12-2013

Minimax Theory for High-dimensional Gaussian Mixtures with Sparse Mean Separation

Martin Azizyan
Carnegie Mellon University, mazizyan@andrew.cmu.edu

Aarti Singh
Carnegie Mellon University, aarti@cs.cmu.edu

Larry Wasserman
Carnegie Mellon University, larry@stat.cmu.edu

Follow this and additional works at: http://repository.cmu.edu/machine_learning

Part of the Theory and Algorithms Commons

Published In
Minimax Theory for High-dimensional Gaussian Mixtures with Sparse Mean Separation

Martin Azizyan  
Machine Learning Department  
Carnegie Mellon University  
mazizyan@cs.cmu.edu

Aarti Singh  
Machine Learning Department  
Carnegie Mellon University  
aarti@cs.cmu.edu

Larry Wasserman  
Department of Statistics  
Carnegie Mellon University  
larry@stat.cmu.edu

Abstract

While several papers have investigated computationally and statistically efficient methods for learning Gaussian mixtures, precise minimax bounds for their statistical performance as well as fundamental limits in high-dimensional settings are not well-understood. In this paper, we provide precise information theoretic bounds on the clustering accuracy and sample complexity of learning a mixture of two isotropic Gaussians in high dimensions under small mean separation. If there is a sparse subset of relevant dimensions that determine the mean separation, then the sample complexity only depends on the number of relevant dimensions and mean separation, and can be achieved by a simple computationally efficient procedure. Our results provide the first step of a theoretical basis for recent methods that combine feature selection and clustering.

1 Introduction

Gaussian mixture models provide a simple framework for several machine learning problems including clustering, density estimation and classification. Mixtures are especially appealing in high dimensional problems. Perhaps the most common use of Gaussian mixtures is for clustering. Of course, the statistical (and computational) behavior of these methods can degrade in high dimensions. Inspired by the success of variable selection methods in regression, several authors have considered variable selection for clustering. However, there appears to be no theoretical results justifying the advantage of variable selection in high dimensional setting.

To see why some sort of variable selection might be useful, consider clustering $n$ subjects using a vector of $d$ genes for each subject. Typically $d$ is much larger than $n$ which suggests that statistical clustering methods will perform poorly. However, it may be the case that there are only a small number of relevant genes in which case we might expect better behavior by focusing on this small set of relevant genes.

The purpose of this paper is to provide precise bounds on clustering error with mixtures of Gaussians. We consider both the general case where all features are relevant, and the special case where only a subset of features are relevant. Mathematically, we model an irrelevant feature by requiring the mean of that feature to be the same across clusters, so that the feature does not serve to differentiate the groups. Throughout this paper, we use the probability of misclustering an observation, relative to the optimal clustering if we had known the true distribution, as our loss function. This is akin to using excess risk in classification.

This paper makes the following contributions:

- We provide information theoretic bounds on the sample complexity of learning a mixture of two isotropic Gaussians with equal weight in the small mean separation setting that precisely captures the dimension dependence, and matches known sample complexity requirements for some existing algorithms. This also debunks the myth that there is a gap between...
statistical and computational complexity of learning mixture of two isotropic Gaussians for small mean separation. Our bounds require non-standard arguments since our loss function does not satisfy the triangle inequality.

- We consider the high-dimensional setting where only a subset of relevant dimensions determine the mean separation between mixture components and show that learning is substantially easier as the sample complexity only depends on the sparse set of relevant dimensions. This provides some theoretical basis for feature selection approaches to clustering.
- We show that a simple computationally feasible procedure nearly achieves the information theoretic sample complexity even in high-dimensional sparse mean separation settings.

**Related Work.** There is a long and continuing history of research on mixtures of Gaussians. A complete review is not feasible but we mention some highlights of the work most related to ours.

Perhaps the most popular method for estimating a mixture distribution is maximum likelihood. Unfortunately, maximizing the likelihood is NP-Hard. This has led to a stream of work on alternative methods for estimating mixtures. These new algorithms use pairwise distances, spectral methods or the method of moments.

Pairwise methods are developed in [Dasgupta (1999)](#), [Schulman and Dasgupta (2000)](#) and [Arora and Kannan (2001)](#). These methods require the mean separation to increase with dimension. The first one requires the separation to be $\sqrt{d}$ while the latter two improve it to $d^{1/4}$. To avoid this problem, [Vempala and Wang (2004)](#) introduced the idea of using spectral methods for estimating mixtures of spherical Gaussians which makes mean separation independent of dimension. The assumption that the components are spherical was removed in [Brubaker and Vempala (2008)](#). Their method only requires the components to be separated by a hyperplane and runs in polynomial time, but requires $n = \Omega(d^4 \log d)$ samples. Other spectral methods include [Kannan et al. (2005)](#), [Achlioptas and McSherry (2005)](#) and [Hsu and Kakade (2013)](#). The latter uses clever spectral decompositions together with the method of moments to derive an effective algorithm.

[Kalai et al. (2012)](#) use the method of moments to get estimates without requiring separation between components of the mixture components. A similar approach is given in [Belkin and Sinha (2010)](#). [Chaudhuri et al. (2009)](#) give a modified k-means algorithm for estimating a mixture of two Gaussians. For the large mean separation setting $\mu > 1$, [Chaudhuri et al. (2009)](#) show that $n = \Omega(d/\mu^2)$ samples are needed. They also provide an information theoretic bound on the necessary sample complexity of any algorithm which matches the sample complexity of their method (up to log factors) in $d$ and $\mu$. If the mean separation is small $\mu < 1$, they show that $n = \Omega(d/\mu^4)$ samples are sufficient. Our results for the small mean separation setting give a matching necessary condition. Assuming the separation between the component means is not too sparse, [Chaudhuri and Rao (2008)](#) provide an algorithm for learning the mixture that has polynomial computational and sample complexity.

Most of these papers are concerned with computational efficiency and do not give precise, statistical minimax upper and lower bounds. None of them deal with the case we are interested in, namely, a high dimensional mixture with sparse mean separation.

We should also point out that the results in different papers are not necessarily comparable since different authors use different loss functions. In this paper we use the probability of misclassifying a future observation, relative to how the correct distribution clusters the observation, as our loss function. This should not be confused with the probability of attributing a new observation to the wrong component of the mixture. The latter loss does not typically tend to zero as the sample size increases. Our loss is similar to the excess risk used in classification where we compare the misclassification rate of a classifier to the misclassification rate of the Bayes optimal classifier.

Finally, we remind the reader that our motivation for studying sparsely separated mixtures is that this provides a model for variable selection in clustering problems. There are some relevant recent papers on this problem in the high-dimensional setting. [Pan and Shen (2007)](#) use penalized mixture models to do variable selection and clustering simultaneously. [Witten and Tibshirani (2010)](#) develop a penalized version of k-means clustering. Related methods include [Raftery and Dean (2006); Sun et al. (2012) and Guo et al. (2010)](#). The applied bioinformatics literature also contains a huge number of heuristic methods for this problem. None of these papers provide minimax bounds for the clustering error or provide theoretical evidence of the benefit of using variable selection in unsupervised problems such as clustering.
2 Problem Setup

In this paper, we consider the simple setting of learning a mixture of two isotropic Gaussians with equal mixing weights. Given \( n \) data points \( X_1, \ldots, X_n \in \mathbb{R}^d \) drawn i.i.d. from a \( d \)-dimensional mixture density function
\[
p_\theta(x) = \frac{1}{2} f(x; \mu_1, \sigma^2 I) + \frac{1}{2} f(x; \mu_2, \sigma^2 I),
\]
where \( f(\cdot; \mu, \Sigma) \) is the density of \( \mathcal{N}(\mu, \Sigma) \), \( \sigma > 0 \) is a fixed constant, and \( \theta := (\mu_1, \mu_2) \in \Theta \). We consider two classes \( \Theta \) of parameters:
\[
\Theta_\lambda = \{ (\mu_1, \mu_2) : \| \mu_1 - \mu_2 \| \geq \lambda \}
\]
\[
\Theta_{\lambda,s} = \{ (\mu_1, \mu_2) : \| \mu_1 - \mu_2 \| \geq \lambda, \| \mu_1 - \mu_2 \|_0 \leq s \} \subseteq \Theta_\lambda.
\]
The first class defines mixtures where the components have a mean separation of at least \( \lambda > 0 \). The second class defines mixtures with mean separation \( \lambda > 0 \) along a sparse set of \( s \in \{1, \ldots, d\} \) dimensions. Also, let \( \theta_0 \) denote the probability measure corresponding to \( p_\theta \).

For a mixture with parameter \( \theta \), the Bayes optimal classification, that is, assignment of a point \( x \in \mathbb{R}^d \) to the correct mixture component, is given by the function
\[
F_\theta(x) = \arg\max_{i \in \{1,2\}} f(x; \mu_i, \sigma^2 I).
\]
Given any other candidate assignment function \( F : \mathbb{R}^d \to \{1,2\} \), we define the loss incurred by \( F \) as
\[
L_\theta(F) = \min_{\pi} P_\theta(\{ x : F_\theta(x) \neq \pi(F(x)) \})
\]
where the minimum is over all assignments \( \pi : \{1,2\} \to \{1,2\} \). This is the probability of misclassifying data relative to an oracle that uses the true distribution to do optimal clustering.

We denote by \( \hat{F}_n \) any assignment function learned from the data \( X_1, \ldots, X_n \), also referred to as an estimator. The goal of this paper is to quantify how the minimax expected loss (worst case expected loss for the best estimator)
\[
R_n = \inf_{\hat{F}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta L_\theta(\hat{F}_n)
\]
scales with number of samples \( n \), the dimension of the feature space \( d \), the number of relevant dimensions \( s \), and the signal-to-noise ratio defined as the ratio of mean separation to standard deviation \( \lambda/\sigma \). We will also demonstrate a specific estimator that achieves the minimax scaling.

For the purposes of this paper, we say that feature \( j \) is irrelevant if \( \mu_1(j) = \mu_2(j) \). Otherwise we say that feature \( j \) is relevant.

3 Minimax Bounds

3.1 Small mean separation setting without sparsity

We begin without assuming any sparsity, that is, all features are relevant. In this case, comparing the projections of the data to the projection of the sample mean onto the first principal component suffices to achieve both minimax optimal sample complexity and clustering loss.

Theorem 1 (Upper bound). Define
\[
\hat{F}_n(x) = \begin{cases} 
1 & \text{if } x^T v_1(\hat{\Sigma}_n) \geq \hat{\mu}_n^T v_1(\hat{\Sigma}_n) \\
2 & \text{otherwise},
\end{cases}
\]
where \( \hat{\mu}_n = n^{-1} \sum_{i=1}^n X_i \) is the sample mean, \( \hat{\Sigma}_n = n^{-1} \sum_{i=1}^n (X_i - \hat{\mu}_n)(X_i - \hat{\mu}_n)^T \) is the sample covariance and \( v_1(\hat{\Sigma}_n) \) denotes the eigenvector corresponding to the largest eigenvalue of \( \hat{\Sigma}_n \). If \( n \geq \max(68, 4d) \), then
\[
\sup_{\theta \in \Theta_\lambda} \mathbb{E}_\theta L_\theta(\hat{F}) \leq 600 \max \left( 4\sigma^2 / \lambda^2, 1 \right) \sqrt{ \frac{d \log(nd)}{n} }.
\]

We believe our results should also hold in the unequal mixture weight setting without major modifications.
Furthermore, if \( \frac{\lambda}{\sigma} \geq 2 \max(80, 14\sqrt{d}) \), then
\[
\sup_{\theta \in \Theta_\lambda} \mathbb{E}_\theta L_\theta(\hat{F}) \leq 17 \exp\left(-\frac{n}{32}\right) + 9 \exp\left(-\frac{\lambda^2}{80\sigma^2}\right).
\]
We note that the estimator in Theorem 1 (and that in Theorem 3) does not use knowledge of \( \sigma^2 \).

**Theorem 2** (Lower bound). Assume that \( d \geq 9 \) and \( \frac{\lambda}{\sigma} \leq 0.2 \). Then
\[
\inf_{\hat{F}_n} \sup_{\theta \in \Theta_\lambda} \mathbb{E}_\theta L_\theta(\hat{F}_n) \geq \frac{1}{500} \min \left\{ \sqrt{\frac{\log 2 \sigma^2}{3}} \sqrt{\frac{d-1}{n}}, \frac{1}{4} \right\}.
\]
We believe that some of the constants (including lower bound on \( d \) and exact upper bound on \( \lambda/\sigma \)) can be tightened, but the results demonstrate matching scaling behavior of clustering error with \( d, n \) and \( \lambda/\sigma \). Thus, we see (ignoring constants and log terms) that
\[
R_n \approx \frac{\sigma^2}{\lambda^2} \sqrt{\frac{d}{n}} \quad \text{or equivalently} \quad n \approx \frac{d}{\lambda^2/\sigma^4} \text{ for a constant target value of } R_n.
\]
The result is quite intuitive: the dependence on dimension \( d \) is as expected. Also we see that the rate depends in a precise way on the signal-to-noise ratio \( \lambda/\sigma \). In particular, the results imply that we need \( d \leq n \).

In modern high-dimensional datasets, we often have \( d > n \) i.e. large number of features and not enough samples. However, inference is usually tractable since not all features are relevant to the learning task at hand. This sparsity of relevant feature set has been successfully exploited in supervised learning problems such as regression and classification. We show next that the same is true for clustering under the Gaussian mixture model.

### 3.2 Sparse and small mean separation setting

Now we consider the case where there are \( s < d \) relevant features. Let \( S \) denote the set of relevant features. We begin by constructing an estimator \( \hat{S}_n \) of \( S \) as follows. Define
\[
\hat{\tau}_n = \min_{1 \leq i \leq d} \sum_{i=1}^{d} \hat{S}_n(i, i), \quad \text{where}
\]
\[
\alpha = \sqrt{\frac{6 \log(nd)}{n}} + \frac{2 \log(nd)}{n}.
\]
Now let
\[
\hat{S}_n = \{i \in \{1, \ldots, d\} : \hat{S}_n(i, i) > \hat{\tau}_n\}.
\]
Now we use the same method as before, but using only the features in \( \hat{S}_n \) identified as relevant.

**Theorem 3** (Upper bound). Define
\[
\hat{F}_n(x) = \begin{cases} 
1 & \text{if } x_{\hat{S}_n}^T v_1(\hat{\Sigma}_{\hat{S}_n}) \geq \hat{\mu}_{\hat{S}_n}^T v_1(\hat{\Sigma}_{\hat{S}_n}) \\
2 & \text{otherwise}
\end{cases}
\]
where \( x_{\hat{S}_n} \) are the coordinates of \( x \) restricted to \( \hat{S}_n \), and \( \hat{\mu}_{\hat{S}_n} \) and \( \hat{\Sigma}_{\hat{S}_n} \) are the sample mean and covariance of the data restricted to \( \hat{S}_n \). If \( n \geq \max(68, 4s) \), \( d \geq 2 \) and \( \alpha \leq \frac{1}{4} \), then
\[
\sup_{\theta \in \Theta_{\lambda,s}} \mathbb{E}_\theta L_\theta(\hat{F}) \leq 603 \max\left\{ \frac{16\sigma^2}{\lambda^2}, 1 \right\} \sqrt{\frac{s \log(ns)}{n}} + 220 \frac{\sigma \sqrt{s}}{\lambda} \left( \frac{\log(nd)}{n} \right)^{\frac{1}{4}}.
\]
Next we find the lower bound.

**Theorem 4** (Lower bound). Assume that \( \frac{\lambda}{\sigma} \leq 0.2 \), \( d \geq 17 \), and that \( 5 \leq s \leq \frac{d+3}{4} \). Then
\[
\inf_{\hat{F}_n} \sup_{\theta \in \Theta_{\lambda,s}} \mathbb{E}_\theta L_\theta(\hat{F}_n) \geq \frac{1}{600} \min \left\{ \sqrt{\frac{8 \sigma^2}{45 \lambda^2}} \sqrt{\frac{s-1}{n}} \log\left( \frac{d-1}{s-1} \right), \frac{1}{2} \right\}.
\]
We remark again that the constants in our bounds can be tightened, but the results suggest that
\[
\frac{\sigma}{\lambda} \left( \frac{s^2 \log d}{n} \right)^{1/4} \geq R_n \geq \frac{\sigma^2}{\lambda^2} \sqrt{\frac{s \log d}{n}},
\]

or \( n = \Omega \left( \frac{s^2 \log d}{\lambda^4/\sigma^4} \right) \) for a constant target value of \( R_n \).

In this case, we have a gap between the upper and lower bounds for the clustering loss. Also, the sample complexity can possibly be improved to scale as \( s \) (instead of \( s^2 \)) using a different method. However, notice that the dimension only enters logarithmically. If the number of relevant dimensions is small then we can expect good rates. This provides some justification for feature selection. We conjecture that the lower bound is tight and that the gap could be closed by using a sparse principal component method as in [Vu and Lei (2012)] to find the relevant features. However, that method is combinatorial and so far there is no known computationally efficient method for implementing it with similar guarantees.

We note that the upper bound is achieved by a two-stage method that first finds the relevant dimensions and then estimates the clusters. This is in contrast to the methods described in the introduction which do clustering and variable selection simultaneously. See the supplementary material for complete proofs of all results.

The proof techniques of the sparse and non-sparse lower bounds are almost identical. The main difference is that in the non-sparse case, we use the Varshamov–Gilbert bound (Lemma 1) to construct a set of sufficiently dissimilar hypotheses, whereas in the sparse case we use an analogous result for sparse hypercubes (Lemma 2). See the supplementary material for complete proofs of all results.

### 4 Proofs of the Lower Bounds

The lower bounds for estimation problems rely on a standard reduction from expected error to hypothesis testing that assumes the loss function is a semi-distance, which the clustering loss isn’t. However, a local triangle inequality-type bound can be shown (Proposition 2). This weaker condition can then be used to lower-bound the expected loss, as stated in Proposition 1 (which follows easily from Fano’s inequality).

The proof techniques of the sparse and non-sparse lower bounds are almost identical. The main difference is that in the non-sparse case, we use the Varshamov–Gilbert bound (Lemma 1) to construct a set of sufficiently dissimilar hypotheses, whereas in the sparse case we use an analogous result for sparse hypercubes (Lemma 2). See the supplementary material for complete proofs of all results.

**Lemma 1** (Varshamov–Gilbert bound). Let \( \Omega = \{0, 1\}^m \) for \( m \geq 8 \). There exists a subset \( \{\omega_0, \ldots, \omega_M\} \subseteq \Omega \) such that \( \omega_0 = (..., 0) \), \( \rho(\omega_i, \omega_j) \geq \frac{\alpha}{M} \) for all \( 0 \leq i < j \leq M \), and \( M \geq 2^{m/8} \), where \( \rho \) denotes the Hamming distance between two vectors (Tsybakov (2009)).

**Lemma 2.** Let \( \omega = \{\omega \in \{0, 1\}^m : \|\omega\|_0 = s\} \) for integers \( m > s \geq 1 \) such that \( s \leq m/4 \). There exist \( \omega_0, \ldots, \omega_M \in \Omega \) such that \( \rho(\omega_i, \omega_j) > s/2 \) for all \( 0 \leq i < j \leq M \), and \( M \geq \left( \frac{m}{s} \right)^{s/5} - 1 \) (Massart (2007), Lemma 4.10).

**Proposition 1.** Let \( \theta_0, \ldots, \theta_M \in \Theta_\lambda \) (or \( \Theta_\lambda, s \)), \( M \geq 2 \), \( 0 < \alpha < 1/8 \), and \( \gamma > 0 \). If for all \( 1 \leq i \leq M \), \( \text{KL}(P_{\theta_i}, P_{\theta_0}) \leq \frac{\alpha \log M}{n} \), and if \( L_{\theta_i}(F) < \gamma \) implies \( L_{\theta_j}(F) \leq \gamma \) for all \( 0 \leq i \neq j \leq M \) and clusterings \( \hat{F} \), then \( \inf_{F_n} \max_{\theta_i \in [0, M]} E_{\theta_i}(L_{\theta_i}(F)) \geq 0.07 \gamma \).

**Proposition 2.** For any \( \theta, \theta' \in \Theta_\lambda, \) and any clustering \( \hat{F} \), let \( \tau = L_{\theta}(\hat{F}) + \sqrt{\text{KL}(P_\theta, P_{\theta'})/2} \). If \( L_{\theta}(F_{\theta'}) + \tau \leq 1/2 \), then \( L_{\theta}(F_{\theta'}) - \tau \leq L_{\theta}(F) \). We will also need the following two results. Let \( \theta = (\mu_0 - \mu/2, \mu_0 + \mu/2) \) and \( \theta' = (\mu_0 - \mu'/2, \mu_0 + \mu'/2) \) for \( \mu_0, \mu, \mu' \in \mathbb{R}^d \) such that \( \|\mu\| = \|\mu'\| \), and let \( \cos \beta = \frac{\|\mu - \mu'\|}{\|\mu\|} \).

**Proposition 3.** Let \( g(x) = \phi(x)/\phi(x) - x\Phi(-x) \). Then \( 2g \left( \frac{\|\mu\|}{2\sigma} \right) \sin \beta \cos \beta \leq L_{\theta}(F_{\theta'}) \leq \frac{\tan \beta}{\pi} \).

**Proposition 4.** Let \( \xi = \frac{\|\mu\|}{2\sigma} \). Then \( \text{KL}(P_{\theta}, P_{\theta'}) \leq \xi^4 (1 - \cos \beta) \).
Proposition 5. Let \( \theta = (\mu_0 - \mu, \mu_0 + \mu) \) for some \( \mu, \mu_0 \in \mathbb{R}^d \) and \( X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_0 \). For any \( \delta > 0 \), we have \( \|\mu_0 - \mu_n\| \geq \sigma \sqrt{\frac{2\max(d, \log \frac{1}{\delta})}{n}} + \|\mu\| \sqrt{\frac{2\log \frac{1}{n}}{n}} \) with probability at least \( 1 - 3\delta \).
Proposition 6. Let \( \theta = (\mu_0 - \mu, \mu_0 + \mu) \) for some \( \mu_0, \mu \in \mathbb{R}^d \) and \( X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_0 \) with \( d > 1 \) and \( n \geq 4d \). Define \( \cos \beta = \frac{v_1(\sigma^2 I + \mu \mu^T)^T v_1(\Sigma_n)}{\|v_1\|} \). For any \( 0 < \delta < \frac{d-1}{\sqrt{\kappa}} \), if

\[
\max\left(\frac{\sigma^2}{\|\mu\|^2}, \frac{\sigma}{\|\mu\|}\right) \sqrt{\frac{\max(d, \delta \log \frac{1}{\delta})}{n}} \leq \frac{1}{160},
\]

then with probability at least \( 1 - 12\delta - 2\exp\left(-\frac{n}{20}\right) \),

\[
\sin \beta \leq 14 \max\left(\frac{\sigma^2}{\|\mu\|^2}, \frac{\sigma}{\|\mu\|}\right) \sqrt{d} \max\left(1, \frac{10}{n} \log \frac{d}{\delta}\right).
\]

Proposition 7. Let \( \theta = (\mu_0 - \mu, \mu_0 + \mu) \), and for some \( x_0, v \in \mathbb{R}^d \) with \( \|v\| = 1 \), let \( \tilde{F}(x) = 1 \) if \( x^T v \geq x_0^T v \), and 2 otherwise. Define \( \cos \beta = \frac{|v|^2 \mu/\|\mu\|}. If \( |(x_0 - \mu_0)^T v| \leq \sigma \epsilon_1 + \|\mu\| \epsilon_2 \) for some \( \epsilon_1 \geq 0 \) and \( 0 \leq \epsilon_2 \leq \frac{1}{2} \), and if \( \sin \beta \leq \frac{1}{\sqrt{\kappa}} \), then

\[
L_\theta(\tilde{F}) \leq \exp\left\{-\frac{1}{2} \max\left(0, \frac{\|\mu\|}{2\sigma} - 2\epsilon_1\right)^2 \right\} \left[2\epsilon_1 + \epsilon_2 \frac{\|\mu\|}{\sigma} + 2 \sin \beta \left(2 \sin \beta \frac{\|\mu\|}{\sigma} + 1\right)\right].
\]

Proof. Let \( r = \left\Vert \frac{(x_0 - \mu_0)^T v}{\cos \beta} \right\Vert \). Since the clustering loss is invariant to rotation and translation,

\[
L_\theta(\tilde{F}) = \left\{ 1 \right\} \left[ \Phi\left(\frac{\|\mu\| + |x| \tan \beta + r}{\|\mu\| - |x| \tan \beta - r} \right) \right] dx
\]

Since \( \tan \beta \leq \frac{1}{2} \) and \( \epsilon_2 \leq \frac{1}{2} \), we have \( r \leq 2\epsilon_1 + 2\|\mu\| \epsilon_2 \), and \( \Phi\left(\frac{\|\mu\| - r}{\sigma} \right) \leq 2 \left(\epsilon_1 + \epsilon_2 \frac{\|\mu\|}{2\sigma} \right) \Phi\left(\max\left(0, \frac{\|\mu\|}{2\sigma} - 2\epsilon_1\right)\right) \). Defining \( A = \left\{ \frac{\|\mu\| - r}{\sigma} \right\} \),

\[
\int_{-\infty}^{\infty} \phi(x) \left[ \Phi\left(\frac{\|\mu\| - r}{\sigma} - |x| \tan \beta \right) \right] dx \leq 2 \int_{0}^{\infty} \Phi\left(\frac{\|\mu\| - r}{\sigma} - |x| \tan \beta \right) dx.
\]

where we used \( u = x \cos \beta - y \sin \beta \) and \( v = x \sin \beta + y \cos \beta \) in the second step. The bound now follows easily.

Proof of Theorem 7. Using Propositions 5 and 6 with \( \delta = \frac{1}{\sqrt{\kappa}} \), Proposition 7 and the fact that \( (C + x) \exp(-\max(0, x - 4)^2/8) \leq (C + x) \exp(-\max(0, x - 4)^2/10) \) for all \( C, x > 0 \),

\[
\mathbb{E}_0 L_\theta(\tilde{F}) \leq 600 \max\left(\frac{4\sigma^2}{X^2}, 1\right) \sqrt{\frac{d \log(nd)}{n}}
\]

(it is easy to verify that the bounds are decreasing with \( \|\mu\| \), so we use \( \|\mu\| = \frac{1}{2} \) to bound the supremum). In the \( d = 1 \) case Proposition 6 and the fact that \( (C + x) \exp(-\max(0, x - 4)^2/8) \leq (C + x) \exp(-\max(0, x - 4)^2/10) \) for all \( C, x > 0 \). \( \Box \)

Proposition 8. Let \( \theta = (\mu_0 - \mu, \mu_0 + \mu) \) for some \( \mu_0, \mu \in \mathbb{R}^d \) and \( X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_0 \). For any \( 0 < \delta < \frac{1}{\sqrt{\kappa}} \) such that \( \sqrt{\frac{6 \log \frac{1}{\delta}}{n}} \leq \frac{1}{2} \), with probability at least \( 1 - 6d\delta \), for all \( i \in [d] \),

\[
|\hat{\Sigma}_n(i, i) - (\sigma^2 + \mu(i)^2)| \leq \sigma^2 \sqrt{\frac{6 \log \frac{1}{\delta}}{n} + 2\sigma |\mu(i)|} \sqrt{\frac{2 \log \frac{1}{\kappa}}{n} + (\sigma + |\mu(i)|)^2 \frac{2 \log \frac{1}{\kappa}}{n}}.
\]

Proposition 9. Let \( \theta = (\mu_0 - \mu, \mu_0 + \mu) \) for some \( \mu_0, \mu \in \mathbb{R}^d \) and \( X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_0 \). Define

\[
S(\theta) = \{i \in [d] : \mu(i) \neq 0\} \quad \text{and} \quad \widetilde{S}(\theta) = \{i \in [d] : |\mu(i)| \geq 4\sigma \sqrt{\kappa}\}.
\]

Assume that \( n \geq 1, d \geq 2, \) and \( \alpha \leq \frac{1}{4} \). Then \( \widetilde{S}(\theta) \subseteq \hat{S}_n \subseteq S(\theta) \) with probability at least \( 1 - \frac{n}{6} \).
**Proof.** By Proposition 8, with probability at least \(1 - \frac{5}{n}\),

\[
|\hat{\Sigma}_n(i, i) - (\sigma^2 + \mu(i))^2| \leq \sigma^2 \sqrt{\frac{6 \log(nd)}{n}} + 2\sigma|\mu(i)| \sqrt{\frac{2 \log(nd)}{n}} + (\sigma + |\mu(i)|)^2 \sqrt{\frac{2 \log(nd)}{n}}
\]

for all \(i \in [d]\). Assume the above event holds. If \(S(\theta) = [d]\), then of course \(\hat{S}_n \subseteq S(\theta)\). Otherwise, for \(i \notin S(\theta)\), we have \((1 - \alpha)\sigma^2 \leq \hat{\Sigma}_n(i, i) \leq (1 + \alpha)\sigma^2\), so it is clear that \(\hat{S}_n \subseteq S(\theta)\). The remainder of the proof is trivial if \(\hat{S}(\theta) = \emptyset\) or \(S(\theta) = \emptyset\). Assume otherwise. For any \(i \in S(\theta)\),

\[
\hat{\Sigma}_n(i, i) \geq (1 - \alpha)\sigma^2 + \left(1 - \frac{2 \log(nd)}{n}\right) \mu(i)^2 - 2\alpha\sigma|\mu(i)|.
\]

By definition, \(|\mu(i)| \geq 4\sigma\sqrt{\alpha}\) for all \(i \in \hat{S}(\theta)\), so \((1 + \alpha)\sigma^2 \leq \hat{\Sigma}_n(i, i)\) and \(i \in \hat{S}_n\) (we ignore strict equality above as a measure 0 event), i.e. \(\hat{S}(\theta) \subseteq \hat{S}_n\), which concludes the proof. 

**Proof of Theorem 3.** Define \(S(\theta) = \{i \in [d] : \mu(i) \neq 0\}\) and \(\hat{S}(\theta) = \{i \in [d] : |\mu(i)| \geq 4\sigma\sqrt{\alpha}\}\).

Assume \(\hat{S}(\theta) \subseteq S_n \subseteq S(\theta)\) (by Proposition 9, this holds with probability at least \(1 - \frac{5}{n}\)). If \(S(\theta) = 0\), then we simply have \(E_\theta L_\theta(\hat{F}) \leq \frac{1}{2}\).

Assume \(\hat{S}(\theta) \neq 0\). Let \(\cos \tilde{\beta} = \frac{v_1(\hat{\Sigma}_n)\Sigma v_1(\Sigma)}{\|v_1(\hat{\Sigma}_n)\Sigma v_1(\Sigma)\|}\), \(\cos \hat{\beta} = \frac{|v_1(\hat{\Sigma}_n)\Sigma v_1(\Sigma)|}{\|v_1(\hat{\Sigma}_n)\Sigma v_1(\Sigma)\|}\), and \(\cos \beta = |v_1(\hat{\Sigma}_n)\Sigma v_1(\Sigma)|\) where \(\Sigma = \sigma^2 I + \mu\mu^T\), and for simplicity we define \(\Sigma_n = \Sigma \hat{S}_n\) and \(\Sigma \hat{S}_n\) to be the same as \(\hat{S}_n\) and \(\Sigma\) in \(\hat{S}_n\), respectively, and 0 elsewhere. Then \(\sin \tilde{\beta} \leq \sin \hat{\beta} + \sin \beta\), and

\[
\sin \beta = \frac{\|\mu - \mu \hat{S}(\theta)\|}{\|\mu\|} \leq \frac{\|\mu - \mu \hat{S}(\theta)\|}{\|\mu\|} \leq \frac{4\sigma\sqrt{\alpha}}{\|\mu\|} \leq 8\sigma\sqrt{\alpha}.
\]

Using the same argument as the proof of Theorem 1 as long as the above bound is smaller than \(\frac{1}{2\sqrt{s}}\),

\[
E_\theta L_\theta(\hat{F}) \leq 600 \max \left(\frac{\sigma^2}{(\frac{1}{2} - 4\sigma\sqrt{\alpha})}, 1\right) \sqrt{\frac{s\log(ns)}{n}} + 104\sigma\frac{\sqrt{s\alpha}}{\lambda} + \frac{3}{n}.
\]

Using the fact \(L_\theta(\hat{F}) \leq \frac{1}{2}\) always, and that \(\alpha \leq \frac{1}{4}\) implies \(\frac{\log(nd)}{n} \leq 1\), the bound follows. 

**6 Conclusion**

We have provided minimax lower and upper bounds for estimating high dimensional mixtures. The bounds show explicitly how the statistical difficulty of the problem depends on dimension \(d\), sample size \(n\), separation \(\lambda\) and sparsity level \(s\).

For clarity, we focused on the special case where there are two spherical components with equal mixture weights. In future work, we plan to extend the results to general mixtures of \(k\) Gaussians.

One of our motivations for this work is the recent interest in variable selection methods to facilitate clustering in high dimensional problems. Existing methods such as Pan and Shen (2007); [Witten and Tibshirani (2010); Raftery and Dean (2006); Sun et al. (2012) and Guo et al. (2010)] provide promising numerical evidence that variable selection does improve high dimensional clustering. Our results provide some theoretical basis for this idea.

However, there is a gap between the results in this paper and the above methodology papers. Indeed, as of now, there is no rigorous proof that the methods in those papers outperform a two stage approach where the first stage screens for relevant features and the second stage applies standard clustering methods on the features found in the first stage. We conjecture that there are conditions under which simultaneous feature selection and clustering outperforms a two stage method. Setting this question will require the aforementioned extension of our results to the general mixture case.

**Acknowledgements**

This research is supported in part by NSF grants IIS-1116458 and CAREER award IIS-1252412, as well as NSF Grant DMS-0806009 and Air Force Grant FA95500910373.
References


Supplementary file for NIPS submission: Minimax Theory for High-dimensional Gaussian Mixtures with Sparse Mean Separation

Martin Azizyan  
Machine Learning Department  
Carnegie Mellon University  
Pittsburgh, PA 15213  
mazizyan@cs.cmu.edu

Aarti Singh  
Machine Learning Department  
Carnegie Mellon University  
Pittsburgh, PA 15213  
aarti@cs.cmu.edu

Larry Wasserman  
Department of Statistics  
Carnegie Mellon University  
Pittsburgh, PA 15213  
larry@stat.cmu.edu

1 Notation

For \( \theta = (\mu_1, \mu_2) \in \mathbb{R}^{2 \times d} \), define

\[
p_\theta(x) = \frac{1}{2} f(x; \mu_1, \sigma^2 I) + \frac{1}{2} f(x; \mu_2, \sigma^2 I),
\]

where \( f(\cdot; \mu, \Sigma) \) is the density of \( \mathcal{N}(\mu, \Sigma) \), \( \sigma > 0 \) is a fixed constant. Let \( \mathcal{P}_\theta \) denote the probability measure corresponding to \( p_\theta \). We consider two classes \( \Theta \) of parameters:

\[
\Theta_\lambda = \{ (\mu_1, \mu_2) : \|\mu_1 - \mu_2\| \geq \lambda \}
\]

\[
\Theta_{\lambda, s} = \{ (\mu_1, \mu_2) : \|\mu_1 - \mu_2\| \geq \lambda, \|\mu_1 - \mu_2\|_0 \leq s \} \subseteq \Theta_\lambda.
\]

Throughout this document, \( \phi \) and \( \Phi \) denote the standard normal density and distribution functions.

For a mixture with parameter \( \theta \), the Bayes optimal classification, that is, assignment of a point \( x \in \mathbb{R}^d \) to the correct mixture component, is given by the function

\[
F_\theta(x) = \arg\max_{i \in \{1, 2\}} f(x; \mu_i, \sigma^2 I).
\]

Given any other candidate assignment function \( F : \mathbb{R}^d \to \{1, 2\} \), we define the loss incurred by \( F \) as

\[
L_\theta(F) = \min_{\pi} P_\theta(\{ x : F_\theta(x) \neq \pi(F(x)) \})
\]

where the minimum is over all permutations \( \pi : \{1, 2\} \to \{1, 2\} \).

For \( X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\theta \), let \( \hat{\mu}_n \) and \( \hat{\Sigma}_n \) be the mean and covariance of the corresponding empirical distribution.

Also, for a matrix \( B \), \( v_i(B) \) and \( \lambda_i(B) \) are the \( i \)’th eigenvector and eigenvalue of \( B \) (assuming \( B \) is symmetric), arranged so that \( \lambda_i(B) \geq \lambda_{i+1}(B) \), and \( \|B\|_2 \) is the spectral norm.

2 Upper bounds

2.1 Standard concentration bounds

2.1.1 Concentration bounds for estimating the mean

Proposition 1. Let \( X \sim \chi^2_d \). Then for any \( \epsilon > 0 \),

\[
\mathbb{P}(X > (1 + \epsilon)d) \leq \exp \left\{ -\frac{d}{2} \left( \epsilon - \log(1 + \epsilon) \right) \right\}.
\]
If $\epsilon < 1$, then

$$\mathbb{P}(X < (1 - \epsilon)d) \leq \exp \left\{ \frac{d}{2}(\epsilon + \log(1 - \epsilon)) \right\}.$$  

**Proof.** Since $\mathbb{E}e^{tX} = (1 - 2t)^{-\frac{d}{2}}$ for $0 < t < \frac{1}{2}$,

$$\mathbb{P}(X > (1 + \epsilon)d) = \mathbb{P}(e^{tX} > e^{(1+\epsilon)d})$$

$$\leq e^{\frac{1}{2}(1+\epsilon)d}(1 - 2t)^{-\frac{d}{4}}$$

$$= \exp \left[ -t(1 + \epsilon)d + \frac{d}{2} \log \frac{1}{1 - 2t} \right].$$

To minimize the right hand side, we differentiate the exponent with respect to $t$ to obtain the equation

$$-t(1 + \epsilon) + \frac{d}{2} \log(1 + 2t) = 0$$

which can be satisfied by setting $t = \frac{1}{2} \left( \frac{1 - \epsilon}{1 - \epsilon} \right) < \frac{1}{2}$ (it is easy to verify that this is a global minimum). Using this value for $t$, the first bound follows.

Also, for $\epsilon < 1$,

$$\mathbb{P}(X < (1 - \epsilon)d) = \mathbb{P}(e^{-tX} > e^{-t(1-\epsilon)d})$$

$$\leq e^{t(1-\epsilon)d}(1 + 2t)^{-\frac{d}{2}}$$

$$= \exp \left[ t(1 - \epsilon)d - \frac{d}{2} \log(1 + 2t) \right]$$

and setting $t = \frac{1}{2}(\frac{1}{1-\epsilon} - 1)$.

$$\mathbb{P}(X < (1 - \epsilon)d) \leq \exp \left\{ \frac{d}{2}(\epsilon + \log(1 - \epsilon)) \right\}.$$  

\[ \square \]

**Proposition 2.** Let $Z_1, \ldots, Z_n \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$. Then for any $\epsilon > 0$,

$$\mathbb{P} \left( \left\| \frac{1}{n} \sum_{i=1}^{n} Z_i \right\| \geq \sqrt{\frac{(1 + \epsilon)d}{n}} \right) \leq \exp \left\{ -\frac{d}{2}(\epsilon - \log(1 + \epsilon)) \right\}.$$  

**Proof.** Using Proposition 1

$$\mathbb{P} \left( \left\| \frac{1}{n} \sum_{i=1}^{n} Z_i \right\| \geq \sqrt{\frac{(1 + \epsilon)d}{n}} \right) = \mathbb{P} \left( \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \right\|^2 \geq (1 + \epsilon)d \right)$$

$$= \mathbb{P}(X \geq (1 + \epsilon)d)$$

$$\leq \exp \left\{ -\frac{d}{2}(\epsilon - \log(1 + \epsilon)) \right\}$$

where $X \sim \chi^2_d$.  

\[ \square \]

2.1.2 Concentration bounds for estimating principal direction

**Proposition 3.** Let $Z_1, \ldots, Z_n \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$ and $\delta > 0$. If $n \geq d$ then with probability at least $1 - 3\delta$,

$$\| \hat{\Sigma}_n - I_d \|_2 \leq \sqrt{\frac{d}{n} \max \left( 1, \frac{8 \log \frac{1}{\delta}}{d} \right) \frac{d}{n}}$$

$$+ \sqrt{\frac{2 \log \frac{1}{\delta}}{d} \max \left( 1, \frac{8 \log \frac{1}{\delta}}{d} \right) \frac{d}{n}}$$

where $\hat{\Sigma}_n$ is the empirical covariance of $Z_i$.  

2
Proof. Let $Z_n = \frac{1}{n} \sum_{i=1}^{n} Z_i$. Then

$$
\left\| \hat{\Sigma}_n - I_d \right\|_2 = \left\| \frac{1}{n} \sum_{i=1}^{n} Z_i Z_i^T - I_d - Z_n Z_n^T \right\|_2 \\
\leq \left\| \frac{1}{n} \sum_{i=1}^{n} Z_i Z_i^T - I_d \right\|_2 + \| Z_n \|_2.
$$

It is well known that for any $\epsilon_1 > 0$,

$$
P \left( \left\| \frac{1}{n} \sum_{i=1}^{n} Z_i Z_i^T - I_d \right\|_2 \geq 3 (1 + \epsilon_1) \frac{\sqrt{d} \max \left( 1, (1 + \epsilon_1) \sqrt{\frac{d}{n}} \right)}{\sqrt{n}} \right) \leq 2 \exp \left\{ - \frac{d \epsilon_1^2}{2} \right\}.
$$

Using this along with Proposition 2, we have for any $\epsilon_2 > 0$,

$$
P \left( \left\| \hat{\Sigma}_n - I_d \right\|_2 \geq 3 (1 + \epsilon_1) \frac{\sqrt{d} \max \left( 1, (1 + \epsilon_1) \sqrt{\frac{d}{n}} \right)}{\sqrt{n}} \right) \leq 2 \exp \left\{ - \frac{d \epsilon_1^2}{2} \right\} + \exp \left\{ - \frac{d}{2} (\epsilon_2 - \log(1 + \epsilon_2)) \right\}.
$$

Setting $\epsilon_1 = \sqrt{\frac{2 \log \frac{1}{d}}{d}}$,

$$
P \left( \left\| \hat{\Sigma}_n - I_d \right\|_2 \geq 3 \left( 1 + \sqrt{\frac{2 \log \frac{1}{d}}{d}} \right) \frac{\sqrt{d} \max \left( 1, (1 + \epsilon_1) \sqrt{\frac{d}{n}} \right)}{\sqrt{n}} \right) \leq 2 \delta + \exp \left\{ - \frac{d}{2} (\epsilon_2 - \log(1 + \epsilon_2)) \right\}
$$

and, setting $\epsilon_2 = \sqrt{\frac{8 \log \frac{1}{d}}{d}} \max \left( 1, \frac{8 \log \frac{1}{d}}{d} \right)$, with probability at least $1 - 3\delta$,

$$
\left\| \hat{\Sigma}_n - I_d \right\|_2 \leq 3 \left( 1 + \sqrt{\frac{2 \log \frac{1}{d}}{d}} \right) \frac{\sqrt{d} \max \left( 1, (1 + \epsilon_1) \sqrt{\frac{d}{n}} \right)}{\sqrt{n}}
$$

and

$$
\left\| \hat{\Sigma}_n - I_d \right\|_2 \leq 3 \left( 1 + \sqrt{\frac{8 \log \frac{1}{d}}{d}} \right) \frac{\sqrt{d} \max \left( 1, \frac{8 \log \frac{1}{d}}{d} \right)}{\sqrt{n}}.
$$

\( \square \)

**Proposition 4.** Let $X_1, Y_1, ..., X_n, Y_n \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. Then for any $\epsilon > 0$,

$$
P \left( \left| \frac{1}{n} \sum_{i=1}^{n} X_i Y_i \right| > \epsilon \right) \leq 2 \exp \left\{ - \frac{n \epsilon \min(1, \epsilon)}{10} \right\}.
$$

**Proof.** Let $Z = XY$ where $X, Y \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. Then for any $t$ such that $|t| < 1$,

$$
\mathbb{E} e^{tZ} = \frac{1}{\sqrt{1 - t^2}}.
$$
So for $0 < t < 1$,

$$
P \left( \frac{1}{n} \sum_{i=1}^{n} X_i Y_i > \epsilon \right) = P \left( \exp \left( \sum_{i=1}^{n} tX_i Y_i \right) > \exp(\text{net}) \right)
\leq E \left( \exp \left( \sum_{i=1}^{n} tX_i Y_i \right) \right) \exp(-\text{net})
= (E \exp(tX_i Y_i))^n \exp(-\text{net})
= (1 - t^2)^{-\frac{n}{2}} \exp(-\text{net})
= \exp \left\{ -\frac{n}{2} (2\epsilon t + \log(1 - t^2)) \right\}.
$$

The bound is minimized by $t = \frac{1}{2} \left( \sqrt{1 + 4\epsilon^2} - 1 \right) < 1$, so

$$
P \left( \frac{1}{n} \sum_{i=1}^{n} X_i Y_i > \epsilon \right) \leq \exp \left\{ -\frac{n}{2} h(2\epsilon) \right\}
$$

where

$$
h(u) = \left( \sqrt{1 + u^2} - 1 \right) + \log \left( 1 - \frac{1}{u^2} \left( \sqrt{1 + u^2} - 1 \right)^2 \right).
$$

Since $h(u) \geq \frac{u}{5} \min(1, u)$,

$$
P \left( \frac{1}{n} \sum_{i=1}^{n} X_i Y_i > \epsilon \right) \leq \exp \left\{ -n \frac{2\epsilon}{5} \min(1, 2\epsilon) \right\}
$$

and the proof is complete by noting that the distribution of $X_i Y_i$ is symmetric. \hfill \Box

### 2.2 Davis–Kahan

**Lemma 1.** Let $A, E \in \mathbb{R}^{d \times d}$ be symmetric matrices, and $u \in \mathbb{R}^{d-1}$ such that

$$
u_i = v_{i+1}(A)^T E v_1(A).
$$

If $\lambda_1(A) - \lambda_2(A) > 0$ and

$$
\|E\|_2 \leq \frac{\lambda_1(A) - \lambda_2(A)}{5}
$$

then

$$
\sqrt{1 - (v_1(A)^T v_1(A + E))^2} \leq \frac{4\|u\|}{\lambda_1(A) - \lambda_2(A)}
$$

(Corollary 8.1.11 of Golub and Van Loan (1996)).

### 2.3 Bounding error in estimating the mean

**Proposition 5.** Let $\theta = (\mu_0 - \mu, \mu_0 + \mu)$ for some $\mu_0, \mu \in \mathbb{R}^d$ and $X_1, ..., X_n \overset{\text{iid}}{\sim} P_\theta$. For any $\delta > 0$,

$$
P \left( \|\mu_0 - \hat{\mu}_n\| \geq \sigma \sqrt{\frac{2 \max(d, 8 \log \frac{1}{\delta})}{n}} + \|\mu\| \sqrt{\frac{2 \log \frac{1}{\delta}}{n}} \right) \leq 3\delta.
$$
Proof. Let $Z_1, ..., Z_n \sim \mathcal{N}(0, I)$ and $Y_1, ..., Y_n$ i.i.d. such that $P(Y_i = -1) = P(Y_i = 1) = \frac{1}{2}$. Then for any $\epsilon_1, \epsilon_2 > 0$,

$$
P \left( \| \mu_0 - \hat{\mu}_n \| \geq \sigma \sqrt{\frac{(1 + \epsilon_1)d}{n}} + \| \mu \| \epsilon_2 \right)
$$

$$
= P \left( \| \mu_0 - \frac{1}{n} \sum_{i=1}^{d} \sigma Z_i + \mu_0 + \mu Y_1 \| \geq \sigma \sqrt{\frac{(1 + \epsilon_1)d}{n}} + \| \mu \| \epsilon_2 \right)
$$

$$
= P \left( \| \frac{1}{n} \sum_{i=1}^{d} Z_i + \frac{1}{n} \sum_{i=1}^{d} Y_i \| \geq \sigma \sqrt{\frac{(1 + \epsilon_1)d}{n}} + \| \mu \| \epsilon_2 \right)
$$

$$
\leq P \left( \left\| \frac{1}{n} \sum_{i=1}^{d} Z_i \right\| + \| \mu \| \left\| \frac{1}{n} \sum_{i=1}^{d} Y_i \right\| \geq \sigma \sqrt{\frac{(1 + \epsilon_1)d}{n}} + \| \mu \| \epsilon_2 \right)
$$

$$
\leq \exp \left\{ -\frac{d}{2} \left( \epsilon_1 - \log(1 + \epsilon_1) \right) \right\} + 2 \exp \left\{ -\frac{\epsilon_2^2}{2} \right\}
$$

where the last step is using Hoeffding’s inequality and Proposition $2$. Setting $\epsilon_2 = \sqrt{\frac{2 \log \frac{1}{\delta}}{d}}$, we have

$$
P \left( \| \mu_0 - \hat{\mu}_n \| \geq \sigma \sqrt{\frac{(1 + \epsilon_1)d}{n}} + \| \mu \| \sqrt{\frac{2 \log \frac{1}{\delta}}{d}} \right)
$$

$$
\leq \exp \left\{ -\frac{d}{2} \left( \epsilon_1 - \log(1 + \epsilon_1) \right) \right\} + 2 \delta.
$$

Since $\epsilon_1 - \log(1 + \epsilon_1) \geq \frac{\epsilon_1}{4} \min(1, \epsilon_1)$,

$$
\exp \left\{ -\frac{d}{2} \left( \epsilon_1 - \log(1 + \epsilon_1) \right) \right\} \leq \exp \left\{ -\frac{d}{8} \epsilon_1 \min(1, \epsilon_1) \right\}.
$$

Setting

$$
\epsilon_1 = \sqrt{\frac{8 \log \frac{1}{\delta}}{d} \max \left( 1, \frac{8 \log \frac{1}{\delta}}{d} \right)},
$$

we have

$$
P \left( \| \mu_0 - \hat{\mu}_n \| \geq \sigma \sqrt{\frac{d}{n} \left( 1 + \sqrt{\frac{8 \log \frac{1}{\delta}}{d} \max \left( 1, \frac{8 \log \frac{1}{\delta}}{d} \right)} \right) + \| \mu \| \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}} \right) \leq 3 \delta
$$

and the bound follows. \qed
2.4 Bounding error in estimating principal direction

**Proposition 6.** Let \( \theta = (\mu_0 - \mu, \mu_0 + \mu) \) for some \( \mu_0, \mu \in \mathbb{R}^d \) and \( X_1, \ldots, X_n \sim_{i.i.d.} P_0 \). If \( n \geq d \) then for any \( \delta, \delta_1 > 0 \), with probability at least \( 1 - 5\delta - 2\delta_1 \),

\[
\| \hat{\Sigma}_n - \left( \sigma^2 I_d + \mu \mu^T \right) \|_2 \\
\leq 3\sigma^2 \left( 1 + \sqrt{\frac{2 \log \frac{1}{\delta}}{d}} \right) \sqrt{\frac{d}{n}} \max \left( 1, \left( 1 + \sqrt{\frac{2 \log \frac{1}{\delta}}{d}} \right) \sqrt{\frac{d}{n}} \right) \\
+ \sigma^2 \left( 1 + \sqrt{\frac{8 \log \frac{1}{\delta}}{d}} \max \left( 1, \frac{8 \log \frac{1}{\delta}}{d} \right) \right) \frac{d}{n} \\
+ 4\sigma \| \mu \| \sqrt{\left( 1 + \sqrt{\frac{8 \log \frac{1}{\delta}}{d}} \max \left( 1, \frac{8 \log \frac{1}{\delta}}{d} \right) \right) \frac{d}{n} + \frac{2\| \mu \|^2 \log \frac{1}{\delta_1}}{n}}
\]

where \( \hat{\Sigma}_n \) is the empirical covariance of \( X_i \).

**Proof.** We can express \( X_i \) as \( X_i = \sigma Z_i + \mu Y_i + \mu_0 \) where \( Z_1, \ldots, Z_n \sim_{i.i.d.} \mathcal{N}(0, I_d) \) and \( Y_1, \ldots, Y_n \) i.i.d. such that \( \mathbb{P}(Y_i = -1) = \mathbb{P}(Y_i = 1) = \frac{1}{2} \). Then

\[
\hat{\Sigma}_n - \left( \sigma^2 I_d + \mu \mu^T \right) = \sigma^2 (\hat{\Sigma}_n^Z - I_d) - \mu \mu^T \mathbf{Y}^2 \\
+ \sigma \left( \frac{1}{n} \sum_{i=1}^{n} Y_i Z_i - \mathbf{Y}^T \mathbf{Z} \right) \mu^T \\
+ \sigma \mu \left( \frac{1}{n} \sum_{i=1}^{n} Y_i Z_i - \mathbf{Y}^T \mathbf{Z} \right)^T
\]

where \( \hat{\Sigma}_n^Z \) is the empirical covariance of \( Z_i \) and \( \mathbf{Y} \) and \( \mathbf{Z} \) are the empirical means of \( Y_i \) and \( Z_i \). So

\[
\| \hat{\Sigma}_n - \left( \sigma^2 I_d + \mu \mu^T \right) \|_2 \leq \sigma^2 \| \hat{\Sigma}_n^Z - I_d \|_2 + \| \mu \|^2 \mathbf{Y}^2 \\
+ 2\sigma \| \mu \| \left( \frac{1}{n} \sum_{i=1}^{n} Y_i Z_i \right) + \| \mathbf{Y} \| \| \mathbf{Z} \|
\]

By Hoeffding’s inequality,

\[
\mathbb{P} \left( \| \mu \|^2 \mathbf{Y}^2 \geq \frac{2\| \mu \|^2 \log \frac{1}{\delta_1}}{n} \right) \leq 2\delta_1.
\]

Since \( |\mathbf{Y}| \leq 1 \) and since \( Y_i Z_i \) has the same distribution as \( Z_i \), by Proposition 2 for any \( \epsilon > 0 \),

\[
\mathbb{P} \left( 2\sigma \| \mu \| \left( \frac{1}{n} \sum_{i=1}^{n} Y_i Z_i \right) + \| \mathbf{Y} \| \| \mathbf{Z} \| \right) \geq 4\sigma \| \mu \| \sqrt{\frac{(1 + \epsilon) d}{n}} \\
\leq 2 \exp \left\{ -\frac{d}{2} \left( \epsilon - \log (1 + \epsilon) \right) \right\} \leq 2 \exp \left\{ -\frac{d}{8} \epsilon \min(1, \epsilon) \right\}
\]

Setting

\[
\epsilon = \sqrt{\frac{8 \log \frac{1}{\delta}}{d}} \max \left( 1, \frac{8 \log \frac{1}{\delta}}{d} \right)
\]

we have

\[
\mathbb{P} \left( 2\sigma \| \mu \| \left( \frac{1}{n} \sum_{i=1}^{n} Y_i Z_i \right) + \| \mathbf{Y} \| \| \mathbf{Z} \| \right) \geq 4\sigma \| \mu \| \sqrt{\left( 1 + \sqrt{\frac{8 \log \frac{1}{\delta}}{d}} \max \left( 1, \frac{8 \log \frac{1}{\delta}}{d} \right) \right) \frac{d}{n}} \\
\leq 2\delta.
\]
Finally, by Proposition 3 with probability at least $1 - 3\delta$,

$$\sigma^2 \left\| \bar{\Sigma}_n - I_d \right\|_2 \leq 3\sigma^2 \left(1 + \sqrt{\frac{2 \log \frac{1}{\delta}}{d}} \right) \sqrt{\frac{d}{n}} \max \left(1, \left(1 + \sqrt{\frac{2 \log \frac{1}{\delta}}{d}} \right) \sqrt{\frac{d}{n}} \right)$$

and we complete the proof by combining the three bounds.

**Proposition 7.** Let $\theta = (\mu_0 - \mu, \mu_0 + \mu)$ for some $\mu_0, \mu \in \mathbb{R}^d$ and $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_0$. If $n \geq d > 1$ then for any $0 < \delta \leq \frac{1}{\sqrt{d}}$ and $i \in [2..d]$, with probability at least $1 - 7\delta$,

$$\left| v_i (\sigma^2 I + \mu \mu^T) (\bar{\Sigma}_n - (\sigma^2 I + \mu \mu^T)) v_i (\sigma^2 I + \mu \mu^T) \right| \leq \sigma^2 \frac{1}{2} \sqrt{\left( \frac{10 \log \frac{1}{\delta}}{n} \right) \max \left(1, \left(1 + \frac{10 \log \frac{1}{\delta}}{n} \right) \right) + \sigma \| \mu \| \sqrt{\left( \frac{2 \log \frac{1}{\delta}}{n} \right) + (\sigma^2 + \sigma \| \mu \|) \left( \frac{2 \log \frac{1}{\delta}}{n} \right) } \right).$$

**Proof.** Let $Z_1, W_1, \ldots, Z_n, W_n \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ and $Y_1, \ldots, Y_n$ i.i.d. such that $\mathbb{P}(Y_i = -1) = \mathbb{P}(Y_i = 1) = \frac{1}{2}$. It is easy to see that the quantity of interest is equal in distribution to

$$\left| \frac{1}{n} \sum_{j=1}^{n} (\sigma Z_i - \sigma Z)(\sigma W_i - \sigma W + \| \mu \| Y_i - \| \mu \| Y) \right|$$

where $Z, W, Y$ are the respective empirical means. Moreover,

$$\left| \frac{1}{n} \sum_{j=1}^{n} (\sigma Z_i - \sigma Z)(\sigma W_i - \sigma W + \| \mu \| Y_i - \| \mu \| Y) \right| \leq \sigma^2 \frac{1}{2} \sqrt{\left( \frac{10 \log \frac{1}{\delta}}{n} \right) \max \left(1, \left(1 + \frac{10 \log \frac{1}{\delta}}{n} \right) \right) + \sigma \| \mu \| \left( \frac{10 \log \frac{1}{\delta}}{n} \right) + \sigma \| \mu \| \left( \frac{2 \log \frac{1}{\delta}}{n} \right) \left( \frac{2 \log \frac{1}{\delta}}{n} \right) } \right|.$$
then with probability at least \(1 - 12\delta - 2\exp(-\frac{n}{20})\),
\[
\sqrt{1 - (v_1(\sigma^2 I + \mu^T)Tv_1(\Sigma_n))^2} \leq 14 \max \left( \frac{\sigma^2}{\|\mu\|^2}, \frac{\sigma}{\|\mu\|} \right) \sqrt{d} \frac{10\log \frac{d}{\sigma^2}}{n} \max \left(1, \frac{10\log \frac{d}{\sigma}}{n} \right).
\]

**Proof.** By Proposition 6 (with \(\delta_1 = \exp(-\frac{n}{20})\)), Proposition 7 (with \(\delta_2 = \frac{2}{\sigma^2}\)), and Lemma 1, with probability at least \(1 - 12\delta - 2\exp(-\frac{n}{20})\),
\[
\sqrt{1 - (v_1(\sigma^2 I + \mu^T)Tv_1(\Sigma_n))^2} \leq 4\sqrt{d - 1} \left( \frac{\sigma^2}{\|\mu\|^2} \frac{10\log \frac{d}{\sigma^2}}{n} \max \left(1, \frac{10\log \frac{d}{\sigma}}{n} \right) + \sigma\|\mu\| \sqrt{\frac{2\log \frac{d}{\sigma}}{n} + (\sigma^2 + \sigma\|\mu\|) \frac{2\log \frac{d}{\sigma}}{n}} \right)
\]
and the result follows after some simplifications.

\(\square\)

### 2.5 General result relating error in estimating mean and principal direction to clustering loss

**Proposition 9.** Let \(\theta = (\mu_0 - \mu, \mu_0 + \mu)\) and let
\[
\hat{F}(x) = \begin{cases} 1 & \text{if } x^Tv \geq x_0^Tv \\ 2 & \text{otherwise} \end{cases}
\]
for some \(x_0, v \in \mathbb{R}^d\), with \(\|v\| = 1\). Define \(\cos \beta = |v^T\mu|/\|\mu\|\). If \(|(x_0 - \mu_0)^Tv| \leq \sigma\varepsilon_1 + \|\mu\|\varepsilon_2\) for some \(\varepsilon_1 \geq 0\) and \(0 \leq \varepsilon_2 \leq \frac{1}{4}\), and if \(\sin \beta \leq \frac{1}{\sqrt{2}}\), then
\[
L_\theta(\hat{F}) \leq \exp \left\{ -\frac{1}{2} \max \left(0, \frac{\|\mu\|}{2\sigma} - 2\varepsilon_1 \right)^2 \right\} \left[ 2\varepsilon_1 + \varepsilon_2 \frac{\|\mu\|}{\sigma} + 2\sin \beta \left( 2\sin \frac{\|\mu\|}{\sigma} + 1 \right) \right].
\]

**Proof.**
\[
L_\theta(\hat{F}) = \min_x P_\theta(\{x : F_\theta(x) \neq \pi(\hat{F}(x))\})
\]
\[
= \min \{ P_\theta(\{x : ((x - \mu_0)^T\mu)((x - x_0)^Tv) \geq 0\}), \]
\[
P_\theta(\{x : ((x - \mu_0)^T\mu)((x - x_0)^Tv) \leq 0\}) \}.
\]
WLOG assume \(v^T\mu \geq 0\) (otherwise we can simply replace \(v\) with \(-v\), which does not affect the bound). Then
\[
L_\theta(\hat{F}) = P_\theta(\{x : ((x - \mu_0)^T\mu)((x - x_0)^Tv) \leq 0\})
\]
\[
= P_\theta(\{x : ((x - \mu_0)^T\mu)((x - \mu_0)^Tv - (x_0 - \mu_0)^Tv) \leq 0\})
\]
\[
= P_\theta \left( \left\{ x : \left( \frac{(x - \mu_0)^T\mu}{\|\mu\|} \right)((x - \mu_0)^Tv - (x_0 - \mu_0)^Tv) \leq 0 \right\} \right).
\]
Define
\[
\tilde{\mu} = \frac{\mu}{\|\mu\|},
\]
\[
\tilde{x} = (x - \mu_0)^T\tilde{\mu},
\]
and
\[
\tilde{y} = (x - \mu_0)^Tv - \tilde{\mu}\tilde{\mu}^Tv
\]
so that
\[
L_\theta(\hat{F}) = P_\theta \left( \{ x : \tilde{x} \tilde{y}\sin \beta + \tilde{x}\cos \beta - (x_0 - \mu_0)^Tv \leq 0 \} \right)
\]
\[
= P_\theta \left( \{ x : \min(0, B(\tilde{y})) \leq \tilde{x} \leq \max(0, B(\tilde{y})) \} \right).
\]

8
where
\[ B(\bar{y}) = \frac{(x_0 - \mu_0)^T v}{\cos \beta} - \bar{y} \tan \beta. \]

Since \( \bar{x} \) and \( \bar{y} \) are projections of \( x - \mu_0 \) onto orthogonal unit vectors, and since \( \bar{x} \) is exactly the component of \( x - \mu_0 \) that lies in the direction of \( \mu \), we can integrate out all other directions and obtain
\[
L_0(\tilde{F}) = \int_{-\infty}^{\infty} \int_{\min(0, B(\bar{y}))}^{\max(0, B(\bar{y}))} \left( \frac{1}{2} \phi_\sigma(\bar{x} + \|\mu\|) + \frac{1}{2} \phi_\sigma(\bar{x} - \|\mu\|) \right) d\bar{x} d\bar{y}
\]
where \( \phi_\sigma \) is the density of \( \mathcal{N}(0, \sigma^2) \). But,
\[
\int_{\min(0, B(\bar{y}))}^{\max(0, B(\bar{y}))} \left( \frac{1}{2} \phi_\sigma(\bar{x} + \|\mu\|) + \frac{1}{2} \phi_\sigma(\bar{x} - \|\mu\|) \right) d\bar{x} = \frac{1}{2} \left( \Phi \left( \frac{\max(0, B(\bar{y})) + \|\mu\|}{\sigma} \right) - \Phi \left( \frac{\min(0, B(\bar{y})) + \|\mu\|}{\sigma} \right) \right)
\]
\[
+ \frac{1}{2} \left( -\Phi \left( \frac{\max(0, B(\bar{y})) + \|\mu\|}{\sigma} \right) + \Phi \left( -\min(0, B(\bar{y})) + \|\mu\| \right) \right)
\]
\[ = \frac{1}{2} \left( \Phi \left( \frac{\|\mu\| + B(\bar{y})}{\sigma} \right) - \Phi \left( \frac{\|\mu\| - B(\bar{y})}{\sigma} \right) \right). \]

Since the above quantity is increasing in \( |B(\bar{y})| \), and since \( |B(\bar{y})| \leq |\bar{y}| \tan \beta + r \) where
\[ r = \left| \frac{(x_0 - \mu_0)^T v}{\cos \beta} \right|, \]
we have that, replacing \( \bar{y} \) by \( x \),
\[
L_0(\tilde{F}) \leq \frac{1}{2} \int_{-\infty}^{\infty} \frac{1}{\sigma} \phi \left( \frac{x}{\sigma} \right) \left[ \Phi \left( \frac{\|\mu\|}{\sigma} + \frac{|x| \tan \beta + r}{\sigma} \right) - \Phi \left( \frac{\|\mu\|}{\sigma} - \frac{|x| \tan \beta - r}{\sigma} \right) \right] dx
\]
\[ \leq \int_{-\infty}^{\infty} \frac{1}{\sigma} \phi \left( \frac{x}{\sigma} \right) \left[ \Phi \left( \frac{\|\mu\|}{\sigma} \right) - \Phi \left( \frac{\|\mu\| - |x| \tan \beta - r}{\sigma} \right) \right] dx
\]
\[ = \int_{-\infty}^{\infty} \phi(x) \left[ \Phi \left( \frac{\|\mu\| - r}{\sigma} \right) - \Phi \left( \frac{\|\mu\| - r}{\sigma} - \frac{|x| \tan \beta}{\sigma} \right) \right] dx
\]
\[ + \Phi \left( \frac{\|\mu\|}{\sigma} \right) - \Phi \left( \frac{\|\mu\|}{\sigma} - \frac{r}{\sigma} \right) \right] . \]

Since \( \tan \beta \leq \frac{1}{2} \), we have that \( r \leq 2(|x_0 - \mu_0|^T v| \leq 2 \sigma \epsilon_1 + 2 \|\mu\| \epsilon_2 \) and
\[
\Phi \left( \frac{\|\mu\|}{\sigma} - \frac{r}{\sigma} \right) \geq \frac{r}{\sigma} \phi \left( \max \left( 0, \frac{\|\mu\| - r}{\sigma} \right) \right)
\]
\[ \leq \left( 2 \epsilon_1 + 2 \epsilon_2 \frac{\|\mu\|}{\sigma} \right) \phi \left( \max \left( 0, \frac{\|\mu\|}{\sigma} - 2 \epsilon_2 \right) \right), \]
and since \( \epsilon_2 \leq \frac{1}{4} \),
\[
\Phi \left( \frac{\|\mu\|}{\sigma} - \frac{r}{\sigma} \right) \leq 2 \left( \epsilon_1 + \epsilon_2 \frac{\|\mu\|}{\sigma} \right) \phi \left( \max \left( 0, \frac{\|\mu\|}{2 \sigma} - 2 \epsilon_1 \right) \right). \]
Defining \( A = \left\lfloor \frac{\|\mu\|-r}{\sigma} \right\rfloor \),
\[
\int_{-\infty}^{\infty} \phi(x) \left[ \Phi \left( \frac{\|\mu\|-r}{\sigma} \right) - \Phi \left( \frac{\|\mu\|-r}{\sigma} - |x| \tan \beta \right) \right] \, dx
\]
\[
\leq 2 \int_{0}^{\infty} \int_{A-x \tan \beta}^{A} \phi(x) \phi(y) \, dy \, dx = 2 \int_{-\infty}^{\infty} A \cos \beta + (x + A \sin \beta) \tan \beta \phi(x) \phi(y) \, dy \, dx
\]
\[
\leq 2 \phi(A \cos \beta) \tan \beta \int_{-A \sin \beta}^{\infty} (x + A \sin \beta) \phi(x) \, dx
\]
\[
= 2 \phi(A \cos \beta) \tan \beta \left( A \sin \beta \Phi(A \sin \beta) + \phi(A \sin \beta) \right)
\]
\[
\leq 2 \phi(A \cos \beta + (A \sin \beta + 1) \tan \beta)
\]
\[
\leq 2 \phi \left( \max \left( 0, \frac{\|\mu\| - r}{\sigma} \right) \right) \tan \beta \left( \frac{\|\mu\| + r}{\sigma} \sin \beta + 1 \right)
\]
and
\[
\int_{-\infty}^{\infty} \phi(x) \left[ \Phi \left( \frac{\|\mu\|-r}{\sigma} \right) - \Phi \left( \frac{\|\mu\|-r}{\sigma} - |x| \tan \beta \right) \right] \, dx
\]
\[
\leq 2 \phi \left( \max \left( 0, \frac{\|\mu\| - 2\epsilon_1}{2\sigma} \right) \right) \tan \beta \left( \frac{2 \|\mu\|}{\sigma} + 2\epsilon_1 \right) \sin \beta + 1ight).
\]
So we have that
\[
L_{\theta}(\hat{F}) \leq 2 \left( \epsilon_1 + \epsilon_2 \frac{\|\mu\|}{2\sigma} \right) \phi \left( \max \left( 0, \frac{\|\mu\|}{2\sigma} - 2\epsilon_1 \right) \right)
\]
\[
+ 2 \phi \left( \max \left( 0, \frac{\|\mu\|}{2\sigma} - 2\epsilon_1 \right) \right) \tan \beta \left( \frac{2 \|\mu\|}{\sigma} + 2\epsilon_1 \right) \sin \beta + 1
\]
\[
\leq \phi \left( \max \left( 0, \frac{\|\mu\|}{2\sigma} - 2\epsilon_1 \right) \right) \times
\]
\[
\left[ 2\epsilon_1 + 2\epsilon_2 \frac{\|\mu\|}{\sigma} + 4 \sin \beta \tan \beta \frac{\|\mu\|}{\sigma} + 4\epsilon_1 \sin \beta \tan \beta + 2 \tan \beta \right]
\]
\[
\leq \exp \left\{ -\frac{1}{2} \max \left( 0, \frac{\|\mu\|}{2\sigma} - 2\epsilon_1 \right)^2 \right\} \left[ 2\epsilon_1 + \epsilon_2 \frac{\|\mu\|}{\sigma} + \tan \beta \left( 2 \sin \beta \frac{\|\mu\|}{\sigma} + 1 \right) \right].
\]
\[
\square
\]

2.6 Non-sparse upper bound

**Theorem 1.** For any \( \theta \in \Theta_{\lambda} \) and \( X_1, ..., X_n \overset{i.i.d.}{\sim} P_0 \), let
\[
\hat{F}(x) = \begin{cases} 
1 & \text{if } x^T v_1(\bar{S}_n) \geq \hat{\mu}_n^T v_1(\bar{S}_n) \\
\frac{1}{2} & \text{otherwise},
\end{cases}
\]
and let \( n \geq \max(68, 4d) \), \( d \geq 1 \).

Then
\[
\sup_{\theta \in \Theta_{\lambda}} \mathbb{E} L_{\theta}(\hat{F}) \leq 600 \max \left( \frac{4\sigma^2}{\lambda^2}, 1 \right) \sqrt{\frac{d \log(nd)}{n}}.
\]

Furthermore, if \( \lambda \geq 2 \max(80, 14\sqrt{5d}) \), then
\[
\sup_{\theta \in \Theta_{\lambda}} \mathbb{E} L_{\theta}(\hat{F}) \leq 17 \exp \left( -\frac{n}{32} \right) + 9 \exp \left( -\frac{\lambda^2}{80\sigma^2} \right).
\]
Proof. Using Propositions 5 and 8 with \( \delta = \frac{1}{\sqrt{n}} \), Proposition 2 and the fact that \((C + x) \exp(-\max(0, x - 4)^2/8) \leq (C + 6) \exp(-\max(0, x - 4)^2/10)\) for all \( C, x > 0 \),

\[
\mathbb{E}L_\theta(\hat{F}) \leq 600 \max \left( \frac{4\sigma^2}{\lambda^2}, 1 \right) \sqrt{\frac{d \log(n d)}{n}}
\]

(it is easy to verify that the bounds are decreasing with \(|\mu|| \), so we use \(|\mu| = \frac{1}{2} \) to bound the supremum). Note that the \( d = 1 \) case must be handled separately, but results in a bound that agrees with the above.

Also, when \( \hat{\delta} \geq 2 \max(80, 14\sqrt{5d}) \), using \( \delta = \exp \left( -\frac{n}{32} \right) \),

\[
\mathbb{E}L_\theta(\hat{F}) \leq 17 \exp \left( -\frac{n}{32} \right) + 9 \exp \left( -\frac{\lambda^2}{80\sigma^2} \right).
\]

\( \Box \)

2.7 Estimating the support in the sparse case

Proposition 10. Let \( \theta = (\mu_0 - \mu, \mu_0 + \mu) \) for some \( \mu_0, \mu \in \mathbb{R}^d \) and \( X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\theta \). For any \( 0 < \delta < \frac{1}{\sqrt{n}} \) such that \( \sqrt{\frac{6 \log \frac{\delta}{n}}{n}} \leq \frac{1}{2} \), with probability at least \( 1 - 6\delta \),

\[
|\hat{\Sigma}_n(i, i) - (\sigma^2 + \mu(i)^2)| \leq \sigma^2 \sqrt{\frac{6 \log \frac{1}{3}}{n}} + 2\sigma|\mu(i)| \sqrt{\frac{2 \log \frac{1}{3}}{n} + \sigma + |\mu(i)|^2} \frac{2 \log \frac{1}{3}}{n}
\]

for all \( i \in [d] \).

Proof. Consider any \( i \in [d] \). Let \( Z_1, \ldots, Z_n \overset{i.i.d.}{\sim} \mathcal{N}(0, 1) \) and \( Y_1, \ldots, Y_n \) i.i.d. such that \( \mathbb{P}(Y_j = -1) = \mathbb{P}(Y_j = 1) = \frac{1}{2} \). Then \( \hat{\Sigma}_n(i, i) \) is equal in distribution to

\[
\frac{1}{n} \sum_{j=1}^{n} (\sigma Z_j + \mu(i) Y_j - \sigma Z - \mu(i) Y)^2
\]

where \( Z \) and \( Y \) are the respective empirical means, and

\[
\frac{1}{n} \sum_{j=1}^{n} (\sigma Z_j + \mu(i) Y_j - \sigma Z - \mu(i) Y)^2 = \frac{1}{n} \sum_{j=1}^{n} (\sigma Z_j + \mu(i) Y_j)^2 - (\sigma Z + \mu(i) Y)^2
\]

\[
= \sigma^2 \frac{1}{n} \sum_{j=1}^{n} Z_j^2 + \mu(i)^2 + 2\sigma \mu(i) \frac{1}{n} \sum_{j=1}^{n} Z_j Y_j
\]

\[
- \sigma^2 Z^2 - \mu(i)^2 Y^2 - 2\sigma \mu(i) Z Y
\]

So, by Hoeffding’s inequality, a Gaussian tail bound, and Proposition 11, we have that for any \( 0 < \delta < \frac{1}{\sqrt{n}} \), with probability at least \( 1 - 6\delta \),

\[
|\hat{\Sigma}_n(i, i) - (\sigma^2 + \mu(i)^2)| \leq \sigma^2 \sqrt{\frac{6 \log \frac{1}{3}}{n}} + 2\sigma|\mu(i)| \sqrt{\frac{2 \log \frac{1}{3}}{n} + \sigma + |\mu(i)|^2} \frac{2 \log \frac{1}{3}}{n}
\]

where we have used the fact that for \( \epsilon \in (0, 0.5] \),

\[
\max \{-\epsilon + \log(1 + \epsilon), \epsilon + \log(1 - \epsilon)\} \leq -\frac{\epsilon^2}{3}
\]

and the result follows easily. \( \Box \)
Proposition 11. Let \( \theta = (\mu_0 - \mu, \mu_0 + \mu) \) for some \( \mu_0, \mu \in \mathbb{R}^d \) and \( X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\theta \). Define
\[
S(\theta) = \{ i \in [d] : \mu(i) \neq 0 \},
\]
\[
\alpha = \sqrt{\frac{6 \log(nd)}{n} + \frac{2 \log(nd)}{n}},
\]
\[
\hat{S}(\theta) = \{ i \in [d] : |\mu(i)| \geq 4\sigma\sqrt{\alpha} \},
\]
\[
\hat{\tau}_n = \frac{1 + \alpha}{1 - \alpha} \min_{i \in [d]} \hat{\Sigma}_n(i, i),
\]
and
\[
\hat{S}_n = \{ i \in [d] : \hat{\Sigma}_n(i, i) > \hat{\tau}_n \}.
\]

Assume that \( n \geq 1, d \geq 2, \) and \( \alpha \leq \frac{1}{4} \). Then \( \hat{S}(\theta) \subseteq \hat{S}_n \subseteq S(\theta) \) with probability at least \( 1 - \frac{6}{n} \).

Proof. By Proposition 10 with probability at least \( 1 - \frac{6}{n} \),
\[
|\hat{\Sigma}_n(i, i) - (\sigma^2 + \mu(i)^2)| \leq \sigma^2 \sqrt{\frac{6 \log(nd)}{n} + 2\sigma|\mu(i)|} \sqrt{\frac{2 \log(nd)}{n}} + (\sigma + |\mu(i)|)^2 \frac{2 \log(nd)}{n}
\]
for all \( i \in [d] \). Assume the above event holds. If \( \hat{S}(\theta) = \emptyset \), then of course \( \hat{S}_n \subseteq S(\theta) \). Otherwise, for \( i \not\in \hat{S}(\theta) \),
\[
(1 - \alpha)\sigma^2 \leq \hat{\Sigma}_n(i, i) \leq (1 + \alpha)\sigma^2
\]
so it is clear that \( \hat{S}_n \subseteq S(\theta) \).

The remainder of the proof is trivial if \( \hat{S}(\theta) = \emptyset \) or \( S(\theta) = \emptyset \). Assume otherwise. For any \( i \in S(\theta) \),
\[
\hat{\Sigma}_n(i, i) \geq (1 - \alpha)\sigma^2 + \mu(i)^2 - 2\sigma|\mu(i)| \sqrt{\frac{2 \log(nd)}{n}} - \mu(i)^2 \frac{2 \log(nd)}{n} - \mu(i)^2 \frac{2 \log(nd)}{n}
\]
\[
\geq (1 - \alpha)\sigma^2 + \left( 1 - \frac{2 \log(nd)}{n} \right) \mu(i)^2 - 2\alpha \sigma|\mu(i)|.
\]

By definition, \( |\mu(i)| \geq 4\sigma\sqrt{\alpha} \) for all \( i \in \hat{S}(\theta) \), so
\[
\frac{(1 + \alpha)^2}{1 - \alpha} \sigma^2 \leq (1 - \alpha)\sigma^2 + \left( 1 - \frac{2 \log(nd)}{n} \right) \mu(i)^2 - 2\alpha \sigma|\mu(i)| \leq \hat{\Sigma}_n(i, i)
\]
and \( i \in \hat{S}_n \) (we ignore strict equality above as a measure 0 event), i.e. \( \hat{S}(\theta) \subseteq \hat{S}_n \), which concludes the proof. \( \square \)

2.8 Sparse upper bound

Theorem 2. For any \( \theta = (\mu_0 - \mu, \mu_0 + \mu) \in \Theta_{\lambda,s} \) and \( X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\theta \) with \( n \geq \max(68, 4s) \) and \( s \geq 1 \), define
\[
\alpha = \sqrt{\frac{6 \log(nd)}{n} + \frac{2 \log(nd)}{n}},
\]
\[
\hat{\tau}_n = \frac{1 + \alpha}{1 - \alpha} \min_{i \in [d]} \hat{\Sigma}_n(i, i),
\]
and
\[
\hat{S}_n = \{ i \in [d] : \hat{\Sigma}_n(i, i) > \hat{\tau}_n \}.
\]

Assume that \( d \geq 2 \), and \( \alpha \leq \frac{1}{4} \). Let
\[
\hat{F}_n(x) = \begin{cases} 1 & \text{if } x^T \hat{\Sigma}_n v_1(\hat{\Sigma}_n) \geq \hat{\mu}_n v_1(\hat{\Sigma}_n) \\ 2 & \text{otherwise} \end{cases}
\]
where \( \hat{\mu}_n \) and \( \hat{\Sigma}_n \) are the empirical mean and covariance of \( X \), for the dimensions in \( \hat{S}_n \), and 0 elsewhere. Then
\[
\sup_{\theta \in \Theta_{\lambda,s}} \mathbb{E}L_\theta(\hat{F}) \leq 603 \max \left( \frac{16\sigma^2}{\lambda^2}, 1 \right) \sqrt{\frac{s \log(ns)}{n} + 220 \frac{\sigma\sqrt{s}}{\lambda} \left( \frac{\log(nd)}{n} \right)^2}.
\]
Proof. Define
\[ S(\theta) = \{ i \in [d] : \mu(i) \neq 0 \} \]
and
\[ \tilde{S}(\theta) = \{ i \in [d] : |\mu(i)| \geq 4\sigma\sqrt{\alpha} \}, \]
Assume \( \tilde{S}(\theta) \subseteq \tilde{S}_n \subseteq S(\theta) \) (by Proposition 11, this holds with probability at least \( 1 - \frac{6}{n} \)). If \( \tilde{S}(\theta) = \emptyset \), then we simply have \( \mathbb{E} L_{\theta}(\hat{F}_n) \leq \frac{1}{2} \).

Assume \( \tilde{S}(\theta) \neq \emptyset \). Let
\[ \cos \hat{\beta} = |v_1(\hat{\Sigma}_{\tilde{S}_n})^Tv_1(\Sigma)|, \]
\[ \cos \tilde{\beta} = |v_1(\Sigma_{\tilde{S}_n})^Tv_1(\Sigma)|, \]
and
\[ \cos \beta = |v_1(\Sigma_{\hat{S}_n})^Tv_1(\Sigma_{\hat{S}_n})| \]
where \( \Sigma = \sigma^2 I + \mu\mu^T \), and \( \Sigma_{\tilde{S}_n} \) is the same as \( \hat{\Sigma}_n \), and 0 elsewhere. Then
\[ \sin \hat{\beta} \leq \sin \tilde{\beta} + \sin \beta. \]

Also
\[ \sin \tilde{\beta} = \frac{\|\mu - \mu_{\tilde{S}(\theta)}\|}{\|\mu\|} \leq \frac{\|\mu - \mu_{\tilde{S}(\theta)}\|}{\|\mu\|} \leq \frac{4\sigma\sqrt{\alpha}}{\|\mu\|} \leq 8\sigma\sqrt{s\alpha} \lambda. \]

Using the same argument as the proof of Theorem 1, we have that as long as the above bound is smaller than \( \frac{1}{2\sqrt{\pi}} \),
\[ \mathbb{E} L_{\theta}(\hat{F}) \leq 600 \max \left( \frac{\sigma^2}{\lambda^2 - 4\sigma\sqrt{s\alpha}}, 1 \right) \sqrt{s \log(ns) \over n} + 104 \frac{\sigma\sqrt{s\alpha}}{\lambda} + 3 \over n \]
\[ \leq 603 \max \left( \frac{16\sigma^2}{\lambda^2}, 1 \right) \sqrt{s \log(ns) \over n} + 104 \frac{\sigma\sqrt{s\alpha}}{\lambda}. \]

However, when \( 8\sigma\sqrt{s\alpha} \lambda > \frac{1}{2\sqrt{\pi}} \), the above bound is bigger than \( \frac{1}{2} \), which is a trivial upper bound on the clustering error, hence the bound can be stated without further conditions. Finally, since \( \alpha \leq \frac{1}{2} \), we must have \( \log(nsd) \over n \leq 1 \), so \( \alpha \leq (\sqrt{6} + 2) \sqrt{\log(nsd) \over n} \), which completes the proof.

\[ \square \]

3 Lower bounds

3.1 Standard tools

Lemma 2. Let \( P_0, P_1, \ldots, P_M \) be probability measures satisfying
\[ \frac{1}{M} \sum_{i=1}^{M} \text{KL}(P_i, P_0) \leq \alpha \log M \]
where \( 0 < \alpha < 1/8 \) and \( M \geq 2 \). Then
\[ \inf_{\psi} \max_{i \in [0..M]} P_i(\psi \neq i) \geq 0.07 \]
\[ (\text{Tsybakov}(2009)). \]
Lemma 3. (Varshamov–Gilbert bound) Let $\Omega = \{0, 1\}^m$ for $m \geq 8$. Then there exists a subset 
\{$\omega_0, ..., \omega_M$\} $\subseteq \Omega$ such that $\omega_0 = (0, ..., 0)$,
\[
\rho(\omega_i, \omega_j) \geq \frac{m}{8}, \quad \forall 0 \leq i < j \leq M,
\]
and
\[
M \geq 2^{m/8},
\]
where $\rho$ denotes the Hamming distance between two vectors [Tsybakov (2009)].

Lemma 4. Let $\Omega = \{\omega \in \{0, 1\}^m : \|\omega\|_0 = s\}$ for integers $m > s \geq 1$. For any $\alpha, \beta \in (0, 1)$ such that $s \leq \alpha \beta m$, there exists $\omega_0, ..., \omega_M \in \Omega$ such that for all $0 \leq i < j \leq M$,
\[
\rho(\omega_i, \omega_j) > 2(1 - \alpha)s
\]
and
\[
\log(M + 1) \geq cs \log \left(\frac{m}{s}\right)
\]
where
\[
c = \frac{\alpha}{-\log(\alpha \beta)}(-\log \beta + \beta - 1).
\]
In particular, setting $\alpha = 3/4$ and $\beta = 1/3$, we have that $\rho(\omega_i, \omega_j) > s/2$, $\log(M + 1) \geq \frac{s}{2} \log \left(\frac{m}{s}\right)$, as long as $s \leq m/4$ [Massart (2007), Lemma 4.10].

3.2 A reduction to hypothesis testing without a general triangle inequality

Proposition 12. Let $\theta_1, ..., \theta_M \in \Theta_{\alpha}$ (or $\Theta_{\alpha,s}$), $M \geq 2$, $0 < \alpha < 1/8$, and $\gamma > 0$. If
\[
\max_{i \in [M]} \text{KL}(P_{\theta_i}, P_{\theta_0}) \leq \frac{\alpha \log M}{n}
\]
and for all $0 \leq i \neq j \leq M$ and clusterings $\hat{F}$,
\[
L_{\theta_i}(\hat{F}) < \gamma \implies L_{\theta_j}(\hat{F}) \geq \gamma,
\]
then
\[
\inf_{\hat{F}_n} \max_{i \in [0..M]} \mathbb{E}_{\theta_i} L_{\theta_i}(\hat{F}_n) \geq 0.07 \gamma.
\]

Proof. Using Markov’s inequality,
\[
\inf_{\hat{F}_n} \max_{i \in [0..M]} \mathbb{E}_{\theta_i} L_{\theta_i}(\hat{F}_n) \geq \gamma \inf_{\hat{F}_n} \max_{i \in [0..M]} P_{\theta_i}^{n} \left(L_{\theta_i}(\hat{F}_n) \geq \gamma\right).
\]
Define $\psi^*(\hat{F}_n) = \arg\min_{i \in [0..M]} L_{\theta_i}(\hat{F}_n)$. By assumption, $L_{\theta_i}(\hat{F}_n) < \gamma$ implies $L_{\theta_j}(\hat{F}_n) \geq \gamma$ for any $j \neq i$, so $L_{\theta_i}(\hat{F}_n) < \gamma$ only when $\psi^*(\hat{F}_n) = i$. Hence,
\[
P_{\theta_i}^{n} \left(\psi^*(\hat{F}_n) = i\right) \geq P_{\theta_i}^{n} \left(L_{\theta_i}(\hat{F}_n) < \gamma\right)
\]
and
\[
\inf_{\hat{F}_n} \max_{i \in [0..M]} P_{\theta_i}^{n} \left(L_{\theta_i}(\hat{F}_n) \geq \gamma\right) \geq \max_{i \in [0..M]} P_{\theta_i}^{n} \left(\psi^*(\hat{F}_n) \neq i\right)
\]
\[
\geq \inf_{\hat{F}_n} \max_{i \in [0..M]} P_{\theta_i}^{n} \left(\hat{F}_n \neq i\right)
\]
\[
\geq 0.07
\]
where the last step is by Lemma 2. 

□
3.3 Properties of the clustering error

**Proposition 13.** For any \( \theta, \theta' \in \Theta_\lambda \), and any clustering \( \widehat{F} \), if

\[
L_0(F_\theta') + L_0(F_\theta) + \sqrt{\frac{\text{KL}(P_\theta, P_{\theta'})}{2}} \leq \frac{1}{2},
\]

then

\[
L_0(F_\theta') - L_0(F_\theta) - \sqrt{\frac{\text{KL}(P_\theta, P_{\theta'})}{2}} \leq L_0(F_\theta') - L_0(\widehat{F}) \leq L_0(F_\theta') + L_0(F_\theta) + \sqrt{\frac{\text{KL}(P_\theta, P_{\theta'})}{2}}.
\]

**Proof.** WLOG assume \( F_\theta, F_{\theta'} \), and \( \widehat{F} \) are such that, using simplified notation,

\[
L_0(F_\theta') = P_0(F_\theta \neq F_{\theta'})
\]

and

\[
L_0(\widehat{F}) = P_0(F_\theta \neq \widehat{F}).
\]

Then

\[
P_0(F_{\theta'} \neq \widehat{F}) = P_0 \left( (F_\theta = F_{\theta'}) \cap (F_\theta \neq \widehat{F}) \cup (F_\theta \neq F_{\theta'}) \cap (F_\theta = \widehat{F}) \right)
\]

\[
= P_0 \left( (F_\theta = F_{\theta'}) \cap (F_\theta \neq \widehat{F}) \right) + P_0 \left( (F_\theta \neq F_{\theta'}) \cap (F_\theta = \widehat{F}) \right).
\]

Since

\[
0 \leq P_0 \left( (F_\theta = F_{\theta'}) \cap (F_\theta \neq \widehat{F}) \right) \leq P_0 \left( F_\theta \neq \widehat{F} \right) = L_0(\widehat{F}),
\]

\[
P_0 \left( (F_\theta \neq F_{\theta'}) \cap (F_\theta = \widehat{F}) \right) \leq P_0 \left( F_\theta \neq F_{\theta'} \right) = L_0(F_{\theta'}),
\]

and

\[
L_0(F_{\theta'}) - L_0(\widehat{F}) = P_0 \left( F_\theta \neq F_{\theta'} \right) - P_0 \left( F_\theta \neq \widehat{F} \right) \leq P_0 \left( (F_\theta \neq F_{\theta'}) \cap (F_\theta = \widehat{F}) \right),
\]

we have that

\[
L_0(F_{\theta'}) - L_0(\widehat{F}) \leq P_0(F_{\theta'} \neq \widehat{F}) \leq L_0(F_{\theta'}) + L_0(\widehat{F})
\]

and

\[
L_0(F_{\theta'}) - L_0(\widehat{F}) - \text{TV}(P_\theta, P_{\theta'}) \leq P_0(F_{\theta'} \neq \widehat{F}) \leq L_0(F_{\theta'}) + L_0(\widehat{F}) + \text{TV}(P_\theta, P_{\theta'}).\]

It is easy to see that if \( L_0(F_{\theta'}) + L_0(\widehat{F}) + \text{TV}(P_\theta, P_{\theta'}) \leq \frac{1}{2} \), then the above bound implies

\[
L_0(F_{\theta'}) - L_0(\widehat{F}) - \text{TV}(P_\theta, P_{\theta'}) \leq L_0(\widehat{F}) \leq L_0(F_{\theta'}) + L_0(\widehat{F}) + \text{TV}(P_\theta, P_{\theta'}).
\]

The final step is to use the fact that \( \text{TV}(P_\theta, P_{\theta'}) \leq \sqrt{\frac{\text{KL}(P_\theta, P_{\theta'})}{2}} \). \( \Box \)

**Proposition 14.** For some \( \mu_0, \mu, \mu' \in \mathbb{R}^d \) such that \( \| \mu \| = \| \mu' \| \), let

\[
\theta = \left( \mu_0 - \frac{\mu}{2}, \mu_0 + \frac{\mu}{2} \right)
\]

and

\[
\theta' = \left( \mu_0 - \frac{\mu'}{2}, \mu_0 + \frac{\mu'}{2} \right).
\]

Then

\[
2g \left( \frac{\| \mu \|}{2\sigma} \right) \sin \beta \cos \beta \leq L_0(F_{\theta'}) \leq \frac{1}{\pi} \tan \beta
\]

where \( \cos \beta = \frac{\mu^T \mu'}{\| \mu \|^2} \) and \( g(x) = \phi(x)(\phi(x) - x\Phi(-x)) \).
Proof. It is easy to see that

\[ L_\theta(F_{\theta'}) = \frac{1}{2} \int_R \frac{1}{\sigma} \phi \left( \frac{x}{\sigma} \right) \left( \Phi \left( \frac{\|\mu\|}{2\sigma} + \frac{|x| \tan \beta}{\sigma} \right) - \Phi \left( \frac{\|\mu\|}{2\sigma} - \frac{|x| \tan \beta}{\sigma} \right) \right) dx. \]

Define \( \xi = \|\mu\|_2 \sigma \). With a change of variables, we have

\[ L_\theta(F_{\theta'}) = \frac{1}{2} \int_R \phi(x) (\Phi(\xi + |x| \tan \beta) - \Phi(\xi - |x| \tan \beta)) dx = \int_0^\infty \phi(x)(\Phi(\xi + x \tan \beta) - \Phi(\xi - x \tan \beta))dx. \]

For any \( a \leq b \), \( \Phi(b) - \Phi(a) \leq \frac{b-a}{\sqrt{2\pi}} \), so

\[ L_\theta(F_{\theta'}) = \int_0^\infty \phi(x)(\Phi(\xi + x \tan \beta) - \Phi(\xi - x \tan \beta))dx \leq \tan \beta \int_0^\infty x \phi(x)dx = \frac{1}{\pi} \tan \beta. \]

Also,

\[ L_\theta(F_{\theta'}) = \int_0^\infty \phi(x)(\Phi(\xi + x \tan \beta) - \Phi(\xi - x \tan \beta))dx \geq 2 \tan \beta \int_0^\infty x \phi(x)\phi(\xi + x \tan \beta)dx \]

\[ = 2 \tan \beta \frac{1}{\sqrt{2\pi}} \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{x^2 + (\xi + x \tan \beta)^2}{2} \right\} dx \]

\[ = 2 \tan \beta \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{\xi^2}{2} \left( 1 - \frac{\tan^2 \beta}{1 + \tan^2 \beta} \right) \right\} \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(x + \xi \tan \beta)^2}{2 \left( \frac{1}{\sqrt{1 + \tan^2 \beta}} \right)^2} \right\} dx \]

\[ \geq 2 \tan \beta \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{\xi^2}{2} \right\} \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(x + \xi \sin \beta \cos \beta)^2}{2 \cos^2 \beta} \right\} dx \]

\[ = 2 \tan \beta \phi(\xi) \left[ \cos^2 \frac{\beta}{\sqrt{2\pi}} \exp \left\{ -\frac{\xi^2}{2} \right\} - \xi \sin \beta \cos^2 \beta \Phi(-\xi \sin \beta) \right] \]

\[ = 2 \sin \beta \cos \phi(\xi) \left[ \phi(\xi \sin \beta) - \xi \sin \beta \Phi(-\xi \sin \beta) \right] \]

\[ \geq 2 \sin \beta \cos \phi(\xi) \left[ \phi(\xi) - \xi \Phi(-\xi) \right]. \]

\( \square \)
3.4 A KL divergence bound of the necessary order

**Proposition 15.** For some \( \mu_0, \mu, \mu' \in \mathbb{R}^d \) such that \( ||\mu|| = ||\mu'|| \), let

\[
\theta = \left( \mu_0 - \frac{\mu}{2}, \mu_0 + \frac{\mu}{2} \right)
\]

and

\[
\theta' = \left( \mu_0 - \frac{\mu'}{2}, \mu_0 + \frac{\mu'}{2} \right).
\]

Then

\[
\text{KL}(P_\theta, P_{\theta'}) \leq \xi^4 (1 - \cos \beta)
\]

where \( \xi = \frac{||\mu||}{2\sigma} \) and \( \cos \beta = \frac{||\mu'\parallel}{||\mu'\parallel ||\mu'||} \).

**Proof.** Since the KL divergence is invariant to affine transformations, it is easy to see that

\[
\text{KL}(P_\theta, P_{\theta'}) = \int_{\mathbb{R}} \int_{\mathbb{R}} p_1(x, y) \log \frac{p_1(x, y)}{p_2(x, y)} \, dx \, dy
\]

where

\[
p_1(x, y) = \frac{1}{2} \phi(x + \xi_x) \phi(y + \xi_y) + \frac{1}{2} \phi(x - \xi_x) \phi(y - \xi_y),
\]

\[
p_2(x, y) = \frac{1}{2} \phi(x + \xi_x) \phi(y - \xi_y) + \frac{1}{2} \phi(x - \xi_x) \phi(y + \xi_y),
\]

\[
\xi_x = \xi \cos \frac{\beta}{2}, \quad \xi_y = \xi \sin \frac{\beta}{2}.
\]

Since

\[
\frac{p_1(x, y)}{p_2(x, y)} = \frac{\phi(x + \xi_x) \phi(y + \xi_y) + \phi(x - \xi_x) \phi(y - \xi_y)}{\phi(x + \xi_x) \phi(y - \xi_y) + \phi(x - \xi_x) \phi(y + \xi_y)}
\]

\[
= \frac{\exp(-x \xi_x - y \xi_y) + \exp(x \xi_x + y \xi_y)}{\exp(-x \xi_x + y \xi_y) + \exp(x \xi_x - y \xi_y)}
\]

we have

\[
\log \frac{p_1(x, y)}{p_2(x, y)} = \log \frac{\cosh(x \xi_x + y \xi_y)}{\cosh(x \xi_x - y \xi_y)}.
\]

Furthermore,

\[
\int_{\mathbb{R}} \int_{\mathbb{R}} \frac{1}{2} \phi(x + \xi_x) \phi(y + \xi_y) \log \frac{\cosh(x \xi_x + y \xi_y)}{\cosh(x \xi_x - y \xi_y)} \, dx \, dy
\]

\[
= \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{1}{2} \phi(-x + \xi_x) \phi(-y + \xi_y) \log \frac{\cosh(-x \xi_x - y \xi_y)}{\cosh(-x \xi_x + y \xi_y)} \, dx \, dy
\]

\[
= \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{1}{2} \phi(x - \xi_x) \phi(y - \xi_y) \log \frac{\cosh(x \xi_x + y \xi_y)}{\cosh(x \xi_x - y \xi_y)} \, dx \, dy
\]

so

\[
\text{KL}(P_\theta, P_{\theta'}) = \int_{\mathbb{R}} \int_{\mathbb{R}} \phi(x - \xi_x) \phi(y - \xi_y) \log \frac{\cosh(x \xi_x + y \xi_y)}{\cosh(x \xi_x - y \xi_y)} \, dx \, dy
\]

\[
= \int_{\mathbb{R}} \int_{\mathbb{R}} \phi(x) \phi(y) \log \frac{\cosh(x \xi_x + \xi_x^2 + y \xi_y + \xi_y^2)}{\cosh(x \xi_x + \xi_x^2 - y \xi_y - \xi_y^2)} \, dx \, dy.
\]
But for any \( x \)

\[
- \int_{\mathbb{R}} \phi(x) \phi(y) \log \cosh(x \xi_x + \xi^2_x - y \xi_y - \xi^2_y) \, dy
\]

\[
= - \int_{\mathbb{R}} \phi(x) \phi(-y) \log \cosh(x \xi_x + \xi^2_x + y \xi_y - \xi^2_y) \, dy
\]

\[
= - \int_{\mathbb{R}} \phi(x) \phi(y) \log \cosh(x \xi_x + \xi^2_x + y \xi_y - \xi^2_y) \, dy,
\]

thus,

\[
\text{KL}(P_\theta, P_{\theta'}) = \int \int_{\mathbb{R}} \phi(x) \phi(y) \log \frac{\cosh(x \xi_x + \xi^2_x + y \xi_y + \xi^2_y)}{\cosh(x \xi_x + \xi^2_x + y \xi_y - \xi^2_y)} \, dxdy
\]

\[
= \int \phi(z) \log \frac{\cosh(z \sqrt{\xi^2_x + \xi^2_y + \xi^2_y + \xi^2_y})}{\cosh(z \sqrt{\xi^2_x + \xi^2_y + \xi^2_y - \xi^2_y})} \, dz
\]

\[
= \int \phi(z) \log \frac{\cosh(z + \xi^2_x + \xi^2_y)}{\cosh(z + \xi^2_x - \xi^2_y)} \, dz
\]

since \( \xi^2_x + \xi^2_y = \xi^2 \). By the mean value theorem and the fact that \( \tanh \) is monotonically increasing,

\[
\log \frac{\cosh(z + \xi^2_x + \xi^2_y)}{\cosh(z + \xi^2_x - \xi^2_y)} \leq 2 \xi^2 \tanh(z + \xi^2_x + \xi^2_y)
\]

\[
= 2 \xi^2 \tanh(z + \xi^2)
\]

for all \( z \). Since \( \tanh \) is an odd function,

\[
\text{KL}(P_\theta, P_{\theta'}) \leq 2 \xi^2 \int \phi(z) \tanh(z + \xi^2) \, dz
\]

\[
= 2 \xi^2 \int \phi(z)(\tanh(z + \xi^2) - \tanh(z)) \, dz.
\]

Using the mean value theorem again,

\[
\tanh(z + \xi^2) - \tanh(z) \leq \xi^2 \max_{x \in [z, z + \xi^2]} (1 - \tanh^2(x))
\]

\[
\leq \xi^2
\]

for all \( z \), so

\[
\text{KL}(P_\theta, P_{\theta'}) \leq 2 \xi^2 \xi^2
\]

\[
= 2 \xi^4 \sin^2 \frac{\beta}{2}
\]

\[
= \xi^4 (1 - \cos \beta).
\]

3.5 Non-sparse lower bound

**Theorem 3.** Assume that \( d \geq 9 \) and \( \frac{\lambda}{2} \leq 0.2 \). Then

\[
\inf_{\tilde{F}_n} \sup_{\theta \in \Theta_\lambda} E_{\theta} L_{\theta}(\tilde{F}_n) \geq \frac{1}{500} \min \left\{ \frac{\sqrt{\log 2 \sigma^2}}{3 \lambda^2 \sqrt{\frac{d-1}{n}}}, \frac{1}{4} \right\}
\]
Proof. Let $\xi = \frac{\lambda}{2\sqrt{d}}$, and define
\[
\epsilon = \min \left\{ \frac{\sqrt{\log 2} \sigma^2}{3}, \frac{1}{\lambda \sqrt{n}}, \frac{\lambda}{4\sqrt{d - 1}} \right\}.
\]
Define $\lambda_0^2 = \lambda^2 - (d-1)\epsilon^2$. Let $\Omega = \{0, 1\}^{d-1}$. For $\omega = (\omega(1), \ldots, \omega(d-1)) \in \Omega$, let $\mu_\omega = \lambda_0 e_d + \sum_{i=1}^{d-1} (2\omega(i) - 1) \epsilon e_i$ (where $\{e_i\}_{i=1}^d$ is the standard basis for $\mathbb{R}^d$). Let $\theta_\omega = \left(-\frac{\mu_\omega}{\lambda_0}, \frac{\mu_\omega}{\lambda_0}\right) \in \Theta_\lambda$.

By Proposition 15, for any $\omega, \nu \in \Omega$,
\[
KL(P_{\theta_\omega}, P_{\theta_\nu}) \leq \xi^4 (1 - \cos \beta_{\omega, \nu})
\]
where
\[
\cos \beta_{\omega, \nu} = \frac{\mu_\omega^T \mu_\nu}{\lambda^2} = 1 - \frac{2\rho(\omega, \nu)\epsilon^2}{\lambda^2}
\]
and $\rho$ is the Hamming distance, so
\[
KL(P_{\theta_\omega}, P_{\theta_\nu}) \leq \xi^4 \frac{2\rho(\omega, \nu)\epsilon^2}{\lambda^2} \leq \xi^4 \frac{2(d-1)\epsilon^2}{\lambda^2}.
\]

By Proposition 14 since $\cos \beta_{\omega, \nu} \geq \frac{1}{2}$,
\[
L_{\theta_\omega}(F_{\theta_\nu}) \leq \frac{1}{\pi} \tan \beta_{\omega, \nu}
\]
\[
\leq \frac{1}{\pi} \frac{\sqrt{1 + \cos \beta_{\omega, \nu}}}{\cos \beta_{\omega, \nu}} \sqrt{1 - \cos \beta_{\omega, \nu}}
\]
\[
\leq \frac{4 \sqrt{d - 1} \epsilon}{\pi \lambda}
\]
and
\[
L_{\theta_\omega}(F_{\theta_\nu}) \geq 2g(\xi) \sin \beta_{\omega, \nu} \cos \beta_{\omega, \nu}
\]
\[
\geq g(\xi) \sin \beta_{\omega, \nu}
\]
\[
\geq \sqrt{2g(\xi)} \frac{\sqrt{\rho(\omega, \nu)\epsilon}}{\lambda}
\]
where $g(x) = \phi(x)(\phi(x) - x\Phi(-x))$. By Lemma 3 there exist $\omega_0, \ldots, \omega_M \in \Omega$ such that $M \geq 2^{(d-1)/8}$ and
\[
\rho(\omega_i, \omega_j) \geq \frac{d - 1}{8}, \quad \forall 0 \leq i < j \leq M.
\]
For simplicity of notation, let $\theta_i = \theta_{\omega_i}$ for all $i \in [0..M]$. Then, for $i \neq j \in [0..M]$,
\[
KL(P_{\theta_i}, P_{\theta_j}) \leq \xi^4 \frac{2(d-1)\epsilon^2}{\lambda^2},
\]
and
\[
L_{\theta_i}(F_{\theta_j}) \leq \frac{4 \sqrt{d - 1} \epsilon}{\pi \lambda}
\]
and
\[
L_{\theta_i}(F_{\theta_j}) \geq \frac{1}{2} g(\xi) \frac{\sqrt{d - 1} \epsilon}{\lambda}.
\]
Define
\[
\gamma = \frac{1}{4} (g(\xi) - 2\xi^2) \frac{\sqrt{d - 1} \epsilon}{\lambda}.
\]
Then for any \( i \neq j \in [0..M] \), and any \( \hat{F} \) such that \( L_{\theta_i}(\hat{F}) < \gamma \),
\[
L_{\theta_i}(F_{\theta_j}) + L_{\theta_j}(\hat{F}) + \sqrt{\frac{\text{KL}(P_{\theta_i}, P_{\theta_j})}{2}} < \left( \frac{4}{\pi} + \frac{1}{4} (g(\xi) - 2\xi^2 + \xi^2) \right) \frac{\sqrt{d-1} \epsilon}{\lambda} \leq \frac{1}{2}
\]
because, for \( \xi \leq 0.1 \), by definition of \( \epsilon \),
\[
\left( \frac{4}{\pi} + \frac{1}{4} (g(\xi) - 2\xi^2 + \xi^2) \right) \frac{\sqrt{d-1} \epsilon}{\lambda} \leq 2 \frac{\sqrt{d-1} \epsilon}{\lambda} \leq \frac{1}{2}.
\]
So, by Proposition \[13\]
\[
L_{\theta_i}(\hat{F}) \geq L_{\theta_i}(F_{\theta_j}) - L_{\theta_i}(\hat{F}) - \sqrt{\frac{\text{KL}(P_{\theta_i}, P_{\theta_j})}{2}} \geq \gamma.
\]
Also,
\[
\max_{i \in [M]} \text{KL}(P_{\theta_i}, P_{\theta_0}) \leq (d-1)\xi^4 2\epsilon^2 \lambda^2 \leq \log M \frac{M}{9n}
\]
because, by definition of \( \epsilon \),
\[
\xi^4 2\epsilon^2 \lambda^2 \leq \frac{\log 2}{72n}.
\]
So by Proposition \[12\] and the fact that \( \xi \leq 0.1 \),
\[
\inf_{\hat{F}_n} \max_{i \in [0..M]} E_\theta L_{\theta_i}(\hat{F}_n) \geq 0.07\gamma
\]
\[
= 0.07 \frac{1}{4} (g(\xi) - 2\xi^2) \frac{\sqrt{d-1} \epsilon}{\lambda}
\]
\[
\geq \frac{1}{500} \min \left\{ \sqrt{\frac{\log 2}{3}} \frac{\sigma^2}{\lambda^2} \sqrt{\frac{d-1}{n}}, \frac{1}{4} \right\}
\]
and to complete the proof we use the fact that
\[
\inf_{\hat{F}_n} \sup_{\theta \in \Theta} E_\theta L_{\theta}(\hat{F}_n) \geq \inf_{\hat{F}_n} \max_{i \in [0..M]} E_\theta L_{\theta_i}(\hat{F}_n).
\]
\[\square\]

### 3.6 Sparse lower bound

**Theorem 4.** Assume that \( \frac{\lambda}{\sigma} \leq 0.2 \), \( d \geq 17 \), and
\[
5 \leq s \leq \frac{d-1}{4} + 1.
\]
Then
\[
\inf_{\hat{F}_n} \sup_{\theta \in \Theta_{\lambda,s}} E_\theta L_{\theta}(\hat{F}_n) \geq \frac{1}{600} \min \left\{ \sqrt{\frac{8}{45}} \frac{\sigma^2}{\lambda^2} \sqrt{\frac{s-1}{n}} \log \left( \frac{d-1}{s-1} \right), \frac{1}{2} \right\}.
\]

**Proof.** For simplicity, we state this proof for \( \Theta_{\lambda,s+1} \), assuming \( 4 \leq s \leq \frac{d-1}{4} \). Let \( \xi = \frac{\lambda}{2\sigma} \), and define
\[
\epsilon = \min \left\{ \sqrt{\frac{8}{45}} \frac{\sigma^2}{\lambda^2} \sqrt{\frac{1}{n}} \log \left( \frac{d-1}{s} \right), \frac{1}{2} \right\}.
\]
Define \( \lambda_0^2 = \lambda^2 - se^2 \). Let \( \Omega = \{ \omega \in \{0,1\}^{d-1} : ||\omega||_0 = s \} \). For \( \omega = (\omega(1), \ldots, \omega(d-1)) \in \Omega \), let \( \mu_{\omega} = \lambda_0 e_d + \sum_{i=1}^{d-1} \omega(i) e_i \) (where \( \{e_i\}_{i=1}^{d} \) is the standard basis for \( \mathbb{R}^d \)). Let \( \theta_\omega = (-\frac{\mu_\omega}{2}, \frac{\mu_\omega}{2}) \in \Theta_{\lambda,s} \).
By Proposition 15,

\[ \text{KL}(P_{\theta_{\omega}}, P_{\theta_{\nu}}) \leq \xi^4 (1 - \cos \beta_{\omega, \nu}) \]

where

\[ \cos \beta_{\omega, \nu} = \frac{|\mu^T_{\omega} \mu_{\nu}|}{\lambda^2} = 1 - \frac{\rho(\omega, \nu)\epsilon^2}{2\lambda^2} \]

and \( \rho \) is the Hamming distance, so

\[ \text{KL}(P_{\theta_{\omega}}, P_{\theta_{\nu}}) \leq \xi^4 \frac{\rho(\omega, \nu)\epsilon^2}{2\lambda^2} \leq \xi^4 \frac{\sqrt{s\epsilon^2}}{\lambda^2}. \]

By Proposition 14, since \( \cos \beta_{\omega, \nu} \geq \frac{1}{2} \),

\[ L_{\theta_{\omega}}(F_{\theta_{\nu}}) \leq \frac{\pi}{\lambda} \tan \beta_{\omega, \nu} \leq \frac{2}{\pi} \sin \beta_{\omega, \nu} \leq \frac{2\sqrt{2}}{\pi} \sqrt{s\epsilon} \]

and

\[ L_{\theta_{\omega}}(F_{\theta_{\nu}}) \geq \frac{g(\xi) \sin \beta_{\omega, \nu} \cos \beta_{\omega, \nu}}{\sqrt{2}} \geq \frac{g(\xi) \sqrt{\rho(\omega, \nu)\epsilon}}{\lambda} \]

where \( g(x) = \phi(x)(\phi(x) - x\Phi(-x)) \). By Lemma 4, there exist \( \omega_0, ..., \omega_M \in \Omega \) such that \( \log(M + 1) \geq \frac{s}{\delta} \log \left( \frac{d-1}{\delta} \right) \) and

\[ \rho(\omega_i, \omega_j) \geq \frac{s}{2}, \quad 0 \leq i < j \leq M. \]

For simplicity of notation, let \( \theta_i = \theta_{\omega_i} \) for all \( i \in [0..M] \). Then, for \( i \neq j \in [0..M] \),

\[ \text{KL}(P_{\theta_i}, P_{\theta_j}) \leq \xi^4 \frac{\sqrt{s\epsilon^2}}{\lambda^2}, \]

and

\[ L_{\theta_i}(F_{\theta_j}) \leq \frac{2\sqrt{2}}{\pi} \sqrt{s\epsilon} \]

and

\[ L_{\theta_i}(F_{\theta_j}) \geq \frac{g(\xi) \sqrt{s\epsilon}}{\lambda}. \]

Define

\[ \gamma = \frac{1}{4} (g(\xi) - \sqrt{2}\xi^2) \frac{\sqrt{s\epsilon}}{\lambda}. \]

Then for any \( i \neq j \in [0..M] \), and any \( \hat{F} \) such that \( L_{\theta_i}(\hat{F}) < \gamma \),

\[ L_{\theta_i}(F_{\theta_j}) + L_{\theta_i}(\hat{F}) + \frac{\text{KL}(P_{\theta_i}, P_{\theta_j})}{2} < \left( \frac{2\sqrt{2}}{\pi} + \frac{1}{4} (g(\xi) - \sqrt{2}\xi^2) + \frac{\xi^2}{\sqrt{2}} \right) \frac{\sqrt{s\epsilon}}{\lambda} \leq \frac{1}{2} \]

because, for \( \xi \leq 0.1 \), by definition of \( \epsilon \),

\[ \left( \frac{2\sqrt{2}}{\pi} + \frac{1}{4} (g(\xi) - \sqrt{2}\xi^2) + \frac{\xi^2}{\sqrt{2}} \right) \frac{\sqrt{s\epsilon}}{\lambda} \leq \frac{\sqrt{s\epsilon}}{\lambda} \leq \frac{1}{2}. \]
So, by Proposition 13

\[ L_{\theta_1}(\hat{F}) \geq L_{\theta_1}(F_{\theta_1}) - L_{\theta_1}(\hat{F}) - \sqrt{\frac{\text{KL}(P_{\theta_1}, P_{\theta_2})}{2}} \geq \gamma. \]

Also,

\[
\max_{i \in [M]} \text{KL}(P_{\theta_i}, P_{\theta_0}) \leq \xi \frac{s \epsilon^2}{\lambda^2} \leq \frac{1}{18n} \log \left( \frac{d - 1}{s} \right)^{\frac{1}{2}} \leq \frac{1}{9n} \log \left( \left( \frac{d - 1}{s} \right)^{\frac{1}{2}} - 1 \right) \leq \log M \frac{1}{9n}
\]

because, by definition of \( \epsilon \),

\[ \xi \frac{s \epsilon^2}{\lambda^2} \leq \frac{s}{90n} \log \left( \frac{d - 1}{s} \right). \]

So by Proposition 12 and the fact that \( \xi \leq 0.1 \),

\[
\inf_{\hat{F}_n} \max_{i \in [0..M]} \mathbb{E}_{\theta_1} L_{\theta_1}(\hat{F}_n) \geq 0.07 \gamma
\]

\[
\geq 0.07 \frac{0.1 \sqrt{s \epsilon}}{4 \lambda}
\]

\[
\geq \frac{1}{600} \min \left\{ \sqrt{\frac{8 \lambda^2}{45 \lambda^2}}, \sqrt{\frac{s \log \left( \frac{d - 1}{s} \right)}{n}} \frac{1}{2} \right\}
\]

and to complete the proof we use the fact that

\[
\inf_{\hat{F}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta L_\theta(\hat{F}_n) \geq \inf_{\hat{F}_n} \max_{i \in [0..M]} \mathbb{E}_{\theta_i} L_{\theta_i}(\hat{F}_n).
\]

\[ \square \]

References

