# On the Role of Channel Capacity in Learning Gaussian Mixture Models

**Elad Romanov** ELAD.ROMANOV@GMAIL.COM
*The Hebrew University of Jerusalem*

**Tamir Bendory** BENDORY@TAUEX.TAU.AC.IL
*Tel Aviv University*

**Or Ordentlich** OR.ORDENTLICH@MAIL.HUJI.AC.IL
*The Hebrew University of Jerusalem*

**Editors:** Po-Ling Loh and Maxim Raginsky

## Abstract

This paper studies the sample complexity of learning the $k$ unknown centers of a balanced Gaussian mixture model (GMM) in $\mathbb{R}^d$ with spherical covariance matrix $\sigma^2\boldsymbol{I}$. In particular, we are interested in the following question: what is the maximal noise level $\sigma^2$, for which the sample complexity is essentially the same as when estimating the centers from labeled measurements? To that end, we restrict attention to a Bayesian formulation of the problem, where the centers are uniformly distributed on the sphere $\sqrt{d}\mathcal{S}^{d-1}$. Our main results characterize the *exact noise threshold* $\sigma^2$ below which the GMM learning problem, in the large system limit $d, k \to \infty$, is as easy as learning from labeled observations, and above which it is substantially harder. The threshold occurs at $\frac{\log k}{d} = \frac{1}{2}\log\left(1+\frac{1}{\sigma^2}\right)$, which is the capacity of the additive white Gaussian noise (AWGN) channel. Thinking of the set of $k$ centers as a code, this noise threshold can be interpreted as the largest noise level for which the error probability of the code over the AWGN channel is small. Previous works on the GMM learning problem have identified the *minimum distance* between the centers as a key parameter in determining the statistical difficulty of learning the corresponding GMM. While our results are only proved for GMMs whose centers are uniformly distributed over the sphere, they hint that perhaps it is the decoding error probability associated with the center constellation as a channel code that determines the statistical difficulty of learning the corresponding GMM, rather than just the minimum distance.

## 1. Introduction

Gaussian mixture models (GMMs) are widely used in statistics and machine learning. Here, we consider the simplest case of a *spherical*, *balanced* $d$-dimensional GMM with $k$-components. Specifically, for centers $\boldsymbol{\mathcal{X}}_k = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k) \in \mathbb{R}^{d\times k}$ and variance $\sigma^2$, the corresponding GMM, denoted by $\mathrm{GMM}_{d,k}(\boldsymbol{\mathcal{X}}_k, \sigma^2)$, is described by the probability distribution $\boldsymbol{Y} \sim \mathrm{GMM}_{d,k}(\boldsymbol{\mathcal{X}}_k, \sigma^2)$:

$$\boldsymbol{Y} = \boldsymbol{X}_\ell + \sigma\boldsymbol{Z}, \quad \ell \sim \mathrm{Unif}([k]),\ \boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \tag{1}$$

where $[k] = \{1, \ldots, k\}$, and $\ell \in [k]$ will sometimes be referred to as the *label* of $\boldsymbol{Y}$ and is statistically independent of $\boldsymbol{Z}$. Our focus is on the classical GMM learning problem, where one observes $n$ independent samples $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \sim \mathrm{GMM}_{d,k}(\boldsymbol{\mathcal{X}}_k, \sigma^2)$, and wishes to recover the unknown centers $\boldsymbol{\mathcal{X}}_k$ (throughout, we always assume that the number of centers $k$ and the variance $\sigma^2$ are known).

This paper is devoted to studying the fundamental information-theoretic limits of the GMM learning problem, namely, the *sample complexity*: what is the smallest number of samples $n$ one

needs to collect in order to recover the centers (to within some prescribed precision)? The main difficulty in learning the GMM centers is that the samples are unlabeled, and the sample complexity is clearly lower bounded by that of the "genie-aided" setup where each sample is labeled. For sufficiently small noise levels the measurements can be accurately clustered, and the problem is as easy as in the "genie-aided" case, while for large enough noise levels reliable clustering is impossible. The main question we seek to answer here is: *what is the critical noise level below which the problem is as statistically easy as in the labeled case, and above which it is significantly harder?*

Past works have shown that the *separation* between the centers $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ has a decisive effect on the statistical difficulty of the problem. Let $\Delta(\boldsymbol{\mathcal{X}}_k) = \min_{1 \le i < j \le k} \|\boldsymbol{X}_i - \boldsymbol{X}_j\|$ be the minimal separation between any two centers. The seminal paper Regev and Vijayaraghavan (2017) has accurately identified the *scaling* of $\Delta(\boldsymbol{\mathcal{X}}_k)$, in the large system limit $k, d \to \infty$, under which one can estimate the centers (say, to within a small constant precision) using only $n = \mathrm{poly}(k, d)$ many samples. They show:[1] 1) *Upper bound:* If $\Delta = \Omega(\sigma\sqrt{\log k})$ then the centers may be estimated with $n = \mathrm{poly}(d, k)$ samples; 2) *Lower bound:* For any $\gamma(k) = o\left(\sigma\sqrt{\log k}\right)$, the class of GMMs with minimum separation $\Delta \ge \sigma\gamma(k)$ is not learnable (in a minimax sense) from $n = \mathrm{poly}(k, d)$ samples. The upper bound was recently improved by Kwon and Caramanis (2020), who showed that when $\Delta = \Omega(\sigma\sqrt{\log k})$, in fact $n = O(\sigma^2 k \cdot \mathrm{polylog}(k))$ samples suffice; this *almost* matches (up to $\mathrm{polylog}(k)$ factors) the sample complexity for the labeled case. Stated differently, the results above identify the critical noise level scaling for the minimax estimation problem as $\sigma^2 \sim \frac{\Delta^2(\boldsymbol{\mathcal{X}}_k)}{\log k}$.

The goal of this paper is to develop a finer grained understanding of the *exact* critical noise level $\sigma$, rather than only its scaling. To tackle this ambitious question, we make two modifications with respect to the setup studied in Regev and Vijayaraghavan (2017) and Kwon and Caramanis (2020): 1) Rather than studying the minimax setting with respect to all sets of centers $\boldsymbol{\mathcal{X}}_k$ with a given $\Delta(\boldsymbol{\mathcal{X}}_k)$, we take a Bayesian approach and assume $\boldsymbol{\mathcal{X}}_k \sim (\mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1}))^{\otimes k}$; 2) We consider a "more forgiving" loss function, which measures the average error in the center reconstruction rather than the maximal error. The rationale behind these modifications will be clarified in the sequel.

Under this setup, we show that the the critical noise level is precisely characterized by the equation $\frac{1}{2}\log\left(1 + \frac{1}{\sigma^2}\right) = \frac{\log k}{d}$, which is, by no accident, the noise level below which a "typical" constellation $\boldsymbol{\mathcal{X}}_k$ constitutes a good error correcting code for the $\mathsf{AWGN}(\sigma^2)$ channel (additive white Gaussian noise, with noise variance $\sigma^2$). Our analysis relies *explicitly* on the decodability properties of $\boldsymbol{\mathcal{X}}_k$, when thought of as a channel code. This is a "global" property of the constellation, compared to the minimum distance (note that it is well-known that at high coding rate, the minimum distance of a code is not entirely predictive of its error probability, see e.g. Barg and Forney (2002)). Regarding the minimum separation, we remark that, as is to be expected, our results are consistent with Regev and Vijayaraghavan (2017) regarding the required scaling of $\Delta(\boldsymbol{\mathcal{X}}_k)$ for statistically-efficient learning. Classical results on sphere packing, e.g., Kabatiansky and Levenshtein (1978), imply that if $\log k/d$ is finite, "typical" constellations under $\boldsymbol{\mathcal{X}}_k \sim (\mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1}))^{\otimes k}$ have minimal separation $\Delta(\boldsymbol{\mathcal{X}}_k) = \Theta(\sqrt{d})$. Thus, 1) When $\log k/d = \Theta(1)$ the critical noise level is at $\sigma^2 = \Theta(1)$, so in terms of minimal separation, $\Delta(\boldsymbol{\mathcal{X}}_k)/\sigma = \Theta(\sqrt{d}) = \Theta(\sqrt{\log k})$; 2) On the other hand, when $\log k/d = o(1)$, the critical noise level is $\sigma^2 = \Theta(d/\log k)$ and so $\Delta(\boldsymbol{\mathcal{X}}_k)/\sigma = \Theta(\sqrt{\log k})$.

Finally, our results *hint* at the possibility of a deeper connection between channel coding and statistical inference: the decodability properties of the set of centers $\boldsymbol{\mathcal{X}}_k$ (as a channel code) may

1. We restrict our attention in this discussion, and throughout the paper, exclusively to an asymptotic regime where $d, k \to \infty$ together with $\limsup_{d,k\to\infty} \frac{\log k}{d} < \infty$.

determine, to an extent, the statistical difficulty of learning the corresponding GMM. The present paper takes a modest first step towards showing such a connection, establishing it for the special case of spherical random codes, whose typical instances posses strong symmetry properties.

### 1.1. Formal Problem Formulation

As mentioned before, we study the large system behavior of the sample complexity under a uniform spherical prior on the centers. Denote the (random) centers by

$$\boldsymbol{\mathcal{X}}_k = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k) \sim \Big(\mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1})\Big)^{\otimes k} . \tag{2}$$

Note that we scale the problem so that $\|\boldsymbol{X}_i\| = \sqrt{d}$ for all $i \in [k]$. We observe $n$ measurements, $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$, sampled from the GMM distribution whose centers are $\boldsymbol{\mathcal{X}}_k$:

$$\Big[\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \,\Big|\, \boldsymbol{\mathcal{X}}_k\Big] \overset{i.i.d.}{\sim} \mathrm{GMM}_{d,k}(\boldsymbol{\mathcal{X}}_k, \sigma^2) , \tag{3}$$

see also (1). Per standard terminology in signal processing, $1/\sigma^2$ may be interpreted as the "signal-to-noise ratio" (SNR) per coordinate. Suppose that $\hat{\boldsymbol{\mathcal{X}}}_k = (\hat{\boldsymbol{X}}_1, \ldots, \hat{\boldsymbol{X}}_k)$ is an estimator of $\boldsymbol{\mathcal{X}}_k$, computed from the measurements. The model admits the following Markov chain structure:

$$\boldsymbol{\mathcal{X}}_k = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k) \longrightarrow (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n) \longrightarrow \hat{\boldsymbol{\mathcal{X}}}_k = (\hat{\boldsymbol{X}}_1, \ldots, \hat{\boldsymbol{X}}_k) . \tag{4}$$

At this point it is instructive to think about the much simpler estimation problem, where each measurement $\boldsymbol{Y}_i$ is observed with its label $\ell_i \in [k]$, and every center is observed *exactly* $n/k$ times. For this problem, the optimal mean squared error (MSE) in the reconstruction of each center is $d^{-1}\mathbb{E}\|\boldsymbol{X}_i - \hat{\boldsymbol{X}}_i\|^2 = k\sigma^2/n$ (to leading order in $k/n$), and is attained for example, by the sample mean. In the GMM estimation problem the samples are not labeled, and furthermore, the number of times each center appears in the measurements is a $\mathrm{Binomial}(n, 1/k)$ random variable. While the mean of this random variable is indeed $n/k$, some centers will appear fewer times. In particular, when $n = o(k \log k)$ some of the centers are likely to not appear even once (coupon collecting). To circumvent the issues arising due to this effect, and focus our study on the problem of dealing with the lack of labels, we measure the discrepancy between $\boldsymbol{\mathcal{X}}_k$ and $\hat{\boldsymbol{\mathcal{X}}}_k$, by the loss function

$$\mathcal{L}_{\mathrm{avg}}(\boldsymbol{\mathcal{X}}_k, \hat{\boldsymbol{\mathcal{X}}}_k) = \frac{1}{k}\sum_{i=1}^{k} d^{-1}\mathrm{dist}^2(\boldsymbol{X}_i, \hat{\boldsymbol{\mathcal{X}}}_k) := \frac{1}{k}\sum_{i=1}^{k} \min_{1\le j\le k} d^{-1}\|\boldsymbol{X}_i - \hat{\boldsymbol{X}}_j\|^2 . \tag{5}$$

In words: the average normalized squared distance between a center $\boldsymbol{X}_i$ and the list $\hat{\boldsymbol{\mathcal{X}}}_k$. As we shall see, under this loss function it is possible to obtain a risk of $k\sigma^2/n$ for $\sigma$ below the critical noise level and $n/k$ large enough. In contrast, the more restrictive max-loss function $\mathcal{L}_{\max}(\boldsymbol{\mathcal{X}}_k, \hat{\boldsymbol{\mathcal{X}}}_k) = \max_{1\le i\le k} d^{-1}\mathrm{dist}^2(\boldsymbol{X}_i, \hat{\boldsymbol{\mathcal{X}}}_k)$ considered in much of the prior work, does not decay with $n$ in the regime $n = o(k \log k)$, regardless of the noise level, due to the non-uniform empirical distribution of the center indices. Under $\mathcal{L}_{\mathrm{avg}}$, on the other hand, to achieve $\varepsilon$ error, it suffices to estimate only a fraction $1 - O(\varepsilon)$ of the centers within error $O(\varepsilon)$, having the remaining centers incur an error $O(1)$. Thus, the effect of non-uniform label empirical distribution is bypassed by this loss function.

Under our formulation of the GMM learning problem, the goal is to construct an estimation rule $\hat{\boldsymbol{\mathcal{X}}}_k : (\mathbb{R}^d)^n \to \mathbb{R}^{d\times k}$ ("algorithm") so to minimize the risk: $\mathbb{E}\mathcal{L}_{\mathrm{avg}}(\boldsymbol{\mathcal{X}}_k, \hat{\boldsymbol{\mathcal{X}}}_k(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n))$.

Importantly, the expectation is taken over the randomness in *both* the sample generating process given the centers (3), *as well as* the center prior distribution (2), whose joint distribution adheres to the Markov chain structure in (4). We study the *information-theoretic limits* of the aforementioned problem. Consider the minimum attainable risk over all estimation laws $\hat{\boldsymbol{\mathcal{X}}}_k$:

$$R_n = \inf_{\hat{\boldsymbol{\mathcal{X}}}_k} \mathbb{E}\mathcal{L}_{\text{avg}}(\boldsymbol{\mathcal{X}}_k, \hat{\boldsymbol{\mathcal{X}}}_k(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)) \,. \tag{6}$$

For a fixed precision level $\varepsilon > 0$, define the *sample complexity*,

$$n^*_\varepsilon = n^*_\varepsilon(d, k, \sigma^2) = \min\left\{n \,:\, R_n \le \varepsilon\right\} \,. \tag{7}$$

Importantly, (6) and (7) make no assumptions about the computational difficulty of implementing $\hat{\boldsymbol{\mathcal{X}}}_k : (\mathbb{R}^d)^n \to \mathbb{R}^{d\times k}$, and in particular are not restricted to computational efficient algorithms ($\text{poly}(d, k)$ runtime). Throughout, computational considerations shall be completely neglected.

### 1.2. Main Results

As our analysis relies on viewing the centers as a code for the AWGN channel, the problem's *rate*

$$\mathsf{R}_{d,k} := \frac{\log k}{d} \,, \tag{8}$$

and the decreasing function $\mathsf{C} : (0,\infty) \to (0,\infty)$

$$\mathsf{C}(\sigma^2) = \frac{1}{2}\log\left(1 + \frac{1}{\sigma^2}\right) \,, \tag{9}$$

characterizing the $\mathsf{AWGN}(\sigma^2)$ channel capacity, will play a key role. Throughout the paper, we couple the noise level $\sigma^2$ to $k$ and $d$ by the parameter $\beta \in (0,\infty)$ via the equation

$$\mathsf{R}_{d,k} = \mathsf{C}(\beta\sigma^2) \,. \tag{10}$$

When $\beta > 1$, the rate is smaller than the capacity; when $\beta < 1$, it is larger. This parametrization will turn out particularly useful in the statement of the results and their derivations.

We restrict attention to the large-system limit, where $d, k \to \infty$, and denote the *limiting rate* by

$$\mathsf{R} = \lim_{d\to\infty} \mathsf{R}_{d,k} \in [0,\infty) \,. \tag{11}$$

We distinguish between two asymptotic regimes:

- **(Positive Rate,** $\mathsf{R} > 0$**)**: $\sigma^2 \in (0,\infty)$ is a fixed constant. In particular, $k = e^{\Theta(d)}$.
- **(Zero Rate,** $\mathsf{R} = 0$**)**: $\sigma^2 \to \infty$. So that also $k \to \infty$, we also require impose $\sigma^2 = o(d)$.

We remark that, since we are interested in estimation to *finite precision* $\varepsilon$ in Theorems 1 and 2 below, the asymptotic regime $\log k = \omega(d)$, namely when the number of centers $k$ is super-exponential in $d$, becomes rather uninteresting. Indeed, for super-exponential $k$ one may simply take $\hat{\boldsymbol{\mathcal{X}}}_k$ to be some fixed $\sqrt{\varepsilon d}$-net of the sphere $\sqrt{d}\mathcal{S}^{d-1}$, which can be of size $(O(1/\varepsilon))^{d/2} \ll k$. Clearly, $\mathcal{L}_{\text{avg}}(\boldsymbol{\mathcal{X}}_k, \hat{\boldsymbol{\mathcal{X}}}_k) \le \varepsilon$ for any $\boldsymbol{\mathcal{X}}_k$, so under such asymptotics $n^*_\varepsilon = 0$ exactly.

Our first main result states that when the rate is below the channel capacity, $\boldsymbol{\mathcal{X}}_k$ is learnable at essentially the same sample complexity as in the labeled case.

**Theorem 1** *Suppose that $\beta > 1$. Then*

$$e^{-2\mathsf{R}} \leq \lim_{\varepsilon \to 0} \lim_{d \to \infty} \frac{n^*_\varepsilon}{(\sigma^2 k/\varepsilon)} \leq 1\,. \tag{12}$$

Theorem 1 implies that when the rate is below the channel capacity, for every *fixed* and *small* precision $\varepsilon > 0$, and for $d$ large, the sample complexity scales like $n = C\sigma^2 k/\varepsilon$, where $C \in [e^{-2\mathsf{R}}, 1]$. Remarkably, when $k$ is sub-exponential in $d$ ($\mathsf{R} = 0$) the pre-factor $C$ is precisely 1. Thus, the sample complexity of the GMM learning problem is *exactly* the same as that of the labeled case, up to lower order terms in $1/\varepsilon$, and asymptotically ($d \to \infty$) vanishing correction terms.

Our second main result states that above the capacity, the sample complexity is super-linear:

**Theorem 2** *Suppose that $\beta < 1$. Then for any fixed sufficiently small $\varepsilon < \varepsilon_0(\mathsf{R})$,*

$$\lim_{d \to \infty} \frac{n^*_\varepsilon}{\sigma^2 k/\varepsilon} = \infty\,. \tag{13}$$

*Moreover, the following quantitative bounds hold for all sufficiently small $\varepsilon < \varepsilon_0(\mathsf{R})$:*

1. *If $\mathsf{R} > 0$ then*

$$\frac{n^*}{\sigma^2 k} = \Omega_{\varepsilon,\beta,\mathsf{R}}\left(\sqrt{\frac{\log k}{\log\log k}}\right)\,. \tag{14}$$

2. *If $\mathsf{R} = 0$ then*

$$\frac{n^*}{\sigma^2 k} = \Omega_{\varepsilon,\beta}\left(\min\left\{\sqrt{\frac{\log k}{\log\log k}}, \sqrt{\frac{d}{\log k}}\right\}\right)\,. \tag{15}$$

Theorems 1 and 2 together reveal a dichotomy: *precisely* at the channel capacity ($\beta = 1$), the large-system behavior of the sample complexity undergoes a *phase-transition*, from a linear growth in $\sigma^2 k$, as in the labeled case, to super-linear growth.

### 1.3. Prior Art

The problem of estimating the parameters of a Gaussian mixture model has a long and rich history, going back to the pioneering work of Pearson (1894). We briefly mention some pointers to the literature, though we emphasize that the list below is *not exhaustive by any means*.

The first work to highlight the importance of minimum separation in learning GMMs is Dasgupta (1999), who gave a poly-time algorithm assuming (in the spherical balanced case) $\Delta = \Omega(\sigma\sqrt{d})$. Subsequent works have gradually improved upon the required bound on $\Delta$. Early incarnations include Arora and Kannan (2001); Vempala and Wang (2004); Achlioptas and McSherry (2005); Dasgupta and Schulman (2007); Kannan et al. (2008), which culminated in a bound $\Delta = \Omega(\sigma k^{1/4})$ as sufficient for estimation in polynomial time. This barrier was broken only fairly recently Diakonikolas et al. (2018); Hopkins and Li (2018); Kothari et al. (2018),who have shown that separation $\Delta = \Omega(\sigma k^{\gamma})$ suffices for polynomial-time learnability, for *any* constant $\gamma > 0$.

As for statistical lower bounds, it is known that in the absence of a separation condition, $n = \exp(k)$ samples are generally necessary to learn the parameters of a GMM Moitra and Valiant (2010); Hardt and Price (2015). The work Regev and Vijayaraghavan (2017) has shown that

separation $\Delta = \Omega(\sigma\sqrt{\log k})$ is a *sufficient and necessary* condition for learning GMMs with $n = \mathrm{poly}(k, d)$ samples; the algorithm they proposed to prove their upper bound has exponential runtime. Kwon and Caramanis (2020) have recently improved their upper bound on the sample complexity, and have show that in fact $n = O(\sigma^2 k \cdot \mathrm{polylog}(k))$ samples suffice, which *almost* matches the trivial lower bound of $n = \Omega(\sigma^2 k)$. Their analysis consists of two components : 1) An exponential-time initialization scheme, that finds points *sufficiently close* to the true centers, based on the results of Ashtiani et al. (2018); 2) New local convergence and finite-sample guarantees for (a slightly modified version of) the well-known Expectation Maximization (EM) algorithm. To our knowledge, the problem of learning $\Delta = \Omega(\sigma\sqrt{\log k})$-separated GMMs in polynomial time, or proving that this cannot be done (the existence of a computational-statistical gap) is still open.

Another line of work circumvents the minimal separation requirement, by instead restricting attention to "typical" problem instances, an approach much in line with the results of the present paper, and in the context of learning GMMs dates, to the best of our knowledge, to the study Srebro et al. (2006). In the papers Hsu and Kakade (2013); Bhaskara et al. (2014); Goyal et al. (2014); Anderson et al. (2014); Anandkumar et al. (2014); Ge et al. (2015), it is shown that when the center configuration satisfies certain algebraic *non-degeneracy* conditions, methods based on tensor decomposition may be used to recover the centers; such non-degenerate configurations are highly abundant when $d$ is large relative to $k$, specifically $k \leq d^{O(1)}$.

Lastly, a different line of work considers learning GMMs by means of density estimation, that is, given samples $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ one has to construct a density $f$ which is close to $\mathrm{GMM}_{d,k}(\boldsymbol{\mathcal{X}}_k, \sigma^2)$ in, e.g., total variation distance. This problem may be considered in either in the setting of proper learning ($f$ has to be a $k$-component GMM) or improper learning (no such restriction), see for example Feldman et al. (2006); Kalai et al. (2010); Chan et al. (2014); Suresh et al. (2014); Li and Schmidt (2017); Diakonikolas et al. (2019); Ashtiani et al. (2018). For well-seaprated spherical GMMs, $\Delta = \Omega(\sigma\sqrt{\log k})$, guarantees for proper distribution estimation may be translated to error bounds on the centers, see Regev and Vijayaraghavan (2017); Kwon and Caramanis (2020).

Our proof program closely follows that of Romanov et al. (2021), which studied the sample complexity of the multi-reference alignment (MRA) problem. MRA is a *particular* instance of a GMM, with exactly $k = d$ components corresponding to different shifted versions of the same signal. While, similarly to Romanov et al. (2021), the proof of our lower bound uses the mutual information method Polyanskiy and Wu (2014), here the mutual information is upper bounded using the I-MMSE relation rather than the Fano-based argument of Romanov et al. (2021). More importantly, the proof of the upper bound here requires overcoming several significant hurdles not present in the MRA model. In particular, while in MRA we always have $k = d$, in the GMM problem $k$ may be much greater, and even exponential in the dimension. Furthermore, in MRA there is a single signal to be estimated and all measurement are informative for its estimation. Here, on the other hand, many centers must be estimated, which significantly complicates the first step of our reconstruction algorithm with respect to that used in Romanov et al. (2021).

**Paper outline.** In Section 2 we provide brief background on channel coding and random spherical codes, which shall be used in the analysis to follow. In Section 3 we outline the proof of the lower bound in Theorems 1 and 2. In Section 4 we outline the proof of the upper bound in Theorem 1. To keep within the space constraint, most of the technical details are deferred to the Appendix.

## 2. Background on Channel Coding

A key message of this paper is the following: the centers $\boldsymbol{\mathcal{X}}_k$ are learnable at linear sample complexity exactly in the regime where the constellation $\boldsymbol{\mathcal{X}}_k = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k)$ defines (with high probability) a good codebook for the AWGN channel with noise variance $\sigma^2$. Throughout the analysis, the connection to the decoding capabilities of $\boldsymbol{\mathcal{X}}_k$ will be instrumental. In this section, we briefly survey the required background from information and coding theory. We refer the reader to Cover and Thomas (2012), Gallager (1968) and Polyanskiy and Wu (2014) for a comprehensive treatment.

A *coding scheme* for sending $\log k$ nats over the $d$-dimensional AWGN channel consists of a codebook and a decoder. The codebook is a set of $k$ codewords $\boldsymbol{\mathcal{C}} = (\boldsymbol{C}_1, \ldots, \boldsymbol{C}_k) \in \mathbb{R}^{d \times k}$, where codeword $\boldsymbol{C}_i$ encodes message $1 \le i \le k$, and all codewords satisfy $\|\boldsymbol{C}_i\|^2 \le d$. The code's rate is $\mathsf{R}_{d,k} = \frac{\log k}{d}$. The decoder $\mathsf{Dec} : \mathbb{R}^d \to [k]$ is a mapping from channel outputs to messages. It is often convenient to allow the decoder to output symbols in $[k] \cup \{\#\}$, where the special symbol $\#$ corresponds to a *declared decoding error*.

The *decoding error* associated with message $i \in [k]$, for a given a codebook-decoder pair, is

$$P_{e,i}(\sigma^2 | \boldsymbol{\mathcal{C}}, \mathsf{Dec}) = \Pr\left(i \neq \mathsf{Dec}(\boldsymbol{X}_i + \sigma \boldsymbol{Z})\right) , \tag{16}$$

and the *average error* over all messages is

$$P_{e,avg}(\sigma^2 | \boldsymbol{\mathcal{C}}, \mathsf{Dec}) := \frac{1}{k} \sum_{i=1}^{k} P_{e,i}(\sigma^2 | \boldsymbol{\mathcal{C}}, \mathsf{Dec}) = \Pr_{\ell \sim \mathrm{Unif}([k])} \left(\ell \neq \mathsf{Dec}(\boldsymbol{X}_\ell + \sigma \boldsymbol{Z})\right) . \tag{17}$$

For a given codebook $\boldsymbol{\mathcal{C}}$, the optimal decoder, in the sense of smallest average error, is clearly given by the maximum a posteriori probability (MAP) rule

$$\mathsf{DecOpt}(\boldsymbol{Y}) = \operatorname*{argmax}_{i \in [k]} \Pr(\ell = i \,|\, \boldsymbol{Y}, \boldsymbol{\mathcal{C}}) = \operatorname*{argmin}_{1 \le i \le k} \|\boldsymbol{Y} - \boldsymbol{C}_i\|^2 , \tag{18}$$

where ties are broken arbitrarily. Accordingly, we define the error of the *codebook* $\boldsymbol{\mathcal{C}}$, and the corresponding individual errors as

$$\rho_{\mathrm{avg}}(\sigma^2 | \boldsymbol{\mathcal{C}}) = P_{e,avg}(\sigma^2 | \boldsymbol{\mathcal{C}}, \mathsf{DecOpt}) , \quad \rho_i(\sigma^2 | \boldsymbol{\mathcal{C}}) = P_{e,i}(\sigma^2 | \boldsymbol{\mathcal{C}}, \mathsf{DecOpt}) . \tag{19}$$

In communication theory, one is interested in designing coding schemes with large rate and small error probability. We say a rate $\mathsf{R} \in (0, \infty)$ is *achievable* if there exists a sequence $(d \to \infty)$ of codebooks $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{d \times k}$ such that $\lim_{d \to \infty} \mathsf{R}_{d,k} = \mathsf{R}$ and $\lim_{d \to \infty} \rho(\sigma^2 | \boldsymbol{\mathcal{C}}) = 0$. Shannon's celebrated channel coding theorem gives a precise characterization of all the achievable rates:

**Theorem 3 (Channel coding theorem, AWGN channel)** *Fix $\sigma^2$, and let $\mathsf{C}(\cdot)$ be given in (9).*

1. *(Achievability). Any rate $\mathsf{R} < \mathsf{C}(\sigma^2)$ is achievable.*
2. *(Converse). No rate $\mathsf{R} > \mathsf{C}(\sigma^2)$ is achievable.*

The achievability part of the channel coding theorem is typically proved using a random coding argument with respect to the ensemble of i.i.d. Gaussian codebooks. However, it can also be proved using the ensemble of random spherical codebooks, $\boldsymbol{\mathcal{C}} = \boldsymbol{\mathcal{X}}_k = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k) \sim \mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1})^{\otimes k}$. In fact, the latter ensemble results in a favorable decay of the error probability with $d$, Shannon (1959). We denote the decoding error, averaged over the codebook ensemble, by

$$\rho_{\mathrm{avg}}(\sigma^2) = \mathbb{E}[\rho_{\mathrm{avg}}(\sigma^2 | \boldsymbol{\mathcal{X}}_k)] \overset{(\star)}{=} \mathbb{E}[\rho_i(\sigma^2 | \boldsymbol{\mathcal{X}}_k)] , \tag{20}$$

where $(\star)$ holds since each $\boldsymbol{X}_i$ has the same distribution.

**Proposition 4** *Let $\beta > 1$ be fixed. Suppose that $d, k \to \infty$, with $\mathsf{R}_{d,k} = \mathsf{C}(\beta\sigma^2)$, so that either: 1) $\sigma^2$ fixed;or 2) $\omega(1) = \sigma^2 = o(d)$. Then $\lim_{d\to\infty} \rho_{avg}(\sigma^2) = 0$.*

While Proposition 4 is well-known when $\sigma^2$ is fixed (positive rate) Shannon (1959), the case of $\omega(1) = \sigma^2 = o(d)$ has not been mainstreamed. We provide a self-contained proof of Proposition 4 in Appendix, Section A, since it will serve as the baseline for the derivations that follow.

## 3. Proof of Lower Bounds

Our proof of the lower bounds in Theorems 1 and 2 uses a standard framework for proving estimation lower bounds (e.g., (Polyanskiy and Wu, 2014, Chapter 28)).

Suppose $\hat{\boldsymbol{\mathcal{X}}}_k = \hat{\boldsymbol{\mathcal{X}}}_k(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ attains $\mathbb{E}\mathcal{L}_{\text{avg}}(\boldsymbol{\mathcal{X}}_k, \hat{\boldsymbol{\mathcal{X}}}_k) \le \varepsilon$. Consider the Markov chain (4). By the data processing inequality (DPI) (Polyanskiy and Wu, 2014, Theorem 2.5),

$$I(\boldsymbol{\mathcal{X}}_k; \hat{\boldsymbol{\mathcal{X}}}_k) \le I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)\,. \tag{21}$$

We lower bound the LHS of (21) in terms of $\varepsilon$ and upper bound the RHS in terms of $n$ and $\sigma^2$. Starting with $I(\boldsymbol{\mathcal{X}}_k; \hat{\boldsymbol{\mathcal{X}}}_k)$, clearly,

$$I(\boldsymbol{\mathcal{X}}_k; \hat{\boldsymbol{\mathcal{X}}}_k) \ge \min_{P_{\boldsymbol{\mathcal{D}}|\boldsymbol{\mathcal{X}}_k} : \mathbb{E}\mathcal{L}_{\text{avg}}(\boldsymbol{\mathcal{X}}_k, \boldsymbol{\mathcal{D}}) \le \varepsilon} I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{\mathcal{D}})\,, \tag{22}$$

where we minimize the mutual information (MI) over all conditional distributions of random variables $\boldsymbol{\mathcal{D}} = (\boldsymbol{D}_1, \ldots, \boldsymbol{D}_k) \in \mathbb{R}^{d\times k}$, under the expected loss constraint $\mathbb{E}\mathcal{L}_{\text{avg}}(\boldsymbol{\mathcal{X}}_k, \boldsymbol{\mathcal{D}}) \le \varepsilon$. The minimization (22) is an instance of a *rate-distortion* problem, typically encountered when studying the information-theoretic limits of lossy compression (Polyanskiy and Wu, 2014, Chapter 25).

One complication that arises when attempting to solve the optimization problem in (22) is that the distortion measure, $\mathcal{L}_{\text{avg}}(\boldsymbol{\mathcal{X}}_k, \boldsymbol{\mathcal{D}}) = \frac{1}{dk}\sum_{i=1}^k \min_{1\le j\le k} \|\boldsymbol{X}_i - \boldsymbol{D}_j\|^2$ is somewhat non-standard. If instead we had used the quadratic loss, $\frac{1}{dk}\|\boldsymbol{\mathcal{X}}_k - \boldsymbol{\mathcal{D}}\|_F^2 = \frac{1}{dk}\sum_{i=1}^k \|\boldsymbol{X}_i - \boldsymbol{D}_i\|^2$, the resulting optimization problem would essentially lend itself to the classical problem of computing the Gaussian quadratic rate-distortion function (RDF), which admits the solution $\frac{dk}{2}\log(1/\varepsilon)$.

The loss $\mathcal{L}_{\text{avg}}$ differs from the standard quadratic loss in that it allows for $k$ additional degrees of freedom: every $i \in [k]$ is matched to the best index $j_i = \operatorname{argmin}_{j\in[k]} \|\boldsymbol{X}_i - \boldsymbol{D}_{j_i}\|$. Since the entropy of the $k$-tuple $(j_1, \ldots, j_k)$ is at most $k\log k$ nats, the RDF for $\mathcal{L}_{\text{avg}}$ must be at most $k\log k$ nats away from the RDF for the standard quadratic loss. We prove in Appendix, Section B.1:

**Lemma 5** *Consider the Markov chain (4), with $\mathbb{E}\mathcal{L}_{avg}(\boldsymbol{\mathcal{X}}_k, \hat{\boldsymbol{\mathcal{X}}}_k) \le \varepsilon$. For universal $c_0 > 0$,*

$$I(\boldsymbol{\mathcal{X}}_k; \hat{\boldsymbol{\mathcal{X}}}_k) \ge \frac{dk}{2}\log(1/\varepsilon) - dk\log\left(1 + c_0(\varepsilon d)^{-1/2}\right) - k\log k\,.$$

Next, we upper bound $I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$, starting with a trivial bound. Let $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_n)$ be the random labels, such that $\boldsymbol{Y}_j = \boldsymbol{X}_{\ell_j} + \sigma\boldsymbol{Z}_j$. By the DPI, $I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n) \le I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n, \boldsymbol{\ell})$. Now, given $\boldsymbol{\ell}$, the mapping $\boldsymbol{\mathcal{X}}_k \mapsto (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ simply corresponds to $k$ parallel Gaussian channels, each used on average $n/k$ times. Thus, as we formally prove in Appendix, Section B.3,

**Lemma 6** *The following holds:*

$$I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n) \le I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n, \boldsymbol{\ell}) \le \frac{dk}{2}\log\left(1 + \frac{n}{k\sigma^2}\right)\,. \tag{23}$$

*Consequently, combining with (21) and Lemma 5,*

$$\lim_{d\to\infty} \frac{n^*_\varepsilon}{k\sigma^2} \geq e^{-2\mathsf{R}}\varepsilon^{-1} - 1\,.$$

The bound (23) misses a *crucial* aspect of our problem: the observations are *not* labeled. We next derive a bound which does capture this effect, though at the loss of the "correct" dependence on $n$.

Observe that $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are conditionally independent given $\boldsymbol{\mathcal{X}}_k$. That is: the "channel" mapping the set of centers to samples is *memoryless*. It is an elementary fact (Polyanskiy and Wu, 2014, Theorem 5.1) that in this case, the MI is subadditive

$$I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n) \leq \sum_{i=1}^{n} I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y}_i) = n \cdot I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y})\,. \tag{24}$$

While this bound fails to correctly capture the dependence of $I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ on $n$ when $n/(k\sigma^2)$ is large, it does suffice for establishing the phase transition of the sample complexity that we seek here. We proceed to bounding the single-sample MI, $I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y})$, a much more manageable object. Let $\ell \sim \mathrm{Unif}([k])$ be the random label of $\boldsymbol{Y}$. Using the MI chain rule both ways,

$$I(\boldsymbol{\mathcal{X}}_k, \ell; \boldsymbol{Y}) = I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y}) + I(\ell; \boldsymbol{Y}|\boldsymbol{\mathcal{X}}_k) = I(\ell; \boldsymbol{Y}) + I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y}|\ell)\,.$$

Now, $I(\ell, \boldsymbol{Y}) = 0$ (since $\{\boldsymbol{X}_i\}_{i=1}^k$ are identically distributed, so $\boldsymbol{Y}$ does not depend on $\ell$). Similarly, $I(\ell; \boldsymbol{Y}|\boldsymbol{\mathcal{X}}_k) = H(\ell|\boldsymbol{\mathcal{X}}_k) - H(\ell|\boldsymbol{\mathcal{X}}_k, \boldsymbol{Y})$, and $H(\ell|\boldsymbol{\mathcal{X}}_k) = H(\ell) = \log k$ Furthermore, $I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y}|\ell) = I(\boldsymbol{X}_\ell; \boldsymbol{Y}|\ell) \leq \mathsf{C}(\sigma^2)d$, as the AWGN channel capacity $\mathsf{C}(\sigma^2)$ upper bounds $I(\boldsymbol{X}; \boldsymbol{X} + \sigma\boldsymbol{Z})/d$ for any random variable on $\mathbb{R}^d$ with second moment $\mathbb{E}\|\boldsymbol{X}\|^2 \leq d$. Combining these equalities and estimates and rearranging, we obtain

$$I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y}) \leq \mathsf{C}(\sigma^2)d - \log k + H(\ell|\boldsymbol{\mathcal{X}}_k, \boldsymbol{Y})\,. \tag{25}$$

In light of (25), it remains to estimate $H(\ell|\boldsymbol{\mathcal{X}}_k, \boldsymbol{Y})$, to be interpreted as the remaining uncertainty in a message $\ell$ that is sent across the channel, given the output $\boldsymbol{Y}$ as well as the known codebook $\boldsymbol{\mathcal{X}}_k$. To that end, consider the non-increasing mapping $\beta \mapsto \varphi(\beta) = H(\ell|\boldsymbol{\mathcal{X}}_k, \boldsymbol{Y})$ (recall that larger $\beta$ corresponds to smaller $\sigma$). Since a typical realization of $\boldsymbol{\mathcal{X}}_k$ results in a code whose error vanishes when $\beta > 1$, Fano's inequality implies that $\varphi(\beta)\big|_{\beta>1} = o(\log(k))$. Thus, for $\beta < 1$, we have that $\varphi(\beta) = -\int_\beta^{1+\delta} \varphi'(s)ds + o(\log(k))$, for any $\delta > 0$. Using the I-MMSE formula Guo et al. (2005), a remarkable connection between information and estimation under Gaussian channels, the derivative $\varphi'(\beta)$ can be expressed as the minimum MSE (MMSE) in estimating $\boldsymbol{X}_\ell$ from $\boldsymbol{Y}$. Finally, we upper bound the MMSE by the optimal MSE for linear estimation, resulting in the following lemma, whose full proof appears in Appendix, Section B.4. We denote by $\mathsf{C}^{-1}$ the inverse of (9), and by $h_b(p) = p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p}$ the binary entropy function.

**Lemma 7** *Suppose that $\beta < 1$, so that $\mathsf{R}_{d,k} = \mathsf{C}(\beta\sigma^2) > \mathsf{C}(\sigma^2)$. For $\delta > 0$, denote the corresponding noise level by $\sigma_0^2(\delta) = \mathsf{C}^{-1}\left((1+\delta)\mathsf{R}_{d,k}\right)$ and let $e(\delta) = \rho_{avg}\left(\mathsf{C}^{-1}\left((1+\delta)\mathsf{R}_{d,k}\right)\right)$ be the ensemble average decoding error, (20), over the $\mathsf{AWGN}(\sigma_0^2)$ channel. We have that*

$$H(\ell|\boldsymbol{\mathcal{X}}_k, \boldsymbol{Y}) \leq \log k - \mathsf{C}(\sigma^2)d + h_b(e(\delta)) + (\delta + e(\delta))\log k\,,$$

*and consequently, using (25),*

$$I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y}) \leq h_b(e(\delta)) + (\delta + e(\delta))\log k\,. \tag{26}$$

As mentioned above, when $\beta < 1$, $e(\delta) = o(1)$ for *all* fixed $\delta > 0$, by Proposition 4; consequently, $I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y}) = o(\log k)$. Combining this with (21), Lemma 5 and (24), assuming sufficiently small $\varepsilon = O_{\mathsf{R}}(1)$, we establish (13):

$$\frac{n^*_\varepsilon}{\sigma^2 k} \geq C(\varepsilon, \mathsf{R}) \cdot \frac{d}{\sigma^2} \cdot (I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y}))^{-1} = \omega\left(\frac{1}{\sigma^2} \cdot \frac{d}{\log k}\right) = \omega\left(\frac{1}{\sigma^2 \mathsf{C}(\beta\sigma^2)}\right) = \omega(1)\,. \tag{27}$$

One can get *quantitative* bounds by carefully setting $\delta = o(1)$, as we do in Appendix, Section B.5:

**Lemma 8** *Suppose that $\beta < 1$. For small enough fixed $\varepsilon \leq \varepsilon_0(\mathsf{R})$:*

1. *(Positive rate). If $\mathsf{R} > 0$ then*

$$\frac{n^*_\varepsilon}{\sigma^2 k} \geq C(\varepsilon, \beta, \mathsf{R}) \sqrt{\frac{\log k}{\log\log k}}\,. \tag{28}$$

2. *(Zero rate). If $\mathsf{R} = 0$ then*

$$\frac{n^*_\varepsilon}{\sigma^2 k} \geq C(\varepsilon, \beta) \min\left\{\sqrt{\frac{\log k}{\log\log k}}, \sqrt{\frac{d}{\log k}}\right\}\,. \tag{29}$$

**Proof** (Of Theorem 2). Directly follows from Lemma 8. ∎

## 4. Proof of Upper Bound

In this section we prove the upper bound of Theorem 1, assuming the rate is smaller than the capacity ($\beta > 1$). The proof is constructive: we propose and analyze an algorithm (which runs in exponential time), whose output $\hat{\boldsymbol{\mathcal{X}}}_k$ satisfies $\mathbb{E}\mathcal{L}_{\text{avg}}(\boldsymbol{\mathcal{X}}_k, \hat{\boldsymbol{\mathcal{X}}}_k) \leq \varepsilon$. It consists of two steps, each using *different* measurements: Step I is allocated $N$ samples, while Step II uses the remaining $\bar{N} = n - N$ samples.

Step I consists of a *brute-force search* over an exponential-sized set of candidate centers. Let $\varepsilon_{\text{I}} > 0$ be a given precision level, and fix $\mathcal{T}$ a $\sqrt{\varepsilon_{\text{I}} d/2}$-net of the sphere $\sqrt{d}\mathcal{S}^{d-1}$. For *each* candidate $\hat{\boldsymbol{X}} \in \mathcal{T}$, we use the measurements $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$ allocated for this step to essentially solve a composite hypothesis testing problem, distinguishing between two alternatives: 1) $\hat{\boldsymbol{X}}$ is $\sqrt{\varepsilon d/2}$-close to *some* center $\boldsymbol{X}_i$; 2) $\hat{\boldsymbol{X}}$ is $\sqrt{\varepsilon d}$-far from *all* the centers. We show that for "typical" center configurations $\boldsymbol{\mathcal{X}}_k$, if $N \gtrsim \sigma^2 k \frac{\log(1/\varepsilon_{\text{I}})}{\varepsilon_{\text{I}}^2}$ then the test correctly throws away *all* the far points, and retains *most* of the close points. Since the true centers $\boldsymbol{\mathcal{X}}_k \sim \text{Unif}(\sqrt{d}\mathcal{S}^{d-1})^{\otimes k}$ are (w.h.p.) $\Omega(\sqrt{d})$-separated, the remaining points in $\mathcal{T}$, that have not been discarded, may be clustered into at most $k$ parts. Step I concludes by returning a list $\tilde{\boldsymbol{\mathcal{X}}}_{\text{I}}$ containing one representation of every cluster.

The dependence of Step I on the precision is sub-optimal: $N$ has to scale like $\frac{\log(1/\varepsilon_{\text{I}})}{\varepsilon_{\text{I}}^2}$ instead of $1/\varepsilon_{\text{I}}$. This sub-optimal rate is mended in Step II. We show that there is a *constant* precision level $\varepsilon_0$, that depends on $\mathsf{R}, \beta > 1$ (namely, *how much* the rate is smaller than the capacity) so that whenever $\varepsilon_{\text{I}} \leq \varepsilon_0$, one can construct a mismatched decoder, using $\tilde{\boldsymbol{\mathcal{X}}}_{\text{I}}$, that *consistently* decodes messages encoded by the true codebook $\boldsymbol{\mathcal{X}}_k$. In other words: given a measurement $\boldsymbol{Y} = \boldsymbol{X}_\ell + \sigma \boldsymbol{Z}$, one can consistently estimate the unknown label $\ell$ (up to a global re-labeling). In Step II we observe $\bar{N} = n - N$ new measurements, and cluster them according to their decoded label. For every cluster

$i \in [k]$, we compute the corresponding sample average $\boldsymbol{A}_i$, and project it onto the ball $\mathcal{B}(\mathbf{0}, \sqrt{d})$ to get our final estimate $\hat{\boldsymbol{X}}_i = \mathcal{P}(\boldsymbol{A}_i)$. Since each label $i$ witnesses, on average, $\bar{N}/k$ measurements, the MSE is, to leading order, $d^{-1}\mathbb{E}\|\boldsymbol{X}_i - \hat{\boldsymbol{X}}_i\|^2 = \sigma^2 k/\bar{N}$. Thus, using $N = C(\mathsf{R}, \beta)\sigma^2 k$ measurements for Step I, and $\bar{N} = \sigma^2 k/\varepsilon$ Step II, yields a list $\hat{\boldsymbol{\mathcal{X}}}_k$ with $\mathbb{E}\mathcal{L}_{\text{avg}}(\boldsymbol{\mathcal{X}}_k, \hat{\boldsymbol{\mathcal{X}}}_k) \le \varepsilon$.

In the remainder of this section, we provide the full details of the strategy outlined above.

### 4.1. Step I: Brute-Force Search

Let $\varepsilon_{\text{I}} \in (0, 1/2)$ a precision parameter. Let $\mathcal{T}$ be a fixed $\sqrt{\varepsilon_{\text{I}} d/2}$-net of $\sqrt{d}\mathcal{S}^{d-1}$, such that $\forall \boldsymbol{X} \in \sqrt{d}\mathcal{S}^{d-1} \exists \hat{\boldsymbol{X}} \in \mathcal{T}$ with $\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|^2 \le \varepsilon_{\text{I}} d/2$ . By standard estimates, e.g. (Wainwright, 2019, Example 5.8), we can assume that $|\mathcal{T}| \le e^{Cd\log(1/\varepsilon_{\text{I}})}$ for some universal $C > 0$. Our goal is to devise a procedure that, given $N$ samples $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_N \sim \text{GMM}_{d,k}(\boldsymbol{\mathcal{X}}_k, \sigma^2)$, will allow us to discard all candidates $\hat{\boldsymbol{X}} \in \mathcal{T}$ that are $\sqrt{\varepsilon_{\text{I}} d}$-far from *all* the centers $\boldsymbol{X}_1, \dots, \boldsymbol{X}_k$, while keeping enough candidates $\hat{\boldsymbol{X}}$ that are $\sqrt{\varepsilon_{\text{I}} d}$-close to some center; ideally, at least one candidate close to almost every $\boldsymbol{X}_i$. Denote the sets, $\mathcal{H}_{\text{Close}}, \mathcal{H}_{\text{Far}} \subset \mathbb{R}^d$

$$\mathcal{H}_{\text{Close}} = \left\{ \hat{\boldsymbol{X}} \,:\, \text{dist}^2(\hat{\boldsymbol{X}}, \boldsymbol{\mathcal{X}}_k) \le \frac{1}{2}\varepsilon_{\text{I}} d \right\}, \; \mathcal{H}_{\text{Far}} = \left\{ \hat{\boldsymbol{X}} \,:\, \text{dist}^2(\hat{\boldsymbol{X}}, \boldsymbol{\mathcal{X}}_k) \ge \varepsilon_{\text{I}} d \right\}. \tag{30}$$

We would like a test that, with high probability: 1) rejects all $\hat{\boldsymbol{X}} \in \mathcal{H}_{\text{Far}} \cap \mathcal{T}$; 2) accepts most of $\hat{\boldsymbol{X}} \in \mathcal{H}_{\text{Close}} \cap \mathcal{T}$. Note that since $\mathcal{T}$ is a $\sqrt{\varepsilon_I d/2}$-cover, then for *every* $1 \le i \le k$, there is some $\hat{\boldsymbol{X}} \in \mathcal{H}_{\text{Close}} \cap \mathcal{T}$ such that in fact $\|\hat{\boldsymbol{X}} - \boldsymbol{X}_i\|^2 \le \varepsilon_{\text{I}} d/2$.

As a first step, we consider a "local test" $\mathsf{Test} : \mathbb{R}^d \times \mathbb{R}^d \to \{0, 1\}$, that takes a candidate $\hat{\boldsymbol{X}} \in \mathcal{T}$ and a single sample $\boldsymbol{Y} \sim \text{GMM}_{d,k}$, and outputs a decision $\in \{0, 1\}$. Consider the quantities:

$$\begin{aligned} \mathsf{q}_{\text{Close}}(\boldsymbol{\mathcal{X}}_k, \mathcal{T}) &= \min_{\hat{\boldsymbol{X}} \in \mathcal{H}_{\text{Close}} \cap \mathcal{T}} \Pr(\mathsf{Test}(\hat{\boldsymbol{X}}, \boldsymbol{Y}) = 1 | \boldsymbol{\mathcal{X}}_k), \\ \mathsf{q}_{\text{Far}}(\boldsymbol{\mathcal{X}}_k, \mathcal{T}) &= \max_{\hat{\boldsymbol{X}} \in \mathcal{H}_{\text{Far}} \cap \mathcal{T}} \Pr(\mathsf{Test}(\hat{\boldsymbol{X}}, \boldsymbol{Y}) = 1 | \boldsymbol{\mathcal{X}}_k). \end{aligned} \tag{31}$$

For a local test $\mathsf{Test} : \mathbb{R}^d \times \mathbb{R}^d \to \{0, 1\}$, a cover $\mathcal{T}$, and $\nu > 0$, define

$$\mathcal{E}_{\mathsf{Test},\mathcal{T},\nu} = \left\{ \boldsymbol{\mathcal{X}}_k \in \left(\sqrt{d}\mathcal{S}^{d-1}\right)^k \,:\, \mathsf{q}_{\text{Close}}(\boldsymbol{\mathcal{X}}_k, \mathcal{T}) \ge \frac{1}{2}k^{-1}, \; \mathsf{q}_{\text{Far}}(\boldsymbol{\mathcal{X}}_k, \mathcal{T}) \le 2k^{-1-\nu} \right\}. \tag{32}$$

We construct a local test with the following properties.

**Lemma 9** *Assume that $\beta > 1$, and fix a cover $\mathcal{T}$ of size $|\mathcal{T}| \le e^{Cd\log(1/\varepsilon_{\text{I}})}$. There are positive constants $\varepsilon_0, c$, that depend on $\mathsf{R}, \beta$, and a local test, $\mathsf{Test} : \mathbb{R}^d \times \mathbb{R}^d \to \{0, 1\}$ (that depends on $d, k, \sigma^2$) such that for every* fixed $\varepsilon_{\text{I}} \in (0, \varepsilon_0)$

$$\lim_{d\to\infty} \Pr\left(\boldsymbol{\mathcal{X}}_k \notin \mathcal{E}_{\mathsf{Test},\mathcal{T},c\varepsilon_{\text{I}}^2}\right) = 0. \tag{33}$$

We propose a local test $\mathsf{Test}$ based on the capacity-achieving decoder used in the proof of Proposition 4. Due to space constraints, the details are deferred to Appendix, Section C.1.

Note that if $\mathsf{q}_{\text{Far}}(\boldsymbol{\mathcal{X}}_k, \mathcal{T}) \ll \mathsf{q}_{\text{Close}}(\boldsymbol{\mathcal{X}}_k, \mathcal{T})$, as is the case for $\boldsymbol{\mathcal{X}}_k \in \mathcal{E}_{\mathsf{Test},\mathcal{T},\nu}$, then by observing the statistics of the $N$ local test outputs $\{\mathsf{Test}(\hat{\boldsymbol{X}}, \boldsymbol{Y}_j)\}_{j=1}^N$, which are i.i.d. Bernoulli random

variables, one can distinguish between $\hat{\boldsymbol{X}} \in \mathcal{H}_{\text{Close}}$ and $\hat{\boldsymbol{X}} \in \mathcal{H}_{\text{Far}}$ with error probability vanishing in $N$. In particular, consider the candidates $\hat{\boldsymbol{X}} \in \mathcal{T}$ that pass the following threshold-based test

$$\mathcal{T}_{\text{Close}} = \left\{ \hat{\boldsymbol{X}} \in \mathcal{T} \quad : \quad \sum_{j=1}^{N} \mathsf{Test}(\hat{\boldsymbol{X}}, \boldsymbol{Y}_j) \geq \frac{1}{4} k^{-1} N \right\}. \tag{34}$$

**Lemma 10** *Fix any $\boldsymbol{\mathcal{X}}_k \in \mathcal{E}_{\mathsf{Test},\mathcal{T},c\varepsilon_{\text{I}}^2}$ and suppose that*

$$N \geq C_1 \sigma^2 k \frac{\log(1/\varepsilon_{\text{I}})}{\varepsilon_{\text{I}}^2} + C_2 k \log(1/\varphi)\,, \tag{35}$$

*where $C_1 = C_1(\mathsf{R}, \beta)$, $C_2 > 0$ is a universal constant and $\varphi \in (0,1)$. Then w.p. $1 - \varphi - o_{\beta,\mathsf{R}}(1)$ over $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \sim \mathrm{GMM}_{d,k}(\boldsymbol{\mathcal{X}}_k, \sigma^2)$, the following event holds:*

1. *(No far candidates). $\mathcal{T}_{\text{Close}} \cap \mathcal{H}_{\text{Far}} \neq \emptyset$.*
2. *(Most centers have a cluster). There is $\mathcal{I} \subseteq [k]$ with $|\mathcal{I}| \geq (1-\varphi)k$ and $\max_{i \in \mathcal{I}} \mathrm{dist}^2(\boldsymbol{X}_i, \mathcal{T}_{\text{Close}}) \leq \varepsilon_{\text{I}} d$.*

The proof of Lemma 10 appears in the Appendix, Section C.2.

To conclude step I, note that if the minimal distance between centers is $> 4\sqrt{\varepsilon_I d}$, then two candidates that are $\sqrt{\varepsilon_{\text{I}} d}$-close to different centers are necessarily $2\sqrt{\varepsilon_{\text{I}} d}$-far from one another.

Let $\tilde{\boldsymbol{\mathcal{X}}}_{\text{I}}$ be any $2\sqrt{\varepsilon_{\text{I}} d}$-separated subset of $\mathcal{T}_{\text{Close}}$ of maximal size. We prove in Appendix, Section C.3 that with high probability, $\boldsymbol{\mathcal{X}}_k$ indeed has $\Omega(\sqrt{d})$ minimal distance, and so:

**Lemma 11** *Assume that $\beta > 1$, $\varepsilon_{\text{I}} \leq \varepsilon_0(\mathsf{R}, \beta)$ is small enough, and $N$ satisfies (35). W.p. $1 - \varphi - o_{\beta,\mathsf{R},\varepsilon_{\text{I}}}(1)$ over* both $\boldsymbol{\mathcal{X}}_k \sim \mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1})^{\otimes k}$ *and* $[\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \,|\, \boldsymbol{\mathcal{X}}_k] \sim \mathrm{GMM}_{d,k}(\boldsymbol{\mathcal{X}}_k, \sigma^2)$, *the following event holds:*

1. *$\tilde{\boldsymbol{\mathcal{X}}}_{\text{I}}$ is a list of size $(1-\varphi)k \leq m \leq k$.*
2. *There is $\mathcal{I} \subseteq [k]$ with $|\mathcal{I}| = m$ so that for every $i \in \mathcal{I}$, there is a* unique *$\tilde{\boldsymbol{X}} \in \tilde{\boldsymbol{\mathcal{X}}}_{\text{I}}$ such that $\|\boldsymbol{X}_i - \tilde{\boldsymbol{X}}\|^2 \leq \varepsilon_{\text{I}} d$.*

### 4.2. Step II: Clustering and Averaging

Upon the successful completion of Step I, Lemma 11, we have produced a partial codebook $\tilde{\boldsymbol{\mathcal{X}}}_{\text{I}}$ of size $m \geq (1-\varphi)k$. Moreover, there is a large subset of messages $\mathcal{I} \subset [k]$, $|\mathcal{I}| = m$ such that for all $i \in \mathcal{I}$, the *true, unknown* codeword $\boldsymbol{X}_i$ is $\sqrt{\varepsilon_{\text{I}} d}$-close to a unique codeword $\tilde{\boldsymbol{X}}_l$ of $\tilde{\boldsymbol{\mathcal{X}}}_{\text{I}}$. Provided that $\varepsilon_{\text{I}} \leq \varepsilon_0(\mathsf{R}, \beta)$ is small enough (but constant), it turns out we can construct a "mismatched decoder", using $\tilde{\boldsymbol{\mathcal{X}}}_{\text{I}}$, that can consistently decode measurements $Y = \boldsymbol{X}_\ell + \sigma \boldsymbol{Z}$ in the following sense: 1) If $\ell \in \mathcal{I}$ then, up to a global relabeling, the decoder returns the correct label $\ell$; 2) If $\ell \notin \mathcal{I}$, the decoder consistently returns an error symbol $\#$. Due to space constraints, we defer all the details to Appendix, Section D.1.

In Step II we are given $\bar{N} = n - N$ new measurements. We use $\tilde{\boldsymbol{\mathcal{X}}}_{\text{I}}$, the codebook from Step I, to decode the corresponding labels; measurements for which the decoder returns $\#$ are *discarded*. We end up with $m$ clusters, and for each cluster $l \in [m]$ we compute the corresponding sample mean $\boldsymbol{A}_l$. Finally, we return the list $\hat{\boldsymbol{\mathcal{X}}}_k = (\hat{\boldsymbol{X}}_1, \ldots, \hat{\boldsymbol{X}}_k)$ such that $\hat{\boldsymbol{X}}_l = \mathcal{P}(\boldsymbol{A}_l)$ for $l \in [m]$, $\mathcal{P}(\cdot)$ being the projection onto the ball $\mathcal{B}(\boldsymbol{0}, \sqrt{d})$, and $\hat{\boldsymbol{X}}_l = \boldsymbol{0}$ for $m+1 \leq l \leq k$.

The following Lemma bounds the error of the *entire* end-to-end procedure, including both Step I and II. The details are deferred to Appendix, Section D.3.

**Lemma 12** *Assume that $\beta > 1$ and $\varepsilon \leq \varepsilon_0(\mathsf{R}, \beta)$ is small enough. Suppose that*

1. *Step I is run with $N \geq C\sigma^2 k + Ck\log(1/\varphi)$ measurements,*
2. *Step II is run with $\bar{N} \geq \frac{k\sigma^2}{\varepsilon} + C\frac{k}{\varepsilon^{1/2}}\log(1/\varphi)$ measurements,*

*where $C = C(\mathsf{R}, \beta)$ is constant and $\varphi \in (0,1)$ is a parameter. Then*

$$\lim_{d\to\infty} \mathbb{E}\mathcal{L}_{avg}(\boldsymbol{\mathcal{X}}_k, \hat{\boldsymbol{\mathcal{X}}}_k) \leq \frac{\varepsilon}{1-\varepsilon^{1/4}} + 12\varphi\,.$$

**Proof** (Of Theorem 1). The claimed lower bound follows from Lemma 6. The upper bound follows by setting, e.g., $\varphi = \varepsilon^2$ in Lemma 12, noting that as $\varepsilon \to 0$, $\frac{\varepsilon}{1-\varepsilon^{1/4}} = \varepsilon + o(\varepsilon)$. ■

## Acknowledgments

We are grateful to Uri Erez for helpful discussions. The work of ER and OO is supported in part by the ISF under grant 1641/21. ER is supported in part by an Einstein-Kaye fellowship from the Hebrew University of Jerusalem. TB is supported in part by the ISF grant no. 1924/21, the BSF grant no. 2020159, and the NSF-BSF grant no. 2019752.

## References

Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *International Conference on Computational Learning Theory*, pages 458–469. Springer, 2005.

Robert J Adler and Jonathan E Taylor. *Random fields and geometry*. Springer Science & Business Media, 2009.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of machine learning research*, 15: 2773–2832, 2014.

Joseph Anderson, Mikhail Belkin, Navin Goyal, Luis Rademacher, and James Voss. The more, the merrier: the blessing of dimensionality for learning large Gaussian mixtures. In *Conference on Learning Theory*, pages 1135–1164. PMLR, 2014.

Sanjeev Arora and Ravi Kannan. Learning mixtures of arbitrary gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 247–257, 2001.

Shiri Artstein-Avidan, Apostolos Giannopoulos, and Vitali D Milman. *Asymptotic geometric analysis, Part I*, volume 202. American Mathematical Soc., 2015.

Hassan Ashtiani, Shai Ben-David, Nicholas JA Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Nearly tight sample complexity bounds for learning mixtures of Gaussians via sample compression schemes. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3416–3425, 2018.

A. Barg and G.D. Forney. Random codes: minimum distances and error exponents. *IEEE Transactions on Information Theory*, 48(9):2568–2573, 2002.

Amir Bennatan, A Robert Calderbank, and Shlomo Shamai. Bounds on the MMSE of "bad" LDPC codes at rates above capacity. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pages 1065–1072, 2008.

Aditya Bhaskara, Moses Charikar, and Aravindan Vijayaraghavan. Uniqueness of tensor decompositions with applications to polynomial identifiability. In *Conference on Learning Theory*, pages 742–778. PMLR, 2014.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Ronit Bustin and Shlomo Shamai. MMSE of "bad" codes. *IEEE Transactions on Information Theory*, 59(2):733–743, 2013.

Siu-On Chan, Ilias Diakonikolas, Rocco A Servedio, and Xiaorui Sun. Efficient density estimation via piecewise polynomial approximation. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 604–613, 2014.

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

Sanjoy Dasgupta. Learning mixtures of Gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 634–644. IEEE, 1999.

Sanjoy Dasgupta and Leonard J Schulman. A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *Journal of Machine Learning Research*, 8:203–226, 2007.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. List-decodable robust mean estimation and learning mixtures of spherical Gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1047–1060, 2018.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

U. Erez and R. Zamir. Achieving 1/2 log (1+SNR) on the AWGN channel with lattice encoding and decoding. *IEEE Transactions on Information Theory*, 50(10):2293–2314, 2004.

Uri Erez. Lecture notes for principles of coding and detection in communication: Capacity of the AWGN channel. URL https://www.eng.tau.ac.il/~anatolyk/courses/Uri/AWGN_sphere_decoder.pdf.

Jon Feldman, Rocco A Servedio, and Ryan O'Donnell. PAC learning axis-aligned mixtures of Gaussians with no separation assumption. In *International Conference on Computational Learning Theory*, pages 20–34. Springer, 2006.

Robert G Gallager. *Information theory and reliable communication*, volume 2. Springer, 1968.

Rong Ge, Qingqing Huang, and Sham M Kakade. Learning mixtures of gaussians in high dimensions. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 761–770, 2015.

Navin Goyal, Santosh Vempala, and Ying Xiao. Fourier PCA and robust tensor decomposition. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 584–593, 2014.

Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, 2005.

Dongning Guo, Shlomo Shamai, and Sergio Verdú. The interplay between information and estimation measures. *Foundations and Trends® in Signal Processing*, 6(4):243–429, 2013.

Moritz Hardt and Eric Price. Tight bounds for learning a mixture of two gaussians. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 753–760, 2015.

Samuel B Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034, 2018.

Daniel Hsu and Sham M Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20, 2013.

Norman Johnson. *Continuous univariate distributions*. Wiley, New York, 1994. ISBN 0471584959.

Grigorii Anatol'evich Kabatiansky and Vladimir Iosifovich Levenshtein. On bounds for packings on a sphere and in space. *Problemy peredachi informatsii*, 14(1):3–25, 1978.

Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two Gaussians. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 553–562, 2010.

Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. *SIAM Journal on Computing*, 38(3):1141–1156, 2008.

Pravesh K Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046, 2018.

Jeongyeol Kwon and Constantine Caramanis. The EM algorithm gives sample-optimality for learning mixtures of well-separated gaussians. In *Conference on Learning Theory*, pages 2425–2487. PMLR, 2020.

Jerry Li and Ludwig Schmidt. Robust and proper learning for mixtures of gaussians via systems of polynomial inequalities. In *Conference on Learning Theory*, pages 1302–1382. PMLR, 2017.

Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 93–102. IEEE, 2010.

Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6(2012-2016):7, 2014.

Oded Regev and Aravindan Vijayaraghavan. On learning mixtures of well-separated gaussians. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 85–96. IEEE, 2017.

Elad Romanov, Tamir Bendory, and Or Ordentlich. Multi-reference alignment in high dimensions: sample complexity and phase transition. *SIAM Journal on Mathematics of Data Science*, 3(2): 494–523, 2021.

Claude E Shannon. Certain results in coding theory for noisy channels. *Information and control*, 1 (1):6–25, 1957.

Claude E Shannon. Probability of error for optimal codes in a Gaussian channel. *Bell System Technical Journal*, 38(3):611–656, 1959.

Nathan Srebro, Gregory Shakhnarovich, and Sam Roweis. An investigation of computational and informational limits in gaussian mixture clustering. In *Proceedings of the 23rd international conference on Machine learning*, pages 865–872, 2006.

Ananda Theertha Suresh, Alon Orlitsky, Jayadev Acharya, and Ashkan Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. *Advances in Neural Information Processing Systems*, 27, 2014.

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

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

## Appendix A. Proof of Proposition 4

As mentioned in the main text, the proof amounts to analyzing a certain sub-optimal decoder for the codebook $\boldsymbol{\mathcal{X}}_k$. While the decoders, and their analysis, are not new, we nonetheless provide all the details here as a "warm-up" for things to come.

We consider two different families of decoders, depending on whether one operates in the zero or positive rate regime.

### A.1. Rate Zero ($\mathsf{R} = 0$)

**The decoder.** Observe that for a spherical code, the MAP decoder (18) reduces to

$$\mathsf{DecOpt}(\boldsymbol{Y}) = \operatorname*{argmax}_{1 \le i \le k} d^{-1}\langle \boldsymbol{Y}, \boldsymbol{X}_i \rangle \,. \tag{36}$$

For the analysis, we consider a sub-optimal decoder, based on thresholding the correlation in (36).

When $\boldsymbol{Y} = \boldsymbol{X}_i + \sigma \boldsymbol{Z}$, clearly, $\mathbb{E}[d^{-1}\langle \boldsymbol{Y}, \boldsymbol{X}_i \rangle] = 1$, while for $j \neq i$, $\mathbb{E}[d^{-1}\langle \boldsymbol{Y}, \boldsymbol{X}_j \rangle] = 0$. Fix thresholds $0 < \eta_1 \le \eta_2$. Consider a decoding rule $\mathsf{DecCORR}_{\eta_1,\eta_2} : \mathbb{R}^d \to [k] \cup \{\#\}$ so that $\mathsf{DecCORR}_{\eta_1,\eta_2}(\boldsymbol{Y}) = i$ if and only if both of the following hold:

1. $d^{-1}\langle \boldsymbol{Y}, \boldsymbol{X}_i \rangle \ge 1 - \eta_1$.
2. For all $j \neq i$, $d^{-1}\langle \boldsymbol{Y}, \boldsymbol{X}_j \rangle < 1 - \eta_2$.

Note that since $\eta_1 \le \eta_2$, *at most* one index $1 \le i \le k$ can satisfy the above. If no such $i$ exists, we set $\mathsf{DecCORR}_{\eta_1,\eta_2}(\boldsymbol{Y}) = \#$.

**Analysis.** We proceed to bound the error of the decoder $\mathsf{DecCORR}_{\eta_1,\eta_2}$.

By symmetry of the codebook generating process, the error probability (averaged over the ensemble) does not depend on the particular transmitted message (index). For convenience, throughout this section, we always assume, without loss of generality, that the transmitted message is $\ell = i$ (and implicitly condition on this event). Thus, the value at the receiver end of the channel is $\boldsymbol{Y} = \boldsymbol{X}_i + \sigma \boldsymbol{Z}$.

**Lemma 13** *Conditioned on any $\boldsymbol{\mathcal{X}}_k \in (\sqrt{d}\mathcal{S}^{d-1})^k$,*

$$\Pr\left(d^{-1}\langle \boldsymbol{Y}, \boldsymbol{X}_i \rangle < 1 - \eta_1 \,|\, \boldsymbol{\mathcal{X}}_k\right) \le e^{-\frac{\eta_1^2}{2\sigma^2} d} \,.$$

**Proof** $d^{-1}\langle \boldsymbol{X}_i + \sigma \boldsymbol{Z}, \boldsymbol{X}_i \rangle \le 1 - \eta_1$ is equivalent to $d^{-1}\langle \sigma \boldsymbol{Z}, \boldsymbol{X}_i \rangle \le -\eta_1$. Since $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, we have $d^{-1}\langle \sigma \boldsymbol{Z}, \boldsymbol{X}_i \rangle \sim \mathcal{N}(0, \sigma^2/d)$, and the bound follows immediately. ∎

**Lemma 14** *For fixed $\eta_1 > 0$, define the set*

$$\mathbb{X}_i = \left\{ \boldsymbol{\mathcal{X}}_k \in (\sqrt{d}\mathcal{S}^{d-1})^k \;:\; \max_{j \in [n] \setminus \{i\}} d^{-1}\langle \boldsymbol{X}_i, \boldsymbol{X}_j \rangle \le \sqrt{\frac{2\log(k-1)}{d} + \frac{\eta_1^2}{\sigma^2}} \right\} .$$

*For $\boldsymbol{\mathcal{X}}_k \sim \mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1})$: $\Pr(\boldsymbol{\mathcal{X}}_k \notin \mathbb{X}_i) \le e^{-\frac{\eta_1^2}{2\sigma^2} d}$.*

**Proof** By the standard tail bound Lemma 35, for $t \geq 0$, $\Pr(d^{-1}\langle \boldsymbol{X}_i, \boldsymbol{X}_j\rangle \geq t) \leq e^{-dt^2/2}$. Taking a union bound over $(k-1)$ choices for $j \neq i$, $\Pr(\max_{j\neq i} d^{-1}\langle \boldsymbol{X}_i, \boldsymbol{X}_j\rangle \geq t) \leq e^{-dt^2/2+\log(k-1)}$. Now set $t = \sqrt{\frac{2\log(k-1)}{d} + \frac{\eta_1^2}{\sigma^2}}$. ∎

**Lemma 15** *Suppose that* $1-\eta_2 \geq \sqrt{\frac{2\log(k-1)}{d} + \frac{\eta_1^2}{\sigma^2}} + \sqrt{\frac{2\sigma^2\log(k-1)}{d}}$. *Then*

$$\Pr\left(\max_{j\neq i} d^{-1}\langle \boldsymbol{Y}, \boldsymbol{X}_j\rangle \geq 1-\eta_2\right) \leq e^{-\frac{d}{2}\left(1-\eta_2-\sqrt{\frac{2\log(k-1)}{d}+\frac{\eta_1^2}{\sigma^2}}-\sqrt{\frac{2\sigma^2\log(k-1)}{d}}\right)^2} + e^{-\frac{\eta_1^2}{2\sigma^2}d}. \quad (37)$$

**Proof** Fix $\boldsymbol{\mathcal{X}}_k \in \mathbb{X}_i$, where the set $\mathbb{X}_i$ is from Lemma 14. Writing $\boldsymbol{Y} = \boldsymbol{X}_i + \sigma\boldsymbol{Z}$, we note that

$$\Pr\left(\max_{j\neq i} d^{-1}\langle \boldsymbol{Y}, \boldsymbol{X}_j\rangle \geq 1-\eta_2 \,\middle|\, \boldsymbol{\mathcal{X}}_k\right) \leq \Pr\left(\max_{j\neq i} d^{-1}\sigma\langle \boldsymbol{Z}, \boldsymbol{X}_j\rangle \geq 1-\eta_2-\sqrt{\frac{2\log(k-1)}{d}+\frac{\eta_1^2}{\sigma^2}} \,\middle|\, \boldsymbol{\mathcal{X}}_k\right).$$

Now, each $d^{-1}\sigma\langle \boldsymbol{Z}, \boldsymbol{X}_j\rangle$ is Gaussian with mean 0 and variance $\sigma^2/d$. By a standard bound on the maximum of Gaussian random variables, Lemma 41, $\mathbb{E}[\max_{j\neq i} d^{-1}\sigma\langle \boldsymbol{Z}, \boldsymbol{X}_j\rangle] \leq \sqrt{\frac{2\sigma^2\log(k-1)}{d}}$. By the Borell-TIS inequality, Lemma 42, we obtain the first term of (37). The second term is just the bound on $\Pr(\boldsymbol{\mathcal{X}}_k \notin \mathbb{X}_i)$ from Lemma 14. ∎

**Proof** (Of Proposition 4, case $\mathsf{R} = 0$.)

Combining Lemmas 13 and 15, for every $\eta_1, \eta_2$ satisfying

$$0 < \eta_1 \leq \eta_2 < 1 - \sqrt{\frac{2\log(k-1)}{d} + \frac{\eta_1^2}{\sigma^2}} - \sqrt{\frac{2\sigma^2\log(k-1)}{d}} \quad (38)$$

the decoder $\mathsf{DecCORR}_{\eta_1,\eta_2}$ attains average error

$$\Pr\left(i \neq \mathsf{DecCORR}_{\eta_1,\eta_2}(\boldsymbol{X}_i + \sigma\boldsymbol{Z})\right) \leq e^{-\frac{d}{2}\left(1-\eta_2-\sqrt{\frac{2\log(k-1)}{d}+\frac{\eta_1^2}{\sigma^2}}-\sqrt{\frac{2\sigma^2\log(k-1)}{d}}\right)^2} + 2e^{-\frac{\eta_1^2}{2\sigma^2}d}. \quad (39)$$

To prove the proposition, it clearly suffices to show that when $\mathsf{R}_{d,k} = \mathsf{C}(\beta\sigma^2)$, $\beta > 1$, then $\sqrt{\frac{2\log(k-1)}{d} + \frac{\eta_1^2}{\sigma^2}} + \sqrt{\frac{2\sigma^2\log(k-1)}{d}}$ is at most a constant, which is strictly smaller than 1. Indeed, since $\sigma^2 = \omega(1)$, the first term is $o(1)$. As for the second term,

$$\sqrt{\frac{2\sigma^2\log(k-1)}{d}} \leq \sqrt{2\sigma^2\mathsf{R}_{d,k}} = \sqrt{2\sigma^2\mathsf{C}(\beta\sigma^2)} \leq \sqrt{1/\beta} < 1,$$

where we used $\mathsf{C}(s) = \frac{1}{2}\log(1+1/s) \leq 1/(2s)$.

∎

### A.2. Positive Rate ($\mathsf{R} > 0$)

**Remark.** The analysis of the previous section ($\mathsf{R} = 0$) unfortunately fails in the positive rate regime, where $\sigma^2$ is constant. To have any hope of finding $\eta_1 \le \eta_2$ that satisfy condition (38), it is necessary that (taking $\eta_1, \eta_2 \to 0$)

$$\sqrt{\frac{2\log(k-1)}{d}} + \sqrt{\frac{2\sigma^2\log(k-1)}{d}} \le 1\,.$$

Since $\log(k-1)/d = \mathsf{R}_{d,k} - O(d^{-1})$, this constrains the rate as $\mathsf{R}_{d,k} \le \frac{1}{\sqrt{2}(1+\sigma)} + O(d^{-1})$. For small $\sigma$, this bound is $\approx 1/\sqrt{2}$, while $\mathsf{C}(\beta\sigma^2) \approx \log(1/\beta\sigma^2)$. Consequently, for $\sigma^2 = O(1)$ this condition fails to hold, and the analysis from Section A.1 is not sufficient for proving the existence of capacity-approaching codes. We note that this is a well-known limitation of the *analysis*; specifically, Lemma 15 is too crude. It estimates the maximum over "noise terms", $\max_{j\neq i} d^{-1}\sigma\langle \boldsymbol{Z}, \boldsymbol{X}_j\rangle$ *as if* they were all independent. In the zero rate regime, different codewords are essentially orthogonal: $d^{-1}\langle \boldsymbol{X}_i, \boldsymbol{X}_j\rangle \lesssim \sqrt{\frac{\log k}{d}} = o(1)$; consequently, by standard results (e.g. (Adler and Taylor, 2009, 2.2.5)), the maximum is indeed very close to the maximum of i.i.d. Gaussians. When $k$ is exponential, however, this is no longer the case, and the correlations between these noise terms can no longer be neglected once $\mathsf{R}$ is sufficiently large. Thus, different techniques are necessary to carry out the analysis (cf. the classical book Gallager (1968)).

**The decoder.** To overcome the obstruction mentioned above, we consider a different, sub-optimal, decoder, which is similar to Shannon's information density threshold decoder Shannon (1957) for a Gaussian i.i.d. codebook, and to that used in Erez and Zamir (2004), see also Erez and Polyanskiy and Wu (2014). Let $\alpha = \frac{1}{1+\sigma^2}$ and $\tau = \sigma^2\alpha = 1-\alpha$. For parameters $\tau \le \tau_1 \le \tau_2$, consider a decoder $\mathsf{DecMMSE}_{\tau_1,\tau_2} : \mathbb{R}^d \to [k]\cup\{\#\}$ so that $\mathsf{DecMMSE}_{\tau_1,\tau_2}(\boldsymbol{Y}) = i$ if and only if both of the following hold:

1. $d^{-1}\|\alpha\boldsymbol{Y} - \boldsymbol{X}_i\|^2 \le \tau_1$.
2. For all $j \neq i$, $d^{-1}\|\alpha\boldsymbol{Y} - \boldsymbol{X}_i\|^2 > \tau_2$.

If no such $1 \le i \le k$ exists, then $\mathsf{DecMMSE}_{\tau_1,\tau_2}(\boldsymbol{Y}) = \#$.

As was before, in the zero rate case, we analyze the error probability conditioned on the transmitted message being some fixed $\ell = i$; by symmetry, the (ensemble-averaged) error probability does not depend on $\ell$. Thus, below, $\boldsymbol{Y} = \boldsymbol{X}_i + \sigma\boldsymbol{Z}$.

To justify the name $\mathsf{DecMMSE}$ recall that the best linear estimator of $\boldsymbol{X}_i$ from $\boldsymbol{Y} = \boldsymbol{X}_i + \sigma\boldsymbol{Z}$, in the sense of smallest MSE (LMMSE), is $\alpha\boldsymbol{Y}$.[2] Note also that $d^{-1}\mathbb{E}\|\alpha\boldsymbol{Y} - \boldsymbol{X}_i\|^2 = \tau$, whereas for $j \neq i$, $d^{-1}\mathbb{E}\|\alpha\boldsymbol{Y} - \boldsymbol{X}_j\|^2 = \alpha^2(1+\sigma^2) + 1 = \alpha + 1 > \tau$.

**Analysis.** We proceed to bound the error of the decoder $\mathsf{DecMMSE}_{\tau_1,\tau_2}$.

**Lemma 16** *For any $\boldsymbol{\mathcal{X}}_k \in (\sqrt{d}\mathcal{S}^{d-1})^k$,*

$$\Pr\left(d^{-1}\|\alpha\boldsymbol{Y} - \boldsymbol{X}_i\|^2 > \tau_1 \,\middle|\, \boldsymbol{\mathcal{X}}_k\right) \le e^{-\frac{1}{2}(1+\sigma^2)(\sqrt{\tau_1/\tau}-1)^2 d}\,.$$

2. When $\boldsymbol{X}_i \sim \mathcal{N}(0,1)^{\otimes d}$ is i.i.d. Gaussian, the LMMSE is actually the MSE-optimal estimator (MMSE). Since we use a spherical prior for $\boldsymbol{X}_i \sim \mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1})$, this is no longer holds exactly, though the discrepancy is negligible when one operates in the regime $\sigma^2 = \Omega(1)$.

**Proof** The mapping $\boldsymbol{Z} \mapsto F(\boldsymbol{Z}) = d^{-1/2}\|\alpha(\boldsymbol{X}_i + \sigma\boldsymbol{Z}) - \boldsymbol{X}_i\|$ is $d^{-1/2}\alpha\sigma$-Lipschitz, with expectation

$$\mathbb{E}F(\boldsymbol{Z}) \le \sqrt{d^{-1}\mathbb{E}\|\alpha(\boldsymbol{X}_i + \sigma\boldsymbol{Z}) - \boldsymbol{X}_i\|^2} \le \sqrt{\tau}\,.$$

Applying the Gaussian Lipschitz concentration inequality, Lemma 34,

$$\begin{aligned}\Pr\left((F(\boldsymbol{Z}))^2 > \tau_1\right) &= \Pr\left(F(\boldsymbol{Z}) > \sqrt{\tau_1}\right)\\ &\le \Pr\left(F(\boldsymbol{Z}) - \mathbb{E}F(\boldsymbol{Z}) > \sqrt{\tau_1} - \sqrt{\tau}\,\middle|\,\boldsymbol{\mathcal{X}}_k\right)\\ &\le e^{-\frac{1}{2}\frac{(\sqrt{\tau_1}-\sqrt{\tau})^2}{(d^{-1/2}\alpha\sigma)^2}}\\ &= e^{-\frac{\tau}{2\alpha^2\sigma^2}(\sqrt{\tau_1/\tau}-1)^2 d}\,.\end{aligned}$$

Now plug $\tau = \alpha\sigma^2$, $\alpha = 1/(1+\sigma^2)$ to get the claimed bound. ■

**Lemma 17** *For $j \neq i$, $\boldsymbol{\mathcal{X}}_k \sim (\mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1}))^k$,*

$$\Pr\left(d^{-1}\|\alpha\boldsymbol{Y}_i - \boldsymbol{X}_j\|^2 \le \tau_2\right) \le \left(1 + \frac{1}{\sigma^2}\right)^{1/2} e^{-\left(\mathsf{C}(\sigma^2) - \frac{1}{2}\log\frac{\tau_2}{\tau}\right)d}\,.$$

**Proof** For a compact convex body $K \subset \mathbb{R}^d$, we denote its boundary by $\partial K$ and surface area by $\mathrm{Surf}(\partial K)$. In addition, we denote the Euclidean ball of radius $r$, centered around $\boldsymbol{a} \in \mathbb{R}^d$, by $\mathcal{B}(\boldsymbol{a}, r)$.

The event above, whose probability we wish to bound, is equivalent to the event $\boldsymbol{X}_j \in \mathcal{B}(\alpha\boldsymbol{Y}_i, \sqrt{\tau_2 d})$. Since $\boldsymbol{X}_j \sim \mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1})$, this probability (conditioned on $\boldsymbol{Y}$), is given by the surface area ratios $\mathrm{Surf}(\sqrt{d}\mathcal{S}^{d-1} \cap \mathcal{B}(\alpha\boldsymbol{Y}, \sqrt{\tau_2 d}))/\mathrm{Surf}(\sqrt{d}\mathcal{S}^{d-1})$. Since $\partial K \cap L \subseteq \partial(K \cap L)$, and the surface area of convex sets is monotonic with respect to containment (e.g., (Artstein-Avidan et al., 2015, Theorem B.1.14)), $\mathrm{Surf}(\sqrt{d}\mathcal{S}^{d-1} \cap \mathcal{B}(\alpha\boldsymbol{Y}, \sqrt{\tau_2 d})) \le \mathrm{Surf}(\partial\mathcal{B}(\alpha\boldsymbol{Y}, \sqrt{\tau_2 d}))$. Consequently, the probability is bounded by

$$\frac{\mathrm{Surf}(\partial\mathcal{B}(\alpha\boldsymbol{Y}_i, \sqrt{\tau_2 d}))}{\mathrm{Surf}(\sqrt{d}\mathcal{S}^{d-1})} = \tau_2^{\frac{d-1}{2}} = \tau_2^{-1/2} e^{-\frac{1}{2}\left(\log\frac{1}{\tau} - \log\frac{\tau_2}{\tau}\right)d}\,.$$

Lastly, use $\tau_2 \ge \tau$ and $1/\tau = 1 + 1/\sigma^2$. ■

**Proof** (Of Proposition 4, case $\mathsf{R} > 0$.)

Combining Lemmas 16 and 17, along with a union bound over all $j \neq i$, the decoding error, averaged over the ensemble $\boldsymbol{\mathcal{X}}_k \sim \mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1})^{\otimes k}$, is bounded as

$$\Pr\left(i \neq \mathsf{DecMMSE}_{\tau_1,\tau_2}(\boldsymbol{Y})\right) \le e^{-\frac{1}{2}(1+\sigma^2)(\sqrt{\tau_1/\tau}-1)^2 d} + (k-1)\left(1 + \frac{1}{\sigma^2}\right)^{1/2} e^{-\left(\mathsf{C}(\sigma^2) - \frac{1}{2}\log\frac{\tau_2}{\tau}\right)d}\,. \tag{40}$$

Choose $\tau_1 = c\tau$, $\tau_2 = c^2\tau$ where $c > 1$ is a sufficiently small constant. In that case, the first term of (40) clearly decays exponentially in $d$.

As for the second term, set $k = e^{d\mathsf{C}(\beta\sigma^2)}$, so the term is bounded like $(1+\frac{1}{\sigma^2})^{1/2} e^{-d\left(\mathsf{C}(\sigma^2) - \mathsf{C}(\beta\sigma^2) - \log(c)\right)}$. Since $\mathsf{C}(\sigma^2) - \mathsf{C}(\beta\sigma^2)$ is a positive constant, if $c > 1$ is small enough then the term decays exponentially fast in $d$. ■

## Appendix B. Proofs from Section 3

### B.1. Proof of Lemma 5

We reduce the calculation into a "standard" rate-distortion function (RDF) under MSE distortion.

For technical reasons, it will be more convenient to work with a Gaussian prior on the source signal, rather than the uniform distribution over the sphere. The reason is that the latter distribution is not absolutely continuous with respect to Lebesgue measure (it is supported on a manifold of positive co-dimension, namely, $\mathcal{S}^{d-1}$), so that its differential entropy (in the usual sense) is not well-defined.

Introduce Gaussian random variables, $\boldsymbol{\mathcal{G}}_k = (\boldsymbol{G}_1, \dots, \boldsymbol{G}_k) \sim \mathcal{N}(0,1)^{\otimes dk}$, so that $\boldsymbol{X}_i = \sqrt{d}\boldsymbol{G}_i/\|\boldsymbol{G}_i\|$. We have the Markov chain,

$$\boldsymbol{\mathcal{G}}_k \longrightarrow \boldsymbol{\mathcal{X}}_k \longrightarrow \hat{\boldsymbol{\mathcal{X}}}_k\,, \tag{41}$$

so by the DPI,

$$I(\boldsymbol{\mathcal{X}}_k; \hat{\boldsymbol{\mathcal{X}}}_k) \geq I(\boldsymbol{\mathcal{G}}_k; \hat{\boldsymbol{\mathcal{X}}}_k)\,. \tag{42}$$

The next Lemma shows that if $\hat{\boldsymbol{\mathcal{X}}}_k$ estimates $\boldsymbol{\mathcal{X}}_k$ with small distortion, then it also estimates $\boldsymbol{\mathcal{G}}_k$ with small comparable distortion:

**Lemma 18** *Suppose the the Markov chain (41) holds. Then for universal constant $c_0$,*

$$\mathbb{E}\mathcal{L}(\boldsymbol{\mathcal{G}}_k, \hat{\boldsymbol{\mathcal{X}}}_k) \leq \left(\sqrt{\mathbb{E}\mathcal{L}(\boldsymbol{\mathcal{X}}_k, \hat{\boldsymbol{\mathcal{X}}}_k)} + c_0 d^{-1/2}\right)^2\,.$$

The proof of Lemma 18 is straightforward, and deferred to Section B.2.

By assumption, $\mathbb{E}\mathcal{L}(\boldsymbol{\mathcal{X}}_k, \hat{\boldsymbol{\mathcal{X}}}_k) \leq \varepsilon$ and therefore, by Lemma 18, $\mathbb{E}\mathcal{L}(\boldsymbol{\mathcal{G}}_k, \hat{\boldsymbol{\mathcal{X}}}_k) \leq (\varepsilon^{1/2} + c_0 d^{-1/2})^2$.

Let $J = (j_1, \dots, j_k) \in [k]^k$ be indices such that $j_i \in \operatorname{argmin}_{1\leq j\leq k} \|\boldsymbol{G}_i - \hat{\boldsymbol{X}}_j\|^2$. In other words, $\mathcal{L}(\boldsymbol{\mathcal{G}}_k, \hat{\boldsymbol{\mathcal{X}}}_k) = (dk)^{-1}\sum_{i=1}^k \|\boldsymbol{G}_i - \hat{\boldsymbol{X}}_{j_i}\|^2$. The random variable $J$ is, clearly, deterministic given $\boldsymbol{\mathcal{G}}_k, \hat{\boldsymbol{\mathcal{X}}}_k$. By the chain rule for mutual information,

$$I(\boldsymbol{\mathcal{G}}_k; \hat{\boldsymbol{\mathcal{X}}}_k) = I(\boldsymbol{\mathcal{G}}_k; \hat{\boldsymbol{\mathcal{X}}}_k, J) - I(\boldsymbol{\mathcal{G}}_k; J|\hat{\boldsymbol{\mathcal{X}}}_k)\,. \tag{43}$$

Since $J$ is a discrete random variable,

$$I(\boldsymbol{\mathcal{G}}_k; J|\hat{\boldsymbol{\mathcal{X}}}_k) := H(J|\hat{\boldsymbol{\mathcal{X}}}_k) - H(J|\boldsymbol{\mathcal{G}}_k, \hat{\boldsymbol{\mathcal{X}}}_k) \leq H(J) \leq \log(k^k) = k\log k\,, \tag{44}$$

where we used the standard facts that the entropy of a discrete variable is non-negative, and that conditioning decreases entropy.

Set $\boldsymbol{D}_i = \boldsymbol{X}_{j_i}$ and $\boldsymbol{\mathcal{D}} = (\boldsymbol{D}_1, \dots, \boldsymbol{D}_k) \in \mathbb{R}^{k\times d}$, so that, by definition,

$$(dk)^{-1}\mathbb{E}\|\boldsymbol{\mathcal{G}}_k - \boldsymbol{\mathcal{D}}\|_F^2 = \mathbb{E}\mathcal{L}_{\text{avg}}(\boldsymbol{\mathcal{G}}_k, \boldsymbol{\mathcal{D}}) \leq (\varepsilon^{1/2} + c_0 d^{-1/2})^2\,.$$

Since $\boldsymbol{\mathcal{D}}$ is a function of $(\hat{\boldsymbol{\mathcal{X}}}_k, J)$, the DPI implies $I(\boldsymbol{\mathcal{G}}_k; \boldsymbol{\mathcal{D}}) \leq I(\boldsymbol{\mathcal{G}}_k; \hat{\boldsymbol{\mathcal{X}}}_k, J)$. Thus,

$$\begin{aligned} I(\boldsymbol{\mathcal{G}}_k; \hat{\boldsymbol{\mathcal{X}}}_k, J) &\geq I(\boldsymbol{\mathcal{G}}_k; \boldsymbol{\mathcal{D}}) \\ &\geq \min_{P_{\tilde{\boldsymbol{\mathcal{D}}}|\boldsymbol{\mathcal{G}}_k}\,:\,(dk)^{-1}\|\boldsymbol{\mathcal{G}}_k - \tilde{\boldsymbol{\mathcal{D}}}\|^2 \leq (\varepsilon^{1/2} + c_0 d^{-1/2})^2} I(\boldsymbol{\mathcal{G}}_k; \tilde{\boldsymbol{\mathcal{D}}}) \\ &= \frac{dk}{2}\log\left(\frac{1}{(\varepsilon^{1/2} + c_0 d^{-1/2})^2}\right)\,, \end{aligned} \tag{45}$$

where (45) is the solution to the classical Gaussian source rate-distortion problem (Polyanskiy and Wu, 2014, Chapter 27). The proof of Lemma 5 concludes by combining (42)-(45). ■

We remark that for sufficiently small $\varepsilon$ the lower bound

$$\min_{P_{\hat{\boldsymbol{\mathcal{X}}}_k|\boldsymbol{\mathcal{G}}_k}:\mathbb{E}\mathcal{L}_{\text{avg}}(\boldsymbol{\mathcal{G}}_k,\boldsymbol{\mathcal{D}})\le\varepsilon} I(\boldsymbol{\mathcal{G}}_k;\hat{\boldsymbol{\mathcal{X}}}_k) \ge \frac{dk}{2}\log\left(\frac{1}{\varepsilon}\right) - k\log k, \tag{46}$$

which we derived within the proof above, is in fact tight (up to the difference between $k\log k$ and $\log|\mathcal{S}_k|$, where $\mathcal{S}_k$ is the symmetric group of permutations on $[k]$). To see this, we consider the Markov chain $\boldsymbol{\mathcal{G}}_k \to \hat{\boldsymbol{\mathcal{G}}}_k \to \hat{\boldsymbol{\mathcal{X}}}_k$, where the channel from $\boldsymbol{\mathcal{G}}_k$ to $\hat{\boldsymbol{\mathcal{G}}}_k$ is the test channel attaining the Gaussian RDF (see e.g., (Cover and Thomas, 2012, Theorem 10.3.2)), and the channel from $\hat{\boldsymbol{\mathcal{G}}}_k \to \hat{\boldsymbol{\mathcal{X}}}_k$ is defined by applying a uniform random permutation $J$ on $\hat{\boldsymbol{\mathcal{G}}}_k$, resulting in $\hat{\boldsymbol{\mathcal{X}}}_k$. Note that

$$I(\boldsymbol{\mathcal{G}}_k;\hat{\boldsymbol{\mathcal{X}}}_k,J) = I(\boldsymbol{\mathcal{G}}_k;\hat{\boldsymbol{\mathcal{G}}}_k) = \frac{dk}{2}\log\left(\frac{1}{\varepsilon}\right) \tag{47}$$

and that

$$I(\boldsymbol{\mathcal{G}}_k;J|\hat{\boldsymbol{\mathcal{X}}}_k) = H(J|\hat{\boldsymbol{\mathcal{X}}}_k) - H(J|\boldsymbol{\mathcal{G}}_k,\hat{\boldsymbol{\mathcal{X}}}_k) = H(J) - H(J|\boldsymbol{\mathcal{G}}_k,\hat{\boldsymbol{\mathcal{X}}}_k) \approx H(J), \tag{48}$$

where the last approximation is due to the fact that for small $\varepsilon$ we can recover $J$ from $\boldsymbol{\mathcal{G}}_k$ and $\hat{\boldsymbol{\mathcal{X}}}_k$. Thus, the approximate tightness of (46) follows from (43).

The subtractive $k\log k$ term we lose here is the reason that the lower bound in Theorem 1 is $e^{-2\mathsf{R}}$ instead of 1. While we believe that 1 is the correct lower bound, this loss seems to be inherent to the mutual information bounding program we follow here.

### B.2. Proof of Lemma 18

It is a well-known fact (Johnson, 1994, Eq. 18.15) that $\mathbb{E}\|\boldsymbol{G}_i\| = \sqrt{d} + O(d^{-1/2})$. Consequently,

$$\mathbb{E}\|\boldsymbol{G}_i - \boldsymbol{X}_i\|^2 = 2d - 2\sqrt{d}\mathbb{E}\|\boldsymbol{G}_i\| = O(1)\,.$$

Let $J = (j_1,\ldots,j_k)$ be $j_i = \text{argmin}_{1\le j\le k}\|\boldsymbol{X}_i - \hat{\boldsymbol{X}}_i\|$. By definition of $\mathcal{L}_{\text{avg}}(\cdot,\cdot)$, (5),

$$\mathcal{L}_{\text{avg}}(\boldsymbol{\mathcal{G}}_k,\hat{\boldsymbol{\mathcal{X}}}_k) \le \frac{1}{dk}\sum_{i=1}^{k}\mathbb{E}\|\boldsymbol{G}_i - \hat{\boldsymbol{X}}_{j_i}\|^2\,,$$

while $\mathcal{L}_{\text{avg}}(\boldsymbol{\mathcal{X}}_k,\hat{\boldsymbol{\mathcal{X}}}_k) = \frac{1}{dk}\sum_{i=1}^{k}\mathbb{E}\|\boldsymbol{X}_i - \hat{\boldsymbol{X}}_{j_i}\|^2$. Moreover, observe that

$$\boldsymbol{\mathcal{D}} = (\boldsymbol{D}_1,\ldots,\boldsymbol{D}_k) \mapsto \left(\frac{1}{dk}\sum_{i=1}^{k}\mathbb{E}\|\boldsymbol{D}_i\|^2\right)^{1/2}$$

defines a semi-norm on $d \times k$ matrices (with square-integrable entries). Thus, by the triangle inequality,

$$\begin{aligned}\left(\mathbb{E}\mathcal{L}_{\text{avg}}(\boldsymbol{\mathcal{G}}_k, \hat{\boldsymbol{\mathcal{X}}}_k)\right)^{1/2} &\leq \left(\frac{1}{dk}\sum_{i=1}^{k}\mathbb{E}\|\boldsymbol{G}_i - \hat{\boldsymbol{X}}_{j_i}\|^2\right)^{1/2} \\ &= \left(\frac{1}{dk}\sum_{i=1}^{k}\mathbb{E}\|(\boldsymbol{G}_i - \boldsymbol{X}_i) + (\boldsymbol{X}_i - \hat{\boldsymbol{X}}_{j_i})\|^2\right)^{1/2} \\ &\leq \underbrace{\left(\frac{1}{dk}\sum_{i=1}^{k}\mathbb{E}\|\boldsymbol{G}_i - \boldsymbol{X}_i\|^2\right)^{1/2}}_{\left(\frac{1}{dk}\sum_{i=1}^{k} O(1)\right)^{1/2} = O(d^{-1/2})} + \underbrace{\left(\frac{1}{dk}\sum_{i=1}^{k}\mathbb{E}\|\boldsymbol{X}_i - \hat{\boldsymbol{X}}_{j_i}\|^2\right)^{1/2}}_{\left(\mathcal{L}_{\text{avg}}(\boldsymbol{\mathcal{X}}_k, \hat{\boldsymbol{\mathcal{X}}}_k)\right)^{1/2}}.\end{aligned}$$

■

### B.3. Proof of Lemma 6

Write $I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n, \boldsymbol{\ell}) = I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{\ell}) + I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n | \boldsymbol{\ell})$, with $I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{\ell}) = 0$. For $1 \leq i \leq k$, let $n_i = n_i(\boldsymbol{\ell}) = |\boldsymbol{\ell}^{-1}(i)|$ be the number of measurements labeled $i$. The proof amounts to the following observation: the desired MI $I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n | \boldsymbol{\ell})$ is simply the cumulative MI across $k$ parallel Gaussians channel, with independent inputs $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$, such that one observes $n_i$ outputs (samples) of each channel $i$. We now quantify this statement.

Let $I(\sigma^2, m) = I(\boldsymbol{X}; \boldsymbol{X} + \sigma\boldsymbol{Z}_1, \ldots, \boldsymbol{X} + \sigma\boldsymbol{Z}_m)$ be the input-output MI between $\boldsymbol{X}$ and $m$ outputs through an $\mathsf{AWGN}(\sigma^2)$ channel. Since the sample mean is a sufficient statistic for the true mean under a Gaussian measurement model, we have

$$I(\sigma^2, m) = I\left(\boldsymbol{X}, \frac{1}{m}\left((\boldsymbol{X} + \sigma\boldsymbol{Z}_1, \ldots, \boldsymbol{X} + \sigma\boldsymbol{Z}_m)\right)\right) = I\left(\frac{\sigma^2}{m}, 1\right) \leq d\mathsf{C}(\sigma^2/m)\,,$$

where $\boldsymbol{C}(\cdot)$ denotes the AWGN channel capacity (9). Thus,

$$I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n | \boldsymbol{\ell}) = \mathbb{E}\left[\sum_{i=1}^{k} I(\sigma^2, n_i(\boldsymbol{\ell}))\right] \leq \mathbb{E}\left[\sum_{i=1}^{k} d\mathsf{C}(\sigma^2/n_i(\boldsymbol{\ell}))\right].$$

One may readily verify that the function $m \mapsto \mathsf{C}(\sigma^2/m)$ is concave. By Jensen's inequality,

$$\sum_{i=1}^{k} d\mathsf{C}(\sigma^2/n(\boldsymbol{\ell})) = k \cdot \frac{1}{k}\sum_{i=1}^{k} d\mathsf{C}(\sigma^2/n(\boldsymbol{\ell})) \leq kd\mathsf{C}\left(\frac{\sigma^2}{\frac{1}{k}\sum_{i=1}^{k} n_i(\boldsymbol{\ell})}\right) = kd\mathsf{C}\left(\sigma^2 k/n\right),$$

and the claimed result follows. ■

We remark that to prove the bound, we *did not* actually need to use the fact that the labels all have the same probability; the calculation above shows that a balanced label distribution in fact maximizes the MI between $\boldsymbol{\mathcal{X}}_k$ and the observations $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ (though this will not be used later).

### B.4. Proof of Lemma 7

The proof relies on the celebrated I-MMSE relation of Guo, Shamai and Verdu (see Guo et al. (2005), Guo et al. (2013), and also the works Bennatan et al. (2008) and Bustin and Shamai (2013) that apply the I-MMSE framework for studying the MSE of estimating the transmitted codeword from the output of the AWGN channel).

Let $\ell \sim \mathrm{Unif}([k])$ and $\boldsymbol{Y}(s) = \boldsymbol{X}_\ell + \sqrt{s}\boldsymbol{Z}$. Denote $I(s) = I(\ell; \boldsymbol{Y}(s)|\boldsymbol{\mathcal{X}}_k) = I(\boldsymbol{X}_\ell; \boldsymbol{Y}(s)|\boldsymbol{\mathcal{X}}_k)$, where equality holds since, with probability one, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ are all distinct. Recall that our goal is to bound $H(\ell|\boldsymbol{Y}(\sigma^2), \boldsymbol{\mathcal{X}}_k) = H(\ell|\boldsymbol{\mathcal{X}}_k) - I(\sigma^2)$.

Clearly, for any $\sigma_0^2 < \sigma^2$,

$$H(\ell|\boldsymbol{Y}(\sigma^2), \boldsymbol{\mathcal{X}}_k) - H(\ell|\boldsymbol{Y}(\sigma_0^2), \boldsymbol{\mathcal{X}}_k) = I(\sigma_0^2) - I(\sigma^2) = -\int_{\sigma_0^2}^{\sigma^2} \frac{d}{ds} I(s) ds \,.$$

Using the I-MMSE relation, Lemma 44, applied pointwise conditioned on $\boldsymbol{\mathcal{X}}_k$,

$$\frac{d}{ds} I(s) = -\frac{1}{2s^2} \mathbb{E}\left[ \|\boldsymbol{X}_\ell - \mathbb{E}(\boldsymbol{X}_\ell|\boldsymbol{Y}(s), \boldsymbol{\mathcal{X}}_k)\|^2 \right] .$$

Since $\mathbb{E}(\boldsymbol{X}_\ell|\boldsymbol{Y}(s), \boldsymbol{\mathcal{X}}_k)$ is the minimum MSE estimator of $\boldsymbol{X}_\ell$ from $(\boldsymbol{Y}(s), \boldsymbol{\mathcal{X}}_k)$, it holds that for any $(\boldsymbol{Y}(s), \boldsymbol{\mathcal{X}}_k)$-measurable random variable $\hat{\boldsymbol{X}} = \hat{\boldsymbol{X}}(\boldsymbol{Y}(s), \boldsymbol{\mathcal{X}}_k)$, we have $-\frac{d}{ds} I(s) \le \frac{1}{2s^2} \mathbb{E}\|\boldsymbol{X}_\ell - \hat{\boldsymbol{X}}\|^2$.

Choose the optimal linear estimator (LMMSE) of $\boldsymbol{X}_\ell$ from $\boldsymbol{Y}(s)$, namely $\hat{\boldsymbol{X}} = \alpha(s)\boldsymbol{Y}(s)$, $\alpha(s) = \frac{1}{1+s}$, so that $\mathbb{E}\|\boldsymbol{X}_\ell - \hat{\boldsymbol{X}}\|^2 = \frac{s}{1+s} d$. One would think, at first sight, that this upper bound should be very loose: after all, the LMMSE is optimal for a Gaussian signal, whereas, conditioned on $\boldsymbol{\mathcal{X}}_k$, the distribution of $\boldsymbol{X}_\ell$ is very much non-Gaussian; it is not even continuous! Recall, however, that we are interested in applying Lemma 7 when the rate $\mathsf{R}_{d,k}$ is *above* the capacity $\mathsf{C}(\sigma^2)$; the key intuition is that when this is the case, the joint statistics of $(\boldsymbol{X}_\ell, \boldsymbol{Y} = \boldsymbol{X}_\ell + \sigma\boldsymbol{Z})$ with $\ell \sim \mathrm{Unif}([k])$, are in some sense "indistinguishable" from those of a joint Gaussian distribution $(\boldsymbol{W}, \boldsymbol{Y} = \boldsymbol{W} + \sigma\boldsymbol{Z})$, $\boldsymbol{W} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, corresponding to the capacity-achieving distribution of the Gaussian channel.

Continuing the calculation,

$$-\int_{\sigma_0^2}^{\sigma^2} \frac{d}{ds} I(s) ds \le \int_{\sigma_0^2}^{\sigma^2} \frac{1}{2s^2} \cdot \frac{s}{1+s} ds \cdot d = \int_{\sigma_0^2}^{\sigma^2} \left(-\mathsf{C}'(s)\right) ds \cdot d = \left(\mathsf{C}(\sigma_0^2) - \mathsf{C}(\sigma^2)\right) d \,,$$

where $\mathsf{C}(s) = \frac{1}{2}\log(1 + 1/s)$ is from (9) and $\mathsf{C}'(s)$ is its derivative. Combining,

$$H(\ell|\boldsymbol{Y}(\sigma^2), \boldsymbol{\mathcal{X}}_k) \le H(\ell|\boldsymbol{Y}(\sigma_0^2), \boldsymbol{\mathcal{X}}_k) - \mathsf{C}(\sigma^2)d + \mathsf{C}(\sigma_0^2)d \,. \tag{49}$$

Now, set $\sigma_0^2 = \mathsf{C}^{-1}\left((1+\delta)\mathsf{R}_{d,k}\right)$. To apply (49), we need to verify that $\sigma_0^2 < \sigma^2$. Applying the decreasing function $\mathsf{C}(\cdot)$, the condition is equivalent to $\mathsf{C}(\sigma^2) < \mathsf{C}(\sigma_0^2) = (1+\delta)\mathsf{R}_{d,k}$, which certainly hold since we assume $\mathsf{R}_{d,k} > \mathsf{C}(\sigma^2)$.

By definition, $\mathsf{C}(\sigma_0^2)d = (1+\delta)\log k$.

Define by $e(\delta)$ the error (averaged over the ensemble $\boldsymbol{\mathcal{X}}_k$) for decoding $\ell$ under $\mathsf{AWGN}(\sigma_0^2)$, using codebook $\boldsymbol{\mathcal{X}}_k$. In other words, it is the error of the MAP estimator for $\ell$ given $(\boldsymbol{Y}(\sigma_0^2), \boldsymbol{\mathcal{X}}_k)$. By Fano's inequality, Lemma 43, $H(\ell|\boldsymbol{Y}(\sigma_0^2), \boldsymbol{\mathcal{X}}_k) \le h_b\left(e(\delta)\right) + e(\delta)\log k$. Combined with (49), we obtain the bound claimed in Lemma 7.

■

### B.5. Proof of Lemma 8

Before getting to the computation, we emphasize that Lemma 7 may be invoked with *any* other upper bound on the ensemble average error $\rho_{\text{avg}}(\cdot)$, that could possibly be obtained through other means, e.g., by analyzing a different decoder than the one from Section A. There is much literature devoted to computing optimal error rates for both the Gaussian i.i.d. and the spherical code ensembles, primary in the regime of *positive rate*. In particular, for rates between the so-called *critical rate* and capacity the exact exponential decay rate is known: $\rho_{\text{avg}}(\sigma^2) = \exp(-E_{SP}^*(\mathsf{R},\sigma^2)d + o(d))$, where $E_{SP}^*(\mathsf{R},\sigma^2)$ is the *sphere-packing error exponent*. See, for example, Shannon (1959); Gallager (1968) for the exact expression. In the analysis that follows, we will need bounds on the error probability in the regime $\mathsf{C}(\sigma^2) - \mathsf{R} = o(1)$. In fact, for the zero rate regime, the capacity itself is $o(1)$, and sometimes it decays even as $o(d^{-1/2})$. In those cases, the sphere packing error exponent is of limited use.

Instead, we use the upper bounds on $\rho_{\text{avg}}(\cdot)$ derived in Section A.

As before, the analysis is divided between the positive ($\mathsf{R} > 0$) and zero ($\mathsf{R} = 0$) rate regimes.

#### B.5.1. Positive Rate

Let us work under the slightly more general regime, where $\mathsf{R}_{d,k}$ is either positive or decays slow enough with $d$, specifically, $\mathsf{R}_{d,k} = \frac{\log k}{d} = \omega(d^{-1/2})$ as $d \to \infty$.

We apply the bound (40) with noise variance

$$\sigma_0^2 = \mathsf{C}^{-1}((1+\delta)\mathsf{R}_{d,k})\,.$$

The second term of (40) is bounded By

$$(1+1/\sigma_0^2)^{1/2} e^{-d\left((1+\delta)\mathsf{R}_{d,k} - \mathsf{R}_{d,k} - \frac{1}{2}\log(\tau_2/\tau)\right)} = O(1)\cdot e^{-d(\delta\mathsf{R}_{d,k} - \frac{1}{2}\log(\tau_2/\tau))}\,.$$

Set $\tau_1 = \tau_2 = (1+\frac{1}{2}\mathsf{R}_{d,k}\delta)\tau$, so that $\log(\tau_2/\tau) \le \frac{1}{2}\mathsf{R}_{d,k}\delta$. Thus,

$$O(1)\cdot e^{-d(\delta\mathsf{R}_{d,k} - \frac{1}{2}\log(\tau_2/\tau))} = O(1)\cdot e^{-d(\delta\mathsf{R}_{d,k} - \frac{1}{4}\mathsf{R}_{d,k}\delta)} \lesssim e^{-C\mathsf{R}_{d,k}\delta d}\,,$$

for some $C > 0$. On the other hand, the first term of (40) is

$$e^{-\frac{1}{2}(1+\sigma_0^2)(\sqrt{\tau_1/\tau}-1)^2 d} \lesssim e^{-C\mathsf{R}_{d,k}^2\delta^2 d} = e^{-C\delta^2\frac{(\log k)^2}{d}}\,.$$

Note that since $\delta = o(1)$, this term is the most significant.

Denote $A = \frac{(\log k)^2}{d}$; recall that for $\mathsf{R}_{d,k} = \omega(d^{-1/2})$, $A = \omega(1)$.

In light of the estimates above, we need to choose $\delta = o(1)$ so to minimize $\delta + e^{-C\delta^2 A}$. Take

$$\delta = C_1\sqrt{\frac{\log A}{A}}\,,$$

for large enough constant $C_1$, which yields

$$\delta + e(\delta) \lesssim \delta + e^{-C\delta^2 A} \lesssim \sqrt{\frac{\log A}{A}}\,.$$

Plugging this into (26),

$$I(\boldsymbol{\mathcal{X}}_k;\boldsymbol{Y}) \lesssim \log k \sqrt{\frac{\log A}{A}}\,.$$

Using (27),

$$\begin{aligned}\frac{n_\varepsilon^*}{\sigma^2 k} &\geq C_1(\varepsilon,\mathsf{R})\cdot\frac{d}{\sigma^2}\cdot\left(I(\boldsymbol{\mathcal{X}}_k;\boldsymbol{Y})\right)^{-1}\\ &\gtrsim \frac{d}{\sigma^2\log k}\sqrt{\frac{A}{\log A}}\end{aligned}$$

where $A = \frac{(\log k)^2}{d}$. Let us understand the asymptotic of this bound as $d\to\infty$ and $\mathsf{R}_{d,k} = \mathsf{C}(\beta\sigma^2) = \frac{\log k}{d} \gg d^{-1/2}$. In that case, $\log A \approx \log\log k$, and so, the above reads

$$\frac{n_\varepsilon^*}{\sigma^2 k} \gtrsim \frac{d}{\sigma^2\log k}\sqrt{\frac{A}{\log A}} \gtrsim \frac{d}{\sigma^2\log k}\sqrt{\frac{\frac{(\log k)^2}{d}}{\log\log k}}\,.$$

Using $1/(\beta\sigma^2) \geq \mathsf{C}(\beta\sigma^2) = (\log k)/d$ (since $\mathsf{C}(s)\leq 1/(2s)$) finally yields

$$\frac{n_\varepsilon^*}{\sigma^2 k} \gtrsim \sqrt{\frac{\log k}{d}}\sqrt{\frac{\log k}{\log\log k}}\,. \tag{50}$$

Finally, note that in the positive rate regime, $\frac{\log k}{d} = \Omega(1)$.

■

### B.5.2. RATE ZERO ($\mathsf{R}=0$)

Assume that $\lim_{d\to\infty}\mathsf{R}_{d,k} = 0$ (including, possibly, $\mathsf{R}_{d,k}\gg d^{-1/2}$).

We would like to use the bound (39) with some $\eta = \eta_1 = \eta_2 = o(1)$ and $\sigma_0^2 = \sigma_0^2(\delta) = \mathsf{C}^{-1}((1+\delta)\mathsf{R}_{d,k})$, for $\delta = o(1)$.

We start with the condition (38), namely,

$$\begin{aligned}0 \leq F &:= 1-\eta-\sqrt{\frac{2\log k}{d}+\frac{\eta^2}{\sigma_0^2(\delta)}}-\sqrt{\frac{2\sigma_0^2(\delta)\log k}{d}}\\ &= 1-\eta-\sqrt{2\mathsf{R}_{d,k}}\sqrt{1+\frac{1}{2\sigma_0^2(\delta)\mathsf{R}_{d,k}}\eta^2}-\sqrt{2\sigma_0^2(\delta)\mathsf{R}_{d,k}}\,.\end{aligned}$$

(we replace $k-1$ with $k$, which yields a stronger condition.)

Use $\mathsf{C}(s) = \frac{1}{2}\log(1+1/s)\leq 1/(2s)$, therefore $\mathsf{C}^{-1}(y)\leq 1/(2y)$, and so $\sigma_0^2(\delta)\leq \frac{1}{2(1+\delta)\mathsf{R}_{d,k}}$:

$$\sqrt{2\sigma_0^2(\delta)\mathsf{R}_{d,k}} \leq \sqrt{\frac{1}{1+\delta}} = 1-\frac{\delta}{2}+O(\delta^2)\,.$$

Moreover,

$$\sqrt{1+\frac{1}{2\sigma_0^2(\delta)\mathsf{R}_{d,k}}\eta^2} \leq 1+\frac{1}{4\sigma_0^2(\delta)\mathsf{R}_{d,k}}\eta^2$$

($\sqrt{1+x} \le 1 + \frac{1}{2}x$ for all $x \ge 0$). Thus,

$$F \ge \delta/2 - O(\delta^2) - \eta - \sqrt{2\mathsf{R}_{d,k}} - 2\sqrt{\mathsf{R}_{d,k}} \cdot \frac{1}{4\sigma_0^2(\delta)\mathsf{R}_{d,k}}\eta^2 \,. \tag{51}$$

Note that $\sigma_0^2(\delta)\mathsf{R}_{d,k} = \Theta(1)$ for any $\delta = o(1)$; to see this, recall that $(1+\delta)\mathsf{R}_{d,k} = \mathsf{C}(\sigma_0^2)$ (by definition), with $\mathsf{C}(\sigma_0^2) = 1/(2\sigma_0^2) + O(1/\sigma_0^4)$ with $\sigma_0^2 \to \infty$. Consequently, the last term of (51) above is necessarily of lower order than either $\sqrt{\mathsf{R}_{d,k}}$ or $\eta$.

Introduce a constant parameter $\nu \in (0, 1/2)$, and set $\eta = (1/2 - \nu)\delta$. Observe that whenever

$$\delta \ge \frac{2\sqrt{2}}{\nu}\sqrt{\mathsf{R}_{d,k}} = \frac{2\sqrt{2}}{\nu}\sqrt{\frac{\log k}{d}}, \quad \delta = o(1)\,, \tag{52}$$

plugging into (51), we have $F \ge \frac{\nu}{2}\delta(1 - o(1)) > 0$.

Let us estimate the terms in (39). The first term is $e^{-\frac{d}{2}F^2} \le e^{-C_1 d\nu^2\delta^2}$. The second term is $2e^{-\frac{\eta^2}{2}\cdot\frac{d}{\sigma_0^2}}$. Using

$$\frac{\eta^2 d}{2\sigma_0^2} \ge \eta^2 d(1+\delta)\mathsf{R}_{d,k} = \eta^2(1+\delta)\log k \ge \eta^2 \log k\,,$$

(we used $\sigma_0^2 \le \frac{1}{2(1+\delta)\mathsf{R}_{d,k}}$), we deduce that the second term is $\le 2e^{-(1/2-\nu)^2\delta^2 \log k}$. Since in the zero rate regime, $d \gg \log k$, we see that the first term is always negligible compared to the second, regardless of how fast $\delta$ decays. Thus, we would like choose $\delta = o(1)$ so to minimize (the asymptotic decay rate of)

$$\delta + e(\delta) \lesssim \underbrace{\delta}_{e_1(\delta)} + \underbrace{e^{-(1/2-\nu)^2\delta^2 \log k}}_{e_2(\delta)}\,. \tag{53}$$

Note that $e_1(\delta)$ is increasing in $\delta$, while $e_2(\delta)$ is decreasing. Denote

$$\delta_1 = \frac{\sqrt{2}}{(1/2-\nu)}\sqrt{\frac{\log\log k}{\log k}}, \quad \delta_2 = \frac{2\sqrt{2}}{\nu}\sqrt{\frac{\log k}{d}}\,, \tag{54}$$

so that $\delta_2$ is the smallest number $\delta$ that satisfies (52).

One may readily verify that $\delta = \delta_1$ optimally balances between $e_1(\delta), e_2(\delta)$, in the sense of asymptotic growth:

$$e_1(\delta_1) \asymp e_2(\delta_1) \asymp \sqrt{\frac{\log\log k}{\log k}}\,, \quad \implies e_1(\delta_1) + e_2(\delta_1) \lesssim \sqrt{\frac{\log\log k}{\log k}}$$

Recall, however, that not all assignments $\delta$ are applicable; we must satisfy the constraint (52), $\delta \ge \delta_2$. If $\delta_2 \le \delta_1$ then there is no problem; on the other hand, if $\delta_2 > \delta_1$, assigning $\delta = \delta_2$,

$$e_1(\delta_2) + e_2(\delta_2) \overset{(i)}{\le} e_1(\delta_2) + e_2(\delta_1) \overset{(ii)}{\lesssim} e_1(\delta_2) + e_1(\delta_1) \overset{(iii)}{\le} 2e_1(\delta_2) \lesssim \sqrt{\frac{\log k}{d}}\,,$$

where we used that: (i) $e_2(\cdot)$ is decreasing; (ii) $e_1(\delta_1) \asymp e_2(\delta_1)$; (iii) $e_1(\cdot)$ is increasing.

Concluding the calculation, using (26), we have

$$I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y}) \lesssim \log k \cdot \max\left\{\sqrt{\frac{\log\log k}{\log k}}, \sqrt{\frac{\log k}{d}}\right\}. \tag{55}$$

Finally, to deduce the lower bound on the sample complexity, use (27):

$$\begin{aligned}\frac{n_\varepsilon^*}{\sigma^2 k} &\geq C_1(\varepsilon) \cdot \frac{d}{\sigma^2} \cdot (I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y}))^{-1} \\ &\geq C_1(\varepsilon) \cdot 2\beta \log k \cdot (I(\boldsymbol{\mathcal{X}}_k; \boldsymbol{Y}))^{-1} \\ &\geq C_2(\varepsilon)\beta \min\left\{\sqrt{\frac{\log k}{\log\log k}}, \sqrt{\frac{d}{\log k}}\right\}. \end{aligned} \tag{56}$$

■

## Appendix C. Proofs for Section 4.1

### C.1. Proof of Lemma 9

As in Section A, we give different constructions between the zero rate ($\mathsf{R} = 0$) and positive rate ($\mathsf{R} > 0$) regimes. The construction for the local test is guided by the form of the capacity-achieving decoder from Section A.

#### C.1.1. Rate Zero ($\mathsf{R} = 0$)

Following the form of the decoder analyzed in Section A.1, we consider a test of the form

$$\mathsf{Test}(\hat{\boldsymbol{X}}, \boldsymbol{Y}) = \mathbb{1}\{d^{-1}\langle \boldsymbol{Y}, \hat{\boldsymbol{X}}\rangle \geq 1 - \eta\}\,, \tag{57}$$

where the choice of $\eta$ will be specified below.

Suppose that $\hat{\boldsymbol{X}} \in \sqrt{d}\mathcal{S}^{d-1}$ is such that, for some particular $i \in [k]$, $d^{-1}\|\hat{\boldsymbol{X}} - \boldsymbol{X}_i\|^2 \leq 0.5\varepsilon_{\mathrm{I}}$. Note that this may be written equivalently as $d^{-1}\langle \boldsymbol{X}_i, \hat{\boldsymbol{X}}\rangle \geq 1 - 0.25\varepsilon_{\mathrm{I}}$. Thus,

$$d^{-1}\langle \boldsymbol{X}_i + \sigma\boldsymbol{Z}, \hat{\boldsymbol{X}}\rangle \geq 1 - 0.25\varepsilon_{\mathrm{I}} + \underbrace{(d^{-1/2}\sigma)\langle \boldsymbol{Z}, d^{-1/2}\hat{\boldsymbol{X}}\rangle}_{\sim\mathcal{N}(0,\sigma^2/d)}\,.$$

Setting

$$\eta = 0.25\varepsilon_{\mathrm{I}}\,, \tag{58}$$

we get

$$\begin{aligned}\Pr(\mathsf{Test}(\hat{\boldsymbol{X}}, \boldsymbol{Y}) = 1) &\geq \frac{1}{k}\Pr(\mathsf{Test}(\hat{\boldsymbol{X}}, \boldsymbol{Y}) = 1 \,|\, \ell = i) \\ &\geq \frac{1}{k}\Pr\left(\mathcal{N}(1 - 0.25\varepsilon_{\mathrm{I}}, \sigma^2/d) \geq 1 - 0.25\varepsilon_{\mathrm{I}}\right) = 0.5/k\,.\end{aligned}$$

Consequently, with probability 1, $\mathsf{q}_{\mathrm{Close}}(\boldsymbol{\mathcal{X}}_k) \geq 0.5k^{-1}$.

The challenging part of the analysis is to control $\mathsf{q}_{\mathrm{Far}}(\boldsymbol{\mathcal{X}}_k)$.

Observe that if $d^{-1}\|\hat{\boldsymbol{X}} - \boldsymbol{X}_i\|^2 \geq \varepsilon_{\mathrm{I}}$ then $d^{-1}\langle \boldsymbol{X}_i, \hat{\boldsymbol{X}}\rangle \leq 1 - 0.5\varepsilon_{\mathrm{I}}$. For $\hat{\boldsymbol{X}} \in \sqrt{d}\mathcal{S}^{d-1}$, denote

$$\begin{aligned}Q_i(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k) &= \Pr_{\boldsymbol{Z}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left(\mathsf{Test}(\hat{\boldsymbol{X}}, \boldsymbol{X}_i + \sigma\boldsymbol{Z}) = 1 \,\middle|\, \boldsymbol{\mathcal{X}}_k\right)\,, \\ \bar{Q}(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k) &= \frac{1}{k}\sum_{i=1}^{k} Q_i(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k)\,.\end{aligned} \tag{59}$$

Note that $Q_i(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k)$ depends on $\boldsymbol{\mathcal{X}}_k$ only through $\boldsymbol{X}_i$.

By definition, $\mathsf{q}_{\mathrm{Far}}(\boldsymbol{\mathcal{X}}_k) = \max_{\hat{\boldsymbol{X}}\in\mathcal{T}\cap\mathcal{H}_{\mathrm{Far}}} \bar{Q}(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k)$. We start with a trivial bound.

**Lemma 19** *Suppose that $d^{-1}\langle \boldsymbol{X}_i, \hat{\boldsymbol{X}}\rangle \leq 1 - \nu - 0.25\varepsilon_{\mathrm{I}}$ for $0 \leq \nu \leq 1 - 0.25\varepsilon_{\mathrm{I}}$. Then*

$$Q_i(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k) \leq k^{-\beta\nu^2}\,.$$

*Consequently, if $\hat{\boldsymbol{X}} \in \mathcal{H}_{\mathrm{Far}}$ then for all $i$, $Q_i(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k) \leq k^{-\frac{\beta}{16}\varepsilon_{\mathrm{I}}^2}$.*

**Proof** $d^{-1}\langle \boldsymbol{X}_i + \sigma \boldsymbol{Z}, \hat{\boldsymbol{X}}\rangle \le 1 - 0.25\varepsilon_{\mathrm{I}} - \nu + \underbrace{d^{-1}\sigma\langle \boldsymbol{Z}, \hat{\boldsymbol{X}}\rangle}_{\sim\mathcal{N}(0,\sigma^2/d)}$. Thus,

$$\Pr(\mathsf{Test}(\hat{\boldsymbol{X}}, \boldsymbol{X}_i + \sigma\boldsymbol{Z}) = 1) \le \Pr\left(\mathcal{N}(0,\sigma^2/d) \ge \nu\right) \le e^{-\frac{d}{2\sigma^2}\nu^2}.$$

Now, $k = e^{d\mathsf{C}(\beta\sigma^2)} \le e^{\frac{d}{2\beta\sigma^2}}$ (since $\mathsf{C}(s) = \frac{1}{2}\log(1+1/s) \le 1/(2s)$), therefore $e^{-\frac{d}{2\sigma^2}\nu^2} \le k^{-\beta\nu^2}$. Finally, if $\hat{\boldsymbol{X}} \in \mathcal{H}_{\mathrm{Far}}$, then $d^{-1}\langle \boldsymbol{X}_i, \hat{\boldsymbol{X}}\rangle \le 1 - \nu - 0.25\varepsilon_{\mathrm{I}}$ with $\nu = 0.25\varepsilon_{\mathrm{I}}$. ■

As mentioned, Lemma 19 gives us the trivial bound $\bar{Q}(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k) \le k^{-\frac{\beta}{16}\varepsilon_{\mathrm{I}}^2}$ for all $\hat{\boldsymbol{X}} \in \mathcal{H}_{\mathrm{Far}}$. This is a highly wasteful bound: it treats $\hat{\boldsymbol{X}}$ as if it is *simultaneously* $\sqrt{\varepsilon_{\mathrm{I}} d}$-close to all of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$. In practice, however, "typical" instances of $\boldsymbol{\mathcal{X}}_k$ create constellations that do not cluster around any particular point; consequently, for most $i \in [k]$, it has to be that, in fact, $d^{-1}\langle \boldsymbol{X}_i, \hat{\boldsymbol{X}}\rangle \approx 0$.

For $t \in (0,1)$, set

$$N_t(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k) = \sum_{i=1}^{k} \mathbb{1}\{d^{-1}\langle \boldsymbol{X}_i, \hat{\boldsymbol{X}}\rangle \ge t\}, \tag{60}$$

the number of centers $\boldsymbol{X}_i$ that have correlation $\ge t$ with $\hat{\boldsymbol{X}}$.

Choose some constants $\varepsilon_0, \nu_0 \in (0,1)$ such that $\nu_0 < 1 - 0.25\varepsilon_0$ and $\beta\nu_0^2 > 1$. This can certainly be done, since $\beta > 1$. By Lemma 19 above, for every $\hat{\boldsymbol{X}} \in \mathcal{H}_{\mathrm{Far}} \cap \mathcal{T}$, assuming $\varepsilon_{\mathrm{I}} \le \varepsilon_0$,

$$\bar{Q}(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k) \le k^{\beta\nu_0^2} + \frac{1}{k}\cdot N_{(1-\nu_0-0.25\varepsilon_0)}(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k)\cdot k^{-\frac{\beta}{16}\varepsilon_{\mathrm{I}}^2}. \tag{61}$$

That is, $\boldsymbol{X}_i$-s whose correlation with $\hat{\boldsymbol{X}}$ is $< 1 - \nu_0 - 0.25\varepsilon_0 < 1 - \nu_0 - 0.25\varepsilon_{\mathrm{I}}$ contribute each at most $Q_i(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k) \le k^{-\beta\nu_0^2}$; on the other hand, centers whose correlations is higher give, at most, the worst-case contribution $Q_i(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k) = k^{-\frac{\beta}{16}\varepsilon_{\mathrm{I}}^2}$. In light of (60), clearly,

$$\mathsf{q}_{\mathrm{Far}}(\boldsymbol{\mathcal{X}}_k) \le k^{\beta\nu_0^2} + \max_{\hat{\boldsymbol{X}}\in\mathcal{T}\cap\mathcal{H}_{\mathrm{Far}}} N_{(1-\nu_0-0.25\varepsilon_0)}(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k)\cdot k^{-1-\frac{\beta}{16}\varepsilon_{\mathrm{I}}^2} \tag{62}$$

Thus, it remains to show that, with high probability, $\max_{\hat{\boldsymbol{X}}\in\mathcal{T}\cap\mathcal{H}_{\mathrm{Far}}} N_{(1-\nu_0-0.25\varepsilon_0)}(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k)$ is small.

**Lemma 20** *Fix any $\hat{\boldsymbol{X}} \in \mathcal{T}$. There are universal $C_1, C_2$ such that whenever $t \ge C_1\sqrt{\frac{\log k}{d}}$, for all $M \ge 1$,*

$$\Pr\left(N_t(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k) \ge M \,\middle|\, \hat{\boldsymbol{X}} \in \mathcal{H}_{\mathrm{Far}}\right) \le (C_2 k e^{-d\frac{t^2}{2}})^M,$$

*where the probability is with respect to $\boldsymbol{\mathcal{X}}_k \sim \mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1})^{\otimes k}$, and conditioned on the event that $\hat{\boldsymbol{X}} \in \mathcal{H}_{\mathrm{Far}}$.*

**Proof** Observe that conditioned on the event $\hat{\boldsymbol{X}} \in \mathcal{H}_{\mathrm{Far}}$, the centers $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ are i.i.d. and $\sim \mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1} \setminus \mathcal{B}(\hat{\boldsymbol{X}}, \sqrt{\varepsilon_{\mathrm{I}} d}))$. For any *non-negative* $f(\cdot)$,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X}_i\sim\mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1}\setminus\mathcal{B}(\hat{\boldsymbol{X}},\sqrt{\varepsilon_{\mathrm{I}}d}))}[f(\boldsymbol{X}_i)] &\le \frac{\mathrm{Surf}(\sqrt{d}\mathcal{S}^{d-1})}{\mathrm{Surf}(\sqrt{d}\mathcal{S}^{d-1}\setminus\mathcal{B}(\hat{\boldsymbol{X}},\sqrt{\varepsilon_{\mathrm{I}}d}))}\cdot\mathbb{E}_{\boldsymbol{X}_i\sim\mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1})}[f(\boldsymbol{X}_i)] \\
&= \frac{1}{1-\varepsilon_{\mathrm{I}}^{d-1}}\cdot\mathbb{E}_{\boldsymbol{X}_i\sim\mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1})}[f(\boldsymbol{X}_i)] \\
&= (1+o(1))\mathbb{E}_{\boldsymbol{X}_i\sim\mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1})}[f(\boldsymbol{X}_i)].
\end{aligned} \tag{63}$$

Consequently, $N_t(\hat{\boldsymbol{X}}, \boldsymbol{\mathcal{X}}_k) \sim \mathrm{Binomial}(n, \mathsf{p})$, with

$$\begin{aligned}\mathsf{p} &= \mathbb{E}_{\boldsymbol{X}_i \sim \mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1} \setminus \mathcal{B}(\hat{\boldsymbol{X}}, \sqrt{\varepsilon_{\mathrm{I}} d}))} \left[ \mathbb{1}\{ d^{-1} \langle \boldsymbol{X}_i, \hat{\boldsymbol{X}} \rangle \geq t \} \right] \\ &= (1 + o(1)) \Pr_{\boldsymbol{X}_i \sim \mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1})} \left( d^{-1} \langle \boldsymbol{X}_i, \hat{\boldsymbol{X}} \rangle \geq t \right) \\ &\leq 2 e^{-d\frac{t^2}{2}} ,\end{aligned}$$

where we used the standard tail bound Lemma 35. We have

$$\Pr\left( X_t(\hat{\boldsymbol{X}} | \boldsymbol{\mathcal{X}}_k) \geq M \,\middle|\, \hat{\boldsymbol{X}} \in \mathcal{H}_{\mathrm{Far}} \right) = \sum_{l=M}^{k} \binom{k}{l} \mathsf{p}^l (1-\mathsf{p})^{k-l} \leq \sum_{l=M}^{k} \left( \frac{ke}{l} \mathsf{p} \right)^l .$$

Assuming $t \geq C\sqrt{\frac{\log k}{d}}$ for large enough (universal) $C > 0$, $\frac{ke}{l}\mathsf{p} \leq \frac{1}{2}$,and so $\sum_{l=M}^{k} \left(\frac{ke}{l}\mathsf{p}\right)^l \leq 2\left(ke\mathsf{p}\right)^M$. ■

**Lemma 21** *There are $C_1, C_2, C_3$ universal such that the following holds.*

*Suppose that $t \geq C_1\sqrt{\frac{\log k}{d}}$, $\varepsilon_{\mathrm{I}} < 1/2$. Then with probability at least $1 - e^{-d}$ over $\boldsymbol{\mathcal{X}}_k \sim \mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1})^{\otimes k}$,*

$$\max_{\hat{\boldsymbol{X}} \in \mathcal{T} \cap \mathcal{H}_{\mathrm{Far}}} N_t(\hat{\boldsymbol{X}} | \boldsymbol{\mathcal{X}}_k) \leq C_2 \frac{\log(1/\varepsilon_{\mathrm{I}})}{\frac{1}{2}t^2 - \frac{\log k + C_3}{d}} =: M_0 .$$

**Proof** Let $B$ be the number of candidates $\hat{\boldsymbol{X}} \in \mathcal{T}$ such that $N_t(\hat{\boldsymbol{X}} | \boldsymbol{\mathcal{X}}_k) > M_0$. Our goal is to show that w.p. $\geq 1 - e^{-d}$, $B = 0$. By Markov's inequality and Lemma 20, assuming $t \gtrsim \sqrt{\frac{\log k}{d}}$,

$$\Pr(B \geq 1) \leq \mathbb{E}[B] = (Cke^{-d\frac{t^2}{2}})^{M_0} |\mathcal{T}| \leq (Cke^{-d\frac{t^2}{2}})^{M_0} e^{Cd\log(1/\varepsilon_{\mathrm{I}})} .$$

Taking $M_0 = C_2 \frac{\log(1/\varepsilon_{\mathrm{I}})}{\frac{1}{2}t^2 - \frac{\log k + C_3}{d}}$ for large enough $C_2, C_3$, the above probability is $\leq e^{-d}$. ■

**Proof** (Of Lemma 9, $\mathsf{R} = 0$.) Recall that, by construction, $\mathsf{q}_{\mathrm{Close}}(\boldsymbol{\mathcal{X}}_k) \geq 0.5k^{-1}$ holds with probability 1.

Recall (62), and apply Lemma 21 with $t = 1 - \nu_0 - 0.25\varepsilon_0$, which is a positive constant. Consequently, with probability $\geq 1 - e^{-d}$,

$$\mathsf{q}_{\mathrm{Far}}(\boldsymbol{\mathcal{X}}_k) \leq k^{-\beta\nu_0^2} + C\log(1/\varepsilon_{\mathrm{I}}) k^{-1-\frac{\beta}{16}\varepsilon_{\mathrm{I}}^2} .$$

Since $\beta\nu_0^2 > 1$, one may indeed choose some $c = c(\beta)$ small enough such that $\mathsf{q}_{\mathrm{Far}}(\boldsymbol{\mathcal{X}}_k) \leq 2k^{-1-c\varepsilon_{\mathrm{I}}^2}$ holds with probability $1 - o_{\beta,\varepsilon_{\mathrm{I}}}(1)$.

■

### C.1.2. POSITIVE RATE ($\mathsf{R} > 0$)

Moving on to the positive rate regime, our construction is guided by the decoder of Section A.2.

Set $\alpha = \frac{1}{1+\sigma^2}, \tau = \frac{\sigma^2}{1+\sigma^2}$.

For fixed $\hat{\boldsymbol{X}}, \boldsymbol{X}_i \in \sqrt{d}\mathcal{S}^{d-1}$, denoting $\boldsymbol{Y} = \boldsymbol{X}_i + \sigma \boldsymbol{Z}$, $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{Z}}\|\alpha \boldsymbol{Y} - \hat{\boldsymbol{X}}\|^2 &= \mathbb{E}_{\boldsymbol{Z}}\|\alpha \boldsymbol{Y} - \boldsymbol{X}_i\|^2 + \|\boldsymbol{X}_i - \hat{\boldsymbol{X}}\|^2 + 2\mathbb{E}_Z\langle \alpha \boldsymbol{Y} - \boldsymbol{X}_i, \boldsymbol{X}_i - \hat{\boldsymbol{X}}\rangle \\ &= \tau d + \|\boldsymbol{X}_i - \hat{\boldsymbol{X}}\|^2 + 2(\alpha - 1)\langle \boldsymbol{X}_i, \boldsymbol{X}_i - \hat{\boldsymbol{X}}\rangle \\ &= \tau d + \alpha \|\boldsymbol{X}_i - \hat{\boldsymbol{X}}\|^2 , \end{aligned} \tag{64}$$

where we used $\|\boldsymbol{X}_i\|^2 = \|\hat{\boldsymbol{X}}\|^2 = d$ and $-\langle \boldsymbol{X}_i, \hat{\boldsymbol{X}}\rangle = \frac{1}{2}(\|\boldsymbol{X}_i - \hat{\boldsymbol{X}}\|^2 - \|\boldsymbol{X}_i\|^2 - \|\hat{\boldsymbol{X}}\|^2)$.

Assume that $\|\boldsymbol{X}_i - \hat{\boldsymbol{X}}\|^2 \le 0.5\varepsilon_{\mathrm{I}} d$. By the Gaussian Lipschitz concentration inequality (Lemma 34), applied for $F(\boldsymbol{Z}) = d^{-1/2}\|\alpha(\boldsymbol{X}_i + \sigma \boldsymbol{Z}) - \hat{\boldsymbol{X}}\|$, which is $(d^{-1/2}\alpha\sigma)$-Lipschitz with expectation $\mathbb{E}_{\boldsymbol{Z}} F(\boldsymbol{Z}) \le \sqrt{\mathbb{E}(F(\boldsymbol{Z}))^2} \le \sqrt{\tau + 0.5\alpha\varepsilon_0}$,

$$\Pr_{\boldsymbol{Z}}\left(d^{-1/2}\|\alpha \boldsymbol{Y}_i - \hat{\boldsymbol{X}}\| \ge \sqrt{\tau + 0.5\alpha\varepsilon_{\mathrm{I}}} + \eta\right) \le e^{-\frac{1}{2}\frac{\eta^2}{\alpha^2\sigma^2}d} .$$

Consider the test

$$\mathsf{Test}(\hat{\boldsymbol{X}}, \boldsymbol{Y}) = \mathbb{1}\{d^{-1/2}\|\alpha \boldsymbol{Y} - \hat{\boldsymbol{X}}\| \le \sqrt{\tau + 0.5\alpha\varepsilon_I} + \eta\}, \quad \eta = \sqrt{\frac{2\alpha^2\sigma^2 \log 2}{d}} = O_{\mathsf{R},\beta}(d^{-1/2}) , \tag{65}$$

so that by construction, $\mathsf{q}_{\mathrm{Close}}(\boldsymbol{\mathcal{X}}_k) \ge 0.5k^{-1}$ holds with probability 1.

It remains to bound $\mathsf{q}_{\mathrm{Far}}(\boldsymbol{\mathcal{X}}_k)$ with high probability. We follow the notation (59), where $\mathsf{Test}(\hat{\boldsymbol{X}}, \boldsymbol{Y})$ that appears in (59) is now defined by (65). Our goal is to bound, with high probability over $\boldsymbol{\mathcal{X}}_k$,

$$\mathsf{q}_{\mathrm{Far}}(\boldsymbol{\mathcal{X}}_k) = \max_{\hat{\boldsymbol{X}} \in \mathcal{T} \cap \mathcal{H}_{\mathrm{Far}}} \bar{Q}_i(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k) = \max_{\hat{\boldsymbol{X}} \in \mathcal{T} \cap \mathcal{H}_{\mathrm{Far}}} \frac{1}{k}\sum_{i=1}^{k} Q_i(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k) .$$

As was in the zero rate case, for any fixed $\hat{\boldsymbol{X}}$, conditioned on the event $\{\hat{\boldsymbol{X}} \in \mathcal{H}_{\mathrm{Far}}\}$, the centers $\boldsymbol{X}_1, \dots, \boldsymbol{X}_k$ are i.i.d. and $\boldsymbol{X}_i \sim \mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1} \setminus \mathcal{B}(\hat{\boldsymbol{X}}, \sqrt{\varepsilon_{\mathrm{I}} d}))$. Consequently, $\bar{Q}(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k)$ is the average of $k$ i.i.d. random variables. We shall show that its expectation is very small, specifically $\mathbb{E}[\bar{Q}(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k)] \le k^{-1-c}$; moreover, we shall show that it concentrates tightly about this expectation, to the extent that the maximum over the net, $\max_{\hat{\boldsymbol{X}} \in \mathcal{T} \cap \mathcal{H}_{\mathrm{Far}}} \bar{Q}_i(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k)$, can be controlled as well. To do this, we use Bernstein's inequality for sums of i.i.d. bounded random variables, Lemma 36.

For brevity, we introduce some notation. For $\hat{\boldsymbol{X}} \in \mathcal{T}$ fixed, denote by $\mathcal{E} = \mathcal{E}(\hat{\boldsymbol{X}})$ the event $\mathcal{E} = \{\hat{\boldsymbol{X}} \in \mathcal{H}_{\mathrm{Far}}\}$ (with respect to the probability on $\boldsymbol{\mathcal{X}}_k \sim \mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1})^{\otimes k}$). Denote by $\mathbb{E}^{\mathcal{E}}[\cdot], \|\cdot\|_\infty^{\mathcal{E}}$ respectively the expectation and $L_\infty$ norm with respect to the conditional measure on $\boldsymbol{\mathcal{X}}_k$; and $\mathbb{P}^{\mathcal{E}}(S) := \mathbb{E}^{\mathcal{E}}[\mathbb{1}_S]$.

To use Bernstein's inequality, we need two components: an $L_\infty$ bound and a bound on the expectation.

We start with the $L_\infty$ bound:

**Lemma 22** *There are $C, \varepsilon_0$, that depend on $\mathsf{R}, \beta$, such that the following holds.*

*For any $\hat{\boldsymbol{X}} \in \mathcal{T}$, whenever $\varepsilon_{\mathrm{I}} < \varepsilon_0$ and $d$ is sufficiently large, $\varepsilon_{\mathrm{I}} = \Omega_{\mathsf{R},\beta}(d^{-1/2})$, then*

$$\|Q_i(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k)\|_\infty^{\mathcal{E}} \le 2k^{-C\varepsilon_I^2} . \tag{66}$$

**Proof** Fix any $\hat{\boldsymbol{X}}, \boldsymbol{\mathcal{X}}_k$ such that $\hat{\boldsymbol{X}} \in \mathcal{H}_{\mathrm{Far}}$. Let $\mu = d^{-1/2}\|\alpha \boldsymbol{X}_i - \hat{\boldsymbol{X}}\| \le 2$. By the rotational invariance of $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$,[3]

$$\|\alpha(\boldsymbol{X}_i + \sigma \boldsymbol{Z}) - \hat{\boldsymbol{X}}\|^2 \stackrel{d}{=} \|\alpha\sigma \boldsymbol{Z} + \mu \boldsymbol{1}\|^2 = \sum_{j=1}^{d} (\alpha\sigma Z_j + \mu)^2 \,,$$

where $\boldsymbol{1} = (1, \dots, 1)$ and $\stackrel{d}{=}$ denotes equality in distribution. The expression above is a sum of i.i.d. sub-Exponential random variables. The sub-Exponential norm, denoted $\|\cdot\|_{\psi_1}$, is upper bounded by

$$\|(\alpha\sigma Z_j + \mu)^2\|_{\psi_1} \lesssim \alpha^2\sigma^2 + \mu^2 = O_{\beta,\mathsf{R}}(1) \,.$$

For background on sub-Exponential random variables, including the definition of the sub-Exponential norm (and Orlicz norms in general), we refer to the book (Vershynin, 2018, Chapter 2). Recall by (64) that

$$d^{-1}\mathbb{E}\|\alpha\sigma \boldsymbol{Z} + \mu \boldsymbol{1}\|^2 = \tau + \alpha d^{-1}\|\boldsymbol{X}_i - \hat{\boldsymbol{X}}\|^2 \ge \tau + \alpha\varepsilon_{\mathrm{I}} \,,$$

and set

$$\begin{aligned} t &= \tau + \alpha\varepsilon_{\mathrm{I}} - \left(\sqrt{\tau + 0.5\alpha\varepsilon_I} + \eta\right)^2 \\ &= \tau + \alpha\varepsilon_{\mathrm{I}} - \left(\sqrt{\tau + 0.5\alpha\varepsilon_I} + O_{\mathsf{R},\beta}(d^{-1/2})\right)^2 \\ &= 0.5\alpha\varepsilon_{\mathrm{I}} + O_{\mathsf{R},\beta}(d^{-1/2}) \,. \end{aligned}$$

Using Bernstein's inequality for sub-Exponential random variables, Lemma 37,

$$\begin{aligned} Q_i(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k) &= \Pr_{\boldsymbol{Z}}(\mathsf{Test}(\hat{\boldsymbol{X}}, \boldsymbol{X}_i + \sigma\boldsymbol{Z}) = 1) \\ &= \Pr\left(d^{-1}\|\alpha(\boldsymbol{X}_i + \sigma\boldsymbol{Z}) - \hat{\boldsymbol{X}}\|^2 \le \left(\sqrt{\tau + 0.5\alpha\varepsilon_I} + \eta\right)^2\right) \\ &\le \Pr\left(d^{-1}\|\alpha(\boldsymbol{X}_i + \sigma\boldsymbol{Z}) - \hat{\boldsymbol{X}}\|^2 - \mathbb{E}_{\boldsymbol{Z}} d^{-1}\mathbb{E}_Z\|\alpha(\boldsymbol{X}_i + \sigma\boldsymbol{Z}) - \hat{\boldsymbol{X}}\|^2 \le -t\right) \\ &\le 2\exp\left(-c\min\left\{\frac{(d\cdot t)^2}{d\cdot\|\alpha\sigma Z_j + \mu\|_{\psi_1}^2}, \frac{d\cdot t}{\|\alpha\sigma Z_j + \mu\|_{\psi_1}}\right\}\right) \\ &\stackrel{(\star)}{\le} 2\exp\left(-C(\mathsf{R},\beta)\min\{\varepsilon_{\mathrm{I}}, \varepsilon_{\mathrm{I}}^2\}\cdot d\right) \\ &= 2\exp\left(-C(\mathsf{R},\beta)\varepsilon_{\mathrm{I}}^2\cdot d\right) \end{aligned}$$

where to get $(\star)$, we used $\|\alpha\sigma Z_j + \mu\|_{\psi_1} = O_{\beta,\mathsf{R}}(1)$ and $\varepsilon_{\mathrm{I}} \gtrsim d^{-1/2}$. Finally, to deduce (66), recall that $k = e^{\mathsf{R}\cdot d}$. ■

Moving on to the expectation:

**Lemma 23** *For any $\varepsilon_{\mathrm{I}} < \varepsilon_0$ sufficiently small and $d$ sufficiently large such that $\varepsilon_{\mathrm{I}} = \Omega_{\mathsf{R},\beta}(d^{-1/2})$,*

$$\mathbb{E}^{\mathcal{E}}[Q_i(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k)] \le 2k^{-1-c} \,,$$

*where $c, \varepsilon_0$ depend on $\mathsf{R}, \beta$.*

3. An alternative method to the one below (which is itself very standard) is to use deviation inequalities for non-central $\chi^2$ random variables, that are readily available in the literature, though somewhat more "messy".

**Proof** Observe that the test, defined in (65), is an orthogonally invariant function of its argument; that is, $\mathsf{Test}(\boldsymbol{x},\boldsymbol{y}) = \mathsf{Test}(R\boldsymbol{x},R\boldsymbol{y})$ for any $R \in O(d)$, where $O(d)$ is the group of $d\times d$ orthogonal matrices. Introduce an independent $R \sim \mathrm{Haar}(O(d))$, and note that, by the orthogonal invariance of $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I})$, $R(\boldsymbol{X}_i + \sigma\boldsymbol{Z}) \overset{d}{=} R\boldsymbol{X}_i + \sigma\boldsymbol{Z}$.

Now, conditioned on $R\boldsymbol{X}_i$, the conditional distribution of $R\hat{\boldsymbol{X}}$ is $\sim \mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1}\backslash\mathcal{B}(R\boldsymbol{X}_i,\sqrt{\varepsilon_{\mathrm{I}}d}))$. Thus, again owing to orthogonal invariance, we have

$$\begin{aligned}\mathbb{E}^{\mathcal{E}}[Q_i(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k)] &= \Pr_{\substack{\boldsymbol{W}\sim\mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1}\backslash\mathcal{B}(\boldsymbol{x},\sqrt{\varepsilon_{\mathrm{I}}d})),\\ \boldsymbol{Z}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}} \left(\|\alpha(\boldsymbol{x}+\sigma\boldsymbol{Z})-\boldsymbol{W}\| \le \sqrt{\tau+0.5\varepsilon_I}+\eta\right)\\ &\overset{(\star)}{\le} \frac{1}{1-\varepsilon_{\mathrm{I}}^{d-1}} \Pr_{\substack{\boldsymbol{W}\sim\mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1}),\\ \boldsymbol{Z}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}} \Pr\left(\|\alpha(\boldsymbol{x}+\sigma\boldsymbol{Z})-\boldsymbol{W}\| \le \sqrt{\tau+0.5\varepsilon_I}+\eta\right),\end{aligned}$$

where $\boldsymbol{x} \in \sqrt{d}\mathcal{S}^{d-1}$ is any fixed vector, and $(\star)$ follows from (63). The probability above has been bounded in Lemma 17, which yields

$$\le (1+o(1))\left(1+\frac{1}{\sigma^2}\right)^{1/2} e^{-\left(\mathsf{C}(\sigma^2)-\frac{1}{2}\log\frac{(\sqrt{\tau+0.5\varepsilon_I}+\eta)^2}{\tau}\right)d}.$$

Since $k = e^{\mathsf{C}(\beta\sigma^2)d}$ for $\beta > 1$ constant (hence $\mathsf{C}(\sigma^2)-\mathsf{C}(\beta\sigma^2)$ is a positive constant), and assuming $\varepsilon_{\mathrm{I}} \le \varepsilon_0$ is small enough, we get that $\mathbb{E}^{\mathcal{E}}[Q_i(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k)] \le 2k^{-1-c}$ for some $c = c(\beta,\mathsf{R})$, for $d$ large enough. ∎

We are ready to bound $\mathsf{q}_{\mathrm{Far}}(\boldsymbol{\mathcal{X}}_k)$:

**Lemma 24** *There are $C,\varepsilon_0$ that depend on $\mathsf{R},\beta$, such that whenever $\varepsilon_{\mathrm{I}} < \varepsilon_0$ then*

$$\mathsf{q}_{\mathrm{Far}}(\boldsymbol{\mathcal{X}}_k) \le 3k^{-1-C\varepsilon_{\mathrm{I}}^2}$$

*holds with probability $1-o_{\varepsilon_{\mathrm{I}},\mathsf{R},\beta}(1)$ over $\boldsymbol{\mathcal{X}}_k \sim \mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1})^{\otimes k}$.*

**Proof** Fix $\hat{\boldsymbol{X}} \in \mathcal{T}$. We start by showing that conditioned on $\hat{\boldsymbol{X}} \in \mathcal{H}_{\mathrm{Far}}$, $\bar{Q}(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k)$ is very small with high probability; to that end, we shall use Bernstein's inequality, Lemma 36.

By Lemma 22, $\|Q_i(\boldsymbol{\mathcal{X}}_k|\hat{\boldsymbol{X}})\|_\infty^{\mathcal{E}} \le 2k^{-C_1\varepsilon_I^2}$. By Lemma 23, $\mathbb{E}^{\mathcal{E}}[Q_i(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k)] \le 2k^{-1-C_2}$. Consequently, for small enough $\varepsilon_I$, $\|Q_i(\boldsymbol{\mathcal{X}}_k|\hat{\boldsymbol{X}}) - \mathbb{E}^{\mathcal{E}}[Q_i(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k)]\|_\infty^{\mathcal{E}} \le 4k^{-C_1\varepsilon_{\mathrm{I}}^2}$. Note moroever that since $Q_i(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k) \ge 0$,

$$\mathrm{Var}^{\mathcal{E}}(Q_i(\boldsymbol{\mathcal{X}}_k|\hat{\boldsymbol{X}})) \le \mathbb{E}^{\mathcal{E}}((Q_i(\boldsymbol{\mathcal{X}}_k|\hat{\boldsymbol{X}})^2) \le \|Q_i(\boldsymbol{\mathcal{X}}_k|\hat{\boldsymbol{X}})\|_\infty^{\mathcal{E}}\mathbb{E}^{\mathcal{E}}[Q_i(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k)] \le 4k^{-1-C_2-C_1\varepsilon_{\mathrm{I}}^2}.$$

Since $Q_1(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k),\dots,Q_k(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k)$ are i.i.d. conditioned on $\mathcal{E}$, by Bernstein's inequality for sums of independent bounded random variables, for some universal $c$,

$$\mathbb{P}^{\mathcal{E}}(\bar{Q}(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k) \ge t + 2k^{-1-C_2}) \le 2\exp\left(-ck\min\left\{\frac{t^2}{\mathrm{Var}^{\mathcal{E}}(Q_i(\boldsymbol{\mathcal{X}}_k|\hat{\boldsymbol{X}}))}, \frac{t}{\|Q_i(\boldsymbol{\mathcal{X}}_k|\hat{\boldsymbol{X}}) - \mathbb{E}^{\mathcal{E}}[Q_i(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k)]\|_\infty^{\mathcal{E}}}\right\}\right).$$

Setting $t = k^{1-C_1\varepsilon_{\mathrm{I}}^2/2}$ and assuming $\varepsilon_{\mathrm{I}}$ is small enough, for $c$ (perhaps other) universal,

$$\mathbb{P}^{\mathcal{E}}(\bar{Q}(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k) \ge 3k^{-1-C_1\varepsilon_{\mathrm{I}}^2/2}) \le \exp\left(-ck^{C_1\varepsilon_{\mathrm{I}}^2/2}\right).$$

So far we have shown that with overwhelming probability over the configuration $\boldsymbol{\mathcal{X}}_k$, conditioned on $\hat{\boldsymbol{X}} \in \mathcal{H}_{\mathrm{Far}}$, $\bar{Q}(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k)$ is very small. We now wish to control $\mathsf{q}_{\mathrm{Far}}(\boldsymbol{\mathcal{X}}_k) = \max_{\hat{\boldsymbol{X}} \in \mathcal{T} \cap \mathcal{H}_{\mathrm{Far}}} \bar{Q}(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k)$.

Let $N = \sum_{\hat{\boldsymbol{X}} \in \mathcal{T}} \mathbb{1}\{\hat{\boldsymbol{X}} \in \mathcal{H}_{\mathrm{Far}} \text{ and } \bar{Q}(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k) \geq 3k^{-1-C_1\varepsilon_{\mathrm{I}}^2/2}\}$. Of course, $N = 0$ implies that $\mathsf{q}_{\mathrm{Far}}(\boldsymbol{\mathcal{X}}_k) \leq 3k^{-1-C_1\varepsilon_{\mathrm{I}}^2/2}$. By Markov's inequality,

$$\Pr(N \geq 1) \leq \mathbb{E}[N] \leq |\mathcal{T}| \mathbb{P}^{\mathcal{E}}(\bar{Q}(\hat{\boldsymbol{X}}|\boldsymbol{\mathcal{X}}_k) \geq 3k^{-1-C_1\varepsilon_{\mathrm{I}}^2/2}) \leq e^{Cd\log(1/\varepsilon_{\mathrm{I}})} \exp\left(-ck^{C_1\varepsilon_{\mathrm{I}}^2/2}\right).$$

Recall that $k$ is exponential in $d$; consequently, for $\varepsilon_{\mathrm{I}} \geq C_2\sqrt{\frac{\log d}{d}}$ with large enough $C_2$ (in particular, whenever $\varepsilon_{\mathrm{I}} > 0$ is constant),

$$e^{Cd\log(1/\varepsilon_{\mathrm{I}})} \exp\left(-ck^{C_1\varepsilon_{\mathrm{I}}^2/2}\right) \leq e^{-Cd\log d} \exp(-cd^2) = o_{\varepsilon_{\mathrm{I}},\beta,\mathsf{R}}(1).$$

■

**Proof** (Of Lemma 9). Follows immediately from Lemma 24, where note that since $\mathsf{q}_{\mathrm{Far}}(\boldsymbol{\mathcal{X}}_k) \leq 1$, we may change the prefactor 3 in Lemma 24 to whatever number $> 1$ we like (at the expense of changing the exponent). We do so for convenience. ■

### C.2. Proof of Lemma 10

We start by showing that w.h.p., $\mathcal{T}_{\mathrm{Close}} \cap \mathcal{H}_{\mathrm{Far}} = \emptyset$, namely, we do not retain candidates which are $\sqrt{\varepsilon_{\mathrm{I}} d}$-far from all centers $\boldsymbol{X}_i$.

Let $\hat{\boldsymbol{X}} \in \mathcal{T} \cap \mathcal{H}_{\mathrm{Far}}$. Conditioned on the event of Lemma 9, the random variables $\{\mathsf{Test}(\hat{\boldsymbol{X}}, \boldsymbol{Y})\}$ are an i.i.d. sequence of $N$ Bernoulli trials, with success probability $\leq 2k^{-1-c\varepsilon_{\mathrm{I}}^2}$. By Chernoff's inequality (Lemma 38) and the estimate of Lemma 39,

$$\Pr(\hat{\boldsymbol{X}} \in \mathcal{T}_{\mathrm{Close}}) \leq e^{-C_1 N \varepsilon_{\mathrm{I}}^2 k^{-1}\log k},$$

for some $C_1 = C_1(\mathsf{R}, \beta)$, and assuming $k$ is large enough. Taking a union bound,

$$\Pr(\mathcal{T}_{\mathrm{Close}} \cap \mathcal{H}_{\mathrm{Far}} \neq \emptyset) \leq |\mathcal{H}_{\mathrm{Far}} \cap \mathcal{T}| e^{-C_1 N \varepsilon_{\mathrm{I}}^2 k^{-1}\log k} \leq e^{Cd\log(1/\varepsilon_{\mathrm{I}})} e^{-C_1 N \varepsilon_{\mathrm{I}}^2 k^{-1}\log k}.$$

Observe that this is $o_{\mathsf{R},\beta}(1)$ whenever $N \geq C_2 \frac{\log(1/\varepsilon_{\mathrm{I}})}{\varepsilon_{\mathrm{I}}^2} k \frac{d}{\log k}$, for some $C_2 = C_2(\mathsf{R}, \beta)$ large enough. Now, when $\mathsf{R} > 0$ then $\frac{d}{\log k} = 1/\mathsf{R}_{d,k} = 1/\mathsf{R}$ is just a constant, and so is $\sigma^2$; so for a suitably modified $C_3 = C_3(\mathsf{R}, \beta)$, $N \geq C_3 \frac{\log(1/\varepsilon_{\mathrm{I}})}{\varepsilon_{\mathrm{I}}^2} k\sigma^2$ suffices. As for the case $\mathsf{R} = 0$, $\mathsf{R}_{d,k} = \mathsf{C}(\beta\sigma^2) = \frac{1}{\beta\sigma^2} + O(\sigma^{-4})$, and so $\frac{d}{\log k} \lesssim \sigma^2$; consequently, $N \geq C_3 \frac{\log(1/\varepsilon_{\mathrm{I}})}{\varepsilon_{\mathrm{I}}^2} k\sigma^2$ suffices. This show the first claim of the Lemma.

Moving on, we need to show that for many $i \in [k]$, $\mathcal{T}_{\mathrm{Close}}$ indeed contains a vector within $\sqrt{\varepsilon_{\mathrm{I}} d}$-distance to $\boldsymbol{X}_i$.

Fix any $\hat{\boldsymbol{X}}_1, \ldots, \hat{\boldsymbol{X}}_k \in \mathcal{T}$ such that $\|\boldsymbol{X}_i - \hat{\boldsymbol{X}}_i\|^2 \leq 0.5\varepsilon_{\mathrm{I}} d$. Since $\mathcal{T}$ is an $\sqrt{0.5\varepsilon_{\mathrm{I}} d}$-net of $\sqrt{d}\mathcal{S}^{d-1}$, there certainly are such vectors in $\mathcal{T}$. We shall show that $|\mathcal{T}_{\mathrm{Close}} \cap \{\hat{\boldsymbol{X}}_1, \ldots, \hat{\boldsymbol{X}}_k\}| \geq (1-\varphi)k$ holds with the claimed probability; this clearly suffices. By the properties of the test, Lemma 9, $\Pr(\mathsf{Test}(\hat{\boldsymbol{X}}_i, \boldsymbol{Y}) = 1) \geq 0.5k^{-1}$. By Chernoff's inequality (Lemma 38),

$$\Pr(\hat{\boldsymbol{X}}_i \notin \mathcal{T}_{\mathrm{Close}}) \leq e^{-cNk^{-1}},$$

for some universal $c$. Consequently, by Markov's inequality,

$$\Pr\left(\sum_{i=1}^{k} \mathbb{1}\{\hat{\boldsymbol{X}}_i \notin \mathcal{T}_{\text{Close}}\} \geq \varphi k\right) \leq \varphi^{-1} e^{-cNk^{-1}} .$$

Thus, when $N \geq Ck\log(1/\varphi)$, for some universal $C$, we get that with probability $\geq 1-\varphi$, we have $|\mathcal{T}_{\text{Close}} \cap \{\hat{\boldsymbol{X}}_1, \ldots, \hat{\boldsymbol{X}}_k\}| \geq (1-\varphi)k$.

■

### C.3. Proof of Lemma 11

As discussed in the main text, it suffices to show that with probability $1-o_{\beta,\mathsf{R}}(1)$, the centers in $\boldsymbol{\mathcal{X}}_k$ have minimal distance $\sqrt{Ld}$ for some $L = L(\beta, \mathsf{R})$. In that case, choosing $\varepsilon_0 \leq L/16$, the required results follows immediately from Lemmas 9 and 10.

The following argument is standard. Sample centers $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ sequentially. Let $E_i$ be the event that $\boldsymbol{X}_i \notin \bigcup_{l=1}^{i-1} \mathcal{B}(\boldsymbol{X}_l, \sqrt{Ld})$ for $i = 2, \ldots, k$. Clearly, $\boldsymbol{\mathcal{X}}_k$ has minimal distance $\geq \sqrt{Ld}$ if and only if $\bigcap_{l=2}^{k} E_i$ holds. Notice that $\Pr(E_i | \bigcap_{l=1}^{i-1} E_i) \geq 1 - L^{d-1}(i-1)$, since $\boldsymbol{X}_i$ has to evade $i-1$ disjoint neighborhoods $\sqrt{d}\mathcal{S}^{d-1} \cap \mathcal{B}(\boldsymbol{X}_i, \sqrt{RLd})$, that amount to total surface area at most $\leq (i-1)\text{Surf}(\partial\mathcal{B}(\boldsymbol{X}_i, \sqrt{Ld})) = (i-1)L^{d-1}\text{Surf}(\sqrt{d}\mathcal{S}^{d-1})$. Somewhat crudely, we lower bound:

$$\Pr\left(\bigcap_{i=2}^{k} E_i\right) = \prod_{i=2}^{k} \Pr(E_l | \bigcap_{l=2}^{i-1} E_l) \geq (1 - L^{d-1}k)^k . \tag{67}$$

Whenever $L^{d-1}k = o(1/k)$, the bound in (67) tends to 1 as $k \to \infty$. When $k = e^{o(d)}$, any constant $L < 1$ will work. When $k = e^{d\mathsf{R}}$, any constant $L < e^{-2\mathsf{R}}$ will work.

■

## Appendix D. Proofs for Section 4.2

### D.1. Decoding Using a Corrupted Codebook

Upon successful completion of Step I, by Lemma 11, we will have constructed a list $\tilde{\boldsymbol{\mathcal{X}}}_{\text{I}} = (\tilde{\boldsymbol{X}}_1, \ldots, \tilde{\boldsymbol{X}}_m)$ of size $m \geq (1-\varphi)k$, such that every member of $\tilde{\boldsymbol{\mathcal{X}}}_{\text{I}}$ is $\sqrt{\varepsilon_{\text{I}} d}$-close to some unique codeword of $\boldsymbol{\mathcal{X}}_k$. In this section, we show that whenever $\varepsilon_{\text{I}}$ is smaller than *some particular threshold* $\varepsilon_0 = \varepsilon_0(\beta, \mathsf{R})$, then $\tilde{\boldsymbol{\mathcal{X}}}_{\text{I}}$ may be used to successfully decode messages encoded with $\boldsymbol{\mathcal{X}}_k$, in the following sense:

- Whenever $\tilde{\boldsymbol{\mathcal{X}}}_{\text{I}}$ contains a point $\tilde{\boldsymbol{X}}_l$ close to $\boldsymbol{X}_{\text{i}(l)} \in \boldsymbol{\mathcal{X}}_k$, applying the decoder on observation $\boldsymbol{Y} = \boldsymbol{X}_{\text{i}(l)} + \sigma\boldsymbol{Z}$ will indeed return, with high probability, the correct index $l$.
- As importantly, whenever $\tilde{\boldsymbol{\mathcal{X}}}_{\text{I}}$ does not contain a keyword which is close to $\boldsymbol{X}_i$, then applying the decoder on observation $\boldsymbol{Y} = \boldsymbol{X}_i + \sigma\boldsymbol{Z}$ will consistently return an error symbol "#"; that is, the decoder will not erroneously assign a sample $\boldsymbol{Y}$ to a label even if it does not have a close approximation for its corresponding center $\boldsymbol{X}_i$.

We now proceed to formalize the discussion above.

Denote by

$$\text{Approx}(\boldsymbol{\mathcal{X}}_k) \subseteq \bigcup_{m=0}^{k} (\sqrt{d}\mathcal{S}^{d-1})^{\otimes m} \tag{68}$$

the set of all lists $\tilde{\boldsymbol{\mathcal{X}}} = (\tilde{\boldsymbol{X}}_1, \ldots, \tilde{\boldsymbol{X}}_m)$ (for *any* $0 \le m \le k$), for which there exists a permutation $\mathsf{i} : [k] \to [k]$ satisfying

$$\|\tilde{\boldsymbol{X}}_l - \boldsymbol{X}_{\mathsf{i}(l)}\|^2 \le \varepsilon d \quad \text{for all } 1 \le l \le m\,. \tag{69}$$

A *family of decoders* is a mapping $\tilde{\boldsymbol{\mathcal{X}}} \mapsto \mathsf{Dec}(\cdot|\tilde{\boldsymbol{\mathcal{X}}})$, mapping codebooks of any length $0 \le m \le k$ to decision rules $\mathbb{R}^d \to [k] \cup \{\#\}$. $\mathsf{Dec}(\cdot|\tilde{\boldsymbol{\mathcal{X}}})$ may depend on $d, k, \sigma^2$ as well, and this shall be implied from now on.

For example, consider the nearest-neighbor family of decoders:

$$\mathsf{DecNN}(\boldsymbol{Y}|\tilde{\boldsymbol{\mathcal{X}}}) = \operatorname*{argmin}_{l\in[m]} \|\boldsymbol{Y} - \tilde{\boldsymbol{X}}_l\|\,. \tag{70}$$

Recall that when $\tilde{\boldsymbol{\mathcal{X}}} = \boldsymbol{\mathcal{X}}_k$, the resulting decoder $\mathsf{DecNN}(\cdot|\boldsymbol{\mathcal{X}}_k)$ is optimal (in the sense of average error) for decoding a message $\ell \sim \mathrm{Unif}([k])$ encoded using $\boldsymbol{\mathcal{X}}_k$. However, when $\tilde{\boldsymbol{\mathcal{X}}}$ is only a partial sub-codebook of $\boldsymbol{\mathcal{X}}_k$, using the decoder (70) might not be a good idea from a practical standing: an observation $Y = \boldsymbol{X}_i + \sigma \boldsymbol{Z}$ corresponding to a codeword $\boldsymbol{X}_i$ which is absent from $\tilde{\boldsymbol{\mathcal{X}}}$ will *necessarily* be decoded into an erroneous message. It is desirable that having identifed such a case, the decoder would instead *declare* an error.

For a codebook $\boldsymbol{\mathcal{X}}_k \in (\sqrt{d}\mathcal{S}^{d-1})^{\otimes k}$ and $\tilde{\boldsymbol{\mathcal{X}}} \in \mathrm{Approx}_\varepsilon(\boldsymbol{\mathcal{X}}_k)$, let $\mathsf{i} : [k] \to [k]$ be the permutation that satisfies (69). When there is more than one such permutation (as will surely be the case when $m \le k-2$), suppose that $\mathsf{i}$ is chosen in some systematic way, such that the assignment $\mathrm{Approx}_\varepsilon(\boldsymbol{\mathcal{X}}_k) \to \mathrm{Sym}(k)$, $\tilde{\boldsymbol{\mathcal{X}}} \mapsto \mathsf{i}$ is well-defined. Note that $\mathsf{i}([m])$ are the indices $\subseteq [k]$ of codewords $\boldsymbol{X}_i$ for which $\tilde{\boldsymbol{\mathcal{X}}}$ contains an $\sqrt{\varepsilon d}$-distance approximation.

Similar to (17), we consider the error probability of decoding a message $i \in [k]$, encoded using codebook $\boldsymbol{\mathcal{X}}_k$, with a decoder $\mathsf{Dec}(\cdot|\tilde{\boldsymbol{\mathcal{X}}})$, $\tilde{\boldsymbol{\mathcal{X}}} \in \mathrm{Approx}_\varepsilon(\boldsymbol{\mathcal{X}}_k)$:

$$P_{approx,i}(\sigma^2|\boldsymbol{\mathcal{X}}_k, \tilde{\boldsymbol{\mathcal{X}}}, \mathsf{Dec}(\cdot|\cdot)) = \begin{cases} \Pr(\mathsf{i}^{-1}(i) \neq \mathsf{Dec}(\boldsymbol{X}_i + \sigma\boldsymbol{Z}|\tilde{\boldsymbol{\mathcal{X}}})) & \text{if } i \in \mathsf{i}([m]) \\ \Pr(\# \neq \mathsf{Dec}(\boldsymbol{X}_i + \sigma\boldsymbol{Z}|\tilde{\boldsymbol{\mathcal{X}}})) & \text{if } i \notin \mathsf{i}([m]) \end{cases}. \tag{71}$$

The "twist" over (17) is that if $i \notin \mathsf{i}([m])$, we consider decoding to be successful if the decoder declares error.

We are ready to state the technical result of this section.

**Proposition 25** *Suppose that $\beta > 1$, $\mathsf{R}_{d,k} = \mathsf{C}(\beta\sigma^2)$ (in either positive or zero rate, as before), $\boldsymbol{\mathcal{X}}_k \sim \mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1})^{\otimes k}$.*

*There is $\varepsilon_0 = \varepsilon_0(\beta, \mathsf{R})$ and a family of decoders $\mathsf{Dec}(\cdot|\cdot)$ such that for all $\varepsilon \le \varepsilon_0$:*

$$\lim_{d\to\infty} \mathbb{E}\left[\sup_{\tilde{\boldsymbol{\mathcal{X}}}\in\mathrm{Approx}_\varepsilon(\boldsymbol{\mathcal{X}}_k)} \frac{1}{k}\sum_{i=1}^{k} P_{approx,i}(\sigma^2|\boldsymbol{\mathcal{X}}_k, \tilde{\boldsymbol{\mathcal{X}}}, \mathsf{Dec}(\cdot|\cdot))\right] = 0\,.$$

In words: Proposition 25 states that provided that $\varepsilon$ deceeds some particular threshold, one can construct a decoder family $\mathsf{Dec}(\cdot|\cdot)$, so that for any *adversarially chosen* $\tilde{\boldsymbol{\mathcal{X}}} \in \mathrm{Approx}_\varepsilon(\boldsymbol{\mathcal{X}}_k)$, the average decoding error (in the sense of (71)) is uniformly small; and that this holds for "most" random codebooks $\boldsymbol{\mathcal{X}}_k$.

### D.2. Proof of Proposition 25

We simply adapt the decoders used in the proof of Proposition 4, and appearing in Section A. We give a different construction at zero ($\mathsf{R} = 0$) and positive ($\mathsf{R} > 0$) rate.

D.2.1. ZERO RATE ($\mathsf{R} = 0$)

We adapt the decoder from Section A.1. Recall the decision rule implemented by this decoder (assuming access to the true codebook $\boldsymbol{\mathcal{X}}_k$): it returns $i \in [k]$ if and only if: 1) $d^{-1}\langle \boldsymbol{Y}, \boldsymbol{X}_i \rangle \geq 1-\eta_1$; 2) For all $j \in [k] \setminus \{i\}$, $d^{-1}\langle \boldsymbol{Y}, \boldsymbol{X}_j \rangle \leq 1 - \eta_2$; if no such $i$ exists it returns an error. Here, $0 < \eta_1 < \eta_2$ are sufficiently small constants.

From now on, suppose without loss of generality that a message $i$ is sent. Denote by $\boldsymbol{Y} = \boldsymbol{X}_i + \sigma \boldsymbol{Z}$, the channel output. Let $\mathcal{S}_i$ the event that the above decoder (which utilizes $\boldsymbol{\mathcal{X}}_k$) succeeds; in other words,

$$\mathcal{S}_i(\eta_1, \eta_2) = \{d^{-1}\langle \boldsymbol{Y}, \boldsymbol{X}_i \rangle \geq 1 - \eta_1\} \cap \bigcap_{j \in [k] \setminus \{i\}} \{d^{-1}\langle \boldsymbol{Y}, \boldsymbol{X}_i \rangle \leq 1 - \eta_2\}\,. \tag{72}$$

Recall: in Section A.1 we proved that $\mathbb{E}_{\boldsymbol{Y}_i, \boldsymbol{\mathcal{X}}_k}[\mathbb{1}_{\mathcal{S}_i(\eta_1,\eta_2)}] = 1 - o(1)$ for all sufficiently small constants $0 < \eta_1 < \eta_2$.

We now adapt the construction of Section A.1 to use the "corrupted" codebook $\tilde{\boldsymbol{\mathcal{X}}}$. For thresholds $\tilde{\eta}_1, \tilde{\eta}_2$, the decoder returns $l$ if and only if $\tilde{\boldsymbol{X}}_l$ is such that $d^{-1}\langle \boldsymbol{Y}, \tilde{\boldsymbol{X}}_l \rangle \geq 1 - \tilde{\eta}_1$, while for all other $j \neq l$, $d^{-1}\langle \boldsymbol{Y}, \tilde{\boldsymbol{X}}_l \rangle < 1 - \tilde{\eta}_2$. If no such codeword exists, it returns "#". Given that message $i$ is sent, the decoder succeeds upon the following event:

- If $i \in \mathsf{i}([m])$, then: 1) $d^{-1}\langle \boldsymbol{Y}, \tilde{\boldsymbol{X}}_{\mathsf{i}^{-1}(i)} \rangle \geq 1 - \tilde{\eta}_1$; 2) For all $l \neq \mathsf{i}^{-1}(i)$, $d^{-1}\langle \boldsymbol{Y}, \tilde{\boldsymbol{X}}_l \rangle < 1 - \tilde{\eta}_2$.
- If $i \notin \mathsf{i}([m])$, then for all $l \in [m]$, $d^{-1}\langle \boldsymbol{Y}, \tilde{\boldsymbol{X}}_l \rangle < 1 - \tilde{\eta}_2$.

From now on, we assume that $i \in \mathsf{i}([m])$; when $i \notin \mathsf{i}([m])$, the analysis follows in a similar manner.

Set $\boldsymbol{e}_l = \tilde{\boldsymbol{X}}_l - \boldsymbol{X}_{\mathsf{i}(l)}$, which may be chosen adversarially (but is independent of the noise $\boldsymbol{Z}$), such that $\|\boldsymbol{e}_l\| \leq \sqrt{\varepsilon d}$ (by definition of $\tilde{\boldsymbol{\mathcal{X}}} \in \mathrm{Approx}_{\varepsilon}(\boldsymbol{\mathcal{X}}_k)$). We have

$$\begin{aligned} d^{-1}\langle \boldsymbol{Y}, \tilde{\boldsymbol{X}}_l \rangle &= d^{-1}\langle \boldsymbol{Y}, \boldsymbol{X}_{\mathsf{i}(l)} \rangle + d^{-1}\langle \boldsymbol{Y}, \boldsymbol{e}_l \rangle \\ &= d^{-1}\langle \boldsymbol{Y}, \boldsymbol{X}_{\mathsf{i}(l)} \rangle + d^{-1}\langle \boldsymbol{X}_i, \boldsymbol{e}_l \rangle + d^{-1}\sigma\langle \boldsymbol{Z}, \boldsymbol{e}_l \rangle\,. \end{aligned}$$

Clearly, $|d^{-1}\langle \boldsymbol{X}_i, \boldsymbol{e}_l \rangle| \leq \sqrt{\varepsilon}$ (Cauchy-Schwartz). Set $M = \max_{l \in [m]} |d^{-1}\sigma\langle \boldsymbol{Z}, \boldsymbol{e}_l \rangle|$. Observe:

$$\begin{aligned} d^{-1}\langle \boldsymbol{Y}, \boldsymbol{X}_i \rangle \geq 1 - \eta_1 \implies d^{-1}\langle \boldsymbol{Y}, \tilde{\boldsymbol{X}}_{\mathsf{i}^{-1}(i)} \rangle \geq 1 - \eta_1 - \sqrt{\varepsilon} - M\,, \\ d^{-1}\langle \boldsymbol{Y}, \boldsymbol{X}_j \rangle \leq 1 - \eta_2 \implies d^{-1}\langle \boldsymbol{Y}, \tilde{\boldsymbol{X}}_{\mathsf{i}^{-1}(j)} \rangle \leq 1 - \eta_2 + \sqrt{\varepsilon} + M\,. \end{aligned} \tag{73}$$

We claim that with probability $1 - 2k^{-5}$, it holds that $M \leq C\sqrt{\varepsilon}$ for some $C = C(\beta)$. Consequently, if $\varepsilon$ is small enough, and the thresholds $\tilde{\eta}_1 < \tilde{\eta}_2$ are chosen such that

$$1 - \eta_1 - \sqrt{\varepsilon} - C\sqrt{\varepsilon} > 1 - \tilde{\eta}_1, \quad 1 - \eta_2 + \sqrt{\varepsilon} + C\sqrt{\varepsilon} < 1 - \tilde{\eta}_2\,,$$

then under the high-probability event $\mathcal{S}_i(\eta_1, \eta_2) \cap \{M \leq C\sqrt{\varepsilon}\}$, the adapted decoder necessarily returns the correct message $\mathsf{i}^{-1}(i)$. Since $1 - \eta_1 > 1 - \eta_2$, then this can clearly be made to hold whenever $\varepsilon < \varepsilon_0$ for small enough constant $\varepsilon_0$. The proof of Proposition 25 therefore concludes by the Lemma below.

■

**Lemma 26** *Assume the conditions of Proposition 25 (with $\mathsf{R} = 0$) and the setup described above. There is $C = C(\beta)$ such that $M \leq C\sqrt{\varepsilon}$ holds with probability $\geq 1 - 2k^{-5}$.*

**Proof** Observe that $d^{-1}\sigma\langle \boldsymbol{Z}, \boldsymbol{e}_l\rangle$ is Gaussian with mean 0 and variance $\leq \varepsilon\frac{\sigma^2}{d}$. and recall that, by definition $M = \max_{l\in[m]} |d^{-1}\sigma\langle \boldsymbol{Z}, \boldsymbol{e}_l\rangle|$.

By standard results on the maxima of Gaussian random variables, Lemmas 41 and 42, there is some universal $c$ such that $M_i \leq \sqrt{\varepsilon\frac{\sigma^2}{d}} \cdot \sqrt{c\log k}$ holds with probability $\geq 1 - 2k^{-5}$.

It remains to observe that

$$\frac{\sigma^2}{d}\log k = \sigma^2 \mathsf{R}_{d,k} = \sigma^2 \mathsf{C}(\beta\sigma^2) \leq \frac{1}{2\beta}\,,$$

where we used $\mathsf{C}(s) = \frac{1}{2}\log(1+1/s) \leq 1/(2s)$. ■

### D.2.2. Positive Rate ($\mathsf{R} > 0$)

We adapt the decoder from Section A.2. Recall the decision rule implemented by this decoder (assuming access to the true codebook $\boldsymbol{\mathcal{X}}_k$): the decoder returns an index $i$ if: 1) $d^{-1/2}\|\alpha\boldsymbol{Y} - \boldsymbol{X}_i\| \leq \sqrt{\tau_1}$; 2) For all $j \neq i$, $d^{-1/2}\|\alpha\boldsymbol{Y} - \boldsymbol{X}_i\| > \sqrt{\tau_2}$; if no such $i$ exists, it returns $\#$. Here $0 < \sqrt{\tau_1} < \sqrt{\tau_2}$ are appropriately chosen thresholds.

We now adapt the aforementioned decoder to use $\tilde{\boldsymbol{\mathcal{X}}}$ instead of $\boldsymbol{\mathcal{X}}_k$.

Observe that by the triangle inequality,

$$d^{-1/2}\|\alpha\boldsymbol{Y} - \boldsymbol{X}_{\mathsf{i}(l)}\| - \sqrt{\varepsilon} \leq d^{-1/2}\|\alpha\boldsymbol{Y} - \tilde{\boldsymbol{X}}_l\| \leq d^{-1/2}\|\alpha\boldsymbol{Y} - \boldsymbol{X}_{\mathsf{i}(l)}\| + \sqrt{\varepsilon}\,.$$

Consequently,

$$\begin{aligned} d^{-1/2}\|\alpha\boldsymbol{Y} - \boldsymbol{X}_{\mathsf{i}(l)}\| \leq \sqrt{\tau_1} &\implies d^{-1/2}\|\alpha\boldsymbol{Y} - \tilde{\boldsymbol{X}}_l\| \leq \sqrt{\tau_1} + \sqrt{\varepsilon}\,, \\ d^{-1/2}\|\alpha\boldsymbol{Y} - \boldsymbol{X}_{\mathsf{i}(l)}\| > \sqrt{\tau_2} &\implies d^{-1/2}\|\alpha\boldsymbol{Y} - \tilde{\boldsymbol{X}}_l\| > \sqrt{\tau_2} - \sqrt{\varepsilon}\,. \end{aligned} \tag{74}$$

The adapted decoder will operate as follows. Assume that $\varepsilon < \varepsilon_0$ for some $2\sqrt{\varepsilon_0} < \sqrt{\tau_2} - \sqrt{\tau_1}$, and set

$$\sqrt{\tilde{\tau}_1} = \sqrt{\tau_1} + \sqrt{\varepsilon_0}, \quad \sqrt{\tilde{\tau}_2} = \sqrt{\tau_2} - \sqrt{\varepsilon_0}\,.$$

The adapted decoder implements the following rule. It returns an index $l \in [m]$ whenever: $\tilde{\boldsymbol{X}}_l$ is such that $\|\alpha\boldsymbol{Y} - \tilde{\boldsymbol{X}}_l\| \leq \sqrt{\tilde{\tau}_1}$; 2) For all $j \in [m]\setminus\{l\}$, $\|\alpha\boldsymbol{Y} - \tilde{\boldsymbol{X}}_j\| > \sqrt{\tilde{\tau}_2}$; if no such $l$ exists, it returns $\#$.

Let us bound the error probability of the decoder. Suppose that a message $i$ was sent, so that $\boldsymbol{Y} = \boldsymbol{X}_i + \sigma\boldsymbol{Z}$, and assume without loss of generality that $i \in \mathsf{i}([m])$; the case $i \notin \mathsf{i}([m])$ follows similarly. Consider the event

$$S(\tau_1,\tau_2) = \{d^{-1/2}\|\alpha\boldsymbol{Y} - \boldsymbol{X}_i\| \leq \sqrt{\tau_1}\} \cap \bigcap_{j\in[k]\setminus\{i\}} \{d^{-1/2}\|\alpha\boldsymbol{Y} - \boldsymbol{X}_j\| > \sqrt{\tau_2}\}\,.$$

In Section A.2 it was shown that, for appropriately chosen $\tau_1, \tau_2$, $S(\tau_1,\tau_2)$ is a high-probability event. Notice that by (74), under the event $S(\tau_1,\tau_2)$, the adapted decoder indeed returns $\mathsf{i}^{-1}(i)$; thus, we have proven that the average error probability is vanishing.

■

### D.3. Proof of Lemma 12

Towards the proof of Lemma 12, we analyze the performance of Step II of the algorithm (Section 4.2) under a slightly more general setting, that does not use the strong symmetry properties that are available (w.h.p.) for a random spherical codebook $\boldsymbol{\mathcal{X}}_k$, and that were crucial for constructing the test of Step I (Section 4.1). Instead, we only assume that one has access to some decoder under which the codebook attains small *average error probability*.

Let $\boldsymbol{\mathcal{X}}_k \subseteq (\sqrt{d}\mathcal{S}^{d-1})^k$ be a fixed codebook. Let $\Psi : \mathbb{R}^d \to [k] \cup \{\#\}$ be a decoder. Denote by $P_{i,j} = \Pr_{\boldsymbol{Z}}(\Psi(\boldsymbol{X}_i + \sigma\boldsymbol{Z}) = j)$ the probability (over the noise $\boldsymbol{Z}$) that $\Psi(\cdot)$ outputs symbol $j \in [k] \cup \{\#\}$ given that the true label was $i$. To keep the presentation light, we start by introducing some notation.

We say that the decoder $\Psi(\cdot)$ satisfies the $(\rho, \varphi)$- *average error probability guarantee* if there exists an index set $\mathcal{I} \subseteq [k]$ of size $|\mathcal{I}| \geq (1-\varphi)k$ such that $\mathrm{Range}(\Psi) \subseteq \mathcal{I} \cup \{\#\}$ and the following holds. Denote by

$$\bar{P}_i = \begin{cases} 1 - P_{i,i} & \text{if } i \in \mathcal{I}, \\ 1 - P_{i,\#} & \text{if } i \notin \mathcal{I} \end{cases} \tag{75}$$

the *error probability* of the $i$-th message. Note that this is the same notion of error probability as in (71) from Section D.1 above. Then we have

$$\frac{1}{k}\sum_{i=1}^{k} \bar{P}_i \leq \rho\,. \tag{76}$$

Note: under the conditions of Lemma 12, upon successful completion of Step I, and with high probability over $\boldsymbol{\mathcal{X}}_k \sim \mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1})^{\otimes k}$, Proposition 25 implies that one may construct a decoder $\Psi(\cdot) = \mathsf{Dec}(\cdot|\tilde{\boldsymbol{\mathcal{X}}}_{\mathrm{I}})$ which satsifes the $(\rho, \varphi)$ average error probability guarantee (up to a global relabeling, which we shall ignore henceforth), for some $\rho = o(1)$.

Let us describe once again in detail the procedure of Step II, stated in terms of the notation above. One has access to a decoder (constructed from $\tilde{\boldsymbol{\mathcal{X}}}_{\mathrm{I}}$), and uses it to label a batch of $\bar{N}$ new samples, $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{\bar{N}}$. Let $S_i \subseteq [\bar{N}]$ be the subset of all measurements that have been assigned label $i$:

$$S_i = \{j \;:\; \Psi(\boldsymbol{Y}_j) = i\}\,.$$

Note that measurements assigned $\#$ are simply discarded. Next, we compute the cluster means:

$$\boldsymbol{A}_i = \frac{1}{|S_i|}\sum_{j \in S_i} \boldsymbol{Y}_j\,,$$

so that $\boldsymbol{A}_i = \boldsymbol{0}$ if $S_i = \emptyset$. The final centers returned by the procedure are simply the projections of $\boldsymbol{A}_i$ onto the ball $\mathcal{B}(\boldsymbol{0}, \sqrt{d})$:[4]

$$\hat{\boldsymbol{X}}_i = \mathcal{P}(\boldsymbol{A}_i)\,.$$

The next Lemma summarizes our guarantees for Step II as described above.

4. Note that we project onto the ball rather than the sphere $\sqrt{d}\mathcal{S}^{d-1}$ since projection onto convex sets is contracting in Euclidean norm. This is not so much the case for projection onto the sphere.

**Lemma 27** *Suppose that $\beta > 1$, and let $\varepsilon, \varphi \in (0,1)$ be constants. Let $\boldsymbol{\mathcal{X}}_k \subset (\sqrt{d}\mathcal{S}^{d-1})^k$ be some fixed center configuration, and suppose that one is given a decoder $\Psi(\cdot)$ satisfying the $(\rho, \varphi)$ average error probability guarantee, for some arbitrary $\rho = o_{\beta,\mathsf{R}}(1)$.*

*Suppose that Step II is run with $\bar{N} \geq \frac{k\sigma^2}{\varepsilon} + C\frac{k}{\varepsilon^{1/2}}\log(1/\varphi)$ for some sufficiently large univeral $C > 0$. Then*

$$\mathbb{E}\mathcal{L}_{avg}(\boldsymbol{\mathcal{X}}_k, \hat{\boldsymbol{\mathcal{X}}}_k) \leq \frac{\varepsilon}{1-\varepsilon^{1/4}} + 8\varphi + o_{\beta,\mathsf{R}}(1)\,,$$

*where the expectation is only taken over the randomness in the measurements $\boldsymbol{Y}_j = \boldsymbol{X}_{\ell_j} + \boldsymbol{Z}_j$, namely $\ell_j \sim \mathrm{Unif}([k])$ and $\boldsymbol{Z}_j \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ (and the rate of decay in the $o(1)$ term depends on $\rho$).*

The proof of Lemma 27 shall be given momentarily, in Section D.4 below. Before getting to it, let us show how it immediately implies Lemma 12:

**Proof** (Of Lemma 12). Note that the diameter of the ball $\mathcal{B}(\boldsymbol{0}, \sqrt{d})$ is $2\sqrt{d}$, so necessarily $\mathrm{dist}^2(\boldsymbol{X}_i, \hat{\boldsymbol{\mathcal{X}}}_k) \leq 4d$ and therefore $\mathcal{L}_{\mathrm{avg}}(\boldsymbol{\mathcal{X}}_k, \hat{\boldsymbol{\mathcal{X}}}_k) \leq 4$. Thus, for any event $\mathcal{E}$,

$$\mathbb{E}\mathcal{L}_{\mathrm{avg}}(\boldsymbol{\mathcal{X}}_k, \hat{\boldsymbol{\mathcal{X}}}_k) \leq \mathbb{E}\left[\mathcal{L}_{\mathrm{avg}}(\boldsymbol{\mathcal{X}}_k, \hat{\boldsymbol{\mathcal{X}}}_k)\mathbb{1}\{\mathcal{E}\}\right] + 4\Pr(\mathcal{E}^c)\,. \tag{77}$$

Let $\tilde{\boldsymbol{\mathcal{X}}}_{\mathrm{I}}$ be the list returned by Step I of the algorithm. Let $\varepsilon_0 = \varepsilon_0(\beta, \mathsf{R})$ be the threshold from Proposition 25, and consider the event

$$\mathcal{E}_I = \left\{\tilde{\boldsymbol{\mathcal{X}}}_{\mathrm{I}} \in \mathrm{Approx}_{\varepsilon_0}(\boldsymbol{\mathcal{X}}_k),\ |\mathcal{I}| \geq (1-\varphi)k\right\}\,,$$

where $\mathcal{I}$ is defined in Lemma 11. By Lemma 11, provided that $N$ is large enough, $\Pr(\mathcal{E}_1^c) \leq \varphi + o_{\beta,\mathsf{R}}(1)$. Let $\Psi(\cdot) = \mathsf{Dec}(\cdot|\tilde{\boldsymbol{\mathcal{X}}}_{\mathrm{I}})$ be the decoder promised by Proposition 25. For a suitably chosen (large enough) $\rho = o_{\beta,\mathsf{R}}(1)$, denote the event

$$\mathcal{E} = \{\Psi(\cdot) \text{ satisfies the } (\rho, \varphi) \text{ average probability guarantee}\}\,.$$

By Proposition 25, along with Markov's inequality, we have $\Pr(\mathcal{E}^c|\mathcal{E}_I) = o_{\beta,\mathsf{R}}(1)$. Thus,

$$\Pr(\mathcal{E}^c) \leq \Pr(\mathcal{E}_I^c) + \Pr(\mathcal{E}^c|\mathcal{E}_I) \leq \varphi + o_{\beta,\mathsf{R}}(1)\,.$$

Use the event $\mathcal{E}$ in (77). By Lemma 27, for large enough $\bar{N}$, $\mathbb{E}\left[\mathcal{L}_{\mathrm{avg}}(\boldsymbol{\mathcal{X}}_k, \hat{\boldsymbol{\mathcal{X}}}_k)\mathbb{1}\{\mathcal{E}\}\right] \leq \frac{\varepsilon}{1-\varepsilon^{1/4}} + 8\varphi + o_{\beta,\mathsf{R}}(1)$, and so Lemma 12 follows. ∎

### D.4. Proof of Lemma 27

Conceptually, the proof of Lemma 27 is quite straightforward. Denote by $L_i \subseteq [\bar{N}]$ the measurements $\boldsymbol{Y}_j = \boldsymbol{X}_{\ell_j} + \boldsymbol{Z}_j$ whose *true label* is $\ell_j = i$. Imagine, for a moment, that we had access to a *genie-aided* decoder, that always assigns measurements to their true labels. In that case, $S_i = L_i$, and so $\boldsymbol{A}_i$ is just the sample mean of $|L_i|$ i.i.d. Gaussian measurements $\mathcal{N}(\boldsymbol{X}_i, \sigma^2\boldsymbol{I})$. Since, *on average*, $|L_i| = \bar{N}/k$, the MSE is $\mathbb{E}\|\boldsymbol{X}_i - \boldsymbol{A}_i\|^2 \approx \sigma^2 k/\bar{N}$, which is $\approx \varepsilon$ when $\bar{N} \approx \sigma^2 k/\varepsilon$. In practice, however, one does not have access to a clairvoyant decoder: we only assume an average error probability guarantee. Another difficulty is that we cannot guarantee that $|L_i| \approx \bar{N}/k$ *simultaneously* for all $i$, unless $\sigma^2 \gtrsim \log k$ (recall the coupon-collecting issue highlighted in the main

paper: we need $\bar{N} \gtrsim k \log k$ to even *observe* a measurement of every label). Consequently, the analysis has to be carried out somewhat delicately.

We start with the trivial observation, that the average error probability guarantee implies that, in fact, *most* individual labels $i \in [k]$ must have a small error probability:

**Lemma 28** *Let $(\boldsymbol{\mathcal{X}}_k, \Psi(\cdot))$ satisfy the $(\rho, \varphi)$ average error probability guarantee (76). There is a set of indices $\mathcal{J}_1 \subseteq [k]$ of size $|\mathcal{J}_1| \geq (1-\rho^{1/2})k$ such that*

$$\bar{P}_i \leq \rho^{1/2} \quad \textit{for all } i \in \mathcal{J}_1 \,. \tag{78}$$

**Proof** An immediate consequence of Markov's inequality. ∎

Aside from lower bounding $|L_i| \gtrsim \bar{N}/k$ in expectation, we shall also need to control the number of measurements that were *erroneously* assigned label $i$ by $\Psi(\cdot)$. Following the notation of (75), let

$$\bar{Q}_i = \frac{1}{k} \sum_{l \in [k] \setminus \{i\}} P_{l,i} \tag{79}$$

be the probability that a random measurement $\boldsymbol{Y} = \boldsymbol{X}_\ell + \sigma \boldsymbol{Z}$, $\ell \sim \mathrm{Unif}([k])$ has true label $\ell \neq i$, but is erroneously assigned label $i$ by $\Psi(\cdot)$.

**Lemma 29** *Let $(\boldsymbol{\mathcal{X}}_k, \Psi(\cdot))$ satisfy the $(\rho, \varphi)$ average error probability guarantee (76). There is a set of indices $\mathcal{J}_2 \subseteq [k]$ of size $|\mathcal{J}_2| \geq (1-\rho^{1/2})k$ such that*

$$\bar{Q}_i \leq \rho^{1/2}/k \quad \textit{for all } i \in \mathcal{J}_2 \,. \tag{80}$$

**Proof** Observe that

$$\frac{1}{k}\sum_{i=1}^{k} \left( \sum_{l \in [k]\setminus\{i\}} P_{l,i} \right) = \frac{1}{k}\sum_{l=1}^{k} \left( \sum_{i \in [k]\setminus\{l\}} P_{l,i} \right) \leq \frac{1}{k}\sum_{l=1}^{k} \bar{P}_l \leq \rho \,.$$

Consequently, by Markov's inequality, there is $\mathcal{J}_2 \subseteq [k]$ of size $|\mathcal{J}_2| \geq (1-\rho^{1/2})k$ such that $\sum_{l \in [k]\setminus\{i\}} P_{l,i} \leq \rho^{1/2}$ for all $i \in \mathcal{J}_2$. ∎

Recall: $L_i \subseteq [\bar{N}]$ are the measurements whose true label is $i$; $S_i \subseteq [\bar{N}]$ are the measurements assigned label $i$ by $\Psi(\cdot)$ (whether truthfully or erroneously). Define the event

$$\mathcal{E}_{i,1} = \left\{ |S_i \setminus L_i| \leq \rho^{1/4} \frac{\bar{N}}{k} \quad \cap \quad |L_i| \geq (1-\varepsilon^{1/4}) \frac{\bar{N}}{k} \right\} . \tag{81}$$

**Lemma 30** *Under the conditions of Lemma 27, for any $i \in \mathcal{J}_2$,*

$$\Pr\left(\mathcal{E}_{i,1}^c\right) \leq \varphi + o(1) \,.$$

**Proof** Let us start by showing that $|S_i \backslash L_i|$ is small with high probability. By definition, $\Pr(\boldsymbol{Y} \in S_i \setminus L_i) = \bar{Q}_i$. Since $i \in \mathcal{J}_2$ this probability is $\leq \rho^{1/2}/k$. Thus, by Markov's inequality,

$$\Pr\left(|S_i \setminus L_i| > \rho^{1/4}\frac{\bar{N}}{k}\right) \leq \frac{\mathbb{E}|S_i \setminus L_i|}{\rho^{1/4}\frac{\bar{N}}{k}} \leq \frac{\rho^{1/2}\frac{\bar{N}}{k}}{\rho^{1/4}\frac{\bar{N}}{k}} = \rho^{1/4} = o(1)\,.$$

Moving on, observe that $|L_i| \sim \mathrm{Binomial}(1/k, \bar{N})$. By Chernoff's inequality, Lemma 40,

$$\Pr\left(|L_i| \leq (1-\varepsilon^{1/4})\frac{\bar{N}}{k}\right) \leq e^{-c\varepsilon^{1/2}\frac{\bar{N}}{k}}\,,$$

which is $\leq \varphi$ for $\bar{N} \geq C\frac{k}{\varepsilon^{1/2}}\log(1/\varphi)$. ■

For $i \in \mathcal{I}$, define the event

$$\mathcal{E}_{i,2} = \left\{|S_i \cap L_i| \geq (1-\rho^{1/4})|L_i|\right\}\,. \tag{82}$$

**Lemma 31** *Under the conditions of Lemma 27, for any $i \in \mathcal{I} \cap \mathcal{J}_1$,*

$$\Pr\left(\mathcal{E}_{i,2}^c\right) = o(1)\,.$$

**Proof** By the definition of $\mathcal{J}_1$, conditioned on $\ell_j = i$, $\Pr(j \notin S_i | \ell_j = i) = \bar{P}_i \leq \rho^{1/2}$. Thus, by Markov's inequality,

$$\Pr\left(|L_i \setminus S_i| \geq \rho^{1/4}|L_i|\right) \leq \rho^{1/4} = o(1)\,.$$

■

We are ready to bound $\mathbb{E}\mathcal{L}_{\mathrm{avg}}(\boldsymbol{\mathcal{X}}_k, \hat{\boldsymbol{\mathcal{X}}}_k)$. Observe that for any events $\mathcal{E}_1, \ldots, \mathcal{E}_k$,

$$\begin{aligned}
\mathbb{E}\mathcal{L}_{\mathrm{avg}}(\boldsymbol{\mathcal{X}}_k, \hat{\boldsymbol{\mathcal{X}}}_k) &= \frac{1}{k}\sum_{i=1}^k d^{-1}\mathbb{E}\mathrm{dist}^2(\boldsymbol{X}_i, \hat{\boldsymbol{\mathcal{X}}}_k) \\
&\overset{(i)}{\leq} \frac{1}{k}\sum_{i=1}^k d^{-1}\mathbb{E}\left[\mathrm{dist}^2(\boldsymbol{X}_i, \hat{\boldsymbol{\mathcal{X}}}_k)\mathbb{1}\left\{\mathcal{E}_i, i \in \mathcal{I}\cap\mathcal{J}_1\cap\mathcal{J}_2\right\}\right] \\
&+ \frac{4}{k}\sum_{i=1}^k \Pr(\mathcal{E}_i^c) + 4\frac{|\mathcal{I}^c|}{k} + 4\frac{|(\mathcal{J}_1\cap\mathcal{J}_2)^c|}{k} \\
&\overset{(ii)}{\leq} \frac{1}{k}\sum_{i=1}^k d^{-1}\mathbb{E}\left[\mathrm{dist}^2(\boldsymbol{X}_i, \hat{\boldsymbol{\mathcal{X}}}_k)\mathbb{1}\left\{\mathcal{E}_i, i \in \mathcal{I}\cap\mathcal{J}_1\cap\mathcal{J}_2\right\}\right] + \frac{4}{k}\sum_{i=1}^k \Pr(\mathcal{E}_i^c) + 4\varphi + o(1)\,,
\end{aligned} \tag{83}$$

where: (i) follows from $\mathrm{dist}(\boldsymbol{\mathcal{X}}_k, \hat{\boldsymbol{\mathcal{X}}}_k) \leq 4d$ (the diameter of the ball is $2\sqrt{d}$); (ii) Follows from $|\mathcal{I}| \leq \varphi k$ (by definition of the $(\rho, \varphi)$ error probability guarantee) and $|(\mathcal{J}_1 \cap \mathcal{J}_2)^c| = o(k)$ follows from Lemmas 28 and 29, with $\rho = o(1)$.

We take the event

$$\mathcal{E}_i = \mathcal{E}_{i,1} \cap \mathcal{E}_{i,2} \cap \mathcal{E}_{i,3}\,, \tag{84}$$

where $\mathcal{E}_{i,1}$ is defined in (81), $\mathcal{E}_{i,2}$ is defined in (82), and the definition of $\mathcal{E}_{i,3}$ shall be deferred for later; its details would be somewhat obtuse at this point in the analysis.

To lighten the notation, introduce

$$\mathbb{E}_{\text{avg},\mathcal{E}}[F_i] = \frac{1}{kd}\sum_{i=1}^{k} \mathbb{E}[F_i \mathbb{1}\left\{\mathcal{E}_i, i \in \mathcal{I} \cap \mathcal{J}_1 \cap \mathcal{J}_2\right\}]\,, \tag{85}$$

where $F_i$ is any sequence indexed by $i \in [k]$.

For $i \in \mathcal{I}$,

$$\operatorname{dist}^2(\boldsymbol{X}_i, \hat{\boldsymbol{\mathcal{X}}}_k) \le \|\boldsymbol{X}_i - \hat{\boldsymbol{X}}_i\|^2 \le \|\boldsymbol{X}_i - \boldsymbol{A}_i\|^2\,,$$

where we used $\hat{\boldsymbol{X}}_i = \mathcal{P}(\boldsymbol{A}_i)$ and that projection onto convex sets is contracting with respect to Euclidean norm. Moreover, note that

$$\boldsymbol{\mathcal{D}} = (\boldsymbol{D}_1, \dots, \boldsymbol{D}_k) \in \mathbb{R}^{d\times k} \quad \mapsto \quad (\mathbb{E}_{\text{avg},\mathcal{E}}[\|\boldsymbol{D}_i\|^2])^{1/2}$$

is a semi-norm, hence satisfies the triangle inequality.

We decompose

$$\begin{aligned}
\boldsymbol{A}_i &= \frac{1}{|S_i|}\sum_{j\in S_i} \boldsymbol{Y}_j \\
&= \frac{1}{|S_i|}\sum_{j\in L_i} \boldsymbol{Y}_j + \frac{1}{|S_i|}\sum_{j\in S_i\setminus L_i} \boldsymbol{Y}_j - \frac{1}{|S_i|}\sum_{j\in L_i\setminus S_i} \boldsymbol{Y}_j \\
&= \frac{1}{|S_i|}\sum_{j\in L_i} \boldsymbol{Y}_i + \frac{1}{|S_i|}\sum_{j\in S_i\setminus L_i} \boldsymbol{X}_i - \frac{1}{|S_i|}\sum_{j\in L_i\setminus S_i} \boldsymbol{X}_{\ell_j} + \frac{\sigma}{|S_i|}\sum_{j\in S_i\setminus L_i} \boldsymbol{Z}_j - \frac{\sigma}{|S_i|}\sum_{j\in L_i\setminus S_i} \boldsymbol{Z}_j\,,
\end{aligned}$$

and accordingly bound the first term of (83):

$$\begin{aligned}
\left(\mathbb{E}_{\text{avg},\mathcal{E}}\operatorname{dist}^2(\boldsymbol{X}_i, \hat{\boldsymbol{\mathcal{X}}}_k)\right)^{1/2} &\le \left(\mathbb{E}_{\text{avg},\mathcal{E}}\|\boldsymbol{X}_i - \boldsymbol{A}_i\|^2\right)^{1/2} \\
&\le \underbrace{\left(\mathbb{E}_{\text{avg},\mathcal{E}}\left\|\boldsymbol{X}_i - \frac{|L_i|}{|S_i|}\boldsymbol{X}_i\right\|^2\right)^{1/2}}_{I_1} + \underbrace{\left(\mathbb{E}_{\text{avg},\mathcal{E}}\left\|\frac{|L_i|}{|S_i|}\boldsymbol{X}_i - \frac{1}{|S_i|}\sum_{j\in L_i}\boldsymbol{Y}_j\right\|^2\right)^{1/2}}_{I_2} \\
&+ \underbrace{\left(\mathbb{E}_{\text{avg},\mathcal{E}}\left\|\frac{1}{|S_i|}\sum_{j\in S_i\setminus L_i}\boldsymbol{X}_i - \frac{1}{|S_i|}\sum_{j\in L_i\setminus S_i}\boldsymbol{X}_{\ell_j}\right\|^2\right)^{1/2}}_{I_3} \\
&+ \underbrace{\left(\mathbb{E}_{\text{avg},\mathcal{E}}\left\|\frac{\sigma}{|S_i|}\sum_{j\in S_i\setminus L_i}\boldsymbol{Z}_j\right\|^2\right)^{1/2}}_{I_4} + \underbrace{\left(\mathbb{E}_{\text{avg},\mathcal{E}}\left\|\frac{\sigma}{|S_i|}\sum_{j\in L_i\setminus S_i}\boldsymbol{Z}_j\right\|^2\right)^{1/2}}_{I_5}\,.
\end{aligned} \tag{86}$$

We proceed to bound the terms above. Starting with $I_1$, observe that

$$I_1^2 = \mathbb{E}_{\text{avg},\mathcal{E}} \left\| \boldsymbol{X}_i - \frac{|L_i|}{|S_i|} \boldsymbol{X}_i \right\|^2 = d \cdot \mathbb{E}_{\text{avg},\mathcal{E}} \left| 1 - \frac{|L_i|}{|S_i|} \right| ,$$

and therefore $I_1 = o(1)$, since under $\mathcal{E}_i$ we must have $(1 - o(1))|L_i| \leq |S_i| \leq (1 + o(1))|L_i|$.

Moving on to $I_2$,

$$I_2^2 = \mathbb{E}_{\text{avg},\mathcal{E}} \left\| \frac{|L_i|}{|S_i|} \boldsymbol{X}_i - \frac{1}{|S_i|} \sum_{j \in L_i} \boldsymbol{Y}_j \right\|^2 = \mathbb{E}_{\text{avg},\mathcal{E}} \left[ \frac{\sigma^2 |L_i|}{|S_i|^2} \left\| \frac{1}{\sqrt{|L_i|}} \sum_{j \in L_i} \boldsymbol{Z}_j \right\|^2 \right] .$$

Noting that $\frac{1}{\sqrt{|L_i|}} \sum_{j \in L_i} \boldsymbol{Z}_j \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and that under $\mathcal{E}_i$, $\frac{|L_i|}{|S_i|^2} \leq (1 + o(1)) \frac{1}{1 - \varepsilon^{1/4}} \frac{k}{\bar{N}}$ we get that $I_2^2 \leq \frac{\varepsilon}{1 - \varepsilon^{1/4}} + o(1)$ since $\bar{N} \geq \sigma^2 k / \varepsilon$.

As for $I_3$,

$$I_3^2 = \mathbb{E}_{\text{avg},\mathcal{E}} \left\| \frac{1}{|S_i|} \sum_{j \in S_i \setminus L_i} \boldsymbol{X}_i - \frac{1}{|S_i|} \sum_{j \in L_i \setminus S_i} \boldsymbol{X}_{\ell_j} \right\|^2 \leq d \cdot \mathbb{E}_{\text{avg},\mathcal{E}} \left( \frac{|S_i \setminus L_i| + |L_i \setminus S_i|}{|S_i|} \right)^2 = o(1) .$$

Bounding the terms $I_4, I_5$ is somewhat more involved, and requires the introduction of a new event, $\mathcal{E}_{i,3}$, whose definition has been deferred up to this point.

**Lemma 32** *Define the event $\mathcal{E}_{i,3}$ in (88). Then $\Pr(\mathcal{E}_{i,3}) = o(1)$ and moreover*

$$I_4, I_5 = o(1) .$$

To keep the narrative flow, we defer the proof of Lemma 32 to Section D.4.1 below.

We are ready to tie all loose ends, and finish the proof of Lemma 27. By (86) and the upper bounds we have shown for $I_1, \ldots, I_5$, we get $\mathbb{E}_{\text{avg},\mathcal{E}} \text{dist}^2(\boldsymbol{X}_i, \hat{\boldsymbol{\mathcal{X}}}_k) \leq \frac{\varepsilon}{1 - \varepsilon^{1/4}} + o(1)$. Combining Lemmas 30, 31 and 32, we get $\Pr(\mathcal{E}_i) \leq \varphi + o(1)$. Plugging these estimates into (83), we finally get the claimed bound of Lemma 27.

### D.4.1. Proof of Lemma 32

The terms $I_4, I_5$ correspond to sums of independent "noise" vectors, whose mean is zero. Therefore, one expects different $\boldsymbol{Z}_j$-s to cancel out one another on average, so that, for example (considering $I_4$),

$$\mathbb{E} \left\| \frac{\sigma}{|S_i|} \sum_{j \in S_i \setminus L_i} \boldsymbol{Z}_j \right\|^2 \approx \frac{\sigma^2}{|S_i|^2} \cdot |S_i \setminus L_i| d$$

rather than $\frac{\sigma^2}{|S_i|^2} \cdot |S_i \setminus L_i|^2 d$ which is what we would have gotten by naive application of the triangle inequality. A subtle point is that the set $S_i$ actually depends on the noise vectors $\boldsymbol{Z}_j$, so one needs to apply some care when taking the expectation above.[5] We propose to overcome this difficulty through a rather crude bound.

5. *A priori*, we cannot discount the possibility that conditioned on $j \in S_i$, the noise $\boldsymbol{Z}_j$ biases towards some particular direction.

For a subset $B \subseteq [n]$ (which itself may be random, but independent of $\{\boldsymbol{Z}_j\}_{j\in[\bar{N}]}$, let

$$\boldsymbol{D}(B,m) = \max_{S\subseteq B, |S|\le m} \left\| \sum_{j\in S} \boldsymbol{Z}_j \right\|^2 .$$

Since under $\mathcal{E}_{i,1}\cap\mathcal{E}_{i,2}\subset\mathcal{E}_i$ we have

$$|S_i \setminus L_i| \le \rho^{1/4}\frac{\bar{N}}{k}, \quad |L_i \setminus S_i| \le \rho^{1/4}|L_i|, \quad |L_i| \ge (1-\varepsilon^{1/4})\frac{\bar{N}}{k},$$

recalling the definition of $I_4, I_5$, (86), clearly,

$$\begin{aligned} I_4^2 &= \sigma^2 \cdot \mathbb{E}_{\text{avg},\mathcal{E}} \left[ \frac{1}{|S_i|^2} \left\| \sum_{j\in S_i\setminus L_i} \boldsymbol{Z}_j \right\|^2 \right] \\ &\lesssim \frac{\sigma^2}{(\bar{N}/k)^2} \cdot \mathbb{E}_{\text{avg},\mathcal{E}} \left[ \boldsymbol{D}([\bar{N}], \rho^{1/4}(\bar{N}/k)) \right], \\ I_5^2 &= \sigma^2 \cdot \mathbb{E}_{\text{avg},\mathcal{E}} \left[ \frac{1}{|S_i|^2} \left\| \sum_{j\in L_i\setminus S_i} \boldsymbol{Z}_j \right\|^2 \right] \\ &\lesssim \sigma^2 \cdot \mathbb{E}_{\text{avg},\mathcal{E}} \left[ \frac{1}{|L_i|^2} \cdot \boldsymbol{D}(L_i, \rho^{1/4}|L_i|) \right] . \end{aligned} \tag{87}$$

We are ready to define the event $\mathcal{E}_{i,3}$, which has been deferred up to this point.

**The event $\mathcal{E}_{i,3}$.** For $C$ a sufficiently large universal constant, define

$$\begin{aligned} \mathcal{E}_{i,3} = &\left\{ \boldsymbol{D}([n], \rho^{1/4}(\bar{N}/k)) \le C \left(\rho^{1/4}(\bar{N}/k)\right)^2 \log \frac{\bar{N}e}{\rho^{1/4}(\bar{N}/k)} + C\rho^{1/4}(\bar{N}/k)\log d \right\}, \\ &\cap \left\{ \boldsymbol{D}(L_i, \rho^{1/4}|L_i|) \le C \left(\rho^{1/4}|L_i|\right)^2 \log \frac{|L_i|e}{\rho^{1/4}|L_i|} + C\rho^{1/4}|L_i|\log d \right\} . \end{aligned} \tag{88}$$

By Lemma 33, given below, $C$ may indeed be chosen so that $\Pr(\mathcal{E}_{i,3}^c) = o(1)$.

**Bounding $I_4$.** Using (87) and (88),

$$I_4^2 \lesssim \frac{\sigma^2}{(\bar{N}/k)^2} \cdot d^{-1} \cdot \left\{ o\left((\bar{N}/k)^2 \log k\right) + o\left((\bar{N}/k)\log d\right) \right\} .$$

The first term is $o(1)$ because $\sigma^2 d^{-1}\log k = O(1)$. The second term is $o(1)$ because $\frac{\sigma^2}{(\bar{N}/k)} d^{-1}\log d = O(d^{-1}\log d) = o(1)$, since $\bar{N} \gtrsim k\sigma^2$. Thus $I_4 = o(1)$.

**Bounding $I_5$.** Using (87) and (88),

$$\begin{aligned} I_5^2 &\lesssim \sigma^2 \mathbb{E}_{\text{avg},\mathcal{E}} \left[ o(1) + o\left(\frac{1}{|L_i|}\log d\right) \right] \\ &\lesssim o\left(\sigma^2 d^{-1}\right) + o\left(\sigma^2 d^{-1} \frac{1}{(\bar{N}/k)} \log d\right) . \end{aligned}$$

The first term is $o(1)$ since, by assumption, $\sigma^2 = o(d)$. Since $\bar{N} \gtrsim k\sigma^2$, the second terms is $o\left(\frac{\log d}{d}\right) = o(1)$. Thus, $I_5 = o(1)$.

This conclude the proof of Lemma 32.

■

**A Technical Lemma.**

**Lemma 33** *Let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ be independent. For a set $S \subseteq [n]$ let $\boldsymbol{W}_S = \sum_{i \in S} \boldsymbol{Z}_i$. There is a univeral $C > 0$ such that for $t \geq 1$,*

$$\Pr\left(\max_{S \subseteq [n], |S| \leq t} \|\boldsymbol{W}_S\| \geq Ct\sqrt{\log \frac{ne}{t}} + C\sqrt{t \log d}\right) \leq d^{-5}\,.$$

**Proof** This is a straightforward application of the well-known Lemmas 41 and Lemma 42, along with a standard "trick".

Let $\mathcal{T}$ be a $1/2$-net of $\mathcal{S}^{d-1}$, of size $\leq 5^d$ (e.g. (Wainwright, 2019, Example 5.8)). It may be readily verified that for any vector $\boldsymbol{x} \in \mathbb{R}^d$, $\|\boldsymbol{x}\| = \max_{\boldsymbol{e} \in \mathcal{S}^{d-1}} \langle \boldsymbol{x}, \boldsymbol{e}\rangle \leq 2 \max_{\boldsymbol{e} \in \mathcal{T}} \langle \boldsymbol{x}, \boldsymbol{e}\rangle$. Set $N_t = |\mathcal{T}| \sum_{s=1}^t \binom{n}{s} \lesssim 5^d (ne/t)^t$. Note also that for any $\boldsymbol{e} \in \mathcal{S}^{d-1}$, $\mathbb{E}\left[\langle \boldsymbol{W}_S, \boldsymbol{e}\rangle^2\right] \leq |S| \leq t$.

By the expectation bound Lemma 41, and the Borell-TIS inequality, Lemma 42, for $x \geq 0$,

$$\begin{aligned}\Pr\left(\max_{S \subseteq [n], |S| \leq t} \|\boldsymbol{W}_S\| \geq 2\sqrt{t}(\sqrt{2 \log N_t} + x)\right) &\leq \Pr\left(\max_{S \subseteq [n], |S| \leq t, \boldsymbol{e} \in \mathcal{T}} \langle \boldsymbol{W}_S, \boldsymbol{e}\rangle \geq \sqrt{t}(\sqrt{2 \log N_t} + x)\right) \\ &\leq e^{-x^2/2}\,.\end{aligned}$$

Set $x = \sqrt{10 \log(d)}$ to get the claimed bound.

■

## Appendix E. Auxiliary Technical Results

### E.1. Concentration Inequalities

The Gaussian Lipschitz concentration inequality (Wainwright, 2019, Theorem 2.25):

**Lemma 34 (Gaussian Lipschitz concentration inequality)** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be $L$-Lipschitz, and $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. For all $t \geq 0$,*

$$\Pr\left(f(\boldsymbol{Z}) \geq \mathbb{E}f(\boldsymbol{Z}) + t\right) \leq e^{-\frac{t^2}{2L^2}}.$$

Standard bound on the measure of a spherical cap (Wainwright, 2019, Eq. (3.33)):

**Lemma 35** *Let $\boldsymbol{Z} \sim \mathrm{Unif}(\mathcal{S}^{d-1})$. For all $\boldsymbol{u} \in \mathcal{S}^{d-1}$ and $t \in (0, 1)$,*

$$\Pr\left(\langle \boldsymbol{u}, \boldsymbol{Z}\rangle \geq t\right) \leq e^{-dt^2/2}.$$

We state Bernstein's inequality for independent bounded random variables (Vershynin, 2018, Theorem 2.8.4)

**Lemma 36** *Let $X_1, \ldots, X_n$ be independent, centered, with $|X_i| \leq K$. Set $S_n = \sum_{i=1}^n X_i$. Then for all $t \geq 0$,*

$$\Pr\left(|S_n| \geq t\right) \leq 2\exp\left(-\frac{t^2/2}{\sum_{i=1}^n \mathrm{Var}(X_i) + Kt/3}\right).$$

The following is Bernstein's inequality for sums of independent sub-Exponential random variables (Vershynin, 2018, Theorem 2.8.1):

**Lemma 37 (Bernstein's inequality, sub-Exponential RVs)** *Let $X_1, \ldots, X_n$ be independent and sub-exponential. Set $S_n = \sum_{i=1}^n X_i$. Then for all $t \geq 0$,*

$$\Pr\left(|S_n - \mathbb{E}[S_n]| \geq t\right) \leq 2\exp\left[-c\min\left(\frac{t^2}{\sum_{i=1}^n \|X_i\|_{\psi_1}^2}, \frac{t}{\max_{1\leq i\leq n}\|X_i\|_{\psi_1}}\right)\right],$$

*where $c > 0$ is an absolute constant.*

We state the following version of the Chernoff bound for Bernoulli random variables. For a citable reference, see for example (Boucheron et al., 2013, Section 2.2):

**Lemma 38 (Chernoff's inequality)** *Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli random variables, with $\mathbb{E}[X_i] = q$. Let $t \leq q \leq p$. Then*

$$\Pr\left(\sum_{i=1}^n X_i \geq p\right) \leq e^{nD_{\mathrm{KL}}(p;q)}, \quad \Pr\left(\sum_{i=1}^n X_i \leq t\right) \leq e^{nD_{\mathrm{KL}}(t;q)},$$

*where*

$$D_{\mathrm{KL}}(p;q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$$

*is the Kullback-Leibler divergence divergence between $\mathrm{Ber}(p)$ and $\mathrm{Ber}(q)$.*

We shall use Lemma 38 with the following easy estimate:

**Lemma 39** *Let $0 < q < p < 1/2$. Then*

$$D_{\mathrm{KL}}(p;q) \geq p \log \frac{p}{q} - 2p\,.$$

**Proof** One may readily verify that $\log(1-p) \geq -2p$ holds for $0 \leq p \leq 1/2$. Thus,

$$D_{\mathrm{KL}}(p;q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} \geq p\log\frac{p}{q} + (1-p)\log(1-p) \overset{(\star)}{\geq} p\log\frac{p}{q} + \log(1-p) \geq p\log\frac{p}{q} - 2p\,,$$

where $(\star)$ holds since $\log(1-p)$ is negative. ■

The following is a version of Chernoff's inequality, specialized for small deviations, and taken from (Vershynin, 2018, Exercise 2.3.5)

**Lemma 40 (Chernoff's inequality; small deviations)** *In the setting of Lemma 38, for $\delta \in (0,1)$,*

$$\Pr\left(\sum_{i=1}^n X_i \geq (1+\delta)qn\right) \leq e^{-c\delta^2 qn}, \quad \Pr\left(\sum_{i=1}^n X_i \leq (1-\delta)qn\right) \leq e^{-c\delta^2 qn}\,,$$

*where $c > 0$ is universal.*

### E.2. Maxima of Gaussian Random Variables

We state two elementary results about the maximum of $n$ Gaussian random variables.

**Lemma 41** *Let $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ be a Gaussian random vector, such that $\mathbb{E}[\boldsymbol{Z}] = \mathbf{0}$ and $\mathbb{E}[Z_i^2] \leq \sigma^2$ for all $i$. Then*

$$\mathbb{E}[\max_{1\leq i\leq n} Z_i] \leq \sqrt{2\sigma^2 \log n}\,.$$

*(When $Z_1, \ldots, Z_n$ are uncorrelated, this is in fact tight to leading order. But we shall not use this stronger fact.)*

**Proof** This is classical. For completeness, we give a one-line proof. For all $\beta > 0$,

$$\mathbb{E}[\max_{1\leq i\leq n} Z_i] \leq \frac{1}{\beta}\mathbb{E}\log\sum_{i=1}^n e^{\beta Z_i} \overset{(\star)}{\leq} \frac{1}{\beta}\log\mathbb{E}\sum_{i=1}^n e^{\beta Z_i} \leq \frac{1}{\beta}(\log n + \frac{1}{2}\sigma^2\beta^2)\,,$$

where $(\star)$ follows from Jensen's inequality. Now set $\beta = \sqrt{2\sigma^2 \log n}$. ■

The following is a special (easy) case of the Borell-TIS inequality, see e.g. (Adler and Taylor, 2009, Theorem 2.1.1). Alternatively, this follows immediately from the Gaussian Lipschitz concentration inequality, Lemma 34:

**Lemma 42 (Borell-TIS)** *Let $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ be a Gaussian random vector with $\mathbb{E}[\boldsymbol{Z}] = \mathbf{0}$. Set $\sigma^2 = \max_{1\leq i\leq n} \mathbb{E}Z_i^2$. Then for $t \geq 0$,*

$$\Pr(\max_{1\leq i\leq n} Z_i \geq \mathbb{E}[\max_{1\leq i\leq n} Z_i] + t) \leq e^{-\frac{t^2}{2\sigma^2}}, \quad \Pr(\max_{1\leq i\leq n} Z_i \leq \mathbb{E}[\max_{1\leq i\leq n} Z_i] - t) \leq e^{-\frac{t^2}{2\sigma^2}}\,.$$

### E.3. Results From Information Theory

The following is Fano's inequality, see e.g. (Polyanskiy and Wu, 2014, Theorem 5.2).

**Lemma 43 (Fano's inequality)** *Let $(\ell, Z, \hat{\ell})$ be random variables such that $\ell, \hat{\ell} \in [k]$, and the Markov chain*

$$\ell \longrightarrow Z \longrightarrow \hat{\ell}$$

*holds. Denote $p_e = (\Pr(\ell \neq \hat{\ell}))$. Then*

$$H(\ell|Z) \leq h_b(p_e) + p_e \log(k-1)\,,$$

*where $h_b(p) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$ is the binary entropy function.*

Lastly is the celebrated I-MMSE relation of (Guo et al., 2005, Theorem 2):

**Lemma 44 (I-MMSE)** *Let $\boldsymbol{X} \in \mathbb{R}^d$ be any random vector with finite second moments, $\mathbb{E}\|\boldsymbol{X}\|^2 < \infty$. Let $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ be indepdent of $\boldsymbol{X}$, and denote $\boldsymbol{Y}(s) = \sqrt{s}\boldsymbol{X} + \boldsymbol{Z}$. Then*

$$\frac{d}{ds} I(\boldsymbol{X}; \boldsymbol{Y}(s)) = \frac{1}{2} \mathbb{E}\left[\|\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}|\boldsymbol{Y}(s))\|^2\right] =: \mathrm{mmse}(s)\,.$$