Adaptive Mean Estimation in the Hidden Markov sub-Gaussian Mixture Model

Vahe Karagulyan VAHE.KARAGULYAN@ESSEC.EDU
Department of Decisions Sciences
ESSEC Business School
Cergy, France

Mohamed Ndaoud NDAOUD@ESSEC.EDU
Department of Decisions Sciences
ESSEC Business School
Cergy, France

Abstract

We investigate the problem of center estimation in the high dimensional binary sub-Gaussian Mixture Model with Hidden Markov structure on the labels. We first study the limitations of existing results in the high dimensional setting and then propose a minimax optimal procedure for the problem of center estimation. Among other findings, we show that our procedure reaches the optimal rate that is of order $\sqrt{\delta d/n} + d/n$ instead of $\sqrt{d/n} + d/n$ where $\delta \in (0, 1/2)$ is a dependence parameter between labels. Along the way, we also develop an adaptive variant of our procedure that is globally minimax optimal. In order to do so, we rely on a more refined and localized analysis of the estimation risk. Overall, leveraging the hidden Markovian dependence between the labels, we show that it is possible to get a strict improvement of the rates adaptively at almost no cost.

Keywords: clustering, sub-Gaussian mixtures, Hidden Markov Models, high dimensional mean estimation, adaptation.

1 Introduction

In the realm of statistical modeling and machine learning, Gaussian Mixture Models (GMMs) have emerged as a versatile and powerful tool. These models find their applications across a wide spectrum of fields, from data analysis to artificial intelligence. GMMs represent a class of probabilistic models that combine multiple Gaussian distributions, each characterized by its own set of parameters. A simple example of a GMM is the symmetric two-component Gaussian mixture model in $\mathbf{R}^d$ with centers $\pm\theta$:

$$\mathbf{P}_\theta = \frac{1}{2}\mathcal{N}(-\theta, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(\theta, \mathbf{I}_d),$$

where $\theta \in \mathbf{R}^d$ is the unknown center of the mixture. Given a sample of n identically distributed observations $(Y_1, Y_2, ..., Y_n) \sim \mathbf{P}_\theta$, the mixture can be represented in the following form. For all $i = 1, 2, ..., n$, we have

$$Y_i = \eta_i\theta + \xi_i,$$

where the labels $\eta_i \in \{-1; +1\}$, and $\xi_1, \xi_2, ..., \xi_n$ are standard independent Gaussian random vectors in $\mathbf{R}^d$. When dealing with data of Gaussian Mixture Models, two primary inference

problems emerge: clustering and parameter estimation. On the one hand, the goal is to recover to which cluster each data point Y_i belongs, which can be seen as the problem of label recovery up to a sign flip. There has been an extensive line of research on this problem, among them some works that studied the performance of Lloyd's algorithm (Lu and Zhou, 2016; Ndaoud, 2022), that of spectral clustering (Abbe et al., 2022; Löffler et al., 2021) and that of SDP relaxation methods (Giraud and Verzelen, 2019; Chen and Yang, 2021), to name a few. On the other hand, parameter estimation focuses primarily on estimating the center θ. There are various works dealing with parameter estimation, for instance (Balakrishnan et al., 2017; Dwivedi et al., 2019, 2018, 2020b,a) and references therein. Most of the existing approaches evolve around variants of MLE such that the Expectation-Maximization (EM) algorithm (see also (Klusowski and Brinda, 2016) or (Wu and Zhou, 2021)). The latter proposes an algorithm which estimates θ under the quadratic loss (up to a sign flip) achieving the minimax error $\sqrt{\frac{d}{n}} \vee \frac{d}{n}$ up to a logarithmic factor, after $\sqrt{n}$ steps of expectation maximizations.

While most of the prior work focused on the case of independent observations i.e. independent labels. In this paper, we will be focusing on parameter estimation of a more sophisticated Gaussian Mixture Model with a Hidden Markovian structure, where observations have memory. Here, the label of each new observation depends on the previous one. More precisely $(\eta_i)_i$ is a Markov chain where every new label gets assigned the opposite sign of its preceding point with fixed probability $\delta \in (0, 1/2)$. As δ varies between 0 to $1/2$ the model interpolates between two better studied models, namely the Gaussian Location Model (when $\delta = 0$) and the Gaussian Mixture Model (when $\delta = 1/2$), which gives heuristics about the minimax rate of our model. A formal description of this model is given in Section 1.3.

1.1 Notation

Throughout the paper we use the following notations. For given quantities a_n and b_n, we write $a_n \lesssim b_n$ ($a_n \gtrsim b_n$) when $a_n \leq cb_n$ ($a_n \geq cb_n$) for some absolute constant $c > 0$. We write $a_n \sim b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. For any $a, b \in \mathbf{R}$, we denote by $a \vee b$ ($a \wedge b$) the maximum (the minimum) of a and b. The operator $\|.\|$ is the ℓ_2-norm in $\mathbf{R}^d$, and $\|M\|_{op}$ is the operator norm of any matrix $M \in \mathbf{R}^{a\times b}$, with respect to the ℓ_2-norm. $M^\top$ is used as the transpose of matrix M, and $\mathbf{I}_d$ the identity matrix of dimension d. For any symmetric matrix A, $\lambda_{max}(A)$ will denote the top eigenvalue of A and $v_{max}(A)$ the corresponding unit eigenvector. Finally c_0, c_1, c are used for positive constants whose values may vary from theorem to theorem.

1.2 Related work and our contributions

Most of the relevant research on clustering mixture with a hidden markov chain structure has been done in the nonparametric setting as in (Gassiat et al., 2023; Abraham et al., 2023; Lehéricy, 2021; Alexandrovich et al., 2015; Castro et al., 2015) and references therein. Our work is specific to the parametric Gaussian mixture model and our results are strongly related to those presented by (Zhang and Weinberger, 2022) who were, to the best of our knowledge, the first to provide statistical results for center estimation of the HMM (Hidden

Markov Model) Gaussian Mixture Model, assuming the knowledge of flip probability δ. More specifically they establish a minimax lower bound (Theorem 2) for the ℓ_2 risk of center estimation. Moreover, they propose a procedure that is minimax optimal up to a logarithmic factor (Theorem 1). A brief summary of their algorithm is based on a divide-and-conquer routine. It starts by dividing the sample into ℓ blocks of length k (depending on the mixing time of the Markov chain), calculating the sample mean of each block separately, then applying a well chosen spectral algorithm on the corresponding means.

The main focus here will be in the high dimensional regime where $d \lesssim n$. In the case $d \gtrsim n$, the Markovian structure does not seem to help as the minimax optimal rate is of order d/n, as one may observe in (1), and does not depend on δ. In what follows we always assume that $d \leq n$.

The lower bound provided in Theorem 1 Zhang and Weinberger (2022) can be summarized as follows

$$\inf_{\tilde{\theta}} \sup_{\|\theta\|^2=M} \mathbf{E}_\theta(\ell^2(\tilde{\theta},\theta)) \gtrsim \begin{cases} \sqrt{\frac{\delta d}{n}} + \frac{d}{n}, & M \leq \sqrt{\frac{\delta d}{n}} \vee \frac{d}{n}, \\ \frac{\frac{\delta d}{n}}{M}, & \sqrt{\frac{\delta d}{n}} \vee \frac{d}{n} \leq M \leq \delta \vee \frac{d}{n}, \\ \frac{d}{n}, & \delta \vee \frac{d}{n} \leq M, \end{cases} \tag{1}$$

up to a logarithmic factor, where the loss function $\ell(\cdot,\cdot)$ is defined as

$$\ell(\tilde{\theta},\theta) := \min\{\|\tilde{\theta}-\theta\|, \|\tilde{\theta}+\theta\|\}.$$

Based on (1) it is straightforward that we also have

$$\inf_{\tilde{\theta}} \sup_{\theta} \mathbf{E}_\theta(\ell^2(\tilde{\theta},\theta)) \gtrsim \sqrt{\frac{\delta d}{n}} + \frac{d}{n}.$$

In order the study the performance of given estimators we define two notions of optimality as follows. Let us first define the rate function $\phi(\cdot)$ such that

$$\forall \theta \in \mathbf{R}^d, \quad \phi(\theta) := \begin{cases} \sqrt{\frac{\delta d}{n}} + \frac{d}{n}, & \|\theta\|^2 \leq \sqrt{\frac{\delta d}{n}} \vee \frac{d}{n}, \\ \frac{\frac{\delta d}{n}}{\|\theta\|^2}, & \sqrt{\frac{\delta d}{n}} \vee \frac{d}{n} \leq \|\theta\|^2 \leq \delta \vee \frac{d}{n}, \\ \frac{d}{n}, & \delta \vee \frac{d}{n} \leq \|\theta\|^2. \end{cases}$$

Definition 1 *Let $\hat{\theta}$ be a measurable estimator of θ. We say that $\hat{\theta}$ is locally minimax optimal, if*

$$\sup_{\theta} \mathbf{P}\left(\ell^2(\hat{\theta},\theta) \gtrsim \phi(\theta)\right) \leq c_1 \cdot e^{-c_0 d}.$$

Moreover, we say that $\hat{\theta}$ is globally minimax optimal, if

$$\sup_{\theta} \mathbf{P}\left(\ell^2(\hat{\theta},\theta) \gtrsim \sqrt{\frac{\delta d}{n}} + \frac{d}{n}\right) \leq c_1 \cdot e^{-c_0 d}.$$

The choice of the deviation bound $c_1 \cdot e^{-c_0 d}$, in Definition 1, is inherited from the simple case of the Gaussian location model ($\delta = 0$), where the rate of estimation is the parametric rate d/n and can be achieved with probability greater than $1 - c_1 \cdot e^{-c_0 d}$. It should be emphasized that local minimax optimal rate $\phi(\theta)$ is determined region by region, whereas the global minimax rate $\sqrt{\delta d/n} + d/n$ is the maximum rate over all these regions, which coincides with the rate of the first region. Consequently one may observe that local minimax optimality leads naturally to global minimax optimality and hence is a stronger notion of optimality. The lower bound (1) is reminiscent, although more general, of the one in Wu and Zhou (2021) who provide a similar lower bound in the independent case ($\delta = 1/2$). The lower bound (1) is less pessimistic than the global minimax lower bound given by $\sqrt{\delta d/n} + d/n$. Indeed, it shows that as $\|\theta\|$ grows the lower bound gets better and eventually we recover the parametric rate d/n given enough separation between the clusters. The notion of local minimax optimality or scaled minimax optimality has also been studied by the second author in the context of linear regression Ndaoud (2019, 2020). We recall the reader here that when $d \gtrsim n$, the lower bound (1) reads as d/n independently on δ.

In terms of the upper bound, Zhang and Weinberger (2022) propose a procedure requiring only knowledge of δ and claim that it is locally minimax optimal in the sense that its risk matches the lower bound (1). After investigating the proof of Theorem 1, we found out what seems to be an issue in their argument. Specifically, in the final part of the proof of Theorem 1, an upper bound is derived using the inequality $\|\theta_*\| \leq \beta(n, d, \delta)$, despite the fact that it concerns to the opposite case. Alongside with this issue, their results sparked our curiosity about the possibility of constructing procedures achieving the minimax optimal rate, without additional logarithmic factors and with tighter probability bounds.

Our contributions can be summarized as follows:

- Focusing on the case when flip probability δ of the model is known, we've improved the results of Zhang and Weinberger (2022) with a more sophisticated choice of the estimator $\hat{\theta}(\ell)$ where the number of buckets ℓ depends on both δ and $\|\theta\|$. As opposed to the procedure given by Zhang and Weinberger (2022), we claim that the optimal procedure has to depend on $\|\theta\|$ as well. We show that the risk of our procedure matches (1) without the logarithmic factor. Our results hold for sub-Gaussian noise, moreover the deviations of our bounds are exponential. Formally for $\ell^* = d \vee \lceil n(\|\theta\|^2 \vee \delta) \rceil \wedge n$, we show that $\hat{\theta}(\ell^*)$ is locally minimax optimal and that

$$\ell^2(\hat{\theta}(\ell^*), \theta) \lesssim \begin{cases} \sqrt{\frac{\delta d}{n}} + \frac{d}{n}, & \|\theta\|^2 \leq \sqrt{\frac{\delta d}{n}} \vee \frac{d}{n} \\ \frac{\frac{\delta d}{n}}{\|\theta\|^2}, & \sqrt{\frac{\delta d}{n}} \vee \frac{d}{n} \leq \|\theta\|^2 \leq \delta \vee \frac{d}{n} \\ \frac{d}{n}, & \delta \vee \frac{d}{n} \leq \|\theta\|^2, \end{cases}$$

  with probability $1 - c_1 e^{-c_0 d}$.

- Furthermore, in Section 3, we propose an adaptive estimator $\hat{\theta}$ that only depends on δ which is also minimax optimal. Along the way, we propose heuristics that estimation of δ might be more complicated that estimation of θ in the high dimensional setting which suggests that fully adaptive procedures should not rely on estimating δ first.

  Our $\|\theta\|$-adaptive procedure is a two-stage algorithm. We start by constructing an estimator using the sub-optimal number of buckets $\ell_1 = d \vee \lceil \delta n \rceil$ which leads to an

estimator $\hat{s} = \|\hat{\theta}(\ell_1)\|$ of the norm $\|\theta\|$. We then update the number of buckets such that $\ell_2 = d \vee \lceil n(\hat{s}^2 \vee \delta)\rceil \wedge n$ before returning the new estimator $\hat{\theta}(\ell_2)$. Hence we show that adaptation to $\|\theta\|$ comes at no cost in this case and that local minimax optimality is possible as long as δ is known.

- Finally, we focus on the full adaptive case. In order to do so, we propose another method that depends weakly on δ in Section 4. This procedure is not locally minimax optimal but achieves the global minimax rate $\sqrt{\delta d/n} + d/n$. Because of its weak dependence on δ, full adaptation is possible in that case. We suggest a final procedure inspired by Lepski's adaptation scheme on the number of buckets ℓ that is globally minimax optimal (up to a logarithmic factor) but not locally minimax optimal.

- As a summary, our fully adaptive procedure reaches the rate $\sqrt{\delta d/n} + d/n$ in the worst case scenario which leads to a strict improvement over the vanilla spectral method that cannot get a rate better than $\sqrt{d/n} + d/n$. It remains an open question whether one can find fully adaptive procedures that are also locally minimax optimal.

All the proofs are deferred to the Appendix.

1.3 Model

We consider the high-dimensional Binary Markov sub-Gaussian Mixture Model. In this context we observe n samples of a d-dimensional vector θ multiplied by a random sign $\eta_i \in \{-1, 1\}, (i = 1, ..., n)$ corrupted with an additive sub-Gaussian noise. Formally the model takes the following form:

$$Y_i = \eta_i \theta + \xi_i, \tag{2}$$

where $Y_i \in \mathbf{R}^d$ are the observations, $\theta \in \mathbf{R}^d, d \geq 1$, η_i-s follow a homogeneous binary symmetric Markov chain with flip probability $\delta \in (0, 1/2)$, namely η_1 is a Rademacher random variable and for all i

$$\eta_{i+1} = \begin{cases} \eta_i & \text{with prob. } 1 - \delta \\ -\eta_i & \text{with prob. } \delta, \end{cases}$$

and ξ_i are i.i.d. isotropic 1-(sub)Gaussian random vectors that are independent from η. We remind the reader that a random vector $\omega \in \mathbf{R}^d$ is said to be 1-(sub)Gaussian if and only if for all $v \in \mathbf{R}^d$ we have $\mathbf{E}(\exp(\langle v, \omega\rangle)) \leq \exp(\|v\|^2/2)$.

For our further analysis we define $X_{i+1} := \eta_{i+1}\eta_i$ or equivalently $\eta_{i+1} = X_{i+1}\eta_i$. Hence

$$X_i = \begin{cases} 1 & \text{with prob. } 1 - \delta \\ -1 & \text{with prob. } \delta. \end{cases}$$

Notice that each X_{i+1} is independent from η_j for all $j \leq i$, since the sign change doesn't depend on the previous state. Moreover X_i-s are i.i.d. We remind the reader that our goal is to construct optimal procedures with respect to the risk defined by $\ell^2(\cdot, \cdot)$.

2 Improved mean estimation for known δ

Let's assume that the flip probability δ is known and that $\delta \in (0, \frac{1}{2})$ without loss of generality, since otherwise we can multiply samples with even indices by -1 and replace the flip probability by $1-\delta$. We begin with a modified procedure initially inherited from Zhang and Weinberger (2022), dividing the sample into ℓ blocks of length k (we'll call them buckets). We will assume without loss of generality that $n = k\ell$. We denote by I_i the set of indices of bucket i. For each bucket $i = 1, 2, ..., \ell$ consider the sample mean of k observations inside the bucket. Namely

$$\tilde{Y_i} = \frac{1}{k}\sum_{j\in I_i} \eta_j\theta + \frac{1}{k}\sum_{j\in I_i} \xi_i,$$

which can be rewritten neatly as

$$\tilde{Y_i} = \bar{\eta}_i\theta + \frac{w_i}{\sqrt{k}},$$

where $\bar{\eta}_i = \frac{1}{k}\sum_{j\in I_i}\eta_j$ and w_i is still an isotropic 1-(sub)Gaussian vector. We combine the ℓ relations in an $\mathbf{R}^{d\times\ell}$ matrix form as follows

$$\tilde{Y} = \theta\bar{\eta}^T + \frac{w}{\sqrt{k}}. \tag{3}$$

Averaging reduces the variance of the noise while weakening the signal where η is now replaced by $\bar{\eta}$. A key observation here is that $\bar{\eta}_i^2$-s are i.i.d., since each $\bar{\eta}_i^2$ only depends on X_j-s of the i-th bucket. Let us define $g(\delta) := \mathbf{E}(\bar{\eta}_i^2)$. The following Lemma gives a hint on the mixing time of the Markov chain.

Lemma 1 *Let $\delta, \delta' \in (0, 1/2)$, then we have that*

$$|g(\delta) - g(\delta')| \leq \frac{2k|\delta-\delta'|}{3}.$$

In particular, and for $\delta' = 0$, it follows that $\mathbf{E}(\bar{\eta}_i^2) \geq 1 - 2k\delta/3$. And hence as long as $k \leq 1/\delta$, $\mathbf{E}(\bar{\eta}_i^2)$ is of constant order and the signal is preserved. In other words, as long as $k \leq 1/\delta$, most of η_i are equal within the same bucket with non-zero probability, hence we can average them without a loss in the signal while reducing the variance of the noise.

Consider the Gram matrix of observations $\tilde{Y}\tilde{Y}^T$ given by

$$\tilde{Y}\tilde{Y}^T = \theta\theta^T\|\bar{\eta}\|^2 + \frac{w\bar{\eta}\cdot\theta^T + \theta\cdot(w\bar{\eta})^T}{\sqrt{k}} + \frac{1}{k}ww^T$$

and its expected value

$$\mathbf{E}[\tilde{Y}\tilde{Y}^T] = \theta\theta^T\mathbf{E}\|\bar{\eta}\|^2 + \frac{\ell}{k}\mathbf{I}_d. \tag{4}$$

For convenience denote $\Sigma := \mathbf{E}\left(\frac{1}{\ell}\tilde{Y}\tilde{Y}^T\right)$ and $\hat{\Sigma} := \frac{1}{\ell}\tilde{Y}\tilde{Y}^T$. Note that θ is proportional to the top eigenvector of Σ, meaning that $\lambda_{max}(\Sigma) = \|\theta\|^2\mathbf{E}\|\bar{\eta}\|^2/\ell + \frac{1}{k}$ and the corresponding unit eigenvector $v_{max}(\Sigma) = \frac{\theta}{\|\theta\|}$. Consequently, we consider the following estimator:

$$\hat{\theta}(\ell) := \sqrt{\frac{\ell}{\mathbf{E}\|\bar{\eta}\|^2}\left(\lambda_{max}(\hat{\Sigma}) - \frac{1}{k}\right)_+} \cdot v_{max}(\hat{\Sigma}). \tag{5}$$

Theorem 2 *For any values of $\ell \leq n$, the estimator $\hat{\theta}(\ell)$ satisfies*

$$\ell(\hat{\theta}(\ell), \theta) \lesssim \begin{cases} \frac{\ell}{\mathbf{E}\|\bar{\eta}\|^2}\left(\sqrt{\frac{d}{\ell}}\|\theta\| + \sqrt{\frac{d}{n}} + \frac{1}{\|\theta\|}\left(\sqrt{\frac{d}{kn}} + \frac{d}{n}\right)\right), & \|\theta\|^2 \leq 1 \\ \sqrt{\frac{d}{n}} + \frac{d}{n}, & \|\theta\|^2 \geq 1 \end{cases} \tag{6}$$

with probability greater than $1 - c_1 \cdot e^{-c_0 d}$, and where we take $\ell = n$ in the case $\|\theta\|^2 \geq 1$.

Next proposition shows that the lower bound in Zhang and Weinberger (2022) is indeed optimal and can be reached using a fine-tuned estimator of the form $\hat{\theta}(\ell)$. To do that, we consider several values of ℓ depending on each particular scenario of $\|\theta\|$, and hence the optimal "oracle" procedure depends on $\|\theta\|$ as opposed to what was claimed in Zhang and Weinberger (2022). We define ℓ^* as follows

$$\ell^* := d \vee \lceil n(\delta \vee \|\theta\|^2) \rceil \wedge n. \tag{7}$$

Proposition 3 *Assume that $d \leq n$. With probability greater than $1 - c_1 \cdot e^{-c_0 d}$, the estimator $\hat{\theta}(\ell^*)$ with ℓ^* defined in (7) satisfies*

$$\ell^2(\hat{\theta}(\ell^*), \theta) \lesssim \begin{cases} \sqrt{\frac{\delta d}{n}} + \frac{d}{n}, & \|\theta\|^2 \leq \sqrt{\frac{\delta d}{n}} \vee \frac{d}{n}, \\ \frac{\frac{\delta d}{n}}{\|\theta\|^2}, & \sqrt{\frac{\delta d}{n}} \vee \frac{d}{n} \leq \|\theta\|^2 \leq \delta \vee \frac{d}{n}, \\ \frac{d}{n}, & \delta \vee \frac{d}{n} \leq \|\theta\|^2. \end{cases} \tag{8}$$

Moreover $\ell^2(\hat{\theta}(\ell^*), \theta) \lesssim \sqrt{\frac{\delta d}{n}} + \frac{d}{n}$, with probability greater than $1 - c_1 \cdot e^{-c_0 d}$.

Remark 4 *The global minimax rate happens around the value $\|\theta\|^2 = \sqrt{\frac{\delta d}{n}} \vee \frac{d}{n}$ in (8), which corresponds to the worst case scenario. It is also clear that for $\delta \leq d/n$, the upper bound in (8) corresponds to the parametric rate of estimation d/n and hence we do not get further improvement in estimation for $\delta \leq d/n$. In summary, the dependence helps get better estimation rates only in the regime $d/n \leq \delta < 1/2$.*

Remark 5 *The choice $k = \lfloor \frac{1}{\delta} \rfloor$ (or $\ell = \lceil n\delta \rceil$) is reasonable, since it is aligned with the mixing time of the Markov Chain η and that is of order $1/\delta$. In other words, within each buckets the labels are more likely to share the same sign as long as the bucket size is smaller than $1/\delta$. So we may want to choose the smallest ℓ such that $\ell \geq n\delta$. Based on (6), it seems that the error term due to the misspecification around $\mathbf{E}\|\bar{\eta}\|$ gets worse as ℓ gets smaller (especially for large values of $\|\theta\|$). This explains why the choice of ℓ needs to depend both on δ and $\|\theta\|$.*

Here the oracle procedure $\hat{\theta}(\ell^*)$ carries a theoretical interest. It succeeds in attaining the local minimax optimal rate, and lays the foundation for constructing adaptive procedures.

Observing the proof one can point out that the optimal choice of ℓ^* depends on the signal strength $\|\theta\|$ and flip probability δ. In practice, neither do we have knowledge about the signal strength $\|\theta\|$, as θ is unknown, nor do we know δ. So the first step to progress is to construct a procedure without assuming any knowledge about θ, which is going to be the goal of next section.

3 Adaptive algorithm for known δ

In order to build a semi adaptive procedure, it is very natural to consider plug-in estimators of $\|\theta\|$. In what follows we take the most reasonable approach, which is estimating $\|\theta\|$ with $\hat{s}$ such that

$$\hat{s} := \|\hat{\theta}(d \vee \lceil n\delta \rceil \wedge n)\|. \tag{9}$$

First, we begin with a simple Lemma and a corresponding Corollary which provides a basis for our θ adaptive procedure.

Lemma 6 *Suppose $\ell = d \vee \lceil n\delta \rceil \wedge n$. Then there exists a constant C large enough such that with probability $1 - c_1 e^{-c_0 d}$ we get the following:*

- *If $\delta \vee Cd/n \leq \|\theta\|^2$, then*

$$\frac{\|\theta\|}{2} \leq \hat{s} \leq \frac{3\|\theta\|}{2}, \tag{10}$$

- *and if $\delta \vee Cd/n > \|\theta\|^2$, then*

$$\hat{s}^2 < 3(\delta \vee Cd/n). \tag{11}$$

As a result, we derive the next Corollary that shows that comparing $\hat{s}$ to $\delta \vee Cd/n$ is similar to comparing $\|\theta\|$ with $\delta \vee Cd/n$.

Corollary 7 *Suppose that $\ell = d \vee \lceil n\delta \rceil \wedge n$. Then for some $C > 0$ large enough, with probability $1 - c_1 e^{-c_0 d}$, we have the following:*

- *If $\hat{s}^2 < 3(\delta \vee Cd/n)$ we get that $\|\theta\|^2 < 12(\delta \vee Cd/n)$,*
- *and if $\hat{s}^2 \geq 3(\delta \vee Cd/n)$ we get that $\|\theta\|^2 \geq \delta \vee Cd/n$.*

Proof We shall treat the two cases separately. Consider first the case $\hat{s}^2 < 3(\delta \vee Cd/n)$. If $\delta \vee Cd/n \leq \|\theta\|^2$ then (10) holds, and we get $\|\theta\|^2 \leq 4\hat{s}^2 < 12(\delta \vee Cd/n)$. If $\|\theta\|^2 < \delta \vee Cd/n$, then obviously $\|\theta\|^2 < 12(\delta \vee Cd/n)$ as well. As for the second case $\hat{s}^2 \geq 3(\delta \vee Cd/n)$, then we must have $\|\theta\|^2 \geq \delta \vee Cd/n$, otherwise we get a contradiction with (11). ■

Having various permissions to replace $\|\theta\|$ with $\hat{s}$, we give an adaptive 2-step algorithm. The pseudocode of the algorithm is given in Algorithm 1.

Algorithm 1 Adaptive mean estimation for known δ

1: **input:** Observations $Y_1, Y_2, ..., Y_n$ from the model (2), and flip probability δ.

2: $\ell_1 \leftarrow d \vee \lceil \delta n \rceil \wedge n$

3: $\hat{s} \leftarrow \sqrt{\frac{\ell_1}{\mathbf{E}\|\bar{\eta}\|^2}\left(\lambda_{max}(\hat{\Sigma}) - \frac{1}{k}\right)_+}$

4: $\ell_2 \leftarrow d \vee \lceil n(\delta \vee \hat{s}^2) \rceil \wedge n$

5: $\hat{\theta}(\ell_2) \leftarrow \sqrt{\frac{\ell_2}{\mathbf{E}\|\bar{\eta}\|^2}\left(\lambda_{max}(\hat{\Sigma}) - \frac{1}{k}\right)_+} \cdot v_{max}(\hat{\Sigma})$

6: **return** $\hat{\theta} = \hat{\theta}(\ell_2)$

In what follows Theorem 8 confirms that our semi adaptive procedure is locally minimax optimal. The proof follows a similar analysis done in Proposition 3 and can be found in the Appendix. We will now define the adaptive choice of ℓ as follows

$$\hat{\ell} := d \vee \lceil n(\delta \vee \hat{s}^2) \rceil \wedge n. \tag{12}$$

Theorem 8 *When $d \leq n$, the proposed estimator $\hat{\theta}(\hat{\ell})$ with $\hat{\ell}$ defined in* (9)-(12) *satisfies*

$$\ell^2(\hat{\theta}(\hat{\ell}), \theta) \lesssim \begin{cases} \sqrt{\frac{\delta d}{n}} + \frac{d}{n}, & \|\theta\|^2 \leq \sqrt{\frac{\delta d}{n}} \vee \frac{d}{n}, \\ \frac{\frac{\delta d}{n}}{\|\theta\|^2}, & \sqrt{\frac{\delta d}{n}} \vee \frac{d}{n} \leq \|\theta\|^2 \leq \delta \vee \frac{d}{n}, \\ \frac{d}{n}, & \delta \vee \frac{d}{n} \leq \|\theta\|^2, \end{cases} \tag{13}$$

with probability larger than $1 - c_1 \log(n/d) e^{-c_0 d}$.
Globally $\ell^2(\hat{\theta}(\hat{\ell}), \theta) \lesssim \sqrt{\frac{\delta d}{n}} + \frac{d}{n}$, with probability larger than $1 - c_1 \log(n/d) e^{-c_0 d}$.

Remark 9 *The result of Theorem 8 holds with a slightly weaker probability. The logarithmic loss is due to the dependence between $\hat{\Sigma}$ and $\hat{\ell}$. One way to get rid of that dependence is by splitting the main sample into two equal sub-samples, and then using one for estimation of $\hat{\ell}$ and the other one for the construction of $\hat{\theta}(\hat{\ell})$.*

We conclude that minimax optimal rates can be achieved adaptively to the signal θ both locally and globally. The optimal procedure depends only on δ through the choice of $\hat{\ell}$ and $g(\delta)$.

4 Mean estimation for unknown δ

Taking a closer look at our estimator $\hat{\theta}(\ell)$ we recall that it depends on $g(\delta) = \mathbf{E}(\|\bar{\eta}\|^2)$, which can be theoretically computed with the knowledge of the flip probability δ. At this stage it

might be tempting to replace δ by some plug-in estimator $\hat{\delta}$. In order to estimate δ observe that

$$\mathbf{E}(\eta_1\eta_2) = (1-2\delta)\|\theta\|^2.$$

Under Gaussian noise, we can show, as in Zhang and Weinberger (2022), that a lower bound of estimation for δ is given by $1/\sqrt{n}$ for any $\|\theta\| \leq 1$. While this lower bound is far from being optimal it shows that in general we can not expect to get $\hat{\delta}$ such that $|\hat{\delta} - \delta| \leq \delta/2$. This heuristic explains that using plug-in estimators of δ might not lead to optimal rates.

An alternative approach is to completely bypass estimation of δ. We start by proposing a modified version of the estimator $\hat{\theta}(\ell)$ where we replace $g(\delta)$ by 1 which in turn leads to a weaker dependence on δ. This choice is equivalent to considering the new signal Σ^* such that

$$\Sigma^* = \theta\theta^T + \frac{\mathbf{I}_d}{k},$$

and

$$\hat{\Sigma} = \frac{1}{\ell}YY^\top = \theta\theta^T + \theta\theta^T\left(\frac{\|\bar{\eta}\|^2}{\ell} - 1\right) + \frac{(w\bar{\eta})\theta^T + \theta(w\bar{\eta})^T}{\sqrt{k\ell}} + \frac{ww^T}{k\ell}.$$

This motivates our new procedure $\tilde{\theta}(\ell)$ defined by

$$\tilde{\theta}(\ell) = \sqrt{\left(\lambda_{max}(\hat{\Sigma}) - \frac{1}{k}\right)_+} \cdot v_{max}(\hat{\Sigma}),$$

where the dependence on $\mathbf{E}\|\bar{\eta}\|^2$ was dropped. The next Theorem highlights the cost of this operation.

Theorem 10 *For any value of $\ell \leq n$, with probability greater than $1 - c_1 \cdot e^{-c_0 d}$, the estimator $\tilde{\theta}(\ell)$ achieves the following:*

$$\ell(\tilde{\theta}(\ell), \theta) \lesssim \begin{cases} \left(\sqrt{\frac{d}{\ell}} + \frac{n\delta}{\ell}\right)\|\theta\| + \sqrt{\frac{d}{n}} + \frac{1}{\|\theta\|}\left(\sqrt{\frac{1}{k}}\sqrt{\frac{d}{n}} + \frac{d}{n}\right), & \|\theta\|^2 \leq 1 \\ \sqrt{\frac{d}{n}} + \frac{d}{n}, & \|\theta\|^2 \geq 1, \end{cases} \tag{14}$$

where we take $\ell = n$ in the case $\|\theta\|^2 \geq 1$.

As expected the additional cost of not knowing $g(\delta)$ is of order $n\delta/\ell$. Next we propose the optimal number of buckets $\ell = \ell^{**}$ for this procedure. It is given by

$$\ell^{**} = d \vee \left\lceil \left(n\delta \vee \frac{\delta^{2/3} n^{4/3} \|\theta\|^{4/3}}{d^{1/3}} \vee n\|\theta\|^2 \right) \right\rceil \wedge n. \tag{15}$$

The corresponding estimation rates take different values according to several scenarios.

Proposition 11 *Assume that $d \leq n$. With probability greater than $1 - c_1 \cdot e^{-c_0 d}$, the estimator $\tilde{\theta}(\ell^{**})$, where ℓ^{**} is defined in (15), achieves the following:*

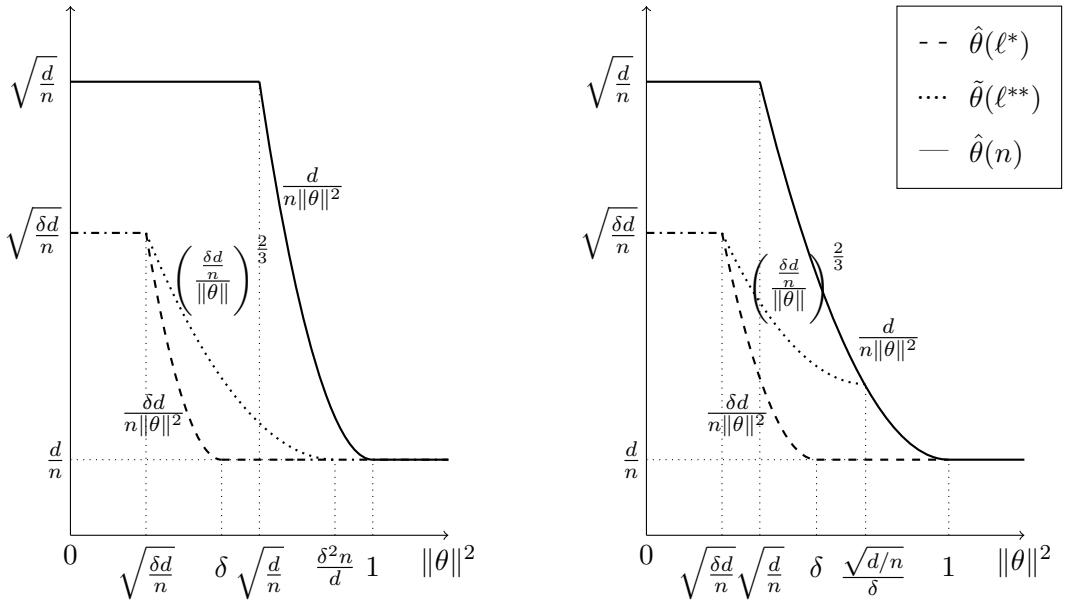
- *When* $\delta \le \sqrt{\frac{d}{n}}$*:*

$$\ell^2(\tilde{\theta}(\ell^{**}),\theta) \lesssim \begin{cases} \sqrt{\frac{\delta d}{n}} + \frac{d}{n}, & \|\theta\|^2 \le \sqrt{\frac{\delta d}{n}} \vee \frac{d}{n} \\ \left(\frac{\frac{\delta d}{n}}{\|\theta\|}\right)^{\frac{2}{3}}, & \sqrt{\frac{\delta d}{n}} \vee \frac{d}{n} \le \|\theta\|^2 \le \frac{\delta^2 n}{d} \vee \frac{d}{n} \\ \frac{d}{n}, & \frac{\delta^2 n}{d} \vee \frac{d}{n} \le \|\theta\|^2. \end{cases} \tag{16}$$

- *When* $\sqrt{\frac{d}{n}} \le \delta$*:*

$$\ell^2(\tilde{\theta}(\ell^{**}),\theta) \lesssim \begin{cases} \sqrt{\frac{\delta d}{n}} + \frac{d}{n}, & \|\theta\|^2 \le \sqrt{\frac{\delta d}{n}} \\ \left(\frac{\frac{\delta d}{n}}{\|\theta\|}\right)^{\frac{2}{3}}, & \sqrt{\frac{\delta d}{n}} \le \|\theta\|^2 \le \frac{\sqrt{\frac{d}{n}}}{\delta} \\ \frac{d}{n} + \frac{\frac{d}{n}}{\|\theta\|^2}, & \frac{\sqrt{\frac{d}{n}}}{\delta} \le \|\theta\|^2. \end{cases} \tag{17}$$

Moreover, it remains true that $\ell^2(\tilde{\theta}(\ell^{**}),\theta) \lesssim \sqrt{\frac{\delta d}{n}} + \frac{d}{n}$*, with probability greater than* $1 - c_1 e^{-c_0 d}$*.*



(a) Case $\delta \le \sqrt{d/n}$ (b) Case $\delta \ge \sqrt{d/n}$

Figure 1: Upper bounds of the risk of estimators $\hat{\theta}(\ell^*)$ (Proposition 3), $\tilde{\theta}(\ell^{**})$ (Proposition 11) and the vanilla spectral method $\hat{\theta}(n)$.

The rates of Proposition 11 are displayed in Figure 1. Based on our proof we can write the estimation error in Proposition 11 in a more compact form given by

$$\ell^2(\tilde{\theta}(\ell^{**}),\theta) \lesssim \left(\sqrt{\frac{\delta d}{n}} + \frac{d}{n}\right) \wedge \left(\frac{\ell^{**} d}{n\|\theta\|^2} + \frac{d}{n}\right).$$

As a result $\tilde{\theta}(\ell^{**})$ is not locally minimax optimal but remains globally minimax optimal. Observe that it is always better than the vanilla spectral algorithm (without bucketing) even locally. Not knowing $g(\delta)$ leads to sub-optimal rates locally which suggests that full adaptation might not be possible for locally minimax optimal procedures. That being said, if the goal is to achieve global minimax optimality then there is still hope. With this in mind the next technical lemma sets a basis for constructing an adaptive algorithm without the knowledge of optimal ℓ^{**}. This is done by proposing a surrogate upper bound that does not depend on δ.

Lemma 12 *For any number of buckets $\ell \geq \ell^{**}$ it holds that*

$$\ell^2(\tilde{\theta}(\ell), \theta) \lesssim \frac{\sqrt{d\ell}}{n} \wedge \left(\frac{\ell d}{n^2\|\theta\|^2} + \frac{d}{n} \right), \tag{18}$$

with probability $1 - c_1 \cdot e^{-c_0 d}$.

While choosing $\ell = \ell^{**}$ results in a small estimation error (even locally), we get higher global rates for $\ell \geq \ell^{**}$. Indeed $\sqrt{\frac{\delta d}{n}} \lesssim \frac{\sqrt{d\ell}}{n}$, under the assumption $\ell \geq \ell^{**}$. This observation shall help rule out values of ℓ for which the estimation rate is too large. In what follows, we focus on the case $\|\theta\|^2 \leq 1$ and propose an adaptive procedure which is inspired from the well-known Lepski's method, widely used for adaptation. The other case $\|\theta\|^2 \geq 1$ is easy since we can simply use the spectral method in that case ($\ell = n$) and get that

$$\ell^2(\tilde{\theta}(\tilde{\ell}), \theta) \lesssim \frac{d}{n},$$

with probability greater than $1 - c_1 \cdot e^{-c_0 d}$. The upper bound (18) will be useful for adaptation since it does not depend on δ but it still depends on $\|\theta\|$. We proceed in two steps. First construct a fully adaptive estimator that is globally minimax optimal. Then, plug-in the norm of the latter estimator replacing $\|\theta\|$ in (18) in order to get better rates locally.

Let us define a grid-line for values of ℓ in the set $\{d, d+1, \dots, n\}$. It is given by

$$G := \{g_i = 2^i d \mid i = 0, 1, \dots, M\},$$

where M is the greatest possible integer that $2^M d \leq n$. We define $\tilde{m}$ and $\tilde{\ell}$ in the following way:

$$\tilde{m} := \min\left\{ m \in \{0, 1, \dots, M\} : \ \ell^2(\tilde{\theta}(g_k), \tilde{\theta}(g_{k-1})) \leq 4C^2 \left(\frac{\sqrt{dg_k}}{n} \wedge \frac{dg_k}{n^2\|\tilde{\theta}(g_k)\|^2} \right) \quad \text{for all } k \geq m+1 \right\},$$

for some C large enough. We then consider the adaptive number of buckets

$$\tilde{\ell} := g_{\tilde{m}} = 2^{\tilde{m}} d. \tag{19}$$

We are now ready to state the first adaptation result of this section.

Theorem 13 *Let $\tilde{\ell}$ defined in* (19)*, $\|\theta\|^2 \leq 1$ and $d \leq n$. Then, the fully adaptive estimator $\tilde{\theta}(\tilde{\ell})$ achieves the following rate*

$$\ell^2(\tilde{\theta}(\tilde{\ell}), \theta) \lesssim \left(\sqrt{\frac{\delta d}{n}} + \frac{d}{n} \right),$$

with probability $1 - c_1 \log(n/d) \cdot e^{-c_0 d}$.

Theorem 13 leads to a fully adaptive estimator that is globally minimax optimal. While it is most likely impossible to achieve local minimax optimal rates adaptively we can still hope to achieve better rates locally. This is done in a more refined construction that we describe as follows. We define $\tilde{m}_2$ and $\tilde{\ell}_2$ in the following way:

$$\tilde{m}_2 := \min\left\{ m \in \{0,1,\dots,M\} : \ \ell^2(\tilde{\theta}(g_k), \tilde{\theta}(g_{k-1})) \le 4C^2 \left(\frac{\sqrt{dg_k}}{n} \wedge \frac{dg_k}{n^2 \|\tilde{\theta}(\tilde{\ell})\|^2} \right) \quad \text{for all } k \ge m+1 \right\},$$

for some C large enough. We then consider the adaptive number of buckets

$$\tilde{\ell}_2 := g_{\tilde{m}_2} = 2^{\tilde{m}_2} d. \tag{20}$$

Observe that, in the definition of $\tilde{m}_2$, the norm of $\|\theta\|$ is replaced by $\|\tilde{\theta}(\tilde{\ell})\|$ and not by $\|\tilde{\theta}(g_k)\|$ as we did in the definition of $\tilde{m}$. We are now ready to state the main result of this section.

Theorem 14 *Let $\tilde{\ell}_2$ defined in (20), $\|\theta\|^2 \le 1$ and $d \le n$. Then, the fully adaptive estimator $\tilde{\theta}(\tilde{\ell}_2)$ achieves the following rate*

$$\ell^2(\tilde{\theta}(\tilde{\ell}_2), \theta) \lesssim \left(\sqrt{\frac{\delta d}{n}} + \frac{d}{n} \right) \wedge \left(\frac{\ell^{**} d}{n^2 \|\theta\|^2} + \frac{d}{n} \right),$$

with probability $1 - c_1 \log(n/d) \cdot e^{-c_0 d}$. In particular, we also have, globally, that

$$\ell^2(\tilde{\theta}(\tilde{\ell}_2), \theta) \lesssim \sqrt{\frac{\delta d}{n}} + \frac{d}{n},$$

with probability $1 - c_1 \log(n/d) \cdot e^{-c_0 d}$.

As a result, we have showed that full adaptation is possible for globally minimax optimal procedures (up to logarithmic factors). One may also observe that the bound in Theorem 14 is exactly the same as the one in Proposition 11 replacing ℓ^{**} by its value. Hence we recover the local estimation rates of $\tilde{\theta}(\ell^{**})$ (cf. Figure 1) in a fully adaptive fashion. In other terms, our adaptive procedure $\tilde{\theta}(\tilde{\ell}_2)$ is strictly better than the vanilla spectral method not only globally but also locally.

5 Conclusion and future work

5.1 Conclusion

As a conclusion, our contribution represents an improvement of previous results when δ is known. Not only we get rid of the logarithmic factors, but we show that our approach achieves optimal rates locally under tighter bounds and holds under (sub)Gaussian noise.

More specifically, we show that the optimal procedure depends on both δ and $\|\theta\|$. We then propose a semi-adaptive approach (adaptive to the signal), showing that minimax optimality requires only knowledge of δ. As for full adaptation, we claim that plug-in estimation of δ might not be optimal and that local minimax optimality is rather difficult without prior knowledge of δ. We finally propose an adaptive procedure that bypasses estimation of δ and that is globally minimax optimal.

In summary, our study refines center estimation in the high-dimensional Markovian binary GMM, and shows that we can reach the improved global rate $\sqrt{\delta d/n} + d/n$ instead of the classical rate $\sqrt{d/n} + d/n$ and that it can be done adaptively. Hence the hidden Markovian structure of the labels can be leveraged to improve center estimation without further information about this structure.

5.2 Future work

As for future research, one interesting direction is to characterize the adaptive locally minimax optimal rates where we conjecture the rates to be different from the non adaptive ones in general. An intriguing phenomenon observed during our study is that, when $d \leq n$, it seems possible to estimate the center of the mixture while impossible to estimate the parameter δ. As opposed to our intuition, the d-dimensional vector θ seems easier to estimate than the scalar δ. A very interesting question is to characterize the minimax rate of estimation for δ in the high dimensional mixture model. The challenging part, would be to capture the right dimension dependence in the lower bound. We believe that this is an interesting direction of research.

Another extension of our work could be a generalization of the current binary Markov chain into more sophisticated graph structures. In the current model, the assumption is that each new observation depends only on the previous one. We can extend the memory by assuming that each signs depends on two or more previous data points as an example.

Acknowledgments and Disclosure of Funding

The work of Mohamed Ndaoud is supported by a Chair of Excellence in Data Science granted by the CY Initiative.

References

Emmanuel Abbe, Jianqing Fan, and Kaizheng Wang. An ℓ_p theory of pca and spectral clustering. *The Annals of Statistics*, 50(4):2359–2385, 2022.

Kweku Abraham, Elisabeth Gassiat, and Zacharie Naulet. Fundamental limits for learning hidden markov model parameters. *IEEE Transactions on Information Theory*, 69(3): 1777–1794, march 2023. ISSN 1557-9654. doi: 10.1109/tit.2022.3213429. URL `http://dx.doi.org/10.1109/TIT.2022.3213429`.

Grigory Alexandrovich, Hajo Holzmann, and Anna Leister. Nonparametric identification and maximum likelihood estimation for hidden markov model, 2015.

Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. 2017.

Yohann De Castro, Élisabeth Gassiat, and Claire Lacour. Minimax adaptive estimation of nonparametric hidden markov models, 2015.

Xiaohui Chen and Yun Yang. Cutoff for exact recovery of gaussian mixture models. *IEEE Transactions on Information Theory*, 67(6):4223–4238, 2021.

Raaz Dwivedi, Koulik Khamaru, Martin J Wainwright, Michael I Jordan, et al. Theoretical guarantees for em under misspecified gaussian mixture models. *Advances in Neural Information Processing Systems*, 31, 2018.

Raaz Dwivedi, Nhat Ho, Koulik Khamaru, Martin J. Wainwright, Michael I. Jordan, and Bin Yu. Challenges with em in application to weakly identifiable mixture models. *ArXiv*, abs/1902.00194, 2019. URL `https://api.semanticscholar.org/CorpusID:59553479`.

Raaz Dwivedi, Nhat Ho, Koulik Khamaru, Michael I. Jordan, Martin J. Wainwright, and Bin Yu. Singularity, misspecification, and the convergence rate of em, 2020a.

Raaz Dwivedi, Nhat Ho, Koulik Khamaru, Martin Wainwright, Michael Jordan, and Bin Yu. Sharp analysis of expectation-maximization for weakly identifiable models. pages 1866–1876, 2020b.

Elisabeth Gassiat, Ibrahim Kaddouri, and Zacharie Naulet. Model-based clustering using non-parametric hidden markov models. *arXiv preprint arXiv:2309.12238*, 2023.

Christophe Giraud and Nicolas Verzelen. Partial recovery bounds for clustering with the relaxed k-means. *Mathematical Statistics and Learning*, 1(3):317–374, 2019.

Jason M Klusowski and WD Brinda. Statistical guarantees for estimating the centers of a two-component gaussian mixture by em. *arXiv preprint arXiv:1608.02280*, 2016.

Luc Lehéricy. Nonasymptotic control of the mle for misspecified nonparametric hidden markov models, 2021.

Matthias Löffler, Anderson Y Zhang, and Harrison H Zhou. Optimality of spectral clustering in the gaussian mixture model. *The Annals of Statistics*, 49(5):2506–2530, 2021.

Yu Lu and Harrison H Zhou. Statistical and computational guarantees of lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*, 2016.

Mohamed Ndaoud. Interplay of minimax estimation and minimax support recovery under sparsity. In *Algorithmic Learning Theory*, pages 647–668. PMLR, 2019.

Mohamed Ndaoud. Scaled minimax optimality in high-dimensional linear regression: A non-convex algorithmic regularization approach. *arXiv preprint arXiv:2008.12236*, 2020.

Mohamed Ndaoud. Sharp optimal recovery in the two component gaussian mixture model. *The Annals of Statistics*, 50(4):2096–2126, 2022.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. 2013.

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

Yihong Wu and Harrison H Zhou. Randomly initialized em algorithm for two-component gaussian mixture achieves near optimality in $o(\sqrt{n})$ iterations. *Mathematical Statistics and Learning*, 4(3), 2021.

Yihan Zhang and Nir Weinberger. Mean estimation in high-dimensional binary markov gaussian mixture models. *Advances in Neural Information Processing Systems*, 35:19673–19686, 2022.

Appendix A. Proofs of main results

A.1 Proof of Theorem 2

Let us consider the following estimator

$$\hat{\theta}(\ell) \coloneqq \sqrt{\frac{\ell}{\mathbf{E}\|\bar{\eta}\|^2}\left(\lambda_{max}(\hat{\Sigma}) - \frac{1}{k}\right)_+} v_{max}(\hat{\Sigma}). \tag{21}$$

Sometimes when the context is clear we will just write $\hat{\theta}$ instead of $\hat{\theta}(\ell)$. We remind the reader that $\Sigma = \frac{\mathbf{E}\|\bar{\eta}\|^2}{\ell}\theta\theta^T + \frac{\mathbf{I}_d}{k}$, it becomes clear that eigenvalues of Σ are $\lambda_{max}(\Sigma) = \|\theta\|^2\frac{\mathbf{E}\|\bar{\eta}\|^2}{\ell} + \frac{1}{k}$ and $\lambda_2 = ... = \lambda_d = \frac{1}{k}$.

We begin the analysis of our estimator by Weyl's inequality that states that

$$\left|\lambda_{max}(\hat{\Sigma}) - \lambda_{max}(\Sigma)\right| \leq \|\hat{\Sigma} - \Sigma\|_{op}. \tag{22}$$

In particular this leads to

$$\|\theta\|^2\frac{\mathbf{E}\|\bar{\eta}\|^2}{\ell} \leq \lambda_{max}(\hat{\Sigma}) - \frac{1}{k} + \|\hat{\Sigma} - \Sigma\|_{op}.$$

In case where $\|\hat{\theta}\| = 0$, we have $\lambda_{max}(\hat{\Sigma}) - \frac{1}{k} \leq 0$, so

$$\|\theta\|^2\frac{\mathbf{E}\|\bar{\eta}\|^2}{\ell} \leq \|\hat{\Sigma} - \Sigma\|_{op}.$$

Hence it holds that

$$|\|\hat{\theta}\| - \|\theta\|| \leq \sqrt{\frac{\ell}{\mathbf{E}\|\bar{\eta}\|^2}}\frac{\|\hat{\Sigma} - \Sigma\|_{op}}{\|\theta\|}. \tag{23}$$

Moreover we also have for $\|\hat{\theta}\| \neq 0$ that

$$\begin{aligned}
\left|\|\hat{\theta}\| - \|\theta\|\right| &= \frac{\left|\|\hat{\theta}\|^2 - \|\theta\|^2\right|}{\|\hat{\theta}\| + \|\theta\|} \\
&\leq \frac{\left|\|\hat{\theta}\|^2 - \|\theta\|^2\right|}{\|\theta\|} \\
&= \frac{\ell}{\mathbf{E}\|\bar{\eta}\|^2}\frac{\left|\lambda_{max}(\hat{\Sigma}) - \lambda_{max}(\Sigma)\right|}{\|\theta\|} \\
&\leq \frac{\ell}{\mathbf{E}\|\bar{\eta}\|^2}\frac{\|\hat{\Sigma} - \Sigma\|_{op}}{\|\theta\|}.
\end{aligned}$$

We conclude that we always have

$$|\|\hat{\theta}\| - \|\theta\|| \leq \frac{\ell}{\mathbf{E}\|\bar{\eta}\|^2}\frac{\|\hat{\Sigma} - \Sigma\|_{op}}{\|\theta\|}, \tag{24}$$

since $\mathbf{E}\|\bar{\eta}\|^2 \leq \ell$. Next observe that the largest eigengap of Σ is $\frac{\mathbf{E}\|\bar{\eta}\|^2}{\ell}\|\theta\|^2$. Hence, and using the Davis-Kahan's perturbation bound, we get that

$$\ell(\hat{v}_{max}, v_{max}) \leq \frac{4\ell}{\mathbf{E}\|\bar{\eta}\|^2}\frac{\|\hat{\Sigma}-\Sigma\|_{op}}{\|\theta\|^2}. \tag{25}$$

Combining (24) and (25) we finally get that

$$\ell(\hat{\theta}, \theta) \leq \left|\|\hat{\theta}\|-\|\theta\|\right| + \|\theta\|\ell(\hat{v}_{max}, v_{max}) \leq \frac{5\ell}{\mathbf{E}\|\bar{\eta}\|^2}\frac{\|\hat{\Sigma}-\Sigma\|_{op}}{\|\theta\|}. \tag{26}$$

Now we continue with the analysis of the difference of covariance and sample covariance matrices. Recall that

$$\hat{\Sigma}-\Sigma = \frac{\theta\theta^T(\|\bar{\eta}\|^2-\mathbf{E}\|\bar{\eta}\|^2)}{\ell} + \frac{w\bar{\eta}\cdot\theta^T + \theta\cdot(w\bar{\eta})^T}{\sqrt{k}\ell} + \frac{1}{k}\left(\frac{ww^T}{\ell}-\mathbf{I}_d\right). \tag{27}$$

Observe that for $\xi := w\bar{\eta}/\|\bar{\eta}\|$ is a $1-$sub-Gaussian vector. Hence

$$\left\|\frac{1}{\sqrt{k}\ell}\left((w\bar{\eta})\theta^T + \theta(w\bar{\eta})^T\right)\right\|_{op} \leq \frac{2}{\sqrt{k}\ell}\|\theta\|\|\bar{\eta}\|\|\xi\|.$$

As a result

$$\|\hat{\Sigma}-\Sigma\|_{op} \leq \frac{\|\theta\|^2|\|\bar{\eta}\|^2-\mathbf{E}\|\bar{\eta}\|^2|}{\ell} + \frac{2}{\sqrt{k}\ell}\|\theta\|\|\bar{\eta}\|\|\xi\| + \frac{1}{k}\left\|\frac{ww^T}{\ell}-I_d\right\|_{op} \tag{28}$$

and using Lemma 17 we conclude that

$$\|\hat{\Sigma}-\Sigma\|_{op} \lesssim \sqrt{\frac{d}{\ell}}\|\theta\|^2 + \frac{\sqrt{\ell d}}{\sqrt{k}\ell}\|\theta\| + \frac{d}{k\ell} + \frac{1}{k}\sqrt{\frac{d}{\ell}} \lesssim \sqrt{\frac{d}{\ell}}\|\theta\|^2 + \sqrt{\frac{d}{n}}\|\theta\| + \left(\frac{d}{n} + \frac{1}{k}\sqrt{\frac{d}{\ell}}\right). \tag{29}$$

Combining (29) and (26) we end up with the desired loss bound

$$\ell(\hat{\theta}, \theta) \lesssim \sqrt{\frac{d}{\ell}}\|\theta\| + \sqrt{\frac{d}{n}} + \frac{1}{\|\theta\|}\left(\sqrt{\frac{d}{kn}} + \frac{d}{n}\right), \tag{30}$$

with probability greater than $1-c_1e^{-c_0 d}$.

For the case $\|\theta\| \geq 1$, and by choosing $\ell = n$ there is no bucketing involved and the procedure boils down to the vanilla spectral method. We repeat the same proof except that now $\mathbf{E}\|\bar{\eta}\|^2 = \ell$ and hence the difference of covariance matrices is decomposed as follows

$$\hat{\Sigma}-\Sigma = \frac{(w\bar{\eta})\theta^T + \theta(w\bar{\eta})^T}{n} + \left(\frac{ww^T}{n}-\mathbf{I}_d\right).$$

It comes out that with probability greater than $1-c_1e^{-c_0 d}$, we have

$$\|\hat{\Sigma}-\Sigma\|_{op} \lesssim \sqrt{\frac{d}{n}}\|\theta\| + \left(\frac{d}{n} + \sqrt{\frac{d}{n}}\right).$$

Lastly, we get, with probability greater than $1-c_1e^{-c_0 d}$, that

$$\ell(\hat{\theta}, \theta) \lesssim \frac{\|\hat{\Sigma}-\Sigma\|_{op}}{\|\theta\|} \lesssim \sqrt{\frac{d}{n}} + \frac{1}{\|\theta\|}\left(\frac{d}{n} + \sqrt{\frac{d}{n}}\right) \lesssim \sqrt{\frac{d}{n}} + \frac{d}{n},$$

since $\|\theta\| \geq 1$.

A.2 Proof of Proposition 3

Recall that $\ell^* := d \vee \lceil n(\delta \vee \|\theta\|^2) \rceil \wedge n$. For the rest of the proof we denote $\hat{\theta} := \hat{\theta}(\ell^*)$. We begin by observing that for the loss function $\ell(\cdot, \cdot)$ we also have the following decomposition

$$\begin{aligned}
\ell^2(\hat{\theta}, \theta) &\leq 2\|\hat{\theta}\|^2 + 2\|\theta\|^2 \\
&\leq 2\left| \|\hat{\theta}\|^2 - \|\theta\|^2 \right| + 4\|\theta\|^2 \\
&\leq \frac{\ell^*}{2\mathbf{E}\|\bar{\eta}\|^2} \|\hat{\Sigma} - \Sigma\|_{op} + 6\|\theta\|^2 \\
&\lesssim \sqrt{\frac{d}{n}}\|\theta\| + \left(\frac{d}{n} + \frac{1}{k^*}\sqrt{\frac{d}{\ell^*}} \right) + \sqrt{\frac{d}{\ell^*}}\|\theta\|^2 + \|\theta\|^2,
\end{aligned}$$

with probability greater than $1 - c_1 e^{-c_0 d}$, where we have used the upper bound (29), and the fact that $\mathbf{E}\|\bar{\eta}\|^2 \geq \ell^*/3$ based on Lemma 1. Hence as long as $\|\theta\|^2 \leq \sqrt{\delta d/n} \vee d/n$ we also have that $\|\theta\|^2 \leq \delta \vee d/n$ and hence $\ell^* = d \vee \lceil n\delta \rceil$. It comes out that

$$\ell^2(\hat{\theta}, \theta) \lesssim \sqrt{\frac{\delta d}{n}} + \frac{d}{n}. \tag{31}$$

Next, in the scenario where $\sqrt{\frac{\delta d}{n}} \vee \frac{d}{n} \leq \|\theta\|^2 \leq \delta \vee \frac{d}{n}$. This case only makes sense for $\delta \geq d/n$. Then, by invoking Theorem 2 with the number of buckets being $\ell^* = \lceil n\delta \rceil$ we get that

$$\ell(\hat{\theta}, \theta) \lesssim \sqrt{\frac{d}{n}} \frac{\|\theta\|}{\sqrt{\delta}} + \sqrt{\frac{d}{n}} + \frac{1}{\|\theta\|}\sqrt{\frac{\delta d}{n}} \lesssim \sqrt{\frac{d}{n}} + \frac{\sqrt{\frac{\delta d}{n}}}{\|\theta\|} \lesssim \frac{\sqrt{\frac{\delta d}{n}}}{\|\theta\|}.$$

For the case $\delta \vee \frac{d}{n} \leq \|\theta\|^2 \leq 1$, then we have $\ell^* = \lceil n\|\theta\|^2 \rceil$. Using Theorem 2 one more time we get

$$\ell(\hat{\theta}, \theta) \lesssim \sqrt{\frac{d}{n\|\theta\|^2}}\|\theta\| + \sqrt{\frac{d}{n}} + \frac{1}{\|\theta\|}\sqrt{\frac{\|\theta\|^2 d}{n}} \lesssim \sqrt{\frac{d}{n}}.$$

Finally, for $\|\theta\| \geq 1$, we get that $\ell^* = n$ and we simply conclude using the corresponding case in Theorem 2. This proof is now complete.

A.3 Proof of Lemma 6

We consider first the case $\|\theta\|^2 \geq \delta \vee Cd/n$. Using (30) we have

$$|\hat{s} - \|\theta\|| \lesssim \sqrt{\frac{d}{\hat{\ell}}}\|\theta\| + \sqrt{\frac{d}{n}} + \frac{1}{\|\theta\|}\left(\sqrt{\frac{1}{\hat{k}}}\sqrt{\frac{d}{n}} + \frac{d}{n} \right). \tag{32}$$

Plugging $\hat{k} \sim \frac{1}{\delta} \wedge \frac{n}{d}$ and using $\|\theta\|^2 \geq \delta \vee Cd/n$ we get

$$|\hat{s} - \|\theta\|| \lesssim \sqrt{\frac{1}{C}}\|\theta\| + \sqrt{\frac{d}{n}}.$$

It comes out, since C is large enough, that

$$|\hat{s} - \|\theta\|| \leq \frac{\|\theta\|}{4} + \frac{\|\theta\|}{4} = \frac{\|\theta\|}{2}.$$

Hence with probability greater than $1 - c_1 e^{-c_0 d}$, we get

$$\frac{\|\theta\|}{2} \leq \hat{s} \leq \frac{3\|\theta\|}{2}.$$

Now suppose $\delta \vee Cd/n > \|\theta\|^2$, and using the control of $\|\hat{\Sigma} - \Sigma\|_{op}$, we get

$$\left|\hat{s}^2 - \|\theta\|^2\right| \lesssim \sqrt{\frac{d}{(n\delta) \vee (Cd)}}\|\theta\|^2 + \sqrt{\frac{d}{n}}\|\theta\| + \sqrt{\frac{(\delta \vee Cd/n)d}{n}}.$$

Then we conclude that

$$\left|\hat{s}^2 - \|\theta\|^2\right| \leq 2(\delta \vee Cd/n). \tag{33}$$

It follows that $\hat{s}^2 < 3(\delta \vee Cd/n)$ with probability greater than $1 - c_1 e^{-c_0 d}$.

A.4 Proof of Theorem 8

For this proof, observe that $\hat{\ell}$ depends on the observations so we cannot apply directly Lemma 17 replacing ℓ by $\hat{\ell}$. Since we only need to know $\hat{\ell}$ up to a multiplicative factor, we will choose $\hat{\ell}$ on a logarithmic grid. Basically we will choose $\hat{\ell} = d2^{\hat{m}} \wedge n$ where $\hat{m}$ is the smallest integer such that $\hat{\ell} \geq \lceil n(3\delta \vee \hat{s}^2) \vee Cd \rceil \wedge n$. Hence and in order to control the deviation terms in Lemma 17 we shall take a supremum for all values of m that are no more than $\log(n/d)$. It comes out that Lemma 17 holds for $\hat{\ell}$ with probability $1 - c_1 \log(n/d) e^{-c_0 d}$.

We recall here that using Corollary 7 we have, with probability $1 - c_1 e^{-c_0 d}$, if $\hat{s}^2 < 3(\delta \vee Cd/n)$ we get that $\|\theta\|^2 < 12(\delta \vee Cd/n)$, and if $\hat{s}^2 \geq 3(\delta \vee Cd/n)$ we get that $\|\theta\|^2 \geq \delta \vee Cd/n$. We analyse the error case by case.

- **Case 1:** $\|\theta\|^2 \leq \sqrt{\frac{\delta d}{n}} \vee \frac{d}{n}$.

  In this case, since $\sqrt{\frac{\delta d}{n}} \leq (\delta \vee Cd/n)$, then we know using Lemma 6 that $\hat{s}^2 < 3\delta$ and $\ell \sim n\delta \vee d$. As shown before in (31) it comes out that $\ell^2(\hat{\theta}(\hat{\ell}), \theta) \lesssim \sqrt{\frac{\delta d}{n}} + \frac{d}{n}$.

- **Case 2:** $\sqrt{\frac{\delta d}{n}} \vee \frac{d}{n} \leq \|\theta\|^2 < \delta \vee \frac{d}{n}$.
  In this case again $\hat{s}^2 < 3\delta$ and $\ell \sim n\delta \vee d$, according to Corollary 7. Continuing with $\ell \sim n\delta \vee d$, and mimicking the corresponding case of Proposition 3, we get

  $$\ell(\hat{\theta}(\hat{\ell}), \theta) \lesssim \sqrt{\frac{d}{n}} \frac{\|\theta\|}{\sqrt{\delta + Cd/n}} + \sqrt{\frac{d}{n}} + \frac{1}{\|\theta\|}\sqrt{\frac{(\delta \vee Cd/n)d}{n}} \lesssim \sqrt{\frac{d}{n}} + \frac{\sqrt{\frac{(\delta + Cd/n)d}{n}}}{\|\theta\|} \lesssim \frac{\sqrt{\frac{(\delta + d/n)d}{n}}}{\|\theta\|}.$$

- **Case 3:** $\delta \vee \frac{d}{n} \leq \|\theta\|^2 < 4$.
  In this case we have two options. Either $12(\delta \vee Cd/n) \leq \|\theta\|^2 \leq 4$ and hence $\hat{s}^2 \geq$

$3(\delta \vee Cd/n)$ and $\ell \sim n\hat{s}^2$. That being the case, similarly to Proposition 3, by taking take $\ell \sim n\|\theta\|^2$ replacing $\|\theta\|$ with $\hat{s}$ we derive

$$\ell(\hat{\theta}(\hat{\ell}), \theta) \lesssim \sqrt{\frac{d}{n\hat{s}^2}}\|\theta\| + \sqrt{\frac{d}{n}} + \frac{\hat{s}}{\|\theta\|}\sqrt{\frac{d}{n}}.$$

Using Lemma 6, we conclude that

$$\ell(\hat{\theta}(\hat{\ell}), \theta) \lesssim \sqrt{\frac{d}{n}} + \sqrt{\frac{d}{n}} + \sqrt{\frac{d}{n}} \lesssim \sqrt{\frac{d}{n}}.$$

In the remaining case where $\delta \vee \frac{d}{n} \leq \|\theta\|^2 \leq 12(\delta \vee C\frac{d}{n})$ we know that $\hat{\ell}$ will be of order $n\delta \vee d$ and the result is straightforward.

- Finally if $\|\theta\|^2 \geq 4$, then $\hat{s}^2 \geq 1$ and $\hat{\ell} = n$. The result follows immediately.

Proof of Theorem 10

Repeating the same decomposition as before in (26) we get

$$\ell(\tilde{\theta}, \theta) \leq \Big|\|\tilde{\theta}\| - \|\theta\|\Big| + \|\theta\|\ell(\hat{v}_{max}, v_{max}) \lesssim \frac{\|\hat{\Sigma} - \Sigma^*\|_{op}}{\|\theta\|}. \tag{34}$$

The covariance decomposition is slightly different this time as we get

$$\hat{\Sigma} - \Sigma^* = \frac{\theta\theta^T(\|\bar{\eta}\|^2 - \mathbf{E}\|\bar{\eta}\|^2 + \mathbf{E}\|\bar{\eta}\|^2 - \ell)}{\ell} + \frac{w\bar{\eta} \cdot \theta^T + \theta \cdot (w\bar{\eta})^T}{\sqrt{k}\ell} + \frac{1}{k}\left(\frac{ww^T}{\ell} - \mathbf{I}_d\right).$$

$$\hat{\Sigma} - \Sigma^* = \hat{\Sigma} - \Sigma + \frac{\theta\theta^T(\mathbf{E}\|\bar{\eta}\|^2 - \ell)}{\ell} \tag{35}$$

Since $|\mathbf{E}\|\bar{\eta}\|^2 - \ell| \leq n\delta$ as claimed in Lemma 1, we get that

$$\ell(\tilde{\theta}, \theta) \lesssim \frac{\|\hat{\Sigma} - \Sigma\|_{op}}{\|\theta\|} + \frac{n\delta}{\ell}\|\theta\|.$$

Recalling our concentration bound as in (30) we finally derive that

$$\ell(\tilde{\theta}, \theta) \lesssim \left(\sqrt{\frac{d}{\ell}} + \frac{n\delta}{\ell}\right)\|\theta\| + \sqrt{\frac{d}{n}} + \frac{1}{\|\theta\|}\left(\sqrt{\frac{1}{k}}\sqrt{\frac{d}{n}} + \frac{d}{n}\right), \tag{36}$$

with probability $1 - c_1 e^{-c_0 d}$.

Proof of Proposition 11

Recall that $\ell^{**} = d \vee \left[\left(n\delta \vee \frac{\delta^{2/3} n^{4/3} \|\theta\|^{4/3}}{d^{1/3}} \vee n\|\theta\|^2\right)\right] \wedge n$. For the rest of the proof, and for simplicity, we denote $\tilde{\theta}(\ell^{**})$ by $\tilde{\theta}$ and ℓ^{**} by ℓ. To begin with we analyse the loss rate when signal strength $\|\theta\|^2$ is small.

Following similar steps as in the proof of the proposition 3 we have

$$\begin{aligned}
\ell^2(\tilde{\theta}, \theta) &\leq 2\left|\|\tilde{\theta}\|^2 - \|\theta\|^2\right| + 4\|\theta\|^2 \\
&\leq 2\,\|\hat{\Sigma} - \Sigma^*\|_{op} \;+\; 4\|\theta\|^2 \\
&\leq 2\,\|\hat{\Sigma} - \Sigma\|_{op} \;+\; \frac{2n\delta}{\ell}\|\theta\| \;+\; 4\|\theta\|^2 \\
&\lesssim \sqrt{\frac{d}{n}}\|\theta\| + \left(\frac{d}{n} + \frac{1}{k}\sqrt{\frac{d}{\ell}}\right) + \sqrt{\frac{d}{\ell}}\|\theta\|^2 + \|\theta\|^2,
\end{aligned}$$

using the covariance decomposition (35), along with the fact that $\frac{n\delta}{\ell} \leq 1$ by the choice of ℓ.

$$\ell^2(\tilde{\theta}, \theta) \lesssim \sqrt{\frac{d}{n}}\|\theta\| + \left(\frac{d}{n} + \frac{1}{k}\sqrt{\frac{d}{\ell}}\right) + \left(1 + \sqrt{\frac{d}{\ell}}\right)\|\theta\|^2, \tag{37}$$

for $\|\theta\|^2 \leq \sqrt{\frac{\delta d}{n}} \vee \frac{d}{n}$. Observe that in that case $\ell^{**} \sim n\delta \vee d$ and hence

$$\begin{aligned}
\ell^2(\tilde{\theta}, \theta) &\lesssim \sqrt{\frac{d}{n}}\|\theta\| + \left(\frac{d}{n} + \delta\sqrt{\frac{d}{n\delta}}\right) + \left(1 + \sqrt{\frac{d}{n\delta}}\right)\|\theta\|^2 \\
&= \sqrt{\frac{d}{n}}\|\theta\| + \frac{d}{n} + \sqrt{\frac{\delta d}{n}} + \|\theta\|^2 + \sqrt{\frac{d}{n\delta}}\|\theta\|^2 \\
&\leq \sqrt{\frac{\delta d}{n}} + \frac{d}{n}.
\end{aligned}$$

Next, let's consider two cases for δ. For the sake of convenience in the subsequent lines the bounds are derived for the loss function itself, not for the square of it.

- Case 1: $\delta \leq \sqrt{\frac{d}{n}}$

  Here, in the scenario when $\sqrt{\frac{\delta d}{n}} \vee \frac{d}{n} \leq \|\theta\|^2 \leq \frac{\delta^2 n}{d} \vee \frac{d}{n}$ we continue the process with the following number of buckets:

  $$\ell \sim \frac{\delta^{\frac{2}{3}} n^{\frac{4}{3}} \|\theta\|^{\frac{4}{3}}}{d^{\frac{1}{3}}}.$$

  Observe that $\ell \leq n$ as $\delta \leq \sqrt{d/n}$. Now let's analyse the loss bound (36) with this choice of ℓ.

  $$\sqrt{\frac{d}{\ell}} \sim \frac{d^{2/3}}{\delta^{1/3} n^{2/3} \|\theta\|^{2/3}}, \;\; \frac{n\delta}{\ell} \sim \frac{\delta^{1/3} d^{1/3}}{n^{1/3}\|\theta\|^{4/3}}, \;\; \sqrt{\frac{d\ell}{n^2}} \sim \frac{\delta^{1/3} d^{1/3} \|\theta\|^{2/3}}{n^{1/3}}.$$

Hence

$$\ell(\tilde{\theta},\theta) \lesssim \frac{d^{2/3}\|\theta\|^{1/3}}{\delta^{1/3}n^{2/3}} + \frac{\delta^{1/3}d^{1/3}}{n^{1/3}\|\theta\|^{1/3}} + \sqrt{\frac{d}{n}} \lesssim \left(\frac{\frac{\delta d}{n}}{\|\theta\|}\right)^{\frac{1}{3}}.$$

Next, if $\frac{\delta^2 n}{d} \vee \frac{d}{n} \leq \|\theta\|^2 \leq 1$ we continue with

$$\ell \sim n\|\theta\|^2,$$

achieving the parametric rate

$$\ell(\tilde{\theta},\theta) \lesssim \sqrt{\frac{d}{n}} + \frac{\delta}{\|\theta\|} + \sqrt{\frac{d}{n}} + \sqrt{\frac{d}{n}} \lesssim \sqrt{\frac{d}{n}}.$$

Lastly when $\|\theta\|^2 \geq 1$ the parametric rate can be recovered with $\ell = n$.

- Case 2 : $\sqrt{\frac{d}{n}} \leq \delta$

  The first scenario is when $\sqrt{\frac{\delta d}{n}} \leq \|\theta\|^2 \leq \frac{\sqrt{\frac{d}{n}}}{\delta}$. Here we have that

  $$\ell \sim \frac{\delta^{\frac{2}{3}} n^{\frac{4}{3}} \|\theta\|^{\frac{4}{3}}}{d^{\frac{1}{3}}}.$$

  Again, plugging in ℓ in (36) and using $\sqrt{\frac{d}{n}} \leq \delta$ we get moreover that

  $$\ell(\tilde{\theta},\theta) \lesssim \frac{d^{2/3}\|\theta\|^{1/3}}{\delta^{1/3}n^{2/3}} + \frac{\delta^{1/3}d^{1/3}}{n^{1/3}\|\theta\|^{1/3}} + \sqrt{\frac{d}{n}} \leq \left(\frac{\delta^{1/3}d^{1/3}}{n^{1/3}\|\theta\|^{1/3}} + \sqrt{\frac{d}{n}}\right) \lesssim \left(\frac{\frac{\delta d}{n}}{\|\theta\|}\right)^{\frac{1}{3}}.$$

  Furthermore, if $\frac{\sqrt{\frac{d}{n}}}{\delta} \leq \|\theta\|^2$ with the choice of $\ell = n$ we recover the usual rate of the vanilla spectral method,

  $$\ell(\tilde{\theta},\theta) \lesssim \sqrt{\frac{d}{n}} + \frac{1}{\|\theta\|}\sqrt{\frac{d}{n}} \lesssim \begin{cases} \frac{\sqrt{\frac{d}{n}}}{\|\theta\|}, & \|\theta\|^2 \leq 1 \\ \sqrt{\frac{d}{n}}, & 1 \leq \|\theta\|^2. \end{cases}$$

  In both cases the compact formula for ℓ is given by

  $$\ell^{**} \sim d \vee \left(n\delta \vee \frac{\delta^{2/3}n^{4/3}\|\theta\|^{4/3}}{d^{1/3}}\right) \wedge n.$$

Remark 15 *It is rather convenient to have one compact form for the number of buckets in both cases of Proposition 11. In that respect, let's notice that in Case 2 when $\|\theta\| \leq 1$ then $\ell^{**} \geq n\|\theta\|^2$. Indeed, since $\delta \geq d/n$, we have that either*

$$\ell_1 = n\delta \geq n\sqrt{\frac{\delta d}{n}} \geq n\|\theta\|^2,$$

or

$$\ell_2 = \frac{\delta^{\frac{2}{3}} n^{\frac{4}{3}} \|\theta\|^{\frac{4}{3}}}{d^{\frac{1}{3}}} \geq \frac{d^{\frac{1}{3}} n^{\frac{4}{3}} \|\theta\|^{\frac{4}{3}}}{n^{\frac{1}{3}} d^{\frac{1}{3}}} = n\|\theta\|^{\frac{4}{3}} \geq n\|\theta\|^2.$$

This means that in Case 2 taking the maximum with additional $n\|\theta\|^2$ won't change anything

$$\left(n\delta \vee \frac{\delta^{2/3} n^{4/3} \|\theta\|^{4/3}}{d^{1/3}} \right) = \left(n\delta \vee \frac{\delta^{2/3} n^{4/3} \|\theta\|^{4/3}}{d^{1/3}} \vee n\|\theta\|^2 \right).$$

Therefore we can say that ℓ^{**} takes the following general form

$$\ell^{**} \sim d \vee \left(n\delta \vee \frac{\delta^{2/3} n^{4/3} \|\theta\|^{4/3}}{d^{1/3}} \vee n\|\theta\|^2 \right) \wedge n.$$

As a consequence of this, let's declare that with this choice we have one that either $\ell^{**} = n$ or the following three inequalities hold together

$$\ell^{**} \geq n\delta \vee d, \qquad \ell^{**} \geq \frac{\delta^{2/3} n^{4/3} \|\theta\|^{4/3}}{d^{1/3}}, \qquad \ell^{**} \geq n\|\theta\|^2. \tag{38}$$

This observation will be useful later on.

A.5 Proof of Theorem 13

For this proof, let us denote $\tilde{\theta} := \tilde{\theta}(\tilde{\ell})$, $\tilde{\theta}_k := \tilde{\theta}(g_k)$ and $\omega_k^2 := 4C^2 \left(\frac{\sqrt{dg_k}}{n} \wedge \frac{dg_k}{n^2 \|\tilde{\theta}(g_k)\|^2} \right)$. We have for any ω,

$$\mathbf{P}(\ell(\tilde{\theta}, \theta) \geq \omega) = \mathbf{P}\left(\ell(\tilde{\theta}, \theta) \geq \omega \ \cap \ \{\tilde{\ell} \leq \ell^{**}\} \right) + \mathbf{P}\left(\ell(\tilde{\theta}, \theta) \geq \omega \ \cap \ \{\tilde{\ell} > \ell^{**}\} \right) := I_1 + I_2. \tag{39}$$

We also recall using (37) that for $g_k \geq n\delta$ we have

$$\left| \|\tilde{\theta}(g_k)\|^2 - \|\theta\|^2 \right| \lesssim \sqrt{\frac{d}{n}} \|\theta\| + \left(\frac{d}{n} + \sqrt{\frac{g_k d}{n^2}} \right) + \left(\frac{n\delta}{g_k} + \sqrt{\frac{d}{g_k}} \right) \|\theta\|^2, \tag{40}$$

with probability $1 - c_1 \cdot e^{-c_0 d}$.

Bounding I_1: We begin by defining the smallest element g_{m_*} from grid G that is greater than ℓ^{**}. Our intention here is to replace the unknown ℓ^{**} with a ℓ in a close range, for which further analysis is tractable.
Formally,

$$m_* := \min\{m \in \{0, 1, \dots, M\} \ : \ \ell^{**} \leq g_m\}.$$

By this definition $g_{m_*-1} < \ell^{**}$, hence g_{m_*} differs from ℓ^{**} by a factor of 2 at most, since $g_{m_*} = 2g_{m_*-1} < 2\ell^{**}$. Let ϵ_* and ϵ_k be signs such that $\ell(\tilde{\theta}_{m_*}, \theta) = \|\epsilon_{m^*} \tilde{\theta}_{m_*} - \theta\|$ and for all $\tilde{m} + 1 \leq k \leq m_*$ we have $\ell(\tilde{\theta}_k, \tilde{\theta}_{k-1}) = \|\epsilon_k \tilde{\theta}_k - \epsilon_{k-1} \tilde{\theta}_{k-1}\|$.

We decompose the difference using triangle inequality by stepping through grid values until we reach the first number of buckets g_{m_*}.

$$\ell(\tilde{\theta}, \theta) \leq \|\tilde{\theta} - \epsilon_{\tilde{m}} \theta\| \leq \|\epsilon_{\tilde{m}} \tilde{\theta}_{g_{\tilde{m}}} - \theta\| \leq \sum_{k=\tilde{m}+1}^{m_*} \|\epsilon_k \tilde{\theta}_k - \epsilon_{k-1} \tilde{\theta}_{k-1}\| + \|\epsilon_{m^*} \tilde{\theta}_{m_*} - \theta\| \leq \sum_{k=1}^{m_*} \omega_k + \ell(\tilde{\theta}_{m_*}, \theta).$$

In order to deal with the sum $\sum_{k=1}^{m_*}\omega_k$ we decompose the sum into two sums $\sum_{k=1}^{m_1}\omega_k$ and $\sum_{k=m_1+1}^{m_*}\omega_k$, where $g_{m_1}\sim n\delta$. It is straightforward, using gemetric series, that

$$\left(\sum_{k=1}^{m_1}\omega_k\right)^2 \lesssim \sqrt{\frac{d\delta}{n}}+\frac{d}{n}.$$

It remains to control $\sum_{k=m_1+1}^{m_*}\omega_k$ where all $g_k\geq n\delta$.

On the one hand, when $\|\theta\|^2\leq \frac{\sqrt{d\ell^{**}}}{n}$, we have that $w_k^2\leq 4C^2\frac{\sqrt{dg_k}}{n}$ and the sum $\sum_{k=\tilde{m}+1}^{m_*}\omega_k$ can be upper bounded by a sum of a geometric progression, and so

$$\sum_{k=\tilde{m}+1}^{m_*}\omega_k\leq\sum_{k=1}^{m_*}\omega_k=2C\frac{\sqrt[4]{dg_1}}{\sqrt{n}}\sum_{k=1}^{m_*}(\sqrt[4]{2})^k=2C\frac{\sqrt[4]{dg_1}}{\sqrt{n}}\frac{(\sqrt[4]{2})^{m_*+1}-1}{\sqrt[4]{2}-1}\leq 2C\frac{\sqrt[4]{dg_{m_*+1}}}{\sqrt{n}}.$$

It comes out that

$$\left(\sum_{k=m_1}^{m_*}\omega_k\right)^2\lesssim\left(\sqrt{\frac{\delta d}{n}}+\frac{d}{n}\right).$$

On the other hand, when $\frac{\sqrt{d\ell^{**}}}{n}\leq\|\theta\|^2\leq 1$, we have, with probability greater than $1-c_1\cdot e^{-c_0 d}$, that

$$\|\tilde{\theta}(g_k)\|^2\gtrsim\|\theta\|^2.$$

It comes out that $w_k^2\lesssim 4C^2\frac{dg_k}{n^2\|\theta\|^2}$ and the sum $\sum_{k=\tilde{m}+1}^{m_*}\omega_k$ can be upper bounded by a sum of a geometric progression, and so

$$\left(\sum_{k=m_1}^{m_*}\omega_k\right)^2\lesssim\frac{d\ell^{**}}{n^2\|\theta\|^2}.$$

It comes out that

$$\left(\sum_{k=m_1}^{m_*}\omega_k\right)^2\lesssim\frac{d\ell^{**}}{n^2\|\theta\|^2}\wedge\left(\sqrt{\frac{\delta d}{n}}+\frac{d}{n}\right).$$

We have by Proposition 11, that, with probability greater than $1-c_1e^{-c_0d}$,

$$\ell^2(\tilde{\theta}_{m_*},\theta)\lesssim\left(\sqrt{\frac{\delta d}{n}}+\frac{d}{n}\right)\wedge\left(\frac{\ell^{**}d}{n\|\theta\|^2}+\frac{d}{n}\right).$$

Hence we get that as long as $\tilde{\ell}\leq\ell^{**}$, then with probability greater than $1-c_1\cdot e^{-c_0d}$, we have

$$\ell^2(\tilde{\theta},\theta)\lesssim\left(\sqrt{\frac{\delta d}{n}}+\frac{d}{n}\right).$$

Bounding I_2: In this case $g_k\geq\ell^{**}$. Notice first that, when $\|\theta\|^2\leq\frac{\sqrt{dg_k}}{n}$, we have that $\|\tilde{\theta}(g_k)\|^2\lesssim\frac{\sqrt{dg_k}}{n}$ and hence $w_k^2\gtrsim\frac{\sqrt{dg_k}}{n}$ and when $\|\theta\|^2\geq\frac{\sqrt{dg_k}}{n}$ then $\|\tilde{\theta}(g_k)\|^2\lesssim\|\theta\|^2$ and $w_k^2\gtrsim\frac{dg_k}{n^2\|\theta\|^2}$. It comes out that as long as $g_k\geq\ell^{**}$ we have that

$$w_k^2\gtrsim\frac{\sqrt{dg_k}}{n}\wedge\frac{dg_k}{n^2\|\theta\|^2}.$$

In this case

$$\begin{aligned} I_2 \le \mathbf{P}(\tilde{\ell} > \ell^{**}) &= \mathbf{P}(\tilde{m} > \ell^{**}) \\ &\le \sum_{m:\ g_m > \ell^{**}} \mathbf{P}(\tilde{\ell} = g_m) \\ &\le \sum_{m:\ g_m > \ell^{**}} \mathbf{P}(\ell(\tilde{\theta}_m, \tilde{\theta}_{m-1}) > \omega_m) \\ &\le \sum_{m:\ g_m > \ell^{**}} \mathbf{P}(\ell(\tilde{\theta}_m, \theta) + \ell(\tilde{\theta}_{m-1}, \theta) > \omega_m) \\ &\le 2 \sum_{m:\ g_m \ge \ell^{**}} \mathbf{P}\left(\ell(\tilde{\theta}_m, \theta) > \frac{\omega_m}{2}\right) \\ &\overset{(*)}{\lesssim} 2|G| e^{-c_0 d} \\ &\lesssim \log\left(\frac{n}{d}\right) e^{-c_0 d}. \end{aligned}$$

where $(*)$ holds, because for every $g_m \ge \ell^{**}$ due to Lemma 12 we have

$$\ell^2(\tilde{\theta}_{g_m}, \theta) \lesssim w_k^2 \qquad \text{w.p.} \quad 1 - c_1 \cdot e^{-c_0 d}.$$

A.6 Proof of Theorem 14

For this proof we will mimick the proof of Theorem 13. Let us denote $\tilde{\theta} := \tilde{\theta}(\tilde{\ell}_2)$, $\tilde{\theta}_k := \tilde{\theta}(g_k)$ and $\omega_k^2 := 4C^2 \left(\frac{\sqrt{d g_k}}{n} \wedge \frac{d g_k}{n^2 \|\tilde{\theta}(\tilde{\ell})\|^2}\right)$. Notice that, with probability $1 - c_1 \log(n/d) e^{-c_0 d}$, we have

$$\left| \|\tilde{\theta}(\tilde{\ell})\| - \|\theta\| \right|^2 \lesssim \frac{d}{n} + \sqrt{\frac{d\delta}{n}}.$$

We have for any ω,

$$\mathbf{P}(\ell(\tilde{\theta}, \theta) \ge \omega) = \mathbf{P}\left(\ell(\tilde{\theta}, \theta) \ge \omega \ \cap\ \{\tilde{\ell}_2 \le \ell^{**}\}\right) + \mathbf{P}\left(\ell(\tilde{\theta}, \theta) \ge \omega \ \cap\ \{\tilde{\ell}_2 > \ell^{**}\}\right) := I_1 + I_2. \tag{41}$$

Bounding I_1: We begin by defining the smallest element g_{m_*} from grid G that is greater than ℓ^{**}. Our intention here is to replace the unknown ℓ^{**} with a ℓ in a close range, for which further analysis is tractable.
Formally,

$$m_* := \min\{m \in \{0, 1, \dots, M\} \ :\ \ell^{**} \le g_m\}.$$

By this definition $g_{m_*-1} < \ell^{**}$, hence g_{m_*} differs from ℓ^{**} by a factor of 2 at most, since $g_{m_*} = 2 g_{m_*-1} < 2\ell^{**}$. Let ϵ_* and ϵ_k be signs such that $\ell(\tilde{\theta}_{m_*}, \theta) = \|\epsilon_{m^*} \tilde{\theta}_{m_*} - \theta\|$ and for all $\tilde{m} + 1 \le k \le m_*$ we have $\ell(\tilde{\theta}_k, \tilde{\theta}_{k-1}) = \|\epsilon_k \tilde{\theta}_k - \epsilon_{k-1} \tilde{\theta}_{k-1}\|$.

We decompose the difference using triangle inequality by stepping through grid values until we reach the first number of buckets g_{m_*}.

$$\ell(\tilde{\theta}, \theta) \le \|\tilde{\theta} - \epsilon_{\tilde{m}} \theta\| \le \|\epsilon_{\tilde{m}} \tilde{\theta}_{g_{\tilde{m}}} - \theta\| \le \sum_{k=\tilde{m}+1}^{m_*} \|\epsilon_k \tilde{\theta}_k - \epsilon_{k-1} \tilde{\theta}_{k-1}\| + \|\epsilon_{m^*} \tilde{\theta}_{m_*} - \theta\| \le \sum_{k=1}^{m_*} \omega_k + \ell(\tilde{\theta}_{m_*}, \theta).$$

On the one hand, when $\|\theta\|^2 \leq \frac{\sqrt{d\ell^{**}}}{n}$, we have that $w_k^2 \leq 4C^2\frac{\sqrt{dg_k}}{n}$ and the sum $\sum_{k=\tilde{m}+1}^{m_*}\omega_k$ can be upper bounded by a sum of a geometric progression, and so

$$\sum_{k=\tilde{m}+1}^{m_*}\omega_k \leq \sum_{k=1}^{m_*}\omega_k = 2C\frac{\sqrt[4]{dg_1}}{\sqrt{n}}\sum_{k=1}^{m_*}(\sqrt[4]{2})^k = 2C\frac{\sqrt[4]{dg_1}}{\sqrt{n}}\frac{(\sqrt[4]{2})^{m_*+1}-1}{\sqrt[4]{2}-1} \leq 2C\frac{\sqrt[4]{dg_{m_*+1}}}{\sqrt{n}}.$$

It comes out that

$$\left(\sum_{k=m_1}^{m_*}\omega_k\right)^2 \lesssim \left(\sqrt{\frac{\delta d}{n}}+\frac{d}{n}\right).$$

On the other hand, when $\frac{\sqrt{d\ell^{**}}}{n} \leq \|\theta\|^2 \leq 1$, we have, with probability greater than $1 - c_1\log(n/d)e^{-c_0 d}$, that

$$\|\tilde{\theta}(\tilde{\ell})\|^2 \lesssim \|\theta\|^2.$$

It comes out that $w_k^2 \lesssim 4C^2\frac{dg_k}{n^2\|\theta\|^2}$ and the sum $\sum_{k=\tilde{m}+1}^{m_*}\omega_k$ can be upper bounded by a sum of a geometric progression, and so

$$\left(\sum_{k=m_1}^{m_*}\omega_k\right)^2 \lesssim \frac{d\ell^{**}}{n^2\|\theta\|^2}.$$

It comes out that

$$\left(\sum_{k=m_1}^{m_*}\omega_k\right)^2 \lesssim \frac{d\ell^{**}}{n^2\|\theta\|^2} \wedge \left(\sqrt{\frac{\delta d}{n}}+\frac{d}{n}\right).$$

We have by Proposition 11, that, with probability greater than $1 - c_1 e^{-c_0 d}$,

$$\ell^2(\tilde{\theta}_{m_*}, \theta) \lesssim \left(\sqrt{\frac{\delta d}{n}}+\frac{d}{n}\right) \wedge \left(\frac{\ell^{**}d}{n\|\theta\|^2}+\frac{d}{n}\right).$$

Hence we get that as long as $\tilde{\ell} \leq \ell^{**}$, then with probability greater than $1 - c_1\log(n/d)e^{-c_0 d}$, we have

$$\ell^2(\tilde{\theta}, \theta) \lesssim \left(\sqrt{\frac{\delta d}{n}}+\frac{d}{n}\right) \wedge \left(\frac{\ell^{**}d}{n\|\theta\|^2}+\frac{d}{n}\right).$$

Bounding I_2: In this case $g_k \geq \ell^{**}$. Notice first that, when $\|\theta\|^2 \leq \frac{\sqrt{dg_k}}{n}$, we have that $\|\tilde{\theta}(\tilde{\ell})\|^2 \lesssim \frac{\sqrt{dg_k}}{n}$ and hence $w_k^2 \gtrsim \frac{\sqrt{dg_k}}{n}$ and when $\|\theta\|^2 \geq \frac{\sqrt{dg_k}}{n}$ then $\|\tilde{\theta}(\tilde{\ell})\|^2 \lesssim \|\theta\|^2$ and $w_k^2 \gtrsim \frac{dg_k}{n^2\|\theta\|^2}$. It comes out that as long as $g_k \geq \ell^{**}$ we have that

$$w_k^2 \gtrsim \frac{\sqrt{dg_k}}{n} \wedge \frac{dg_k}{n^2\|\theta\|^2}.$$

In this case

$$I_2 \leq \mathbf{P}(\tilde{\ell}_2 > \ell^{**}) = \mathbf{P}(\tilde{m} > \ell^{**})$$

$$
\begin{aligned}
&\leq \sum_{m:\ g_m>\ell^{**}} \mathbf{P}(\tilde{\ell}_2 = g_m)\\
&\leq \sum_{m:\ g_m>\ell^{**}} \mathbf{P}(\ell(\tilde{\theta}_m, \tilde{\theta}_{m-1}) > \omega_m)\\
&\leq \sum_{m:\ g_m>\ell^{**}} \mathbf{P}(\ell(\tilde{\theta}_m, \theta) + \ell(\tilde{\theta}_{m-1}, \theta) > \omega_m)\\
&\leq 2 \sum_{m:\ g_m\geq\ell^{**}} \mathbf{P}\left(\ell(\tilde{\theta}_m, \theta) > \frac{\omega_m}{2}\right)\\
&\overset{(*)}{\lesssim} 2|G|e^{-c_0 d}\\
&\lesssim \log\left(\frac{n}{d}\right) e^{-c_0 d}.
\end{aligned}
$$

where $(*)$ holds, because for every $g_m \geq \ell^{**}$ due to Lemma 12 we have

$$\ell^2(\tilde{\theta}_{g_m}, \theta) \leq w_k^2 \qquad \text{w.p.} \quad 1 - c_1 \cdot e^{-c_0 d}.$$

Appendix B. Technical Lemmas

Lemma 16 *Assume that $k \geq 1$. Let $\delta, \delta' \in (0, 1/2)$, then we have that*

$$|g(\delta) - g(\delta')| \leq \frac{2k|\delta - \delta'|}{3}.$$

Proof Let us denote $\rho := 1 - 2\delta$. It is easy to observe that $\mathbf{E}(X_i) = \rho$ for all $i \geq 2$. Let us also define $X_1 = \eta_1$. For $0 < \delta$, we have

$$
\begin{aligned}
k^2 g(\delta) &= \mathbf{E}\left(\sum_{s=1}^{k}\prod_{i=1}^{s} X_i\right)^2\\
&= 2\sum_{1\leq s<s'\leq k} \mathbf{E}\left(\prod_{i=1}^{s} X_i^2 \prod_{i=s+1}^{s'} X_i\right) + \sum_{1\leq s\leq k} \mathbf{E}\left(\prod_{i=1}^{s} X_i^2\right)\\
&= k + 2\sum_{1\leq s<s'\leq k} \rho^{s'-s}.
\end{aligned}
$$

It comes out that $g(\delta)$ is non-increasing as a function of δ. Hence

$$|g(\delta) - g(\delta')| \leq \left|\lim_{\delta\to 0} g'(\delta)\right| |\delta - \delta'|.$$

Moreover we have that

$$k^2 g(\delta) = k + 2\sum_{s'=2}^{k} \frac{\rho - \rho^{s'}}{1-\rho} = \frac{k(1+\rho)}{1-\rho} - 2\frac{\rho - \rho^{k+1}}{(1-\rho)^2} = \frac{2k(1-\delta)\delta - 1 + 2\delta + (1-2\delta)^{k+1}}{2\delta^2}.$$

Using Taylor series approximation, we have that

$$(1-2\delta)^{k+1} = 1 - 2(k+1)\delta + 2k(k+1)\delta^2 - 4k(k+1)(k-1)\delta^3/3 + o(\delta^3).$$

Hence

$$g(\delta) = 1 - \frac{2(k^2-1)}{3k}\delta + o(\delta).$$

It comes out that

$$\left|\lim_{\delta\to 0} g'(\delta)\right| \leq 2k/3.$$

This concludes the proof. ■

Lemma 17 *Assume that $d \leq \ell \leq n$, $n\delta \leq \ell$ and ξ a sub-Gaussian isotropic vector. Then*

$$\mathbb{P}\left(|\|\bar{\eta}\|^2 - \mathbf{E}(\|\bar{\eta}\|^2)| \geq \sqrt{d\ell}\right) \geq 1 - 2e^{-cd},$$

$$\mathbb{P}\left(\frac{\ell}{4} \leq \|\bar{\eta}\|^2 \leq \ell\right) \geq 1 - e^{-c\ell},$$

$$\mathbb{P}\left(\left\|\frac{ww^T}{\ell} - I_d\right\|_{op} \leq C\frac{d}{\ell} + C\sqrt{\frac{d}{\ell}}\right) \geq 1 - e^{-cd},$$

and

$$\mathbb{P}\left(\|\xi\| \leq C\sqrt{d}\right) \geq 1 - e^{-cd},$$

for absolute constants $c > 0$ and $C > 0$.

Proof First, since $0 \leq \bar{\eta}_i^2 \leq 1$ and $\bar{\eta}_i^2$ are i.i.d., Hoeffding's inequality can be applied for $-\bar{\eta}_i^2$-s, namely for every $t > 0$

$$\mathbb{P}\left(|\|\bar{\eta}\|^2 - \mathbb{E}\|\bar{\eta}\|^2| \geq t\right) \leq 2e^{\frac{-t^2}{2\ell}}. \tag{42}$$

Hence we get the first inequality for $t = \sqrt{d\ell}$. Moreover

$$\mathbb{P}\left(\|\bar{\eta}\|^2 > \mathbb{E}\|\bar{\eta}\|^2 - t\right) \geq 1 - e^{\frac{-t^2}{2\ell}} \tag{43}$$

Recalling Lemma 1, we have that $\mathbb{E}[\bar{\eta}_i^2] \geq 1 - 2\delta k/3$ and for $\delta k \leq 1$ (or $\ell \geq \delta n$) we get $\mathbb{E}[\bar{\eta}_i^2] \geq \frac{1}{3}$ and $\mathbb{E}[\|\bar{\eta}\|^2] \geq \frac{\ell}{3}$. Consequently, taking $t = \frac{\ell}{12}$ we obtain that with probability greater than $1 - e^{\frac{-\ell}{288}}$

$$\|\bar{\eta}\|^2 \geq \frac{\ell}{4}$$

On the other hand (with probability 1)

$$\|\bar{\eta}\|^2 = \sum_{i=1}^{\ell} \bar{\eta}_i^2 \leq \ell$$

which summarizes the first concentration.

Second, by sub-Gaussian covariance estimation from (Vershynin, 2010), for every $t > 0$ with probability greater than $1 - \exp(-t)$ it holds that

$$\left\| \frac{ww^T}{\ell} - I_d \right\|_{op} \leq C \left(\frac{t}{\ell} + \frac{d}{\ell} + \sqrt{\frac{t}{\ell}} + \sqrt{\frac{d}{\ell}} \right)$$

By taking $t = d$ we obtain the desired result.

Third, using Hanson-Wright inequality from (Rudelson and Vershynin, 2013) for sub-Gaussian vectors we get the following tail bound for every $t > 0$

$$\mathbb{P}\left(\|\xi\|^2 \geq d + C(t + \sqrt{dt}) \right) \leq e^{-t}$$

Then taking $t = d$ we conclude our third concentration. ■

Proof of Lemma 12

Let's consider two scenarios for $\|\theta\|^2$.
For a small signal strength $\|\theta\|^2 \leq \frac{\sqrt{d\ell}}{n}$, we undertake analogous bounding procedure as in Proposition 3.

$$\begin{aligned}
\ell^2(\tilde{\theta}, \theta) &\lesssim \|\hat{\theta}\|^2 + \|\theta\|^2 \\
&\lesssim \left| \|\hat{\theta}\|^2 - \|\theta\|^2 \right| + \|\theta\|^2 \\
&\lesssim \|\hat{\Sigma} - \Sigma\|_{op} + \|\theta\|^2 \\
&\lesssim \sqrt{\frac{d}{n}} \|\theta\| + \left(\frac{d}{n} + \frac{\sqrt{d\ell}}{n} \right) + \left(1 + \sqrt{\frac{d}{\ell}} \right) \|\theta\|^2 \\
&\lesssim \sqrt{\frac{d}{n}} \frac{\sqrt[4]{d\ell}}{\sqrt{n}} + \frac{\sqrt{d\ell}}{n} + \left(1 + \sqrt{\frac{d}{\ell}} \right) \frac{\sqrt{d\ell}}{n} \\
&\lesssim \frac{\sqrt{d\ell}}{n} + \frac{\sqrt{d\ell}}{n} + \frac{\sqrt{d\ell}}{n} \\
&\lesssim \frac{\sqrt{d\ell}}{n} \wedge \left(\frac{d\ell}{\|\theta\|^2 n^2} + \frac{d}{n} \right),
\end{aligned}$$

where, in the last inequality, we use the fact that $\frac{d}{n} \leq \frac{\sqrt{d\ell}}{n} \leq \frac{d\ell}{\|\theta\|^2 n^2}$ in this regime .

For further analysis let's remind ourselves that $\ell \geq \ell^{**}$ and $\|\theta\|^2 \geq \frac{\sqrt{d\ell}}{n}$. Hence $\frac{\sqrt{d\ell}}{n} \geq \frac{d\ell}{n\|\theta\|^2} + \frac{d}{n}$ because $d \leq \ell$. According to Theorem 10 we know that

$$\ell^2(\tilde{\theta}, \theta) \lesssim \left(\frac{d}{\ell} + \frac{n^2\delta^2}{\ell^2} \right) \|\theta\|^2 + \frac{d}{n} + \frac{1}{\|\theta\|^2} \frac{d\ell}{n^2}.$$

It remains to show that

$$\left(\frac{d}{\ell} + \frac{n^2\delta^2}{\ell^2} \right) \|\theta\|^2 \leq \frac{d}{n} + \frac{1}{\|\theta\|^2} \frac{d\ell}{n^2},$$

in the case where $\|\theta\|^2 \leq 1$, since when $\|\theta\|^2 \geq 1$ we have that $\ell^{**} = n$ and

$$\ell^2(\tilde{\theta}, \theta) \lesssim \frac{d}{n}.$$

To prove the latter let's observe that , using (38), for $\ell \geq \ell^{**}$ we have

$$\ell \geq n\delta \vee d, \qquad \ell \geq \frac{\delta^{2/3} n^{4/3} \|\theta\|^{4/3}}{d^{1/3}}, \qquad \ell \geq n\|\theta\|^2. \tag{44}$$

It comes out that

$$\frac{n^2\delta^2}{\ell^2}\|\theta\|^2 \leq \frac{1}{\|\theta\|^2}\frac{d\ell}{n^2},$$

and that

$$\frac{d}{\ell}\|\theta\|^2 \leq \frac{1}{\|\theta\|^2}\frac{d\ell}{n^2}.$$

Thus

$$\ell^2(\tilde{\theta}, \theta) \lesssim \frac{d\ell}{n^2\|\theta\|^2}.$$

In summary we always have that

$$\ell^2(\tilde{\theta}, \theta) \lesssim \frac{\sqrt{d\ell}}{n} \wedge \left(\frac{d\ell}{n^2\|\theta\|^2} + \frac{d}{n}\right).$$