Subspace clustering in high-dimensions: Phase transitions & Statistical-to-Computational gap

Anonymous Author(s)
Affiliation
Address
email

Abstract

A simple model to study subspace clustering is the high-dimensional k-Gaussian mixture model where the cluster means are sparse vectors. Here we provide an exact asymptotic characterization of the statistically optimal reconstruction error in this model in the high-dimensional regime with extensive sparsity, i.e. when the fraction of non-zero components of the cluster means ρ, as well as the ratio α between the number of samples and the dimension are fixed, while the dimension diverges. We identify the information-theoretic threshold below which obtaining a positive correlation with the true cluster means is statistically impossible. Additionally, we investigate the performance of the approximate message passing (AMP) algorithm analyzed via its state evolution, which is conjectured to be optimal among polynomial algorithm for this task. We identify in particular the existence of a statistical-to-computational gap between the algorithm that require a signal-to-noise ratio $\lambda_{\text{alg}} \geq k/\sqrt{\alpha}$ to perform better than random, and the information theoretic threshold at $\lambda_{\text{it}} \approx \sqrt{-k\rho \log \rho}/\sqrt{\alpha}$. Finally, we discuss the case of sub-extensive sparsity ρ by comparing the performance of the AMP with other sparsity-enhancing algorithms, such as sparse-PCA and diagonal thresholding.

1 Introduction

With the growing size of modern data, clustering techniques play an important role in reducing the dimensionality of the features used in modern Machine Learning pipelines. Indeed, in many tasks of interest ranging from DNA sequence analysis to image classification, the relevant features are known to live in a lower-dimensional space (intrinsic dimension) than their raw acquisition format (extrinsic dimension) [1]. In these cases, identifying these features can help saving computational resources while significantly improving on learning performance. But given a corrupted embedding of low-dimensional features in a high-dimensional space, is it always *statistically possible* to retrieve them? And if yes - how can reconstruction be achieved *efficiently* in practice? In this manuscript we address these two fundamental questions in a simple model for subspace clustering: a k-cluster Gaussian mixture model with sparse centroids. In this model, the low-dimensional hidden features are given by the sparse centroids, which are embedded in a higher dimensional space and corrupted by additive Gaussian noise. We assume that the number of non-zero components of the centroids as well as the number of samples scales linearly with the dimension of the embedding space. Given a finite sample from the mixture, the goal of the statistician is to cluster the data, i.e. estimate the centroids (or features) as well as possible.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

2 Model & setting

Let $\boldsymbol{x}_\nu \in \mathbb{R}^d$, $\nu \in [n] := \{1, \cdots, n\}$ denote n i.i.d. data points drawn from an isotropic k-cluster Gaussian mixture:

$$\boldsymbol{x}_\nu \sim_{i.i.d.} \sum_{c \in \mathcal{C}} p_c \mathcal{N}\left(\sqrt{\lambda/s}\boldsymbol{v}_c, \mathrm{I}_d\right), \qquad \nu \in [n] \tag{1}$$

where p_c are the class probabilities, $\mathcal{C}$ is the index set ($|\mathcal{C}| = k$), $\boldsymbol{v}_c \in \mathbb{R}^d$ are the s-*sparse* vectors representing the means of the clusters and λ is a measure of the signal-to-noise ratio (SNR) for this model. In the following, we focus in the balanced case for which $p_c = 1/k$. We consider the means $\{\boldsymbol{v}_c\}_{c \in \mathcal{C}}$ distributed according to a Gauss-Bernoulli distribution:

$$\boldsymbol{v}_c \sim_{i.i.d.} \rho\mathcal{N}(0, \mathrm{I}_d) + (1-\rho)\delta_0, \qquad c \in \mathcal{C} \tag{2}$$

where $\rho := s/d$ is the density of non-zero elements. For the convenience of the theoretical analysis that follows, it will be useful to work with a particular encoding of the class labels $\mathcal{C}$. For a given sample $\nu \in [n]$ belonging to the class $c \in \mathcal{C}$, define the following label indicator vector $\boldsymbol{u}_c^\nu \in \{-1, k-1\}^k$:

$$\boldsymbol{u}_c^\nu = \frac{1}{k}(-1, \ldots, -1, \underbrace{k-1}_{c\text{-th element}}, -1, \ldots, -1) \tag{3}$$

For a given draw $(\boldsymbol{x}_\nu)_{\nu \in [n]}$, define the matrices $\mathrm{X} \in \mathbb{R}^{d \times n}$, $\mathrm{V} \in \mathbb{R}^{d \times k}$ with columns given by $\boldsymbol{x}_\nu$ and $\boldsymbol{v}_c$ and $\mathrm{U} \in \mathbb{R}^{n \times k}$ with rows given by $\boldsymbol{u}_c^\nu$.

A crucial point that we shall exploit in this paper is that, with this notation, our model for subspace clustering can be written as a matrix factorization problem, where the data has been generated as:

$$\mathrm{X} = \sqrt{\frac{\lambda}{s}}\mathrm{V}\mathrm{U}^\top + \mathrm{W} \tag{4}$$

with $\mathrm{W} \in \mathbb{R}^{d \times n}$ a Gaussian matrix with elements $W_{i\nu} \sim \mathcal{N}(0, 1)$.

As discussed in the introduction, the s-sparse class means $\{\boldsymbol{v}_c\}_{c \in \mathcal{C}}$ can be thought of as low-dimensional features embedded in a higher-dimensional space $\mathbb{R}^d$, that have been corrupted by isotropic additive Gaussian noise with variance $\sim \lambda^{-1}$. Given a finite draw $(\boldsymbol{x}_\nu)_{\nu \in [n]}$ generated from this model with class means $\mathrm{V}_\star$ and labels $\mathrm{U}_\star$, the goal of the statistician is to perform *clustering*, i.e. to estimate $\mathrm{U}_\star$ from X, which is equivalent to retrieving the class label for each sample in the data. However, note that there is a clear class symmetry is this problem: the labels can only be estimated up to a permutation $\pi \in \mathrm{Sym}(\mathcal{C})$ reshuffling the columns of $\mathrm{U}_\star$. Taking this symmetry into account we can assess the performance of an estimator $\hat{\mathrm{U}}$ through the averaged symmetrized mean-squared error:

$$\mathrm{MSE}(\hat{\mathrm{U}}) = \min_{\pi \in \mathrm{Sym}(\mathcal{C})} \frac{1}{n}\mathbb{E}||\pi(\hat{\mathrm{U}}) - \mathrm{U}_\star||_F \tag{5}$$

where $||\cdot||_F$ denotes the matrix Frobenius norm. In particular, we will be interested in characterizing reconstruction in the *proportional high-dimensional limit* where the number of samples n, the ambient dimension d and the sparsity level s go to infinity $n, d, s \to \infty$ at fixed density $\rho = s/d$, sample complexity $\alpha = n/d$, number of clusters $k \geq 2$ and signal strength $\lambda > 0$.

Note that the clustering problem above is closely related to the problem of estimating the class means / features $\mathrm{V}_\star$. Indeed, written as in eq. (4) the problem of estimating both the labels and centroids $(\mathrm{U}_\star, \mathrm{V}_\star)$ boils down to a low-rank matrix factorization problem. In this manuscript, we have chosen to frame the discussion in terms of the clustering, but all our results could be presented also in terms of the reconstruction of the class means.

In this manuscript we provide a sharp asymptotic characterization of when reconstruction, as measured by positive correlation with the ground truth, is possible in high-dimensions for this model, both *statistically* and *algorithmically*. In particular, our **main contributions** are:

- We map the subspace clustering problem to an asymmetric matrix factorization problem. We compute the closed-form asymptotic formula for the performance of the Bayesian-optimal estimator in the high-dimensional limit, building on a statitiscal physics inspired approach [2] that has been rigorously proven in this case [3, 4] . This allows us to provide a sharp threshold bellow which reconstruction of the features is *statistically* impossible as a function of the parameters of the model.

• To estimate the algorithmic limitations of reconstruction, we analyse a tailored approximate message passing (AMP) algorithm [2, 5, 6] approximating the posterior marginals in this problem, and derive the associated *state evolution equations* characterising its asymptotic performance. Such algorithms are optimal among first order methods [7, 8]) and therefore their reconstruction threshold provides a bound on the algorithmic complexity clustering in our model.
• The two results above allow us to paint a full picture of the *statistical-to-algorithmic* trade-offs in high-dimensional subspace clustering for the sparse k-Gaussian mixture model, and in particular to identify an *algorithmically hard* region of the sparsity level $(1-\rho)$ vs. signal-to-noise ratio (λ) plane for which reconstruction is possible statistically but not algorithmically, see Fig. 1. Further, we provide a detailed analysis in the high sparsity regime ($\rho \to 0^+$) of how the algorithmically hard region grows as we increase the number of clusters and the sparsity level. In particular, the *information theoretical* transition arise at $\lambda_{\text{it}} \approx \sqrt{-k\rho\log\rho}/\sqrt{\alpha}$, and the *algorithmic* one at $\lambda_{\text{alg}} \geq k/\sqrt{\alpha}$.
• The analyis for AMP optimality relies on the finite ρ assumption as $n, d \to \infty$. We thus also investigate the complementary case when the number of non-zero components is of the order $s \lesssim n$ and indeed see that sparse principal component analysis (SPCA) and Diagonal thresholding [9] can perform better than random in this region. We find, however, that this requires $s \leq \sqrt{n}$ (thus $\rho = o(1)$). We rephrase our findings in terms of existing literature on the subspace clustering for two-classes Gaussian mixtures [10, 11].

Related works: Subspace clustering is a well-studied topic in classical Statistics, with a wide range of methods used in practice, see [12] for a review. Closer to this work is the theoretical line of work studying the limitations of clustering in high-dimensions. Baik, Ben Arous and Péché have shown that PCA for Gaussian mixture clustering fails to correlate with the mixture means below certain threshold known as the *BBP transition* [13]. For dense Gaussian means, the statistical and algorithmic limitations of clustering were analysed in different regimes of interest. Our approach to study subspace clustering relies on a mapping to a low-rank matrix factorization problem. Low-rank matrix factorization has been extensively studied in the literature, and its asymptotic Bayes-optimal performance was characterized in [2–5, 14–16]. The construction of AMP algorithms and the associated state evolutions for matrix factorization was done in [2, 5, 6, 17]. In this work, we leverage these general results on matrix factorization. The closest to our work is perhaps [2, 18], where the authors characterize the asymptotic reconstruction thresholds for the dense case ($\rho = 1$) in the proportional limit where the number of samples and input dimensions diverge at a constant rate. Non-asymptotic results were also discussed in [19] which considered a modification of Lloyd's algorithm [20] achieving minimax optimal rate and proving a computational lower bound. To the best of our knowledge, the case in which the means are sparse has only been analysed in the regime where the number of non-zero components is sub-extensive with respect to the input dimension. Lower bounds for the statistically optimal recovery threshold in this regime were given in [21, 22], while computational lower bounds are studied in [23, 24]. Clustering of two component Gaussian mixtures has been studied in order to shed light on comparison between statistical and algorithmical tractability in a sparse scenario in [10, 11, 25]. In particular, [10] conjectured and proved [11] algorithmic bounds for this problem. They claim that even below the BBP threshold they can build an algorithm achieving exponentially small misclustering error, given that (up-to log factors) we have $s \lesssim \sqrt{n}$. We relate our findings to their work exploring the extreme sparsity regime in detail.

3 Main theoretical results

In this section, we introduce the two main technical results allowing us to characterize the limitations of clustering reconstruction (both statistically and algorithmically) for the model introduced above in the proportional high-dimensional limit.

Statistical reconstruction: First, note that up to the permutation symmetry, the estimator minimizing the averaged mean-squared error in eq. (5) admits a closed-form solution given by the marginals of the posterior distribution:

$$\hat{\mathrm{U}}_{\text{bo}} = \underset{\mathrm{U}\in\mathbb{R}^{n\times k}}{\arg\min}\, \mathrm{MSE}(\mathrm{U}) = \mathbb{E}\left[\mathrm{U}|\mathrm{X}\right] \tag{6}$$

where the posterior distribution for the model defined in eq. (4) explicitly reads:

$$P(U|X) = \frac{1}{Z(X)} \prod_{\nu=1}^{n} P_u(\boldsymbol{u}_\nu) \int_{\mathbb{R}^k} \prod_{i=1}^{d} \left(\mathrm{d}v_i P_v(\boldsymbol{v}_i)\right) \prod_{\nu=1}^{n}\prod_{i=1}^{d} e^{-\frac{1}{2}\left(X_{\nu i} - \sqrt{\frac{\lambda}{s}}\boldsymbol{u}_\nu^\top \boldsymbol{v}_i\right)^2} \tag{7}$$

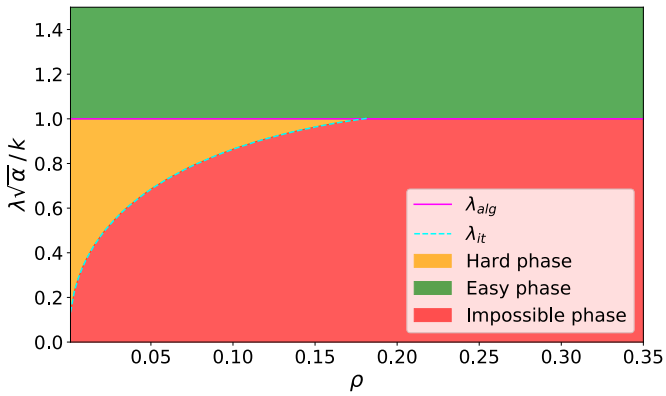

Figure 1: Phase diagram for the subspace clustering of two-clusters GMM at fixed $\alpha = 2$. We plot the SNR λ as a function of ρ and we rescale the y-axis by $\sqrt{\alpha}/k$. We colour different region of the figure according to the associated phase. The algorithmic threshold λ_{alg} is the solid line in magenta while the information-theoretic threshold λ_{it} is the dashed line in cyan. In the *impossible* region, no method can perform better than a random guess. In the *hard* region, a partial reconstruction of the signal is theoretically possible, but we do not know of any polynomial time algorithm that can do it. In the *easy* region, however, AMP, can achieve positive correlation with the ground truth (and actually achieves Bayes MMSE, except very close to the tri-critical point, see the discussion in App. A)

and for convenience we defined the vectors $\boldsymbol{u}_\nu, \boldsymbol{v}_i \in \mathbb{R}^k$ which are the rows of U, V, and with the prior distribution P_u being the uniform distribution over the indicator vectors defined in eq. (3) and P_v given by the Gauss-Bernouilli distribution defined in eq. (2).

Although in principle it would be possible to compute the minimum mean-squared error (MMSE) given by the Bayes-optimal estimator in eq. (6) by sampling from the posterior from eq. (7), this is unpractical when d, n are large. For instance, simply computing the different integrals involved in the evidence Z scale exponentially with the dimensions. The first result consists precisely in a closed-form solution for the asymptotic performance of the Bayes-optimal estimator:

Main theoretical result 1 *In the proportional high-dimensional limit where $n, d, s \to \infty$ with fixed ratios $\rho = {}^{s}/_{d}$, $\alpha = {}^{n}/_{d}$ and fixed λ, k, the minimum mean-squared error for the reconstruction of $U \in \mathbb{R}^{n\times k}$ is given by:*

$$\lim_{n\to\infty} MMSE = \frac{k-1}{k} - \operatorname{Tr} M_u^\star \tag{8}$$

where $M_u^\star \in \mathbb{R}^{k\times k}$ is the solution of the following minimization problem:

$$M_u^\star = \underset{M_u \in \mathbb{R}^{k\times k}}{\arg\min} \left\{ \max_{M_v \in \mathbb{R}^{k\times k}} \Phi_{rs}\left(M_u, M_v\right) \right\}. \tag{9}$$

with:

$$\begin{aligned}\Phi_{rs}(M_u, M_v) =& \frac{\alpha\lambda \operatorname{Tr} M_u M_v}{2\rho} - \mathbb{E}_{\boldsymbol{v}_*,\boldsymbol{w}}\left[\log Z_v\left(\frac{\alpha\lambda M_u}{\rho}, \frac{\alpha\lambda M_u \boldsymbol{u}_*}{\rho} + \sqrt{\frac{\alpha\lambda M_u}{\rho}}\boldsymbol{w}\right)\right] \\ &- \alpha\, \mathbb{E}_{\boldsymbol{u}_*,\boldsymbol{w}}\left[\log Z_v\left(\frac{\lambda M_v}{\rho}, \frac{\lambda M_v \boldsymbol{v}_*}{\rho} + \sqrt{\frac{\lambda M_v}{\rho}}\boldsymbol{w}\right)\right]\end{aligned} \tag{10}$$

and:

$$Z_u(A, \boldsymbol{b}) = \frac{1}{k}\sum_{c\in\mathcal{C}} \exp\left(\boldsymbol{b}^\top \boldsymbol{u}_c - \frac{1}{2}\boldsymbol{u}_c^\top A \boldsymbol{u}_c\right) \tag{11}$$

$$Z_v(A, \boldsymbol{b}) = 1 - \rho + \rho \exp\left(\frac{\boldsymbol{b}^\top (I_k + A)^{-1}\boldsymbol{b}}{2}\right)\sqrt{\det\left(I_k + A\right)^{-1}} \tag{12}$$

Result 1 follows from our mapping of the subspace clustering problem introduced in Sec. 2 to a low-rank matrix factorization form eq. (4). Indeed, this mapping allow us to leverage a closed-form formula characterizing the asymptotic MMSE for low-rank matrix estimation with generic priors that was derived heuristically [2, 18] using the replica method from Statistical Physics and was rigorously proven in a series of works [3–5, 14] to the context of subspace clustering. Our contribution resides in making this connection and drawing the consequences for the subspace clustering problem, a non-trivial endeavour given the complexity of the resulting formulas. Note that the minimization problem eq. (38) is fundamentally different from the one in eq. (6). Indeed, the first involves only scalar variables, and can be easily solved in a computer, while the latter is a high-dimensional problem which is computationally intractable for large d, n. The other parameter $\mathrm{M}_v^\star \in \mathbb{R}^{k\times k}$ solving eq. (38) can be used to characterize the MMSE reconstruction error on V.

Algorithmic reconstruction: While result 1 allow us to sharply characterize when clustering is statistically possible, it does not provide us a practical way to estimate the true class labels $\mathrm{U}_\star \in \mathbb{R}^{n\times k}$ from the data $\mathrm{X} \in \mathbb{R}^{d\times n}$. In order to provide a bound in the algorithmic complexity clustering, we consider an *approximate message passing* (AMP) algorithm for our problem. Message passing algorithms are a class of first order algorithms (scaling as $O(nd)$, the dimensions of the input matrix X) designed to approximate the marginals of a target posterior distribution, and which have two very important features. First, for a large class of random problems (such as the clustering problem studied here) AMP provides the best known first order method in terms of estimation performance [26–30], and has been rigorously proven to be the optimal for certain problems [7, 8]. Second, the asymptotic performance of AMP can be tracked by a set of low-dimensional *state evolution equations* [31], meaning that its reconstruction performance can be sharply computed without having to run a high-dimensional instance of the problem. For low-rank matrix factorization problems an associated AMP algorithm can be derived [5, 6, 17, 32, 33]. Therefore, yet again we leverage the mapping of subspace clustering to a low-rank matrix factorization problem to derive the associated AMP algorithm 3 with denoising functions η_v, η_u:

$$\eta_u(\mathrm{A}, \boldsymbol{b}) = \frac{1}{\sum_{c=1}^k \exp\left(\boldsymbol{b}^\top \boldsymbol{u}_c - \frac{\boldsymbol{u}_c^\top \mathrm{A} \boldsymbol{u}_c}{2}\right)} \sum_{c=1}^k \boldsymbol{u}_c \exp\left(\boldsymbol{b}^\top \boldsymbol{u}_c - \frac{\boldsymbol{u}_c^\top \mathrm{A} \boldsymbol{u}_c}{2}\right) \tag{13}$$

$$\eta_v(\mathrm{A}, \boldsymbol{b}) = \frac{(\mathrm{I}_k + A)^{-1}\boldsymbol{b}}{\rho + (1-\rho)\sqrt{\det(\mathrm{I}_k + \mathrm{A})}\exp\left(\frac{-\boldsymbol{b}^\top(\mathrm{I}_k+\mathrm{A})^{-1}\boldsymbol{b}}{2}\right)} \tag{14}$$

As mentioned above, one of they key features of Algorithm 3 is that its asymptotic performance can be tracked exactly by a set of low-dimensional equations.

Main theoretical result 2 *In the proportional high-dimensional limit where $n, d, s \to \infty$ with fixed ratios $\rho = s/d$, $\alpha = n/d$ and fixed λ, k, the correlation between the ground truth $(U_\star, V_\star)$ and the AMP estimators $(\hat{U}^t_{amp}, \hat{V}^t_{amp})$ at iterate t,*

$$M_u^t = \frac{1}{n} U_\star^\top \hat{U}^t_{amp}, \qquad M_v^t = \frac{1}{n} V_\star^\top \hat{V}^t_{amp} \tag{15}$$

satisfy the following state evolution equations*:*

$$M_u^{t+1} = \mathbb{E}_{\boldsymbol{u}_* \sim P_u, \boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0}_k, I_k)}\left[\eta_u\left(\frac{\alpha\lambda M_u}{\rho}, \frac{\alpha\lambda M_u \boldsymbol{u}_*}{\rho} + \sqrt{\frac{\alpha\lambda M_u}{\rho}}\boldsymbol{w}\right)\boldsymbol{u}_*^\top\right] \tag{16}$$

$$M_v^{t+1} = \mathbb{E}_{\boldsymbol{v}_* \sim P_v, \boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0}_k, I_k)}\left[\eta_v\left(\frac{\lambda M_v}{\rho}, \frac{\lambda M_v \boldsymbol{v}_*}{\rho} + \sqrt{\frac{\lambda M_v}{\rho}}\boldsymbol{w}\right)\boldsymbol{v}_*^\top\right] \tag{17}$$

Moreover, the asymptotic performance of is simply given by:

$$\lim_{n\to\infty} MSE(\hat{U}^t_{amp}) = \frac{k-1}{k} - \mathrm{Tr}\, M_u^t \tag{18}$$

Result 2 is a consequence from the general theory connecting AMP algorithms to their state evolution (SE) [31]. In the context of low-rank matrix factorization, a derivation of the state evolution equations

Algorithm 1 low-rAMP

Input: Data $\mathrm{X} \in \mathbb{R}^{d\times n}$

Initialize $\hat{\boldsymbol{v}}_i^{t=0}, \hat{\boldsymbol{u}}_\nu^{t=0} \sim \mathcal{N}(\mathbf{0}_k, \epsilon \mathrm{I}_k)$, $\hat{\sigma}_{u,\nu}^{t=0} = 0_{k\times k}$, $\hat{\sigma}_{v,i}^{t=0} = 0_{k\times k}$.

for $t \leq t_{\max}$ **do**

$$\mathrm{A}_u^t = \tfrac{\lambda}{s}\left(\hat{\mathrm{U}}^t\right)^\top \hat{\mathrm{U}}, \qquad A_v^t = \tfrac{\lambda}{s}\left(\hat{\mathrm{V}}^t\right)^\top \hat{\mathrm{V}}$$

$$\mathrm{B}_v^t = \sqrt{\tfrac{\lambda}{s}} X \hat{\mathrm{U}}^t - \tfrac{\lambda}{s}\sum_{\nu=1}^{n} \sigma_{u,\nu}^t \hat{\mathrm{V}}^{t-1}, \quad \mathrm{B}_u^t = \sqrt{\tfrac{\lambda}{s}} X^\top V - \tfrac{\lambda}{s}\sum_{i=1}^{d} \sigma_{v,i}^t \hat{\mathrm{U}}^{t-1}$$

Take $\{\boldsymbol{b}_{v,i}^t \in \mathbb{R}^k\}_{i=1}^d, \{\boldsymbol{b}_{u,\nu}^t \in \mathbb{R}^k\}_{\nu=1}^n$ rows of $\mathrm{B}_v^t, \mathrm{B}_u^t$

$$\hat{\boldsymbol{v}}_i^{t+1} = \eta_v(\mathrm{A}_v^t, \boldsymbol{b}_{v,i}^t), \qquad \hat{\boldsymbol{u}}_\nu^{t+1} = \eta_u(\mathrm{A}_u^t, \boldsymbol{b}_{u,\nu}^t)$$

$$\hat{\sigma}_{v,i}^{t+1} = \partial_{\boldsymbol{b}}\eta_v(\mathrm{A}_v^t, \boldsymbol{b}_{v,i}^t), \qquad \hat{\sigma}_{u,\nu}^{t+1} = \partial_{\boldsymbol{b}}\eta_u(\mathrm{A}_u^t, \boldsymbol{b}_{v,\nu}^t)$$

Here $\hat{\mathrm{U}}^t \in \mathbb{R}^{n\times k}, \hat{\mathrm{V}}^t \in \mathbb{R}^{d\times k}, \mathrm{B}_u^t \in \mathbb{R}^{n\times k}, \mathrm{B}_v^t \in \mathbb{R}^{d\times k}, \mathrm{A}_u^t \in \mathbb{R}^{k\times k}, \mathrm{A}_v^t \in \mathbb{R}^{k\times k}$

end for

Return: Estimators $\hat{\boldsymbol{v}}_{\mathrm{amp},i}, \hat{\boldsymbol{u}}_{\mathrm{amp},\nu} \in \mathbb{R}^k, \hat{\sigma}_{u,\nu}, \hat{\sigma}_{v,i} \in \mathbb{R}^{k\times k}$

above from Algorithm 3 were first provided for the rank-one case in [5, 6] and were extended to general rank and denoising functions in [2, 18]. A crucial observation is that Result 2 is intimately related to Result 1. Indeed, the state evolution equations (17) coincide exactly with running gradient descent on the potential defined in eq. (40)! However, while the performance of the statistically optimal estimator $\hat{\mathrm{U}}_{\mathrm{bo}}$ is given by the global minima of the potential Φ_{rs}, the performance of the AMP estimator $\hat{\mathrm{U}}_{\mathrm{amp}}$ is given by the minima achieved by gradient descent: i.e. the one closest the initialization $\mathrm{M}_u^{t=0}$. Therefore, studying both the statistical and algorithmic limitations of clustering in high-dimensions boils down to the study of the minima of the potential Φ_{rs}, see App. A for further details.

4 Reconstruction limits for sparse clustering

These results allow us to paint a full picture for when sparse subspace clustering is possible in the model defined in Sec. 2 as a function of the quantity of data α, the sparsity $1-\rho$, the number of clusters k and the SNR λ. Figure 1 summarizes the different reconstruction regimes in the (ρ, λ) plane for a two-clusters problem at fixed sample complexity $\alpha = 2$, also known as a *phase diagram*. Moreover the general considerations on the reconstruction limits for sparse clustering, given by analyzing the two-clusters problem, are easily generalizable to the general mixture case and not restricted to that particular model, see Sec. 5. For a fixed sparsity $1-\rho$, we identify the following regions in Fig. 1:

Impossible phase: There is not enough information in the data matrix X handled to the statistician to assign cluster membership better than chance for *any algorithm*. The Bayes-optimal MMSE is not better than a random guess. Clustering (reconstruction of $\mathrm{U}_\star$ better than chance) is impossible.

Hard phase: The MMSE is non-trivial, and clustering is statistically possible to some extent, but the best known polynomial time algorithm, AMP, fails to correlate better than chance with the true cluster assignment $\mathrm{U}_\star$. Any polynomial-time algorithm is conjectured to fail in this region.

Easy phase: In the easy phase, not only clustering is statistically possible, but AMP is able to achieve positive correlation with $\mathrm{U}_\star$. One can also invextigate when AMP achieves the Bayes-optimal MMSE (instead of just positive correlation) leading to the same transition (except from a subtle correction very close to the tri-critical point, see the discussion in App. A).

In Fig. 2, we investigate the different phases presented above by varying the sparsity level $1-\rho$ and SNR λ at fixed α. We compare the performance of AMP with popular spectral algorithms: Principal Component Analysis (PCA) and Sparse Principal Component Analysis (SPCA). We can initialize AMP and the SE equations in two different ways: we call *uninformative initialization* a choice for the first iterates of AMP and SE which assumes no knowledge on the ground truth values; conversely with *informative initialization* we consider that the statistician has some prior knowledge on the the ground truth signal. Note that in a real-life scenario the statistician doesn't have access the ground truth. Yet, as a theoretical tool the informative initialization provides important information about the algorithmically hard phase, see App A for a discussion. The initialization strategy we considered in the uninformed case for Algorithm 3 is not the only possible choice, there are smarter ways of initialising which can lead to a considerable improvement without explicitly assuming any

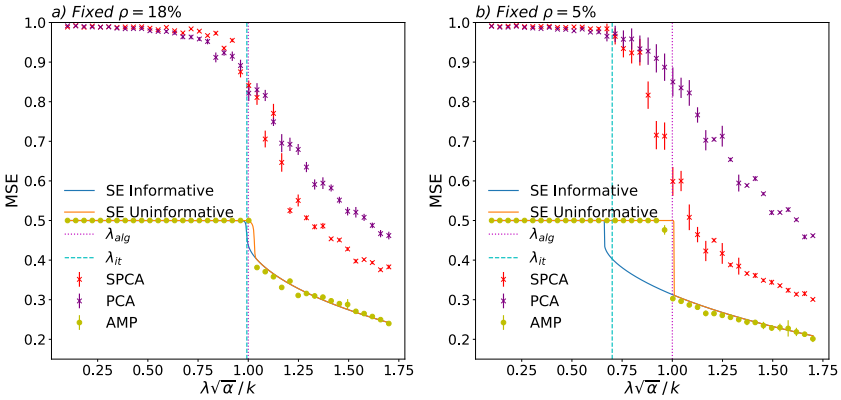

Figure 2: We compare the performance for clustering of two-classes GMM, as measured by the MSE of AMP, SPCA, PCA and SE informed and uninformed. We plot the MSE as a function of the SNR λ and we rescale the x-axis by $\sqrt{\alpha}/k$. For each algorithm considered the error bars are built using the standard deviation over fifty runs with parameters $(n = 8000, d = 4000)$, i.e. $\alpha = 2$. We plot in vertical line the theoretical values for the Information-Theoretic threshold λ_{it} (dashed cyan line), and the algorithmic threshold λ_{alg} (dotted line in magenta). The theoretical values coincide with the experimental one. The SE with uninformed initialization follows AMP as expected. *Left*: The sparsity is fixed with parameter $\rho = 18\%$. Both SPCA and PCA have a worse performance with respect to AMP and in this sparsity regime we have only a marginal advantage by using SPCA with respect to PCA. *Right*: The sparsity is fixed with parameter $\rho = 5\%$. Increasing the sparsity level the width of the algorithmically hard phase becomes bigger and SPCA performs clearly better than PCA.

information about the signal, e.g. spectral initialization [34]. Looking at Fig. 2 we see that SE with uninformed initialization tracks AMP. Morover we note that increasing the sparsity level, i.e. decreasing ρ, the problem becomes algorithmically harder. This is reflected in a discontinous jump in the MSE at λ_{alg} which becomes larger. We observe along the same lines a neat advantage by imposing the sparsity constraint in the spectral algorithm (SPCA) with respect to vanilla one (PCA) as the sparsity grows. We discuss the details on the numerical simulations in App. D and the code is available at https://github.com/anonymous.

5 Stability analysis and algorithmic threshold

In this section we provide a detailed analytical derivation of the threshold λ_{alg} characterizing the algorithmic reconstruction as a function of the number of clusters and the sparsity. First, note due to the permutation symmetry of the clusters the overlap matrices admit the following parametrization:

$$\mathrm{M}_u^t = \frac{m_u^t}{k}\mathrm{I}_k - \frac{m_u^t}{k^2}\mathrm{J}_k \qquad\qquad \mathrm{M}_v^t = m_v^t\mathrm{I}_k - \frac{m_v^t}{k}\mathrm{J}_k \tag{19}$$

where J_k is the $k \times k$-matrix with all elements equal to one. This parametrization is preserved under the SE iterations. Therefore, inserting it in the SE equations yield equations for (m_u^t, m_v^t):

$$m_u^{t+1} = f_u^{(k)}(\lambda m_v^t/\rho) \qquad\qquad m_v^{t+1} = f_v^{(k)}(\alpha\lambda m_u^t/\rho) \tag{20}$$

where we introduced the following update functions:

$$f_u^{(k)}(z) := \frac{k}{k-1}\,\mathbb{E}_{\boldsymbol{\omega}\sim\mathcal{N}(\mathbf{0}_k,\mathrm{I}_k)}\left[\frac{e^{z+w_1\sqrt{z}}}{e^{\lambda m_v^t + w_1\sqrt{z}} + \sum_{l=2}^k e^{w_l\sqrt{z}}}\right] - \frac{1}{k-1} \tag{21}$$

$$f_v^{(k)}(z) := \frac{\rho^2 z}{k(k+z)}\int_0^{+\infty}\frac{S_{k-1}}{(2\pi)^{\frac{k}{2}}}\frac{\xi^{k+1}e^{-\xi^2/2}}{\rho + (1-\rho)(\frac{k+z}{k})^{\frac{k}{2}}e^{-\frac{\xi^2 z}{2k}}}\,\mathrm{d}\xi \tag{22}$$

where S_{k-1} is the surface of the k-dimensional unitary hypersphere. Recall that $(m_u^t, m_v^t) \in [0,1]^2$ fully characterize the reconstruction performance, with $(m_u^\infty, m_v^\infty) = (0,0)$ corresponding to the performance of a random guess. Conversely, $(m_u^\infty, m_v^\infty) = (1,\rho)$ corresponds to perfect reconstruction of the cluster membership and the sparse cluster means. One can check that the *trivial fixed point*

$(m_u, m_v) = (0,0)$ is always a fixed point of the SE equations. Moreover, note that its stability is crucially connected to the algorithmic threshold λ_{alg}. Indeed, if the trivial fixed point $(m_u, m_v) = (0,0)$ is stable, AMP with an uninformed initialization will always converge to this point - and therefore achieve random guessing performance. To study its stability, we expand the update functions around $(m_u, m_v) = (0,0)$ up to the second order. Expressing everything in terms of m_u, we obtain:

$$m_u^{t+1} = F_{\text{se}}\left(m_u^t\right) = \frac{\lambda^2\alpha}{k^2} m_u^t + \left(\frac{(k-4)\alpha^2\lambda^4}{2k^4} - \frac{\lambda^3\alpha^2}{k^3}\right)\left(m_u^t\right)^2 + o\left((m_u^t)^2\right) \tag{23}$$

This immediately tells us that the trivial fixed point becomes unstable at the *algorithmic threshold* $\lambda_{\text{alg}} = {}^{k}/\sqrt{\alpha}$. This transition is a well-known result in random matrix theory and goes under the name of BBP transition [13]. Despite the fact that we expand around the trivial fixed point, the perturbative method also provides information also about the Bayes-optimal performance, thanks to the general properties that the phase diagrams in Bayes-optimal inference problems must respect [35]. Indeed, a sufficient criterion for the presence of an algorithmically hard phase requires at the algorithmic threshold: $F''_{\text{se}}(0) > 0$. The local study predicts that the phase diagram will present an algorithmically hard region for $k > k_{\text{hard}} = 4 + 2\sqrt{\alpha}$. The criterion nevertheless is not necessary in this setting, as we immediately see from the phase diagram for the two-components GMM in Fig. 1: we clearly observe an algorithmically hard region for high-sparsity although $k < k_{\text{hard}}$. In fact, the *information-theoretic threshold* λ_{it}, defined as the value of the SNR at which the problem becomes statistically possible, cannot be found simply thanks to the expansion around the trivial fixed point. In App. B we discuss how to compute λ_{it}, which we see plotted in Fig. 1 for the two-classes case. A detailed discussion on the thresholds is given in App. A.

6 Large sparsity regime

We characterize in this section the scaling of the thresholds in the very sparse (and most interesting) regime when $\rho \to 0^+$. We highlight the main passages of the computation and focus on the principal results, see Appendix C for a detailed analysis. The starting point is the following change of variables:

$$m_u = \tilde{m}_u \sqrt{\frac{-\rho \log \rho}{\alpha}} \qquad m_v = \tilde{m}_v \rho \qquad \lambda = C(k) k \sqrt{\frac{-\rho \log \rho}{\alpha}} \tag{24}$$

Inserting these expressions into eq. (20), we obtain simplified SE equations for the rescaled overlaps $(\tilde{m}_u, \tilde{m}_v)$ without any residual dependence on (ρ, α):

$$\tilde{m}_u = C(k)\tilde{m}_v \qquad \tilde{m}_v = T_k\left(C(k)\tilde{m}_u\right) \tag{25}$$

where we introduced the auxiliary function $T_k(\cdot)$ defined as:

$$T_k(z) = \int_0^{+\infty} \frac{S_{k-1}}{k(2\pi)^{k/2}} \xi^{k+1} e^{-\xi^2/2} \Theta\left(\frac{z\xi^2}{2} - 1\right) d\xi$$

where $\Theta(\cdot)$ is the Heaviside theta function. By considering the large k expansion of $T_k(\cdot)$, we can derive the scaling of the IT threshold with (k, ρ, α), see Appendix C for more details. Putting together with our previous result for λ_{alg} from Sec. 5, we obtain the following fundamental result:

$$\lambda_{\text{it}} \approx \sqrt{\frac{-k\rho \log \rho}{\alpha}} \qquad \lambda_{\text{alg}} = \frac{k}{\sqrt{\alpha}}. \tag{26}$$

These equations completely characterize the behavior at large rank & small (but finite) sparsity. The statistical-to-computational gap, where AMP is not able to exploit the information on the sparse nature of the cluster means, grows with both the sparsity and the rank. In App. C, we further show that the large rank expression for the IT threshold λ_{it} is accurate already at moderate $k \approx 10$, see Fig. 8.

It is interesting at this point to reframe our results on the high sparsity regime in the context of the existing literature. The results for the IT transition for the detection of two-classes sparse GMM were discussed in [22, 23], and we obtain a consistent scaling in the small ρ behaviour. Despite of this fact, the algorithmic bounds of different relevant works for the same problem [10, 11, 23] seem, at first sight, to not agree with our findings. In particular [11] proves rigorously the existence of an algorithm that achieves minimax rate under the BBP threshold for clustering of two-classes sparse GMM. This apparent inconsistency is related to different sparsity regimes analyzed. Indeed, here we investigate the extensive sparsity regime, i.e. $\rho = O(1)$, while the guarantees for the efficient algorithms working

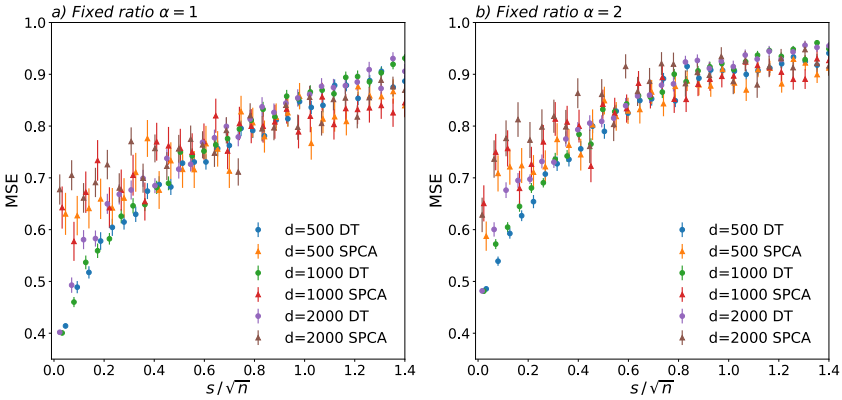

Figure 3: We compare the performance of diagonal thresholding (DT) and Sparse PCA (SPCA), as measured by the MSE, for clustering of two-classes GMM for two different parameter α. We plot the MSE vs the number of non-zero component s and we rescale the x-axis by the square-root of the number of samples. The left plot is done at $\alpha = 1$ while the right one is for $\alpha = 2$.The SNR is tuned such that we are always under the BBP threshold. With this choice we always work in the sub-extensive sparsity regime $\rho = o(1)$ and we can verify numerically what has been claimed in the literature [10, 11]: efficient algorithms in the high dimensional limit need a number of non-zero component (up-to log factors) $s \lesssim \sqrt{n}$ in order to beat random guessing under the BBP threshold.

under the BBP threshold require a very high sparsity level, i.e. $\rho = o(1)$. This difference is crucial. In Fig. 3 we illustrate it by considering the performance of two popular algorithms for this problem: *Diagonal Thresholding* (DT) [9] and SPCA. For extremely large sparsity, i.e. $s = O(1)$, these algorithms indeed provide estimators with positive correlation with the true classes below the BBP threshold! However, as soon as we increase the density of non-zero components the performance strongly deteriorates. In fact, the transition to random chance performance takes place as the number of non-zero components approach $s \lesssim \sqrt{n}$, in agreement with the literature [10, 11].

7 Conclusion

In this work we considered the problem of clustering k homogeneous Gaussian mixtures with sparse means. Mapping this to a low-rank matrix factorization problem, we have provided an exact asymptotic characterization of the MMSE in the high dimensional limit. The Bayes-optimal performance was compared to AMP, the best known polynomial time algorithm for this problem in the studied regime. In the large sparsity regime, we uncovered a large statistical-to-computational gap as the sparsity level grows, and unveiled the existence of a computationally *hard phase*. In particular, we have shown that the SNR threshold below which recovery is statistically impossible is given by $\lambda_{\text{it}} \approx \sqrt{-k\rho \log \rho}/\sqrt{\alpha}$, while the one for which AMP positively correlates with the ground truth classes is given by $\lambda_{\text{alg}} \geq k/\sqrt{\alpha}$. Our result for the existence of an algorithmically hard region was compared with the existing literature for this problem, solving an apparent contradiction due to the scaling assumption of the sparsity level with the dimension of the features. We corroborated our findings with the help of algorithms for subspace clustering such as sparse principal component analysis and diagonal thresholding.

References

[1] Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning, 2021.

[2] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Constrained low-rank matrix estimation: phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7):073403, Jul 2017.

[3] Marc Lelarge and Léo Miolane. Fundamental limits of symmetric low-rank matrix estimation. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, volume 65 of *Proceedings of Machine Learning Research*, pages 1297–1301. PMLR, 2017.

[4] Léo Miolane. Fundamental limits of low-rank matrix estimation: the non-symmetric case, 2017.

[5] Yash Deshpande and Andrea Montanari. Information-theoretically optimal sparse pca. In *2014 IEEE International Symposium on Information Theory*, pages 2197–2201. IEEE, 2014.

[6] Alyson K Fletcher and Sundeep Rangan. Iterative reconstruction of rank-one matrices in noise. *Information and Inference: A Journal of the IMA*, 7(3):531–562, 2018.

[7] Michael Celentano, Andrea Montanari, and Yuchen Wu. The estimation error of general first order methods, 2020.

[8] Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data, 2021.

[9] Iain M. Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009. PMID: 20617121.

[10] Jiashun Jin, Zheng Tracy Ke, and Wanjie Wang. Phase transitions for high dimensional clustering and related problems, 2016.

[11] Matthias Löffler, Alexander S. Wein, and Afonso S. Bandeira. Computationally efficient sparse clustering, 2021.

[12] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *SIGKDD Explor.*, 6(1):90–105, 2004.

[13] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643 – 1697, 2005.

[14] Mohamad Dia, Nicolas Macris, Florent Krzakala, Thibault Lesieur, Lenka Zdeborová, et al. Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. *Advances in Neural Information Processing Systems*, 29, 2016.

[15] Ahmed El Alaoui and Michael I Jordan. Detection limits in the high-dimensional spiked rectangular model. In *Conference On Learning Theory*, pages 410–438. PMLR, 2018.

[16] Ahmed El Alaoui, Florent Krzakala, and Michael Jordan. Fundamental limits of detection in the spiked wigner model. *The Annals of Statistics*, 48(2):863–885, 2020.

[17] Ryosuke Matsushita and Toshiyuki Tanaka. Low-rank matrix reconstruction and clustering via approximate message passing. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[18] Thibault Lesieur, Caterina de Bacco, Jess Banks, Florent Krzakala, Cris Moore, and Lenka Zdeborova. Phase transitions and optimal algorithms in high-dimensional gaussian mixture clustering. *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sep 2016.

[19] Mohamed Ndaoud. Sharp optimal recovery in the two-component gaussian mixture model, 2020.

[20] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[21] Martin Azizyan, Aarti Singh, and Larry Wasserman. Minimax theory for high-dimensional gaussian mixtures with sparse mean separation, 2013.

[22] Nicolas Verzelen and Ery Arias-Castro. Detection and feature selection in sparse mixture models, 2016.

[23] Jianqing Fan, Han Liu, Zhaoran Wang, and Zhuoran Yang. Curse of heterogeneity: Computational barriers in sparse mixture models and phase retrieval, 2018.

[24] Matthew Brennan and Guy Bresler. Average-case lower bounds for learning sparse mixtures, robust estimation and semirandom adversaries, 2019.

[25] Martin Azizyan, Aarti Singh, and Larry Wasserman. Efficient sparse clustering of high-dimensional non-spherical gaussian mixtures, 2014.

[26] David L. Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[27] F. Krzakala, M. Mézard, F. Sausset, Y. F. Sun, and L. Zdeborová. Statistical-physics-based reconstruction in compressed sensing. *Phys. Rev. X*, 2:021005, May 2012.

[28] Jean Barbier, Florent Krzakala, Nicolas Macris, LÃ©o Miolane, and Lenka ZdeborovÃ¡. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

[29] Benjamin Aubin, Bruno Loureiro, Antoine Maillard, Florent Krzakala, and Lenka Zdeborová. The spiked matrix model with generative priors. *IEEE Transactions on Information Theory*, 67(2):1156–1181, 2021.

[30] Benjamin Aubin, Bruno Loureiro, Antoine Baker, Florent Krzakala, and Lenka Zdeborová. Exact asymptotics for phase retrieval and compressed sensing with random generative priors. In Jianfeng Lu and Rachel Ward, editors, *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, volume 107 of *Proceedings of Machine Learning Research*, pages 55–73. PMLR, 20–24 Jul 2020.

[31] Adel Javanmard and Andrea Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling, 2012.

[32] Ryosuke Matsushita and Toshiyuki Tanaka. Low-rank matrix reconstruction and clustering via approximate message passing. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[33] Andrea Montanari and Ramji Venkataramanan. Estimation of low-rank matrices via approximate message passing. *The Annals of Statistics*, 49(1):321 – 345, 2021.

[34] Marco Mondelli, Christos Thrampoulidis, and Ramji Venkataramanan. Optimal combination of linear and spectral estimators for generalized linear models, 2020.

[35] Federico Ricci-Tersenghi, Guilhem Semerjian, and Lenka Zdeborová. Typology of phase transitions in bayesian inference problems. *Physical Review E*, 99(4), Apr 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3.
   (b) Did you include complete proofs of all theoretical results? [N/A] Some of the results regard low-rank matrix factorization problems, we use in this case formulas which have been proven rigorously in [3, 4]. On the other hand, heuristic results in the manuscript have been dutifully checked resorting to extensive numerical simulations.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code will be released in a non-anonymous version of the manuscript.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]