon^2}{2}}\cosh\left(\sum_{i\in\overline{G}_k}\varepsilon Y_{i,d}\right)\ .$$

Combining these two equalities in (28), we end up with

$$\frac{d\,\mathbb{P}_{\rho,G}}{d\,\mathbb{P}_{\rho,\overline{G}}}(Y) = \prod_{d=1}^{p}\prod_{k\in[1,K]}\frac{\cosh\left(\sum_{i\in G_k}\varepsilon Y_{i,d}\right)}{\cosh\left(\sum_{i\in\overline{G}_k}\varepsilon Y_{i,d}\right)}\ . \tag{29}$$

We denote ϕ the standard Gaussian density $\phi(x)=\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}$. We recall that m_k is the size of G_k (and also of $|\overline{G}_k|$) and $m_{kl}=|\overline{G}_k\cap G_l|$. Under the law $\mathbb{P}_{\rho,G}$, conditionally on $\mu_1,\dots\mu_K\sim\rho$, we have that:

- $\sum_{i\in G_k}Y_{i,d}\sim\mathcal{N}(m_k\mu_{k,d},m_k)$,
- $\sum_{i\in\overline{G}_k}Y_{i,d}\sim\mathcal{N}(\sum_{l\in[1,K]}m_{kl}\mu_l,m_k)$.

Plugging these two points, together with equality (29), in the definition of the Kullback-Leibler divergence leads to

$$\begin{aligned}\frac{1}{p}KL(\mathbb{P}_{\rho,G},\mathbb{P}_{\rho,\overline{G}}) &= \sum_{k\in[1,K]}\mathbb{E}_\rho\left[\int\log\cosh(\varepsilon(m_k\mu_{k,1}+\sqrt{m_k}x))\phi(x)dx\right]\\ &\quad-\sum_{k\in[1,K]}\mathbb{E}_\rho\left[\int\log\cosh(\varepsilon(\sum_{l\in[1,K]}m_{kl}\mu_{l,1}+\sqrt{m_k}x))\phi(x)dx\right]\ .\end{aligned}\tag{30}$$

By symmetry, it is sufficient to upper-bound the term corresponding to the first group $\overline{G}_1$ in the sum above which is equal to $S_1=S_{11}-S_{12}$, with

$$\begin{aligned}S_{11} &= \mathbb{E}_\rho\left[\int\log\cosh(\varepsilon(m_1\mu_{1,1}+\sqrt{m_1}x))\phi(x)dx\right],\\ \text{and}\quad S_{12} &= \mathbb{E}_\rho\left[\int\log\cosh(\varepsilon(\sum_{l\in[1,K]}m_{1l}\mu_{l,1}+\sqrt{m_1}x))\phi(x)dx\right].\end{aligned}$$

First, we upper-bound the term S_{11}. We will use the following inequality

$$\frac{x^2}{2}-\frac{x^4}{12}\le\log\cosh(x)\le\frac{x^2}{2},\ \ \forall x\in\mathbb{R}\ . \tag{31}$$

Proof of inequality (31). Let us first prove the upper-bound $\log\cosh(x) \leq \frac{x^2}{2}$. For $x \in \mathbb{R}$, $\cosh(x) = \sum_{t\in\mathbb{N}} \frac{x^{2t}}{(2t)!} \leq \sum_{t\in\mathbb{N}} \frac{x^{2t}}{t!2^t} = \exp\left(\frac{t^2}{2}\right)$. Applying the logarithmic function leads to $\log\cosh(x) \leq \frac{x^2}{2}$.

Let us now prove the lower-bound $\log\cosh(x) \geq \frac{x^2}{2} - \frac{x^4}{12}$. To do so, we write, for $x \geq 0$, $f(x) = \tanh(x)$ and $g(x) = x - \frac{x^3}{3}$. We have $f'(x) = 1 - (f(x))^2$. Besides, $g'(x) = 1 - x^2 \leq 1 - (g(x))^2$. Together with the fact that $f(0) = g(0) = 0$, this implies

$$f(x) \geq g(x), \ \ \forall x \geq 0 \ .$$

Integrating these function leads to

$$\log\cosh(x) \geq \frac{x^2}{2} - \frac{x^4}{12}, \ \ \forall x \geq 0 \ .$$

By parity of these functions, this last inequality is satisfied for all $x \in \mathbb{R}$. □

Inequality (31), together with the independence of x and $\mu_{1,1}$, imply that

$$S_{11} \leq \frac{1}{2}\,\mathbb{E}_\rho\left[\int (\varepsilon(m_1\mu_{1,1} + \sqrt{m_1}x))^2\phi(x)dx\right] \leq \frac{1}{2}\varepsilon^2\left(m_1 + m_1^2\varepsilon^2\right) \ . \tag{32}$$

We arrive at

$$S_{11} \leq \frac{1}{2}m_1\varepsilon^2(1 + m_1\varepsilon^2) \ . \tag{33}$$

Let us now lower-bound S_{12}. Inequality (31) induces

$$\begin{aligned}
S_{12} \geq & \frac{1}{2}\,\mathbb{E}_\rho\left[\int \varepsilon^2\Big(\sum_{l\in[1,K]} m_{1l}\mu_{l,1} + \sqrt{m_1}x\Big)^2\phi(x)dx\right] \\
& - \frac{1}{12}\,\mathbb{E}_\rho\left[\int \varepsilon^4\Big(\sum_{l\in[1,K]} m_{1l}\mu_{l,1} + \sqrt{m_1}x\Big)^4\phi(x)dx\right] \\
\geq & \frac{1}{2}\varepsilon^2\left(\sum_{l\in[1,K]} m_{1l}^2\varepsilon^2 + m_1\right) - \frac{1}{12}\varepsilon^4\,\mathbb{E}_\rho\left[\Big(\sum_{l\in[1,K]} m_{1l}\mu_{l,1}\Big)^4\right] - \frac{1}{12}\varepsilon^4 3m_1^2 \\
& - \frac{1}{2}\varepsilon^4 m_1\,\mathbb{E}_\rho[(\sum_{l\in[1,K]} m_{1l}\mu_{l,1})^2] \\
\geq & \frac{1}{2}\varepsilon^2\left(\sum_{l\in[1,K]} m_{1l}^2\varepsilon^2 + m_1\right) - \frac{1}{12}\varepsilon^8\left(\sum_{l\in[1,K]} m_{1l}^4 + 6\left(\sum_{l\in[1,K]} m_{1l}^2\right)^2\right) - \frac{1}{4}\varepsilon^4 m_1^2 \\
& - \frac{1}{2}\varepsilon^6 m_1\sum_{l\in[1,K]} m_{1l}^2 \ .
\end{aligned}$$

We end up with

$$S_{12} \geq \frac{1}{2}m_1\varepsilon^2 - \frac{1}{4}m_1^2\varepsilon^4 - \frac{1}{2}\varepsilon^6 m_1\sum_{l\in[1,K]} m_{1,l}^2 - \frac{1}{12}\varepsilon^8\left(\sum_{l\in[1,K]} m_{1l}^4 + 6\left(\sum_{l\in[1,K]} m_{1l}^2\right)^2\right) \ . \tag{34}$$

Combining inequalities (33) and (34), together with the equality $m_1 = \sum_{l\in[1,K]} m_{1l}$, leads to

$$\begin{aligned}
S_1 \leq & \frac{1}{2}m_1\varepsilon^2(1+m_1\varepsilon^2) - \frac{1}{2}m_1\varepsilon^2 + \frac{1}{4}m_1^2\varepsilon^4 + \frac{1}{2}\varepsilon^6 m_1 \sum_{l\in[1,K]} m_{1,l}^2 \\
& + \frac{1}{12}\varepsilon^8 \left(\sum_{l\in[1,K]} m_{1l}^4 + 6\left(\sum_{l\in[1,K]} m_{1l}^2 \right)^2 \right) \\
\leq & \frac{3}{4}m_1^2\varepsilon^4 + \frac{7}{12}\varepsilon^8 m_1^4 + \frac{1}{2}\varepsilon^6 m_1^3 \\
\leq & c_4 \frac{n^2}{K^2}\varepsilon^4 \left(1 + \varepsilon^2 \frac{n}{K} + \varepsilon^4 \frac{n^2}{K^2} \right) \ ,
\end{aligned}$$

where c_4 is a numerical constant, obtained using the fact that $m_1 \leq \frac{3}{2}\frac{n}{K}$. Summing over $k \in [1,K]$ in (30) leads to

$$KL(\mathbb{P}_{\rho,G}, \mathbb{P}_{\rho,\overline{G}}) \leq c_4 p \frac{n^2}{K}\varepsilon^4 \left(1 + \varepsilon^2 \frac{n}{K} + \varepsilon^4 \frac{n^2}{K^2} \right) \ .$$

From the definition of ε, we have $p\frac{n^2}{K}\varepsilon^4 = 4\frac{n^2}{Kp}\bar{\Delta}^4$. The hypothesis $\bar{\Delta}^2 \leq c_2\sqrt{\frac{p}{n}K\log(K)}$ leads to $p\frac{n}{K}\varepsilon^4 \leq 4c_2^2 n\log(K)$. Moreover, the hypothesis $c_1\log(K) \leq \bar{\Delta}^2 \leq c_2\sqrt{\frac{p}{n}K\log(K)}$ implies $\varepsilon^2\frac{n}{K} = \frac{2n}{pK}\frac{\bar{\Delta}^4}{\bar{\Delta}^2} \leq 2\frac{c_2^2}{c_1} \leq 1$, if c_2 is small enough with respect to c_1. Thus, there exists a numerical constant $c > 0$ such that, when c_2 is small enough,

$$KL(\mathbb{P}_{\rho,G}, \mathbb{P}_{\rho,\overline{G}}) \leq cc_2^2 n\log(K) \ .$$

This concludes the proof of the lemma.

C.5 Proof of Lemma 20

We suppose that there exists a probability distribution ρ on $\mathbb{R}^{P\times K}$ and $C > 0$ satisfying

$$\inf_{\hat{G}} \sup_{G\in\mathcal{P}_\alpha} \mathbb{E}_{\rho,G}\left[err(\hat{G}, G) \right] - \rho(\mathbb{R}^{p\times K} \setminus \Theta_{\bar{\Delta}}) \geq C \ .$$

Given an estimator $\hat{G}$, there exists a partition $G \in \mathcal{P}_\alpha$ such that $\mathbb{E}_{\rho,G}\left[err(\hat{G}, G) \right] - \rho(\mathbb{R}^{p\times K} \setminus \Theta_{\bar{\Delta}}) \geq C$. Since $\mathbb{E}_{\rho,G}\left[err(\hat{G}, G) \right] = \int_{\mathbb{R}^{p\times K}} \mathbb{E}_{\mu,G}\left[err(\hat{G}, G) \right] d\rho(\mu)$ and $\mathbb{E}_{\mu,G}\left[err(\hat{G}, G) \right]$ is upper bounded by 1, we end up with $\int_{\Theta_{\bar{\Delta}}} \mathbb{E}_{\mu,G}\left[err(\hat{G}, G) \right] d\rho(\mu) \geq C$.

This implies the existence of $\mu \in \Theta_{\bar{\Delta}}$ such that $\mathbb{E}_{\mu,G}\left[err(\hat{G}, G) \right] \geq C$. A fortiori,

$$\sup_{\mu\in\Theta_{\bar{\Delta}}} \sup_{G\in\mathcal{P}_\alpha} \mathbb{E}_{\mu,G}\left[err(\hat{G}, G) \right] \geq C \ .$$

This last inequality being true for all estimator $\hat{G}$, we get the sought inequality

$$\inf_{\hat{G}} \sup_{\mu\in\Theta_{\bar{\Delta}}} \sup_{G\in\mathcal{P}_\alpha} \mathbb{E}_{\mu,G}\left[err(\hat{G}, G) \right] \geq C \ .$$

C.6 Proof of Lemma 21

We recall that $\overline{G}$ is the partition of $[1,n]$ which is defined by $i \in \overline{G}_k$ if and only if $i \equiv k\ [K]$. Throughout the proof of this lemma, we denote $m = \lfloor \frac{n}{K} \rfloor$. We define $V \subset \mathcal{P}_{\frac{3}{2}}$, the set of partitions G which satisfy

- the restriction of G to $[1, K\lfloor \frac{m}{2} \rfloor] \cup [Km+1, n]$ is equal to the restriction of $\overline{G}$ on the same set,
- $|G_k| = |\overline{G}_k|$ for all $k \in [1, K]$.

The number of partitions in V is equal to the number of partitions $[K\lfloor \frac{m}{2} \rfloor + 1, Km]$ in K groups of size $m - \lfloor \frac{m}{2} \rfloor$. For the $\lceil \frac{K}{2} \rceil$ first groups, we choose $m - \lfloor \frac{m}{2} \rfloor$ elements amongst at least $\lfloor \frac{K}{2} \rfloor (m - \lfloor \frac{m}{2} \rfloor)$ elements. Hence, the number of such partitions is lower-bounded by $\binom{\lfloor \frac{K}{2} \rfloor (m - \lfloor \frac{m}{2} \rfloor)}{m - \lfloor \frac{m}{2} \rfloor}^{\lceil \frac{K}{2} \rceil}$. We arrive at $\log |V| \geq \frac{K}{2} \log \binom{\lfloor \frac{K}{2} \rfloor (m - \lfloor \frac{m}{2} \rfloor)}{m - \lfloor \frac{m}{2} \rfloor}$. Besides, if the constant K_0 is large enough, we get that, when $K \geq K_0$, $\binom{\lfloor \frac{K}{2} \rfloor (m - \lfloor \frac{m}{2} \rfloor)}{m - \lfloor \frac{m}{2} \rfloor} \geq \left(\frac{Km}{8}\right)^{m - \lfloor \frac{m}{2} \rfloor} \frac{1}{\left(m - \lfloor \frac{m}{2} \rfloor\right)!}$. Besides, $\log \left(m - \lfloor \frac{m}{2} \rfloor\right)! \leq \left(m - \lfloor \frac{m}{2} \rfloor\right) \log \left(m - \lfloor \frac{m}{2} \rfloor\right)$. We arrive at $\log \left(\binom{\lfloor \frac{K}{2} \rfloor (m - \lfloor \frac{m}{2} \rfloor)}{m - \lfloor \frac{m}{2} \rfloor} \right) \geq \frac{m}{2} \left(\log \left(\frac{Km}{8} \right) - \log \left(m - \lfloor \frac{m}{2} \rfloor \right) \right)$. This implies the existence of a numerical constant $c_3 > 0$ such that, if $K \geq K_0$ with K_0 large enough,

$$\log |V| \geq c_3 n \log(K) \ . \tag{35}$$

Let S be a maximal subset of V satisfying; for all $G, G' \in S$, the proportion of misclassified points between these two partitions $err(G, G')$ is lower-bounded by a, with $a > 0$ a numerical constant that we will choose small enough. As a consequence, for all $G \in V$, there exists $G' \in S$ such that $err(G, G') \leq a$. For $G \in S$, we denote $B_G \subset V$ the set of partitions $G' \in V$ that satisfy $err(G, G') \leq a$. Then,

$$|S| \geq \frac{|V|}{\max_{G \in S} |B_G|} \ . \tag{36}$$

Given $G \in S$, let us upper-bound $|B_G|$. For $G' \in V$, we define $E(G') = \{i \in [1, n], \ \exists k \neq l \in [1, K], \ i \in G_k \cap G'_l\} \subset [K\lfloor \frac{m}{2} \rfloor + 1, Km]$. We recall the definition

$$err(G, G') = \min_{\pi \in \mathcal{S}_K} \frac{1}{2n} \sum_{k=1}^{K} |G_k \Delta G'_{\pi(k)}| \ .$$

For $k \in [1, K]$, if $\pi(k) \neq k$, then $|G_k \Delta G'_{\pi(k)}| \geq 2 |\lfloor \frac{m}{2} \rfloor| \geq |G'_k \cap E(G')|$. If $\pi(k) = k$, then $|G_k \Delta G'_{\pi(k)}| \geq |G'_k \cap E(G')|$. Plugging this in the definition of $err(G, G')$ leads to

$$err(G, G') \geq \frac{1}{2n} E(G') \ .$$

Hence, if $G' \in B_G$, then $E(G') \leq 2an$. For choosing a partition G' such that $E(G') \leq 2an$, we choose $2an$ points amongst $[K\lfloor \frac{m}{2} \rfloor + 1, Km]$ points and, for each of these points, we choose the group G'_k which will contain it. We arrive at $|\{G' \in V, \ E(G') \leq 2an\}| \leq \binom{K(m - \lfloor \frac{m}{2} \rfloor)}{2an} K^{2an}$. We end up with $\log |B_G| \leq 2an \log(K) + \log \binom{K(m - \lfloor \frac{m}{2} \rfloor)}{2an} \leq 2an \log(K) + 2an \log \left(\frac{1}{2a} \right)$. Combining this last inequality with inequalities (35) and (36) leads to

$$\log |S| \geq c_3 n \log(K) - 2an \log(K) - 2an \log(\frac{1}{2a}) \ .$$

If the constant a is small enough and the constant K_0 large enough, we have, when $K \geq K_0$, and a fortiori $n \geq 2K_0$, $\frac{c_3}{2} n \log(K) - 2an \log(\frac{1}{2a}) \geq \frac{c_3 n}{4} \log(K)$.

Besides, if a is small enough, we also have $\frac{c_3}{2} n \log(K) - 2an \log(K) \geq \frac{c_3 n}{4} \log(K)$. This leads to the sought inequality

$$\log |S| \geq \frac{c_3 n}{2} \log(K) \ .$$

For $G \in S$ and $k \in [1, K]$, we have $|G_k| = |\overline{G}_k| \in [\lfloor \frac{n}{K} \rfloor, \lfloor \frac{n}{K} \rfloor + 1]$. Hence, $S \subset \mathcal{P}_{\frac{3}{2}}$. By definition of S, for $G \neq G' \in S$, we have the lower-bound $err(G, G') \geq a$. Hence, the set S satisifies all the conditions of Lemma 21. This concludes the proof of the lemma.

C.7 Proof of Lemma 22

We suppose that $\frac{p}{4}\log(2) \geq \log(K)$ and we want to find vectors $\mu_1, \dots, \mu_K$ in $\mathbb{R}^p$ such that $\frac{1}{2}\min_{l\neq r}\|\mu_r - \mu_l\|^2 \geq \bar{\Delta}^2$ and $\frac{1}{2}\max_{l\neq r}\|\mu_l - \mu_r\|^2 \leq 4\bar{\Delta}^2$.

Denote $\mathcal{H} = \bar{\Delta}\sqrt{\frac{2}{p}}\{-1,1\}^p$. There are 2^p elements in $\mathcal{H}$. Consider H a maximal subset of $\mathcal{H}$ such that the following holds. For θ and θ' two distinct elements of H, there exists at least $\frac{p}{4}$ index such that $\theta_i \neq \theta'_i$.

If θ and θ' are two distinct elements of H, we have from the definitions of $\mathcal{H}$ and H that

$$\bar{\Delta}^2 \leq \frac{1}{2}\|\theta - \theta'\|^2 \leq 4\bar{\Delta}^2 \ .$$

It remains to lower-bound the cardinality of H. Given $\theta \in H$, denote B_θ the subset of $\mathcal{H}$ made of the elements θ' that have at least $\lfloor \frac{3p}{4} \rfloor$ index such that $\theta_i = \theta'_i$. From the definition of H, we deduce that $\mathcal{H} \subset \bigcup_{\theta\in H} B_\theta$. Since the B_θ's are all of the same cardinality, we get that, for $\theta \in H$, $|\mathcal{H}| \leq |H||B_\theta|$.

Let us upper-bound $|B_\theta|$. Given a fixed set of index of cardinal $\lfloor \frac{3p}{4} \rfloor$, there are at most $2^{\frac{p}{4}}$ points in $\mathcal{H}$ that are equal to θ on these index. Hence, the cardinal of B_θ is upper-bounded by $\binom{p}{\lfloor \frac{3p}{4} \rfloor}2^{\frac{p}{4}}$. Plugging this inequality in $|\mathcal{H}| \leq |H||B_\theta|$ leads to $\log(|H|) \geq p\log(2) - \log\binom{p}{\lfloor \frac{3p}{4} \rfloor} - \frac{p}{4}\log(2)$. Using the inequality $\log\binom{p}{\lfloor \frac{3p}{4} \rfloor} \leq \log\binom{p}{\lfloor \frac{p}{2} \rfloor} \leq \frac{p}{2}\log(2)$, we end up with $\log(|H|) \geq \frac{p}{4}\log(2)$. This implies from our hypothesis on p that $|H| \geq K$. Any K-tuple of vectors $\mu_1, \dots, \mu_K$ taken from $|H|$ satisfies the sought conditions. This concludes the proof of the lemma.

C.8 Proof of Lemma 24

Let $p > 1$ and $K \geq 2$. We lower bound in this section $\min_{k\neq l}\|\mu_k - \mu_l\|$ when $\mu \sim \rho$. Let $k \neq l \in [1, K]$. Then, $\|\mu_k - \mu_l\|^2 = 4\varepsilon^2 \sum_{d=1}^p \mathbf{1}_{\mu_{k,d}\neq\mu_{l,d}}$. Hence, $\frac{1}{4\varepsilon^2}\|\mu_k - \mu_l\|^2$ is a sum of p independent Bernoulli random variables of parameter $\frac{1}{2}$. In particular, $\frac{1}{4\varepsilon^2}\|\mu_k - \mu_l\|^2$ is the sum of p independent random variables of mean $\frac{1}{2}$ and bounded in absolute value by 1. Using Hoeffding's inequality leads to

$$\begin{aligned}
\mathbb{P}\left[\frac{1}{4\varepsilon^2}\|\mu_k - \mu_l\|^2 \leq \frac{p}{4}\right] &= \mathbb{P}\left[\frac{1}{4\varepsilon^2}\|\mu_k - \mu_l\|^2 - \mathbb{E}\left[\frac{1}{4\varepsilon^2}\|\mu_k - \mu_l\|^2\right] \leq -\frac{p}{4}\right] \\
&\leq \exp\left(\frac{-2\left(\frac{p}{4}\right)^2}{p}\right) \\
&\leq \exp\left(-\frac{p}{8}\right) \ .
\end{aligned}$$

Besides, $\frac{1}{4\varepsilon^2}\|\mu_k - \mu_l\|^2 = \frac{p}{8\bar{\Delta}^2}\|\mu_k - \mu_l\|^2$. Hence,

$$\mathbb{P}\left[\frac{1}{2}\|\mu_k - \mu_l\|^2 \leq \bar{\Delta}^2\right] \leq \exp\left(-\frac{p}{8}\right) \ .$$

Using an union bound on the set of pairs $k \neq l \in [1, K]$ leads to

$$\mathbb{P}[\exists k \neq l \in [1,K],\ \frac{1}{2}\|\mu_k - \mu_l\|^2 \leq \bar{\Delta}^2] \leq \frac{K(K-1)}{2}\exp\left(-\frac{p}{8}\right) \ .$$

This concludes the proof of the lemma.

D Proof of Theorem 4

Without loss of generality, we suppose throughout this proof that $\sigma = 1$. We suppose that $n \geq n_0$, with n_0 a constant that we will choose large enough. We also suppose that $\alpha \geq \frac{3}{2}$ and $n \geq 9K/2$.

As in Section C, we will distinguish two cases. In a first time, we will prove that there exists numerical constants c_1 and C such that when $\bar{\Delta}^2 \leq c_1 \log(n)$,

$$\inf_{\hat{G}} \sup_{\mu \in \Theta_{\bar{\Delta}}} \sup_{G \in \mathcal{P}_\alpha} \mathbb{P}_{\mu,G}(\hat{G} \neq G) > C \ .$$

Then, in a second case, we will show that there exists a numerical constant c_2, such that this also holds when $c_1 \log(n) \leq \bar{\Delta}^2 \leq c_2 \sqrt{\frac{pK}{n} \log(n)}$.

In the following of this proof, we consider a partition $G^{(0)} \in \mathcal{P}_{\frac{3}{2}}$. In particular, $G^{(0)} \in \mathcal{P}_\alpha$. The existence of such a partition is ensured by the hypothesis $n \geq 2K$ (for example, we can take the partition $\overline{G}$ defined in the proof of Theorem 3 by $i \in \overline{G}_k$ if and only if $i \equiv k \ [K]$).

D.1 $\bar{\Delta}^2 \leq c_1 \log(n)$.

Let us suppose that $\bar{\Delta}^2 \leq c_1 \log(n)$ for c_1 a positive numerical constant, that we will choose small enough later.

Let e be a unit vector of $\mathbb{R}^p$ and define $\mu_k = \sqrt{2} k \bar{\Delta} e$ for $k \in [1, K]$. It is clear that $\mu = \mu_1, \ldots, \mu_K \in \Theta_{\bar{\Delta}}$. We will prove our statement using Fano's Lemma that is recalled page 37. Our goal will be to find different partitions $G^{(1)}, ..., G^{(M)} \in \mathcal{P}_\alpha$, with M as large as possible, such that $KL(\mathbb{P}_{\mu,G^{(r)}}, \mathbb{P}_{\mu,G^{(0)}})$ remains small for all $r \in [1, M]$.

Given $k \in [1, K-1]$, $i \in G^{(0)}_k$ and $j \in G^{(0)}_{k+1}$, we define the partition $G^{(i,j)}$ as follows. For $l \in [1, K]$ distinct both from k and $k+1$, we take $G^{(i,j)}_l = G^{(0)}_{l'}$. Besides, we take $G^{(i,j)}_k = \left(G^{(0)}_k \setminus \{i\}\right) \cup \{j\}$ and $G^{(0)}_{k+1} = \left(G^{(0)}_{k+1} \setminus \{i\}\right) \cup \{j\}$. This partition corresponds to the partition $G^{(0)}$ after shifting the points i and j. We denote $Sh(G^{(0)})$ the set of all these partitions

$$Sh(G^{(0)}) = \{G^{(i,j)}, \ \ i \in G^{(0)}_k, j \in G^{(0)}_{k+1}, \ \ k \in [1, K-1]\} \ . \tag{37}$$

For $G^{(i,j)} \in Sh(G^{(0)})$, the groups of $G^{(i,j)}$ are of the same size than the groups of $G^{(0)}$. Thus, $G^{(i,j)} \in \mathcal{P}_\alpha$. And, all the groups of $G^{(0)}$ are of size at least 3. This implies that the $G^{(i,j)}$'s are all distinct.

For $G^{(i,j)} \in Sh(G^{(0)})$, let us compute $KL(\mathbb{P}_{\mu,G^{(i,j)}}, \mathbb{P}_{\mu,G^{(0)}})$. Given $i' \in [1, n]$, we denote $\mathbb{P}_{\mu,G^{(i,j)}}(Y_{i'})$ the marginal law of the vector $Y_{i'}$ under the joint law $\mathbb{P}_{\mu,G^{(i,j)}}$. Using the independence of the $Y_{i'}$'s for $i' \in [1, n]$, we get

$$\begin{aligned}
KL(\mathbb{P}_{\mu,G^{(i,j)}}, \mathbb{P}_{\mu,G}) =& \sum_{i'=1}^{n} KL(\mathbb{P}_{\mu,G^{(i,j)}}(Y_{i'}), \mathbb{P}_{\mu,G^{(0)}}(Y_{i'})) \\
=& KL\left(\mathbb{P}_{\mu,G^{(i,j)}}(Y_i), \mathbb{P}_{\mu,G^{(0)}}(Y_i)\right) + KL\left(\mathbb{P}_{\mu,G^{(i,j)}}(Y_j), \mathbb{P}_{\mu,G^{(0)}}(Y_j)\right) \\
=& KL\left(\mathcal{N}(\mu_{k+1}, I_p), \mathcal{N}(\mu_k, I_p)\right) + KL\left(\mathcal{N}(\mu_k, I_p), \mathcal{N}(\mu_{k+1}, I_p)\right) \\
=& 2\bar{\Delta}^2 \ \leq \ 2c_1^2 \log(n) \ ,
\end{aligned}$$

where we used the property $KL\left(\mathcal{N}(f, I_p), \mathcal{N}(g, I_p)\right) = \frac{\|f-g\|^2}{2}$, for all $f, g \in \mathbb{R}^p$.

For any estimator $\hat{G}$, we associate $\hat{J}$ the estimator that gives (i, j) if $\hat{G} = G^{(i,j)}$ and $(1, 2)$ elsewhere (we can suppose that $G^{(1,2)} \in Sh(G^{(0)})$). Lemma 25 implies that, for all estimator $\hat{G}$, the

corresponding estimator $\hat{J}$ satisfies

$$\frac{1}{|Sh(G^{(0)})|}\sum_{G^{(i,j)}\in Sh(G^{(0)})}\mathbb{P}_{\mu,G^{(i,j)}}(\hat{J}\neq(i,j))\geq 1-\frac{1+2c_1^2\log(n)}{\log(|Sh(G^{(0)})|)}\ .$$

For all $i\in G_k$, if $j\in G_{k+1}$, then $G^{(i,j)}\in Sh(G^{(0)})$. The groups of $G^{(0)}$ are of size at least $\frac{2n}{3K}$. Hence, since $K\geq 2$ and $n\geq 9K/2$,

$$|Sh(G^{(0)})|\geq (K-1)\left(\frac{2n}{3K}\right)^2\geq\frac{4}{18}\frac{n^2}{K}\geq n\ . \tag{38}$$

We arrive at

$$\frac{1}{|Sh(G^{(0)})|}\sum_{G^{(i,j)}\in Sh(G^{(0)})}\mathbb{P}_{\mu,G^{(i,j)}}(\hat{J}\neq(i,j))\geq 1-\frac{1+2c_1^2\log(n)}{\log(n)}\ .$$

If c_1 is small enough and n_0 large enough, there exists a constant $C>0$ such that, for all integers $n\geq n_0$, we have $1-\frac{1+2c_1^2\log(n)}{\log(n)}\geq C$. This implies

$$\frac{1}{|Sh(G^{(0)})|}\sum_{G^{(i,j)}\in Sh(G^{(0)})}\mathbb{P}_{\mu,G^{(i,j)}}(\hat{J}\neq(i,j))\geq C\ .$$

For any estimator $\hat{G}$ and its corresponding estimator $\hat{J}$, for any $G^{(i,j)}\in Sh(G^{(0)})$, we have $\mathbb{P}_{\mu,G^{(i,j)}}(\hat{G}=G^{(i,j)})\leq\mathbb{P}_{\mu,G^{(i,j)}}(\hat{J}=(i,j))$. Thus, we arrive at

$$\frac{1}{|Sh(G^{(0)})|}\sum_{G^{(i,j)}\in Sh(G^{(0)})}\mathbb{P}_{\mu,G^{(i,j)}}(\hat{G}\neq G^{(i,j)})\geq C\ .$$

This, with the fact that, for all estimator $\hat{G}$,

$$\sup_{\mu\in\theta_{\bar{\Delta}}}\sup_{G\in\mathcal{P}_\alpha}\mathbb{P}_{\mu,G}(\hat{G}\neq G)\geq\frac{1}{|Sh(G^{(0)})|}\sum_{G^{(i,j)}\in Sh(G^{(0)})}\mathbb{P}_{\mu,G^{(i,j)}}(\hat{G}\neq G^{(i,j)})\geq C\ .$$

This concludes the proof of the theorem in the regime where $\bar{\Delta}^2\geq c_1\log(n)$.

D.2 $c_1\log(n)\leq\bar{\Delta}^2\leq c_2\sqrt{\frac{p}{n}K\log(n)}$.

We suppose that $c_1\log(n)\leq\bar{\Delta}^2\leq c_2\sqrt{\frac{p}{n}K\log(n)}$, with c_1 the numerical constant chosen just above and c_2 another numerical constant that we will choose small enough. Given ρ a probability distribution on $(\mathbb{R}^p)^K$ and a partition G of $[1,n]$ in K groups, we define the probability distribution on $(\mathbb{R}^p)^n$ by

$$\mathbb{P}_{\rho,G}(B)=\int\mathbb{P}_{\mu,G}(B)d\rho(\mu)\ .$$

The proof of Theorem 4 in this regime uses the following lemma. It is a reduction lemma which plays the same role as Lemma 20 in the proof of Theorem 3.

Lemma 26. *We suppose that there exists a probability distribution ρ on $(\mathbb{R}^p)^K$ and $a>0$ such that*

$$\inf_{\hat{G}}\sup_{G\in\mathcal{P}_\alpha}\mathbb{P}_{\rho,G}(\hat{G}\neq G)-\rho((\mathbb{R}^p)^K\setminus\Theta_{\bar{\Delta}})>a\ .$$

Then, we have

$$\inf_{\hat{G}}\sup_{\mu\in\Theta_{\bar{\Delta}}}\sup_{G\in\mathcal{P}_\alpha}\mathbb{P}_{\mu,G}(\hat{G}\neq G)>a\ .$$

We refer to Section D.4 for a proof of this lemma. We will consider the same distribution on $(\mathbb{R}^p)^K$ as in Section C. We take $\varepsilon = \sqrt{\frac{2}{p}}\bar{\Delta}$ and ρ the uniform distribution on the hypercube $\mathcal{E} = \{-\varepsilon, \varepsilon\}^{p\times K}$.

We will use Fano's Lemma to lower bound $\inf_{\hat{G}} \sup_{G\in\mathcal{P}_\alpha} \mathbb{P}_{\rho,G}(\hat{G} \neq G)$. To do so, we need to find many partitions $G^{(1)}, \dots, G^{(M)} \in \mathcal{P}_\alpha$, with M large, such that $KL(\mathbb{P}_{\rho,G^{(l)}}, \mathbb{P}_{\rho,G^{(0)}})$ remains small. Again, we use the set of partitions $Sh(G^{(0)})$, defined by (37), for $G^{(0)} \in \mathcal{P}_{\frac{3}{2}}$. For such a partition $G \in Sh(G^{(0)})$, the next lemma controls the quantity

$$KL(\mathbb{P}_{\rho,G}, \mathbb{P}_{\rho,G^{(0)}}) = \int \log(\frac{d\,\mathbb{P}_{\rho,G}}{d\,\mathbb{P}_{\rho,G^{(0)}}}) d\,\mathbb{P}_{\rho,G} \ .$$

We refer to Section D.3 for a proof of this lemma.

Lemma 27. *We suppose that c_2 is small enough with respect to c_1. Then, there exists a numerical constant $c > 0$ such that, for all $G \in Sh(G^{(0)})$, we have the inequality*

$$KL(\mathbb{P}_{\rho,G}, \mathbb{P}_{\rho,G^{(0)}}) \leq cc_2^2 \log(n) \ .$$

Together with Lemma 25 applied to the set $Sh(G^{(0)})$, this lemma induces

$$\inf_{\hat{G}} \frac{1}{|Sh(G^{(0)})|} \sum_{G\in Sh(G^{(0)})} \mathbb{P}_{\rho,G^{(0)}}[\hat{G} \neq G^{(0)}] \geq 1 - \frac{1 + cc_2^2 \log(n)}{\log(|Sh(G^{(0)})|)} \ .$$

Since for all estimator $\hat{G}$, the inequality

$$\frac{1}{|Sh(G^{(0)})|} \sum_{G\in Sh(G^{(0)})} \mathbb{P}_{\rho,G}[\hat{G} \neq G] \leq \sup_{G\in\mathcal{P}_\alpha} \mathbb{P}_{\rho,G}[\hat{G} \neq G]$$

is satisfied, we have the following inequality

$$\inf_{\hat{G}} \sup_{G\in\mathcal{P}_\alpha} \mathbb{P}_{\rho,G}[\hat{G} \neq G] \geq 1 - \frac{1 + cc_2^2 \log(n)}{\log(|Sh(G^{(0)})|)} \ .$$

Finally, referring to equation (38), we have $\log(|Sh(G^{(0)})|) \geq \log(n)$. Thus, if we choose c_2 small enough and if n_0 large enough, there exists a numerical constant $b > 0$ satisfying

$$\inf_{\hat{G}} \sup_{G\in\mathcal{P}_\alpha} \mathbb{P}_{\rho,G}[\hat{G} \neq G] \geq b \ .$$

In order to apply Lemma 26, it remains to control $\rho(\mathbb{R}^{p\times K} \setminus \Theta_{\bar{\Delta}})$. Lemma 24 states that $\rho(\mathbb{R}^{p\times K} \setminus \Theta_{\bar{\Delta}}) \leq K(K-1)\exp(-p/8)$. Moreover, since $c_1 \log(n) \leq c_2\sqrt{\frac{p}{n}K\log(n)}$, we have $p \geq \frac{c_1^2}{c_2^2}\frac{n}{K}\log(n)$. This implies that $\rho(\mathbb{R}^p \setminus \Theta_{\bar{\Delta}}) \leq \frac{b}{2}$, provided c_2 is small enough with respect to c_1. Combining this inequality with Lemma 26 leads to

$$\inf_{\hat{G}} \sup_{G\in\mathcal{P}_\alpha} \sup_{\mu\in\Theta_{\bar{\Delta}}} \mathbb{P}_{\rho,\mu}[\hat{G} \neq G] \geq \frac{b}{2} \ .$$

This concludes the proof of Theorem 4.

D.3 Proof of Lemma 27

By symmetry, we can suppose that $1 \in G_1^{(0)}$ and $2 \in G_2^{(0)}$ and we compute $KL(\mathbb{P}_{\rho,G}, \mathbb{P}_{\rho,G^{(0)}})$ for $G = G^{(1,2)}$. We proceed similarly as for the proof of Lemma 23, except that here we will use an

upper-bound and a lower-bound of increments of the function $\log\cosh$, instead of bounding the function $\log\cosh$ itself. That will allow us to have a sharper bound on $KL(\mathbb{P}_{\rho,G^{(1,2)}}, \mathbb{P}_{\rho,G^{(0)}})$.

In the following, in order to ease the computations, we denote by ρ' the probability distribution on $(\mathbb{R})^{p\times K}$ that satisfies; if $(\mu_1,\dots,\mu_K)\sim\rho'$, all the μ_i's are independent, $\mu_1=\mu_2=0$ and all the other μ_i's are drawn uniformly on the set $\{-\varepsilon,+\varepsilon\}^p$. We recall that $\varepsilon^2=\frac{2}{p}\bar{\Delta}$ and that $c_1\log(n)\le\bar{\Delta}^2\le c_2\sqrt{\frac{p}{n}K\log(n)}$, with c_2 that we will choose small enough.

First, we compute the quantity

$$\frac{d\,\mathbb{P}_{\rho,G^{(1,2)}}}{d\,\mathbb{P}_{\rho,G^{(0)}}}=\frac{\frac{d\,\mathbb{P}_{\rho,G^{(1,2)}}}{d\,\mathbb{P}_{\rho',G^{(1,2)}}}}{\frac{d\,\mathbb{P}_{\rho,G^{(0)}}}{d\,\mathbb{P}_{\rho',G^{(1,2)}}}}=\frac{\frac{d\,\mathbb{P}_{\rho,G^{(1,2)}}}{d\,\mathbb{P}_{\rho',G^{(1,2)}}}}{\frac{d\,\mathbb{P}_{\rho,G^{(0)}}}{d\,\mathbb{P}_{\rho',G^{(0)}}}}\ , \tag{39}$$

where the second equality comes from the fact that $\mathbb{P}_{\rho',G^{(0)}}=\mathbb{P}_{\rho',G^{(1,2)}}$. Given a probability distribution $\mathbb{P}$ on some Euclidean space, which is absolutely continuous with respect to the Lebesgue measure, we write $d\,\mathbb{P}$ for the density of this distribution with respect to the Lebesgue measure. For the numerator in (39), we have

$$\begin{aligned}\frac{d\,\mathbb{P}_{\rho,G^{(1,2)}}}{d\,\mathbb{P}_{\rho',G^{(1,2)}}}(Y)=&\frac{\mathbb{E}_\rho\left[d\,\mathbb{P}_{\mu,G^{(1,2)}}(Y)\right]}{\mathbb{E}_{\rho'}\left[d\,\mathbb{P}_{\mu,G^{(1,2)}}(Y)\right]}\\ =&\frac{\mathbb{E}_\rho\left[\prod_{k\in[1,K]}\prod_{i\in G_k^{(1,2)}}\exp\left(-\frac{1}{2}\|Y_i-\mu_k\|^2\right)\right]}{\mathbb{E}_{\rho'}\left[\prod_{k\in[1,K]}\prod_{i\in G_k^{(1,2)}}\exp\left(-\frac{1}{2}\|Y_i-\mu_k\|^2\right)\right]}\\ =&\frac{\mathbb{E}_\rho\left[\prod_{d\in[1,p]}\prod_{k\in[1,K]}\prod_{i\in G_k^{(1,2)}}\exp\left(-\frac{1}{2}\left(Y_{i,d}-\mu_{k,d}\right)^2\right)\right]}{\mathbb{E}_{\rho'}\left[\prod_{d\in[1,p]}\prod_{k\in[1,K]}\prod_{i\in G_k^{(1,2)}}\exp\left(-\frac{1}{2}\left(Y_{i,d}-\mu_{k,d}\right)^2\right)\right]}\ .\end{aligned}$$

Using the independence of the $\mu_{k,d}$'s both for the law ρ and ρ' together with the fact that, when $k>3$, μ_k has the same distribution under ρ and ρ', we get that

$$\begin{aligned}\frac{d\,\mathbb{P}_{\rho,G^{(1,2)}}}{d\,\mathbb{P}_{\rho',G^{(1,2)}}}(Y)=&\prod_{d\in[1,p]}\prod_{k\in[1,K]}\frac{\mathbb{E}_\rho\left[\prod_{i\in G_k^{(1,2)}}\exp\left(-\frac{1}{2}(Y_{i,d}-\mu_{k,d})^2\right)\right]}{\mathbb{E}_{\rho'}\left[\prod_{i\in G_k^{(1,2)}}\exp\left(-\frac{1}{2}(Y_{i,d}-\mu_{k,d})^2\right)\right]}\\ =&\prod_{d\in[1,p]}\prod_{k\in\{1,2\}}\frac{\mathbb{E}_\rho\left[\prod_{i\in G_k^{(1,2)}}\exp\left(-\frac{1}{2}(Y_{i,d}-\mu_{k,d})^2\right)\right]}{\mathbb{E}_{\rho'}\left[\prod_{i\in G_k^{(1,2)}}\exp\left(-\frac{1}{2}(Y_{i,d}-\mu_{k,d})^2\right)\right]}\\ =&\prod_{d\in[1,p]}\prod_{k\in\{1,2\}}\mathbb{E}_\rho\left[\prod_{i\in G_k^{(1,2)}}\exp\left(-\frac{1}{2}\left((Y_{i,d}-\mu_{k,d})^2-(Y_{i,d})^2\right)\right)\right]\\ =&\prod_{d\in[1,p]}\prod_{k\in\{1,2\}}\mathbb{E}_\rho\left[\prod_{i\in G_k^{(1,2)}}\exp\left(Y_{i,d}\mu_{k,d}-\frac{\varepsilon^2}{2}\right)\right]\\ =&\prod_{d\in[1,p]}e^{\frac{-|G_1^{(1,2)}|\varepsilon^2}{2}}\cosh\left(\sum_{i\in G_1^{(1,2)}}\varepsilon Y_{i,d}\right)e^{\frac{-|G_2^{(1,2)}|\varepsilon^2}{2}}\cosh\left(\sum_{i\in G_2^{(1,2)}}\varepsilon Y_{i,d}\right),\end{aligned}$$

where the third equality comes from the fact that, for all $d\in[1,p]$ and when $i\le 2$, $\mu_{i,d}=0$ almost surely under the law ρ'. Similarly, we get that

$$\frac{d\,\mathbb{P}_{\rho,G^{(0)}}}{d\,\mathbb{P}_{\rho',G^{(0)}}}(Y)=\prod_{d\in[1,p]}e^{\frac{-|G_1^{(0)}|\varepsilon^2}{2}}\cosh\left(\sum_{i\in G_1^{(0)}}\varepsilon Y_{i,d}\right)e^{\frac{-|G_2^{(0)}|\varepsilon^2}{2}}\cosh\left(\sum_{i\in G_2^{(0)}}\varepsilon Y_{i,d}\right).$$

Combining these two equalities in (39), and using the fact that the groups of $G^{(0)}$ and $G^{(1,2)}$ are of the same size, we get

$$\frac{d\,\mathbb{P}_{\rho,G^{(1,2)}}}{d\,\mathbb{P}_{\rho,G^{(0)}}}(Y) = \prod_{d=1}^{p} \frac{\cosh\left(\sum_{i\in G_1^{(1,2)}} \varepsilon Y_{i,d}\right)\cosh\left(\sum_{i\in G_2^{(1,2)}} \varepsilon Y_{i,d}\right)}{\cosh\left(\sum_{i\in G_1^{(1,2)}} \varepsilon Y_{i,d}\right)\cosh\left(\sum_{i\in G_2^{(0)}} \varepsilon Y_{i,d}\right)} \quad . \tag{40}$$

Plugging equality (40) in the definition of the Kullback-Leibler divergence leads to

$$\begin{aligned} KL(\mathbb{P}_{\rho,G^{(1,2)}}, \mathbb{P}_{\rho,G^{(0)}}) =& \mathbb{E}_{\rho,G^{(1,2)}}\left[\log\left(\frac{d\,\mathbb{P}_{\rho,G^{(1,2)}}}{d\,\mathbb{P}_{\rho,G^{(0)}}}\right)\right] \\ =& \sum_{d\in[1,p]} \mathbb{E}_{\rho,G^{(1,2)}}\left[\log\cosh\left(\varepsilon\sum_{i\in G_1^{(1,2)}} Y_{i,d}\right) + \log\cosh\left(\varepsilon\sum_{i\in G_2^{(1,2)}} Y_{i,d}\right)\right] \\ &- \sum_{d\in[1,p]} \mathbb{E}_{\rho,G^{(1,2)}}\left[\log\cosh\left(\varepsilon\sum_{i\in G_1^{(0)}} Y_{i,d}\right) + \log\cosh\left(\varepsilon\sum_{i\in G_2^{(0)}} Y_{i,d}\right)\right]. \end{aligned}$$

We recall that ϕ is the standard Gaussian density $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$. We denote by m_1 the size of $G_1^{(0)}$ and m_2 the size of $G_2^{(0)}$. Under the law $\mathbb{P}_{\rho,G^{(1,2)}}$, conditionally on $\mu_1, \dots \mu_K \sim \rho$, we have

- $\sum_{i\in G_1^{(1,2)}} Y_{i,d} \sim \mathcal{N}(m_1\mu_{1,d}, m_1)$,
- $\sum_{i\in G_2^{(1,2)}} Y_{i,d} \sim \mathcal{N}(m_2\mu_{2,d}, m_2)$,
- $\sum_{i\in G_1^{(0)}} Y_{i,d} \sim \mathcal{N}((m_1-1)\mu_{1,d} + \mu_{2,d}, m_1)$,
- $\sum_{i\in G_2^{(0)}} Y_{i,d} \sim \mathcal{N}((m_2-1)\mu_{2,d} + \mu_{1,d}, m_2)$.

These four points, together with the fact that the $\mu_{k,d}$'s are identically distributed, lead to

$$\begin{aligned} KL(\mathbb{P}_{\rho,G^{(1,2)}}, \mathbb{P}_{\rho,G^{(0)}}) =& p\,\mathbb{E}_\rho\left[\int \log\cosh(\varepsilon(m_1\mu_{1,1} + \sqrt{m_1}x))\phi(x)dx\right] \\ &+ p\,\mathbb{E}_\rho\left[\int \log\cosh(\varepsilon(m_2\mu_{2,1} + \sqrt{m_2}x))\phi(x)dx\right] \\ &- p\,\mathbb{E}_\rho\left[\int \log\cosh(\varepsilon((m_1-1)\mu_{1,1} + \mu_{2,1} + \sqrt{m_1}x))\phi(x)dx\right] \\ &- p\,\mathbb{E}_\rho\left[\int \log\cosh(\varepsilon((m_2-1)\mu_{2,1} + \mu_{1,1} + \sqrt{m_2}x))\phi(x)dx\right] \quad . \end{aligned} \tag{41}$$

First, let us upper-bound the term $\mathbb{E}_\rho\left[\int \log\cosh(\varepsilon(m_1\mu_{1,1} + \sqrt{m_1}x))\phi(x)dx\right]$. We denote $u = \varepsilon((m_1-1)\mu_{1,1} + \sqrt{m_1}x)$ and $h = \varepsilon\mu_{1,1}$. Then, we have

$$E_\rho\left[\int \log\cosh(\varepsilon(m_1\mu_{1,1} + \sqrt{m_1}x))\phi(x)dx\right] = \mathbb{E}_\rho\left[\int \log\cosh(u+h)\phi(x)dx\right] \quad .$$

We will use the Taylor expansion of the function $\log\cosh$ around u. We compute the following derivatives:

- For all $x \in \mathbb{R}$, $\log\cosh'(x) = \tanh(x)$,
- For all $x \in \mathbb{R}$, $\log\cosh''(x) = 1 - \tanh^2(x)$ which is bounded by 2 in absolute value.

Hence, Taylor-Lagrange inequality implies

$$|\log\cosh(x+y) - \log\cosh(x) - \tanh(x)y| \leq y^2, \ \ \forall (x,y) \in \mathbb{R}^2 \ . \tag{42}$$

Plugging this inequality leads to

$$\mathbb{E}_\rho\left[\int \log\cosh(u+h)\phi(x)dx\right] \leq \mathbb{E}_\rho\left[\log\cosh(u)\right] + \mathbb{E}_\rho\left[\int \tanh(u)h\phi(x)dx\right] + \mathbb{E}_\rho(h^2) \ . \tag{43}$$

First, since $h^2 = \varepsilon^4$ almost surely, we have $\mathbb{E}_\rho(h^2) = \varepsilon^4$. Now, we need to upper bound $\mathbb{E}_\rho\left[\int \tanh(u)h\phi(x)dx\right]$. For any $y \in \mathbb{R}$, we have $\tanh'(y) = 1 - \tanh^2(y)$ and $\tanh''(y) = -2\tanh(y)(1-\tanh(y)^2)$, which is bounded by 4 in absolute value. Hence, Taylor-Lagrange inequality taken at 0 leads to

$$|\tanh(y) - y| \leq 2y^2 \ \ , \forall y \in \mathbb{R} \ .$$

This leads to $\mathbb{E}_\rho\left[\int \tanh(u)h\phi(x)dx\right] \leq \mathbb{E}_\rho\left[\int uh\phi(x)dx\right] + 2\,\mathbb{E}_\rho\left[\int u^2|h|\phi(x)dx\right]$. On the one hand,

$$\begin{aligned}\mathbb{E}_\rho\left[\int uh\phi(x)dx\right] =& \varepsilon^2\,\mathbb{E}_\rho\left[\int ((m_1-1)\mu_{1,1} + \sqrt{m_1}x)\mu_{1,1}\phi(x)dx\right] \\ =& \varepsilon^2\,\mathbb{E}_\rho\left[(m_1-1)\mu_{1,1}^2\right] \\ =& (m_1-1)\varepsilon^4 \ .\end{aligned}$$

On the other hand,

$$\begin{aligned}\mathbb{E}_\rho\left[\int u^2|h|\phi(x)dx\right] =& \varepsilon^4\,\mathbb{E}_\rho\left[\int ((m_1-1)\mu_{1,1} + \sqrt{m_1}x)^2\phi(x)dx\right] \\ =& \varepsilon^4\left((m_1-1)^2\varepsilon^2 + m_1\right) \ .\end{aligned}$$

Plugging these inequalities in (43) leads to

$$\mathbb{E}_\rho\left[\int \log\cosh(u+h)\phi(x)dx\right] \leq \mathbb{E}_\rho\left[\log\cosh(u)\right] + (m_1-1)\varepsilon^4 + 2\varepsilon^4\left((m_1-1)^2\varepsilon^2 + m_1\right) + \varepsilon^4. \tag{44}$$

Now, let us lower-bound the term

$$\mathbb{E}_\rho\left[\int \log\cosh(\varepsilon((m_1-1)\mu_{1,1} + \mu_{2,1} + \sqrt{m_1}x))\phi(x)dx\right] = \mathbb{E}_\rho\left[\int \log\cosh(u+h')\phi(x)dx\right],$$

where we define $h' = \varepsilon\mu_{2,1}$, which is independent of u. Using inequality (42) together with the independence of u and h' leads to

$$\mathbb{E}_\rho\left[\int \log\cosh(u+h')\phi(x)dx\right] \geq \mathbb{E}_\rho\left[\log\cosh u\right] + \mathbb{E}_\rho\left[\int \tanh(u)\phi(x)dx\right]\mathbb{E}_\rho\left[h'\right] - \mathbb{E}_\rho\left[h'^2\right].$$

Since $\mathbb{E}_\rho\left[h'\right] = 0$ and $\mathbb{E}_\rho\left[h'^2\right] = \varepsilon^4$, we have

$$\mathbb{E}_\rho\left[\int \log\cosh(u+h')\phi(x)dx\right] \geq \mathbb{E}_\rho\left[\log\cosh u\right] - \varepsilon^4 \ . \tag{45}$$

Similarly, for the other terms in the equality (41), we have, denoting $u_2 = \varepsilon((m_2-1)\mu_{2,1} + \sqrt{m_2}x)$, $h_2 = \varepsilon\mu_{2,1}$ and $h_2' = \varepsilon\mu_{1,1}$:

$$\begin{aligned}&\mathbb{E}_\rho\left[\int \log\cosh(u_2+h_2)\phi(x)dx\right] \leq \mathbb{E}_\rho\left[\log\cosh(u_2)\right] + (m_2-1)\varepsilon^4 + 2\varepsilon^4((m_2-1)^2\varepsilon^2 + m_2) + \varepsilon^4 \\ &\mathbb{E}_\rho\left[\int \log\cosh(u_2+h_2')\phi(x)dx\right] \geq \mathbb{E}_\rho\left[\log\cosh(u_2)\right] - \varepsilon^4 \ .\end{aligned}$$

We denote $m = m_1 + m_2$. Plugging these two inequalities, together with inequalities (45) and (44), in equality (41) leads to

$$\begin{aligned} KL(\mathbb{P}_{\rho,G^{(1,2)}}, \mathbb{P}_{\rho,G^{(0)}}) \leq & p\left(\varepsilon^4(3m+2) + \varepsilon^6\left((m_1-1)^2 + (m_2-1)^2\right)\right) \\ \leq & 4pm\varepsilon^4(1+m\varepsilon^2) \ . \end{aligned}$$

Since $\varepsilon^2 = \frac{2}{p}\bar{\Delta}^2$, we have

$$KL(\mathbb{P}_{\rho,G^{(1,2)}}, \mathbb{P}_{\rho,G^{(0)}}) \leq 16\bar{\Delta}^4 \frac{m}{p}(1 + \frac{2m}{p}\bar{\Delta}^2) \ .$$

Besides, $G^{(0)} \in \mathcal{P}_{\frac{3}{2}}$. Thus, all the groups of $G^{(0)}$ are of size at most $\frac{3}{2}\frac{n}{K}$. Hence, $m \leq 3\frac{n}{K}$. We arrive at

$$KL(\mathbb{P}_{\rho,G^{(1,2)}}, \mathbb{P}_{\rho,G^{(0)}}) \leq 48\bar{\Delta}^4 \frac{n}{Kp}(1 + 6\frac{n}{Kp}\bar{\Delta}^2) \ .$$

The hypothesis $c_1 \log(n) \leq \bar{\Delta}^2 \leq c_2\sqrt{\frac{p}{n}K\log(n)}$ leads to

$$\bar{\Delta}^4 \frac{n}{Kp} \leq c_2^2 \log(n) \ ,$$

and to

$$\frac{n}{Kp}\bar{\Delta}^2 = \frac{n}{Kp}\bar{\Delta}^4 \frac{1}{\bar{\Delta}^2} \leq \frac{c_2^2\log(n)}{c_1\log(n)} \leq 1 \ ,$$

when c_2 is chosen small enough with respect to c_1. Thus, there exists a numerical constant c such that

$$KL(\mathbb{P}_{\rho,G^{(1,2)}}, \mathbb{P}_{\rho,G^{(0)}}) \leq cc_2^2 \log(n) \ .$$

By symmetry, this inequality is satisfied by all partition $G \in Sh(G^{(0)})$. This concludes the proof of the lemma.

D.4 Proof of Lemma 26

Let us suppose that there exists a probability distribution ρ on $(\mathbb{R}^p)^K$ and $a > 0$ such that

$$\inf_{\hat{G}} \sup_{G\in\mathcal{P}_\alpha} \mathbb{P}_{\rho,G}(\hat{G} \neq G) - \rho(\mathbb{R}^{p\times K} \setminus \Theta_{\bar{\Theta}}) \geq a \ .$$

Given an estimator $\hat{G}$, the previous hypothesis directly implies that there exists $G \in \mathcal{P}_\alpha$ such that $\mathbb{P}_{\rho,G}(\hat{G} \neq G) - \rho(\mathbb{R}^{p\times K} \setminus \Theta_{\bar{\Theta}}) \geq a$.

By definition, $\mathbb{P}_{\rho,G}(\hat{G} \neq G) = \int \mathbb{P}_{\mu,G}(\hat{G} \neq G)d\rho(\mu)$. The quantity $\mathbb{P}_{\mu,G}(\hat{G} \neq G)$ being bounded by 1, we have $\mathbb{P}_{\rho,G}(\hat{G} \neq G) \leq \int_{\Theta_{\bar{\Theta}}} \mathbb{P}_{\mu,G}(\hat{G} \neq G)d\rho(\mu) + \rho(\mathbb{R}^{p\times K} \setminus \Theta_{\bar{\Theta}})$. Therefore, $\int_{\Theta_{\bar{\Theta}}} \mathbb{P}_{\mu,G}(\hat{G} \neq G)d\rho(\mu) \geq a$. This implies the existence of $\mu \in \Theta_\Delta$ such that $\mathbb{P}_{\mu,G}(\hat{G} \neq G) \geq a$.

This being true for all estimator $\hat{G}$, we get the following inequality that concludes the proof of the lemma

$$\inf_{\hat{G}} \sup_{\mu\in\Theta_{\bar{\Theta}}} \sup_{G\in\mathcal{P}_\alpha} \mathbb{P}_{\mu,G}(\hat{G} \neq G) \geq a \ .$$

E Proof of Corollary 1

We suppose $n \geq cK^2\log(n)$, with $c > 0$ a numerical constant that we will choose large enough, and $p \geq n$. Let $M^{\hat{G}}$ be the estimator of M^* induced by the exact K-means estimator $\hat{G}$. Again, we suppose without loss of generality that $\sigma = 1$.

We will show that, if $\bar{\Delta}^4 \geq c' \frac{pK}{n} \log(n)$, for c' a numerical constant chosen large enough, the conditions of Theorem 5 will be satisfied. These conditions are a condition of balanceness of the partition G^* and a condition on the separation of the μ_k's.

First, Lemma 28 states that the partition induced by the k_i's balanced. We refer to Section E.1 for a proof of this lemma.

Lemma 28. *We consider the partition G^* induced by the k_i's by the relation $G_k^* = \{i \in [1,n],\ \ k_i = k\}$. Then, there exists a numerical constant $c > 0$, such that if $n \geq cK^2 \log(n)$, the following holds with probability at least $1 - \frac{2}{n^2}$. For all $k \in [1,K]$, the size of G_k^* satisfies $\frac{n}{2K} \leq |G_k^*| \leq \frac{3n}{2K}$.*

Hence, conditionally on the μ_k's, Theorem 5 implies the existence of a constant $c_1 > 0$ such that, if $\min_{k \neq l} \frac{1}{4}\|\mu_k - \mu_l\|^4 \geq c_1 \frac{pK}{n} \log(n)$, the partition $\hat{G}$ recovers exactly the partition G^*, with probability larger than $1 - \frac{c_2}{n^2}$, with $c_2 > 0$ a numerical constant. The next lemma shows that this condition on the separation of the μ_k's is satisfied with high probability. We refer to Section E.2 for a proof of this lemma.

Lemma 29. *We suppose $p \geq n$. There exists numerical constants $c_3 > 0$ and $c_4 > 0$ such that, if $\bar{\Delta}^4 \geq c_3 \frac{pK}{n} \log(n)$, the following holds. With probability at least $1 - \frac{c_4}{n^2}$, the separation between the clusters satisfies $\Delta^4 = \min_{k \neq l} \frac{1}{4}\|\mu_k - \mu_l\|^4 \geq c_1 \frac{pK}{n} \log(n)$.*

Combining Lemma 28, Lemma 29 together with Theorem 5 leads to the following statement. If $n \geq cK^2 \log(n)$, $p \geq n$ and $\bar{\Delta}^4 \geq c_3 \frac{pK}{n} \log(n)$, the partition $\hat{G}$ recovers exactly the partition G^* with probability at least $\frac{C}{n^2}$, with C a numerical constant. This induces $\mathbb{P}[M^{\hat{G}} \neq M^*] \leq \frac{C}{n^2}$, and thus

$$\mathbb{E}[\|M^{\hat{G}} \neq M^*\|_F^2] \leq \frac{C}{n^2} \ .$$

This concludes the proof of the corollary.

E.1 Proof of Lemma 28

Let us denote $N_k = |\{i \in [1,n],\ \ k_i = k\}|$, for $k \in [1,K]$. We prove in this section that, if $n \geq cK^2 \log(n)$, for $c > 0$ a numerical constant that we will choose large enough, then, with probability at least $\frac{2}{n^2}$, simultaneously on all $k \in [1,K]$, $|N_k - \frac{n}{K}| \leq \frac{n}{2K}$.

Let $k \in [1,K]$. Then, $N_k = \sum_{i \in [1,n]} \mathbf{1}_{k_i = k}$ is a sum of n independent Bernoulli random variables of parameter $\frac{1}{K}$. Hence, $\mathbb{E}[N_k] = \frac{n}{K}$ and Hoeffding's inequality implies

$$\mathbb{P}[|N_k - \frac{n}{K}| \geq \frac{n}{2K}] \leq 2\exp\left(-\frac{2\frac{n^2}{4K^2}}{n}\right) \leq 2\exp\left(-\frac{n}{2K^2}\right) \ .$$

Moreover, if the numerical constant c such that $n \geq cK^2 \log(n)$ is large enough, we have $\exp\left(-\frac{n}{2K^2}\right) \leq \frac{1}{n^2 K}$. An union bound on $k \in [1,K]$ induces that

$$\mathbb{P}\left[\exists k \in [1,K],\ \ |N_k - \frac{n}{K}| \geq \frac{n}{2K}\right] \leq \frac{2}{n^2} \ .$$

This concludes the proof of the lemma.

E.2 Proof of Lemma 29

We proceed as for the proof of Lemma 24 in Section C.8. Let $k \neq l \in [1,K]$. Then, $\|\mu_k - \mu_l\|^2 = 4\varepsilon^2 \sum_{d \in [1,p]} \mathbf{1}_{\mu_{k,d} \neq \mu_{l,d}}$. Hence, $\frac{1}{4\varepsilon^2}\|\mu_k - \mu_l\|^2$ is a sum of p independent Bernoulli random variable

of parameter $\frac{1}{2}$. Using Hoeffding's inequality leads to

$$\mathbb{P}\left[\frac{1}{4\varepsilon^2}\|\mu_k-\mu_l\|^2 \le \frac{p}{4}\right] \le \exp\left(-\frac{p}{8}\right) \ .$$

Since $p \ge n$, there exists a numerical constant $c_4 > 0$ such that $\exp\left(-\frac{p}{2}\right) \le \frac{c_4}{K^2n^2}$. Using an union bound on the different pairs $k \ne l \in [1,K]$ implies that the following holds with probability at least $1-\frac{c_4}{n^2}$. For all $k \ne l \in [1,K]$, $\frac{1}{2}\|\mu_k-\mu_l\|^2 \ge p\varepsilon^2 = \bar{\Delta}^2$. This concludes the proof of the lemma.

F Hierarichical Clustering with single linkage

For sake of completeness, we provide in this section an analysis of hierarchical clustering with single linkage in the isotropic Gaussian setup. We recall our setup: for a hidden partition G^* and hidden vectors $\mu_1,\dots,\mu_K \in \mathbb{R}^p$, the Y_i's are drawn independently and, if $i \in G_k^*$, $Y_i \sim \mathcal{N}\left(\mu_k, \sigma^2 I_p\right)$.

Let us describe the algorithm considered. We build recursively a sequence of partitions as follows. Initially, we take the partition $G^0 = \{\{1\},\dots,\{n\}\}$. Then, as long as $G^{(t)}$ has more than K groups, we construct the partition $G^{(t+1)}$ by merging two groups of $G^{(t)}$ with the two closest points. The algorithm stops when the number of groups of the partition $G^{(t)}$ is K, which occurs when $t = n-K$. Let us write more precisely this algorithm. We define the linkage function between two subsets $A, B \subset [1,n]$ as $l(A,B) = \min_{(i,j)\in A\times B}\|Y_i - Y_j\|^2$. Here is the hierarchical clustering algorithm considered.

Algorithm 1: Hierarchical Clustering algorithm with single linkage

Data: $Y_1,\dots,Y_n$
$t \leftarrow 0$;
$G^{(0)} \leftarrow \{\{1\},\dots,\{n\}\}$;
while $t < n-K$ **do**
  Find $\hat{a},\hat{b}$ minimizing $l\left(G_{\hat{a}}^{(t)}, G_{\hat{b}}^{(t)}\right)$;
  Build $G^{(t+1)}$ by merging the groups $G_{\hat{a}}^{(t)}$ and $G_{\hat{b}}^{(t)}$, the other groups remaining unchanged;
  $t \leftarrow t+1$;
end
Result: The partition $G^{(n-K)}$.

The next result gives a sufficient condition on the separation Δ for recovering exactly the partition G^* with high probability using Algorithm 1.

Proposition 4. *There exists numerical constants c_1 and c_2 such that the following holds. If $\Delta^2 \ge c_1\left(\log(n)+\sqrt{p\log(n)}\right)$, hierarchical clustering recovers exactly the partition G^* with probability at least $1-\frac{c_2}{n^2}$.*

Proof of Proposition 4. Without loss of generality, we suppose that $\sigma = 1$. We recall that, by definition, for $i \in G_k^*$, we have $k_i^* = k$. Let $i \ne j \in [1,n]$. Then

$$\|Y_i - Y_j\|^2 = \|E_i - E_j\|^2 + 2\langle E_i - E_j, \mu_{k_i^*} - \mu_{k_j^*}\rangle + \|\mu_{k_i^*} - \mu_{k_j^*}\|^2 \ .$$

In order to prove Proposition 4, we shall prove that, with high probability, the above quantity is uniformly smaller when $k_i^* = k_j^*$ than when $k_i^* \ne k_j^*$. For $i \ne j \in [1,n]$, using Lemma 16, we get that for some numerical constants $c > 0$ and for all $x > 0$,

$$\mathbb{P}[|\|E_i - E_j\|^2 - 2p| > c\left(\sqrt{px} + x\right)] \leq 2e^{-x} \ .$$

Setting $e^{-x} = \frac{1}{n^4}$ and doing an union bound on all possible couples $i \neq j$, we get

$$\mathbb{P}\left[\forall i \neq j \in [1,n],\ \ |\|E_i - E_j\|^2 - 2p| > 4c\left(\sqrt{p\log(n)} + \log(n)\right)\right] \leq \frac{1}{n^2} \ .$$

Let us know control the cross term $\langle E_i - E_j, \mu_{k_i^*} - \mu_{k_j^*}\rangle$ uniformly on all $i \neq j \in [1,n]$. For such $i \neq j \in [1,n]$, $\langle E_i - E_j, \mu_{k_i^*} - \mu_{k_j^*}\rangle \sim \sqrt{2}\|\mu_{k_i^*} - \mu_{k_j^*}\|\mathcal{N}(0, I_p)$. Hence, for some numerical constant $c' > 0$, with probability at least $1 - e^{-x^2}$, we have $\langle E_i - E_j, \mu_{k_i^*} - \mu_{k_j^*}\rangle \geq -c'x\|\mu_{k_i^*} - \mu_{k_j^*}\|$. Setting $e^{-x^2} = \frac{1}{n^4}$ and doing an union bound on all $i \neq j$, we end up with

$$\mathbb{P}\left[\forall i \neq j,\ \ \langle E_i - E_j, \mu_{k_i^*} - \mu_{k_j^*}\rangle \geq -2c'\sqrt{\log(n)}\|\mu_{k_i^*} - \mu_{k_j^*}\|\right] \geq 1 - \frac{1}{n^2} \ .$$

Hence, with probability at least $1 - \frac{2}{n^2}$, simultaneously on all $i \neq j \in [1,n]$, we have the two inequalities

$$\begin{aligned}\langle E_i - E_j, \mu_{k_i^*} - \mu_{k_j^*}\rangle &\geq -2c'\sqrt{\log(n)}\|\mu_{k_i^*} - \mu_{k_j^*}\| \ ; \\ |\|E_i - E_j\|^2 - 2p| &\leq 4c\left(\sqrt{p\log(n)} + \log(n)\right) \ .\end{aligned}$$

Let us restrict ourselves to this event of probability at least $1 - \frac{2}{n^2}$ on which these two inequalities are satisfied. For $i \neq j \in [1,n]$,

- If $k_i^* = k_j^*$, then $\|Y_i - Y_j\|^2 = \|E_i - E_j\|^2 \leq 2p + 4c\left(\sqrt{p\log(n)} + \log(n)\right)$,
- If $k_i^* \neq k_j^*$, then $\|Y_i - Y_j\|^2 \geq 2p - 4c\left(\sqrt{p\log(n)} + \log(n)\right) - 4c'\sqrt{\log(n)}\|\mu_{k_i^*} - \mu_{k_j^*}\| + \|\mu_{k_i^*} - \mu_{k_j^*}\|^2$.

Thus, if $\Delta^2 \geq c_1\left(\log(n) + \sqrt{p\log(n)}\right)$, for c_1 a numerical constant chosen large enough, we get, for all $i \neq j$,

- If $k_i^* = k_j^*$, then $\|Y_i - Y_j\|^2 \leq 2p + \frac{\Delta^2}{3}$,
- If $k_i^* \neq k_j^*$, then $\|Y_i - Y_j\|^2 \geq 2p + \frac{2\Delta^2}{3}$.

Therefore, for all $i \neq j$ such that $k_i^* = k_j^*$ and $i' \neq j'$ such that $k_{i'}^* \neq k_{j'}^*$, we have with probability at least $1 - \frac{2}{n^2}$

$$\|Y_i - Y_j\|^2 < \|Y_{i'} - Y_{j'}\|^2 \ . \tag{46}$$

In other words, Algorithm 1 will always choose, when it is possible, to merge groups that both intersect a same cluster of the partition G^*. By induction on $t \in [0, n-K]$, we deduce from this that $G^{(t)}$ is a subpartition of G^*, ie that each group of $G^{(t)}$ is a subset of a group of G^* .

Initialization: The partition $G^{(0)} = \{1\}, \ldots, \{n\}$ is indeed a subpartition of G^*.

Induction step: Let $t \in [0, n-K-1]$ and let us suppose that $G^{(t)}$ is a subpartition of G^* and let us prove that so is $G^{(t+1)}$. Since $t \leq n-K-1$, $|G^{(t)}| \geq K+1$. Hence, there exists at least two groups of $G^{(t)}$ that are subsets of the same group of G^*. Equation (46) ensures that Algorithm 1 will choose to merge such a pair of groups. Hence, $G^{(t+1)}$ is also a subpartition of G^*. This concludes the induction.

In particular, the output partition $G^{(n-K)}$ is a subpartition of G^*. Combining this with $|G^{(n-K)}| = K$ leads to $G^{(n-K)} = G^*$. This concludes the proof of the proposition. □