Supplementary Material for Theoretical Insights Into Multiclass Classification: A High-dimensional Asymptotic View

Anonymous Author(s)
Affiliation
Address
email

Abstract

Contemporary machine learning applications often involve classification tasks with many classes. Despite their extensive use, a precise understanding of the statistical properties and behavior of classification algorithms is still missing, especially in modern regimes where the number of classes is rather large. In this paper, we take a step in this direction by providing the first asymptotically precise analysis of linear multiclass classification. Our theoretical analysis allows us to precisely characterize how the test error varies over different training algorithms, data distributions, problem dimensions as well as number of classes, inter/intra class correlations and class priors. Specifically, our analysis reveals that the classification accuracy is highly distribution-dependent with different algorithms achieving optimal performance for different data distributions and/or training/features sizes. Unlike linear regression/binary classification, the test error in multiclass classification relies on intricate functions of the trained model (e.g., correlation between some of the trained weights) whose asymptotic behavior is difficult to characterize. This challenge is already present in simple classifiers, such as those minimizing a square loss. Our novel theoretical techniques allow us to overcome some of these challenges. The insights gained may pave the way for a precise understanding of other classification algorithms beyond those studied in this paper.

Roadmap to the supplementary

This supplementary material contains a more complete version of the paper together with all the formal proofs. In particular, compared to the main Neurips submission this supplementary contains formal statements about the Weighted Least-squares approach for the Gaussian Mixture Model (see Section 3.3) and the Multinomial Logistic Model (see Section 4.3). We also carryout additional numerical experiments in Section A. Finally, we formally prove all of our results in the appendix of this supplementary.

1 Introduction

Multiclass classification is fundamental to a large number of real-world machine learning applications that demand the ability to automatically distinguish between thousands of different classes. Applications include essentially any problem with categorical outputs spanning natural language processing [62], where a seq2seq decoder has to choose the correct word token, reinforcement learning [28, 46], where the agent has to choose the correct action, to recommendation systems, where the model should recommend the correct movie out of many other options. For instance, YouTube's recommendation

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

system is modeled as an extreme multiclass problem with more than a million classes where each video corresponds to a viable class [12].

The growing list of applications motivate an in-depth exploration of multiclass classification algorithms. Despite their extensive use however, a precise understanding of the statistical properties and behavior of classification algorithms is still missing with many open questions: *What is the total and per class test accuracy? How does this quantity depend on various problem parameters such as data distributions, problem dimensions, etc.? What is the highest test accuracy achievable by any algorithm? What is the best algorithm for each scenario? Which algorithm achieves the highest accuracy on rare or minority classes? How does the answer to the above question change in modern regimes where the number of classes are large?*

Asymptotic analysis in modern high-dimensional regimes where the number of training data and feature sizes grow in tandem with each other provides a promising setting for precisely quantifying the accuracy of classification algorithms as a function of problem variables and resolving the questions above. However, despite the rich literature on precise high-dimensional estimation and more recently binary classification, multiclass classification is an under-explored venue possibly due to the difficulty of capturing the intricate dependencies between the classes even for relatively simple linear classifiers.

Contributions. We initiate a precise asymptotic study of linear multiclass classification in the modern high-dimensional regime, where the sizes of the training data and of the feature vectors grow large at a proportional rate. A key promise of such a precise analysis is that it allows us to accurately compare between different classification algorithms and data models. Compared to linear regression/binary classification, we identify the following crucial challenge: the test accuracy in multiclass classification relies on intricate cross-correlations between the trained weights of the classifier. This has two consequences that drive our analysis. First, in order to obtain sharp asymptotics on the test error of any classifier, it is a prerequisite to precisely quantify the asymptotics of these cross-correlations. Second, the test error does not depend on the correlations in closed-form expressions. Thus, to compare between different classifiers, we need efficient numerical and analytic means to evaluate the test error in terms of the correlation matrices. Interestingly, we show that these challenges are already present in simple classifiers, such as minimizing the square loss, and in stylized distributional settings, such as Gaussian features. Our contributions are as follows:

- We study two different data models: a Gaussian Mixtures Model (GMM) and a Multinomial Logit Model (MLM) with Gaussian features. For each one of them, we provide a precise characterization of total and class-wise test accuracy for three different training algorithms: a least-squares based classifier, a weighted least-squares based classifier, and a simple per class averaging estimator. For the least-squares based classifiers, we develop a new technique to overcome the technical challenge of characterizing the limiting behavior of the weights' cross-correlations. For the per class averaging classifier, we show that it is Bayes optimal for a GMM with equal priors.
- We discuss efficient means of evaluating the test accuracy as a function of the weights' cross-correlations. This, together with the derived asymptotic formulae for the latter, lead to the first precise high-dimensional characterization of how the total/class-wise accuracy varies for different training algorithms, data distributions, problem dimensions as well as number of classes, the inter/intra class correlations and class priors. For special problem geometries, we derive precise conditions on the data distribution and on the relative size of the training set over which each of the two studied algorithms dominates.
- We present and discuss extensive numerical simulations that corroborate our theoretical findings. For instance, with an eye towards making classification algorithms more fair/equitable, we use our precise characterization of the class-wise accuracy to demonstrate how different algorithms behave in the presence of rare/minority classes. We also empirically compare the algorithms studied in this paper to other popular losses such as cross-entropy minimization. This allows us to better understand the performance of various algorithms in modern regimes of large number of classes.

Related Work. There is a classical body of algorithmic work on multiclass classification, e.g., [13, 36, 68, 5, 16] and several empirical studies of their comparative performance [56, 20, 1, 52]. A more recent extension of this line of work investigates the effect of the loss function in deep neural networks [25, 21, 33, 4]. Algorithms for extreme multiclass problems with huge number of classes

has also been studied in several [10, 70, 14, 55, 34] works. On the theory front, numerous works have investigated consistency [71, 36, 64, 54, 53] and finite-sample behavior [32, 23, 1, 40, 11, 37, 45, 38] of multiclass classification algorithms. Our work differs from this literature in that we are interested in *precise* characterizations of the test accuracy rather than order-wise bounds. Here we focus on linear classifiers, but we consider the modern high-dimensional regime in which both the sample size and the features' dimension are large.

Specifically, our theoretical approach to linear multiclass classification fits in the rapidly growing literature on *sharp* high-dimensional asymptotics of convex optimization-based estimators [18, 59, 3, 2, 60, 51, 66, 31, 19, 17, 50, 48, 67, 8, 26, 6, 29]. Most of this line of work studies linear models and regression problems. More recently there has been a surge of interest in sharp analysis of a variety of methods tailored to binary classification models [65, 27, 7, 61, 44, 43, 30, 57, 63, 15, 49, 41, 47, 63, 42]. Nevertheless, none of these prior works have yet considered multiclass classification settings. Our paper unveils the salient features of the multiclass setting and shows that corresponding results from the binary setting do not directly apply here. We emphasize that this is the case even for seemingly simple one-vs-all (OVA) classifiers, such as minimizing the square-loss, that involve training a single binary classifier per class [56]. The key technical tool behind our sharp analysis is the convex Gaussian min-max Theorem (CGMT) [66, 60, 50])). However, a "naive" application of the CGMT on the original optimization of the classifier does not allow us to compute all the necessary correleations between the classfier's weights to precisely capture the total/class-wise errors. Instead, our key idea is to formulate an artificial optimization problem, which captures the missing correlations and at the same time conveniently allows us to leverage the CGMT.

Notation. We use $[k]$ to denote $\{1,\dots,k\}$. We write $\boldsymbol{e}_\ell$ for the ℓ-th standard basis vector in $\mathbb{R}^k$. We also write $\boldsymbol{I}_k, \boldsymbol{0}_{k\times k}$ and $\boldsymbol{1}_k$ for the $k\times k$ identity and all-zeros matrices and the $k\times 1$ all-ones vectors. For a vector $\boldsymbol{c}\in\mathbb{R}^k$ we write $\arg\max \boldsymbol{c}$ to denote the index of its largest entry, i.e., $\arg\max \boldsymbol{c} = \arg\max_{j\in[k]} \boldsymbol{c}_i$. We use $Q(x)$ for the tail of a standard Gaussian (Q-function). Finally, we reserve variables $G_0, G_1, \dots, G_k \overset{iid}{\sim} \mathcal{N}(0,1)$ to denote i.i.d. standard Gaussians.

2 Problem formulation

In this paper we focus on multiclass classification problems with k classes. Specifically, we assume the training data consists of n feature/label pairs $\{(\boldsymbol{x}_i, Y_i)\}_{i=1}^n$ with $\boldsymbol{x}_i\in\mathbb{R}^d$ representing the features and $Y_i\in\{1,2,\dots,k\}$ the associated labels representing one of k classes. It will be convenient to also model the labels as one-hot encoded vectors $\boldsymbol{y}_i\in\mathbb{R}^k$ representing one of k classes with one-hot encoding i.e. $\boldsymbol{y}_i = \boldsymbol{e}_{Y_i}$. Therefore, when convenient we shall use $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ to represent the training data. Throughout this paper we shall use $\boldsymbol{X} = [\boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \dots \quad \boldsymbol{x}_n]\in\mathbb{R}^{d\times n}$ to denote the matrix of features with the corresponding labels aggregated into a matrix of the form $\boldsymbol{Y} = [\boldsymbol{y}_1 \quad \boldsymbol{y}_2 \quad \dots \quad \boldsymbol{y}_n]\in\mathbb{R}^{k\times n}$. We shall also use $\boldsymbol{Y}_\ell\in\mathbb{R}^n$ to denote the ℓ-th row of $\boldsymbol{Y}$. In our analysis we focus on training linear classifiers. We shall use $\boldsymbol{W}\in\mathbb{R}^{k\times d}$ and $\boldsymbol{b}\in\mathbb{R}^k$ to denote the weights and biases of this linear model. The overall input-output relationship of the classifier in this case is a function that maps an input vector $\boldsymbol{x}\in\mathbb{R}^d$ into an output of size k via $\boldsymbol{x}\mapsto \boldsymbol{W}\boldsymbol{x}+\boldsymbol{b}\in\mathbb{R}^k$ where a training algorithm is used to train the corresponding weights $\boldsymbol{W}\in\mathbb{R}^{k\times d}$ and biases $\boldsymbol{b}\in\mathbb{R}^k$. Next we detail the data models and training algorithms that are formally studied in this paper. We end this section by discussing how the test error can be calculated for the different data models.

2.1 Data Models

In our theoretical analysis we assume the training data $\{(\boldsymbol{x}_i, Y_i)\}_{i=1}^n$ (alternatively $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$) are generated i.i.d. according to $(\boldsymbol{x}, Y)$/$(\boldsymbol{x}, \boldsymbol{y})$. We consider two models for the distribution of $(\boldsymbol{x}, \boldsymbol{y})$ which we detail next. In both models we shall use mean/regressor vectors $\{\boldsymbol{\mu}_\ell\}_{\ell=1}^k\in\mathbb{R}^d$ and aggregate them into columns of a matrix of the form $\boldsymbol{M} := [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_k]\in\mathbb{R}^{d\times k}$. In the first model, these vectors represent the mean of the features conditioned on the class, i.e. $\boldsymbol{\mu}_\ell = \mathbb{E}\left[\boldsymbol{x}|Y=\ell\right]$, where as in the second model these vectors can be viewed as regressor coefficients. We shall refer to $\{\boldsymbol{\mu}_\ell\}_{\ell=1}^k$/$\boldsymbol{M}$ as "mean" vectors/matrix in both models. We use $\Sigma_{\boldsymbol{\mu},\boldsymbol{\mu}} = \boldsymbol{M}^T\boldsymbol{M}$ to denote the Grammian matrix of means. Furthermore, we shall use $\mu_\ell := \|\boldsymbol{\mu}_\ell\|_{\ell_2}$ to denote the norm of the mean vector $\boldsymbol{\mu}_\ell$.

Gaussian Mixture Model (GMM). In this model each example $(\boldsymbol{x}, Y)$ belongs to class $\ell \in [k]$ with probability π_ℓ, i.e., $\mathbb{P}\{Y = \ell\} = \pi_\ell$. We let $\boldsymbol{\pi} = \begin{bmatrix}\pi_1 & \pi_2 & \dots & \pi_k\end{bmatrix}^T \in \mathbb{R}^k$ denote the vector of priors which of course obeys $\boldsymbol{\pi} \geq \mathbf{0}$ and $\mathbf{1}^T\boldsymbol{\pi} = 1$. Also, we model the class conditional density of an example in class ℓ with an isotropic Gaussian centered at a mean vector $\boldsymbol{\mu}_\ell$. In particular, we say that a data point $(\boldsymbol{x}, Y)$ (or its one-hot encoded representation $(\boldsymbol{x}, \boldsymbol{y})$) follows the GMM model when

$$\mathbb{P}\{Y = \ell\} = \pi_\ell \qquad \text{and} \qquad \boldsymbol{x} = \boldsymbol{\mu}_Y + \boldsymbol{z},\ \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_d). \tag{2.1}$$

We note that for a training set summarized by the feature and label matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ with columns generated i.i.d. according to the above distribution we have: $\boldsymbol{X} = \boldsymbol{M}\boldsymbol{Y} + \boldsymbol{Z}$ where $\boldsymbol{Z} \in \mathbb{R}^{d\times n}$ is a Gaussian noise matrix with i.i.d. $\mathcal{N}(0, \sigma^2)$ entries.

Multinomial Logit Model (MLM). In this model we assume that feature vectors $\boldsymbol{x}$ are distributed i.i.d. $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ and that the conditional density of the class labels is given by the soft-max function. Concretely, we say that a data point $(\boldsymbol{x}, Y)$ (or its one-hot encoded representation $(\boldsymbol{x}, \boldsymbol{y})$) follows the multinomial logit model when

$$\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d) \qquad \text{and} \qquad \mathbb{P}\{Y = \ell \mid \boldsymbol{x}\} = e^{\langle \boldsymbol{\mu}_\ell, \boldsymbol{x}\rangle} \Big/ \sum_{j\in[k]} e^{\langle \boldsymbol{\mu}_j, \boldsymbol{x}\rangle}. \tag{2.2}$$

2.2 Classification algorithms

As mentioned earlier in this paper we focus on training linear classifiers of the form $\boldsymbol{x} \mapsto \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}$ with $\boldsymbol{W} \in \mathbb{R}^{k\times d}$ denoting the weights and $\boldsymbol{b} \in \mathbb{R}^k$ the offset values.

Least-squares (LS). In this approach '

Class averaging (Avg). This approach uses the following weight and offset values

$$\widehat{\boldsymbol{W}} := \frac{1}{n}\boldsymbol{Y}\boldsymbol{X}^T \qquad \text{and} \qquad \widehat{\boldsymbol{b}} := \frac{1}{n}\boldsymbol{Y}\mathbf{1}. \tag{2.3}$$

Let n_ℓ be the number of training data from class ℓ then, equivalently,

$$\widehat{\boldsymbol{w}}_\ell = \frac{n_\ell}{n}\left(\frac{1}{n_\ell}\sum_{i:\, Y_i=\ell}^{n} \boldsymbol{x}_i\right) \quad \text{and} \quad \widehat{\boldsymbol{b}}_\ell = \frac{n_\ell}{n}.$$

Therefore, this classifier picks weights according to the empirical mean of features of each class multiplied by the relative frequency of that class and the offset value as the fraction of data points from that class. We note that this algorithm has the same classification performance as the outcome of the ridge-regularized least-squares with infinite regularization.

Weighted Least-squares (WLS). This is a variation of the Least-squares approach where we fit a weighted least squares loss of the form

$$(\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{b}}) := \arg\min_{\boldsymbol{W},\boldsymbol{b}} \frac{1}{2n}\left\|\left(\boldsymbol{W}\boldsymbol{X} + \boldsymbol{b}\mathbf{1}_n^T - \boldsymbol{Y}\right)\boldsymbol{D}\right\|_F^2. \tag{2.4}$$

Here, $\boldsymbol{D} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with the ith diagonal entry equal to $D_{ii} = \omega_\ell$ when the i-th data point is from class ℓ (i.e. $Y_i = \ell$) and $\omega_\ell \geq 0,\ \ell \in [k]$ denote the weights. Aggregating the weights into a vector of the form $\boldsymbol{\omega} = \begin{bmatrix}\omega_1 & \omega_2 & \dots & \omega_k\end{bmatrix}^T \in \mathbb{R}^k$ we can rewrite $\boldsymbol{D}$ in the form

$$\boldsymbol{D} = \operatorname{diag}\left(\boldsymbol{Y}^T\boldsymbol{\omega}\right).$$

In this approach the loss associated to data points to class ℓ are weighted by a factor ω_ℓ^2. For instance, if the class priors are known, a natural choice would be $\omega_\ell = 1/\sqrt{\pi_\ell}$. Such a weighted approach allows the classification algorithm to focus on rare/minority classes which are not well represented in the training data.

Cross-entropy (CE). In this approach the best weight/offset values are determined by fitting a cross entropy loss

$$(\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{b}}) := \arg\min_{\boldsymbol{W},\boldsymbol{b}} \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{\sum_{\ell=1}^{k} e^{\langle \widehat{\boldsymbol{w}}_\ell, \boldsymbol{x}_i\rangle + \boldsymbol{b}_\ell}}{e^{\langle \widehat{\boldsymbol{w}}_{Y_i}, \boldsymbol{x}_i\rangle + \boldsymbol{b}_{Y_i}}}\right).$$

Theoretical analysis for CE is substantially more involved and we defer it to future work. Nevertheless, we compare with this classifier in our numerical simulations.

2.3 Class-wise and total test classification error in the high-dimensional regime

Let $\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{b}}$ denote the parameters of a trained classifier. Now consider a fresh data sample $(\boldsymbol{x}, Y)$ generated according to the same distribution as the training data. The class predicted by the classifier in this case is given by

$$\widehat{Y} = \arg\max_{j\in[k]} \langle \widehat{\boldsymbol{w}}_j, \boldsymbol{x}\rangle + \widehat{\boldsymbol{b}}_j.$$

Therefore, the classification error condition on the the true label being c (which we shall refer to as the class-wise test error) is equal to

$$\mathbb{P}_{e|c} := \mathbb{P}\left\{\widehat{Y} \neq Y \middle| Y = c\right\} = \mathbb{P}\left\{\langle \widehat{\boldsymbol{w}}_c, \boldsymbol{x}\rangle + \widehat{\boldsymbol{b}}_c \leq \max_{j\neq c} \langle \widehat{\boldsymbol{w}}_j, \boldsymbol{x}\rangle + \widehat{\boldsymbol{b}}_j\right\}. \tag{2.5}$$

The total classification error is given by

$$\mathbb{P}_e := \mathbb{P}\left\{\widehat{Y} \neq Y\right\} = \mathbb{P}\left\{\arg\max_{j\in[k]}\{\langle \widehat{\boldsymbol{w}}_j, \boldsymbol{x}\rangle + \widehat{\boldsymbol{b}}_j\} \neq Y\}\right\} = \mathbb{P}\left\{\langle \widehat{\boldsymbol{w}}_Y, \boldsymbol{x}\rangle + \widehat{\boldsymbol{b}}_Y \leq \max_{j\neq Y} \langle \widehat{\boldsymbol{w}}_j, \boldsymbol{x}\rangle + \widehat{\boldsymbol{b}}_j\right\}. \tag{2.6}$$

For both the GMM and MLM, the classification error depends on the vector of intercepts $\widehat{\boldsymbol{b}} \in \mathbb{R}^k$ and the following key "correlation" matrices:

$$\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}} := \widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^T \quad \text{and} \quad \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}} := \widehat{\boldsymbol{W}}\boldsymbol{M}.$$

GMM: $$\mathbb{P}_e = \mathbb{P}\left\{\arg\max \left(\sigma \boldsymbol{g} + \widehat{\boldsymbol{b}} + \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}}\boldsymbol{e}_Y\right) \neq Y\right\}, \text{ where } \boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}}). \tag{2.7}$$

MLM: $$\mathbb{P}_e = \mathbb{P}\left\{\arg\max \left(\boldsymbol{g} + \widehat{\boldsymbol{b}}\right) \neq Y(\boldsymbol{h})\right\}, \quad \text{where } \begin{bmatrix}\boldsymbol{g}\\ \boldsymbol{h}\end{bmatrix} \sim \mathcal{N}\Big(\boldsymbol{0}, \begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}} & \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}}\\ \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}}^T & \boldsymbol{\Sigma}_{\boldsymbol{\mu},\boldsymbol{\mu}}\end{bmatrix}\Big), \tag{2.8}$$

and $\mathbb{P}\{Y(\boldsymbol{h}) = \ell\} = e^{\boldsymbol{h}_\ell} / \sum_{j\in[k]} e^{\boldsymbol{h}_j}, \ \ell \in [k]$.

Calculating the class-wise/total misclassifcation errors. The identities (2.7) and (2.8) as well as similar ones for the class-wise test error demonstrate that the total/class-wise errors only depend on the correlation matrices $\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}}$, the offset values $\widehat{\boldsymbol{b}}$ and the the class conditional means. For instance, as we show in the supplementary for GMM the class-wise errors are given by

$$\mathbb{P}_{e|c} = 1 - \mathbb{P}\left\{\boldsymbol{S}_c^{1/2}\boldsymbol{z} \geq \boldsymbol{t}_c\right\}, \tag{2.9}$$

where $\boldsymbol{z}$ is a Gaussian random vector distributed as $\mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{k-1})$, $\boldsymbol{S}_c \in \mathbb{R}^{(k-1)\times(k-1)}$ is a symmetric matrix such that its i,j element is given by $[\boldsymbol{S}_c]_{ij} := \langle \widehat{\boldsymbol{w}}_c - \widehat{\boldsymbol{w}}_j, \widehat{\boldsymbol{w}}_c - \widehat{\boldsymbol{w}}_i\rangle$ and $\boldsymbol{t}_c \in \mathbb{R}^{k-1}$ a vector with entries $[\boldsymbol{t}_c]_i := \langle \widehat{\boldsymbol{w}}_i - \widehat{\boldsymbol{w}}_c, \boldsymbol{\mu}_c\rangle + (\widehat{\boldsymbol{b}}_i - \widehat{\boldsymbol{b}}_c)$. Similarly, based on (2.9) the total classification error in GMM is equal to $\mathbb{P}_e = \sum_{\ell=1}^k \pi_\ell \mathbb{P}_{e|c} = 1 - \sum_{\ell=1}^k \pi_\ell \mathbb{P}\left\{\boldsymbol{S}_c^{1/2}\boldsymbol{z} \geq \boldsymbol{t}_c\right\}$. As also detailed in the supplementary, the class-wise/total test errors for MLM similarly depends on quantities of the form $\mathbb{P}\{\boldsymbol{A}\boldsymbol{z} \geq \boldsymbol{t}\}$ with $\boldsymbol{z}$ a standard Gaussian random vector, $\boldsymbol{A}$ and $\boldsymbol{t}$ depending only on correlation matrices, conditional means and classifier offset-values. There are a variety of algorithmic approaches to calculate $\mathbb{P}\{\boldsymbol{A}\boldsymbol{z} \geq \boldsymbol{t}\}$ once $\boldsymbol{A}/\boldsymbol{t}$ is known based on MCMC methods. Analytic bounds on this quantity have also been studied in the literature [24, 58]; see more details in the supplementary.

High-dimensional regime. The main technical contribution of this paper involves sharp asymptotic formulae for the class-wise and total classification error of averaging and (weighted) LS algorithms for GMM and MLM. Due to space constraints, we defer the statements of our main results for WLS, as well as, all our proofs to the supplementary. All our results hold in the following high-dimensional regime with finite k.

Assumption 1 *We focus on a double asymptotic regime where $n, d \to \infty$ at a fixed ratio $\gamma = d/n > 0$.*

For the (weighted) least-squares classifier, we focus here in the overdetermined regime $\gamma < 1$. However, our approach is also directly applicable to regularized (or min-norm) LS/WLS in the overparameterized regime $\gamma > 1$. For a sequence of random variables $\mathcal{X}_{n,d}$ that converges in probability to some constant c in the limit above, we simply write $X_{n,d} \xrightarrow{P} c$.

3 Results for Gaussian Mixture Model

In this section we discuss the asymptotics of the intercepts/correlation matrices for the averaging and the LS classifiers for the Gaussian Mixture Model. We end this section by also characterizing the Bayes optimal estimator in this model when the priors are balanced $\pi_\ell = 1/k$ for $\ell \in [k]$.

3.1 Class averaging classifier

Proposition 3.1 *Consider data generated according to GMM in an asymptotic regime with any $\gamma > 0$. For the averaging estimator discussed in Section 2.2, the following high-dimensional limits hold*

$$\widehat{\boldsymbol{b}} \xrightarrow{P} \boldsymbol{\pi}\,, \quad \boldsymbol{\Sigma}_{\boldsymbol{w}\boldsymbol{\mu}} \xrightarrow{P} \mathit{diag}(\boldsymbol{\pi}) \cdot \boldsymbol{\Sigma}_{\boldsymbol{\mu},\boldsymbol{\mu}}\,, \tag{3.1a}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}} \xrightarrow{P} \gamma\sigma^2 \cdot \mathit{diag}(\boldsymbol{\pi}) + \mathit{diag}(\boldsymbol{\pi}) \cdot \boldsymbol{\Sigma}_{\boldsymbol{\mu},\boldsymbol{\mu}} \cdot \mathit{diag}(\boldsymbol{\pi})\,. \tag{3.1b}$$

The above result allows us to precisely characterize the behavior of the averaging estimator in the high-dimensional regime. Let us consider a few special cases.

Two classes. Consider the special case with two classes with class priors $\pi_1 = 1-\pi_2 =: \pi$. In this case we can compute the class-wise misclassification probabilities $\mathbb{P}_{e|1}$ and $\mathbb{P}_{e|2}$ explicitly. Specifically using (3.1), we have $\boldsymbol{S}_1 = \|\pi\boldsymbol{\mu}_1 - (1-\pi)\boldsymbol{\mu}_2\|_{\ell_2}^2 + \gamma\sigma^2$ and $t_1 = (1-2\pi) + (1-\pi)\langle\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\rangle - \pi\,\|\boldsymbol{\mu}_1\|_{\ell_2}^2$. Substituting the latter two in (2.9) we arrive at $\mathbb{P}_{e|1} \xrightarrow{P} Q\big(\frac{\pi\|\boldsymbol{\mu}_1\|_{\ell_2}^2 - (1-\pi)\langle\boldsymbol{\mu}_1,\boldsymbol{\mu}_2\rangle + 2\pi - 1}{\sqrt{\|\pi\boldsymbol{\mu}_1 - (1-\pi)\boldsymbol{\mu}_2\|_{\ell_2}^2 + \gamma\sigma^2}}\big)$. In the case of equal priors $\pi = \pi_1 = \pi_2 = 1/2$, antipodal and equal energy of the means, i.e. $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2$ and $\mu := \|\boldsymbol{\mu}_1\|_{\ell_2} = \|\boldsymbol{\mu}_2\|_{\ell_2}$, we can use the above to conclude that $\mathbb{P}_{e|1} = \mathbb{P}_{e|2} = \frac{1}{2}\,\mathbb{P}_e = \frac{1}{2}Q\big(\sqrt{\frac{\mu^2}{\mu^2+\gamma\sigma^2}}\big)$. This formula recovers the result of [47] for this special case. Also, as mentioned in [47], the formula matches the Bayes optimal error computed in [39] for Gaussian mean vectors. This shows that the class averaging method is Bayes optimal in this very simple setting. In Section 3.4, we generalize this result to multiple classes: we show that the average estimator is (asymptotically) Bayes optimal for balanced classes and equal-energy Gaussian means for any $k \geq 2$.

Orthogonal means, equal priors and equal energy. Next we focus on a special case with orthogonal means $\langle\boldsymbol{\mu}_i, \boldsymbol{\mu}_j\rangle = 0,\ i \neq j \in [k]$ of equal energy $\mu^2 := \|\boldsymbol{\mu}_i\|_{\ell_2}^2$ and of equal priors $\pi_i = \pi = 1/k$ for $i \in [k]$. In this case, the class-wise miss-classification error converges to $\mathbb{P}_{e|c} \xrightarrow{P} 1 - \mathbb{P}\{\boldsymbol{S}_c^{1/2}\,\boldsymbol{z} > \boldsymbol{t}\}$, where $\boldsymbol{S}_c = \pi(\pi\mu^2 + \gamma\sigma^2)(\boldsymbol{I}_{k-1} + \boldsymbol{1}_{k-1}\boldsymbol{1}_{k-1}^T)$ and $\boldsymbol{t} = -\pi\mu^2\boldsymbol{1}_{k-1}$. Defining $u_{\mathrm{Avg}} := \frac{\mu^2}{\sigma}\sqrt{\frac{1}{\mu^2+k\gamma\sigma^2}}$ after some algebraic manipulations the total classification error of the averaging estimator in this case is given by $\mathbb{P}_{c|e} = \mathbb{P}_{e,\mathrm{Avg}} \xrightarrow{P} \mathbb{P}\left\{G_0 + \max_{j\in[k-1]} G_j \geq u_{\mathrm{Avg}}\right\}$.

3.2 Least-squares classifier

This section focuses on characterizing the intercepts and correlation matrices for the least-squares classifier. To present our results, we assume that the Grammian matrix has eigenvalue decomposition

$$\boldsymbol{\Sigma}_{\boldsymbol{\mu},\boldsymbol{\mu}} = \boldsymbol{M}^T\boldsymbol{M} = \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^T, \qquad \boldsymbol{\Sigma} \succ 0 \in \mathbb{R}^{r\times r},\ \boldsymbol{V} \in \mathbb{R}^{k\times r},\ r \leq k. \tag{3.2}$$

with $\boldsymbol{\Sigma}$ a diagonal positive-definite matrix and $\boldsymbol{V}$ an orthonormal matrix obeying $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_r$.

Theorem 3.2 *Consider data generated according to GMM in an asymptotic regime with $\gamma < 1$. In addition to (3.2), define the following two positive (semi)-definite matrices: $\boldsymbol{P} := \mathit{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T \succeq \boldsymbol{0}_{k\times k}$ and $\boldsymbol{\Delta} := \sigma^2\boldsymbol{I}_r + \boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{P}\boldsymbol{V}\boldsymbol{\Sigma} \succ \boldsymbol{0}_{r\times r}$. Then, for the least-squares linear classifier $\big(\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{b}}\big)$ the following limits are true asymptotically*

$$\widehat{\boldsymbol{b}} \xrightarrow{P} \boldsymbol{\pi} - \boldsymbol{P}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{\pi}\,, \quad \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}} \xrightarrow{P} \boldsymbol{P}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\,, \tag{3.3a}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}} \xrightarrow{P} \frac{\gamma}{(1-\gamma)\sigma^2}\boldsymbol{P} + \boldsymbol{P}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\Big(\boldsymbol{\Delta}^{-1} - \frac{\gamma}{(1-\gamma)\sigma^2}\boldsymbol{I}_r\Big)\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{P}\,. \tag{3.3b}$$

The above result allows us to precisely characterize the behavior of the least-squares classifier in the high-dimensional regime. In the supplementary, we specialize (3.3a) to the case of orthogonal

means. Compared to the weight vectors $\widehat{\boldsymbol{w}}_i, i \in [k]$ of the class averaging classifier that are also (asymptotically) orthogonal when means are orthogonal, this is *not* the case for LS. We show next that these spurious correlations only hurt the classification error when classes are balanced.

Proposition 3.3 *Consider the case of orthogonal, equal energy-means $\boldsymbol{\Sigma}_{\boldsymbol{\mu},\boldsymbol{\mu}} = \mu \boldsymbol{I}_k$, balanced priors $\pi_i = 1/k,\ i \in [k]$ and $\gamma < 1$. Setting $u_{\rm LS} := \frac{\mu^2}{\sigma}\sqrt{\frac{1-\gamma}{\mu^2+k\gamma\sigma^2}}$, it holds that $\mathbb{P}_{e,\rm LS} \xrightarrow{P} \mathbb{P}\left\{G_0 + \max_{j\in[k-1]} G_j \geq u_{\rm LS}\right\}$. Specifically, since $u_{\rm LS} = u_{\rm Avg}\sqrt{1-\gamma} < u_{\rm Avg}$, the averaging estimator strictly outperforms LS for all $0 < \gamma < 1$ and $k \geq 2$ in this setting.*

3.3 Weighted least-squares classifier

We now focus on characterizing the intercepts/correlation matrices for the WLS classifier.

Theorem 3.4 *Consider data generated according to GMM and $\gamma < 1$. Consider a weighted LS classifier with weights $\boldsymbol{D} = diag(\omega_1, \ldots, \omega_k)$ and let η be the unique solution to $\sum_{\ell=1}^k \frac{\pi_\ell \omega_\ell^2}{\omega_\ell^2+\eta} = \gamma$. Also define $\boldsymbol{P} := diag(\widetilde{\boldsymbol{\pi}}) - \widetilde{\boldsymbol{\pi}}\widetilde{\boldsymbol{\pi}}^T \succeq \boldsymbol{0}_{k\times k}$ and $\boldsymbol{\Delta} := \sigma^2 \boldsymbol{I}_r + \boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{P}\boldsymbol{V}\boldsymbol{\Sigma} \succ \boldsymbol{0}_{r\times r}$ with the entries of $\widetilde{\boldsymbol{\pi}}$ given by $\widetilde{\pi}_\ell = \frac{1}{\gamma}\frac{\pi_\ell\omega_\ell^2}{\omega_\ell^2+\eta}$. Then, for the WLS linear classifier $\left(\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{b}}\right)$ the following asymptotic limits hold*

$$\widehat{\boldsymbol{b}} \xrightarrow{P} \widetilde{\boldsymbol{\pi}} - \boldsymbol{P}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\widetilde{\boldsymbol{\pi}}\,, \quad \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}} \xrightarrow{P} \boldsymbol{P}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\,, \tag{3.4a}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}} \xrightarrow{P} \frac{\zeta}{\sigma^2}\boldsymbol{P} + \boldsymbol{P}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\Big(\boldsymbol{\Delta}^{-1} - \frac{\zeta}{\sigma^2}\boldsymbol{I}_r\Big)\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{P} + \frac{\eta\zeta}{\sigma^2}\boldsymbol{Q}\,. \tag{3.4b}$$

Here, $\zeta := \frac{\gamma}{\eta\left(\sum_{\ell=1}^k \frac{\pi_\ell\omega_\ell^2}{(\omega_\ell^2+\eta)^2}\right)}$ and $\boldsymbol{Q} \in \mathbb{R}^{k\times k}$ is a known matrix depending on various problem parameters (precise value discussed in the proof).

Surprisingly, the effect of the weights is essentially equivalent to adjusting the class priors from $\boldsymbol{\pi}$ to $\widetilde{\boldsymbol{\pi}}$ defined in the theorem (modulo the extra additive term in the cross correlation matrix $\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}}$). This shows that weighted LS has similar performance to an un-weighted LS applied to a model with different class priors $\widetilde{\boldsymbol{\pi}}$. This characterization allows us to precisely understand how different weighting schemes can alter test accuracy for rare/minority classes.

3.4 Bayes estimator for the balanced Gaussian Mixture Model

To check how far the above algorithms are from the lowest misclassification error achievable by any algorithm in this section, we consider a Bayesian setting with Gaussian mean vectors and we derive the Bayes-optimal risk for the case of equal priors. Recall that the Bayes estimator $\hat{Y} = \arg\max_{\ell\in[k]} \mathbb{P}\{Y = \ell \mid \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{x}\}$ minimizes the risk $\mathbb{P}_e = \mathbb{P}\{\hat{Y} \neq Y\} = \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{x},Y}\left[\mathbb{1}[\hat{Y} \neq Y]\right]$.

Proposition 3.5 *Consider $\boldsymbol{\mu}_i \overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \frac{\mu^2}{d}\mathrm{I}_d)$ and $\pi_i = 1/k$ for all $i \in [k]$. Set $u_{\rm Bayes} := \frac{\mu^2}{\sigma}\frac{1}{\sqrt{\mu^2+k\gamma\sigma^2}}$. Then, the Bayes risk converges to $\mathbb{P}\left\{G_0 + \max_{\ell\in[k-1]} G_\ell \geq u_{\rm Bayes}\right\}$.*

Under the Gaussian prior, the means are asymptotically orthogonal and equal-energy. As shown earlier, in this setting, $\mathbb{P}_{e,\rm Avg} \xrightarrow{P} \mathbb{P}\left\{G_0 + \max_{\ell\in[k-1]} G_\ell \geq u_{\rm Avg}\right\}$. But, $u_{\rm Avg} = u_{\rm Bayes}$. Thus, the averaging method is (asymptotically) Bayes optimal for equal-norm, orthogonal means and balanced classes. An analogous result was recently derived in [39, 47], but only for binary classification.

4 Results for Multinomial Logit Model

In this section we discuss the asymptotics of the intercepts/correlation matrices for MLM. We present results for arbitrary mean-vectors as well as special cases where the means are mutually orthogonal. Recall the eigenvalue decomposition of the Grammian $\boldsymbol{\Sigma}_{\boldsymbol{\mu},\boldsymbol{\mu}} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^T$ in (3.2). In order to state our results, it is convenient to introduce the following probability vectors in $\mathbb{R}^k$ and $\mathbb{R}^{k^2}$:

$$\boldsymbol{\pi} := \mathbb{E}\Big[\frac{e^{\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}}{\boldsymbol{1}_k^T e^{\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}}\Big] \in \mathbb{R}^k \ \text{ and } \ \boldsymbol{\Pi} := \mathbb{E}\Big[\frac{\left(e^{\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}\right)\left(e^{\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}\right)^T}{\left(\boldsymbol{1}_k^T e^{\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}\right)^2}\Big] \in \mathbb{R}^{k\times k}, \ \text{ where } \boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}_r, \boldsymbol{I}_r). \tag{4.1}$$

Note that $\boldsymbol{\pi}$ and $\boldsymbol{\Pi}$ are the first and second moments of the soft-max mapping of $\boldsymbol{V\Sigma g} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\mu},\boldsymbol{\mu}})$. In fact, for the MLM in (2.2) it holds that $\mathbb{P}\{Y = \ell\} = \mathbb{E}[\mathbb{P}\{Y = \ell \,|\, \boldsymbol{x}\}] = \mathbb{E}[\frac{e^{\boldsymbol{e}_\ell^T \boldsymbol{V\Sigma g}}}{\mathbf{1}_k^T e^{\boldsymbol{V\Sigma g}}}] = \boldsymbol{\pi}_\ell,\ \ell \in [k]$ since $\boldsymbol{M}^T\boldsymbol{x}$ is distributed as $\boldsymbol{V\Sigma g}$. Thus, $\boldsymbol{\pi}$ is the vector of class priors (which explains the slight abuse of notation here in relation to our notation for the class priors of the GMM).

4.1 Class averaging classifier

Proposition 4.1 *Consider data generated according to MLM in an asymptotic regime with any $\gamma > 0$. For the averaging classifier, the following high-dimensional limits hold*

$$\widehat{\boldsymbol{b}} \xrightarrow{P} \boldsymbol{\pi}\,, \quad \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}} \xrightarrow{P} (\mathit{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi}) \cdot \boldsymbol{\Sigma}_{\boldsymbol{\mu},\boldsymbol{\mu}}\,, \tag{4.2a}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}} \xrightarrow{P} \gamma \cdot \mathit{diag}(\boldsymbol{\pi}) + (\mathit{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi}) \cdot \boldsymbol{\Sigma}_{\boldsymbol{\mu},\boldsymbol{\mu}} (\mathit{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi})\,. \tag{4.2b}$$

Using Gaussian decomposition in (2.8) and checking from (4.2) that $\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}} - \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}}\boldsymbol{\Sigma}^{\dagger}_{\boldsymbol{\mu},\boldsymbol{\mu}}\boldsymbol{\Sigma}^T_{\boldsymbol{w},\boldsymbol{\mu}} \xrightarrow{P} \gamma \cdot \mathrm{diag}(\boldsymbol{\pi})$ the test error obtains the following explicit form:

$$\mathbb{P}_{e,\mathrm{Avg}} \xrightarrow{P} \mathbb{P}\left\{\arg\max\left\{\sqrt{\gamma} \cdot \mathrm{diag}(\sqrt{\boldsymbol{\pi}}) \cdot \widetilde{\boldsymbol{g}} + (\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi}) \cdot \boldsymbol{V\Sigma} \cdot \boldsymbol{g} + \boldsymbol{\pi}\right\} \neq Y(\boldsymbol{g})\right\}, \tag{4.3}$$

where $\widetilde{\boldsymbol{g}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_k)$, $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_r)$ and $\mathbb{P}\{Y(\boldsymbol{g}) = c\} = e^{\boldsymbol{e}_c^T \boldsymbol{V\Sigma g}} \big/ \sum_{j\in[k]} e^{\boldsymbol{e}_j^T \boldsymbol{V\Sigma g}}$, $c \in [k]$.

4.2 Least-squares classifier

This section focuses on characterizing the intercepts and correlation matrices for the least-squares classifier. We also use the result to characterize conditions under which LS outperforms averaging.

Theorem 4.2 *Consider data generated according to MLM in an asymptotic regime with $0 < \gamma < 1$. Recall the notation in (4.1). For the LS classifier, the following high-dimensional limits hold.*

$$\widehat{\boldsymbol{b}} \xrightarrow{P} \boldsymbol{\pi}\,, \quad \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}} \xrightarrow{P} (\mathit{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi}) \cdot \boldsymbol{\Sigma}_{\boldsymbol{\mu},\boldsymbol{\mu}}\,, \tag{4.4a}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}} \xrightarrow{P} \frac{\gamma}{1-\gamma} \cdot \left(\mathit{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T\right) + \frac{1-2\gamma}{1-\gamma} \cdot (\mathit{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi}) \cdot \boldsymbol{\Sigma}_{\boldsymbol{\mu},\boldsymbol{\mu}} \cdot (\mathit{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi})\,. \tag{4.4b}$$

It is interesting to observe that (4.4a) is identical to (4.2a). However, the cross-correlations in $\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}}$ differ. We prove below that this leads to an improved performance of the LS classifier for large sample sizes. First, Theorem 4.2 can be used to check that

$$\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}} - \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}}\boldsymbol{\Sigma}^{\dagger}_{\boldsymbol{\mu},\boldsymbol{\mu}}\boldsymbol{\Sigma}^T_{\boldsymbol{w},\boldsymbol{\mu}} \xrightarrow{P} \frac{\gamma}{1-\gamma}\left(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T - (\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi})\,\boldsymbol{\Sigma}_{\boldsymbol{\mu},\boldsymbol{\mu}}\,(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi})\right).$$

Thus, the only change in the test error formula compared to (4.3) is the term $\gamma \cdot \mathrm{diag}(\boldsymbol{\pi})$ substituted by the matrix above.

Proposition 4.3 *Assume orthogonal, equal-energy means $\boldsymbol{\Sigma}_{\boldsymbol{\mu},\boldsymbol{\mu}} = \mu^2 \boldsymbol{I}_k$, $k \geq 2$. Let $\gamma_\star = \frac{\mu^2 k}{(k-1)^2}\big(1 - k\,\mathbb{E}\big[\frac{e^{2\mu G_1}}{(\sum_{\ell\in[k]} e^{\mu G_\ell})^2}\big]\big)^2 \in (0,1)$. Then, with probability 1 as $n \to \infty$, $\mathbb{P}_{e,\mathrm{LS}} < \mathbb{P}_{e,\mathrm{Avg}} \iff \gamma < \gamma_\star$.*

4.3 Weighted least-squares classifier

Theorem 4.4 predicts the asymptotic performance of *weighted* least-squares for data generated according to MLM.

Theorem 4.4 *Consider data generated according to MLM and $\gamma < 1$. Consider a weighted LS classifier with weights $\boldsymbol{D} = \mathit{diag}(\omega_1, \ldots, \omega_k)$ and let η be the unique solution to $\sum_{\ell=1}^k \frac{\pi_\ell \omega_\ell^2}{\omega_\ell^2 + \eta} = \gamma$. Also define vector $\boldsymbol{\nu} \in \mathbb{R}^k$ with entries given by $\boldsymbol{\nu}_\ell = \frac{1}{\gamma}\frac{\omega_\ell^2}{\omega_\ell^2+\eta}$ and positive semidefinite matrix*

$$\boldsymbol{\Delta} = \mathbb{E}\left[\left(\boldsymbol{\nu}^T\boldsymbol{v}\right)\boldsymbol{g}\boldsymbol{g}^T\right] - \boldsymbol{\Sigma}\boldsymbol{V}^T\,(\mathit{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi})\,\boldsymbol{\nu}\boldsymbol{\nu}^T\,(\mathit{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi})\,\boldsymbol{V\Sigma} \succ \mathbf{0}_{r\times r}, \tag{4.5}$$

where $\boldsymbol{v} \in \mathbb{R}^k$ *is a random vector with entries* $V_\ell = \frac{e^{\boldsymbol{e}_\ell^T \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}}{\sum_{\ell' \in [k]} e^{\boldsymbol{e}_{\ell'}^T \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}}$ *for* $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_r)$*. Then, for the WLS linear classifier* $\left(\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{b}}\right)$ *the following asymptotic limits hold*

$$\widehat{\boldsymbol{b}} \xrightarrow{P} diag(\boldsymbol{\nu})\boldsymbol{\pi} - diag(\boldsymbol{\nu})\left(\boldsymbol{I}_k - \boldsymbol{\pi}\boldsymbol{\nu}^T\right)(diag(\boldsymbol{\pi}) - \boldsymbol{\Pi})\,\boldsymbol{V}\boldsymbol{\Sigma}\widetilde{\boldsymbol{\Delta}}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\,(diag(\boldsymbol{\pi}) - \boldsymbol{\Pi})\,\boldsymbol{\nu}. \tag{4.6a}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}} \xrightarrow{P} diag(\boldsymbol{\nu})\left(\boldsymbol{I}_k - \boldsymbol{\pi}\boldsymbol{\nu}^T\right)(diag(\boldsymbol{\pi}) - \boldsymbol{\Pi})\,\boldsymbol{V}\boldsymbol{\Sigma}\widetilde{\boldsymbol{\Delta}}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{4.6b}$$

The corresponding formula for the asymptotic limit of the cross-correlation matrix $\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}}$ *is given in* (J.39) *in Section J.*

To see how the theorem above includes Theorem 4.2 as a special case, consider the case $\omega_i = 1,\ i \in [k]$. We show how this recovers the solution for (un-weighted) LS. First, note that in this case solving for η simply gives $\eta = \frac{1}{\gamma} - 1$. Thus, $\boldsymbol{\nu} = \boldsymbol{1}_k$ and $\widetilde{\boldsymbol{\pi}} = \boldsymbol{\pi}$. Also, observe in (4.1) that $(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi})\,\boldsymbol{1}_k = \boldsymbol{0}$ and $\boldsymbol{1}^T\boldsymbol{v} = 1$. Thus, (4.5) reduces to $\boldsymbol{\Delta} = \mathbb{E}[\boldsymbol{g}\boldsymbol{g}^T] = \boldsymbol{I}_r$. With these, it can be readily checked that (4.6a) and (4.6b) simplify to the expressions in (4.4a).

5 Numerical Results

This section validates our theory via numerical experiments and provides further insights on multiclass classification. See also the supplementary for more extensive experiments. We study the class-wise/total test misclassification error in both GMM and MLM for different sample sizes, number of classes and class priors. In line with Section 2.2 we consider four algorithms: (i) Averaging (Avg), (ii) LS, (iii) Weighted LS (WLS) with the ith class weighted by $\omega_\ell^2 = 1/\pi_\ell$, (iv) Cross-Entropy (CE).

Figures 1 and 2 focus on GMM with $k = 9$ classes, $d = 300$ and $\|\mu_i\|_{\ell_2}^2 = 15$. To model different class prior probabilities, we use the distribution $\pi_1 = \pi_2 = \pi_3 = 0.5\pi_4 = 0.5\pi_5 = 0.5\pi_6 = 0.25\pi_7 = 0.25\pi_8 = 0.25\pi_9 = 1/21$. We consider three scenarios: (a) orthogonal means, equal prior ($\pi_i = 1/9$); (b) orthogonal means, different prior; (c) correlated means with pairwise correlation coefficient equal to 0.5 (i.e., $\langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_j \rangle / (\|\boldsymbol{\mu}_i\|_{\ell_2} \|\boldsymbol{\mu}_j\|_{\ell_2}) = 0.5$ for $i \neq j$) and different priors as discussed above. Figure 1 shows the test miss-classification errors as a function of $\gamma := d/n$. In all scenarios our theoretical predictions are a near perfect match to the empirical performance. In scenario (a), class-wise averaging achieves the lowest error as predicted by Proposition 3.5. However, in scenario (b) where the means have different norms the averaging method has higher misclassification error compared with CE, LS and WLS for large sample sizes (small γ). We note that both LS and WLS achieve lower errors compared with CE as the sample size grows. Scenario (c) is similar to (b) however due to class correlations, the errors are uniformly higher. Figure 2 shows the corresponding class-wise miss-classification errors for the smallest γ in Figure 1 ($\gamma = 0.117$). In scenario (a), errors are equal which is expected given the equal class priors. In scenarios (b) and (c) however, due to different priors, large classes 7,8,9 achieve best accuracy. The performance difference is most visible for the averaging approach. LS mitigates this issue to some extent, while WLS creates the flattest class-wise errors suggesting that it can reduce the miss-classification error on small/minority classes.

Figure 3 focuses on orthogonal classes with varying number of classes k where $\|\mu_i\|_{\ell_2}^2 = 15$ and $d \in \{50, 100, 200\}$ with $kd/n = k\gamma$ fixed at $k\gamma = 20/11$. It plots the ratio of the empirical error probability and our theoretical prediction as k grows until $k = d$. Two observations are worth mentioning here. (1) The accuracy of our predictions noticeably improves as the problem dimension d, n grow as expected given the asymptotic nature of our analysis. Interestingly, the convergence appears to be noticeably faster (as a function of d) for the LS rather than the Averaging classifier. (2) Our theoretical results formally require that k is fixed while d (and n) grow large. Yet, the presented experimental results suggest that they might also hold for large k under the shown scaling. This is a fascinating research question that we believe is worth investigating further.

Figure 4 provides experiments on the MLM model with $k = 9$ orthogonal classes. Unlike GMM, CE achieves the best performance in MLM. In Figure 4 (a), classes have same norms $\|\mu_i\|_{\ell_2} = 10$, while in Figure 4 (b) we have quadrupled the norms of classes 7,8,9 and doubled the norms of classes 4,5,6. This disparity between the norms seems to help improve the CE accuracy, but hurt LS/averaging accuracy for small γ. Finally, Figure 4 (c) shows the class-wise probability of error associated with (b) for $\gamma = 0.117$ and demonstrates that LS outperforms averaging.

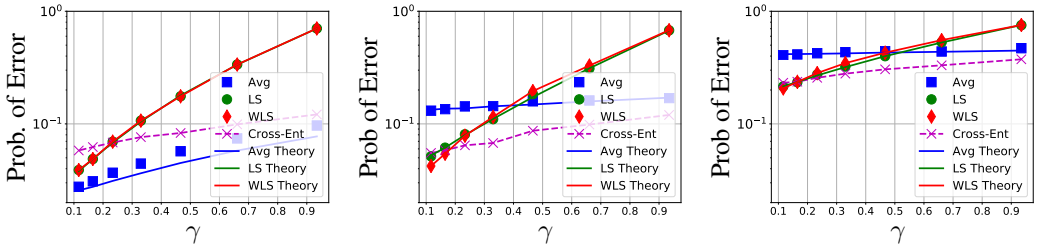

Figure 1: GMM with $k = 9, d = 300$. (a) orthogonal, equal prior, (b) orthogonal, different prior, (c) correlated, different prior.

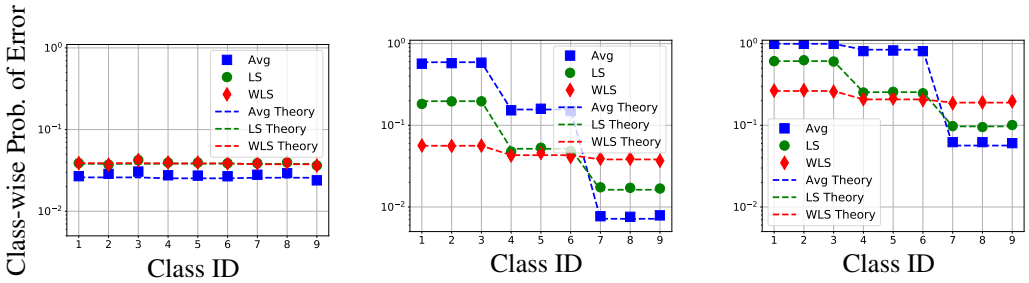

Figure 2: Class-wise probability of errors corresponding to Figure 1 with $\gamma = 0.117$.

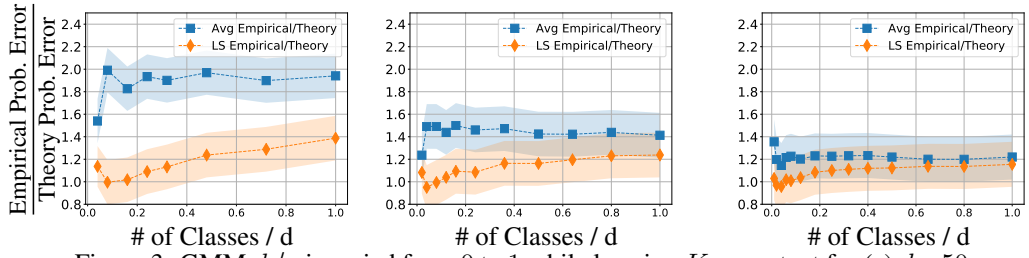

Figure 3: GMM, k/p is varied from 0 to 1 while keeping $K\gamma$ constant for (a) $d = 50$, (b) $d = 100$, (c) $d = 200$.

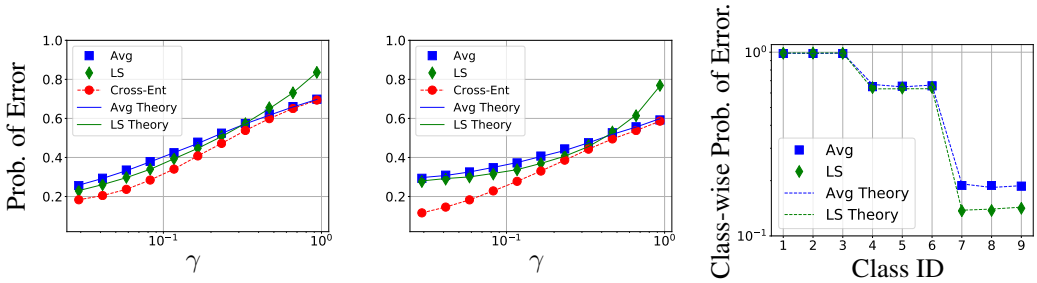

Figure 4: MLM with orthogonal means for (a) equal norms and (b) different norms. (c) Class-wise probability errors for (b).

6 Future Directions

This work aims at initiating a precise asymptotic study of multiclass classifiers that provides a promising setting for resolving a rich set of open questions regarding the (comparative) performance of classification algorithms as a function of the involved problem variables. As mentioned, even understanding the statistical performance of one-vs-all multiclass classifiers does not follow directly from the existing literature on binary classifiers. Extending the results of this paper to the one-vs-all logistic and SVM classifiers would allow for a principled comparison among these different choices. A possibly more challenging, albeit mathematically intriguing and practically relevant task, is characterizing the asymptotics of more complicated (non-separable) losses, such as the cross-entropy loss. For this, even characterizing the asymptotic behavior of the correlations $\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}}$

requires new ideas. The previously mentioned study of "extreme multiclass classification" in which the number of classes k is very large is another very fascinating direction.

Broader Impact

In this paper we develop a precise and asymptotically exact understanding of the statistical behavior of a variety of classification algorithms. In particular we precisely, characterize how the total and class-wise accuracy varies under different training algorithms, data distributions, problem dimensions, inter/intra class correlations and class priors. Despite being theoretical/foundational in nature it has potential for broader practical impact. In particular, our precise characterization of class-wise accuracy allows us to understand how different training algorithms impact accuracy of machine learning algorithms on rare/minority classes. Such a precise understanding may help guide the development of more fair/equitable algorithms. On the flip side, such insights may potentially also be used nefariously enabling the marginalization of rare/minority classes by developing algorithms that reduce their class-wise accuracy.

References

[1] ALLWEIN, E. L., SCHAPIRE, R. E., AND SINGER, Y. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of machine learning research 1*, Dec (2000), 113–141.

[2] AMELUNXEN, D., LOTZ, M., MCCOY, M. B., AND TROPP, J. A. Living on the edge: A geometric theory of phase transitions in convex optimization. *arXiv preprint arXiv:1303.6672* (2013).

[3] BAYATI, M., AND MONTANARI, A. The lasso risk for gaussian matrices. *Information Theory, IEEE Transactions on 58*, 4 (2012), 1997–2017.

[4] BOSMAN, A. S., ENGELBRECHT, A., AND HELBIG, M. Visualising basins of attraction for the cross-entropy and the squared error neural network loss functions. *Neurocomputing* (2020).

[5] BREDENSTEINER, E. J., AND BENNETT, K. P. Multicategory classification by support vector machines. In *Computational Optimization*. Springer, 1999, pp. 53–79.

[6] BU, Z., KLUSOWSKI, J., RUSH, C., AND SU, W. Algorithmic analysis and statistical estimation of slope via approximate message passing. In *Advances in Neural Information Processing Systems* (2019), pp. 9361–9371.

[7] CANDÈS, E. J., AND SUR, P. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *arXiv preprint arXiv:1804.09753* (2018).

[8] CELENTANO, M., AND MONTANARI, A. Fundamental barriers to high-dimensional regression with convex penalties. *arXiv preprint arXiv:1903.10603* (2019).

[9] CHANDRASEKARAN, V., RECHT, B., PARRILO, P. A., AND WILLSKY, A. S. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics 12*, 6 (2012), 805–849.

[10] CHOROMANSKA, A., AGARWAL, A., AND LANGFORD, J. Extreme multi class classification. In *NIPS Workshop: eXtreme Classification, submitted* (2013).

[11] CORTES, C., KUZNETSOV, V., MOHRI, M., AND YANG, S. Structured prediction theory based on factor graph complexity. In *Advances in Neural Information Processing Systems* (2016), pp. 2514–2522.

[12] COVINGTON, P., ADAMS, J., AND SARGIN, E. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems* (2016), pp. 191–198.

[13] CRAMMER, K., AND SINGER, Y. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research 2*, Dec (2001), 265–292.

[14] DEMIRKAYA, A., CHEN, J., AND OYMAK, S. Exploring the role of loss functions in multiclass classification. In *2020 54th Annual Conference on Information Sciences and Systems (CISS)* (2020), IEEE, pp. 1–5.

[15] DENG, Z., KAMMOUN, A., AND THRAMPOULIDIS, C. A model of double descent for high-dimensional binary linear classification. *arXiv preprint arXiv:1911.05822* (2019).

[16] DIETTERICH, T. G., AND BAKIRI, G. Solving multiclass learning problems via error-correcting output codes. *Journal of artificial intelligence research 2* (1994), 263–286.

[17] DONOHO, D., AND MONTANARI, A. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields 166*, 3-4 (2016), 935–969.

[18] DONOHO, D. L., MALEKI, A., AND MONTANARI, A. The noise-sensitivity phase transition in compressed sensing. *Information Theory, IEEE Transactions on 57*, 10 (2011), 6920–6941.

[19] EL KAROUI, N. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields 170*, 1-2 (2018), 95–175.

[20] FÜRNKRANZ, J. Round robin classification. *Journal of Machine Learning Research 2*, Mar (2002), 721–747.

[21] GAJOWNICZEK, K., CHMIELEWSKI, L. J., ORŁOWSKI, A., AND ZĄBKOWSKI, T. Generalized entropy cost function in neural networks. In *International Conference on Artificial Neural Networks* (2017), Springer, pp. 128–136.

[22] GORDON, Y. *On Milman's inequality and random subspaces which escape through a mesh in* $\mathbb{R}^n$. Springer, 1988.

[23] GUERMEUR, Y. Combining discriminant models with new multi-class svms. *Pattern Analysis & Applications 5*, 2 (2002), 168–179.

[24] HASHORVA, E., AND HÜSLER, J. On multivariate gaussian tails. *Annals of the Institute of Statistical Mathematics 55*, 3 (2003), 507–522.

[25] HOU, L., YU, C.-P., AND SAMARAS, D. Squared earth mover's distance-based loss for training deep neural networks. *arXiv preprint arXiv:1611.05916* (2016).

[26] HU, H., AND LU, Y. M. Asymptotics and optimal designs of slope for sparse linear regression. *arXiv preprint arXiv:1903.11582* (2019).

[27] HUANG, H. Asymptotic behavior of support vector machine for spiked population model. *The Journal of Machine Learning Research 18*, 1 (2017), 1472–1492.

[28] JANG, E., GU, S., AND POOLE, B. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).

[29] JAVANMARD, A., SOLTANOLKOTABI, M., AND HASSANI, H. Precise tradeoffs in adversarial training for linear regression. *arXiv preprint arXiv:2002.10477* (2020).

[30] KAMMOUN, A., AND ALOUINI, M.-S. On the precise error analysis of support vector machines. *Submitted to IEEE Transactions on information theory* (2019).

[31] KAROUI, N. E. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445* (2013).

[32] KOLTCHINSKII, V., PANCHENKO, D., ET AL. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics 30*, 1 (2002), 1–50.

[33] KUMAR, H., AND SASTRY, P. Robust loss functions for learning multi-class classifiers. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (2018), IEEE, pp. 687–692.

[34] KUZNETSOV, V., MOHRI, M., AND SYED, U. Rademacher complexity margin bounds for learning with a large number of classes. In *ICML Workshop on Extreme Classification: Learning with a Very Large Number of Labels* (2015).

[35] LEDOUX, M., AND TALAGRAND, M. *Probability in Banach Spaces: isoperimetry and processes*, vol. 23. Springer, 1991.

[36] LEE, Y., LIN, Y., AND WAHBA, G. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association 99*, 465 (2004), 67–81.

[37] LEI, Y., DOGAN, U., BINDER, A., AND KLOFT, M. Multi-class svms: From tighter data-dependent generalization bounds to novel algorithms. In *Advances in Neural Information Processing Systems* (2015), pp. 2035–2043.

[38] LEI, Y., DOGAN, Ü., ZHOU, D.-X., AND KLOFT, M. Data-dependent generalization bounds for multi-class classification. *IEEE Transactions on Information Theory 65*, 5 (2019), 2995–3021.

[39] LELARGE, M., AND MIOLANE, L. Asymptotic bayes risk for gaussian mixture in a semi-supervised setting. *arXiv preprint arXiv:1907.03792* (2019).

[40] LI, J., LIU, Y., YIN, R., ZHANG, H., DING, L., AND WANG, W. Multi-class learning: from theory to algorithm. In *Advances in Neural Information Processing Systems* (2018), pp. 1586–1595.

[41] LIANG, T., AND SUR, P. A precise high-dimensional asymptotic theory for boosting and min-l1-norm interpolated classifiers. *arXiv preprint arXiv:2002.01586* (2020).

[42] LOLAS, P. Regularization in high-dimensional regression and classification via random matrix theory. *arXiv preprint arXiv:2003.13723* (2020).

[43] MAI, X., LIAO, Z., AND COUILLET, R. A large scale analysis of logistic regression: asymptotic performance and new insights. In *ICASSP* (2019).

[44] MAI, X., LIAO, Z., AND COUILLET, R. A large scale analysis of logistic regression: Asymptotic performance and new insights. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019), IEEE, pp. 3357–3361.

[45] MAURER, A. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory* (2016), Springer, pp. 3–17.

[46] MEI, J., XIAO, C., SZEPESVARI, C., AND SCHUURMANS, D. On the global convergence rates of softmax policy gradient methods. *arXiv preprint arXiv:2005.06392* (2020).

[47] MIGNACCO, F., KRZAKALA, F., LU, Y. M., AND ZDEBOROVÁ, L. The role of regularization in classification of high-dimensional noisy gaussian mixture. *arXiv preprint arXiv:2002.11544* (2020).

[48] MIOLANE, L., AND MONTANARI, A. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *arXiv preprint arXiv:1811.01212* (2018).

[49] MONTANARI, A., RUAN, F., SOHN, Y., AND YAN, J. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544* (2019).

[50] OYMAK, S., RECHT, B., AND SOLTANOLKOTABI, M. Sharp time–data tradeoffs for linear inverse problems. *IEEE Transactions on Information Theory 64*, 6 (2017), 4129–4158.

[51] OYMAK, S., THRAMPOULIDIS, C., AND HASSIBI, B. The squared-error of generalized lasso: A precise analysis. *arXiv preprint arXiv:1311.0830* (2013).

[52] PAL, M., AND MATHER, P. Support vector machines for classification in remote sensing. *International journal of remote sensing 26*, 5 (2005), 1007–1011.

[53] PIRES, B. Á., AND SZEPESVÁRI, C. Multiclass classification calibration functions. *arXiv preprint arXiv:1609.06385* (2016).

[54] PIRES, B. A., SZEPESVARI, C., AND GHAVAMZADEH, M. Cost-sensitive multiclass classification risk bounds. In *International Conference on Machine Learning* (2013), pp. 1391–1399.

[55] RAWAT, A. S., CHEN, J., YU, F. X. X., SURESH, A. T., AND KUMAR, S. Sampled softmax with random fourier features. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 13834–13844.

[56] RIFKIN, R., AND KLAUTAU, A. In defense of one-vs-all classification. *Journal of machine learning research 5*, Jan (2004), 101–141.

[57] SALEHI, F., ABBASI, E., AND HASSIBI, B. The impact of regularization on high-dimensional logistic regression. *arXiv preprint arXiv:1906.03761* (2019).

[58] SATHE, Y., AND LINGRAS, S. A note on the inequalities for tail probability of the multivariate normal distribution. *Communications in Statistics-Theory and Methods 9*, 7 (1980), 711–715.

[59] STOJNIC, M. Various thresholds for ℓ_1-optimization in compressed sensing. *arXiv preprint arXiv:0907.3666* (2009).

[60] STOJNIC, M. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291* (2013).

[61] SUR, P., AND CANDÈS, E. J. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences 116*, 29 (2019), 14516–14525.

[62] SUTSKEVER, I., VINYALS, O., AND LE, Q. V. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (2014), pp. 3104–3112.

[63] TAHERI, H., PEDARSANI, R., AND THRAMPOULIDIS, C. Sharp asymptotics and optimal performance for inference in binary models. *arXiv preprint arXiv:2002.07284* (2020).

[64] TEWARI, A., AND BARTLETT, P. L. On the consistency of multiclass classification methods. *Journal of Machine Learning Research 8*, May (2007), 1007–1025.

[65] THRAMPOULIDIS, C., ABBASI, E., AND HASSIBI, B. Lasso with non-linear measurements is equivalent to one with linear measurements. In *Advances in Neural Information Processing Systems* (2015), pp. 3420–3428.

[66] THRAMPOULIDIS, C., OYMAK, S., AND HASSIBI, B. Regularized linear regression: A precise analysis of the estimation error. In *Proceedings of The 28th Conference on Learning Theory* (2015), pp. 1683–1709.

[67] WANG, S., WENG, H., AND MALEKI, A. Does slope outperform bridge regression? *arXiv preprint arXiv:1909.09345* (2019).

[68] WESTON, J., AND WATKINS, C. Multi-class support vector machines. Tech. rep., Citeseer, 1998.

[69] WISLER, A., BERISHA, V., WEI, D., RAMAMURTHY, K., AND SPANIAS, A. Empirically-estimable multi-class classification bounds. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2016), IEEE, pp. 2594–2598.

[70] YEN, I. E.-H., HUANG, X., RAVIKUMAR, P., ZHONG, K., AND DHILLON, I. Pd-sparse: A primal and dual sparse approach to extreme multiclass and multilabel classification. In *International Conference on Machine Learning* (2016), pp. 3069–3077.

[71] ZHANG, T. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research 5*, Oct (2004), 1225–1251.

Contents

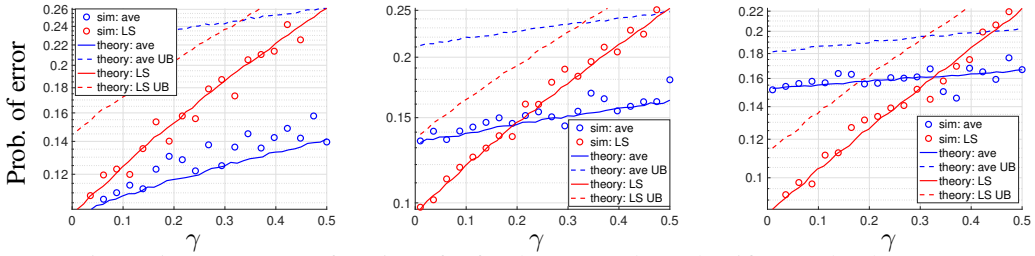

Figure 5: Test error a a function of γ for the Avg and LS classifiers under the GMM. Precise theoretical predictions compared to a theoretical upper bound (UB) obtained by Union bound that does not require knowledge of the off-diagonal entries of $\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}}$. (a) $\pi_1 = \pi_2 = \pi_3 = 1/3$; (b) $2\pi_1 = \pi_2 = 2\pi_3$; (c) $4\pi_1 = \pi_2 = 4\pi_3$

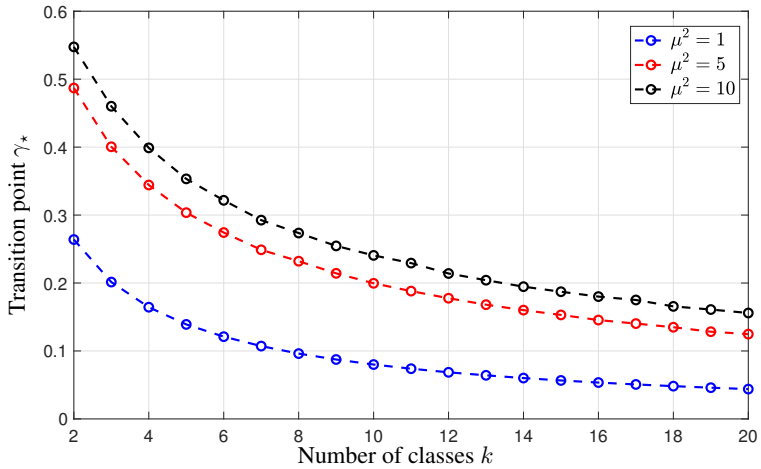

Figure 6: The threshold $\gamma_\star$ of Proposition 4.3 as a function of the number of classes k and the means' energy μ. LS provably outperforms class-averaging for $\gamma < \gamma_\star$.

A Additional Numerical Results

In this section, we provide further numerical experiments. The first fundamental question we ask is

To what extent pairwise class correlations are necessary for performance prediction?

In Figure 5, we attempt to answer this question in a GMM setup with $k = 3$ classes. The solid lines are exact performance predictions based on our theory for averaging and least-squares estimators. The dashed lines are the theoretical upper bounds which does not require the knowledge of cross-correlations between the classes (i.e. off-diagonal entries of $\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}}$ are unknown). These bounds are calculated by considering the *worst case correlation matrix* $\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}}$ given $\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}}$ and the self-correlations of the classes i.e. the diagonal entries of $\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}}$. Both of these information can be obtained by studying the properties of isolated least-squares on individual classes without understanding their pairwise relations.

Figure 5 has two important takeaways. The first takeaway is the fact that pairwise correlations are indeed critical for exact asymptotic analysis and naive approaches cannot reproduce the results of our novel multiclass learning framework. There is a visible gap between upper and exact bounds and this gap is particularly more visible for the averaging estimator. The second takeaway is that our upper bounds are surprisingly strong. In most regimes, our upper bound is within a factor of 1.5 to 2 of the actual bound. These observations suggest a simpler method to calculate reasonably strong theoretical bounds.

The second fundamental question is

When does least-squares provably outperform averaging?

Our Proposition 4.3 provides a fundamental transition point in sample complexity above which least-squares is provably better than averaging under MLM. In Figure 6, we visualize $\gamma_\star$ as a function of different number of classes as well as different levels of mean energy. Least-squares outperform averaging in the region below the lines displayed in Figure 6. Our key message is that least-squares work better when the sample complexity is higher and the problem is less noisy. As the number of classes k increase, the problem becomes more difficult/noisy and we require a larger sample complexity to ensure classifier achieves a similar amount of accuracy as small k. Following this intuition, as k increases, $\gamma_\star$ shifts smaller due to larger sample requirement. Similarly energy μ directly controls the noise level of the problem i.e. larger μ results in a larger signal-to-noise ratio. Thus, as we increase μ, $\gamma_\star$ increases as well because same test accuracy can be achieved with smaller sample size.

B Preliminaries

In this section we gather a few preliminary results that will be used later on in our proofs.

B.1 Slepian's inequality

Lemma B.1 (Slepian's inequality [35]) *Let $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{S})$ and $\tilde{\boldsymbol{g}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{R})$ such that for all $i, j \in [k]$:*

$$\boldsymbol{S}_{ii} = \boldsymbol{R}_{ii}, \qquad \text{and} \qquad \boldsymbol{S}_{ij} \leq \boldsymbol{R}_{ij}.$$

Then, for any $\boldsymbol{t} \in \mathbb{R}^k$ it holds that

$$\mathbb{P}\left\{\bigcup_{j\in[k]} \{\boldsymbol{g}_j \geq \boldsymbol{t}_j\}\right\} \leq \mathbb{P}\left\{\bigcup_{j\in[k]} \{\tilde{\boldsymbol{g}}_j \geq \boldsymbol{t}_j\}\right\}.$$

or equivalently,

$$\mathbb{P}\left\{\bigcap_{j\in[k]} \{\boldsymbol{g}_j \leq \boldsymbol{t}_j\}\right\} \geq \mathbb{P}\left\{\bigcap_{j\in[k]} \{\tilde{\boldsymbol{g}}_j \leq \boldsymbol{t}_j\}\right\}.$$

B.2 Useful results for MLM

Lemma B.2 (Gaussian integration by parts) *Let $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}_r, \boldsymbol{I}_r)$ and random vector $\boldsymbol{v} = [V_1, V_2, \ldots, V_k]^T$ with entries:*

$$\boldsymbol{v} = \frac{e^{\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}}{\mathbf{1}_k^T e^{\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}}, \qquad V_\ell = \frac{e^{\boldsymbol{e}_\ell^T \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}}{\sum_{j\in[k]} e^{\boldsymbol{e}_j^T \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}},\ \ell \in [k]. \tag{B.1}$$

Further recall the notation of $\boldsymbol{\pi}$ and $\boldsymbol{\Pi}$ in (4.1). The following statements are true:

(i) $\mathbb{E}[\boldsymbol{v}] = \boldsymbol{\pi}$.

(ii) For all $i \in [r], \ell \in [k]$:

$$\mathbb{E}[\boldsymbol{g}_i V_\ell] = (\boldsymbol{e}_i^T \boldsymbol{\Sigma}\boldsymbol{V}^T \boldsymbol{e}_\ell)\, \boldsymbol{\pi}_\ell - \boldsymbol{e}_i^T \boldsymbol{\Sigma}\boldsymbol{V}^T \sum_{j\in[k]} \boldsymbol{e}_j \boldsymbol{\Pi}_{ij},$$

and in matrix form:

$$\mathbb{E}\left[\boldsymbol{v}\boldsymbol{g}^T\right] = (diag(\boldsymbol{\pi}) - \boldsymbol{\Pi})\, \boldsymbol{V}\boldsymbol{\Sigma}$$

(ii). $\mathbb{E}\left[\boldsymbol{v}\boldsymbol{g}^T\right] = (diag(\boldsymbol{\pi}) - \boldsymbol{\Pi})\, \boldsymbol{V}\boldsymbol{\Sigma}$.

B.3 Block matrix inversions

Lemma B.3 (Block matrix inversion) *Let $\boldsymbol{T} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{b} \\ \boldsymbol{b}^T & \delta \end{bmatrix}$ be an invertible block matrix. Then*

$$\boldsymbol{T}^{-1} \begin{bmatrix} \mathbf{f} \\ \epsilon \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Delta}^{-1}\left(\mathbf{f} - \frac{\epsilon}{\delta}\boldsymbol{b}\right) \\ \frac{\epsilon}{\delta} - \frac{1}{\delta}\boldsymbol{b}^T \boldsymbol{\Delta}^{-1}\left(\mathbf{f} - \frac{\epsilon}{\delta}\boldsymbol{b}\right) \end{bmatrix} \tag{B.2}$$

where $\boldsymbol{\Delta} = \boldsymbol{A} - \frac{1}{\delta}\boldsymbol{b}\boldsymbol{b}^T \succ \mathbf{0}$ is the Schur complement.

C Calculating and bounding the missclassification error

C.1 MLM

C.1.1 Class-wise miss-classification error

The class-wise misclassification error can be calculated as follows

$$
\begin{aligned}
\mathbb{P}\left\{\widehat{Y} \neq Y | Y = c\right\} &= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{P}\left\{\widehat{Y} \neq Y | Y = c, \boldsymbol{x}\right\}\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\frac{\mathbb{P}\left\{\widehat{Y} \neq Y, Y = c | \boldsymbol{x}\right\}}{\mathbb{P}\left\{Y = c | \boldsymbol{x}\right\}}\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\frac{\mathbb{P}\left\{\widehat{Y} \neq c | \boldsymbol{x}\right\}}{\mathbb{P}\left\{Y = c | \boldsymbol{x}\right\}}\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\frac{\mathbb{P}\left\{\arg\max(\widehat{\boldsymbol{W}}\boldsymbol{x} + \widehat{\boldsymbol{b}}) \neq c | \boldsymbol{x}\right\}}{\mathbb{P}\left\{Y = c | \boldsymbol{x}\right\}}\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\frac{1 - \prod_{\ell \neq c} \mathbb{1}_{\{\widehat{\boldsymbol{w}}_\ell^T \boldsymbol{x} + \widehat{b}_\ell \leq \widehat{\boldsymbol{w}}_c^T \boldsymbol{x} + \widehat{b}_c\}}}{\mathbb{P}\left\{Y = c | \boldsymbol{x}\right\}}\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\left(1 + \sum_{j \neq c} e^{(\boldsymbol{\mu}_j - \boldsymbol{\mu}_c)^T \boldsymbol{x}}\right)\left(1 - \prod_{\ell \neq c} \mathbb{1}_{\{\widehat{\boldsymbol{w}}_\ell^T \boldsymbol{x} + \widehat{b}_\ell \leq \widehat{\boldsymbol{w}}_c^T \boldsymbol{x} + \widehat{b}_c\}}\right)\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[1 - \prod_{\ell \neq c} \mathbb{1}_{\{\widehat{\boldsymbol{w}}_\ell^T \boldsymbol{x} + \widehat{b}_\ell \leq \widehat{\boldsymbol{w}}_c^T \boldsymbol{x} + \widehat{b}_c\}}\right] + \sum_{j \neq c} \mathbb{E}_{\boldsymbol{x}}\left[e^{(\boldsymbol{\mu}_j - \boldsymbol{\mu}_c)^T \boldsymbol{x}}\left(1 - \prod_{\ell \neq c} \mathbb{1}_{\{\widehat{\boldsymbol{w}}_\ell^T \boldsymbol{x} + \widehat{b}_\ell \leq \widehat{\boldsymbol{w}}_c^T \boldsymbol{x} + \widehat{b}_c\}}\right)\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[1 - \prod_{\ell \neq c} \mathbb{1}_{\{\widehat{\boldsymbol{w}}_\ell^T \boldsymbol{x} + \widehat{b}_\ell \leq \widehat{\boldsymbol{w}}_c^T \boldsymbol{x} + \widehat{b}_c\}}\right] + \sum_{j \neq c} e^{\frac{\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_c\|_{\ell_2}^2}{2}} \mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_c, \boldsymbol{I})}\left[1 - \prod_{\ell \neq c} \mathbb{1}_{\{\widehat{\boldsymbol{w}}_\ell^T \boldsymbol{x} + \widehat{b}_\ell \leq \widehat{\boldsymbol{w}}_c^T \boldsymbol{x} + \widehat{b}_c\}}\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[1 - \prod_{\ell \neq c} \mathbb{1}_{\{\widehat{\boldsymbol{w}}_\ell^T \boldsymbol{x} + \widehat{b}_\ell \leq \widehat{\boldsymbol{w}}_c^T \boldsymbol{x} + \widehat{b}_c\}}\right] \\
&\quad + \sum_{j \neq c} e^{\frac{\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_c\|_{\ell_2}^2}{2}} \mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})}\left[1 - \prod_{\ell \neq c} \mathbb{1}_{\{\widehat{\boldsymbol{w}}_\ell^T (\boldsymbol{\mu}_j - \boldsymbol{\mu}_c + \boldsymbol{x}) + \widehat{b}_\ell \leq \widehat{\boldsymbol{w}}_c^T (\boldsymbol{\mu}_j - \boldsymbol{\mu}_c + \boldsymbol{x}) + \widehat{b}_c\}}\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[1 - \prod_{\ell \neq c} \mathbb{1}_{\{(\widehat{\boldsymbol{w}}_\ell - \widehat{\boldsymbol{w}}_c)^T \boldsymbol{x} \leq \widehat{b}_c - \widehat{b}_\ell\}}\right] \\
&\quad + \sum_{j \neq c} e^{\frac{\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_c\|_{\ell_2}^2}{2}} \mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})}\left[1 - \prod_{\ell \neq c} \mathbb{1}_{\{(\widehat{\boldsymbol{w}}_\ell - \widehat{\boldsymbol{w}}_c)^T \boldsymbol{x} \leq (\widehat{\boldsymbol{w}}_\ell - \widehat{\boldsymbol{w}}_c)^T (\boldsymbol{\mu}_c - \boldsymbol{\mu}_j) + \widehat{b}_c - \widehat{b}_\ell\}}\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\prod_{\ell \neq c} \mathbb{1}_{\{(\widehat{\boldsymbol{w}}_\ell - \widehat{\boldsymbol{w}}_c)^T \boldsymbol{x} \geq \widehat{b}_c - \widehat{b}_\ell\}}\right] \\
&\quad + \sum_{j \neq c} e^{\frac{\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_c\|_{\ell_2}^2}{2}} \mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})}\left[\prod_{\ell \neq c} \mathbb{1}_{\{(\widehat{\boldsymbol{w}}_\ell - \widehat{\boldsymbol{w}}_c)^T \boldsymbol{x} \geq (\widehat{\boldsymbol{w}}_\ell - \widehat{\boldsymbol{w}}_c)^T (\boldsymbol{\mu}_c - \boldsymbol{\mu}_j) + \widehat{b}_c - \widehat{b}_\ell\}}\right]
\end{aligned}
$$

To continue define $\boldsymbol{S}_c \in \mathbb{R}^{(k-1)\times(k-1)}$ to be a symmetric matrix such that its i, j element is $[\boldsymbol{S}_c]_{ij} := \langle \widehat{\boldsymbol{w}}_c - \widehat{\boldsymbol{w}}_j, \widehat{\boldsymbol{w}}_c - \widehat{\boldsymbol{w}}_i \rangle$ and $\boldsymbol{t}_c^{(j)} \in \mathbb{R}^{k-1}$ a vector with entries $[\boldsymbol{t}_c^{(j)}]_i := (\widehat{\boldsymbol{w}}_i - \widehat{\boldsymbol{w}}_c)^T (\boldsymbol{\mu}_c - \boldsymbol{\mu}_j) + \widehat{\boldsymbol{b}}_c - \widehat{\boldsymbol{b}}_i$.

Thus,

$$\mathbb{P}_{e|c} := \mathbb{P}\Big\{\widehat{Y} \neq Y | Y = c\Big\} = \sum_{j=1}^{k} e^{\frac{\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_c\|_{\ell_2}^2}{2}} \mathbb{P}\Big\{\boldsymbol{S}_c^{\frac{1}{2}} \boldsymbol{z} \geq \boldsymbol{t}_c^{(j)}\Big\}$$

Also note that

$$\pi_c = \mathbb{P}\{Y = c\} = \mathbb{E}_{\boldsymbol{x}}\Big[\mathbb{P}\Big\{Y = c | \boldsymbol{x}\Big\}\Big] = \mathbb{E}_{\boldsymbol{x}}\left[\frac{1}{\left(1 + \sum_{j \neq c} e^{(\boldsymbol{\mu}_j - \boldsymbol{\mu}_c)^T \boldsymbol{x}}\right)}\right]$$

C.1.2 A useful expression for the total miss-classification error

We can rewrite (2.8) as follows for $\widetilde{\boldsymbol{g}} \sim \mathcal{N}(\boldsymbol{0}_k, \boldsymbol{I}_k)$ and $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0}_k, \boldsymbol{\Sigma}_{\boldsymbol{\mu},\boldsymbol{\mu}})$

$$\begin{aligned}
\mathbb{P}_e &= \mathbb{P}\left\{\arg\max_{j \in [k]} \ (\boldsymbol{g}_j + \widehat{\boldsymbol{b}}_j) \neq Y(\boldsymbol{h})\right\}, \\
&= \mathbb{P}\left\{\arg\max \ \left(\left(\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}} - \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}} \boldsymbol{\Sigma}_{\boldsymbol{\mu},\boldsymbol{\mu}}^{\dagger} \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}}\right)^{1/2} \widetilde{\boldsymbol{g}} + \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}} \boldsymbol{\Sigma}_{\boldsymbol{\mu},\boldsymbol{\mu}}^{\dagger} \boldsymbol{h} + \widehat{\boldsymbol{b}}\right) \neq Y(\boldsymbol{h})\right\} \\
&= \mathbb{P}\left\{\arg\max \ \left(\left(\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}} - \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}} \boldsymbol{\Sigma}_{\boldsymbol{\mu},\boldsymbol{\mu}}^{\dagger} \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}}\right)^{1/2} \widetilde{\boldsymbol{g}} + \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}} \boldsymbol{\Sigma}_{\boldsymbol{\mu},\boldsymbol{\mu}}^{\dagger} \boldsymbol{h} + \widehat{\boldsymbol{b}}\right) \neq Y(\boldsymbol{h})\right\} \\
&= 1 - \sum_{c \in [k]} \mathbb{E}\left[\frac{e^{\boldsymbol{h}_c}}{\sum_{\ell \in [k]} e^{\boldsymbol{h}_\ell}} \mathbb{P}\left\{\boldsymbol{S}_c^{1/2} \boldsymbol{z} \geq \boldsymbol{t}(\boldsymbol{h})\right\}\right],
\end{aligned} \tag{C.1}$$

where in the last line $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{k-1})$:

$$\begin{aligned}
[\boldsymbol{t}(\boldsymbol{h})]_j &= \widehat{\boldsymbol{b}}_j - \widehat{\boldsymbol{b}}_c + [\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}} \boldsymbol{\Sigma}_{\boldsymbol{\mu},\boldsymbol{\mu}}^{\dagger} \boldsymbol{h}]_j - [\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}} \boldsymbol{\Sigma}_{\boldsymbol{\mu},\boldsymbol{\mu}}^{\dagger} \boldsymbol{h}]_c, \quad j \in [k-1] \\
[\boldsymbol{S}_c]_{i,j} &= (\boldsymbol{e}_c - \boldsymbol{e}_j)^T \left(\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}} - \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}} \boldsymbol{\Sigma}_{\boldsymbol{\mu},\boldsymbol{\mu}}^{\dagger} \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}}\right) (\boldsymbol{e}_c - \boldsymbol{e}_i), \quad i,j \in [k-1].
\end{aligned}$$

C.2 Class-wise and total miss-classification error for GMM

The class-wise miss-classification error is given by

$$\mathbb{P}_{e|c} = \mathbb{P}\left(\exists j \neq c \ : \ \langle \widehat{\boldsymbol{w}}_c - \widehat{\boldsymbol{w}}_j, \boldsymbol{z} \rangle \leq \langle \widehat{\boldsymbol{w}}_j - \widehat{\boldsymbol{w}}_c, \boldsymbol{\mu}_c \rangle + (\widehat{\boldsymbol{b}}_j - \widehat{\boldsymbol{b}}_c)\right) \tag{C.2}$$

$$= 1 - \mathbb{P}\left(\forall j \neq c \ : \ \langle \widehat{\boldsymbol{w}}_c - \widehat{\boldsymbol{w}}_j, \boldsymbol{z} \rangle \geq \langle \widehat{\boldsymbol{w}}_j - \widehat{\boldsymbol{w}}_c, \boldsymbol{\mu}_c \rangle + (\widehat{\boldsymbol{b}}_j - \widehat{\boldsymbol{b}}_c)\right). \tag{C.3}$$

where we used that $\boldsymbol{x} = \boldsymbol{\mu}_c + \boldsymbol{z}$. Let $\boldsymbol{S}_c \in \mathbb{R}^{(k-1)\times(k-1)}$ be a symmetric matrix such that its i,j element is $[\boldsymbol{S}_c]_{ij} := \langle \widehat{\boldsymbol{w}}_c - \widehat{\boldsymbol{w}}_j, \widehat{\boldsymbol{w}}_c - \widehat{\boldsymbol{w}}_i \rangle$ and $\boldsymbol{t}_c \in \mathbb{R}^{k-1}$ a vector with entries $[\boldsymbol{t}_c]_i := \langle \widehat{\boldsymbol{w}}_i - \widehat{\boldsymbol{w}}_c, \boldsymbol{\mu}_c \rangle + (\widehat{\boldsymbol{b}}_i - \widehat{\boldsymbol{b}}_c)$. Then, we can rewrite (C.3) as

$$\mathbb{P}_{e|c} := 1 - \mathbb{P}\left\{\boldsymbol{S}_c^{1/2} \boldsymbol{z} \geq \boldsymbol{t}_c\right\}, \tag{C.4}$$

where the inequality in the rightmost expression applies entry-wise.

Next we focus on calculating the total miss-classification error. To this aim by using the law of total probability we have

$$\mathbb{P}_e = \sum_{c=1}^{k} \pi_c \, \mathbb{P}_{e|c} = 1 - \sum_{c=1}^{k} \pi_c \, \mathbb{P}\left\{\boldsymbol{S}_c^{1/2} \boldsymbol{z} \geq \boldsymbol{t}_c\right\}.$$

C.3 Evaluating and bounding tail probabilities of multivariate Gaussians

In Sections C.1.2 and C.2, we expressed the class-wise probability of missclassification error for both GMM and MLM in the following convenient form

$$1 - \mathbb{P}\{\boldsymbol{A}\boldsymbol{g} \leq \boldsymbol{t}\} = \mathbb{P}\left\{\bigcup_{i \in [k-1]} \{\boldsymbol{a}_i^T \boldsymbol{g} \geq \boldsymbol{t}_i\}\right\}. \tag{C.5}$$

Here, $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{k-1})$ and $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_{k-1} \end{bmatrix}^T \in \mathbb{R}^{(k-1)\times(k-1)}, \boldsymbol{t} \in \mathbb{R}^{k-1}$ are appropriate coefficient matrices; see (C.4) and (C.1).

The formulation above is convenient both in our theoretical analysis, as well as, in simulations. In the rest of this section, we briefly discuss some relevant tools that allow to further simplify or bound expressions in the form of (C.5).

C.3.1 A special case: Rank-one update of Identity

First, we discuss the case where the coefficient matrix $\boldsymbol{A}$ and vector $\boldsymbol{t}$ in (C.5) take the special form $\boldsymbol{A} \propto \boldsymbol{I} + \mathbf{1}\mathbf{1}^T$ and $\boldsymbol{t} \propto \mathbf{1}$. This special case often appears in some of the stylized symmetric problem settings studied in this paper, such as classification problems with orthogonal and equally-balanced means.

Lemma C.1 *Let $\boldsymbol{A} = \boldsymbol{I}_k + \mathbf{1}_k\mathbf{1}_k^T$ and $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{A})$. Then, for any $t \in \mathbb{R}$,*

$$1 - \mathbb{P}\{\boldsymbol{g} \le t\mathbf{1}_k\} = \mathbb{P}\{G_0 + \max_{i\in[k]} G_i \ge t\}, \quad G_0, G_1, \ldots, G_k \overset{iid}{\sim} \mathcal{N}(0,1)\,. \tag{C.6}$$

Proof It is easily checked that we can decompose for each $i \in [n]$:

$$\boldsymbol{g}_i = G_0 + G_i,$$

where $G_0, G_1, \ldots, G_k$ are iid standard normals. Indeed, it can be readily checked from this that $\mathbb{E}[\boldsymbol{g}_i^2] = 2$ and $\mathbb{E}[\boldsymbol{g}_i\boldsymbol{g}_j] = \mathbb{E}[G_0^2] = 1,\ i \neq j$, which is consistent with $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{A})$.

Thus, we can write

$$1 - \mathbb{P}\{\boldsymbol{g} \le t\mathbf{1}_k\} = \mathbb{P}\{\max_{i\in[k]} \boldsymbol{g}_i \ge t\} = \mathbb{P}\{G_0 + \max_{i\in[k]} G_i \ge t\},$$

to complete the proof. ∎

C.3.2 Slepian's bound

When the matrix $\boldsymbol{A}$ does not have the special structure assumed by Lemma C.1, it is not possible in general to provide simple expressions as the one in (C.6). Yet, it might be possible to obtain upper bounds of the same simple form. Such simple bounds can be useful for theoretical interpretations of otherwise complicated formulae, or can provide efficient means for quick (but, non-tight) implementations.

In this section, we discuss Slepian's inequality (see B.1) as a useful tool in this direction. Assume for simplicity that $a = \min_{i\neq j\in[k]} \boldsymbol{A}_{ij} \ge 0$ and note that

$$\boldsymbol{A} \ge (\mathrm{diag}(\boldsymbol{A}) - a\boldsymbol{I}) + a\mathbf{1}\mathbf{1}^T,$$

where the inequality above holds element-wise and equality is true for the diagonal elements. Then, one can apply Slepian's Lemma B.1 to upper bound the conditional probability of error in (C.5) with the following simple bound:

$$\begin{aligned} 1 - \mathbb{P}\{\boldsymbol{A}\boldsymbol{g} \le \boldsymbol{t}\} &\le 1 - \mathbb{P}\left\{\left((\mathrm{diag}(\boldsymbol{A}) - a\boldsymbol{I}) + a\mathbf{1}\mathbf{1}^T\right)^{1/2}\boldsymbol{g} \le \boldsymbol{t}\right\} \\ &\le \mathbb{P}\left\{\bigcup_{j\in[k]} \left\{G_0 + G_j\sqrt{[\boldsymbol{A}]_{jj}/a} \ge [\boldsymbol{t}]_j/a\right\}\right\}, \qquad G_0, G_1, \ldots, G_k \overset{iid}{\sim} \mathcal{N}(0,1)\,. \end{aligned}$$

In the second line above, we used the Gaussian decomposition of Lemma C.1.

C.3.3 Union bound

Of course, it is also possible to apply (a simpler) union bound to upper bound the tail probability in (C.5). Here, we show explicitly the result of applying union bound to the class-wise error probabilities of the GMM model. Specifically, consider (C.2). An application of the union bound leads to the following:

$$\begin{aligned} \mathbb{P}_{e|c} &\le \sum_{j\neq c} \mathbb{P}_{G\sim\mathcal{N}(0,1)}\left\{\|\widehat{\boldsymbol{w}}_j - \widehat{\boldsymbol{w}}_c\|_{\ell_2}\, G \le \langle \widehat{\boldsymbol{w}}_j - \widehat{\boldsymbol{w}}_c, \boldsymbol{\mu}_c\rangle + (\widehat{\boldsymbol{b}}_j - \widehat{\boldsymbol{b}}_c)\right\} \\ &= \sum_{j\neq c} Q\left(\langle \frac{\widehat{\boldsymbol{w}}_c - \widehat{\boldsymbol{w}}_j}{\|\widehat{\boldsymbol{w}}_c - \widehat{\boldsymbol{w}}_j\|_{\ell_2}}, \boldsymbol{\mu}_c\rangle + \frac{\widehat{\boldsymbol{b}}_c - \widehat{\boldsymbol{b}}_j}{\|\widehat{\boldsymbol{w}}_c - \widehat{\boldsymbol{w}}_j\|_{\ell_2}}\right) = \sum_{j\neq c} Q\left(\frac{[\boldsymbol{t}_c]_j}{\sqrt{[\boldsymbol{S}_c]_{j,j}}}\right) \\ &\le (k-1)\cdot Q\left(d_{\min}\right), \end{aligned} \tag{C.7}$$

where in the last line, we denote $d_{\min} := \min_{j\neq c}\left\{[\boldsymbol{t}_c]_j \Big/ \sqrt{[\boldsymbol{S}_c]_{j,j}}\right\}$. It is worth noting that the upper bound given in (C.7) still requires knowledge of the cross-correlations $\langle \widehat{\boldsymbol{w}}_\ell, \widehat{\boldsymbol{w}}_c\rangle,\ \ell \neq c$.

Oracle bound. For completeness, we briefly discuss an oracle lower bound for the class-wise probability of error in GMM. Specifically, assume that the means $\boldsymbol{\mu}_i,\ i \in [n]$ are known. Then the problem of classifying a new sample $\boldsymbol{x}$ is a k-ary hypothesis testing problem with Gaussian conditionals. Denote $\mathbb{P}_{\text{genie,Bayes}}$ the Bayes error of this hypothesis testing problem. Clearly $\mathbb{P}_{\text{genie,Bayes}}$ is a lower bound on the error of any classifier that is trained on data. The paper [69] further lower bounds $\mathbb{P}_{\text{genie,Bayes}}$ in terms of the Bayesian probability of errors between every two classes as follows:

$$\begin{aligned}\mathbb{P}_e \geq \mathbb{P}_{\text{genie,Bayes}} &\geq \frac{2}{k}\sum_{i\neq j}\pi_i\,\mathbb{P}_{\text{genie,ij}} \qquad \text{(C.8)}\\ &= \frac{2}{k}\sum_{i\neq j}\frac{\pi_i}{\pi_i+\pi_j}\left\{\pi_i\cdot Q\left(\frac{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|_{\ell_2}}{2}+\frac{\log(\pi_i/\pi_j)}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|_{\ell_2}}\right)+\pi_j\cdot Q\left(\frac{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|_{\ell_2}}{2}-\frac{\log(\pi_i/\pi_j)}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|_{\ell_2}}\right)\right\}.\end{aligned}$$

where $\mathbb{P}_{\text{genie,ij}}$ is the Bayesian error between classes i and j with priors $\frac{\pi_i}{\pi_i+\pi_j}$ and $\frac{\pi_j}{\pi_i+\pi_j}$. For the last equality we have used the well-known formula for the Bayesian probability of binary Gaussian hypothesis testing. In the case of equal-priors the genie lower bound above simplifies to:

$$\mathbb{P}_e \geq \frac{2}{k}\sum_{i\neq j}\pi_i\cdot Q\left(\frac{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|_{\ell_2}}{2}\right). \qquad \text{(C.9)}$$

Note that, in contrast to (C.9), our analysis allows for precise evaluations of the missclassification error $\mathbb{P}_e$.

D The Class-averaging estimator

D.1 Proofs for GMM

D.1.1 GMM: Proof of Proposition 3.1

The first statement (3.1a) follows directly from the fact that $\frac{1}{n}\mathbf{1}^T\boldsymbol{Y}_i = \frac{n_i}{n} \xrightarrow{P} \pi_i$. For the next two statements note that

$$\widehat{\boldsymbol{w}}_i = \frac{1}{n}\boldsymbol{M}\boldsymbol{Y}\boldsymbol{Y}_i + \frac{1}{n}\boldsymbol{Z}\boldsymbol{Y}_i = \frac{1}{n}\sum_{j=1}^{k}\boldsymbol{\mu}_j(\boldsymbol{Y}_j^T\boldsymbol{Y}_i) + \frac{1}{n}\boldsymbol{Z}\boldsymbol{Y}_i = \frac{\|\boldsymbol{Y}_i\|_{\ell_2}^2}{n}\boldsymbol{\mu}_i + \frac{1}{n}\boldsymbol{Z}\boldsymbol{Y}_i, \qquad \text{(D.1)}$$

where in the last line we used orthogonality of the rows $\boldsymbol{Y}_j$ of the matrix $\boldsymbol{Y}$:

$$\langle \boldsymbol{Y}_i, \boldsymbol{Y}_j\rangle = 0,\ \forall i\neq j\in[k]. \qquad \text{(D.2)}$$

To conclude simply use the facts that for all $i \in [k]$:

(i) $\frac{\|\boldsymbol{Y}_i\|_{\ell_2}^2}{n} = \frac{n_i}{n} \xrightarrow{P} \pi_i$.

(ii) $\boldsymbol{Z}\boldsymbol{Y}_i \sim \sigma\|\boldsymbol{Y}_i\|_{\ell_2}\boldsymbol{g}_i$ with $\boldsymbol{g}_i \overset{iid}{\sim} \mathcal{N}(0, I_d)$ because of (D.2)

(iii) $\frac{\|\boldsymbol{g}_i\|_{\ell_2}}{\sqrt{n}} \xrightarrow{P} \sqrt{\gamma}$ and $\frac{1}{\sqrt{n}}\langle \boldsymbol{g}_i, \boldsymbol{\mu}_j\rangle \xrightarrow{P} 0$.

D.2 Proofs for MLM

D.2.1 Proof of Proposition 4.1

Let us define $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}_r, \boldsymbol{I}_r)$ and random vector $\boldsymbol{v} = [V_1, V_2, \ldots, V_k]^T$ with entries:

$$\boldsymbol{v} = \frac{e^{\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}}{\mathbf{1}_k^T e^{\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}}, \qquad V_i = \frac{e^{\boldsymbol{e}_i^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}}{\sum_{j\in[k]} e^{\boldsymbol{e}_j^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}},\ i\in[k]. \qquad \text{(D.3)}$$

We will prove the following three statements:

$$\widehat{\boldsymbol{b}} \xrightarrow{P} \mathbb{E}\left[\boldsymbol{v}\right] \tag{D.4a}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}} \xrightarrow{P} \mathbb{E}\left[\boldsymbol{v}\boldsymbol{g}^T\right]\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{D.4b}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}} \xrightarrow{P} \gamma \cdot \mathrm{diag}(\mathbb{E}\left[\boldsymbol{v}\right]) + \mathbb{E}\left[\boldsymbol{v}\boldsymbol{g}^T\right] \cdot \mathbb{E}\left[\boldsymbol{g}\boldsymbol{v}^T\right] \tag{D.4c}$$

To see how these lead to (4.2), note that $\mathbb{E}[\boldsymbol{v}] = \boldsymbol{\pi}$ by the definition in (4.1) and an application of Gaussian integration by parts shows that $\mathbb{E}[\boldsymbol{v}\boldsymbol{g}^T] = (\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi})\boldsymbol{V}\boldsymbol{\Sigma}$. Therefore, in what follows, we prove (D.4)

For the intercepts $\widehat{\boldsymbol{b}}_\ell,\ \ell \in [k]$ it holds that

$$\widehat{\boldsymbol{b}}_\ell = \frac{1}{n}\mathbf{1}^T\boldsymbol{Y}_\ell \xrightarrow{P} \mathbb{P}\left(Y = \ell\right) = \mathbb{E}\left[\frac{e^{\boldsymbol{h}_\ell}}{\sum_{j\in[k]} e^{\boldsymbol{h}_j}}\right],$$

where $\boldsymbol{h} \sim \mathcal{N}\left(\mathbf{0}_k, \boldsymbol{\Sigma}_{\boldsymbol{\mu}\boldsymbol{\mu}}\right)$. To deduce the first statement in (D.4a), note that $\boldsymbol{h} \overset{(\mathrm{D})}{=} \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}$.

Continuing with the vectors $\widehat{\boldsymbol{w}}_\ell,\ \ell \in [k]$, recall that $\boldsymbol{w}_\ell = \frac{1}{n}\boldsymbol{X}\boldsymbol{Y}_\ell = \frac{1}{n}\sum_{i_\ell\in[n]} \boldsymbol{x}_{i_\ell}[\boldsymbol{Y}_\ell]_{i_\ell}$. Consider the singular decomposition

$$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = \begin{bmatrix}\boldsymbol{u}_1 & \boldsymbol{u}_2 & \dots & \boldsymbol{u}_r\end{bmatrix}\mathrm{diag}(\sigma_1, \sigma_2, \dots, \sigma_r)\begin{bmatrix}\boldsymbol{v}_1^T \\ \boldsymbol{v}_2^T \\ \dots \\ \boldsymbol{v}_r^T\end{bmatrix},$$

with $\boldsymbol{U} \in \mathbb{R}^{d\times r}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$, and $\boldsymbol{V} \in \mathbb{R}^{k\times r}$ where $r = \mathrm{rank}(\boldsymbol{M}) \le k$. Decompose $\boldsymbol{X} \in \mathbb{R}^{d\times n}$ as $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{X} + \boldsymbol{P}^\perp\boldsymbol{X}$ with $\boldsymbol{P}^\perp = \boldsymbol{I}_d - \boldsymbol{U}\boldsymbol{U}^T$. With this notation we compute

$$\begin{aligned}\langle \boldsymbol{w}_\ell, \boldsymbol{\mu}_c\rangle = \frac{1}{n}\sum_{i\in[n]} \boldsymbol{x}_i^T\boldsymbol{\mu}_c[\boldsymbol{Y}_\ell]_i &= \frac{1}{n}\sum_{j=1}^r\sum_{i\in[n]}(\boldsymbol{x}_i^T\boldsymbol{u}_j)\cdot(\boldsymbol{\mu}_c^T\boldsymbol{u}_j)[\boldsymbol{Y}_\ell]_i + \frac{1}{n}\sum_{j=1}^r\sum_{i\in[n]}\cdot(\boldsymbol{\mu}_c^T\boldsymbol{P}^\perp\boldsymbol{x}_i)[\boldsymbol{Y}_\ell]_i \\ &\xrightarrow{P} \sum_{j=1}^r\left(\boldsymbol{e}_c^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{e}_j\right)\mathbb{E}\left[\boldsymbol{g}_j\frac{\left(e^{\boldsymbol{e}_\ell^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}\right)}{\sum_{\ell'\in[k]} e^{\boldsymbol{e}_{\ell'}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}}\right] \\ &= \sum_{j=1}^r \mathbb{E}[V_\ell\boldsymbol{g}_j]\left(\boldsymbol{e}_j^T\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{e}_c\right).\end{aligned} \tag{D.5}$$

Here, we have recognized that for every $i \in [n] : \boldsymbol{U}^T\boldsymbol{x}_i \sim \boldsymbol{g}$, and also, conditioned on $\boldsymbol{x}_i$: $[\boldsymbol{Y}_\ell]_i \sim \mathrm{Bern}\left(e^{\boldsymbol{\mu}_\ell^T\boldsymbol{x}_i}\big/\sum_{\ell'} e^{\boldsymbol{\mu}_{\ell'}^T\boldsymbol{x}_i}\right)$ and $\boldsymbol{\mu}_\ell^T\boldsymbol{x}_i = \boldsymbol{e}_\ell^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^T\boldsymbol{x}_i \sim \boldsymbol{e}_\ell^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}, \ell \in [k]$. This shows the second statement in (D.4b) when expressed in matrix form.

We proceed similarly with the proof of the last statement in (4.2b) as follows:

$$\begin{aligned}\langle \boldsymbol{w}_\ell, \boldsymbol{w}_c\rangle &= \frac{1}{n^2}\sum_{i_\ell\in[n], i_c\in[n]} \boldsymbol{x}_{i_\ell}^T\boldsymbol{x}_{i_c}[\boldsymbol{Y}_\ell]_{i_\ell}[\boldsymbol{Y}_c]_{i_c} \\ &= \frac{1}{n^2}\sum_{j=1}^r\sum_{i_\ell\in[n], i_c\in[n]}(\boldsymbol{x}_{i_\ell}^T\boldsymbol{u}_j)(\boldsymbol{x}_{i_c}^T\boldsymbol{u}_j)[\boldsymbol{Y}_\ell]_{i_\ell}[\boldsymbol{Y}_c]_{i_c} + \frac{1}{n^2}\sum_{i_\ell\in[n], i_c\in[n]}\left(\boldsymbol{P}^\perp\boldsymbol{x}_{i_\ell}\right)^T\left(\boldsymbol{P}^\perp\boldsymbol{x}_{i_c}\right)[\boldsymbol{Y}_\ell]_{i_\ell}[\boldsymbol{Y}_c]_{i_c}\end{aligned}$$

For $i_\ell = i_c = i \in [n]$ note that

$$\frac{1}{n^2}\sum_{i\in[n]}\sum_{j=1}^r(\boldsymbol{x}_i^T\boldsymbol{u}_j)(\boldsymbol{x}_i^T\boldsymbol{u}_j)[\boldsymbol{Y}_\ell]_i[\boldsymbol{Y}_c]_i \xrightarrow{P} 0,$$

while, for $i_\ell \neq i_c$,

$$\begin{aligned}\frac{1}{n^2}\sum_{i_\ell\neq i_c\in[n]}\sum_{j=1}^r(\boldsymbol{x}_{i_\ell}^T\boldsymbol{u}_j)[\boldsymbol{Y}_\ell]_{i_\ell}(\boldsymbol{x}_{i_c}^T\boldsymbol{u}_j)[\boldsymbol{Y}_c]_{i_c} &\xrightarrow{P} \sum_{j=1}^r\mathbb{E}\left[\boldsymbol{g}_j\cdot\frac{\left(e^{\boldsymbol{e}_\ell^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}\right)}{\sum_{\ell'\in[k]} e^{\boldsymbol{e}_{\ell'}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}}\right]\mathbb{E}\left[\boldsymbol{g}_j\cdot\frac{\left(e^{\boldsymbol{e}_c^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}\right)}{\sum_{\ell'\in[k]} e^{\boldsymbol{e}_{\ell'}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}}\right] \\ &= \sum_{j=1}^r\mathbb{E}[V_\ell\boldsymbol{g}_j]\,\mathbb{E}[\boldsymbol{g}_jV_c] = \boldsymbol{e}_\ell^T\boldsymbol{E}[\boldsymbol{v}\boldsymbol{g}^T]\cdot\boldsymbol{E}[\boldsymbol{g}\boldsymbol{v}^T]\boldsymbol{e}_c^T\end{aligned}$$

Furthermore,

$$\frac{1}{n}\sum_{i\in[n]}\|\boldsymbol{P}^{\perp}\boldsymbol{x}_i\|_{\ell_2}^2\,[\boldsymbol{Y}_\ell]_i[\boldsymbol{Y}_c]_i^2 \xrightarrow{P} \gamma\cdot\mathbb{1}_{\ell,c}\cdot\mathbb{E}\left[\left(\frac{e^{\boldsymbol{e}_\ell^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}}{\sum_{\ell'\in[k]}e^{\boldsymbol{e}_{\ell'}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}}\right)\right] = \gamma\cdot\boldsymbol{e}_\ell^T\mathrm{diag}(\mathbb{E}[\boldsymbol{v}])\boldsymbol{e}_c.$$

Combining the last two displays results in (4.2b), as desired.

D.2.2 Orthogonal means

Here, we specialize the general result of Proposition 4.1 to the special case of orthogonal and means: $\langle\boldsymbol{\mu}_i,\boldsymbol{\mu}_j\rangle = 0, \forall i\neq j$. Recall the notation $\mu_i = \|\boldsymbol{\mu}_i\|_{\ell_2}, i\in[k]$. Then, in this case the parameters in (4.1) are simply given by the following

$$\boldsymbol{\pi}_i := \mathbb{E}\big[\frac{e^{\mu_i G_i}}{\sum_{\ell\in[k]}e^{\mu_\ell G_\ell}}\big], i\in[k] \quad\text{and}\quad \boldsymbol{\Pi}_{ij} := \mathbb{E}\big[\frac{e^{\mu_i G_i}e^{\mu_j G_j}}{\big(\sum_{\ell\in[k]}e^{\mu_\ell G_\ell}\big)^2}\big],\ i,j\in[k]. \tag{D.6}$$

In this case (4.3) can be equivalently expressed as

$$\mathbb{P}_{e,\mathrm{Avg}} \xrightarrow{P} \mathbb{P}\left(\arg\max_{\ell\in[k]}\{\gamma\cdot\mathrm{diag}(\boldsymbol{\pi})\cdot\widetilde{\boldsymbol{g}} + (\mathrm{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi})\cdot\boldsymbol{\Sigma}\boldsymbol{g}\}\neq Y(\boldsymbol{g})\right) \tag{D.7}$$

$$= \mathbb{P}\Big(\bigcup_{j\neq Y}\left\{\gamma\cdot\boldsymbol{\pi}_\ell\cdot\widetilde{\boldsymbol{g}}_\ell \geq \gamma\cdot\boldsymbol{\pi}_Y\cdot\widetilde{\boldsymbol{g}}_Y + (\boldsymbol{e}_Y-\boldsymbol{e}_\ell)^T(\mathrm{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi})\boldsymbol{\Sigma}\boldsymbol{g} + (\boldsymbol{\pi}_Y-\boldsymbol{\pi}_\ell)\right\}\Big), \tag{D.8}$$

where $\boldsymbol{g},\widetilde{\boldsymbol{g}}\overset{iid}{\sim}\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_k)$, $\mathbb{P}(Y(\boldsymbol{g})=c) = \frac{e^{\mu_c\boldsymbol{g}_c}}{\sum_{\ell\in[k]}e^{\mu_\ell\boldsymbol{g}_\ell}}$ and $\boldsymbol{\Sigma} = \mathrm{diag}(\mu_1,\ldots,\mu_k)$.

E On the Bayes risk of GMM: Proof of Proposition 3.5

Without loss of generality in this proof we assume $\sigma = 1$. The general result follows by simply replacing (μ,σ) with $(\frac{\mu}{\sigma},1)$ and using the proof for $\sigma=1$. Recall that the feature vectors $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ of the training data set are given by:

$$\boldsymbol{x}_i = \boldsymbol{M}\boldsymbol{y}_i + \boldsymbol{z}_i,\quad i\in[n],$$

where the matrix of means $\boldsymbol{M}\in\mathbb{R}^{d\times k}$ has iid Gaussian entries with variance μ^2/d, $\boldsymbol{z}_i\overset{iid}{\sim}\mathcal{N}(0,I_d)$ and $\boldsymbol{y}_i\overset{iid}{\sim}\mathrm{Unif}(\mathbf{e}_1,\ldots,\mathbf{e}_k)$ with $\mathbf{e}_j$ denoting the j^{th} canonical vector in $\mathbb{R}^k$. By definition here the Bayes estimator is the maximum-likelihood (ML) estimator. By applying the law of total probability and by successive application of the Bayes rule we have the following chain of reformulations of the ML:

$$\begin{aligned}
\hat{\boldsymbol{y}}_{n+1} &= \arg\max_{\boldsymbol{e}_j,\,j\in[k]} P(\boldsymbol{y}=\boldsymbol{e}_j\mid\boldsymbol{X},\boldsymbol{Y},\boldsymbol{x}_{n+1})\\
&= \arg\max_{\boldsymbol{e}_j,\,j\in[k]}\int P(\boldsymbol{y}=\boldsymbol{e}_j\mid\boldsymbol{M},\boldsymbol{X},\boldsymbol{Y},\boldsymbol{x}_{n+1})\,P(\boldsymbol{M}\mid\boldsymbol{X},\boldsymbol{Y},\boldsymbol{x}_{n+1})\,\mathrm{d}\boldsymbol{M}\\
&= \arg\max_{\boldsymbol{e}_j,\,j\in[k]}\int \frac{P(\boldsymbol{x}_{n+1}\mid\boldsymbol{y}=\boldsymbol{e}_j,\boldsymbol{M},\boldsymbol{X},\boldsymbol{Y})\cdot P(\boldsymbol{y}=\boldsymbol{e}_j\mid\boldsymbol{M},\boldsymbol{X},\boldsymbol{Y})}{P(\boldsymbol{x}_{n+1}|\boldsymbol{M},\boldsymbol{X},\boldsymbol{Y})}P(\boldsymbol{M}\mid\boldsymbol{X},\boldsymbol{Y},\boldsymbol{x}_{n+1})\,\mathrm{d}\boldsymbol{M}\\
&= \arg\max_{\boldsymbol{e}_j,\,j\in[k]}\int P(\boldsymbol{x}_{n+1}\mid\boldsymbol{y}=\boldsymbol{e}_j,\boldsymbol{M})\,\frac{P(\boldsymbol{M}\mid\boldsymbol{X},\boldsymbol{Y},\boldsymbol{x}_{n+1})}{P(\boldsymbol{x}_{n+1}|\boldsymbol{M},\boldsymbol{X},\boldsymbol{Y})}\,\mathrm{d}\boldsymbol{M} \qquad \text{(E.1)}\\
&= \arg\max_{\boldsymbol{e}_j,\,j\in[k]}\int P(\boldsymbol{x}_{n+1}\mid\boldsymbol{y}=\boldsymbol{e}_j,\boldsymbol{M})\,\frac{P(\boldsymbol{M}\mid\boldsymbol{X},\boldsymbol{Y})}{P(\boldsymbol{x}_{n+1}|\boldsymbol{X},\boldsymbol{Y})}\,\mathrm{d}\boldsymbol{M}\\
&= \arg\max_{\boldsymbol{e}_j,\,j\in[k]}\int P(\boldsymbol{x}_{n+1}\mid\boldsymbol{y}=\boldsymbol{e}_j,\boldsymbol{M})\,P(\boldsymbol{M}\mid\boldsymbol{X},\boldsymbol{Y})\,\mathrm{d}\boldsymbol{M} \qquad \text{(E.2)}\\
&= \arg\max_{\boldsymbol{e}_j,\,j\in[k]}\int P(\boldsymbol{x}_{n+1}\mid\boldsymbol{y}=\boldsymbol{e}_j,\boldsymbol{M})\,P(\boldsymbol{X}\mid\boldsymbol{M},\boldsymbol{Y})\,P(\boldsymbol{M})\,\mathrm{d}\boldsymbol{M} \qquad \text{(E.3)}
\end{aligned}$$

To arrive in (E.1) we used that $P(\boldsymbol{y}=\boldsymbol{e}_j\mid\boldsymbol{M},\boldsymbol{X},\boldsymbol{Y}) = \pi,\ \forall j\in[k]$ and $P(\boldsymbol{x}_{n+1}\mid\boldsymbol{y}=\boldsymbol{e}_j,\boldsymbol{M},\boldsymbol{X},\boldsymbol{Y}) = P(\boldsymbol{x}_{n+1}\mid\boldsymbol{y}=\boldsymbol{e}_j,\boldsymbol{M})$. Also, (E.2) follows by recognizing

that $P(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}, \boldsymbol{Y}) > 0$ is independent of the variable of integration $\boldsymbol{M}$ and of the optimization variable j. For the same reasons, in (E.3) we have ignored the normalizing term $P(\boldsymbol{X}|\boldsymbol{Y})$.

Recalling that $\boldsymbol{z}_{n+1} \sim \mathcal{N}(0, I_d)$, we have that $P(\boldsymbol{x}_{n+1} \mid \boldsymbol{y} = \boldsymbol{e}_j, \boldsymbol{M}) \propto \exp\left(-\|\boldsymbol{x}_{n+1} - \boldsymbol{\mu}_j\|_{\ell_2}^2/2\right)$ where $\propto$ hides constant positive terms. Moreover, the posterior probability of the mean matrix given the training data is given by

$$\begin{aligned} P(\boldsymbol{X} \mid \boldsymbol{M}, \boldsymbol{Y}) \cdot P(\boldsymbol{M}) &\propto \exp\left(-\frac{\|\boldsymbol{X} - \boldsymbol{M}\boldsymbol{Y}\|_{\ell_2}^2}{2}\right) \cdot \exp\left(-\frac{\|\boldsymbol{M}\|_{\ell_2}^2}{2(\mu^2/d)}\right) \\ &\propto \prod_{c=1}^{k} \left\{ \exp\left(-\frac{\|\boldsymbol{\mu}_c\|_{\ell_2}^2}{2(\mu^2/d)}\right) \cdot \prod_{i \in \mathcal{C}_c} \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{\mu}_c\|_{\ell_2}^2}{2}\right) \right\}, \end{aligned} \tag{E.4}$$

where we denote by $\mathcal{C}_c$ the collection of training samples that belong to class $c \in [k]$, i.e. $\mathcal{C}_c = \{i \in [n] \mid \boldsymbol{y}_i = \boldsymbol{e}_c\}$.

With these the objective function of the ML rule in (E.3) becomes:

$$\hat{\boldsymbol{y}}_{n+1} = \arg\max_{j \in [k]} \mathcal{I}(\ell, \mathcal{C}_j \cap n+1) \cdot \prod_{\substack{c=1 \\ c \neq j}}^{k} \mathcal{I}(c, \mathcal{C}_c) \tag{E.5}$$

where for $\ell \in [k]$ and a subset $\mathcal{A} \subset [n+1]$ we denote

$$\mathcal{I}(\ell, \mathcal{A}) := \int \mathrm{d}\boldsymbol{\mu}_c \, \exp\left(-\frac{\|\boldsymbol{\mu}_c\|_{\ell_2}^2}{2(\mu^2/d)}\right) \cdot \exp\left(-\sum_{i \in \mathcal{A}} \frac{\|\boldsymbol{x}_i - \boldsymbol{\mu}_c\|_{\ell_2}^2}{2}\right).$$

By completing the squares and invoking a gaussian integral it can be shown that

$$\begin{aligned} \mathcal{I}(\ell, \mathcal{A}) &:= \sqrt{\frac{(d/\mu^2 + |\mathcal{A}|)}{(2\varpi)^d}} \exp\left(-\frac{\left(1 - \frac{1}{d/\mu^2 + |\mathcal{A}|}\right)}{2} \sum_{i \in \mathcal{A}} \|\boldsymbol{x}_i\|_{\ell_2}^2 + \frac{1}{2\left(d/\mu^2 + |\mathcal{A}|\right)} \sum_{i \in \mathcal{A}} \langle \boldsymbol{x}_i, \sum_{\substack{j \in \mathcal{A} \\ j \neq i}} \boldsymbol{x}_j \rangle \right) \\ &:= \sqrt{\frac{(d/\mu^2 + |\mathcal{A}|)}{(2\varpi)^d}} \exp\left(-\frac{1}{2\left(\frac{d}{\mu^2} + |\mathcal{A}|\right)} \left(\left(\frac{d}{\mu^2} + |\mathcal{A}| - 1\right) \sum_{i \in \mathcal{A}} \|\boldsymbol{x}_i\|_{\ell_2}^2 - \sum_{i \in \mathcal{A}} \langle \boldsymbol{x}_i, \sum_{\substack{j \in \mathcal{A} \\ j \neq i}} \boldsymbol{x}_j \rangle \right) \right) \\ &:= \sqrt{\frac{(d/\mu^2 + |\mathcal{A}|)}{(2\varpi)^d}} \exp\left(-\frac{1}{2\left(\frac{d/n}{\mu^2} + |\mathcal{A}|/n\right)} \left(\left(\frac{d/n}{\mu^2} + \frac{|\mathcal{A}|}{n} - \frac{1}{n}\right) \sum_{i \in \mathcal{A}} \|\boldsymbol{x}_i\|_{\ell_2}^2 - \frac{1}{n} \sum_{i \in \mathcal{A}} \langle \boldsymbol{x}_i, \sum_{\substack{j \in \mathcal{A} \\ j \neq i}} \boldsymbol{x}_j \rangle \right) \right). \end{aligned}$$

Using this in (E.5) we have that

$$\hat{\boldsymbol{y}}_{n+1} = \arg\max_{j \in [k]} \mathcal{I}(j) \cdot \exp\left(-\frac{1}{2\left(\frac{d/n}{\mu^2} + \frac{n_j+1}{n}\right)} \left(\left(\frac{d/n}{\mu^2} + \frac{n_j}{n}\right) \|\boldsymbol{x}_{n+1}\|_{\ell_2}^2 - \frac{2}{n} \langle \boldsymbol{x}_{n+1}, \sum_{\ell \in \mathcal{C}_j} \boldsymbol{x}_\ell \rangle \right) \right), \tag{E.6}$$

where $\xi(n_c) := \frac{d/n}{\mu^2} + \frac{n_c}{n}$, $c \in [k]$ and

$$\mathcal{I}(j) := \prod_{\substack{c=1 \\ c \neq j}}^{k} e^{-\frac{1}{2\xi(n_c)} \left((\xi(n_c) - \frac{1}{n}) \sum_{i \in \mathcal{C}} \|\boldsymbol{x}_i\|_{\ell_2}^2 - \frac{1}{n} \sum_{i \in \mathcal{C}} \langle \boldsymbol{x}_i, \sum_{\substack{\ell \in \mathcal{C} \\ \ell \neq i}} \boldsymbol{x}_\ell \rangle \right)} \cdot e^{-\frac{1}{2(\xi(n_j) + \frac{1}{n})} \left(\xi(n_j) \sum_{i \in \mathcal{C}_j} \|\boldsymbol{x}_i\|_{\ell_2}^2 - \frac{1}{n} \sum_{i \in \mathcal{C}_j} \langle \boldsymbol{x}_i, \sum_{\substack{\ell \in \mathcal{C}_j \\ \ell \neq i}} \boldsymbol{x}_\ell \rangle \right)}.$$

Since the term $\mathcal{I}$ is independent of j in the maximization in (E.6), we conclude that

$$\begin{aligned} \hat{\boldsymbol{y}}_{n+1} &= \arg\max_{j \in [k]} \; \log\left(\mathcal{I}(j)\right) - \frac{1}{2\left(\frac{d/n}{\mu^2} + \frac{n_j+1}{n}\right)} \left\{ \left(\frac{d/n}{\mu^2} + \frac{n_j}{n}\right) \|\boldsymbol{x}_{n+1}\|_{\ell_2}^2 - \frac{2}{n} \langle \boldsymbol{x}_{n+1}, \sum_{\ell \in \mathcal{C}_j} \boldsymbol{x}_\ell \rangle \right\} \\ &= \arg\max_{j \in [k]} \; \log\left(\mathcal{I}(j)\right) + \frac{1}{2\left(\frac{d/n}{\mu^2} + \frac{n_j+1}{n}\right)} \left\{ \frac{2}{n} \langle \boldsymbol{x}_{n+1}, \sum_{\ell \in \mathcal{C}_j} \boldsymbol{x}_\ell \rangle \right\} \end{aligned} \tag{E.7}$$

Next, we evaluate the objective in (E.7) in the asymptotic limit $n, d \to \infty, n/d = \gamma$. First, since $n_c/n \xrightarrow{P} \pi$, note that $\mathcal{I}(j) - \mathcal{I}(\ell) \xrightarrow{P} 0$ for all $i, j \in [k]$. Moreover, note that

$$\begin{aligned}\frac{1}{n}\langle \boldsymbol{x}_{n+1}, \sum_{\ell \in \mathcal{C}_j} \boldsymbol{x}_\ell \rangle &= \frac{1}{n}\langle \boldsymbol{M}\boldsymbol{y}_{n+1} + \boldsymbol{z}_{n+1}, n_j\, \boldsymbol{\mu}_j + \sum_{\ell \in \mathcal{C}_j} \boldsymbol{z}_\ell \rangle \\ &= \frac{n_j}{n}\langle \boldsymbol{M}\boldsymbol{y}_{n+1}, \boldsymbol{\mu}_j \rangle + \frac{n_j}{n}\langle \boldsymbol{z}_{n+1}, \boldsymbol{\mu}_j \rangle + \frac{1}{n}\sum_{\ell \in \mathcal{C}_j} \langle \boldsymbol{M}\boldsymbol{y}_{n+1}, \boldsymbol{z}_\ell \rangle + \frac{1}{n}\sum_{\ell \in \mathcal{C}_j} \langle \boldsymbol{z}_{n+1}, \boldsymbol{z}_\ell \rangle \end{aligned} \tag{E.8}$$

For each one of the four terms in (E.8), we have the following by the CLT:

$$\begin{aligned}&\frac{n_j}{n}\langle \boldsymbol{M}\boldsymbol{y}_{n+1}, \boldsymbol{\mu}_\ell \rangle \xrightarrow{P} \pi\mu^2 \langle \boldsymbol{y}_{n+1}, \boldsymbol{e}_j \rangle \\ &\frac{n_j}{n}\langle \boldsymbol{z}_{n+1}, \boldsymbol{\mu}_j \rangle \xrightarrow{(D)} \mathcal{N}(0, \pi^2 r^2) \\ &\frac{1}{n}\sum_{\ell \in \mathcal{C}_j} \langle \boldsymbol{M}\boldsymbol{y}_{n+1}, \boldsymbol{z}_\ell \rangle \xrightarrow{P} 0 \\ &\frac{1}{n}\sum_{\ell \in \mathcal{C}_j} \langle \boldsymbol{z}_{n+1}, \boldsymbol{z}_\ell \rangle \xrightarrow{(D)} \mathcal{N}(0, \pi\gamma),\end{aligned}$$

where in the last line we used the fact that $\frac{1}{\sqrt{n_j}} \sum_{\ell \in \mathcal{C}_j} \frac{\langle \boldsymbol{z}_{n+1}, \boldsymbol{z}_\ell \rangle}{\sqrt{n}} \xrightarrow{(D)} \mathcal{N}(0, \gamma)$.

Therefore, in the asymptotic limit, the Bayes estimator is the solution to:

$$\hat{\boldsymbol{y}}_{n+1} = \arg\max_{\boldsymbol{e}_j, j \in [k]} \pi\mu^2 \langle \boldsymbol{y}_{n+1}, \boldsymbol{e}_j \rangle + \sqrt{\pi\left(\pi\mu^2 + \gamma\right)}\, G_j, \quad G_1, \dots, G_k \overset{iid}{\sim} \mathcal{N}(0,1). \tag{E.9}$$

As such, the probability of error is

$$\begin{aligned}\mathbb{P}_e &= \mathbb{P}\left(\hat{\boldsymbol{y}}_{n+1} \neq \boldsymbol{y}_{n+1}\right) = \mathbb{P}\left(\pi\mu^2 + \sqrt{\pi\left(\pi\mu^2 + \gamma\right)}\, G_0 \leq \max_{\ell \in [k-1]} \sqrt{\pi\left(\pi\mu^2 + \gamma\right)}\, G_\ell\right) \\ &= \mathbb{P}\left(G_0 + \max_{\ell \in [k-1]} G_\ell \geq \mu^2 \sqrt{\frac{\pi}{\pi\mu^2 + \gamma}}\right),\end{aligned} \tag{E.10}$$

F Proof outline for (weighted) least-squares: key ideas and challenges

In this section, we provide a proof sketch for the analysis of the multiclass least-squares (LS) classifier.

Specifically, we discuss our approach towards specifying the high-dimensional limits of the key quantities needed to evaluate the classification error: $\boldsymbol{b}, \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}}$, and $\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}}$. We note that our proofs for the Weighted Least-Squares (WLS) classifiers follow the same general strategy, but in some parts require significantly more involved and intricate analysis and derivations.

Our proofs follow the following general steps. See proof of each theorem for complete details and derivations.

Step I: Decomposing the loss across classes. Recall from Section 2.2 that the multiclass LS classifier produces a linear classifier $\boldsymbol{x} \mapsto \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}$ via a least-squares fit to the training data:

$$(\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{b}}) := \frac{1}{2n} \left\| \boldsymbol{W}\boldsymbol{X} + \boldsymbol{b}\mathbf{1}_n^T - \boldsymbol{Y} \right\|_F^2. \tag{F.1}$$

Notice that the objective function above is separable. That is,

$$\frac{1}{2n} \left\| \boldsymbol{W}\boldsymbol{X} + \boldsymbol{b}\mathbf{1}_n^T - \boldsymbol{Y} \right\|_F^2 = \frac{1}{2n} \sum_{\ell=1}^{k} \left\| \boldsymbol{X}^T\boldsymbol{w}_\ell + b_\ell \mathbf{1}_n - \boldsymbol{Y}_\ell \right\|_{\ell_2}^2.$$

Hence, for each $\ell \in [k]$,

$$(\widehat{\boldsymbol{w}}_\ell, \widehat{\boldsymbol{b}}_\ell) = \arg\min_{\boldsymbol{w}_\ell, \boldsymbol{b}_\ell} \frac{1}{2n} \left\| \boldsymbol{X}^T\boldsymbol{w}_\ell + \boldsymbol{b}_\ell \mathbf{1}_n - \boldsymbol{Y}_\ell \right\|_{\ell_2}^2. \tag{F.2}$$

This decomposition is convenient for analysis as it is easier to compute the statistical properties of the simple single-output LS in (F.2) compared to the multi-output objective in (F.1). Indeed, as we show, this simplification will eventually allow us to compute the high-dimensional behavior of the intercepts $\widehat{\boldsymbol{b}}_\ell,\ \ell \in [k]$, the mean-correlations $\langle \widehat{\boldsymbol{w}}_\ell, \boldsymbol{\mu}_c \rangle,\ \ell, c \in [k]$, the norms $\|\widehat{\boldsymbol{w}}_\ell\|_{\ell_2},\ \ell \in [k]$ and the LS training loss $\left\|\boldsymbol{X}^T \widehat{\boldsymbol{w}}_\ell + \widehat{\boldsymbol{b}}_\ell \mathbf{1}_n - \boldsymbol{Y}_\ell\right\|_{\ell_2}$.

Step II: Reduction to an Auxiliary Optimization (AO) problem via CGMT. To calculate the high-dimensional behavior (F.2) we use the Convex Gaussian min-max Theorem (CGMT) [60, 66] framework. We provide a brief introduction of the CGMT machinery in Section F.1. Roughly stated this framework allows us to replace a *Primary Optimization* (PO) problem of the form (F.2) with an *Auxiliary Optimization* (AO) problem that is simpler to analyze, but is predictive of the behavior of the latter. For instance for the PO in (F.2) in the GMM after some algebraic manipulations the AO problem takes the form

$$\frac{1}{2}\left(\min_{\boldsymbol{w}_\ell, b_\ell} \frac{1}{\sqrt{n}} \left\|\sigma \|\boldsymbol{w}_\ell\|_{\ell_2} \boldsymbol{g} + \boldsymbol{Y}^T \boldsymbol{M}^T \boldsymbol{w}_\ell + b_\ell \mathbf{1}_n - \boldsymbol{Y}_\ell\right\|_{\ell_2} + \frac{1}{\sqrt{n}} \sigma \boldsymbol{h}^T \boldsymbol{w}_\ell\right)_+^2 \tag{F.3}$$

where $(x)_+ := \max(0, x)$ and $\boldsymbol{g} \in \mathbb{R}^n$ and $\boldsymbol{h} \in \mathbb{R}^d$ are two independent Gaussian random vectors distributed as $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$ and $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$.

Step III: Simplification of the AO and computing $\boldsymbol{\Sigma}_{w,\mu}$ and b. In this step we carry out a series of intricate calculations to further simplify (F.3) and characterize its various asymptotic properties. In particular, in this step we compute the high-dimensional behavior of the intercepts $\widehat{\boldsymbol{b}}_\ell,\ \ell \in [k]$, the mean-correlations $\langle \widehat{\boldsymbol{w}}_\ell, \boldsymbol{\mu}_c \rangle,\ \ell, c \in [k]$, the norms $\|\widehat{\boldsymbol{w}}_\ell\|_{\ell_2},\ \ell \in [k]$ and the LS training loss $\left\|\boldsymbol{X}^T \widehat{\boldsymbol{w}}_\ell + \widehat{\boldsymbol{b}}_\ell \mathbf{1}_n - \boldsymbol{Y}_\ell\right\|_{\ell_2}$. For instance, for GMM these calculations allow us to conclude that

$$\widehat{\boldsymbol{b}} \xrightarrow{P} \boldsymbol{\pi} - \boldsymbol{P}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{\pi} \quad \text{and} \quad \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}} \xrightarrow{P} \boldsymbol{P}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T,$$

where $\boldsymbol{P} := \operatorname{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T \succeq \mathbf{0}_{k\times k}$ and $\boldsymbol{\Delta} := \sigma^2 \boldsymbol{I}_r + \boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{P}\boldsymbol{V}\boldsymbol{\Sigma} \succ \mathbf{0}_{r\times r}$.

Step IV: Computing $\boldsymbol{\Sigma}_{w,w}$ and capturing cross-correlations. The final and most involved part of our analysis is characterizing the asymptotic behavior of $\boldsymbol{\Sigma}_{w,w}$. To see why this is particularly challenging note that the reduction from (F.1) to (F.2) "breaks" the dependence of all $\widehat{\boldsymbol{w}}_1, \widehat{\boldsymbol{w}}_2, \ldots, \widehat{\boldsymbol{w}}_k$ on the *same* feature matrix $\boldsymbol{X}$. Capturing this dependence is crucial in determining the "cross-correlations" $\langle \widehat{\boldsymbol{w}}_\ell, \widehat{\boldsymbol{w}}_c \rangle,\ \ell \neq c$. As noted in Section 2.3 the matrix $\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}}$ is needed to calculate the class-wise and total miss-classification errors. Unfortunately, the CGMT is *not* directly applicable to the multi-output LS optimization in (F.1). Our idea to circumvent this challenge builds on the following simple observation: the vector $\widehat{\boldsymbol{w}}_{\ell,c} = \widehat{\boldsymbol{w}}_\ell + \widehat{\boldsymbol{w}}_c$ is itself the solution to another simple single-output LS problem.

Lemma F.1 *For $\ell \neq c \in [k]$ let $\widehat{\boldsymbol{w}}_\ell$, $\widehat{\boldsymbol{w}}_j$ be the ℓ and j-th row of $\widehat{\boldsymbol{W}}$ which is the solution to the multi-output least-squares minimization (F.1). Denote $\widehat{\boldsymbol{w}}_{\ell,c} := \widehat{\boldsymbol{w}}_\ell + \widehat{\boldsymbol{w}}_c$. Then, $\widehat{\boldsymbol{w}}_{\ell,c}$ is a minimizer in the following single-output least-squares problem:*

$$\widehat{\boldsymbol{w}}_{\ell,j} = \arg\min_{\boldsymbol{w},b} \frac{1}{2n} \left\|\boldsymbol{Y}_\ell + \boldsymbol{Y}_j - \boldsymbol{X}^T\boldsymbol{w} - b\mathbf{1}_n\right\|_{\ell_2}^2. \tag{F.4}$$

Thanks to Lemma I.1, we can use the CGMT to characterize the limiting behavior of $\|\widehat{\boldsymbol{w}}_\ell + \widehat{\boldsymbol{w}}_c\|_{\ell_2}$. These calculations are similar to (but, in certain cases, such as for weighted least-squares, more involved than) those in Steps II and III above. Now note that an asymptotic characterization of $\|\widehat{\boldsymbol{w}}_\ell + \widehat{\boldsymbol{w}}_c\|_{\ell_2}$ immediately yields the asymptotic characterization of $\langle \widehat{\boldsymbol{w}}_\ell, \widehat{\boldsymbol{w}}_c \rangle$ as

$$\langle \widehat{\boldsymbol{w}}_\ell, \widehat{\boldsymbol{w}}_c \rangle = \frac{\|\widehat{\boldsymbol{w}}_\ell + \widehat{\boldsymbol{w}}_c\|_{\ell_2}^2 - \|\widehat{\boldsymbol{w}}_\ell\|_{\ell_2}^2 - \|\widehat{\boldsymbol{w}}_c\|_{\ell_2}^2}{2}, \tag{F.5}$$

and $\|\widehat{\boldsymbol{w}}_\ell\|_{\ell_2}, \|\widehat{\boldsymbol{w}}_c\|_{\ell_2}$ are already computed in Step IV. For instance, for GMM the analysis in this step allow us to calculate the asymptotic behavior of $\boldsymbol{\Sigma}_{w,w}$ as follows

$$\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}} \xrightarrow{P} \frac{\gamma}{(1-\gamma)\sigma^2}\boldsymbol{P} + \boldsymbol{P}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\Big(\boldsymbol{\Delta}^{-1} - \frac{\gamma}{(1-\gamma)\sigma^2}\boldsymbol{I}_r\Big)\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{P}.$$

F.1 Background on CGMT

The CGMT is an extension of Gordon's Gaussian min-max inequality (GMT) [22]. In the context of high-dimensional inference problems, Gordon's inequality was first successfully used in the study oh sharp phase-transitions in noiseless Compressed Sensing [59, 9, 2, 59]. More recently, [60] (see also [2, Sec. 10.3]) discovered that Gordon's inequality is essentially tight for certain convex problems. A concrete and general formulation of this idea was given by [66] and was called the CGMT.

In order to summarize the essential ideas, consider the following two Gaussian processes:

$$X_{\boldsymbol{w},\boldsymbol{u}} := \boldsymbol{u}^T \boldsymbol{G}\boldsymbol{w} + \psi(\boldsymbol{w},\boldsymbol{u}), \tag{F.6a}$$

$$Y_{\boldsymbol{w},\boldsymbol{u}} := \|\boldsymbol{w}\|_{\ell_2} \boldsymbol{g}^T \boldsymbol{u} + \|\boldsymbol{u}\|_{\ell_2} \boldsymbol{h}^T \boldsymbol{w} + \psi(\boldsymbol{w},\boldsymbol{u}), \tag{F.6b}$$

where: $\boldsymbol{G} \in \mathbb{R}^{n\times d}$, $\boldsymbol{g} \in \mathbb{R}^n$, $\boldsymbol{h} \in \mathbb{R}^d$, they all have entries iid Gaussian; the sets $\mathcal{S}_{\boldsymbol{w}} \subset \mathbb{R}^d$ and $\mathcal{S}_{\boldsymbol{u}} \subset \mathbb{R}^n$ are compact; and, $\psi : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}$. For these two processes, define the following (random) min-max optimization programs, which are refered to as the *primary optimization* (PO) problem and the *auxiliary optimization* AO:

$$\Phi(\boldsymbol{G}) = \min_{\boldsymbol{w}\in\mathcal{S}_{\boldsymbol{w}}} \max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}} X_{\boldsymbol{w},\boldsymbol{u}}, \tag{F.7a}$$

$$\phi(\boldsymbol{g},\boldsymbol{h}) = \min_{\boldsymbol{w}\in\mathcal{S}_{\boldsymbol{w}}} \max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}} Y_{\boldsymbol{w},\boldsymbol{u}}. \tag{F.7b}$$

if the sets $\mathcal{S}_{\boldsymbol{w}}$ and $\mathcal{S}_{\boldsymbol{u}}$ are convex and *bounded*, and ψ is continuous *convex-concave* on $\mathcal{S}_{\boldsymbol{w}} \times \mathcal{S}_{\boldsymbol{u}}$, then, for any $\nu \in \mathbb{R}$ and $t > 0$, it holds

$$\mathbb{P}\left(|\Phi(\boldsymbol{G}) - \nu| > t\right) \le 2\mathbb{P}\left(|\phi(\boldsymbol{g},\boldsymbol{h}) - \nu| > t\right). \tag{F.8}$$

In words, concentration of the optimal cost of the AO problem around q^* implies concentration of the optimal cost of the corresponding PO problem around the same value q^*. Asymptotically, if we can show that $\phi(\boldsymbol{g},\boldsymbol{h}) \xrightarrow{P} q^*$, then we can conclude that $\Phi(\boldsymbol{G}) \xrightarrow{P} q^*$. Moreover, starting from (F.8) and under appropriate strict convexity conditions, the CGMT shows that concentration of the optimal solution of the AO problem implies concentration of the optimal solution of the PO around the same value. For example, if minimizers of (F.7b) satisfy $\|\boldsymbol{w}_\phi(\boldsymbol{g},\boldsymbol{h})\|_{\ell_2} \xrightarrow{P} \alpha^*$ for some $\alpha^* > 0$, then, the same holds true for the minimizers of (F.7a): $\|\boldsymbol{w}_\Phi(\boldsymbol{G})\|_{\ell_2} \xrightarrow{P} \alpha^*$. Thus, one can analyze the AO to infer corresponding properties of the PO, the premise being of course that the former is simpler to handle than the latter. In [50], the authors introduce a principled machinery that allows to (a) express a quite general family of convex inference optimization problems in the form of the PO and (b) simplify the AO from a (random) optimization over vector variables to an easier optimization over only few scalar variables, termed the "scalarized AO".

G Least-squares for GMM

G.1 Proof of Theorem 3.2

G.1.1 Computing $\Sigma_{w,\mu}$

$$\begin{aligned}\min_{\boldsymbol{W}\in\mathbb{R}^{k\times d},\, \boldsymbol{b}\in\mathbb{R}^k} \quad \frac{1}{2n}\left\|\boldsymbol{W}\boldsymbol{X} + \boldsymbol{b}\boldsymbol{1}_n^T - \boldsymbol{Y}\right\|_F^2 &= \sum_{\ell=1}^k \min_{\boldsymbol{w}_\ell,\boldsymbol{b}_\ell} \frac{1}{2n}\left\|\boldsymbol{X}^T\boldsymbol{w}_\ell + b_\ell\boldsymbol{1}_n - \boldsymbol{Y}_\ell\right\|_{\ell_2}^2 \\ &= \sum_{\ell=1}^k \min_{\boldsymbol{w}_\ell,\boldsymbol{b}_\ell} \frac{1}{2n}\left\|\boldsymbol{Y}^T\boldsymbol{M}^T\boldsymbol{w}_\ell + \boldsymbol{Z}^T\boldsymbol{w}_\ell + b_\ell\boldsymbol{1}_n - \boldsymbol{Y}_\ell\right\|_{\ell_2}^2\end{aligned}$$

Define

$$\mathcal{L}_{PO}\left(\boldsymbol{w}_\ell, b_\ell\right) := \frac{1}{2n}\left\|\boldsymbol{Y}^T\boldsymbol{M}^T\boldsymbol{w}_\ell + \boldsymbol{Z}^T\boldsymbol{w}_\ell + b_\ell\boldsymbol{1}_n - \boldsymbol{Y}_\ell\right\|_{\ell_2}^2 \tag{G.1}$$

Identifying the AO. To continue further note that by duality we have

$$\min_{\boldsymbol{w}_\ell,\boldsymbol{b}_\ell} \mathcal{L}_{PO}\left(\boldsymbol{w}_\ell, b_\ell\right) = \min_{\boldsymbol{w}_\ell,\boldsymbol{b}_\ell} \max_{\boldsymbol{s}} \frac{1}{n}\left(\boldsymbol{s}^T\boldsymbol{Y}^T\boldsymbol{M}^T\boldsymbol{w}_\ell + \boldsymbol{s}^T\boldsymbol{Z}^T\boldsymbol{w}_\ell + b_\ell\boldsymbol{s}^T\boldsymbol{1}_n - \boldsymbol{s}^T\boldsymbol{Y}_\ell - \frac{\|\boldsymbol{s}\|_{\ell_2}^2}{2}\right)$$

Note that the above is jointly convex in $(\boldsymbol{w}_\ell, b_\ell)$ and concave in $\boldsymbol{s}$. We consider the Auxiliary Optimization (AO) problem

$$\min_{\boldsymbol{w}_\ell, \boldsymbol{b}_\ell} \max_{\boldsymbol{s}} \frac{1}{n}\left(\boldsymbol{s}^T\boldsymbol{Y}^T\boldsymbol{M}^T\boldsymbol{w}_\ell + \sigma\left\|\boldsymbol{w}_\ell\right\|_{\ell_2}\boldsymbol{g}^T\boldsymbol{s} + \sigma\left\|\boldsymbol{s}\right\|_{\ell_2}\boldsymbol{h}^T\boldsymbol{w}_\ell + b_\ell\boldsymbol{s}^T\mathbf{1}_n - \boldsymbol{s}^T\boldsymbol{Y}_\ell - \frac{\left\|\boldsymbol{s}\right\|_{\ell_2}^2}{2}\right),$$

where $\boldsymbol{g} \in \mathbb{R}^n$ and $\boldsymbol{h} \in \mathbb{R}^d$ are independent Gaussian random vectors with i.i.d. $\mathcal{N}(0,1)$ entries. Maximizing over the direction of $\boldsymbol{s}$ and its norm $\beta = \left\|\boldsymbol{s}\right\|_{\ell_2}$ we arrive at

$$\begin{aligned}
&\min_{\boldsymbol{w}_\ell, \boldsymbol{b}_\ell} \max_{\beta\geq 0} \frac{1}{n}\left(\beta\left\|\sigma\left\|\boldsymbol{w}_\ell\right\|_{\ell_2}\boldsymbol{g} + \boldsymbol{Y}^T\boldsymbol{M}^T\boldsymbol{w}_\ell + b_\ell\mathbf{1}_n - \boldsymbol{Y}_\ell\right\|_{\ell_2} + \beta\sigma\boldsymbol{h}^T\boldsymbol{w}_\ell - \frac{\beta^2}{2}\right)\\
&\qquad = \min_{\boldsymbol{w}_\ell, \boldsymbol{b}_\ell} \frac{1}{2n}\left(\left\|\sigma\left\|\boldsymbol{w}_\ell\right\|_{\ell_2}\boldsymbol{g} + \boldsymbol{Y}^T\boldsymbol{M}^T\boldsymbol{w}_\ell + b_\ell\mathbf{1}_n - \boldsymbol{Y}_\ell\right\|_{\ell_2} + \sigma\boldsymbol{h}^T\boldsymbol{w}_\ell\right)_+^2\\
&\qquad = \frac{1}{2}\left(\min_{\boldsymbol{w}_\ell, \boldsymbol{b}_\ell} \frac{1}{\sqrt{n}}\left\|\sigma\left\|\boldsymbol{w}_\ell\right\|_{\ell_2}\boldsymbol{g} + \boldsymbol{Y}^T\boldsymbol{M}^T\boldsymbol{w}_\ell + b_\ell\mathbf{1}_n - \boldsymbol{Y}_\ell\right\|_{\ell_2} + \frac{1}{\sqrt{n}}\sigma\boldsymbol{h}^T\boldsymbol{w}_\ell\right)_+^2
\end{aligned}$$

Scalarization of the AO. For convenience, define

$$\bar{\phi}_{AO,\ell} := \min_{\boldsymbol{w}_\ell, \boldsymbol{b}_\ell} \frac{1}{\sqrt{n}}\left\|\sigma\left\|\boldsymbol{w}_\ell\right\|_{\ell_2}\boldsymbol{g} + \boldsymbol{Y}^T\boldsymbol{M}^T\boldsymbol{w}_\ell + b_\ell\mathbf{1}_n - \boldsymbol{Y}_\ell\right\|_{\ell_2} + \frac{\sigma}{\sqrt{n}}\boldsymbol{h}^T\boldsymbol{w}_\ell. \tag{G.2}$$

To continue, consider the singular value decomposition

$$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = \begin{bmatrix}\boldsymbol{u}_1 & \boldsymbol{u}_2 & \dots & \boldsymbol{u}_r\end{bmatrix}\operatorname{diag}(\sigma_1, \sigma_2, \dots, \sigma_r)\begin{bmatrix}\boldsymbol{v}_1^T\\ \boldsymbol{v}_2^T\\ \dots\\ \boldsymbol{v}_r^T\end{bmatrix}, \tag{G.3}$$

with $\boldsymbol{U} \in \mathbb{R}^{d\times r}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$, and $\boldsymbol{V} \in \mathbb{R}^{k\times r}$ where $r = \text{rank}(\boldsymbol{M}) \leq k$. We further decompose $\boldsymbol{w}_\ell$ in its projections on the orthogonal columns $\boldsymbol{u}_1, \dots, \boldsymbol{u}_r$ of $\boldsymbol{U}$:

$$\boldsymbol{w}_\ell = \sum_{i=1}^r \alpha_i\boldsymbol{u}_i + \alpha_0\boldsymbol{w}_\ell^\perp,$$

where $\left\|\boldsymbol{w}_\ell^\perp\right\|_{\ell_2} = 1$ and $\boldsymbol{U}^T\boldsymbol{w}_\ell^\perp = \mathbf{0}$, $\alpha_0 \geq 0$ and we denote

$$\alpha_i := \boldsymbol{u}_i^T\boldsymbol{w}_\ell, i \in [r]. \tag{G.4}$$

We also define $\boldsymbol{\alpha} = \begin{bmatrix}\alpha_1 & \alpha_2 & \dots & \alpha_k\end{bmatrix}^T$. In this notation, we have

$$\begin{aligned}
\bar{\phi}_{AO,\ell}(\boldsymbol{g},\boldsymbol{h}) := &\min_{\alpha_0\geq 0,\, \boldsymbol{\alpha}\in\mathbb{R}^r,\, \boldsymbol{b}_\ell} \frac{1}{\sqrt{n}}\left\|\sigma\sqrt{\alpha_0^2 + \left\|\boldsymbol{\alpha}\right\|_{\ell_2}^2}\boldsymbol{g} + \boldsymbol{Y}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha} + b_\ell\mathbf{1}_n - \boldsymbol{Y}_\ell\right\|_{\ell_2}\\
&\qquad + \sum_{i=1}^r \alpha_i\sigma\frac{\boldsymbol{h}^T\boldsymbol{u}_i}{\sqrt{n}} + \frac{\alpha_0\sigma}{\sqrt{n}}\min_{\boldsymbol{w}_\ell^\perp}\left(\boldsymbol{h}^T\boldsymbol{w}_\ell^\perp\right)\\
= &\min_{\alpha_0\geq 0,\, \boldsymbol{\alpha}\in\mathbb{R}^r,\, \boldsymbol{b}_\ell} \frac{1}{\sqrt{n}}\left\|\sigma\sqrt{\alpha_0^2 + \left\|\boldsymbol{\alpha}\right\|_{\ell_2}^2}\boldsymbol{g} + \boldsymbol{Y}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha} + b_\ell\mathbf{1}_n - \boldsymbol{Y}_\ell\right\|_{\ell_2}\\
&\qquad + \sum_{i=1}^r \alpha_i\sigma\frac{\boldsymbol{h}^T\boldsymbol{u}_i}{\sqrt{n}} - \alpha_0\sigma\frac{\left\|\boldsymbol{h}^\perp\right\|_{\ell_2}}{\sqrt{n}},
\end{aligned} \tag{G.5}$$

where in the second line we denote $\boldsymbol{h}^\perp$ the projection of $\boldsymbol{h}$ onto the complement subspace of the span of $\boldsymbol{u}_1, \dots, \boldsymbol{u}_r$ and we recalled that $\left\|\boldsymbol{w}_\ell^\perp\right\|_{\ell_2} = 1$ and $\langle\boldsymbol{w}_\ell^\perp, \boldsymbol{u}_i\rangle = 0, i \in [r]$.

Convergence of the AO. First, note that

$$\begin{aligned}
\frac{1}{n}\left\|\boldsymbol{Y}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha} + b_\ell\mathbf{1}_n - \boldsymbol{Y}_\ell\right\|_{\ell_2}^2 =& \frac{1}{n}\left\|\boldsymbol{Y}^T\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha} - \boldsymbol{e}_\ell\right) + b_\ell\mathbf{1}_n\right\|_{\ell_2}^2\\
=& \frac{1}{n}\left\|\boldsymbol{Y}^T\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha} - \boldsymbol{e}_\ell\right)\right\|_{\ell_2}^2 + b_\ell^2 + \frac{2}{n}b_\ell\mathbf{1}_n^T\boldsymbol{Y}^T\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha} - \boldsymbol{e}_\ell\right)\\
=& \text{trace}\left(\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha} - \boldsymbol{e}_\ell\right)^T\operatorname{diag}\left(\frac{n_1}{n}, \frac{n_2}{n}, \dots, \frac{n_k}{n}\right)\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha} - \boldsymbol{e}_\ell\right)\right)\\
&+ b_\ell^2 + 2b_\ell\begin{bmatrix}\frac{n_1}{n} & \frac{n_2}{n} & \dots & \frac{n_k}{n}\end{bmatrix}\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha} - \boldsymbol{e}_\ell\right)
\end{aligned}$$

Thus

$$\begin{aligned}\frac{1}{n}\left\|\boldsymbol{Y}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}+b_\ell\mathbf{1}_n-\boldsymbol{Y}_\ell\right\|_{\ell_2}^2 \xrightarrow{P} &\operatorname{trace}\left(\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}-\boldsymbol{e}_\ell\right)^T\operatorname{diag}\left(\boldsymbol{\pi}\right)\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}-\boldsymbol{e}_\ell\right)\right)\\ &+b_\ell^2+2b_\ell\boldsymbol{\pi}^T\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}-\boldsymbol{e}_\ell\right)\\ =&\boldsymbol{\alpha}^T\left(\boldsymbol{\Sigma}\boldsymbol{V}^T\operatorname{diag}(\boldsymbol{\pi})\boldsymbol{V}\boldsymbol{\Sigma}\right)\boldsymbol{\alpha}-2\pi_\ell\boldsymbol{\alpha}^T\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{e}_\ell+2b_\ell\boldsymbol{\alpha}^T\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{\pi}\\ &+b_\ell^2-2b_\ell\pi_\ell+\pi_\ell\end{aligned}$$

At this point, observe that we have reduced the AO to an optimization problem over only $r+2$ scalar variables. Using the law of large numbers, the fact that $\|\boldsymbol{h}^\perp\|_{\ell_2}$ concentrates around $\sqrt{d-r}$ and $(d-r)/n \xrightarrow{P} \gamma$, as well as the limit calculation above it is not hard to see that for fixed α_0, b_ℓ and $\boldsymbol{\alpha}=[\alpha_1,\ldots,\alpha_r]^T\in\mathbb{R}^r$, the objective function in (G.5) converges to the following:

$$\begin{aligned}&\mathcal{D}_\ell(\alpha_0,\boldsymbol{\alpha},b_\ell)\\ &:=\sqrt{\alpha_0^2\sigma^2+\boldsymbol{\alpha}^T\left(\sigma^2\boldsymbol{I}_r+\boldsymbol{\Sigma}\boldsymbol{V}^T\operatorname{diag}(\boldsymbol{\pi})\boldsymbol{V}\boldsymbol{\Sigma}\right)\boldsymbol{\alpha}-2\boldsymbol{\alpha}^T\left(\pi_\ell\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{e}_\ell-b_\ell\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{\pi}\right)+b_\ell^2-2b_\ell\pi_\ell+\pi_\ell}\\ &\quad-\alpha_0\sigma\sqrt{\gamma},\end{aligned}\tag{G.6}$$

We will show in the next paragraph that the argument inside the square-root in (G.6) is a convex quadratic over $(\alpha_0,\boldsymbol{\alpha},b_\ell)$ (see (G.19). Thus, the function $\mathcal{D}_\ell(\alpha_0,\boldsymbol{\alpha},b_\ell)$ is jointly convex. Using uniform convergence of convex functions over compact sets, we arrive at

$$\bar{\phi}_{AO,\ell}(\boldsymbol{g},\boldsymbol{h})\xrightarrow{P}\min_{\alpha_0\geq 0,\boldsymbol{\alpha},b_\ell}\mathcal{D}_\ell(\alpha_0,\boldsymbol{\alpha},b_\ell).\tag{G.7}$$

Deterministic Analysis. Here, we analyze the deterministic scalar minimization on the RHS of (G.7). Define

$$\boldsymbol{A}:=\begin{bmatrix}\sigma^2\boldsymbol{I}_r+\boldsymbol{\Sigma}\boldsymbol{V}^T\operatorname{diag}(\boldsymbol{\pi})\boldsymbol{V}\boldsymbol{\Sigma} & \boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{\pi}\\ \boldsymbol{\pi}^T\boldsymbol{V}\boldsymbol{\Sigma} & 1\end{bmatrix}\quad\text{and}\quad\boldsymbol{c}_\ell=\begin{bmatrix}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{e}_\ell\\ 1\end{bmatrix}\tag{G.8}$$

and observe that we can write

$$\mathcal{D}_\ell(\alpha_0,\boldsymbol{\alpha},b_\ell)=\sqrt{\alpha_0^2\sigma^2+\pi_\ell+\begin{bmatrix}\boldsymbol{\alpha}^T & b_\ell\end{bmatrix}\boldsymbol{A}\begin{bmatrix}\boldsymbol{\alpha}\\ b_\ell\end{bmatrix}-2\pi_\ell\boldsymbol{c}_\ell^T\begin{bmatrix}\boldsymbol{\alpha}\\ b_\ell\end{bmatrix}}-\alpha_0\sigma\sqrt{\gamma}\tag{G.9}$$

First, note that the matrix $\boldsymbol{A}$ is positive definite. This can be checked by computing the Schur complement of $\boldsymbol{A}$:

$$\boldsymbol{\Delta}:=\sigma^2\boldsymbol{I}_r+\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{P}\boldsymbol{V}\boldsymbol{\Sigma}:=\sigma^2\boldsymbol{I}_r+\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\pi}\boldsymbol{\pi}^T\right)\boldsymbol{V}\boldsymbol{\Sigma}\succ\mathbf{0}_{r\times r}.\tag{G.10}$$

Positive definiteness above holds because $\boldsymbol{P}:=\left(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\pi}\boldsymbol{\pi}^T\right)\succeq\mathbf{0}_{k\times k}$. Thus the term under the square-root in (G.9) is a strictly convex quadratic. Thus, $\mathcal{D}_\ell$ is jointly convex in its arguments.

To simplify the RHS of (G.7) we proceed by minimizing $\mathcal{D}_\ell(\alpha_0,\boldsymbol{\alpha},b_\ell)$ over $(\boldsymbol{\alpha},b_\ell)$ which is equal to

$$\begin{aligned}\begin{bmatrix}\widehat{\boldsymbol{\alpha}}\\ \widehat{b}_\ell\end{bmatrix}=\pi_\ell\boldsymbol{A}^{-1}\boldsymbol{c}_\ell&=\pi_\ell\begin{bmatrix}\boldsymbol{I} & \mathbf{0}\\ -\boldsymbol{\pi}^T\boldsymbol{V}\boldsymbol{\Sigma} & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{\Delta}^{-1} & \mathbf{0}\\ \mathbf{0}^T & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{I} & -\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{\pi}\\ \mathbf{0}^T & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{e}_\ell\\ 1\end{bmatrix}\\ &=\pi_\ell\begin{bmatrix}\boldsymbol{I} & \mathbf{0}\\ -\boldsymbol{\pi}^T\boldsymbol{V}\boldsymbol{\Sigma} & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{\Delta}^{-1} & \mathbf{0}\\ \mathbf{0}^T & 1\end{bmatrix}\begin{bmatrix}-\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\boldsymbol{\pi}-\boldsymbol{e}_\ell\right)\\ 1\end{bmatrix}\\ &=\pi_\ell\begin{bmatrix}-\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{\pi}-\boldsymbol{e}_\ell)\\ 1+\boldsymbol{\pi}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{\pi}-\boldsymbol{e}_\ell)\end{bmatrix}\end{aligned}\tag{G.11}$$

Thus, the minimum value attained is

$$-\pi_\ell^2\begin{bmatrix}-\left(\boldsymbol{\pi}-\boldsymbol{e}_\ell\right)^T\boldsymbol{V}\boldsymbol{\Sigma} & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{\Delta}^{-1} & \mathbf{0}\\ \mathbf{0}^T & 1\end{bmatrix}\begin{bmatrix}-\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\boldsymbol{\pi}-\boldsymbol{e}_\ell\right)\\ 1\end{bmatrix}=-\pi_\ell^2\left(1+\left(\boldsymbol{\pi}-\boldsymbol{e}_\ell\right)^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\boldsymbol{\pi}-\boldsymbol{e}_\ell\right)\right).$$

Using the above (G.7) reduces to

$$\bar{\phi}_{AO,\ell}(\boldsymbol{g},\boldsymbol{h})\xrightarrow{P}\min_{\alpha_0\geq 0}\sqrt{\alpha_0^2\sigma^2+\pi_\ell-\pi_\ell^2\left(1+\left(\boldsymbol{\pi}-\boldsymbol{e}_\ell\right)^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\boldsymbol{\pi}-\boldsymbol{e}_\ell\right)\right)}-\alpha_0\sigma\sqrt{\gamma}.\tag{G.12}$$

Setting the derivative with respect to α_0 to zero we arrive at

$$\frac{\alpha_0\sigma^2}{\sqrt{\alpha_0^2\sigma^2+\pi_\ell-\pi_\ell^2\left(1+(\boldsymbol{\pi}-\boldsymbol{e}_\ell)^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{\pi}-\boldsymbol{e}_\ell)\right)}}=\sigma\sqrt{\gamma}.$$

Thus,

$$\widehat{\alpha}_0=\frac{1}{\sigma}\sqrt{\frac{\gamma}{1-\gamma}}\sqrt{\pi_\ell(1-\pi_\ell)-\pi_\ell^2(\boldsymbol{\pi}-\boldsymbol{e}_\ell)^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{\pi}-\boldsymbol{e}_\ell)}. \tag{G.13}$$

Plugging the latter into (G.12) we arrive at

$$\bar{\phi}_{AO,\ell}(\boldsymbol{g},\boldsymbol{h})\xrightarrow{P}\sqrt{1-\gamma}\sqrt{\pi_\ell(1-\pi_\ell)-\pi_\ell^2(\boldsymbol{\pi}-\boldsymbol{e}_\ell)^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{\pi}-\boldsymbol{e}_\ell)}$$

Asymptotic predictions. First, from (G.11) the bias term converges as follows:

$$\widehat{b}_\ell\xrightarrow{P}\pi_\ell\left(1+\boldsymbol{\pi}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{\pi}-\boldsymbol{e}_\ell)\right).$$

Thus,

$$\widehat{\boldsymbol{b}}\xrightarrow{P}\mathrm{diag}(\boldsymbol{\pi})\left(\mathbf{1}_k+\left(\boldsymbol{\pi}\mathbf{1}_k^T-\boldsymbol{I}_k\right)\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{\pi}\right)$$

Recall from (G.4) that $\boldsymbol{\alpha}=\boldsymbol{U}^T\boldsymbol{w}_\ell$. Thus, the correlations $\langle\boldsymbol{\mu}_i,\boldsymbol{w}_\ell\rangle,\ i\in[k]$ converge as follows:

$$\boldsymbol{M}^T\boldsymbol{w}_\ell=\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^T\boldsymbol{w}_\ell\xrightarrow{P}\boldsymbol{V}\boldsymbol{\Sigma}\widehat{\boldsymbol{\alpha}}=-\pi_\ell\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{\pi}-\boldsymbol{e}_\ell). \tag{G.14}$$

Here, convergence applies element-wise to the entries of the involved random vectors. Moreover, from the analysis above we can predict the limit of the norm $\|\boldsymbol{w}_\ell\|_{\ell_2}$. For this, note that $\|\boldsymbol{w}_\ell\|_{\ell_2}^2=\widehat{\alpha}_0^2+\widehat{\boldsymbol{\alpha}}^T\widehat{\boldsymbol{\alpha}}$. Thus,

$$\|\boldsymbol{w}_\ell\|_{\ell_2}^2\xrightarrow{P}\frac{\gamma}{(1-\gamma)\sigma^2}\pi_\ell(1-\pi_\ell)+\pi_\ell^2(\boldsymbol{\pi}-\boldsymbol{e}_\ell)^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\left(\boldsymbol{\Delta}^{-1}-\frac{\gamma}{(1-\gamma)\sigma^2}\boldsymbol{I}_r\right)\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{\pi}-\boldsymbol{e}_\ell) \tag{G.15}$$

G.1.2 Computing $\Sigma_{w,w}$

In the previous section we used the CGMT to predict the bias $\widehat{b}_\ell$, the correlations $\langle\boldsymbol{\mu}_i,\widehat{\boldsymbol{w}}_\ell\rangle,\ i[k]$ and the norm $\|\widehat{\boldsymbol{w}}_\ell\|_{\ell_2}$ for all $\ell\in[k]$ members of the multi-output classifier. Here, we show how to compute the limits of the cross-correlations $\langle\widehat{\boldsymbol{w}}_\ell,\widehat{\boldsymbol{w}}_j\rangle,\ell\neq j\in[k]$.

Lemma G.1 *For $\ell\neq j\in[k]$ let $\widehat{\boldsymbol{w}}_\ell$ $\widehat{\boldsymbol{w}}_j$ be solutions to the least-squares minimization*

$$(\widehat{\boldsymbol{w}}_\ell,\widehat{\boldsymbol{w}}_j,\widehat{b}_\ell,\widehat{b}_j)=\arg\min_{\boldsymbol{w}_\ell,\boldsymbol{w}_j,b_\ell,b_j}\left\{\frac{1}{2n}\left\|\boldsymbol{Y}_\ell-\boldsymbol{X}^T\boldsymbol{w}_\ell-b_\ell\mathbf{1}_n\right\|_{\ell_2}^2+\frac{1}{2n}\left\|\boldsymbol{Y}_j-\boldsymbol{X}^T\boldsymbol{w}_j-b_j\mathbf{1}_n\right\|_{\ell_2}^2\right\}.$$

Denote $\widehat{\boldsymbol{w}}_{\ell,j}:=\widehat{\boldsymbol{w}}_\ell+\widehat{\boldsymbol{w}}_j$ and $\widehat{b}_{\ell,j}:=\widehat{b}_\ell+\widehat{b}_j$. Then, $(\widehat{\boldsymbol{w}}_{\ell,j},\widehat{b}_{\ell,j})$ is a minimizer in the following least-squares problem:

$$(\widehat{\boldsymbol{w}}_{\ell,j},\widehat{b}_{\ell,j})=\arg\min_{\boldsymbol{w},b}\frac{1}{2n}\left\|\boldsymbol{Y}_\ell+\boldsymbol{Y}_j-\boldsymbol{X}^T\boldsymbol{w}-b\mathbf{1}_n\right\|_{\ell_2}^2 \tag{G.16}$$

Proof Clearly the minimization in (G.16) is convex. Thus, it suffices to prove that $\widehat{\boldsymbol{w}}_\ell+\widehat{\boldsymbol{w}}_j$ satisfies the KKT conditions. First, by optimality of $\widehat{\boldsymbol{w}}_\ell$, we have that

$$\boldsymbol{X}\left(\boldsymbol{Y}_\ell-\boldsymbol{X}^T\widehat{\boldsymbol{w}}_\ell-\widehat{b}_\ell\mathbf{1}_n\right)=0$$

Similarly, for $\widehat{\boldsymbol{w}}_j$:

$$\boldsymbol{X}\left(\boldsymbol{Y}_j-\boldsymbol{X}^T\widehat{\boldsymbol{w}}_j-\widehat{b}_j\mathbf{1}_n\right)=0.$$

Adding the equations on the above displays we find that

$$\boldsymbol{X}\left(\boldsymbol{Y}_\ell+\boldsymbol{Y}_j-\boldsymbol{X}^T(\widehat{\boldsymbol{w}}_j+\widehat{\boldsymbol{w}}_\ell)-(\widehat{b}_j+\widehat{b}_\ell)\mathbf{1}_n\right)=0.$$

Recognize that this coincides with the optimality condition for (G.16). Thus, the proof is complete. ■

Thanks to Lemma I.1, we can use the CGMT to characterize the limiting behavior of $\|\widehat{\boldsymbol{w}}_\ell + \widehat{\boldsymbol{w}}_j\|_{\ell_2}$. Observe that this immediately gives the limit of $\langle \widehat{\boldsymbol{w}}_\ell, \widehat{\boldsymbol{w}}_j \rangle$ since

$$\langle \widehat{\boldsymbol{w}}_\ell, \widehat{\boldsymbol{w}}_j \rangle = \frac{\|\widehat{\boldsymbol{w}}_\ell + \widehat{\boldsymbol{w}}_j\|_{\ell_2}^2 - \|\widehat{\boldsymbol{w}}_\ell\|_{\ell_2}^2 - \|\widehat{\boldsymbol{w}}_j\|_{\ell_2}^2}{2}. \tag{G.17}$$

The analysis of (G.16) is very similar to that of (I.1); thus, most details are omitted. Similar to (G.5) we can relate (G.16) with the following AO problem:

$$\begin{aligned}\bar{\phi}_{AO,\ell,j}(\boldsymbol{g},\boldsymbol{h}) \coloneqq \min_{\beta_0\geq 0,\boldsymbol{\beta}\in\mathbb{R}^r,b_{\ell,j}} \frac{1}{\sqrt{n}} \left\| \sigma\sqrt{\beta_0^2+\|\boldsymbol{\beta}\|_{\ell_2}^2}\,\boldsymbol{g} + \boldsymbol{Y}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\beta} + b_{\ell,j}\mathbf{1}_n - \boldsymbol{Y}_\ell - \boldsymbol{Y}_j \right\|_{\ell_2} \\ + \sigma\sum_{i=1}^r \beta_i \frac{\boldsymbol{h}^T\boldsymbol{u}_i}{\sqrt{n}} - \sigma\beta_0 \frac{\|\boldsymbol{h}^\perp\|_{\ell_2}}{\sqrt{n}},\end{aligned} \tag{G.18}$$

where we have decomposed

$$\boldsymbol{w}_{\ell,j} = \sum_{i=1}^r \beta_i \boldsymbol{u}_i + \beta_0 \boldsymbol{w}_{\ell,j}^\perp,$$

with $\left\|\boldsymbol{w}_{\ell,j}^\perp\right\|_{\ell_2} = 1$ and $\boldsymbol{U}^T\boldsymbol{w}_{\ell,j}^\perp = \mathbf{0}_r$.

Using a calculation similar to the one leading to (G.9) we can show that (G.18) converges point-wise in $\beta_0, \boldsymbol{\beta} = [\beta_1, \ldots, \beta_r], b$ to the following:

$$\mathcal{D}_\ell(\beta_0,\boldsymbol{\beta},b_{\ell,j}) = \sqrt{\beta_0^2\sigma^2 + \pi_\ell + \pi_j + \begin{bmatrix}\boldsymbol{\beta}^T & b_{\ell,j}\end{bmatrix}\boldsymbol{A}\begin{bmatrix}\boldsymbol{\beta}\\ b_{\ell,j}\end{bmatrix} - 2\boldsymbol{d}_{\ell,j}^T\begin{bmatrix}\boldsymbol{\beta}\\ b_{\ell,j}\end{bmatrix}} - \beta_0\sigma\sqrt{\gamma}, \tag{G.19}$$

where $\boldsymbol{A}$ is as in (G.8) and we have further defined

$$\boldsymbol{d}_{\ell,j} \coloneqq \begin{bmatrix}\pi_\ell\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{e}_\ell + \pi_j\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{e}_j \\ \pi_\ell + \pi_j\end{bmatrix}.$$

Thus, similar to (G.11) we can compute the minimizer of the deterministic

$$\begin{bmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{b}_{\ell,j}\end{bmatrix} = \begin{bmatrix}-\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\pi_\ell(\boldsymbol{\pi}-\boldsymbol{e}_\ell)+\pi_j(\boldsymbol{\pi}-\boldsymbol{e}_j)\right)\\ \pi_\ell+\pi_j+\boldsymbol{\pi}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\pi_\ell(\boldsymbol{\pi}-\boldsymbol{e}_\ell)+\pi_j(\boldsymbol{\pi}-\boldsymbol{e}_j)\right)\end{bmatrix} \tag{G.20}$$

and

$$\widehat{\beta}_0 = \frac{1}{\sigma}\sqrt{\frac{\gamma}{1-\gamma}}\sqrt{\pi_\ell+\pi_j-(\pi_\ell+\pi_j)^2-\left(\pi_\ell\left(\boldsymbol{\pi}-\boldsymbol{e}_\ell\right)+\pi_j\left(\boldsymbol{\pi}-\boldsymbol{e}_j\right)\right)^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\pi_\ell\left(\boldsymbol{\pi}-\boldsymbol{e}_\ell\right)+\pi_j\left(\boldsymbol{\pi}-\boldsymbol{e}_j\right)\right)}, \tag{G.21}$$

where recall that $\boldsymbol{\Delta}$ is as in (G.10).

From the CGMT, we have that $\|\widehat{\boldsymbol{w}}_\ell + \widehat{\boldsymbol{w}}_j\|_{\ell_2}^2 \xrightarrow{P} \widehat{\boldsymbol{\beta}}_0^2 + \|\boldsymbol{\beta}\|_{\ell_2}^2$. Combining this with the calculations above, we conclude that

$$\begin{aligned}&\|\widehat{\boldsymbol{w}}_\ell + \widehat{\boldsymbol{w}}_j\|_{\ell_2}^2 \xrightarrow{P} \\ &\frac{\gamma}{(1-\gamma)\sigma^2}\left(\pi_\ell+\pi_j\right)\left(1-\pi_\ell-\pi_j\right) \\ &+\left(\pi_\ell\left(\boldsymbol{\pi}-\boldsymbol{e}_\ell\right)+\pi_j\left(\boldsymbol{\pi}-\boldsymbol{e}_j\right)\right)^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\left(\boldsymbol{\Delta}^{-1}-\frac{\gamma}{(1-\gamma)\sigma^2}\boldsymbol{I}_r\right)\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\pi_\ell\left(\boldsymbol{\pi}-\boldsymbol{e}_\ell\right)+\pi_j\left(\boldsymbol{\pi}-\boldsymbol{e}_j\right)\right)\end{aligned} \tag{G.22}$$

Finally, using (G.22) and (G.15) in (G.17) it follows that

$$\langle \boldsymbol{w}_\ell, \boldsymbol{w}_j \rangle \xrightarrow{P} \pi_\ell\pi_j\left(-\frac{\gamma}{(1-\gamma)\sigma^2}+\left(\boldsymbol{\pi}-\boldsymbol{e}_\ell\right)^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\left(\boldsymbol{\Delta}^{-1}-\frac{\gamma}{(1-\gamma)\sigma^2}\boldsymbol{I}_r\right)\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\boldsymbol{\pi}-\boldsymbol{e}_j\right)\right). \tag{G.23}$$

G.2 Orthogonal means

Here, we specialize the asymptotic predictions of Theorem 3.2 to the case of orthogonal means $\langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_j \rangle = 0,\ i \neq j$.

Corollary G.2 (Orthogonal means) *Consider the case of orthogonal means, i.e. $\langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_j \rangle = 0, \forall i \neq j$ and $\gamma < 1$ with Euclidean norms given by $\mu_i = \|\boldsymbol{\mu}_i\|_{\ell_2}$. Define the following parameters for $i \in [k]$:*

$$\rho_i := \pi_i \sigma^2 \big/ \left(\sigma^2 + \pi_i \mu_i^2\right) \quad \text{and} \quad \beta_i = \rho_i \sigma^2 \big/ \Big(\sigma^2 - \sum_{i=1}^{k} \pi_i \rho_i \mu_i^2\Big).$$

Then, the following asymptotic limits hold for the least-squares classifier, for all $i, j \in [k]$:

$$\widehat{\boldsymbol{b}}_i \xrightarrow{P} \beta_i\,, \qquad \langle \widehat{\boldsymbol{w}}_i, \boldsymbol{\mu}_j \rangle \xrightarrow{P} \frac{1}{\sigma}(\mathbb{1}_{ij} - \beta_i)\rho_j \mu_j\,, \tag{G.24a}$$

$$\langle \widehat{\boldsymbol{w}}_i, \widehat{\boldsymbol{w}}_j \rangle \xrightarrow{P} \frac{1}{\sigma^4}\beta_i\beta_j \sum_{\ell=1}^{k} \rho_\ell^2 \mu_\ell^2 - \frac{1}{\sigma^4}\beta_i \rho_j^2 \mu_j^2 - \frac{1}{\sigma^4}\beta_j \rho_i^2 \mu_i^2 - \frac{\gamma \beta_i \rho_j}{(1-\gamma)\sigma^2} + \frac{\mathbb{1}_{ij}}{\sigma^2}\Big(\frac{\gamma}{(1-\gamma)}\rho_i + \frac{1}{\sigma^2}\rho_i^2 \mu_i^2\Big) \tag{G.24b}$$

Furthermore, if the means have equal norms $\mu := \mu_i$ and the classes are balanced: $\pi_i = 1/k,\ i \in [k]$, then, setting $u_{\mathrm{LS}} := \frac{\mu^2}{\sigma}\sqrt{\frac{1-\gamma}{\mu^2 + k\gamma\sigma^2}}$, it holds that

$$\mathbb{P}_e = \mathbb{P}\left\{G_0 + \max_{j \in [k-1]} G_j \geq u_{\mathrm{LS}}\right\}, \quad G_0, G_1, \ldots, G_{k-1} \overset{iid}{\sim} \mathcal{N}(0,1)\,. \tag{G.25}$$

Proof This is a direct corollary of Theorem 3.2. Indeed, (G.24) can be derived from (3.3a) after substituting $\boldsymbol{V} = \boldsymbol{I}_k, \boldsymbol{\Sigma} = \mathrm{diag}(\mu_1, \mu_2, \ldots, \mu_k)$ and some algebra steps that we omit for brevity.

Instead, we outline below how to conclude (G.26) from (G.24). Assume that $\mu_i = \mu, \forall i \in [k]$ and $\pi_i = \pi = 1/k, \forall i \in [k]$. Recall from (2.9) that $\mathbb{P}\left(\text{error} \,|\, \boldsymbol{y} = \mathbf{e}_c\right) = 1 - \mathbb{P}\left(\boldsymbol{S}_c^{1/2}\boldsymbol{z} > \boldsymbol{t}\right)$, and using (G.24) it can be checked that

$$\boldsymbol{S}_c = \frac{\pi}{1+\pi\mu^2}\left(\frac{\pi\mu^2}{1+\pi\mu^2} + \frac{\gamma}{1-\gamma}\right)(\boldsymbol{I}_k + \mathbf{1}_k \mathbf{1}_k^T) \qquad \text{and} \qquad \boldsymbol{t} = -\frac{\pi\mu^2}{1+\pi\mu^2}\mathbf{1}.$$

Thus, setting

$$u_{\mathrm{LS}} := \mu^2 \sqrt{\frac{\pi}{\pi\mu^2 + \left(\frac{\gamma}{1-\gamma}\right)(1+\pi\mu^2)}} = \mu^2 \sqrt{\frac{1-\gamma}{\mu^2 + \gamma/\pi}}, \tag{G.26}$$

and applying Lemma (C.1), the probability of error is given by the advertised expression. ∎

H Least-squares for MLM

H.1 Proof of Theorem 4.2

H.1.1 Computing $\Sigma_{w,\mu}$

Assume that $\boldsymbol{X}, \boldsymbol{Y}$ are generated from the MLM.

Fix any $\ell \in [k]$. The classifier parameters $\widehat{\boldsymbol{w}}_\ell, \widehat{\boldsymbol{b}}_\ell$ minimize the following objective function $\mathcal{L}_{PO}\left(\boldsymbol{w}_\ell, b_\ell\right) := \frac{1}{2n}\left\|\boldsymbol{X}^T\boldsymbol{w}_\ell + b_\ell \mathbf{1}_n - \boldsymbol{Y}_\ell\right\|_{\ell_2}^2$.

Identifying the AO. To continue further note that by duality we have

$$\min_{\boldsymbol{w}_\ell, \boldsymbol{b}_\ell} \mathcal{L}_{PO}\left(\boldsymbol{w}_\ell, b_\ell\right) = \min_{\boldsymbol{w}_\ell, \boldsymbol{b}_\ell} \max_{\boldsymbol{s}} \frac{1}{n}\left(\boldsymbol{s}^T\boldsymbol{X}^T\boldsymbol{w}_\ell + b_\ell \boldsymbol{s}^T\mathbf{1}_n - \boldsymbol{s}^T\boldsymbol{Y}_\ell - \frac{\|\boldsymbol{s}\|_{\ell_2}^2}{2}\right) \tag{H.1}$$

and the optimization is jointly convex in $(\boldsymbol{w}_\ell, b_\ell)$ and concave in $\boldsymbol{s}$. Here, note that $\boldsymbol{Y}_\ell$ depends on the Gaussian matrix $\boldsymbol{X}$. Thus, before applying the CGMT, we need to break this dependence as follows. Consider the singular value decomposition

$$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = [\boldsymbol{u}_1 \quad \boldsymbol{u}_2 \quad \dots \quad \boldsymbol{u}_r]\,\mathrm{diag}(\sigma_1, \sigma_2, \dots, \sigma_r)\begin{bmatrix}\boldsymbol{v}_1^T\\ \boldsymbol{v}_2^T\\ \dots\\ \boldsymbol{v}_r^T\end{bmatrix}, \tag{H.2}$$

with $\boldsymbol{U} \in \mathbb{R}^{d\times r}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$, and $\boldsymbol{V} \in \mathbb{R}^{k\times r}$ where $r = \mathrm{rank}(\boldsymbol{M}) \le k$. For every $i \in [n]$, we decompose $\boldsymbol{x}_i$ in its projection on the subspace spanned orthogonal columns $\boldsymbol{u}_1, \dots, \boldsymbol{u}_r$ as follows:

$$\boldsymbol{x}_i = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{X}_i + \boldsymbol{P}^\perp\boldsymbol{X}_i = \boldsymbol{U}\tilde{\boldsymbol{g}}_i + \boldsymbol{P}^\perp\boldsymbol{x}_i,$$

where $\boldsymbol{P}^\perp = \boldsymbol{I}_r - \boldsymbol{U}\boldsymbol{U}^T$, and we denote

$$\widetilde{\boldsymbol{G}} := [\tilde{\boldsymbol{g}}_1 \quad \tilde{\boldsymbol{g}}_2 \quad \dots \quad \tilde{\boldsymbol{g}}_n], \quad \tilde{\boldsymbol{g}}_i := \boldsymbol{U}^T\boldsymbol{x}_i \in \mathbb{R}^r, i \in [n]. \tag{H.3}$$

Recalling that $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}_d, \boldsymbol{I}_d)$ note that

$$\tilde{\boldsymbol{g}}_i \sim \mathcal{N}(\boldsymbol{0}_r, \boldsymbol{I}_r) \quad \text{and} \quad \tilde{\boldsymbol{g}}_i \perp \boldsymbol{P}^\perp\boldsymbol{x}_i. \tag{H.4}$$

Further recall that for all $i \in [n]$, conditioned on $\boldsymbol{x}_i$:

$$[\boldsymbol{Y}_\ell]_i \sim \mathrm{Bern}\left(\frac{e^{\boldsymbol{\mu}_\ell^T\boldsymbol{x}_i}}{\sum_{\ell'\in[k]} e^{\boldsymbol{\mu}_{\ell'}^T\boldsymbol{x}_i}}\right) \sim \mathrm{Bern}\left(V_\ell\right), \tag{H.5}$$

where we used (H.3) and the SVD decomposition of $\boldsymbol{M}$. In this notation, we can rewrite the PO as follows:

$$\min_{\boldsymbol{w}_\ell, \boldsymbol{b}_\ell} \max_{\boldsymbol{s}} \frac{1}{n}\left(\boldsymbol{s}^T\boldsymbol{X}^T\boldsymbol{P}^\perp\boldsymbol{w}_\ell + \boldsymbol{s}^T\widetilde{\boldsymbol{G}}^T\boldsymbol{U}^T\boldsymbol{w}_\ell + b_\ell\boldsymbol{s}^T\boldsymbol{1}_n - \boldsymbol{s}^T\boldsymbol{Y}_\ell - \frac{\|\boldsymbol{s}\|_{\ell_2}^2}{2}\right)$$

From (H.4) and (H.5) notice that $\boldsymbol{Y}_\ell$ depends only on $\widetilde{\boldsymbol{G}}$ and $\widetilde{\boldsymbol{G}}$ is independent of $\boldsymbol{X}^T\boldsymbol{P}^\perp$. Therefore, the corresponding Auxiliary Optimization (AO) problem becomes

$$\min_{\boldsymbol{w}_\ell, \boldsymbol{b}_\ell} \max_{\boldsymbol{s}} \frac{1}{n}\left(\|\boldsymbol{P}^\perp\boldsymbol{w}_\ell\|_{\ell_2}\boldsymbol{g}^T\boldsymbol{s} + \|\boldsymbol{s}\|_{\ell_2}\boldsymbol{h}^T\boldsymbol{P}^\perp\boldsymbol{w}_\ell + \boldsymbol{s}^T\widetilde{\boldsymbol{G}}^T\boldsymbol{U}^T\boldsymbol{w}_\ell + b_\ell\boldsymbol{s}^T\boldsymbol{1}_n - \boldsymbol{s}^T\boldsymbol{Y}_\ell - \frac{\|\boldsymbol{s}\|_{\ell_2}^2}{2}\right), \tag{H.6}$$

where $\boldsymbol{g} \in \mathbb{R}^n$ and $\boldsymbol{h} \in \mathbb{R}^d$ are iid Gaussian vectors independent of everything else.

Scalarization of the AO. Maximizing over the direction of $\boldsymbol{s}$ and denoting its norm $\beta = \|\boldsymbol{s}\|_{\ell_2} \ge 0$ we arrive at

$$\begin{aligned}\min_{\boldsymbol{w}_\ell, \boldsymbol{b}_\ell} \max_{\beta\ge 0} \frac{1}{n}&\left(\beta\left\|\|\boldsymbol{P}^\perp\boldsymbol{w}_\ell\|_{\ell_2}\boldsymbol{g} + \boldsymbol{G}^T\boldsymbol{U}^T\boldsymbol{w}_\ell + b_\ell\boldsymbol{1}_n - \boldsymbol{Y}_\ell\right\|_{\ell_2} + \beta\boldsymbol{h}^T\boldsymbol{P}^\perp\boldsymbol{w}_\ell - \frac{\beta^2}{2}\right)\\ &= \min_{\boldsymbol{w}_\ell, \boldsymbol{b}_\ell} \frac{1}{2n}\left(\left\|\|\boldsymbol{P}^\perp\boldsymbol{w}_\ell\|_{\ell_2}\boldsymbol{g} + \widetilde{\boldsymbol{G}}^T\boldsymbol{U}^T\boldsymbol{w}_\ell + b_\ell\boldsymbol{1}_n - \boldsymbol{Y}_\ell\right\|_{\ell_2} + \boldsymbol{h}^T\boldsymbol{P}^\perp\boldsymbol{w}_\ell\right)_+^2\\ &= \frac{1}{2}\left(\min_{\boldsymbol{w}_\ell, \boldsymbol{b}_\ell} \frac{1}{\sqrt{n}}\left\|\|\boldsymbol{P}^\perp\boldsymbol{w}_\ell\|_{\ell_2}\boldsymbol{g} + \widetilde{\boldsymbol{G}}^T\boldsymbol{U}^T\boldsymbol{w}_\ell + b_\ell\boldsymbol{1}_n - \boldsymbol{Y}_\ell\right\|_{\ell_2} + \frac{1}{\sqrt{n}}\boldsymbol{h}^T\boldsymbol{P}^\perp\boldsymbol{w}_\ell\right)_+^2\end{aligned} \tag{H.7}$$

In the remaining, we focus in the inner minimization above. Let us denote

$$\boldsymbol{a} := \boldsymbol{U}^T\boldsymbol{w}_\ell \quad \text{and} \quad \alpha_0 = \|\boldsymbol{P}^\perp\boldsymbol{w}_\ell\|_{\ell_2}.$$

Notice that $\boldsymbol{a} \perp \boldsymbol{P}^\perp\boldsymbol{w}_\ell$ and thus the orthogonal decomposition $\boldsymbol{w}_\ell = \boldsymbol{U}\boldsymbol{a} + \boldsymbol{P}^\perp\boldsymbol{w}_\ell$. With this observation, we can optimize over the direction of $\boldsymbol{P}^T\boldsymbol{w}_\ell$ in (H.7) by aligning it with $\boldsymbol{P}^T\boldsymbol{h}$. With this, the minimization in (H.7) reduces to the following

$$\min_{\boldsymbol{a}, \alpha_0\ge 0, \boldsymbol{b}_\ell} \frac{1}{\sqrt{n}}\left\|\alpha_0\boldsymbol{g} + \widetilde{\boldsymbol{G}}^T\boldsymbol{a} + b_\ell\boldsymbol{1}_n - \boldsymbol{Y}_\ell\right\|_{\ell_2} - \alpha_0\frac{1}{\sqrt{n}}\|\boldsymbol{P}^\perp\boldsymbol{h}\|_{\ell_2}. \tag{H.8}$$

Convergence of the AO. First, we argue on point-wise convergence of the objective function in (H.8). Fix $\boldsymbol{a}, \alpha_0$ and $\boldsymbol{b}_\ell$. From the WLLN, $\frac{1}{\sqrt{n}} \left\| \boldsymbol{P}^\perp \boldsymbol{h} \right\|_{\ell_2} \xrightarrow{P} \sqrt{\gamma}$ and

$$\frac{1}{n} \left\| \alpha_0 \boldsymbol{g} + \widetilde{\boldsymbol{G}}^T \boldsymbol{a} + b_\ell \boldsymbol{1}_n - \boldsymbol{Y}_\ell \right\|_{\ell_2}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\alpha_0 \boldsymbol{g}_i + \boldsymbol{a}^T \tilde{\boldsymbol{g}}_i + b_\ell - [\boldsymbol{Y}_\ell]_i \right)^2 \xrightarrow{P} \mathbb{E}\left[\left(\alpha_0 G_0 + \boldsymbol{a}^T \boldsymbol{g} + \boldsymbol{b}_\ell - Y_\ell \right)^2 \right], \tag{H.9}$$

where the expectation is over $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}_r, \boldsymbol{I}_r)$ (with some abuse of notation) and

$$Y_\ell \sim \text{Bern}(V_\ell) \quad \text{and} \quad V_\ell = \frac{e^{\boldsymbol{e}_\ell^T \boldsymbol{V} \boldsymbol{\Sigma} \boldsymbol{g}}}{\sum_{\ell'=1}^{k} e^{\boldsymbol{e}_{\ell'} \boldsymbol{V} \boldsymbol{\Sigma} \boldsymbol{g}}}. \tag{H.10}$$

Therefore, point-wise on $\boldsymbol{a}, \alpha_0$ and $\boldsymbol{b}_\ell$, the objective of the AO converges to

$$\mathcal{D}_\ell(\alpha_0, \boldsymbol{\alpha}, b_\ell) := \sqrt{\mathbb{E}\left[\left(\alpha_0 G_0 + \boldsymbol{a}^T \boldsymbol{g} + \boldsymbol{b}_\ell - Y_\ell \right)^2 \right]} - \alpha_0 \sqrt{\gamma}. \tag{H.11}$$

Next, with an argument based on convexity and compactness similar to that in "Convergence analysis of the AO" in Section G it can be argued that the convergence above is uniform. Thus,

$$\text{(H.8)} \xrightarrow{P} \min_{\alpha_0 \geq 0, \boldsymbol{\alpha}, b_\ell} \mathcal{D}_\ell(\alpha_0, \boldsymbol{\alpha}, b_\ell). \tag{H.12}$$

Deterministic analysis of the AO. Here, we solve the deterministic minimization problem in (H.12). Optimization over $\boldsymbol{b}_\ell$ is straightforward. By setting

$$\widetilde{\boldsymbol{b}}_\ell = \mathbb{E}[Y_\ell] = \mathbb{E}[V_\ell],$$

we now have to optimize

$$\min_{\alpha_0 \geq 0, \boldsymbol{\alpha}} \sqrt{\alpha_0^2 + \mathbb{E}\left[\left(\boldsymbol{a}^T \boldsymbol{g} - Y_\ell \right)^2 \right] - \left(\mathbb{E}[V_\ell] \right)^2} - \alpha_0 \sqrt{\gamma}. \tag{H.13}$$

By direct differentiation and first-order optimality, we compute the optimal values as follows:

$$\widetilde{\boldsymbol{a}}_j = \mathbb{E}[\boldsymbol{g}_j Y_\ell] = \mathbb{E}[\boldsymbol{g}_j V_\ell], \; j \in [r] \tag{H.14}$$

$$\widetilde{\alpha}_0^2 = \frac{\gamma}{1-\gamma} \left(\text{Var}[Y_\ell] - \sum_{j=1}^{r} \left(\mathbb{E}[\boldsymbol{g}_j V_\ell] \right)^2 \right) = \frac{\gamma}{1-\gamma} \left(\mathbb{E}[V_\ell] - \left(\mathbb{E}[V_\ell] \right)^2 - \sum_{j=1}^{r} \left(\mathbb{E}[\boldsymbol{g}_j V_\ell] \right)^2 \right) \tag{H.15}$$

Asymptotic Predictions. From the analysis above, we conclude with the following limits about the solution $\widehat{\boldsymbol{b}}_\ell, \widehat{\boldsymbol{w}}_\ell$ of the PO:

$$\widehat{\boldsymbol{b}}_\ell \xrightarrow{P} \mathbb{E}[V_\ell] \tag{H.16a}$$

$$\langle \boldsymbol{\mu}_c, \widehat{\boldsymbol{w}}_\ell \rangle \xrightarrow{P} \boldsymbol{e}_c^T \boldsymbol{V} \boldsymbol{\Sigma} \, \mathbb{E}[\boldsymbol{g} V_\ell], \quad c \in [k] \tag{H.16b}$$

$$\| \widehat{\boldsymbol{w}}_\ell \|_{\ell_2}^2 \xrightarrow{P} \sum_{j=1}^{r} \left(\mathbb{E}[\boldsymbol{g}_j V_\ell] \right)^2 + \frac{\gamma}{1-\gamma} \left(\mathbb{E}[V_\ell] - \left(\mathbb{E}[V_\ell] \right)^2 - \sum_{j=1}^{r} \left(\mathbb{E}[\boldsymbol{g}_j V_\ell] \right)^2 \right) \tag{H.16c}$$

$$= \frac{\gamma}{1-\gamma} \left(\mathbb{E}[V_\ell] - \left(\mathbb{E}[V_\ell] \right)^2 \right) + \frac{1-2\gamma}{1-\gamma} \sum_{j=1}^{r} \left(\mathbb{E}[\boldsymbol{g}_j V_\ell] \right)^2 \tag{H.16d}$$

Recall the notation in (4.1). Note that $\mathbb{E}[V_\ell] = \boldsymbol{\pi}_\ell$. Moreover, using Gaussian integration by parts Lemma B.2, it can be shown that $\mathbb{E}[V_\ell \boldsymbol{g}] = \boldsymbol{\Sigma} \boldsymbol{V}^T \left(\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi} \right) \boldsymbol{e}_\ell$. Using these and writing and in matrix form, we arrive at (4.4a)

H.1.2 Computing $\Sigma_{w,w}$

Here, we prove (4.4b). Specifically, we compute the correlations $\langle \widehat{\boldsymbol{w}}_\ell, \widehat{\boldsymbol{w}}_c \rangle, \; \ell \neq c \in [k]$ by following the strategy of Section G.1.2. Specifically, in view of Lemma I.1 we need to study the following PO:

$$\min_{\boldsymbol{w}, \boldsymbol{b}} \max_{\boldsymbol{s}} \frac{1}{n} \left(\boldsymbol{s}^T \boldsymbol{X}^T \boldsymbol{w} + b \boldsymbol{s}^T \boldsymbol{1}_n - \boldsymbol{s}^T (\boldsymbol{Y}_\ell + \boldsymbol{Y}_c) - \frac{\| \boldsymbol{s} \|_{\ell_2}^2}{2} \right) \tag{H.17}$$

which is minimized by $\widehat{\boldsymbol{w}}_\ell + \widehat{\boldsymbol{w}}_c$. Thus the analysis will lead us to an asymptotic formula for $\|\widehat{\boldsymbol{w}}_\ell + \widehat{\boldsymbol{w}}_c\|_{\ell_2}$. This when combined with the formulae for $\|\widehat{\boldsymbol{w}}_\ell\|_{\ell_2}$ and $\|\widehat{\boldsymbol{w}}_c\|_{\ell_2}$ in (H.16d) will give the desired.

The analysis of (H.17) is almost identical to the analysis of (H.1) in the previous section. Specifically, without repeating all the details for brevity, it can be shown that the AO of (H.17) converges to the following (cf. (H.11)):

$$\mathcal{D}_\ell(\alpha_0, \boldsymbol{\alpha}, b_\ell) := \sqrt{\mathbb{E}\left[\left(\alpha_0 G_0 + \boldsymbol{a}^T\boldsymbol{g} + \boldsymbol{b}_\ell - Y_{\ell,c}\right)^2\right]} - \alpha_0\sqrt{\gamma}, \tag{H.18}$$

where as before $G_0 \sim \mathcal{N}(0,1), \boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}_r, \boldsymbol{I}_r)$, only now (H.10) is modified to:

$$Y_{\ell,c} \sim \text{Bern}(V_c + V_\ell) \quad \text{and as before:} \quad V_\ell = \frac{e^{\boldsymbol{e}_\ell^T \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}}{\sum_{\ell'=1}^k e^{\boldsymbol{e}_{\ell'}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}}. \tag{H.19}$$

With these, it can be shown that

$$\|\widehat{\boldsymbol{w}}_\ell + \widehat{\boldsymbol{w}}_c\|_{\ell_2}^2 \xrightarrow{P} \sum_{j=1}^r \left(\mathbb{E}[\boldsymbol{g}_j(V_c + V_\ell)]\right)^2 + \frac{\gamma}{1-\gamma}\left(\mathbb{E}[V_c + V_\ell] - \left(\mathbb{E}[V_c + V_\ell]\right)^2 - \sum_{j=1}^r \left(\mathbb{E}[\boldsymbol{g}_j(V_c + V_\ell)]\right)^2\right)$$

Combining this with (H.16d), we conclude that for $\ell \neq c \in [k]$:

$$\langle \widehat{\boldsymbol{w}}_\ell, \widehat{\boldsymbol{w}}_c \rangle \xrightarrow{P} \frac{1-2\gamma}{1-\gamma} \sum_{j=1}^r \mathbb{E}[\boldsymbol{g}_j V_c]\,\mathbb{E}[\boldsymbol{g}_j V_\ell] - \frac{\gamma}{1-\gamma}\,\mathbb{E}[V_c]\,\mathbb{E}[V_\ell]. \tag{H.20}$$

This shows (4.4b) after applying Gaussian integration by parts and expressing it in matrix form.

H.2 Orthogonal means and equal-energy

Here, we use Theorem 4.2 to prove that, in contrast to the GMM, in the MLM under orthogonal and equal-energy means: LS outperforms the averaging classifier for large enough sample sizes.

$$\boldsymbol{\pi} = \pi_1 \boldsymbol{1}_k = (1/k)\boldsymbol{1}_k, \tag{H.21}$$

$$\boldsymbol{\Pi} = (\boldsymbol{\Pi}_{11} - \boldsymbol{\Pi}_{12})\boldsymbol{I}_k + \boldsymbol{\Pi}_{12}\boldsymbol{1}_k\boldsymbol{1}_k^T \;\text{ with }\; \boldsymbol{\Pi}_{12} = \frac{1 - k^2\boldsymbol{\Pi}_{11}^2}{k(k-1)} \;\text{ and }\; \boldsymbol{\Pi}_{11} = \mathbb{E}\Big[\frac{e^{2\mu G_1}}{\left(\sum_{\ell\in[k]} e^{\mu G_\ell}\right)^2}\Big].$$

Then,

$$\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}} - \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}}\boldsymbol{\Sigma}_{\boldsymbol{\mu},\boldsymbol{\mu}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}}^T \xrightarrow{P} \frac{\gamma}{1-\gamma}\cdot\left(p\boldsymbol{I}_k - q\boldsymbol{1}_k\boldsymbol{1}_k\right), \tag{H.22}$$

where we defined

$$p := \pi_1 - \mu^2(\pi_1 - \boldsymbol{\Pi}_{11} + \boldsymbol{\Pi}_{12})^2 \text{ and } q := (\pi_1^2 + \boldsymbol{\Pi}_{12}^2\mu^2 k - 2\mu^2\boldsymbol{\Pi}_{12}(\pi_1 - \boldsymbol{\Pi}_{11} + \boldsymbol{\Pi}_{12})). \tag{H.23}$$

Thus, similar to (4.3) and with the same notation,

$$\mathbb{P}_{e,\text{LS}} \xrightarrow{P} \mathbb{P}\left\{\arg\max_{\ell\in[k]}\left\{\sqrt{\frac{\gamma}{1-\gamma}}\cdot\left(p\boldsymbol{I}_k - q\boldsymbol{1}_k\boldsymbol{1}_k^T\right)^{1/2}\cdot\widetilde{\boldsymbol{g}} + \mu\cdot(\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi})\,\boldsymbol{g}\right\} \neq Y(\boldsymbol{g})\right\}. \tag{H.24}$$

In (H.24) (as well as in (4.3)), note that the matrices multiplying $\widetilde{\boldsymbol{g}}$ and $\boldsymbol{g}$ have all the form of a rank one update of a (scaled) identity matrix. It turns out that we can exploit this structure to simplify the formulae for the test error even further. Importantly, this lets us directly compare $\mathbb{P}_{e,\text{LS}}$ and $P_{e,\text{Avg}}$ of the two classifiers. These are detailed in Section H.3.

H.3 Proof of Proposition 4.3

In (D.7) and (H.24), we showed the following limits for orthogonal means of equal-energy $\mu > 0$:

$$\mathbb{P}_{e,\text{Avg}} \xrightarrow{P} \mathbb{P}\left(\arg\max_{\ell\in[k]}\left\{\gamma\cdot\boldsymbol{\pi}_1\cdot\boldsymbol{I}_k\cdot\widetilde{\boldsymbol{g}} + \mu\left((\boldsymbol{\pi}_1 - \boldsymbol{\Pi}_{11})\cdot\boldsymbol{I}_k + \boldsymbol{\Pi}_{12}\boldsymbol{1}_k\boldsymbol{1}_k^T\right)\cdot\boldsymbol{g}\right\} \neq Y(\boldsymbol{g})\right)$$

$$\mathbb{P}_{e,\text{LS}} \xrightarrow{P} \mathbb{P}\left(\arg\max_{\ell\in[k]}\left\{\sqrt{\frac{\gamma}{1-\gamma}}\cdot\left(p\boldsymbol{I}_k - q\boldsymbol{1}_k\boldsymbol{1}_k^T\right)^{1/2}\cdot\widetilde{\boldsymbol{g}} + \mu\left((\boldsymbol{\pi}_1 - \boldsymbol{\Pi}_{11})\cdot\boldsymbol{I}_k + \boldsymbol{\Pi}_{12}\boldsymbol{1}_k\boldsymbol{1}_k^T\right)\cdot\boldsymbol{g}\right\} \neq Y(\boldsymbol{g})\right),$$

where $\mathbb{P}(Y(\boldsymbol{g}) = \ell) = \frac{e^{\mu \boldsymbol{g}_\ell}}{\sum_{j\in[k]} e^{\mu \boldsymbol{g}_j}}$ and we have further used (H.21) and the notation in (H.23).

We compare the expression on the RHS in the above display by applying Lemma H.1 below with the following substitutions

$$\boldsymbol{g} \leftarrow \widetilde{\boldsymbol{g}}, \quad \boldsymbol{h} \leftarrow \boldsymbol{g}, \quad c(\boldsymbol{h}) \leftarrow Y(\boldsymbol{g})$$
$$p_2 \leftarrow \frac{\gamma}{1-\gamma}\left(\boldsymbol{\pi}_1 - \mu^2(\boldsymbol{\pi}_1 - \boldsymbol{\Pi}_{11} + \boldsymbol{\Pi}_{12})^2\right), \quad q_2 \leftarrow \frac{\gamma}{1-\gamma}(\boldsymbol{\pi}_1^2 + \boldsymbol{\Pi}_{12}^2\mu^2 k - 2\mu^2\boldsymbol{\Pi}_{12}(\boldsymbol{\pi}_1 - \boldsymbol{\Pi}_{11} + \boldsymbol{\Pi}_{12})),$$
$$p_1 \leftarrow \gamma\boldsymbol{\pi}_1, \qquad q_1 \leftarrow 0.$$

This shows that with probability 1, $\mathbb{P}_{e,\mathrm{LS}} < \mathbb{P}_{e,\mathrm{Avg}}$ if and only if $p_2 < p_1 \iff \gamma < \gamma_\star = \mu^2(\boldsymbol{\pi}_1 - \boldsymbol{\Pi}_{11} + \boldsymbol{\Pi}_{12})^2 \big/ \boldsymbol{\pi}_1$. To retrieve , recall that $\boldsymbol{\pi}_1 = 1/k$ and $k\boldsymbol{\Pi}_{11} + (k^2-k)\boldsymbol{\Pi}_{12} = 1$. The only thing left to prove is that $\gamma_\star < 1$. To see this note that $p_2 > 0$ from positive semi-definiteness of the Schur matrix in (H.23). It takes simple algebra to conclude that $p_2 > 0 \implies \gamma_\star < 1$.

Lemma H.1 *Let $k \geq 2$, $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}_k, \boldsymbol{I}_k)$ $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0}_k, \boldsymbol{I}_k)$, and discrete random variable $c(\boldsymbol{h})$ such that $\mathbb{P}(c(\boldsymbol{h}) = \ell) = e^{\boldsymbol{h}_\ell} \big/ \sum_{j\in[k]} e^{\boldsymbol{h}_j}$. Consider the function $F : \mathbb{R}_{>0} \times \mathbb{R} \to [0,1]$ defined as follows*

$$F(p,q) = \mathbb{P}\left(\arg\max\left\{\left(p\boldsymbol{I}_k - q\boldsymbol{1}_k\boldsymbol{1}_k^T\right)^{1/2}\boldsymbol{g} + \left(\alpha\boldsymbol{I}_k - \beta\boldsymbol{1}_k\boldsymbol{1}_k^T\right)^{1/2}\boldsymbol{h}\right\} \neq c(\boldsymbol{h})\right),$$

such that $p\boldsymbol{I}_k - q\boldsymbol{1}_k\boldsymbol{1}_k^T \succ 0$ and fixed $\alpha\boldsymbol{I}_k - \beta\boldsymbol{1}_k\boldsymbol{1}_k^T \succ 0$. Then, the following statements are true.

1. $F(p,q) = \mathbb{P}\left(\arg\max\left\{\sqrt{p}\cdot\boldsymbol{g} + \sqrt{\alpha}\,\boldsymbol{h}\right\} \neq c(\boldsymbol{h})\right)$.
2. *For $0 < p_2 < p_1$ and any $q_1 < \frac{p_1}{k}, q_2 < \frac{p_2}{k}$, it holds that $F(p_2,q_2) < F(p_1,q_1)$.*

Proof [of Lemma H.1] Fix any $p > 0, q \leq \frac{p}{k}$. Denote $\boldsymbol{T} := \left(p\boldsymbol{I}_k - q\boldsymbol{1}_k\boldsymbol{1}_k^T\right)^{1/2}$ and $\boldsymbol{S} := \left(\alpha\boldsymbol{I}_k - \beta\boldsymbol{1}_k\boldsymbol{1}_k^T\right)^{1/2}$ for convenience. It can be checked that $\boldsymbol{T} := \left(\sqrt{p}\boldsymbol{I}_k + \frac{\sqrt{p-qk}-\sqrt{p}}{k}\right)\boldsymbol{1}_k\boldsymbol{1}_k^T$ and $\boldsymbol{S} := \left(\sqrt{\alpha}\boldsymbol{I}_k + \frac{\sqrt{\alpha-\beta k}-\sqrt{\alpha}}{k}\right)\boldsymbol{1}_k\boldsymbol{1}_k^T$. From these, it follows that

$$F(p,q) = \mathbb{P}\left(\arg\max\left\{\sqrt{p}\cdot\boldsymbol{g} + \sqrt{\alpha}\,\boldsymbol{h}\right\} \neq c(\boldsymbol{h})\right).$$

This shows the first statement.

Next, we show the second statement. Using the distribution of $c(\boldsymbol{h})$ and symmetry we have the following chain of inequalities:

$$\begin{aligned}
1 - F(p,q) &= \mathbb{P}\left(\arg\max_{j\in[k]}\left\{\sqrt{p}\cdot\boldsymbol{g} + \sqrt{\alpha}\,\boldsymbol{h}\right\} = c(\boldsymbol{h})\right) \\
&= k\cdot\mathbb{E}\left[\frac{e^{\boldsymbol{h}_k}}{\sum_{j\in[k]} e^{\boldsymbol{h}_j}}\cdot\mathbb{1}\left\{\arg\max\left\{\sqrt{p}\cdot\boldsymbol{g} + \sqrt{\alpha}\,\boldsymbol{h}\right\} = k\right\}\right] \\
&= k\cdot\mathbb{E}\left[\frac{e^{\boldsymbol{h}_k}}{\sum_{j\in[k]} e^{\boldsymbol{h}_j}}\cdot\prod_{j\in[k-1]}\mathbb{1}\left\{\sqrt{p}\cdot\boldsymbol{g}_j + \sqrt{\alpha}\,\boldsymbol{h}_j < \sqrt{p}\cdot\boldsymbol{g}_k + \sqrt{\alpha}\,\boldsymbol{h}_k\right\}\right] \\
&= k\cdot\mathbb{E}\left[\frac{e^{\boldsymbol{h}_k}}{\sum_{j\in[k]} e^{\boldsymbol{h}_j}}\cdot\prod_{j\in[k-1]}\mathbb{1}\left\{\boldsymbol{g}_j < \boldsymbol{g}_k + \frac{\sqrt{\alpha}\,\boldsymbol{h}_k - \sqrt{\alpha}\,\boldsymbol{h}_j}{\sqrt{p}}\right\}\right] \\
&= k\cdot\mathbb{E}\left[\frac{e^{\boldsymbol{h}_k}}{\sum_{j\in[k]} e^{\boldsymbol{h}_j}}\cdot\prod_{j\in[k-1]} Q\left(\boldsymbol{g}_k + \frac{\sqrt{\alpha}\,\boldsymbol{h}_j - \sqrt{\alpha}\,\boldsymbol{h}_k}{\sqrt{p}}\right)\right] =: k\cdot G(\sqrt{p}), \qquad \text{(H.25)}
\end{aligned}$$

where in the last line we used the rotational symmetry of the Gaussian distribution:

$$\mathbb{P}\left\{\boldsymbol{g}_j < \boldsymbol{g}_k + \frac{\sqrt{\alpha}\,\boldsymbol{h}_k - \sqrt{\alpha}\,\boldsymbol{h}_j}{\sqrt{p}}\,|\,\boldsymbol{h}_1,\ldots,\boldsymbol{h}_k\right\} = \mathbb{P}\left\{\boldsymbol{g}_j > \boldsymbol{g}_k + \frac{\sqrt{\alpha}\,\boldsymbol{h}_j - \sqrt{\alpha}\,\boldsymbol{h}_k}{\sqrt{p}}\,|\,\boldsymbol{h}_1,\ldots,\boldsymbol{h}_k\right\},$$

and the fact that $\boldsymbol{g}_1,\ldots,\boldsymbol{g}_{k-1}$ are independent.

Next, we will show that the function $\mathcal{G}(\cdot)$ defined above is strictly decreasing in $(0,\infty)$. Towards this goal, using $Q'(x) = -\frac{1}{\sqrt{2\pi}} e^{-x^2/2} = -\phi(x)$ and using the shorthand

$$H_{kj} = \boldsymbol{h}_j - \boldsymbol{h}_k,\ j \in [k],$$

we may compute the derivative of $\mathcal{G}$ at any $s > 0$ as follows:

$$\begin{aligned}
\frac{\mathrm{d}\mathcal{G}(s)}{\mathrm{d}s} &= \sum_{i\in[k-1]} \mathbb{E}\left[\frac{e^{\boldsymbol{h}_k}}{\sum_{j\in[k]} e^{\boldsymbol{h}_j}} \cdot \phi\Big(\boldsymbol{g}_k + \frac{\sqrt{\alpha}\, H_{ki}}{s}\Big) \cdot \frac{\sqrt{\alpha}\, H_{ki}}{s^2} \cdot \prod_{j\neq i\in[k-1]} Q\left(\boldsymbol{g}_k + \frac{\sqrt{\alpha}\, H_{kj}}{s}\right)\right] \\
&= \sum_{i\in[k-1]} \mathbb{E}\left[\frac{1}{\sum_{j\in[k]} e^{H_{kj}}} \cdot \phi\Big(\boldsymbol{g}_k + \frac{\sqrt{\alpha}\, H_{ki}}{s}\Big) \cdot \frac{\sqrt{\alpha}\, H_{ki}}{s^2} \cdot \prod_{j\neq i\in[k-1]} Q\left(\boldsymbol{g}_k + \frac{\sqrt{\alpha}\, H_{kj}}{s}\right)\right] \\
&= \sum_{i\in[k-1]} \frac{\sqrt{\alpha}}{s^2}\, \mathbb{E}\left[H_{ki} \cdot \mathcal{A}_i\left(\boldsymbol{g}_k, \{H_{kj}\}_{j\in[k-1]}\right)\right], \qquad \text{(H.26)}
\end{aligned}$$

where in the last line we have defined

$$\mathcal{A}_i\left(\boldsymbol{g}_k, \{H_{kj}\}_{j\in[k-1]}\right) := \frac{1}{1+\sum_{j\in[k-1]} e^{H_{kj}}} \cdot \phi\Big(\boldsymbol{g}_k + \frac{\sqrt{\alpha}\, H_{ki}}{s}\Big) \cdot \prod_{j\neq i\in[k-1]} Q\left(\boldsymbol{g}_k + \frac{\sqrt{\alpha}\, H_{kj}}{s}\right),$$

Next, we use Gaussian integration by parts (GIBP) to further simplify the expression in (H.26). Fix any $i \in [k-1]$. Then, by (GIBP):

$$\begin{aligned}
A_i &:= \mathbb{E}\left[H_{ki} \cdot \mathcal{A}_i\left(H_{k1}, H_{k2}, \ldots, H_{kk}\right)\right] \qquad \text{(H.27)} \\
&= \mathbb{E}[H_{ki}^2]\, \mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}H_{ki}} \mathcal{A}_i\left(\boldsymbol{g}_k, \{H_{kj}\}_{j\in[k-1]}\right)\right] + \sum_{\substack{\ell\in[k-1]\\ \ell\neq i}} \mathbb{E}[H_{ki}\cdot H_{k\ell}]\, \mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}H_{k\ell}} \mathcal{A}_i\left(\boldsymbol{g}_k, \{H_{kj}\}_{j\in[k-1]}\right)\right] \\
&= \underbrace{2\, \mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}H_{ki}} \mathcal{A}_i\left(\boldsymbol{g}_k, \{H_{kj}\}_{j\in[k-1]}\right)\right]}_{\text{TermI}} + \underbrace{\sum_{\substack{\ell\in[k-1]\\ \ell\neq i}} \mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}H_{k\ell}} \mathcal{A}_i\left(\boldsymbol{g}_k, \{H_{kj}\}_{j\in[k-1]}\right)\right]}_{\text{TermII}}, \qquad \text{(H.28)}
\end{aligned}$$

where in the second line, we used the fact that $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k)$ to compute

$$\mathbb{E}[H_{ki}^2] = 2, \quad \text{and} \quad \mathbb{E}[H_{ki} H_{k\ell}] = 1,\ \ell \neq i,\ \ell \in [k-1].$$

We now compute the derivatives in (H.28). First, for any $\ell \in [k-1],\ \ell \neq i$,

$$\begin{aligned}
&\frac{\mathrm{d}\mathcal{A}_i\left(\boldsymbol{g}_k, \{H_{kj}\}_{j\in[k-1]}\right)}{\mathrm{d}H_{k\ell}} = -\frac{e^{H_{k\ell}}}{\left(1+\sum_{j\in[k-1]} e^{H_{kj}}\right)^2} \cdot \phi\Big(\boldsymbol{g}_k + \frac{\sqrt{\alpha}\, H_{ki}}{s}\Big) \cdot \prod_{j\neq i\in[k-1]} Q\left(\boldsymbol{g}_k + \frac{\sqrt{\alpha}\, H_{kj}}{s}\right) \quad =: \text{TermII(a)}_\ell \\
&- \frac{\sqrt{\alpha}}{s} \cdot \frac{1}{1+\sum_{j\in[k-1]} e^{H_{kj}}} \cdot \phi\Big(\boldsymbol{g}_k + \frac{\sqrt{\alpha}\, H_{ki}}{s}\Big) \cdot \phi\Big(\boldsymbol{g}_k + \frac{\sqrt{\alpha}\, H_{k\ell}}{s}\Big) \cdot \prod_{j\neq(i,\ell)\in[k-1]} Q\left(\boldsymbol{g}_k + \frac{\sqrt{\alpha}\, H_{kj}}{s}\right) \quad =: \text{TermII(b)}_\ell \\
&\qquad = \text{TermII(a)}_\ell + \text{TermII(b)}_\ell \qquad \text{(H.29)}
\end{aligned}$$

Thus,

$$\text{TermII} = \sum_{\ell\neq i\in[k-1]} \mathbb{E}\left[\text{TermII(a)}_\ell\right] + \mathbb{E}\left[\text{TermII(b)}_\ell\right] =: \mathbb{E}\left[\text{TermII(a)}\right] + N_i, \qquad \text{(H.30)}$$

where we defined

$$N_i = -\frac{\sqrt{\alpha}}{s} \cdot \mathbb{E}\left[\frac{\phi\Big(\boldsymbol{g}_k + \frac{\sqrt{\alpha}\, H_{ki}}{s}\Big)}{1+\sum_{j\in[k-1]} e^{H_{kj}}} \cdot \sum_{\ell\neq i\in[k-1]} \left\{\phi\Big(\boldsymbol{g}_k + \frac{\sqrt{\alpha}\, H_{k\ell}}{s}\Big) \cdot \prod_{j\neq(i,\ell)\in[k-1]} Q\left(\boldsymbol{g}_k + \frac{\sqrt{\alpha}\, H_{kj}}{s}\right)\right\}\right] < 0. \qquad \text{(H.31)}$$

and we remark for later use that

$$\mathbb{E}\left[\text{TermII(a)}\right] = \sum_{\ell\neq i\in[k-1]} \mathbb{E}\left[\text{TermII(a)}_\ell\right] < 0. \qquad \text{(H.32)}$$

Second, it holds that

$$\begin{aligned}
\frac{\mathrm{d}\mathcal{A}_i\left(\boldsymbol{g}_k, \{H_{kj}\}_{j\in[k-1]}\right)}{\mathrm{d}H_{ki}} &= -\frac{e^{H_{ki}}}{\left(1+\sum_{j\in[k-1]} e^{H_{kj}}\right)^2}\cdot\phi\Big(\boldsymbol{g}_k+\frac{\sqrt{\alpha}\,H_{ki}}{s}\Big)\cdot\prod_{j\neq i\in[k-1]} Q\left(\boldsymbol{g}_k+\frac{\sqrt{\alpha}\,H_{kj}}{s}\right)\\
&\quad+\frac{\mathrm{d}\phi\Big(\boldsymbol{g}_k+\frac{\sqrt{\alpha}\,H_{ki}}{s}\Big)}{\mathrm{d}H_{ki}}\cdot\frac{1}{1+\sum_{j\in[k-1]} e^{H_{kj}}}\cdot\prod_{j\neq i\in[k-1]} Q\left(\boldsymbol{g}_k+\frac{\sqrt{\alpha}\,H_{kj}}{s}\right)\\
&= -\frac{e^{H_{ki}}}{\left(1+\sum_{j\in[k-1]} e^{H_{kj}}\right)^2}\cdot\phi\Big(\boldsymbol{g}_k+\frac{\sqrt{\alpha}\,H_{ki}}{s}\Big)\cdot\prod_{j\neq i\in[k-1]} Q\left(\boldsymbol{g}_k+\frac{\sqrt{\alpha}\,H_{kj}}{s}\right) \quad =:\mathrm{TermI(a)}\\
&\quad-\frac{\sqrt{\alpha}}{s}\left(\boldsymbol{g}_k+\frac{\sqrt{\alpha}\,H_{ki}}{s}\right)\phi\Big(\boldsymbol{g}_k+\frac{\sqrt{\alpha}\,H_{ki}}{s}\Big)\cdot\frac{1}{1+\sum_{j\in[k-1]} e^{H_{kj}}}\cdot\prod_{j\neq i\in[k-1]} Q\left(\boldsymbol{g}_k+\frac{\sqrt{\alpha}\,H_{kj}}{s}\right) \quad =:\mathrm{TermI(b)},\\
&= \mathrm{TermI(a)}+\mathrm{TermI(b)},
\end{aligned} \tag{H.33}$$

where in the penultimate line we used the fact that $\phi'(x) = -x\phi(x)$. Consider the two terms in (H.33). Clearly,

$$\mathbb{E}[\mathrm{TermI(a)}] < 0. \tag{H.34}$$

For the second term we observe that:

$$\mathbb{E}\left[\mathrm{TermI(b)}\right] = -\frac{\sqrt{\alpha}}{s}\,\mathbb{E}\left[\left(\boldsymbol{g}_k+\frac{\sqrt{\alpha}\,H_{ki}}{s}\right)\cdot\mathcal{A}_i\left(\boldsymbol{g}_k,\{H_{kj}\}_{j\in[k-1]}\right)\right] \tag{H.35}$$

$$\begin{aligned}
&= -\frac{\alpha}{s^2}\cdot\mathbb{E}\left[H_{ki}\cdot\mathcal{A}_i\left(\boldsymbol{g}_k,\{H_{kj}\}_{j\in[k-1]}\right)\right]-\frac{\sqrt{\alpha}}{s}\,\mathbb{E}\left[\boldsymbol{g}_k\cdot\mathcal{A}_i\left(\boldsymbol{g}_k,\{H_{kj}\}_{j\in[k-1]}\right)\right]\\
&= -\frac{\alpha}{s^2}\cdot A_i-\frac{\sqrt{\alpha}}{s}\,\mathbb{E}\left[\boldsymbol{g}_k\cdot\mathcal{A}_i\left(\boldsymbol{g}_k,\{H_{kj}\}_{j\in[k-1]}\right)\right].
\end{aligned} \tag{H.36}$$

Moreover, using again GIBP, $\mathbb{E}[\boldsymbol{g}_k^2] = 1$, $\mathbb{E}[\boldsymbol{g}_k H_{kj}] = 0,\ j\in[k]$ and the fact that $\phi'(x) = -x\phi(x)$,

$$\begin{aligned}
&\mathbb{E}\left[\boldsymbol{g}_k\cdot\mathcal{A}_i\left(\boldsymbol{g}_k,\{H_{kj}\}_{j\in[k-1]}\right)\right] = \mathbb{E}\left[\frac{\mathrm{d}\mathcal{A}_i\left(\boldsymbol{g}_k,\{H_{kj}\}_{j\in[k]}\right)}{\mathrm{d}\boldsymbol{g}_k}\right]\\
&= -\,\mathbb{E}\left[\left(\boldsymbol{g}_k+\frac{\sqrt{\alpha}\,H_{ki}}{s}\right)\cdot\mathcal{A}_i\left(\boldsymbol{g}_k,\{H_{kj}\}_{j\in[k-1]}\right)\right]\\
&\quad-\sum_{\ell\neq i\in[k-1]}\mathbb{E}\left[\phi\Big(\boldsymbol{g}_k+\frac{\sqrt{\alpha}\,H_{k\ell}}{s}\Big)\cdot\frac{1}{1+\sum_{j\in[k-1]} e^{H_{k\ell}}}\cdot\phi\Big(\boldsymbol{g}_k+\frac{\sqrt{\alpha}\,H_{ki}}{s}\Big)\cdot\prod_{j\neq(i,\ell)\in[k-1]} Q\left(\boldsymbol{g}_k+\frac{\sqrt{\alpha}\,H_{kj}}{s}\right)\right]\\
&= \frac{s}{\sqrt{\alpha}}\,\mathbb{E}\left[\mathrm{TermI(b)}\right]+\frac{s}{\sqrt{\alpha}}N_i,
\end{aligned} \tag{H.37}$$

where, we have recalled (H.35) and (H.31). Using (H.37) in (H.36), we find that

$$\mathbb{E}\left[\mathrm{TermI(b)}\right] = -\frac{\alpha}{s^2}\cdot A_i-\mathbb{E}\left[\mathrm{TermI(b)}\right]-N_i \implies \mathbb{E}\left[\mathrm{TermI(b)}\right] = -\frac{\alpha}{2s^2}\cdot A_i-\frac{N_i}{2}. \tag{H.38}$$

We are now ready to put things together:

$$\begin{aligned}
A_i &= 2\cdot\mathrm{TermI}+\mathrm{TermII} && \text{by (H.28)}\\
&= 2\,\mathbb{E}[\mathrm{TermI(a)}]+2\,\mathbb{E}[\mathrm{TermI(b)}]+\mathbb{E}[\mathrm{TermII(a)}]+N_i && \text{by (H.30)}\\
&= 2\,\mathbb{E}[\mathrm{TermI(a)}]-\frac{\alpha}{s^2}\cdot A_i-N_i+\mathbb{E}[\mathrm{TermII(a)}]+N_i && \text{by (H.38)}\\
\implies A_i &= \frac{s^2}{s^2+\alpha}\left(2\,\mathbb{E}[\mathrm{TermI(a)}]+\mathbb{E}[\mathrm{TermII(a)}]\right)\\
&< 0 && \text{by (H.34) and (H.32).}
\end{aligned}$$

From this, (H.26) and (H.27), we have shown that $\mathcal{G}$ is strictly decreasing in $(0,\infty)$. Recalling the definition of $\mathcal{G}$ in (H.25), this implies that $F(p,q)$ is strictly increasing in $p > 0$, as desired to complete the proof. ∎

I Weighted LS for GMM (Proof of Theorem 3.4)

I.1 Computing $\Sigma_{w,\mu}$

$$\min_{\boldsymbol{W}\in\mathbb{R}^{k\times d},\, \boldsymbol{b}\in\mathbb{R}^k} \quad \frac{1}{2n}\left\|\left(\boldsymbol{W}\boldsymbol{X}+\boldsymbol{b}\boldsymbol{1}_n^T-\boldsymbol{Y}\right)\boldsymbol{D}\right\|_F^2 = \sum_{\ell=1}^k \min_{\boldsymbol{w}_\ell,\boldsymbol{b}_\ell} \frac{1}{2n}\left\|\boldsymbol{D}\left(\boldsymbol{X}^T\boldsymbol{w}_\ell+b_\ell\boldsymbol{1}_n-\boldsymbol{Y}_\ell\right)\right\|_{\ell_2}^2$$
$$= \sum_{\ell=1}^k \min_{\boldsymbol{w}_\ell,\boldsymbol{b}_\ell} \frac{1}{2n}\left\|\boldsymbol{D}\left(\boldsymbol{Y}^T\boldsymbol{M}^T\boldsymbol{w}_\ell+\boldsymbol{Z}^T\boldsymbol{w}_\ell+b_\ell\boldsymbol{1}_n-\boldsymbol{Y}_\ell\right)\right\|_{\ell_2}^2$$

Define

$$\mathcal{L}_{PO}\left(\boldsymbol{w}_\ell,b_\ell\right) := \frac{1}{2n}\left\|\boldsymbol{D}\left(\boldsymbol{Y}^T\boldsymbol{M}^T\boldsymbol{w}_\ell+\boldsymbol{Z}^T\boldsymbol{w}_\ell+b_\ell\boldsymbol{1}_n-\boldsymbol{Y}_\ell\right)\right\|_{\ell_2}^2 \tag{I.1}$$

Identifying the AO. To continue further note that by duality we have

$$\min_{\boldsymbol{w}_\ell,\boldsymbol{b}_\ell}\mathcal{L}_{PO}\left(\boldsymbol{w}_\ell,b_\ell\right) = \min_{\boldsymbol{u},\boldsymbol{w}_\ell,\boldsymbol{b}_\ell}\max_{\boldsymbol{s}}\frac{1}{n}\left(\boldsymbol{s}^T\boldsymbol{D}\boldsymbol{Y}^T\boldsymbol{M}^T\boldsymbol{w}_\ell+\boldsymbol{s}^T\boldsymbol{D}\boldsymbol{Z}^T\boldsymbol{w}_\ell+b_\ell\boldsymbol{s}^T\boldsymbol{D}\boldsymbol{1}_n-\boldsymbol{s}^T\boldsymbol{D}\boldsymbol{Y}_\ell-\boldsymbol{s}^T\boldsymbol{u}+\frac{\|\boldsymbol{u}\|_{\ell_2}^2}{2}\right)$$

Note that the above is jointly convex in $(\boldsymbol{u},\boldsymbol{w}_\ell,b_\ell)$ and concave in $\boldsymbol{s}$. We consider the Auxiliary Optimization (AO) problem

$$\min_{\boldsymbol{u},\boldsymbol{w}_\ell,\boldsymbol{b}_\ell}\max_{\boldsymbol{s}}\frac{1}{n}\left(\boldsymbol{s}^T\boldsymbol{D}\boldsymbol{Y}^T\boldsymbol{M}^T\boldsymbol{w}_\ell+\sigma\|\boldsymbol{w}_\ell\|_{\ell_2}\boldsymbol{g}^T\boldsymbol{D}\boldsymbol{s}+\sigma\|\boldsymbol{D}\boldsymbol{s}\|_{\ell_2}\boldsymbol{h}^T\boldsymbol{w}_\ell+b_\ell\boldsymbol{s}^T\boldsymbol{D}\boldsymbol{1}_n-\boldsymbol{s}^T\boldsymbol{D}\boldsymbol{Y}_\ell-\boldsymbol{s}^T\boldsymbol{u}+\frac{\|\boldsymbol{u}\|_{\ell_2}^2}{2}\right),$$

where $\boldsymbol{g}\in\mathbb{R}^n$ and $\boldsymbol{h}\in\mathbb{R}^d$ are independent Gaussian random vectors with i.i.d. $\mathcal{N}(0,1)$ entries. To continue further we carry out a change of variable $\boldsymbol{s}\to\boldsymbol{D}\boldsymbol{s}$ to arrive at

$$\min_{\boldsymbol{u},\boldsymbol{w}_\ell,\boldsymbol{b}_\ell}\max_{\boldsymbol{s}}\frac{1}{n}\left(\boldsymbol{s}^T\boldsymbol{Y}^T\boldsymbol{M}^T\boldsymbol{w}_\ell+\sigma\|\boldsymbol{w}_\ell\|_{\ell_2}\boldsymbol{g}^T\boldsymbol{s}+\sigma\|\boldsymbol{s}\|_{\ell_2}\boldsymbol{h}^T\boldsymbol{w}_\ell+b_\ell\boldsymbol{s}^T\boldsymbol{1}_n-\boldsymbol{s}^T\boldsymbol{Y}_\ell-\boldsymbol{s}^T\boldsymbol{D}^{-1}\boldsymbol{u}+\frac{\|\boldsymbol{u}\|_{\ell_2}^2}{2}\right),$$

Maximizing over the direction of $\boldsymbol{s}$ and its norm $\beta=\|\boldsymbol{s}\|_{\ell_2}$ we arrive at

$$\min_{\boldsymbol{u},\boldsymbol{w}_\ell,\boldsymbol{b}_\ell}\max_{\beta\geq0}\max_{\boldsymbol{s}:\|\boldsymbol{s}\|_{\ell_2}=1}\frac{1}{n}\left(\beta\boldsymbol{s}^T\boldsymbol{Y}^T\boldsymbol{M}^T\boldsymbol{w}_\ell+\sigma\beta\|\boldsymbol{w}_\ell\|_{\ell_2}\boldsymbol{g}^T\boldsymbol{s}+\sigma\beta\boldsymbol{h}^T\boldsymbol{w}_\ell+b_\ell\beta\boldsymbol{s}^T\boldsymbol{1}_n-\beta\boldsymbol{s}^T\boldsymbol{Y}_\ell-\beta\boldsymbol{s}^T\boldsymbol{D}^{-1}\boldsymbol{u}+\frac{\|\boldsymbol{u}\|_{\ell_2}^2}{2}\right)$$
$$=\min_{\boldsymbol{u},\boldsymbol{w}_\ell,\boldsymbol{b}_\ell}\max_{\beta\geq0}\frac{1}{n}\left(\beta\left\|\boldsymbol{Y}^T\boldsymbol{M}^T\boldsymbol{w}_\ell+\sigma\|\boldsymbol{w}_\ell\|_{\ell_2}\boldsymbol{g}+b_\ell\boldsymbol{1}_n-\boldsymbol{Y}_\ell-\boldsymbol{D}^{-1}\boldsymbol{u}\right\|_{\ell_2}+\sigma\beta\boldsymbol{h}^T\boldsymbol{w}_\ell+\frac{\|\boldsymbol{u}\|_{\ell_2}^2}{2}\right)$$
$$=\min_{\boldsymbol{u},\boldsymbol{b}_\ell}\max_{\beta\geq0}\min_{\boldsymbol{w}_\ell}\frac{1}{n}\left(\beta\left\|\boldsymbol{Y}^T\boldsymbol{M}^T\boldsymbol{w}_\ell+\sigma\|\boldsymbol{w}_\ell\|_{\ell_2}\boldsymbol{g}+b_\ell\boldsymbol{1}_n-\boldsymbol{Y}_\ell-\boldsymbol{D}^{-1}\boldsymbol{u}\right\|_{\ell_2}+\sigma\beta\boldsymbol{h}^T\boldsymbol{w}_\ell+\frac{\|\boldsymbol{u}\|_{\ell_2}^2}{2}\right)$$

To continue, consider the singular value decomposition

$$\boldsymbol{M}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T=\begin{bmatrix}\boldsymbol{u}_1 & \boldsymbol{u}_2 & \dots & \boldsymbol{u}_r\end{bmatrix}\operatorname{diag}(\sigma_1,\sigma_2,\dots,\sigma_r)\begin{bmatrix}\boldsymbol{v}_1^T\\ \boldsymbol{v}_2^T\\ \dots\\ \boldsymbol{v}_r^T\end{bmatrix}, \tag{I.2}$$

with $r:=rank(\boldsymbol{M})\leq k$ and define the variable $\boldsymbol{\alpha}=\boldsymbol{U}^T\boldsymbol{w}_\ell$ and $\boldsymbol{\alpha}_\perp=\boldsymbol{U}_\perp^T\boldsymbol{w}_\ell$ where $\boldsymbol{U}_\perp$ is the orthogonal complement of the columns of $\boldsymbol{U}$. With these definitions the above optimization problem reduces to

$$\min_{\boldsymbol{u},\boldsymbol{b}_\ell}\max_{\beta\geq0}\min_{\boldsymbol{\alpha}}\min_{\boldsymbol{\alpha}_\perp}\frac{1}{n}\left(\beta\left\|\boldsymbol{Y}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}+\sigma\sqrt{\|\boldsymbol{\alpha}\|_{\ell_2}^2+\|\boldsymbol{\alpha}_\perp\|_{\ell_2}^2}\boldsymbol{g}+b_\ell\boldsymbol{1}_n-\boldsymbol{Y}_\ell-\boldsymbol{D}^{-1}\boldsymbol{u}\right\|_{\ell_2}+\sigma\beta\boldsymbol{h}^T\boldsymbol{U}\boldsymbol{\alpha}+\sigma\beta\boldsymbol{h}^T\boldsymbol{U}_\perp\boldsymbol{\alpha}_\perp+\frac{\|\boldsymbol{u}\|_{\ell_2}^2}{2}\right)$$

Decomposing the optimization over $\boldsymbol{\alpha}_\perp$ in terms of its direction and norm $\alpha_0=\|\boldsymbol{\alpha}_\perp\|_{\ell_2}$ we arrive at

$$\min_{\boldsymbol{u},\boldsymbol{b}_\ell}\max_{\beta\geq0}\min_{\boldsymbol{\alpha}}\min_{\alpha_0\geq0}\frac{1}{n}\left(\beta\left\|\boldsymbol{Y}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}+\sigma\sqrt{\|\boldsymbol{\alpha}\|_{\ell_2}^2+\alpha_0^2}\boldsymbol{g}+b_\ell\boldsymbol{1}_n-\boldsymbol{Y}_\ell-\boldsymbol{D}^{-1}\boldsymbol{u}\right\|_{\ell_2}+\sigma\beta\boldsymbol{h}^T\boldsymbol{U}\boldsymbol{\alpha}-\sigma\alpha_0\beta\left\|\boldsymbol{U}_\perp^T\boldsymbol{h}\right\|_{\ell_2}+\frac{\|\boldsymbol{u}\|_{\ell_2}^2}{2}\right)$$

Since $\boldsymbol{U}^T\boldsymbol{h}$ is $r \le k$ dimensional in our asymptotic regime the $\boldsymbol{h}^T\boldsymbol{U}\boldsymbol{\alpha}$ term can be ignored. Also replacing β with $\beta/\sqrt{n}$ we thus arrive at

$$\begin{aligned}
&\min_{\boldsymbol{u},\boldsymbol{b}_\ell}\ \max_{\beta\ge 0}\ \min_{\boldsymbol{\alpha}}\ \min_{\alpha_0\ge 0}\ \frac{\beta}{\sqrt{n}}\left\|\boldsymbol{Y}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}+\sigma\sqrt{\|\boldsymbol{\alpha}\|_{\ell_2}^2+\alpha_0^2}\boldsymbol{g}+b_\ell\mathbf{1}_n-\boldsymbol{Y}_\ell-\boldsymbol{D}^{-1}\boldsymbol{u}\right\|_{\ell_2}-\frac{1}{\sqrt{n}}\sigma\alpha_0\beta\left\|\boldsymbol{U}_\perp^T\boldsymbol{h}\right\|_{\ell_2}+\frac{\|\boldsymbol{u}\|_{\ell_2}^2}{2n}\\
&=\min_{\boldsymbol{u},\boldsymbol{b}_\ell}\ \max_{\beta\ge 0}\ \min_{\boldsymbol{\alpha}}\ \min_{\alpha_0\ge 0}\ \min_{\tau\ge 0}\ \frac{\beta}{2n\tau}\left\|\boldsymbol{Y}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}+\sigma\sqrt{\|\boldsymbol{\alpha}\|_{\ell_2}^2+\alpha_0^2}\boldsymbol{g}+b_\ell\mathbf{1}_n-\boldsymbol{Y}_\ell-\boldsymbol{D}^{-1}\boldsymbol{u}\right\|_{\ell_2}^2+\frac{\beta\tau}{2}\\
&\qquad\qquad-\frac{1}{\sqrt{n}}\sigma\alpha_0\beta\left\|\boldsymbol{U}_\perp^T\boldsymbol{h}\right\|_{\ell_2}+\frac{\|\boldsymbol{u}\|_{\ell_2}^2}{2n}\\
&=\min_{\boldsymbol{b}_\ell}\ \max_{\beta\ge 0}\ \min_{\boldsymbol{\alpha}}\ \min_{\alpha_0\ge 0}\ \min_{\tau\ge 0}\ \min_{\boldsymbol{u}}\ \frac{\beta}{2n\tau}\left\|\boldsymbol{Y}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}+\sigma\sqrt{\|\boldsymbol{\alpha}\|_{\ell_2}^2+\alpha_0^2}\boldsymbol{g}+b_\ell\mathbf{1}_n-\boldsymbol{Y}_\ell-\boldsymbol{D}^{-1}\boldsymbol{u}\right\|_{\ell_2}^2+\frac{\beta\tau}{2}\\
&\qquad\qquad-\frac{1}{\sqrt{n}}\sigma\alpha_0\beta\left\|\boldsymbol{U}_\perp^T\boldsymbol{h}\right\|_{\ell_2}+\frac{\|\boldsymbol{u}\|_{\ell_2}^2}{2n}
\end{aligned}$$

Setting the derivative with respect to $\boldsymbol{u}$ to zero we arrive at

$$\boldsymbol{u}=\frac{\beta}{\tau}\boldsymbol{D}^{-1}\left(\boldsymbol{I}+\frac{\beta}{\tau}\boldsymbol{D}^{-2}\right)^{-1}\left(\boldsymbol{Y}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}+\sigma\sqrt{\|\boldsymbol{\alpha}\|_{\ell_2}^2+\alpha_0^2}\boldsymbol{g}+b_\ell\mathbf{1}_n-\boldsymbol{Y}_\ell\right)$$

Plugging the latter into the above the AO simplifies to

$$\min_{\boldsymbol{b}_\ell}\ \max_{\beta\ge 0}\ \min_{\boldsymbol{\alpha}}\ \min_{\alpha_0\ge 0}\ \min_{\tau\ge 0}\ \frac{\beta}{2\tau n}\mathrm{trace}\left(\boldsymbol{t}^T\left(\boldsymbol{I}+\frac{\beta}{\tau}\boldsymbol{D}^{-2}\right)^{-1}\boldsymbol{t}\right)-\frac{1}{\sqrt{n}}\sigma\alpha_0\beta\left\|\boldsymbol{U}_\perp^T\boldsymbol{h}\right\|_{\ell_2}+\frac{\beta\tau}{2}$$

where

$$\boldsymbol{t}:=\boldsymbol{Y}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}+\sigma\sqrt{\|\boldsymbol{\alpha}\|_{\ell_2}^2+\alpha_0^2}\boldsymbol{g}+b_\ell\mathbf{1}_n-\boldsymbol{Y}_\ell=\boldsymbol{Y}^T\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}-\boldsymbol{e}_\ell\right)+\sigma\sqrt{\|\boldsymbol{\alpha}\|_{\ell_2}^2+\alpha_0^2}\boldsymbol{g}+b_\ell\mathbf{1}_n$$

To continue note that in our asymptotic regime we have

$$\frac{1}{\sqrt{n}}\left\|\boldsymbol{U}_\perp^T\boldsymbol{h}\right\|_{\ell_2}\ \xrightarrow{P}\ \sqrt{\gamma}$$

and the cross terms can be ignored so that in an asymptotic sense

$$\begin{aligned}
&\frac{1}{n}\mathrm{trace}\left(\boldsymbol{t}^T\left(\boldsymbol{I}+\frac{\beta}{\tau}\boldsymbol{D}^{-2}\right)^{-1}\boldsymbol{t}\right)\\
&\quad=\frac{1}{n}\sigma^2\left(\|\boldsymbol{\alpha}\|_{\ell_2}^2+\alpha_0^2\right)\mathrm{trace}\left(\left(\boldsymbol{I}+\frac{\beta}{\tau}\boldsymbol{D}^{-2}\right)^{-1}\right)\\
&\qquad+\frac{1}{n}\mathrm{trace}\left(\left(\boldsymbol{Y}^T\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}-\boldsymbol{e}_\ell\right)+b_\ell\mathbf{1}_n\right)^T\left(\boldsymbol{I}+\frac{\beta}{\tau}\boldsymbol{D}^{-2}\right)^{-1}\left(\boldsymbol{Y}^T\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}-\boldsymbol{e}_\ell\right)+b_\ell\mathbf{1}_n\right)\right)
\end{aligned}$$

Therefore we arrive at

$$\begin{aligned}
&\min_{\boldsymbol{b}_\ell}\ \max_{\beta\ge 0}\ \min_{\boldsymbol{\alpha}}\ \min_{\tau\ge 0}\ \min_{\alpha_0\ge 0}\ \frac{\beta}{2\tau n}\sigma^2\left(\|\boldsymbol{\alpha}\|_{\ell_2}^2+\alpha_0^2\right)\mathrm{trace}\left(\left(\boldsymbol{I}+\frac{\beta}{\tau}\boldsymbol{D}^{-2}\right)^{-1}\right)-\sigma\alpha_0\beta\sqrt{\gamma}+\frac{\beta\tau}{2}\\
&\qquad\qquad+\frac{\beta}{2\tau n}\mathrm{trace}\left(\left(\boldsymbol{Y}^T\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}-\boldsymbol{e}_\ell\right)+b_\ell\mathbf{1}_n\right)^T\left(\boldsymbol{I}+\frac{\beta}{\tau}\boldsymbol{D}^{-2}\right)^{-1}\left(\boldsymbol{Y}^T\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}-\boldsymbol{e}_\ell\right)+b_\ell\mathbf{1}_n\right)\right)
\end{aligned}$$

which can be rewritten in the form

$$\begin{aligned}
&\min_{\boldsymbol{b}_\ell}\ \max_{\beta\ge 0}\ \min_{\boldsymbol{\alpha}}\ \min_{\tau\ge 0}\ \frac{\beta}{2\tau n}\sigma^2\|\boldsymbol{\alpha}\|_{\ell_2}^2\mathrm{trace}\left(\left(\boldsymbol{I}+\frac{\beta}{\tau}\boldsymbol{D}^{-2}\right)^{-1}\right)+\frac{\beta\tau}{2}\\
&\qquad\qquad+\frac{\beta}{2\tau n}\mathrm{trace}\left(\left(\boldsymbol{Y}^T\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}-\boldsymbol{e}_\ell\right)+b_\ell\mathbf{1}_n\right)^T\left(\boldsymbol{I}+\frac{\beta}{\tau}\boldsymbol{D}^{-2}\right)^{-1}\left(\boldsymbol{Y}^T\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}-\boldsymbol{e}_\ell\right)+b_\ell\mathbf{1}_n\right)\right)\\
&\qquad\qquad+\frac{\beta}{2\tau n}\sigma^2\alpha_0^2\mathrm{trace}\left(\left(\boldsymbol{I}+\frac{\beta}{\tau}\boldsymbol{D}^{-2}\right)^{-1}\right)-\alpha_0\sigma\beta\sqrt{\gamma}
\end{aligned}$$

To continue further we shall assume $\boldsymbol{D} = \operatorname{diag}\left(\boldsymbol{Y}^T\boldsymbol{\omega}\right)$. Note that in this case

$$\frac{1}{n}\operatorname{trace}\left(\left(\boldsymbol{I}+\frac{\beta}{\tau}\boldsymbol{D}^{-2}\right)^{-1}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{(\boldsymbol{y}_i^T\boldsymbol{\omega})^2}{(\boldsymbol{y}_i^T\boldsymbol{\omega})^2+\frac{\beta}{\tau}} = \sum_{\ell=1}^{k}\frac{n_\ell}{n}\frac{\omega_\ell^2}{\omega_\ell^2+\frac{\beta}{\tau}} \xrightarrow{P} \sum_{\ell=1}^{k}\frac{\pi_\ell\omega_\ell^2}{\omega_\ell^2+\frac{\beta}{\tau}}$$

Also

$$\begin{aligned}
&\frac{1}{n}\operatorname{trace}\left(\left(\boldsymbol{Y}^T\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}-\boldsymbol{e}_\ell\right)+b_\ell\mathbf{1}_n\right)^T\left(\boldsymbol{I}+\frac{\beta}{\tau}\boldsymbol{D}^{-2}\right)^{-1}\left(\boldsymbol{Y}^T\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}-\boldsymbol{e}_\ell\right)+b_\ell\mathbf{1}_n\right)\right)\\
&= \frac{1}{n}\operatorname{trace}\left(\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}-\boldsymbol{e}_\ell\right)^T\boldsymbol{Y}\left(\boldsymbol{I}+\frac{\beta}{\tau}\boldsymbol{D}^{-2}\right)^{-1}\boldsymbol{Y}^T\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}-\boldsymbol{e}_\ell\right)\right)\\
&\quad+\frac{2}{n}b_\ell\mathbf{1}_n^T\left(\boldsymbol{I}+\frac{\beta}{\tau}\boldsymbol{D}^{-2}\right)^{-1}\boldsymbol{Y}^T\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}-\boldsymbol{e}_\ell\right)+\frac{b_\ell^2}{n}\operatorname{trace}\left(\left(\boldsymbol{I}+\frac{\beta}{\tau}\boldsymbol{D}^{-2}\right)^{-1}\right)\\
&\xrightarrow{P} \operatorname{trace}\left(\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}-\boldsymbol{e}_\ell\right)^T\operatorname{diag}\left(\frac{\pi_1\omega_1^2}{\omega_1^2+\frac{\beta}{\tau}},\frac{\pi_2\omega_2^2}{\omega_2^2+\frac{\beta}{\tau}},\ldots,\frac{\pi_k\omega_k^2}{\omega_k^2+\frac{\beta}{\tau}}\right)\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}-\boldsymbol{e}_\ell\right)\right)\\
&\quad+2b_\ell\begin{bmatrix}\frac{\pi_1\omega_1^2}{\omega_1^2+\frac{\beta}{\tau}} & \frac{\pi_2\omega_2^2}{\omega_2^2+\frac{\beta}{\tau}} & \ldots & \frac{\pi_k\omega_k^2}{\omega_k^2+\frac{\beta}{\tau}}\end{bmatrix}\left(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\alpha}-\boldsymbol{e}_\ell\right)+b_\ell^2\left(\sum_{\ell=1}^{k}\frac{\pi_\ell\omega_\ell^2}{\omega_\ell^2+\frac{\beta}{\tau}}\right)
\end{aligned}$$

Next define

$$\begin{aligned}
\boldsymbol{A}(\eta) &:= \begin{bmatrix}\sigma^2\left(\boldsymbol{\pi}^T\boldsymbol{\nu}(\eta)\right)\boldsymbol{I}+\boldsymbol{\Sigma}\boldsymbol{V}^T\operatorname{diag}(\boldsymbol{\pi})\operatorname{diag}(\boldsymbol{\nu}(\eta))\boldsymbol{V}\boldsymbol{\Sigma} & \boldsymbol{\Sigma}\boldsymbol{V}^T\operatorname{diag}(\boldsymbol{\nu}(\eta))\boldsymbol{\pi}\\ \boldsymbol{\pi}^T\operatorname{diag}(\boldsymbol{\nu}(\eta))\boldsymbol{V}\boldsymbol{\Sigma} & \boldsymbol{\pi}^T\boldsymbol{\nu}(\eta)\end{bmatrix}\\
\boldsymbol{c}_\ell &:= \begin{bmatrix}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{e}_\ell\\ 1\end{bmatrix}
\end{aligned} \tag{I.3}$$

where

$$\boldsymbol{\nu}(\eta) = \frac{1}{\gamma}\begin{bmatrix}\frac{\omega_1^2}{\omega_1^2+\eta}\\ \frac{\omega_2^2}{\omega_2^2+\eta}\\ \ldots\\ \frac{\omega_k^2}{\omega_k^2+\eta}\end{bmatrix} \tag{I.4}$$

We thus arrive at

$$\begin{aligned}
\min_{\boldsymbol{\alpha}}\min_{\boldsymbol{b}_\ell}\max_{\alpha_0\geq 0}\max_{\beta\geq 0}\min_{\tau\geq 0}\ &\frac{\gamma\beta}{2\tau}\left(\pi_\ell\boldsymbol{\nu}_\ell\left(\frac{\beta}{\tau}\right)+\begin{bmatrix}\boldsymbol{\alpha}^T & b_\ell\end{bmatrix}\boldsymbol{A}\left(\frac{\beta}{\tau}\right)\begin{bmatrix}\boldsymbol{\alpha}\\ b_\ell\end{bmatrix}-2\pi_\ell\boldsymbol{\nu}_\ell\left(\frac{\beta}{\tau}\right)\boldsymbol{c}_\ell^T\begin{bmatrix}\boldsymbol{\alpha}\\ b_\ell\end{bmatrix}\right)\\
&+\frac{\gamma\beta}{2\tau}\sigma^2\left(\boldsymbol{\pi}^T\boldsymbol{\nu}\left(\frac{\beta}{\tau}\right)\right)\alpha_0^2-\alpha_0\sigma\beta\sqrt{\gamma}+\frac{\beta\tau}{2}
\end{aligned}$$

Setting the derivative of the above with respect to α_0 to zero we arrive at

$$\frac{\gamma\beta}{\tau}\sigma^2\left(\boldsymbol{\pi}^T\boldsymbol{\nu}\left(\frac{\beta}{\tau}\right)\right)\alpha_0-\sigma\beta\sqrt{\gamma}=0 \quad\Rightarrow\quad \alpha_0=\frac{\tau}{\sigma\sqrt{\gamma}\left(\boldsymbol{\pi}^T\boldsymbol{\nu}\left(\frac{\beta}{\tau}\right)\right)}$$

Note that the above objective has the form

$$f\left(\frac{\beta}{\tau}\right)-\alpha_0\sigma\beta\sqrt{\gamma}+\frac{\beta\tau}{2}$$

with

$$f(\eta) := \frac{\eta\gamma}{2}\left(\pi_\ell\boldsymbol{\nu}_\ell(\eta)+\begin{bmatrix}\boldsymbol{\alpha}^T & b_\ell\end{bmatrix}\boldsymbol{A}(\eta)\begin{bmatrix}\boldsymbol{\alpha}\\ b_\ell\end{bmatrix}-2\pi_\ell\boldsymbol{\nu}_\ell(\eta)\boldsymbol{c}_\ell^T\begin{bmatrix}\boldsymbol{\alpha}\\ b_\ell\end{bmatrix}\right)+\frac{\gamma\eta}{2}\sigma^2\left(\boldsymbol{\pi}^T\boldsymbol{\nu}(\eta)\right)\alpha_0^2.$$

Thus, the derivatives with respect to β and τ to zero we have

$$\frac{1}{\tau}f'\left(\frac{\beta}{\tau}\right)-\alpha_0\sigma\sqrt{\gamma}+\frac{\tau}{2}=0 \quad\Rightarrow\quad f'\left(\frac{\beta}{\tau}\right)-\alpha_0\sigma\sqrt{\gamma}\tau+\frac{\tau^2}{2}=0 \quad\Rightarrow\quad f'\left(\frac{\beta}{\tau}\right)=\tau^2\left(\frac{1}{\boldsymbol{\pi}^T\boldsymbol{\nu}\left(\frac{\beta}{\tau}\right)}-\frac{1}{2}\right)$$

and

$$-\frac{\beta}{\tau^2}f'\left(\frac{\beta}{\tau}\right)+\frac{\beta}{2}=0 \quad \Rightarrow \quad \tau^2=2f'\left(\frac{\beta}{\tau}\right)$$

Combining the latter two we conclude that $\boldsymbol{\pi}^T\boldsymbol{\nu}\left(\frac{\beta}{\tau}\right)=1$. Thus, $\eta=\frac{\beta}{\tau}$ is the solution to $\boldsymbol{\pi}^T\boldsymbol{\nu}(\eta)=1$. To calculate τ and hence α_0 we calculate f' which is equal to

$$\begin{aligned}f'(\eta)=&\frac{\gamma}{2}\left(\pi_\ell\boldsymbol{\nu}_\ell(\eta)+\begin{bmatrix}\boldsymbol{\alpha}^T & b_\ell\end{bmatrix}\boldsymbol{A}(\eta)\begin{bmatrix}\boldsymbol{\alpha}\\ b_\ell\end{bmatrix}-2\pi_\ell\boldsymbol{\nu}_\ell(\eta)\boldsymbol{c}_\ell^T\begin{bmatrix}\boldsymbol{\alpha}\\ b_\ell\end{bmatrix}\right)+\frac{\gamma}{2}\sigma^2\left(\boldsymbol{\pi}^T\boldsymbol{\nu}(\eta)\right)\alpha_0^2+\frac{\gamma\eta}{2}\sigma^2\alpha_0^2(\boldsymbol{\pi}^T\boldsymbol{\nu}'(\eta))\\&+\frac{\gamma\eta}{2}\left(\pi_\ell\boldsymbol{\nu}_\ell'(\eta)+\begin{bmatrix}\boldsymbol{\alpha}^T & b_\ell\end{bmatrix}\boldsymbol{A}'(\eta)\begin{bmatrix}\boldsymbol{\alpha}\\ b_\ell\end{bmatrix}-2\pi_\ell\boldsymbol{\nu}_\ell'(\eta)\boldsymbol{c}_\ell^T\begin{bmatrix}\boldsymbol{\alpha}\\ b_\ell\end{bmatrix}\right)\end{aligned}$$

Where

$$\begin{aligned}\boldsymbol{\nu}'(\eta)=&-\frac{1}{\gamma}\begin{bmatrix}\frac{\omega_1^2}{(\omega_1^2+\eta)^2}\\ \frac{\omega_2^2}{(\omega_2^2+\eta)^2}\\ \cdots\\ \frac{\omega_k^2}{(\omega_k^2+\eta)^2}\end{bmatrix}\\ \boldsymbol{A}'(\eta):=&\begin{bmatrix}\sigma^2\left(\boldsymbol{\pi}^T\boldsymbol{\nu}'(\eta)\right)\boldsymbol{I}+\boldsymbol{\Sigma}\boldsymbol{V}^T\mathrm{diag}\left(\boldsymbol{\pi}\right)\mathrm{diag}\left(\boldsymbol{\nu}'(\eta)\right)\boldsymbol{V}\boldsymbol{\Sigma} & \boldsymbol{\Sigma}\boldsymbol{V}^T\mathrm{diag}\left(\boldsymbol{\nu}'(\eta)\right)\boldsymbol{\pi}\\ \boldsymbol{\pi}^T\mathrm{diag}\left(\boldsymbol{\nu}'(\eta)\right)\boldsymbol{V}\boldsymbol{\Sigma} & \boldsymbol{\pi}^T\boldsymbol{\nu}'(\eta)\end{bmatrix}\\ \boldsymbol{c}_\ell:=&\begin{bmatrix}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{e}_\ell\\ 1\end{bmatrix}\end{aligned}$$

Now note that at the optimal point we have

$$f'(\eta)=\frac{\tau^2}{2}=\frac{\gamma}{2}\sigma^2\left(\boldsymbol{\pi}^T\boldsymbol{\nu}(\eta)\right)\alpha_0^2$$

Thus from the above we can conclude that

$$\begin{aligned}\alpha_0^2=&-\frac{1}{\eta\sigma^2(\boldsymbol{\pi}^T\boldsymbol{\nu}'(\eta))}\left(\pi_\ell\boldsymbol{\nu}_\ell(\eta)+\begin{bmatrix}\boldsymbol{\alpha}^T & b_\ell\end{bmatrix}\boldsymbol{A}(\eta)\begin{bmatrix}\boldsymbol{\alpha}\\ b_\ell\end{bmatrix}-2\pi_\ell\boldsymbol{\nu}_\ell(\eta)\boldsymbol{c}_\ell^T\begin{bmatrix}\boldsymbol{\alpha}\\ b_\ell\end{bmatrix}\right)\\&-\frac{1}{\sigma^2(\boldsymbol{\pi}^T\boldsymbol{\nu}'(\eta))}\left(\pi_\ell\boldsymbol{\nu}_\ell'(\eta)+\begin{bmatrix}\boldsymbol{\alpha}^T & b_\ell\end{bmatrix}\boldsymbol{A}'(\eta)\begin{bmatrix}\boldsymbol{\alpha}\\ b_\ell\end{bmatrix}-2\pi_\ell\boldsymbol{\nu}_\ell'(\eta)\boldsymbol{c}_\ell^T\begin{bmatrix}\boldsymbol{\alpha}\\ b_\ell\end{bmatrix}\right)\end{aligned}$$

Thus the AO optimization problem reduces to

$$\min_{\boldsymbol{b}_\ell}\ \min_{\boldsymbol{\alpha}}\ \frac{\eta\gamma}{2}\left(\pi_\ell\boldsymbol{\nu}_\ell+\begin{bmatrix}\boldsymbol{\alpha}^T & b_\ell\end{bmatrix}\boldsymbol{A}\begin{bmatrix}\boldsymbol{\alpha}\\ b_\ell\end{bmatrix}-2\pi_\ell\boldsymbol{\nu}_\ell\boldsymbol{c}_\ell^T\begin{bmatrix}\boldsymbol{\alpha}\\ b_\ell\end{bmatrix}\right)$$

where η is the solution to

$$\sum_{\ell=1}^{k}\frac{\pi_\ell\omega_\ell^2}{\omega_\ell^2+\eta}=\gamma$$

and

$$\begin{aligned}\boldsymbol{A}(\eta):=&\begin{bmatrix}\sigma^2\boldsymbol{I}_r+\boldsymbol{\Sigma}\boldsymbol{V}^T\mathrm{diag}\left(\boldsymbol{\pi}\odot\boldsymbol{\nu}\right)\boldsymbol{V}\boldsymbol{\Sigma} & \boldsymbol{\Sigma}\boldsymbol{V}^T\left(\boldsymbol{\pi}\odot\boldsymbol{\nu}\right)\\ \left(\boldsymbol{\pi}\odot\boldsymbol{\nu}\right)^T\boldsymbol{V}\boldsymbol{\Sigma} & 1\end{bmatrix}\\ \boldsymbol{c}_\ell:=&\begin{bmatrix}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{e}_\ell\\ 1\end{bmatrix}\end{aligned}\tag{I.5}$$

where

$$\boldsymbol{\nu}:=\frac{1}{\gamma}\begin{bmatrix}\frac{\omega_1^2}{\omega_1^2+\eta}\\ \frac{\omega_2^2}{\omega_2^2+\eta}\\ \cdots\\ \frac{\omega_k^2}{\omega_k^2+\eta}\end{bmatrix}$$

First, note that the matrix $\boldsymbol{A}$ is positive definite. This can be checked by computing the Schur complement of $\boldsymbol{A}$:

$$\boldsymbol{\Delta} \coloneqq \sigma^2 \boldsymbol{I}_r + \boldsymbol{\Sigma V}^T \boldsymbol{P V \Sigma} \coloneqq \sigma^2 \boldsymbol{I}_r + \boldsymbol{\Sigma V}^T \left(\operatorname{diag}(\boldsymbol{\pi} \odot \boldsymbol{\nu}) - (\boldsymbol{\pi} \odot \boldsymbol{\nu}) (\boldsymbol{\pi} \odot \boldsymbol{\nu})^T \right) \boldsymbol{V \Sigma} \succ \boldsymbol{0}_{r \times r}. \quad \text{(I.6)}$$

Positive definiteness above holds because $\boldsymbol{P} \coloneqq \left(\operatorname{diag}(\boldsymbol{\pi} \odot \boldsymbol{\nu}) - (\boldsymbol{\pi} \odot \boldsymbol{\nu}) (\boldsymbol{\pi} \odot \boldsymbol{\nu})^T \right) \succeq \boldsymbol{0}_{k \times k}$. Thus the objective is a strictly convex quadratic and is jointly convex in its arguments. We proceed by minimizing the objective over $(\boldsymbol{\alpha}, b_\ell)$ which is equal to

$$\begin{bmatrix} \widehat{\boldsymbol{\alpha}} \\ \widehat{b}_\ell \end{bmatrix} = \pi_\ell \nu_\ell \boldsymbol{A}^{-1} \boldsymbol{c}_\ell = \pi_\ell \nu_\ell \begin{bmatrix} -\boldsymbol{\Delta}^{-1} \boldsymbol{\Sigma V}^T (\boldsymbol{\pi} \odot \boldsymbol{\nu} - \boldsymbol{e}_\ell) \\ 1 + (\boldsymbol{\pi} \odot \boldsymbol{\nu})^T \boldsymbol{V \Sigma \Delta}^{-1} \boldsymbol{\Sigma V}^T (\boldsymbol{\pi} \odot \boldsymbol{\nu} - \boldsymbol{e}_\ell) \end{bmatrix} \quad \text{(I.7)}$$

Thus, the minimum value attained is

$$\begin{aligned} & -\pi_\ell^2 \nu_\ell^2 \begin{bmatrix} -(\boldsymbol{\pi} \odot \boldsymbol{\nu} - \boldsymbol{e}_\ell)^T \boldsymbol{V \Sigma} & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\Delta}^{-1} & \boldsymbol{0} \\ \boldsymbol{0}^T & 1 \end{bmatrix} \begin{bmatrix} -\boldsymbol{\Sigma V}^T (\boldsymbol{\pi} \odot \boldsymbol{\nu} - \boldsymbol{e}_\ell) \\ 1 \end{bmatrix} \quad \text{(I.8)} \\ & = -\pi_\ell^2 \nu_\ell^2 \left(1 + (\boldsymbol{\pi} \odot \boldsymbol{\nu} - \boldsymbol{e}_\ell)^T \boldsymbol{V \Sigma \Delta}^{-1} \boldsymbol{\Sigma V}^T (\boldsymbol{\pi} \odot \boldsymbol{\nu} - \boldsymbol{e}_\ell) \right). \end{aligned}$$

Thus the objective reduces to

$$\frac{\eta\gamma}{2} \left(\pi_\ell \nu_\ell (1 - \pi_\ell \nu_\ell) - \pi_\ell^2 \nu_\ell^2 (\boldsymbol{\pi} \odot \boldsymbol{\nu} - \boldsymbol{e}_\ell)^T \boldsymbol{V \Sigma \Delta}^{-1} \boldsymbol{\Sigma V}^T (\boldsymbol{\pi} \odot \boldsymbol{\nu} - \boldsymbol{e}_\ell) \right). \quad \text{(I.9)}$$

Therefore,

$$\begin{aligned} \alpha_0^2 = & -\frac{1}{\eta\sigma^2(\boldsymbol{\pi}^T \boldsymbol{\nu}'(\eta))} \left(\pi_\ell \nu_\ell (1 - \pi_\ell \nu_\ell) - \pi_\ell^2 \nu_\ell^2 (\boldsymbol{\pi} \odot \boldsymbol{\nu} - \boldsymbol{e}_\ell)^T \boldsymbol{V \Sigma \Delta}^{-1} \boldsymbol{\Sigma V}^T (\boldsymbol{\pi} \odot \boldsymbol{\nu} - \boldsymbol{e}_\ell) \right) \\ & - \frac{1}{\sigma^2(\boldsymbol{\pi}^T \boldsymbol{\nu}'(\eta))} \left(\pi_\ell \boldsymbol{\nu}_\ell'(\eta) + \begin{bmatrix} \boldsymbol{\alpha}^T & b_\ell \end{bmatrix} \boldsymbol{A}'(\eta) \begin{bmatrix} \boldsymbol{\alpha} \\ b_\ell \end{bmatrix} - 2\pi_\ell \boldsymbol{\nu}_\ell'(\eta) \boldsymbol{c}_\ell^T \begin{bmatrix} \boldsymbol{\alpha} \\ b_\ell \end{bmatrix} \right) \end{aligned}$$

where

$$\begin{bmatrix} \widehat{\boldsymbol{\alpha}} \\ \widehat{b}_\ell \end{bmatrix} = \pi_\ell \nu_\ell \begin{bmatrix} -\boldsymbol{\Delta}^{-1} \boldsymbol{\Sigma V}^T (\boldsymbol{\pi} \odot \boldsymbol{\nu} - \boldsymbol{e}_\ell) \\ 1 + (\boldsymbol{\pi} \odot \boldsymbol{\nu})^T \boldsymbol{V \Sigma \Delta}^{-1} \boldsymbol{\Sigma V}^T (\boldsymbol{\pi} \odot \boldsymbol{\nu} - \boldsymbol{e}_\ell) \end{bmatrix} \quad \text{(I.10)}$$

and

$$\boldsymbol{\nu}'(\eta) = -\frac{1}{\gamma} \begin{bmatrix} \frac{\omega_1^2}{(\omega_1^2+\eta)^2} \\ \frac{\omega_2^2}{(\omega_2^2+\eta)^2} \\ \cdots \\ \frac{\omega_k^2}{(\omega_k^2+\eta)^2} \end{bmatrix}$$

$$\boldsymbol{A}'(\eta) \coloneqq \begin{bmatrix} \sigma^2 \left(\boldsymbol{\pi}^T \boldsymbol{\nu}'(\eta) \right) \boldsymbol{I} + \boldsymbol{\Sigma V}^T \operatorname{diag}(\boldsymbol{\pi}) \operatorname{diag}(\boldsymbol{\nu}'(\eta)) \boldsymbol{V \Sigma} & \boldsymbol{\Sigma V}^T \operatorname{diag}(\boldsymbol{\nu}'(\eta)) \boldsymbol{\pi} \\ \boldsymbol{\pi}^T \operatorname{diag}(\boldsymbol{\nu}'(\eta)) \boldsymbol{V \Sigma} & \boldsymbol{\pi}^T \boldsymbol{\nu}'(\eta) \end{bmatrix}$$

To continue note that

$$\begin{aligned} & \pi_\ell \boldsymbol{\nu}_\ell'(\eta) + \begin{bmatrix} \boldsymbol{\alpha}^T & b_\ell \end{bmatrix} \boldsymbol{A}'(\eta) \begin{bmatrix} \boldsymbol{\alpha} \\ b_\ell \end{bmatrix} - 2\pi_\ell \boldsymbol{\nu}_\ell'(\eta) \boldsymbol{c}_\ell^T \begin{bmatrix} \boldsymbol{\alpha} \\ b_\ell \end{bmatrix} \\ & \qquad = \pi_\ell \boldsymbol{\nu}_\ell'(\eta) + \pi_\ell^2 \nu_\ell^2 \boldsymbol{c}_\ell^T \boldsymbol{A}^{-1} \boldsymbol{A}'(\eta) \boldsymbol{A}^{-1} \boldsymbol{c}_\ell - 2\pi_\ell^2 \nu_\ell' \nu_\ell \boldsymbol{c}_\ell^T \boldsymbol{A}^{-1} \boldsymbol{c}_\ell \end{aligned}$$

Thus,

$$\begin{aligned} \alpha_0^2 = & -\frac{1}{\eta\sigma^2(\boldsymbol{\pi}^T \boldsymbol{\nu}'(\eta))} \left(\pi_\ell \nu_\ell (1 - \pi_\ell \nu_\ell) - \pi_\ell^2 \nu_\ell^2 (\boldsymbol{\pi} \odot \boldsymbol{\nu} - \boldsymbol{e}_\ell)^T \boldsymbol{V \Sigma \Delta}^{-1} \boldsymbol{\Sigma V}^T (\boldsymbol{\pi} \odot \boldsymbol{\nu} - \boldsymbol{e}_\ell) \right) \\ & - \frac{1}{\sigma^2(\boldsymbol{\pi}^T \boldsymbol{\nu}'(\eta))} \left(\pi_\ell \boldsymbol{\nu}_\ell'(\eta) + \pi_\ell^2 \nu_\ell^2 \boldsymbol{c}_\ell^T \boldsymbol{A}^{-1} \boldsymbol{A}'(\eta) \boldsymbol{A}^{-1} \boldsymbol{c}_\ell - 2\pi_\ell^2 \nu_\ell' \nu_\ell \boldsymbol{c}_\ell^T \boldsymbol{A}^{-1} \boldsymbol{c}_\ell \right) \\ = & \frac{\gamma}{\eta\sigma^2 \left(\sum_{\ell=1}^k \frac{\pi_\ell \omega_\ell^2}{(\omega_\ell^2+\eta)^2} \right)} \left(\pi_\ell \nu_\ell (1 - \pi_\ell \nu_\ell) - \pi_\ell^2 \nu_\ell^2 (\boldsymbol{\pi} \odot \boldsymbol{\nu} - \boldsymbol{e}_\ell)^T \boldsymbol{V \Sigma \Delta}^{-1} \boldsymbol{\Sigma V}^T (\boldsymbol{\pi} \odot \boldsymbol{\nu} - \boldsymbol{e}_\ell) \right) \\ & + \frac{\gamma}{\sigma^2 \left(\sum_{\ell=1}^k \frac{\pi_\ell \omega_\ell^2}{(\omega_\ell^2+\eta)^2} \right)} \left(\pi_\ell \boldsymbol{\nu}_\ell'(\eta) + \pi_\ell^2 \nu_\ell^2 \boldsymbol{c}_\ell^T \boldsymbol{A}^{-1} \boldsymbol{A}'(\eta) \boldsymbol{A}^{-1} \boldsymbol{c}_\ell - 2\pi_\ell^2 \nu_\ell' \nu_\ell \boldsymbol{c}_\ell^T \boldsymbol{A}^{-1} \boldsymbol{c}_\ell \right) \\ \coloneqq & \frac{\zeta}{\sigma^2} \left(\pi_\ell \nu_\ell (1 - \pi_\ell \nu_\ell) - \pi_\ell^2 \nu_\ell^2 (\boldsymbol{\pi} \odot \boldsymbol{\nu} - \boldsymbol{e}_\ell)^T \boldsymbol{V \Sigma \Delta}^{-1} \boldsymbol{\Sigma V}^T (\boldsymbol{\pi} \odot \boldsymbol{\nu} - \boldsymbol{e}_\ell) \right) \\ & + \frac{\zeta\eta}{\sigma^2} \left(\pi_\ell \boldsymbol{\nu}_\ell'(\eta) + \pi_\ell^2 \nu_\ell^2 \boldsymbol{c}_\ell^T \boldsymbol{A}^{-1} \boldsymbol{A}'(\eta) \boldsymbol{A}^{-1} \boldsymbol{c}_\ell - 2\pi_\ell^2 \nu_\ell' \nu_\ell \boldsymbol{c}_\ell^T \boldsymbol{A}^{-1} \boldsymbol{c}_\ell \right) \end{aligned}$$

where $\zeta := \frac{\gamma}{\eta\left(\Sigma_{\ell=1}^{k} \frac{\pi_\ell \omega_\ell^2}{(\omega_\ell^2+\eta)^2}\right)}$.

Asymptotic predictions. First, from (I.7) the bias term converges as follows:

$$\widehat{b}_\ell \xrightarrow{P} \pi_\ell \nu_\ell \left(1 + (\boldsymbol{\pi} \odot \boldsymbol{\nu})^T \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{\pi} \odot \boldsymbol{\nu} - \boldsymbol{e}_\ell)\right).$$

Thus,

$$\widehat{\boldsymbol{b}} \xrightarrow{P} \left(\boldsymbol{I}_k - \boldsymbol{P}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\right)(\boldsymbol{\pi} \odot \boldsymbol{\nu})$$

Recall that $\boldsymbol{\alpha} = \boldsymbol{U}^T\boldsymbol{w}_\ell$. Thus, the correlations $\langle \boldsymbol{\mu}_i, \boldsymbol{w}_\ell \rangle,\ i \in [k]$ converge as follows:

$$\boldsymbol{M}^T\boldsymbol{w}_\ell = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^T\boldsymbol{w}_\ell \xrightarrow{P} \boldsymbol{V}\boldsymbol{\Sigma}\widehat{\boldsymbol{\alpha}} = -\pi_\ell\nu_\ell \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{\pi} \odot \boldsymbol{\nu} - \boldsymbol{e}_\ell). \tag{I.11}$$

Here, convergence applies element-wise to the entries of the involved random vectors. Moreover, from the analysis above we can predict the limit of the norm $\|\boldsymbol{w}_\ell\|_{\ell_2}$. For this, note that $\|\boldsymbol{w}_\ell\|_{\ell_2}^2 = \widehat{\alpha}_0^2 + \widehat{\boldsymbol{\alpha}}^T\widehat{\boldsymbol{\alpha}}$. Thus,

$$\begin{aligned}\|\boldsymbol{w}_\ell\|_{\ell_2}^2 \xrightarrow{P} &\frac{\zeta}{\sigma^2}\pi_\ell\nu_\ell(1-\pi_\ell\nu_\ell) + \pi_\ell^2\nu_\ell^2 (\boldsymbol{\pi} \odot \boldsymbol{\nu} - \boldsymbol{e}_\ell)^T \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\left(\boldsymbol{\Delta}^{-1} - \frac{\zeta}{\sigma^2}\boldsymbol{I}_r\right)\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{\pi} \odot \boldsymbol{\nu} - \boldsymbol{e}_\ell) \\ &+ \frac{\eta\zeta}{\sigma^2}\left(\pi_\ell\boldsymbol{\nu}_\ell'(\eta) + \pi_\ell^2\nu_\ell^2\boldsymbol{c}_\ell^T\boldsymbol{A}^{-1}\boldsymbol{A}'\boldsymbol{A}^{-1}\boldsymbol{c}_\ell - 2\pi_\ell^2\nu_\ell'\nu_\ell\boldsymbol{c}_\ell^T\boldsymbol{A}^{-1}\boldsymbol{c}_\ell\right)\end{aligned} \tag{I.12}$$

I.2 Computing $\Sigma_{w,w}$

In the previous section we used the CGMT to predict the bias $\widehat{b}_\ell$, the correlations $\langle \boldsymbol{\mu}_i, \widehat{\boldsymbol{w}}_\ell \rangle,\ i[k]$ and the norm $\|\widehat{\boldsymbol{w}}_\ell\|_{\ell_2}$ for all $\ell \in [k]$ members of the multi-output classifier. Here, we show how to compute the limits of the cross-correlations $\langle \widehat{\boldsymbol{w}}_\ell, \widehat{\boldsymbol{w}}_j \rangle, \ell \neq j \in [k]$.

Lemma I.1 *For $\ell \neq j \in [k]$ let $\widehat{\boldsymbol{w}}_\ell$ $\widehat{\boldsymbol{w}}_j$ be solutions to the least-squares minimization (2.4), i.e.,*

$$\begin{aligned}&(\widehat{\boldsymbol{w}}_\ell, \widehat{\boldsymbol{w}}_j, \widehat{b}_\ell, \widehat{b}_j) \\ &\quad = \arg\min_{\boldsymbol{w}_\ell, \boldsymbol{w}_j, b_\ell, b_j} \left\{\frac{1}{2n}\left\|\boldsymbol{D}\left(\boldsymbol{Y}_\ell - \boldsymbol{X}^T\boldsymbol{w}_\ell - b_\ell\boldsymbol{1}_n\right)\right\|_{\ell_2}^2 + \frac{1}{2n}\left\|\boldsymbol{D}\left(\boldsymbol{Y}_j - \boldsymbol{X}^T\boldsymbol{w}_j - b_j\boldsymbol{1}_n\right)\right\|_{\ell_2}^2\right\}.\end{aligned}$$

Denote $\widehat{\boldsymbol{w}}_{\ell,j} := \widehat{\boldsymbol{w}}_\ell + \widehat{\boldsymbol{w}}_j$ and $\widehat{b}_{\ell,j} := \widehat{b}_\ell + \widehat{b}_j$. Then, $(\widehat{\boldsymbol{w}}_{\ell,j}, \widehat{b}_{\ell,j})$ is a minimizer in the following least-squares problem:

$$(\widehat{\boldsymbol{w}}_{\ell,j}, \widehat{b}_{\ell,j}) = \arg\min_{\boldsymbol{w},b} \frac{1}{2n}\left\|\boldsymbol{D}\left(\boldsymbol{Y}_\ell + \boldsymbol{Y}_j - \boldsymbol{X}^T\boldsymbol{w} - b\boldsymbol{1}_n\right)\right\|_{\ell_2}^2 \tag{I.13}$$

Proof Clearly the minimization in (I.13) is convex. Thus, it suffices to prove that $\widehat{\boldsymbol{w}}_\ell + \widehat{\boldsymbol{w}}_j$ satisfies the KKT conditions. First, by optimality of $\widehat{\boldsymbol{w}}_\ell$, we have that

$$\boldsymbol{X}\boldsymbol{D}^2\left(\boldsymbol{Y}_\ell - \boldsymbol{X}^T\widehat{\boldsymbol{w}}_\ell - \widehat{b}_\ell\boldsymbol{1}_n\right) = 0$$

Similarly, for $\widehat{\boldsymbol{w}}_j$:

$$\boldsymbol{X}\boldsymbol{D}^2\left(\boldsymbol{Y}_j - \boldsymbol{X}^T\widehat{\boldsymbol{w}}_j - \widehat{b}_j\boldsymbol{1}_n\right) = 0.$$

Adding the equations on the above displays we find that

$$\boldsymbol{X}\boldsymbol{D}^2\left(\boldsymbol{Y}_\ell + \boldsymbol{Y}_j - \boldsymbol{X}^T(\widehat{\boldsymbol{w}}_j + \widehat{\boldsymbol{w}}_\ell) - (\widehat{b}_j + \widehat{b}_\ell)\boldsymbol{1}_n\right) = 0.$$

Recognize that this coincides with the optimality condition for (I.13). Thus, the proof is complete. ■

Thanks to Lemma I.1, we can use the CGMT to characterize the limiting behavior of $\|\widehat{\boldsymbol{w}}_\ell + \widehat{\boldsymbol{w}}_j\|_{\ell_2}$. Observe that this immediately gives the limit of $\langle \widehat{\boldsymbol{w}}_\ell, \widehat{\boldsymbol{w}}_j \rangle$ since

$$\langle \widehat{\boldsymbol{w}}_\ell, \widehat{\boldsymbol{w}}_j \rangle = \frac{\|\widehat{\boldsymbol{w}}_\ell + \widehat{\boldsymbol{w}}_j\|_{\ell_2}^2 - \|\widehat{\boldsymbol{w}}_\ell\|_{\ell_2}^2 - \|\widehat{\boldsymbol{w}}_j\|_{\ell_2}^2}{2}. \tag{I.14}$$

The analysis of (I.13) is very similar to that of (I.1). In particular we use the following decomposition

$$\boldsymbol{w}_{\ell,j} = \sum_{i=1}^{r} \beta_i \boldsymbol{u}_i + \beta_0 \boldsymbol{w}_{\ell,j}^{\perp},$$

with $\left\|\boldsymbol{w}_{\ell,j}^{\perp}\right\|_{\ell_2} = 1$ and $\boldsymbol{U}^T \boldsymbol{w}_{\ell,j}^{\perp} = \boldsymbol{0}_r$. This allows us to arrive at

$$\min_{\boldsymbol{\beta}} \min_{\boldsymbol{b}_{\ell,j}} \max_{\alpha_0 \geq 0} \max_{\beta \geq 0} \min_{\tau \geq 0}$$
$$\frac{\gamma\beta}{2\tau}\left(\pi_\ell \boldsymbol{\nu}_\ell\left(\frac{\beta}{\tau}\right) + \pi_j \boldsymbol{\nu}_j\left(\frac{\beta}{\tau}\right) + \begin{bmatrix}\boldsymbol{\beta}^T & b_{\ell,j}\end{bmatrix} \boldsymbol{A}\left(\frac{\beta}{\tau}\right)\begin{bmatrix}\boldsymbol{\beta} \\ b_{\ell,j}\end{bmatrix} - 2\left(\pi_\ell \boldsymbol{\nu}_\ell\left(\frac{\beta}{\tau}\right)\boldsymbol{c}_\ell + \pi_j \boldsymbol{\nu}_j\left(\frac{\beta}{\tau}\right)\boldsymbol{c}_j\right)^T \begin{bmatrix}\boldsymbol{\beta} \\ b_{\ell,j}\end{bmatrix}\right)$$
$$+ \frac{\gamma\beta}{2\tau}\sigma^2\left(\boldsymbol{\pi}^T \boldsymbol{\nu}\left(\frac{\beta}{\tau}\right)\right)\beta_0^2 - \beta_0\sigma\beta\sqrt{\gamma} + \frac{\beta\tau}{2}$$

where $\boldsymbol{A}(\eta)$ and $\boldsymbol{c}_\ell$ are as in (I.3) and $\boldsymbol{\nu}(\eta)$ is as in (I.4).

Setting the derivative of the above with respect to α_0 to zero we arrive at

$$\frac{\gamma\beta}{\tau}\sigma^2\left(\boldsymbol{\pi}^T \boldsymbol{\nu}\left(\frac{\beta}{\tau}\right)\right)\beta_0 - \sigma\beta\sqrt{\gamma} = 0 \quad \Rightarrow \quad \beta_0 = \frac{\tau}{\sigma\sqrt{\gamma}\left(\boldsymbol{\pi}^T \boldsymbol{\nu}\left(\frac{\beta}{\tau}\right)\right)}$$

Note that the above objective has the form

$$g\left(\frac{\beta}{\tau}\right) - \beta_0\sigma\beta\sqrt{\gamma} + \frac{\beta\tau}{2}$$

with

$$g(\eta) := \frac{\eta\gamma}{2}\left(\pi_\ell \boldsymbol{\nu}_\ell(\eta) + \pi_j \boldsymbol{\nu}_j(\eta) + \begin{bmatrix}\boldsymbol{\beta}^T & b_{\ell,j}\end{bmatrix} \boldsymbol{A}(\eta)\begin{bmatrix}\boldsymbol{\beta} \\ b_{\ell,j}\end{bmatrix} - 2\left(\pi_\ell \boldsymbol{\nu}_\ell(\eta)\boldsymbol{c}_\ell + \pi_j \boldsymbol{\nu}_j(\eta)\boldsymbol{c}_j\right)^T \begin{bmatrix}\boldsymbol{\beta} \\ b_{\ell,j}\end{bmatrix}\right)$$
$$+ \frac{\gamma\eta}{2}\sigma^2\left(\boldsymbol{\pi}^T \boldsymbol{\nu}(\eta)\right)\beta_0^2.$$

Thus, the derivatives with respect to β and τ to zero we have

$$\frac{1}{\tau}g'\left(\frac{\beta}{\tau}\right) - \beta_0\sigma\sqrt{\gamma} + \frac{\tau}{2} = 0 \quad \Rightarrow \quad g'\left(\frac{\beta}{\tau}\right) - \beta_0\sigma\sqrt{\gamma}\tau + \frac{\tau^2}{2} = 0 \quad \Rightarrow \quad g'\left(\frac{\beta}{\tau}\right) = \tau^2\left(\frac{1}{\boldsymbol{\pi}^T \boldsymbol{\nu}\left(\frac{\beta}{\tau}\right)} - \frac{1}{2}\right)$$

and

$$-\frac{\beta}{\tau^2}g'\left(\frac{\beta}{\tau}\right) + \frac{\beta}{2} = 0 \quad \Rightarrow \quad \tau^2 = 2g'\left(\frac{\beta}{\tau}\right)$$

Combining the latter two we conclude that $\boldsymbol{\pi}^T \boldsymbol{\nu}\left(\frac{\beta}{\tau}\right) = 1$. Thus, $\eta = \frac{\beta}{\tau}$ is the solution to $\boldsymbol{\pi}^T \boldsymbol{\nu}(\eta) = 1$.
To calculate τ and hence β_0 we calculate g' which is equal to

$$g'(\eta) = \frac{\gamma}{2}\left(\pi_\ell \boldsymbol{\nu}_\ell(\eta) + \pi_j \boldsymbol{\nu}_j(\eta) + \begin{bmatrix}\boldsymbol{\beta}^T & b_{\ell,j}\end{bmatrix} \boldsymbol{A}(\eta)\begin{bmatrix}\boldsymbol{\beta} \\ b_{\ell,j}\end{bmatrix} - 2\left(\pi_\ell \boldsymbol{\nu}_\ell(\eta)\boldsymbol{c}_\ell + \pi_j \boldsymbol{\nu}_j(\eta)\boldsymbol{c}_j\right)^T \begin{bmatrix}\boldsymbol{\beta} \\ b_{\ell,j}\end{bmatrix}\right)$$
$$+ \frac{\gamma}{2}\sigma^2\left(\boldsymbol{\pi}^T \boldsymbol{\nu}(\eta)\right)\beta_0^2 + \frac{\gamma\eta}{2}\sigma^2\beta_0^2(\boldsymbol{\pi}^T \boldsymbol{\nu}'(\eta))$$
$$+ \frac{\gamma\eta}{2}\left(\pi_\ell \boldsymbol{\nu}_\ell'(\eta) + \pi_j \boldsymbol{\nu}_j'(\eta) + \begin{bmatrix}\boldsymbol{\beta}^T & b_{\ell,j}\end{bmatrix} \boldsymbol{A}'(\eta)\begin{bmatrix}\boldsymbol{\beta} \\ b_{\ell,j}\end{bmatrix} - 2\left(\pi_\ell \boldsymbol{\nu}_\ell'(\eta)\boldsymbol{c}_\ell + \pi_j \boldsymbol{\nu}_j'(\eta)\boldsymbol{c}_j\right)^T \begin{bmatrix}\boldsymbol{\beta} \\ b_{\ell,j}\end{bmatrix}\right)$$

Where

$$\boldsymbol{\nu}'(\eta) = -\frac{1}{\gamma}\begin{bmatrix}\frac{\omega_1^2}{(\omega_1^2+\eta)^2} \\ \frac{\omega_2^2}{(\omega_2^2+\eta)^2} \\ \cdots \\ \frac{\omega_k^2}{(\omega_k^2+\eta)^2}\end{bmatrix}$$

$$\boldsymbol{A}'(\eta) := \begin{bmatrix}\sigma^2\left(\boldsymbol{\pi}^T \boldsymbol{\nu}'(\eta)\right)\boldsymbol{I} + \boldsymbol{\Sigma}\boldsymbol{V}^T \mathrm{diag}(\boldsymbol{\pi})\,\mathrm{diag}(\boldsymbol{\nu}'(\eta))\,\boldsymbol{V}\boldsymbol{\Sigma} & \boldsymbol{\Sigma}\boldsymbol{V}^T \mathrm{diag}(\boldsymbol{\nu}'(\eta))\,\boldsymbol{\pi} \\ \boldsymbol{\pi}^T \mathrm{diag}(\boldsymbol{\nu}'(\eta))\,\boldsymbol{V}\boldsymbol{\Sigma} & \boldsymbol{\pi}^T \boldsymbol{\nu}'(\eta)\end{bmatrix}$$

$$\boldsymbol{c}_\ell := \begin{bmatrix}\boldsymbol{\Sigma}\boldsymbol{V}^T \boldsymbol{e}_\ell \\ 1\end{bmatrix}$$

Now note that at the optimal point we have

$$g'(\eta) = \frac{\tau^2}{2} = \frac{\gamma}{2}\sigma^2\left(\boldsymbol{\pi}^T\boldsymbol{\nu}(\eta)\right)\beta_0^2$$

Thus from the above we can conclude that

$$\begin{aligned}\beta_0^2 =& -\frac{1}{\eta\sigma^2(\boldsymbol{\pi}^T\boldsymbol{\nu}'(\eta))}\left(\pi_\ell\boldsymbol{\nu}_\ell(\eta)+\pi_j\boldsymbol{\nu}_j(\eta)+\begin{bmatrix}\boldsymbol{\beta}^T & b_{\ell,j}\end{bmatrix}\boldsymbol{A}(\eta)\begin{bmatrix}\boldsymbol{\beta}\\ b_{\ell,j}\end{bmatrix}-2\left(\pi_\ell\boldsymbol{\nu}_\ell(\eta)\boldsymbol{c}_\ell+\pi_j\boldsymbol{\nu}_j(\eta)\boldsymbol{c}_j\right)^T\begin{bmatrix}\boldsymbol{\beta}\\ b_{\ell,j}\end{bmatrix}\right)\\ &-\frac{1}{\sigma^2(\boldsymbol{\pi}^T\boldsymbol{\nu}'(\eta))}\left(\pi_\ell\boldsymbol{\nu}_\ell'(\eta)+\pi_j\boldsymbol{\nu}_j'(\eta)+\begin{bmatrix}\boldsymbol{\beta}^T & b_{\ell,j}\end{bmatrix}\boldsymbol{A}'(\eta)\begin{bmatrix}\boldsymbol{\beta}\\ b_{\ell,j}\end{bmatrix}-2\left(\pi_\ell\boldsymbol{\nu}_\ell'(\eta)\boldsymbol{c}_\ell+\pi_j\boldsymbol{\nu}_j'(\eta)\boldsymbol{c}_j\right)^T\begin{bmatrix}\boldsymbol{\beta}\\ b_{\ell,j}\end{bmatrix}\right)\end{aligned}$$

Thus the AO optimization problem reduces to

$$\min_{\boldsymbol{b}_{\ell,j}}\ \min_{\boldsymbol{\beta}}\ \frac{\eta\gamma}{2}\left(\pi_\ell\boldsymbol{\nu}_\ell+\pi_j\boldsymbol{\nu}_j+\begin{bmatrix}\boldsymbol{\beta}^T & b_{\ell,j}\end{bmatrix}\boldsymbol{A}\begin{bmatrix}\boldsymbol{\beta}\\ b_{\ell,j}\end{bmatrix}-2\left(\pi_\ell\boldsymbol{\nu}_\ell\boldsymbol{c}_\ell+\pi_j\boldsymbol{\nu}_j\boldsymbol{c}_j\right)^T\begin{bmatrix}\boldsymbol{\beta}\\ b_{\ell,j}\end{bmatrix}\right)$$

where η is the solution to

$$\sum_{\ell=1}^{k}\frac{\pi_\ell\omega_\ell^2}{\omega_\ell^2+\eta} = \gamma$$

Thus, similar to (I.10) we can compute the minimizer of the deterministic

$$\begin{bmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{b}_{\ell,j}\end{bmatrix} = \begin{bmatrix}-\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\pi_\ell\nu_\ell(\boldsymbol{\pi}\odot\boldsymbol{\nu}-\boldsymbol{e}_\ell)+\pi_j\nu_j(\boldsymbol{\pi}\odot\boldsymbol{\nu}-\boldsymbol{e}_j)\right)\\ \pi_\ell\nu_\ell+\pi_j\nu_j+(\boldsymbol{\pi}\odot\boldsymbol{\nu})^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\pi_\ell\nu_\ell(\boldsymbol{\pi}\odot\boldsymbol{\nu}-\boldsymbol{e}_\ell)+\pi_j\nu_j(\boldsymbol{\pi}\odot\boldsymbol{\nu}-\boldsymbol{e}_j)\right)\end{bmatrix} \tag{I.15}$$

and

$$\boldsymbol{\nu}'(\eta) = -\frac{1}{\gamma}\begin{bmatrix}\frac{\omega_1^2}{(\omega_1^2+\eta)^2}\\ \frac{\omega_2^2}{(\omega_2^2+\eta)^2}\\ \cdots\\ \frac{\omega_k^2}{(\omega_k^2+\eta)^2}\end{bmatrix}$$

$$\boldsymbol{A}'(\eta) := \begin{bmatrix}\sigma^2\left(\boldsymbol{\pi}^T\boldsymbol{\nu}'(\eta)\right)\boldsymbol{I}+\boldsymbol{\Sigma}\boldsymbol{V}^T\mathrm{diag}(\boldsymbol{\pi})\,\mathrm{diag}(\boldsymbol{\nu}'(\eta))\,\boldsymbol{V}\boldsymbol{\Sigma} & \boldsymbol{\Sigma}\boldsymbol{V}^T\mathrm{diag}(\boldsymbol{\nu}'(\eta))\,\boldsymbol{\pi}\\ \boldsymbol{\pi}^T\mathrm{diag}(\boldsymbol{\nu}'(\eta))\,\boldsymbol{V}\boldsymbol{\Sigma} & \boldsymbol{\pi}^T\boldsymbol{\nu}'(\eta)\end{bmatrix}$$

To continue note that

$$\begin{aligned}&\pi_\ell\boldsymbol{\nu}_\ell'(\eta)+\pi_j\boldsymbol{\nu}_j'(\eta)+\begin{bmatrix}\boldsymbol{\beta}^T & b_{\ell,j}\end{bmatrix}\boldsymbol{A}'(\eta)\begin{bmatrix}\boldsymbol{\beta}\\ b_{\ell,j}\end{bmatrix}-2\left(\pi_\ell\boldsymbol{\nu}_\ell'(\eta)\boldsymbol{c}_\ell+\pi_j\boldsymbol{\nu}_j'(\eta)\boldsymbol{c}_j\right)^T\begin{bmatrix}\boldsymbol{\beta}\\ b_{\ell,j}\end{bmatrix}\\ &\quad=\pi_\ell\boldsymbol{\nu}_\ell'(\eta)+\pi_j\boldsymbol{\nu}_j'(\eta)+\left(\pi_\ell\boldsymbol{\nu}_\ell'(\eta)\boldsymbol{c}_\ell+\pi_j\boldsymbol{\nu}_j'(\eta)\boldsymbol{c}_j\right)^T\boldsymbol{A}^{-1}\boldsymbol{A}'(\eta)\boldsymbol{A}^{-1}\left(\pi_\ell\boldsymbol{\nu}_\ell'(\eta)\boldsymbol{c}_\ell+\pi_j\boldsymbol{\nu}_j'(\eta)\boldsymbol{c}_j\right)\\ &\qquad-2\left(\pi_\ell\boldsymbol{\nu}_\ell'(\eta)\boldsymbol{c}_\ell+\pi_j\boldsymbol{\nu}_j'(\eta)\boldsymbol{c}_j\right)^T\boldsymbol{A}^{-1}\left(\pi_\ell\boldsymbol{\nu}_\ell'(\eta)\boldsymbol{c}_\ell+\pi_j\boldsymbol{\nu}_j'(\eta)\boldsymbol{c}_j\right)\end{aligned}$$

Thus,

$$\begin{aligned}\beta_0^2 =& -\frac{1}{\eta\sigma^2(\boldsymbol{\pi}^T\boldsymbol{\nu}'(\eta))}\left(\pi_\ell\boldsymbol{\nu}_\ell(\eta)+\pi_j\boldsymbol{\nu}_j(\eta)+\begin{bmatrix}\boldsymbol{\beta}^T & b_{\ell,j}\end{bmatrix}\boldsymbol{A}(\eta)\begin{bmatrix}\boldsymbol{\beta}\\ b_{\ell,j}\end{bmatrix}-2\left(\pi_\ell\boldsymbol{\nu}_\ell(\eta)\boldsymbol{c}_\ell+\pi_j\boldsymbol{\nu}_j(\eta)\boldsymbol{c}_j\right)^T\begin{bmatrix}\boldsymbol{\beta}\\ b_{\ell,j}\end{bmatrix}\right)\\ &-\frac{1}{\sigma^2(\boldsymbol{\pi}^T\boldsymbol{\nu}'(\eta))}\left(\pi_\ell\boldsymbol{\nu}_\ell'(\eta)+\pi_j\boldsymbol{\nu}_j'(\eta)+\begin{bmatrix}\boldsymbol{\beta}^T & b_{\ell,j}\end{bmatrix}\boldsymbol{A}'(\eta)\begin{bmatrix}\boldsymbol{\beta}\\ b_{\ell,j}\end{bmatrix}-2\left(\pi_\ell\boldsymbol{\nu}_\ell'(\eta)\boldsymbol{c}_\ell+\pi_j\boldsymbol{\nu}_j'(\eta)\boldsymbol{c}_j\right)^T\begin{bmatrix}\boldsymbol{\beta}\\ b_{\ell,j}\end{bmatrix}\right)\\ =&\frac{\zeta}{\sigma^2}\Bigg(\pi_\ell\nu_\ell+\pi_j\nu_j-\left(\pi_\ell\nu_\ell+\pi_j\nu_j\right)^2\\ &\qquad-\left(\pi_\ell\nu_\ell(\boldsymbol{\pi}\odot\boldsymbol{\nu}-\boldsymbol{e}_\ell)+\pi_j\nu_j(\boldsymbol{\pi}\odot\boldsymbol{\nu}-\boldsymbol{e}_j)\right)^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\pi_\ell\nu_\ell(\boldsymbol{\pi}\odot\boldsymbol{\nu}-\boldsymbol{e}_\ell)+\pi_j\nu_j(\boldsymbol{\pi}\odot\boldsymbol{\nu}-\boldsymbol{e}_j)\right)\Bigg)\\ &+\frac{\zeta\eta}{\sigma^2}\Bigg(\pi_\ell\boldsymbol{\nu}_\ell'(\eta)+\pi_j\boldsymbol{\nu}_j'(\eta)+\left(\pi_\ell\boldsymbol{\nu}_\ell'(\eta)\boldsymbol{c}_\ell+\pi_j\boldsymbol{\nu}_j'(\eta)\boldsymbol{c}_j\right)^T\boldsymbol{A}^{-1}\boldsymbol{A}'(\eta)\boldsymbol{A}^{-1}\left(\pi_\ell\boldsymbol{\nu}_\ell'(\eta)\boldsymbol{c}_\ell+\pi_j\boldsymbol{\nu}_j'(\eta)\boldsymbol{c}_j\right)\\ &\qquad-2\left(\pi_\ell\boldsymbol{\nu}_\ell'(\eta)\boldsymbol{c}_\ell+\pi_j\boldsymbol{\nu}_j'(\eta)\boldsymbol{c}_j\right)^T\boldsymbol{A}^{-1}\left(\pi_\ell\boldsymbol{\nu}_\ell'(\eta)\boldsymbol{c}_\ell+\pi_j\boldsymbol{\nu}_j'(\eta)\boldsymbol{c}_j\right)\Bigg)\end{aligned}$$

where $\zeta := \frac{\gamma}{\eta\left(\Sigma_{\ell=1}^{k} \frac{\pi_\ell \omega_\ell^2}{(\omega_\ell^2+\eta)^2}\right)}$. From the CGMT, we have that $\|\widehat{\boldsymbol{w}}_\ell + \widehat{\boldsymbol{w}}_j\|_{\ell_2}^2 \xrightarrow{P} \widehat{\beta}_0^2 + \|\boldsymbol{\beta}\|_{\ell_2}^2$. Combining this with the calculations above, we conclude that

$$
\begin{aligned}
&\|\widehat{\boldsymbol{w}}_\ell + \widehat{\boldsymbol{w}}_j\|_{\ell_2}^2 \\
&\xrightarrow{P} \frac{\zeta}{\sigma^2}\left(\pi_\ell\nu_\ell + \pi_j\nu_j\right)\left(1 - \pi_\ell\nu_\ell - \pi_j\nu_j\right) \\
&+ \left(\pi_\ell\nu_\ell\left(\boldsymbol{\pi}\odot\boldsymbol{\nu} - \boldsymbol{e}_\ell\right) + \pi_j\nu_j\left(\boldsymbol{\pi}\odot\boldsymbol{\nu} - \boldsymbol{e}_j\right)\right)^T \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\left(\boldsymbol{\Delta}^{-1} - \frac{\eta}{\sigma^2}\boldsymbol{I}_r\right)\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\pi_\ell\nu_\ell\left(\boldsymbol{\pi}\odot\boldsymbol{\nu} - \boldsymbol{e}_\ell\right) + \pi_j\nu_j\left(\boldsymbol{\pi}\odot\boldsymbol{\nu} - \boldsymbol{e}_j\right)\right) \\
&+ \frac{\zeta\eta}{\sigma^2}\Bigg(\pi_\ell\boldsymbol{\nu}_\ell'(\eta) + \pi_j\boldsymbol{\nu}_j'(\eta) + \left(\pi_\ell\boldsymbol{\nu}_\ell'(\eta)\,\boldsymbol{c}_\ell + \pi_j\boldsymbol{\nu}_j'(\eta)\,\boldsymbol{c}_j\right)^T \boldsymbol{A}^{-1}\boldsymbol{A}'(\eta)\boldsymbol{A}^{-1}\left(\pi_\ell\boldsymbol{\nu}_\ell'(\eta)\,\boldsymbol{c}_\ell + \pi_j\boldsymbol{\nu}_j'(\eta)\,\boldsymbol{c}_j\right) \\
&\qquad - 2\left(\pi_\ell\boldsymbol{\nu}_\ell'(\eta)\,\boldsymbol{c}_\ell + \pi_j\boldsymbol{\nu}_j'(\eta)\,\boldsymbol{c}_j\right)^T \boldsymbol{A}^{-1}\left(\pi_\ell\boldsymbol{\nu}_\ell'(\eta)\,\boldsymbol{c}_\ell + \pi_j\boldsymbol{\nu}_j'(\eta)\,\boldsymbol{c}_j\right)\Bigg)
\end{aligned}
\tag{I.16}
$$

Finally, using (I.16) and (I.12) in (I.14) it follows that

$$
\begin{aligned}
&\langle \boldsymbol{w}_\ell, \boldsymbol{w}_j\rangle \xrightarrow{P} \\
&\pi_\ell\nu_\ell\pi_j\nu_j\left(-\frac{\zeta}{\sigma^2} + \left(\boldsymbol{\pi}\odot\boldsymbol{\nu} - \boldsymbol{e}_\ell\right)^T \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\left(\boldsymbol{\Delta}^{-1} - \frac{\zeta}{\sigma^2}\boldsymbol{I}_r\right)\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\boldsymbol{\pi}\odot\boldsymbol{\nu} - \boldsymbol{e}_j\right)\right) \\
&+ \frac{\zeta\eta}{\sigma^2}\pi_\ell\nu_\ell'\pi_j\nu_j'\left(\boldsymbol{c}_j^T\boldsymbol{A}^{-1}\boldsymbol{A}'\boldsymbol{A}^{-1}\boldsymbol{c}_\ell - 2\boldsymbol{c}_j^T\boldsymbol{A}^{-1}\boldsymbol{c}_\ell\right) \\
&= \pi_\ell\nu_\ell\pi_j\nu_j\left(-\frac{\zeta}{\sigma^2} + \left(\boldsymbol{\pi}\odot\boldsymbol{\nu} - \boldsymbol{e}_\ell\right)^T \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\left(\boldsymbol{\Delta}^{-1} - \frac{\zeta}{\sigma^2}\boldsymbol{I}_r\right)\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\boldsymbol{\pi}\odot\boldsymbol{\nu} - \boldsymbol{e}_j\right)\right) \\
&+ \frac{\zeta\eta}{\sigma^2}\pi_\ell\nu_\ell'\pi_j\nu_j'\left(\boldsymbol{e}_j^T\begin{bmatrix}\boldsymbol{\Sigma}\boldsymbol{V}^T \\ \boldsymbol{1}^T\end{bmatrix}^T \boldsymbol{A}^{-1}\boldsymbol{A}'\boldsymbol{A}^{-1}\begin{bmatrix}\boldsymbol{\Sigma}\boldsymbol{V}^T \\ \boldsymbol{1}^T\end{bmatrix}\boldsymbol{e}_\ell - 2\boldsymbol{e}_j^T\begin{bmatrix}\boldsymbol{\Sigma}\boldsymbol{V}^T \\ \boldsymbol{1}^T\end{bmatrix}^T \boldsymbol{A}^{-1}\begin{bmatrix}\boldsymbol{\Sigma}\boldsymbol{V}^T \\ \boldsymbol{1}^T\end{bmatrix}\boldsymbol{e}_\ell\right)
\end{aligned}
\tag{I.17}
$$

Putting everything together we arrive at

$$
\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}} \xrightarrow{P} \frac{\zeta}{\sigma^2}\boldsymbol{P} + \boldsymbol{P}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\left(\boldsymbol{\Delta}^{-1} - \frac{\zeta}{\sigma^2}\boldsymbol{I}_r\right)\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{P} + \frac{\zeta\eta}{\sigma^2}\boldsymbol{M}
$$

where

$$
\boldsymbol{Q} := \operatorname{diag}\left(\boldsymbol{\pi}\odot\boldsymbol{\nu}'\right) + \operatorname{diag}\left(\boldsymbol{\pi}\odot\boldsymbol{\nu}'\right)\begin{bmatrix}\boldsymbol{\Sigma}\boldsymbol{V}^T \\ \boldsymbol{1}^T\end{bmatrix}^T\left(\boldsymbol{A}^{-1}\boldsymbol{A}'\boldsymbol{A}^{-1} - 2\boldsymbol{A}^{-1}\right)\begin{bmatrix}\boldsymbol{\Sigma}\boldsymbol{V}^T \\ \boldsymbol{1}^T\end{bmatrix}\operatorname{diag}\left(\boldsymbol{\pi}\odot\boldsymbol{\nu}'\right)
$$

and as mentioned earlier

$$
\begin{aligned}
\boldsymbol{A}' := \boldsymbol{A}'(\eta) &:= \begin{bmatrix}\sigma^2\left(\boldsymbol{\pi}^T\boldsymbol{\nu}'(\eta)\right)\boldsymbol{I} + \boldsymbol{\Sigma}\boldsymbol{V}^T\operatorname{diag}\left(\boldsymbol{\pi}\right)\operatorname{diag}\left(\boldsymbol{\nu}'(\eta)\right)\boldsymbol{V}\boldsymbol{\Sigma} & \boldsymbol{\Sigma}\boldsymbol{V}^T\operatorname{diag}\left(\boldsymbol{\nu}'(\eta)\right)\boldsymbol{\pi} \\ \boldsymbol{\pi}^T\operatorname{diag}\left(\boldsymbol{\nu}'(\eta)\right)\boldsymbol{V}\boldsymbol{\Sigma} & \boldsymbol{\pi}^T\boldsymbol{\nu}'(\eta)\end{bmatrix} \\
\boldsymbol{A} := \boldsymbol{A}(\eta) &:= \begin{bmatrix}\sigma^2\left(\boldsymbol{\pi}^T\boldsymbol{\nu}(\eta)\right)\boldsymbol{I} + \boldsymbol{\Sigma}\boldsymbol{V}^T\operatorname{diag}\left(\boldsymbol{\pi}\right)\operatorname{diag}\left(\boldsymbol{\nu}(\eta)\right)\boldsymbol{V}\boldsymbol{\Sigma} & \boldsymbol{\Sigma}\boldsymbol{V}^T\operatorname{diag}\left(\boldsymbol{\nu}(\eta)\right)\boldsymbol{\pi} \\ \boldsymbol{\pi}^T\operatorname{diag}\left(\boldsymbol{\nu}(\eta)\right)\boldsymbol{V}\boldsymbol{\Sigma} & \boldsymbol{\pi}^T\boldsymbol{\nu}(\eta)\end{bmatrix}
\end{aligned}
$$

Let us end by simplifying $\boldsymbol{M}$ to this aim define $\widetilde{\boldsymbol{\pi}}' = \boldsymbol{\pi}\odot\boldsymbol{\nu}'$ note that

$$
\begin{aligned}
\boldsymbol{A}^{-1}\begin{bmatrix}\boldsymbol{\Sigma}\boldsymbol{V}^T \\ \boldsymbol{1}^T\end{bmatrix}\operatorname{diag}\left(\boldsymbol{\pi}\odot\boldsymbol{\nu}'\right) &= \begin{bmatrix}\boldsymbol{I} & \boldsymbol{0} \\ -\widetilde{\boldsymbol{\pi}}^T\boldsymbol{V}\boldsymbol{\Sigma} & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{\Delta}^{-1} & \boldsymbol{0} \\ \boldsymbol{0}^T & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{I} & -\boldsymbol{\Sigma}\boldsymbol{V}^T\widetilde{\boldsymbol{\pi}} \\ \boldsymbol{0}^T & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}\boldsymbol{V}^T \\ \boldsymbol{1}^T\end{bmatrix}\operatorname{diag}\left(\boldsymbol{\pi}\odot\boldsymbol{\nu}'\right) \\
&= \begin{bmatrix}\boldsymbol{I} & \boldsymbol{0} \\ -\widetilde{\boldsymbol{\pi}}^T\boldsymbol{V}\boldsymbol{\Sigma} & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{\Delta}^{-1} & \boldsymbol{0} \\ \boldsymbol{0}^T & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\boldsymbol{I} - \widetilde{\boldsymbol{\pi}}\boldsymbol{1}^T\right) \\ \boldsymbol{1}^T\end{bmatrix}\operatorname{diag}\left(\boldsymbol{\pi}\odot\boldsymbol{\nu}'\right) \\
&= \begin{bmatrix}\boldsymbol{I} & \boldsymbol{0} \\ -\widetilde{\boldsymbol{\pi}}^T\boldsymbol{V}\boldsymbol{\Sigma} & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\boldsymbol{I} - \widetilde{\boldsymbol{\pi}}\boldsymbol{1}^T\right) \\ \boldsymbol{1}^T\end{bmatrix}\operatorname{diag}\left(\boldsymbol{\pi}\odot\boldsymbol{\nu}'\right) \\
&= \begin{bmatrix}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\boldsymbol{I} - \widetilde{\boldsymbol{\pi}}\boldsymbol{1}^T\right)\operatorname{diag}\left(\boldsymbol{\pi}\odot\boldsymbol{\nu}'\right) \\ -\widetilde{\boldsymbol{\pi}}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\boldsymbol{I} - \widetilde{\boldsymbol{\pi}}\boldsymbol{1}^T\right) + \boldsymbol{1}^T\end{bmatrix}\operatorname{diag}\left(\boldsymbol{\pi}\odot\boldsymbol{\nu}'\right) \\
&= \begin{bmatrix}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\operatorname{diag}(\widetilde{\boldsymbol{\pi}}') - \widetilde{\boldsymbol{\pi}}\widetilde{\boldsymbol{\pi}}'^T\right) \\ -\widetilde{\boldsymbol{\pi}}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\operatorname{diag}(\widetilde{\boldsymbol{\pi}}') - \widetilde{\boldsymbol{\pi}}\widetilde{\boldsymbol{\pi}}'^T\right) + \widetilde{\boldsymbol{\pi}}'^T\end{bmatrix}
\end{aligned}
$$

Using the above we arrive at

$$\boldsymbol{Q} = \mathrm{diag}(\widetilde{\boldsymbol{\pi}}') + \begin{bmatrix} \boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\mathrm{diag}(\widetilde{\boldsymbol{\pi}}') - \widetilde{\boldsymbol{\pi}}\widetilde{\boldsymbol{\pi}}'^T\right) \\ -\widetilde{\boldsymbol{\pi}}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\mathrm{diag}(\widetilde{\boldsymbol{\pi}}') - \widetilde{\boldsymbol{\pi}}\widetilde{\boldsymbol{\pi}}'^T\right) + \widetilde{\boldsymbol{\pi}}'^T \end{bmatrix}^T (\boldsymbol{A}' - 2\boldsymbol{A}) \begin{bmatrix} \boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\mathrm{diag}(\widetilde{\boldsymbol{\pi}}') - \widetilde{\boldsymbol{\pi}}\widetilde{\boldsymbol{\pi}}'^T\right) \\ -\widetilde{\boldsymbol{\pi}}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\mathrm{diag}(\widetilde{\boldsymbol{\pi}}') - \widetilde{\boldsymbol{\pi}}\widetilde{\boldsymbol{\pi}}'^T\right) + \widetilde{\boldsymbol{\pi}}'^T \end{bmatrix}$$

where

$$\boldsymbol{A}' = \begin{bmatrix} \sigma^2\left(\widetilde{\boldsymbol{\pi}}'^T\mathbf{1}\right)\boldsymbol{I} + \boldsymbol{\Sigma}\boldsymbol{V}^T\mathrm{diag}\left(\widetilde{\boldsymbol{\pi}}'\right)\boldsymbol{V}\boldsymbol{\Sigma} & \boldsymbol{\Sigma}\boldsymbol{V}^T\mathrm{diag}\left(\widetilde{\boldsymbol{\pi}}'\right) \\ \widetilde{\boldsymbol{\pi}}'^T\boldsymbol{V}\boldsymbol{\Sigma} & \widetilde{\boldsymbol{\pi}}'^T\mathbf{1} \end{bmatrix}$$

$$\boldsymbol{A} = \begin{bmatrix} \sigma^2\boldsymbol{I} + \boldsymbol{\Sigma}\boldsymbol{V}^T\mathrm{diag}\left(\widetilde{\boldsymbol{\pi}}\right)\boldsymbol{V}\boldsymbol{\Sigma} & \boldsymbol{\Sigma}\boldsymbol{V}^T\mathrm{diag}\left(\widetilde{\boldsymbol{\pi}}\right) \\ \widetilde{\boldsymbol{\pi}}^T\boldsymbol{V}\boldsymbol{\Sigma} & 1 \end{bmatrix}$$

Using the above the cross correlation matrix is given by

$$\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}} \xrightarrow{P} \frac{\zeta}{\sigma^2}\boldsymbol{P} + \boldsymbol{P}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\Big(\boldsymbol{\Delta}^{-1} - \frac{\zeta}{\sigma^2}\boldsymbol{I}_r\Big)\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{P} + \frac{\zeta\eta}{\sigma^2}\boldsymbol{Q}$$

J Weighted LS for MLM (Proof of Theorem 4.4)

Let $\boldsymbol{D} := \boldsymbol{D}^{(n)} := \mathrm{diag}(d_1, \ldots, d_n)$ be a diagonal matrix with non-zero diagonal entries. In particular, assume that the entries of $\boldsymbol{D}$ are distributed $D_i \overset{iid}{\sim} D$ where the random variable D may depend on the entries of the matrix of response variables $\boldsymbol{Y}$. Here, we focus on the following setting:

$$\boldsymbol{D} = \sum_{j\in[k]} \mathrm{diag}(\omega_j \boldsymbol{Y}_\ell), \quad \omega_j \geq 0,\ j \in [k]. \tag{J.1}$$

Specifically, for (J.1), we have $D_i \overset{iid}{\sim} D$ with $D = \omega_\ell Y_\ell + \sum_{i\neq\ell\in[k]} \omega_i Y_i$, where for all $c \in [k]$:

$$\mathbb{P}\left([Y_1, Y_2, \ldots, Y_k]^T = \boldsymbol{e}_c\right) = V_c = \frac{e^{\boldsymbol{e}_c^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}}{\sum_{\ell'=1}^k e^{\boldsymbol{e}_{\ell'}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}}. \tag{J.2}$$

With these, we consider the weighted least-squares (WLS) solution for all $\ell \in [k]$:

$$(\widehat{\boldsymbol{w}}_\ell, \widehat{b}) = \arg\min_{\boldsymbol{w},b} \mathcal{L}_{PO}(\boldsymbol{w}, b) := \frac{1}{2n}\left\|\boldsymbol{D}\left(\boldsymbol{X}^T\boldsymbol{w} + b\mathbf{1}_n - \boldsymbol{Y}_\ell\right)\right\|_{\ell_2}^2,$$

where $\boldsymbol{D}$ is as in (J.1). In fact, it is convenient to rewrite the above as follows:

$$(\widehat{\boldsymbol{w}}_\ell, \widehat{b}) = \arg\min_{\boldsymbol{w},b,\boldsymbol{u}} \max_{\boldsymbol{s}} \frac{1}{n}\left(\boldsymbol{s}^T\boldsymbol{D}\boldsymbol{X}^T\boldsymbol{w} + b\boldsymbol{s}^T\boldsymbol{D}\mathbf{1}_n\boldsymbol{s}^T\boldsymbol{D}\boldsymbol{Y}_\ell - \boldsymbol{s}^T\boldsymbol{u} + \frac{\|\boldsymbol{u}\|_{\ell_2}^2}{2}\right). \tag{J.3}$$

Identifying the AO. The PO in (J.3) is very similar to (H.1). In particular, following step by step the same decomposition trick as in Section G.1.1, it can be shown that the AO corresponding to (J.3) becomes (cf. (H.6))

$$\min_{\boldsymbol{w}_\ell,\boldsymbol{b}_\ell,\boldsymbol{u}} \max_{\boldsymbol{s}} \frac{1}{n}\left(\|\boldsymbol{P}^\perp\boldsymbol{w}_\ell\|_{\ell_2}\boldsymbol{g}^T\boldsymbol{D}\boldsymbol{s} + \|\boldsymbol{D}\boldsymbol{s}\|_{\ell_2}\boldsymbol{h}^T\boldsymbol{P}^\perp\boldsymbol{w}_\ell + \boldsymbol{s}^T\boldsymbol{D}\widetilde{\boldsymbol{G}}^T\boldsymbol{U}^T\boldsymbol{w}_\ell + b_\ell\boldsymbol{s}^T\boldsymbol{D}\mathbf{1}_n - \boldsymbol{s}^T\boldsymbol{D}\boldsymbol{Y}_\ell - \boldsymbol{u}^T\boldsymbol{s} + \frac{\|\boldsymbol{u}\|_{\ell_2}^2}{2}\right),$$

where we use the same notation as in Section G.1.1 for $\boldsymbol{P}^\perp, \boldsymbol{U}, \widetilde{\boldsymbol{G}}, \boldsymbol{g}$ and $\boldsymbol{h}$. Recall also the relation of $\boldsymbol{Y}_\ell$ to $\widetilde{\boldsymbol{G}}$ in (H.5).

Scalarization of the AO. We start the process of simplifying the AO by setting $\beta := \|\boldsymbol{D}\boldsymbol{s}\|_{\ell_2}/\sqrt{n}$ and optimizing over the direction of $\boldsymbol{D}\boldsymbol{s}$ to equivalently write the AO as

$$\min_{\boldsymbol{w}_\ell,\boldsymbol{b}_\ell,\boldsymbol{u}} \max_{\beta\geq 0} \frac{1}{\sqrt{n}}\left(\beta\left\|\|\boldsymbol{P}^\perp\boldsymbol{w}_\ell\|_{\ell_2}\boldsymbol{g} + \widetilde{\boldsymbol{G}}^T\boldsymbol{U}^T\boldsymbol{w}_\ell + b_\ell\mathbf{1}_n - \boldsymbol{Y}_\ell - \boldsymbol{D}^{-1}\boldsymbol{u}\right\|_{\ell_2} + \beta\boldsymbol{h}^T\boldsymbol{P}^\perp\boldsymbol{w}_\ell\right) + \frac{\|\boldsymbol{u}\|_{\ell_2}^2}{2n}, \tag{J.4}$$

Next, focus on the minimization over $\boldsymbol{w}_\ell$. Let us denote

$$\boldsymbol{a} \coloneqq \boldsymbol{U}^T \boldsymbol{w}_\ell \quad \text{and} \quad \alpha_0 = \left\| \boldsymbol{P}^\perp \boldsymbol{w}_\ell \right\|_{\ell_2}.$$

Notice that $\boldsymbol{a} \perp \boldsymbol{P}^\perp \boldsymbol{w}_\ell$ and thus the orthogonal decomposition $\boldsymbol{w}_\ell = \boldsymbol{U}\boldsymbol{a} + \boldsymbol{P}^\perp \boldsymbol{w}_\ell$. With this observation, note that the optimal direction of $\boldsymbol{P}^T \boldsymbol{w}_\ell$ in (J.4) aligns with $\boldsymbol{P}^T \boldsymbol{h}$ for all values of β. Therefore, (J.4) reduces to

$$\min_{\boldsymbol{a}, \alpha_0 \geq 0, \boldsymbol{b}_\ell, \boldsymbol{u}} \max_{\beta \geq 0} \frac{1}{\sqrt{n}} \left(\beta \left\| \alpha_0 \boldsymbol{g} + \widetilde{\boldsymbol{G}}^T \boldsymbol{a} + b_\ell \mathbf{1}_n - \boldsymbol{Y}_\ell - \boldsymbol{D}^{-1} \boldsymbol{u} \right\|_{\ell_2} - \beta \alpha_0 \left\| \boldsymbol{P}^\perp \boldsymbol{h} \right\|_{\ell_2} \right) + \frac{\|\boldsymbol{u}\|_{\ell_2}^2}{2n}, \quad \text{(J.5)}$$

Continuing let us denote $\boldsymbol{t} \coloneqq \alpha_0 \boldsymbol{g} + \widetilde{\boldsymbol{G}}^T \boldsymbol{a} + b_\ell \mathbf{1}_n - \boldsymbol{Y}_\ell$ for convenience and rewrite $\left\| \boldsymbol{t} - \boldsymbol{D}^{-1} \boldsymbol{u} \right\|_{\ell_2}$ as follows

$$\frac{\left\| \boldsymbol{t} - \boldsymbol{D}^{-1} \boldsymbol{u} \right\|_{\ell_2}}{\sqrt{n}} = \min_{\tau > 0} \frac{\tau}{2} + \frac{\left\| \boldsymbol{t} - \boldsymbol{D}^{-1} \boldsymbol{u} \right\|_{\ell_2}^2}{2\tau n}.$$

Note that the resulting minimization is convex in $\boldsymbol{u}$ and concave in β. Also, by considering the bounded AO (such that β is bounded; see [15, Sec. A]), we can flip the order of min-max and optimize over $\boldsymbol{u}$ first. In particular, $\boldsymbol{u}$ minimizes the following strictly convex quadratic

$$\min_{\boldsymbol{u}} \left\{ \frac{1}{n} \left(\frac{\beta}{2\tau} \left\| \boldsymbol{D}^{-1} \boldsymbol{u} \right\|_{\ell_2} + \frac{1}{2} \|\boldsymbol{u}\|_{\ell_2}^2 - \frac{\beta}{\tau} \boldsymbol{t}^T \boldsymbol{D}^{-1} \boldsymbol{u} \right) = \frac{1}{2n} \boldsymbol{u}^T \left(\frac{\beta}{\tau} \boldsymbol{D}^{-2} + \boldsymbol{I}_n \right) \boldsymbol{u} - \frac{\beta}{\tau n} \boldsymbol{t}^T \boldsymbol{D}^{-1} \boldsymbol{u} \right\}.$$

In particular,

$$\boldsymbol{u} = \frac{\beta}{\tau} \left(\frac{\beta}{\tau} \boldsymbol{D}^{-2} + \boldsymbol{I} \right)^{-1} \boldsymbol{D}^{-1} \boldsymbol{t} = \left(\boldsymbol{D}^{-1} + \frac{\tau}{\beta} \boldsymbol{D} \right)^{-1} \left(\alpha_0 \boldsymbol{g} + \widetilde{\boldsymbol{G}}^T \boldsymbol{a} + b_\ell \mathbf{1}_n - \boldsymbol{Y}_\ell \right)$$

Putting things together, the new objective function of (J.5) becomes

$$\min_{\boldsymbol{a}, \alpha_0 \geq 0, b_\ell, \tau > 0} \max_{\beta \geq 0} \mathcal{R}(\boldsymbol{a}, \alpha_0, b_\ell, \tau, \beta) \quad \text{(J.6)}$$

$$\text{where } \mathcal{R}(\boldsymbol{a}, \alpha_0, b_\ell, \tau, \beta) \coloneqq \frac{\beta\tau}{2n} + \frac{\beta}{2\tau n} \|\boldsymbol{t}\|_{\ell_2}^2 - \frac{\beta}{2\tau n} \boldsymbol{t}^T \left(\boldsymbol{I} + \frac{\tau}{\beta} \boldsymbol{D}^2 \right)^{-1} \boldsymbol{t} - \frac{\beta \alpha_0}{\sqrt{n}} \left\| \boldsymbol{P}^\perp \boldsymbol{h} \right\|_{\ell_2}.$$

Convergence of the AO After having simplified the AO into an optimization problem over $r + 4$ variables, we are ready to study its asymptotic behavior. First, we argue on point-wise convergence of $\mathcal{R}$ in (J.6). Fix $\boldsymbol{a}, \alpha_0, \boldsymbol{b}_\ell, \tau$ and β. From the WLLN, $\frac{1}{\sqrt{n}} \left\| \boldsymbol{P}^\perp \boldsymbol{h} \right\|_{\ell_2} \xrightarrow{P} \sqrt{\gamma}$ and as in (H.9)

$$\frac{1}{n} \|\boldsymbol{t}\|_{\ell_2}^2 = \frac{1}{n} \sum_{i=1}^n \left(\alpha_0 \boldsymbol{g}_i + \boldsymbol{a}^T \tilde{\boldsymbol{g}}_i + b_\ell - [\boldsymbol{Y}_\ell]_i \right)^2 \xrightarrow{P} \mathbb{E} \left[\left(\alpha_0 G_0 + \boldsymbol{a}^T \boldsymbol{g} + \boldsymbol{b}_\ell - Y_\ell \right)^2 \right],$$

where the expectation is over $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}_r, \boldsymbol{I}_r)$ (with some abuse of notation) and

$$Y_\ell \sim \text{Bern}(V_\ell) \quad \text{and} \quad V_\ell = \frac{e^{\boldsymbol{e}_\ell^T \boldsymbol{V} \boldsymbol{\Sigma} \boldsymbol{g}}}{\sum_{\ell'=1}^r e^{\boldsymbol{e}_{\ell'} \boldsymbol{V} \boldsymbol{\Sigma} \boldsymbol{g}}}. \quad \text{(J.7)}$$

Furthermore,

$$\frac{1}{n} \boldsymbol{t}^T \left(\boldsymbol{I} + \frac{\tau}{\beta} \boldsymbol{D}^2 \right)^{-1} \boldsymbol{t} = \frac{1}{n} \sum_{i=1}^n \frac{\left(\alpha_0 \boldsymbol{g}_i + \boldsymbol{a}^T \tilde{\boldsymbol{g}}_i + b_\ell - [\boldsymbol{Y}_\ell]_i \right)^2}{1 + \frac{\tau}{\beta} d_i^2} \xrightarrow{P} \mathbb{E} \left[\frac{\left(\alpha_0 G_0 + \boldsymbol{a}^T \boldsymbol{g} + \boldsymbol{b}_\ell - Y_\ell \right)^2}{1 + \frac{\tau}{\beta} D^2} \right]$$

Therefore, point-wise on $\boldsymbol{a}, \alpha_0, \boldsymbol{b}_\ell, \tau$ and β, the objective $\mathcal{R}$ of the AO converges to

$$\begin{aligned} \mathcal{D}_\ell(\alpha_0, \boldsymbol{\alpha}, b_\ell, \tau, \beta) &\coloneqq \frac{\beta\tau}{2} + \frac{\beta}{2\tau} \mathbb{E} \left[\left(\alpha_0 G_0 + \boldsymbol{a}^T \boldsymbol{g} + \boldsymbol{b}_\ell - Y_\ell \right)^2 \right] - \frac{\beta}{2\tau} \mathbb{E} \left[\frac{\left(\alpha_0 G_0 + \boldsymbol{a}^T \boldsymbol{g} + \boldsymbol{b}_\ell - Y_\ell \right)^2}{1 + \frac{\tau}{\beta} D^2} \right] - \beta \alpha_0 \sqrt{\gamma} \\ &= \frac{\beta\tau}{2} + \frac{1}{2} \mathbb{E} \left[\frac{D^2 \left(\alpha_0 G_0 + \boldsymbol{a}^T \boldsymbol{g} + \boldsymbol{b}_\ell - Y_\ell \right)^2}{1 + \frac{\tau}{\beta} D^2} \right] - \beta \alpha_0 \sqrt{\gamma} \\ &= \frac{\beta\tau}{2} + \frac{1}{2} \mathbb{E} \left[\frac{\left(\alpha_0 G_0 + \boldsymbol{a}^T \boldsymbol{g} + \boldsymbol{b}_\ell - Y_\ell \right)^2}{D^{-2} + (\tau/\beta)} \right] - \beta \alpha_0 \sqrt{\gamma}. \end{aligned} \quad \text{(J.8)}$$

We note that the function above is jointly convex in $(\alpha_0, \boldsymbol{\alpha}, b_\ell, \tau)$ and concave in β.

J.1 Deterministic Analysis of the AO

It can be checked that the first order optimality conditions of $\mathcal{D}_\ell(\alpha_0, \boldsymbol{\alpha}, b_\ell, \tau, \beta)$ with respect to β and $\tau > 0$ are given as follows:

$$\beta^2 = \mathbb{E}\left[\frac{\left(\alpha_0 G_0 + \boldsymbol{a}^T \boldsymbol{g} + \boldsymbol{b}_\ell - Y_\ell\right)^2}{\left(D^{-2} + (\tau/\beta)\right)^2}\right] \quad \text{or} \quad \beta = 0, \tag{J.9}$$

$$\alpha_0\sqrt{\gamma} = \frac{\tau}{2} + \frac{\tau}{2\beta^2} \cdot \mathbb{E}\left[\frac{\left(\alpha_0 G_0 + \boldsymbol{a}^T \boldsymbol{g} + \boldsymbol{b}_\ell - Y_\ell\right)^2}{\left(D^{-2} + (\tau/\beta)\right)^2}\right]. \tag{J.10}$$

Thus, at optimality either $\beta = 0$ or $\tau = \alpha_0\sqrt{\gamma}$. In what follows, consider the solution $\tau = \alpha_0\sqrt{\gamma}$. We will show that this leads to the true saddle point of $\mathcal{D}$.

Moreover, by denoting $\eta := \frac{\beta}{\tau}$ and recalling from (J.7) that $Y_\ell = \mathrm{Bern}(V_\ell)$, we can express $\mathcal{D}_\ell(\alpha_0, \boldsymbol{\alpha}, b_\ell, \tau, \beta)$ as follows

$$\frac{\beta\tau}{2} + \frac{\alpha_0^2}{2}\mathbb{E}\left[\frac{1}{D^{-2} + 1/\eta}\right] - \beta\alpha_0\sqrt{\gamma} + \frac{1}{2}\begin{bmatrix}\boldsymbol{a}^T & \boldsymbol{b}_\ell\end{bmatrix} \cdot \boldsymbol{A}(\eta) \cdot \begin{bmatrix}\boldsymbol{a} \\ \boldsymbol{b}_\ell\end{bmatrix} - \boldsymbol{c}_\ell^T(\eta)\begin{bmatrix}\boldsymbol{a}^T \\ \boldsymbol{b}_\ell\end{bmatrix} + \frac{1}{2}\mathbb{E}\left[\frac{Y_\ell^2}{D^{-2} + 1/\eta}\right],$$

where

$$\boldsymbol{A}(\eta) := \begin{bmatrix}\mathbb{E}\left[\frac{\boldsymbol{g}\boldsymbol{g}^T}{D^{-2}+1/\eta}\right] & \mathbb{E}\left[\frac{\boldsymbol{g}}{D^{-2}+1/\eta}\right] \\ \mathbb{E}\left[\frac{\boldsymbol{g}^T}{D^{-2}+1/\eta}\right] & \mathbb{E}\left[\frac{1}{D^{-2}+1/\eta}\right]\end{bmatrix} \tag{J.11a}$$

$$\boldsymbol{c}_\ell(\eta) := \begin{bmatrix}\mathbb{E}\left[\frac{\boldsymbol{g}Y_\ell}{D^{-2}+1/\eta}\right]\end{bmatrix} \tag{J.11b}$$

we have the following first-order optimality conditions for $\alpha_0, \boldsymbol{a}$ and $\boldsymbol{b}_\ell$:

$$\begin{bmatrix}\boldsymbol{a} \\ \boldsymbol{b}_\ell\end{bmatrix} = \boldsymbol{A}^{-1}(\eta) \cdot \boldsymbol{c}_\ell(\eta) \tag{J.12}$$

$$\alpha_0 = \beta\sqrt{\gamma} \Big/ \mathbb{E}\left[\frac{1}{D^{-2} + 1/\eta}\right]. \tag{J.13}$$

Rearranging (J.13) and using $\tau = \alpha_0\sqrt{\gamma}$ gives the following equation for η:

$$\frac{\alpha_0\sqrt{\gamma}}{\beta}\mathbb{E}\left[\frac{1}{D^{-2} + 1/\eta}\right] = \gamma \overset{\tau=\alpha_0\sqrt{\gamma}}{\Longrightarrow} \mathbb{E}\left[\frac{1/\eta}{D^{-2} + 1/\eta}\right] = \gamma. \tag{J.14}$$

Thus, the optimal values of $\boldsymbol{a}$ and $\boldsymbol{b}_\ell$ are found by (J.12) for η the positive solution of the equation in (J.14). To solve for α_0, we combine (J.13) and (J.9) which leads to

$$\alpha_0^2\left(\gamma\eta^2 - \mathbb{E}\left[\left(\frac{1}{D^{-2} + 1/\eta}\right)^2\right]\right) = \mathbb{E}\left[\frac{\left(\boldsymbol{a}^T\boldsymbol{g} + \boldsymbol{b}_\ell - Y_\ell\right)^2}{\left(D^{-2} + 1/\eta\right)^2}\right], \tag{J.15}$$

where we have also used the RHS of (J.14). Next, we specialize these findings to the special structures of the weighting matrix $\boldsymbol{D}$ (J.1), respectively.

Applying weighting (J.1). Assume (J.1) holds. In this case, Equation (J.14) that determines the value of $\eta > 0$ becomes

$$F(\eta) := \sum_{i\in[k]} \frac{\boldsymbol{\pi}_i\omega_i^2}{\omega_i^2 + \eta} = \gamma, \tag{J.16}$$

where we have recalled the notation in (4.1) $\boldsymbol{\pi}_i := \mathbb{E}[V_i] > 0,\ i \in [k]$. It can be easily checked by direct differentiation that $\eta \mapsto F$ is strictly decreasing in $(0, \infty)$. Also, using $\sum_{i\in[k]} \boldsymbol{\pi}_i = 1$ the range of F in $(0, \infty)$ is $(0, 1)$. Thus, it follows that (J.16) has a unique solution for all $\gamma \in (0, 1)$.

Also, in this case we can write (J.11) in the following more convenient form:

$$\boldsymbol{A}(\eta) := \sum_{i\in[k]} \left(\frac{\omega_i^2\eta}{\omega_i^2+\eta}\right) \underbrace{\mathbb{E}\left[\begin{bmatrix}\boldsymbol{g}\\1\end{bmatrix}\begin{bmatrix}\boldsymbol{g}^T & 1\end{bmatrix} V_i\right]}_{=:\widetilde{\boldsymbol{A}}_i} \tag{J.17}$$

$$\boldsymbol{c}_\ell(\eta) := \left(\frac{\omega_\ell^2\eta}{\omega_\ell^2+\eta}\right) \underbrace{\mathbb{E}\left[\begin{bmatrix}\boldsymbol{g}\\1\end{bmatrix} V_\ell\right]}_{=:\widetilde{\boldsymbol{c}}_\ell}. \tag{J.18}$$

For convenience let us define vectors $\boldsymbol{\nu} := \boldsymbol{\nu}(\eta), \widetilde{\boldsymbol{\pi}} = \widetilde{\boldsymbol{\pi}}(\eta) \in \mathbb{R}^k$ with entries:

$$\widetilde{\boldsymbol{\pi}}_\ell := \boldsymbol{\pi}_i \left(\frac{1}{\gamma}\cdot\frac{\omega_i^2}{\omega_i^2+\eta}\right) =: \boldsymbol{\pi}_i\cdot\boldsymbol{\nu}_i \tag{J.19}$$

Because of (J.16), notice that $\widetilde{\boldsymbol{\pi}}$ is a probability vector, i.e. $\widetilde{\boldsymbol{\pi}}^T\mathbf{1}_k = 1$. Moreover, for η satisfying (J.16), it also holds that $\boldsymbol{\pi}^T\boldsymbol{\nu} = 1$.

With the notation above, it holds

$$\begin{aligned}
\boldsymbol{A}(\eta) &= \gamma\cdot\eta\cdot\begin{bmatrix}\sum_{i\in[k]}\boldsymbol{\nu}_i\cdot\mathbb{E}[V_i\boldsymbol{g}\boldsymbol{g}^T] & \sum_{i\in[k]}\boldsymbol{\nu}_i\cdot\mathbb{E}[V_i\boldsymbol{g}]\\ \sum_{i\in[k]}\boldsymbol{\nu}_i\cdot\mathbb{E}[V_i\boldsymbol{g}^T] & 1\end{bmatrix}\\
&= \gamma\cdot\eta\cdot\begin{bmatrix}\sum_{i\in[k]}\boldsymbol{\nu}_i\cdot\mathbb{E}[V_i\boldsymbol{g}\boldsymbol{g}^T] & \boldsymbol{\Sigma}\boldsymbol{V}^T(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi})\,\boldsymbol{\nu}\\ \boldsymbol{\nu}^T(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi})\boldsymbol{V}\boldsymbol{\Sigma} & 1\end{bmatrix} &\text{(J.20)}\\
&= \gamma\cdot\eta\cdot\begin{bmatrix}\mathbb{E}\left[\left(\boldsymbol{\nu}^T\boldsymbol{v}\right)\boldsymbol{g}\boldsymbol{g}^T\right] & \boldsymbol{\Sigma}\boldsymbol{V}^T(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi})\,\boldsymbol{\nu}\\ \boldsymbol{\nu}^T(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi})\boldsymbol{V}\boldsymbol{\Sigma} & 1\end{bmatrix} &\text{(J.21)}\\
\boldsymbol{c}_\ell(\eta) &= \gamma\cdot\eta\cdot\begin{bmatrix}\boldsymbol{\nu}_\ell\,\mathbb{E}[V_\ell\boldsymbol{g}]\\ \widetilde{\boldsymbol{\pi}}_\ell\end{bmatrix}\\
&= \gamma\cdot\eta\cdot\begin{bmatrix}\boldsymbol{\Sigma}\boldsymbol{V}^T(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi})\,\boldsymbol{\nu}_\ell\,\boldsymbol{e}_\ell\\ \widetilde{\boldsymbol{\pi}}_\ell\end{bmatrix} &\text{(J.22)}
\end{aligned}$$

where we have also used the fact that $\mathbb{E}[V_i\boldsymbol{g}] = \boldsymbol{\Sigma}\boldsymbol{V}^T(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi})\,\boldsymbol{e}_i,\ i\in[k]$ and recalled the notation

$$\boldsymbol{v} = [V_1,\dots,V_k]^T.$$

Using (J.30) and (J.31), we conclude from (J.12) the following expressions for $\boldsymbol{a}$ and $\boldsymbol{b}$:

$$\boldsymbol{a} = \boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi})\cdot\boldsymbol{\nu}_\ell\cdot(\boldsymbol{e}_\ell-\boldsymbol{\pi}_\ell\boldsymbol{\nu}), \tag{J.23}$$

$$\boldsymbol{b}_\ell = \widetilde{\boldsymbol{\pi}}_\ell - \boldsymbol{\nu}^T(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi})\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi})\cdot\boldsymbol{\nu}_\ell\cdot(\boldsymbol{e}_\ell-\boldsymbol{\pi}_\ell\boldsymbol{\nu}), \tag{J.24}$$

where we defined

$$\boldsymbol{\Delta} = \mathbb{E}\left[\left(\boldsymbol{\nu}^T\boldsymbol{v}\right)\boldsymbol{g}\boldsymbol{g}^T\right] - \boldsymbol{\Sigma}\boldsymbol{V}^T(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi})\,\boldsymbol{\nu}\boldsymbol{\nu}^T(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi})\,\boldsymbol{V}\boldsymbol{\Sigma} \succ \mathbf{0}_{r\times r}. \tag{J.25}$$

Finally, we show how to compute α_0 using (J.15). The RHS in (J.15) can be computed as

$$\begin{aligned}
&\sum_{i\neq\ell\in[k]} \frac{\begin{bmatrix}\boldsymbol{a}^T & \boldsymbol{b}_\ell\end{bmatrix}\widetilde{\boldsymbol{A}}_i\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}_\ell\end{bmatrix}}{\left(\omega_i^{-2}+1/\eta\right)^2} + \frac{\begin{bmatrix}\boldsymbol{a}^T & \boldsymbol{b}_\ell\end{bmatrix}\widetilde{\boldsymbol{A}}_\ell\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}_\ell\end{bmatrix} - 2\begin{bmatrix}\boldsymbol{a}^T & \boldsymbol{b}_\ell\end{bmatrix}\widetilde{\boldsymbol{c}}_\ell + \boldsymbol{\pi}_\ell}{\left(\omega_\ell^{-2}+1/\eta\right)^2}\\
&\qquad = \eta^2\cdot\gamma^2\cdot\left\{\begin{bmatrix}\boldsymbol{a}^T & \boldsymbol{b}_\ell\end{bmatrix}\left(\sum_{i\in[k]}\boldsymbol{\nu}_i^2\widetilde{\boldsymbol{A}}_i\right)\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}_\ell\end{bmatrix} - 2\begin{bmatrix}\boldsymbol{a}^T & \boldsymbol{b}_\ell\end{bmatrix}\boldsymbol{\nu}_\ell^2\widetilde{\boldsymbol{c}}_\ell + \boldsymbol{\pi}_\ell\boldsymbol{\nu}_\ell^2\right\},
\end{aligned}$$

where $\boldsymbol{a}, \boldsymbol{b}_\ell$ are as in (J.23) and (J.24). Also, note that

$$\mathbb{E}\left[\left(\frac{1}{D^{-2}+1/\eta}\right)^2\right] = \eta^2\sum_{i\in[k]}\frac{\boldsymbol{\pi}_i\omega_i^4}{\left(\omega_i^2+\eta\right)^2} = \eta^2\cdot\gamma^2\cdot\boldsymbol{\pi}^T\operatorname{diag}(\boldsymbol{\nu})\boldsymbol{\nu} = \eta^2\cdot\gamma^2\cdot\widetilde{\boldsymbol{\pi}}^T\boldsymbol{\nu}.$$

Put together, we have the following expression for α_0:

$$\alpha_0^2 = \frac{1}{(1/\gamma - \widetilde{\boldsymbol{\pi}}^T\boldsymbol{\nu})} \cdot \left\{ \begin{bmatrix} \boldsymbol{a}^T & \boldsymbol{b}_\ell \end{bmatrix} \left(\sum_{i\in[k]} \boldsymbol{\nu}_i^2 \widetilde{\boldsymbol{A}}_i \right) \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b}_\ell \end{bmatrix} - 2 \begin{bmatrix} \boldsymbol{a}^T & \boldsymbol{b}_\ell \end{bmatrix} \boldsymbol{\nu}_\ell^2 \widetilde{\boldsymbol{c}}_\ell + \boldsymbol{\pi}_\ell \boldsymbol{\nu}_\ell^2 \right\}$$
$$= \frac{1}{(1/\gamma - \widetilde{\boldsymbol{\pi}}^T\boldsymbol{\nu})} \cdot \left\{ \begin{bmatrix} \boldsymbol{a}^T & \boldsymbol{b}_\ell \end{bmatrix} \boldsymbol{A}' \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b}_\ell \end{bmatrix} - 2 \begin{bmatrix} \boldsymbol{a}^T & \boldsymbol{b}_\ell \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}\boldsymbol{V}^T \left(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi}\right) \boldsymbol{\nu}_\ell^2 \boldsymbol{e}_\ell \\ \widetilde{\boldsymbol{\pi}}_\ell \cdot \boldsymbol{\nu}_\ell \end{bmatrix} + \widetilde{\boldsymbol{\pi}}_\ell \cdot \boldsymbol{\nu}_\ell \right\}, \tag{J.26}$$

where $\boldsymbol{a}, \boldsymbol{b}_\ell$ are as in (J.23), (J.24) and we have also defined

$$\boldsymbol{A}' = \begin{bmatrix} \mathbb{E}\left[\left(\boldsymbol{\nu}^T \mathrm{diag}(\boldsymbol{\nu}) \boldsymbol{v} \right) \boldsymbol{g}\boldsymbol{g}^T \right] & \boldsymbol{\Sigma}\boldsymbol{V}^T \left(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi}\right) \mathrm{diag}(\boldsymbol{\nu}) \boldsymbol{\nu} \\ \boldsymbol{\nu}^T \mathrm{diag}(\boldsymbol{\nu}) \left(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi}\right) \boldsymbol{V}\boldsymbol{\Sigma} & \boldsymbol{\nu}^T \mathrm{diag}(\boldsymbol{\nu}) \boldsymbol{\pi} \end{bmatrix}. \tag{J.27}$$

Asymptotic Predictions. Writing (J.24) in vector form we find that

$$\widehat{\boldsymbol{b}} \xrightarrow{P} \widetilde{\boldsymbol{\pi}} - \mathrm{diag}(\boldsymbol{\nu}) \left(\boldsymbol{I}_k - \boldsymbol{\pi}\boldsymbol{\nu}^T \right) \left(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi}\right) \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T \left(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi}\right) \boldsymbol{\nu}. \tag{J.28}$$

Also, recalling that $\boldsymbol{e}_\ell^T \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}} = \widehat{\boldsymbol{w}}_\ell \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \xrightarrow{P} \boldsymbol{a}\boldsymbol{\Sigma}\boldsymbol{V}^T$ and using (J.23):

$$\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}} \xrightarrow{P} \mathrm{diag}(\boldsymbol{\nu}) \left(\boldsymbol{I}_k - \boldsymbol{\pi}\boldsymbol{\nu}^T \right) \left(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi}\right) \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{J.29}$$

Finally, for the magnitudes of the weight vectors, recall that $\|\widehat{\boldsymbol{w}}_\ell\|_{\ell_2}^2 \xrightarrow{P} \|\boldsymbol{a}\|_{\ell_2}^2 + \alpha_0^2$. Thus, to find the limiting values of the norms, we can combine (J.26) and (J.23)-(J.24). For convenience, we summarize the final expression here. Define the following[1]

$$\boldsymbol{A} := \begin{bmatrix} \mathbb{E}\left[\left(\boldsymbol{\nu}^T \boldsymbol{v} \right) \boldsymbol{g}\boldsymbol{g}^T \right] & \boldsymbol{\Sigma}\boldsymbol{V}^T \left(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi}\right) \boldsymbol{\nu} \\ \boldsymbol{\nu}^T \left(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi}\right) \boldsymbol{V}\boldsymbol{\Sigma} & 1 \end{bmatrix} \tag{J.30}$$

$$\boldsymbol{c}_\ell = \begin{bmatrix} \boldsymbol{\Sigma}\boldsymbol{V}^T \left(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi}\right) \boldsymbol{\nu}_\ell \boldsymbol{e}_\ell \\ \widetilde{\boldsymbol{\pi}}_\ell \end{bmatrix} \tag{J.31}$$

Further recall the matrix $\boldsymbol{A}'$ in (J.27).

$$\|\widehat{\boldsymbol{w}}_\ell\|_{\ell_2}^2 \xrightarrow{P} \left\| \boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T \left(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\Pi}\right) \cdot \boldsymbol{\nu}_\ell \cdot (\boldsymbol{e}_\ell - \boldsymbol{\pi}_\ell \boldsymbol{\nu}) \right\|_{\ell_2}^2$$
$$+ \frac{1}{(1/\gamma - \widetilde{\boldsymbol{\pi}}^T\boldsymbol{\nu})} \cdot \left\{ \boldsymbol{c}_\ell^T \boldsymbol{A}^{-1} \boldsymbol{A}' \boldsymbol{A}^{-1} \boldsymbol{c}_\ell - 2\boldsymbol{\nu}_\ell \boldsymbol{c}_\ell^T \boldsymbol{A}^{-1} \boldsymbol{c}_\ell + \widetilde{\boldsymbol{\pi}}_\ell \cdot \boldsymbol{\nu}_\ell \right\}, \tag{J.32}$$

Remark J.2 *Consider the special case $\omega_i = 1$, $i \in [k]$. We show how the above recovers the solution for (un-weighted) LS. First, note that in this case (J.16) simply gives $\eta = \frac{1}{\gamma} - 1$. Thus, $\boldsymbol{\nu} = \mathbf{1}_k$ and $\widetilde{\boldsymbol{\pi}} = \boldsymbol{\pi}$. Also, recall that $(diag(\boldsymbol{\pi}) - \boldsymbol{\Pi})\,\mathbf{1}_k = \mathbf{0}$ and $\mathbf{1}^T\boldsymbol{v} = 1$. Thus, (J.25) simply gives $\boldsymbol{\Delta} = \mathbb{E}[\boldsymbol{g}\boldsymbol{g}^T] = \boldsymbol{I}_r$. With these, it can be readily checked that (J.28) and (J.23) simplify to the expressions in (4.4a). Similarly, $\boldsymbol{A} = \boldsymbol{A}' = \boldsymbol{I}_{r+1}$ and (J.26) reduces in this case to (H.1.1). For general weight coefficients, such simplifications do not seem possible and one needs to compute the matrix $\mathbb{E}\left[\left(\boldsymbol{\nu}^T \boldsymbol{v} \right) \boldsymbol{g}\boldsymbol{g}^T \right]$ that appears in the definitions of $\boldsymbol{\Delta}, \boldsymbol{A}$ and $\boldsymbol{A}'$. We note that this calculation can be somewhat simplified by applying Gaussian integration by parts similar to lemma B.2.*

J.3 Computing cross-correlations $\Sigma_{w,w}$

In this section, we use Lemma I.1 to compute the cross-correlations $\langle \widehat{\boldsymbol{w}}_\ell, \widehat{\boldsymbol{w}}_j \rangle$, $j \neq \ell \in [k]$. Specifically, the analysis of (I.13) is almost identical to the analysis of (J.3) in the previous section. Specifically, without repeating all the details for brevity, it can be shown that the AO of (I.13) converges to $\min_{\boldsymbol{a},\alpha_0\geq 0,b_\ell,\tau>0} \max_{\beta\geq 0} \mathcal{D}(\boldsymbol{a},\alpha_0,b_\ell,\tau,\beta)$ where $\mathcal{D}(\boldsymbol{a},\alpha_0,b_\ell,\tau,\beta)$ is as in (J.8) only with Y_ℓ substituted by $Y_{\ell,c}$:

$$Y_{\ell,c} \sim \mathrm{Bern}(V_c + V_\ell) \quad \text{and as before:} \quad V_i = \frac{e^{\boldsymbol{e}_i^T \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}}{\sum_{\ell'=1}^r e^{\boldsymbol{e}_{\ell'} \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{g}}},\ i = \ell, c. \tag{J.33}$$

[1] Note the slight abuse of notation compared to the definitions in (J.30) and (J.30). This "renaming" should not be confusing as the constant $\gamma \cdot \eta$ (that is different between the two definitions) cancels in the evaluation of $\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b}_\ell \end{bmatrix} = \boldsymbol{A}^{-1}\boldsymbol{c}_\ell$ (see (J.12)).

Thus, what changes in the calculations above is in (J.18) and (J.35), where we now have instead

$$\boldsymbol{c}(\eta) := \left(\frac{1}{\frac{1}{\omega_\ell^2}+1/\eta}\right)\underbrace{\mathbb{E}\left[\begin{bmatrix}\boldsymbol{g}\\1\end{bmatrix}V_\ell\right]}_{=:\widetilde{\boldsymbol{c}}_\ell} + \left(\frac{1}{\frac{1}{\omega_c^2}+1/\eta}\right)\underbrace{\mathbb{E}\left[\begin{bmatrix}\boldsymbol{g}\\1\end{bmatrix}V_c\right]}_{=:\widetilde{\boldsymbol{c}}_c} \tag{J.34}$$

and

$$\sum_{i\neq\{\ell,c\}\in[k]} \frac{\begin{bmatrix}\boldsymbol{a}^T & \boldsymbol{b}_\ell\end{bmatrix}\widetilde{\boldsymbol{A}}_i\begin{bmatrix}\boldsymbol{a}\\\boldsymbol{b}_\ell\end{bmatrix}}{\left(\omega_i^{-2}+1/\eta\right)^2} + \frac{\begin{bmatrix}\boldsymbol{a}^T & \boldsymbol{b}_\ell\end{bmatrix}\widetilde{\boldsymbol{A}}_\ell\begin{bmatrix}\boldsymbol{a}\\\boldsymbol{b}_\ell\end{bmatrix} - 2\begin{bmatrix}\boldsymbol{a}^T & \boldsymbol{b}_\ell\end{bmatrix}\widetilde{\boldsymbol{c}}_\ell + \pi_\ell}{\left(\omega_\ell^{-2}+\eta\right)^2} + \frac{\begin{bmatrix}\boldsymbol{a}^T & \boldsymbol{b}_\ell\end{bmatrix}\widetilde{\boldsymbol{A}}_c\begin{bmatrix}\boldsymbol{a}\\\boldsymbol{b}_\ell\end{bmatrix} - 2\begin{bmatrix}\boldsymbol{a}^T & \boldsymbol{b}_\ell\end{bmatrix}\widetilde{\boldsymbol{c}}_c + \pi_c}{\left(\omega_c^{-2}+1/\eta\right)^2}, \tag{J.35}$$

respectively. With these and following mutatis-mutandis the steps and the notation in the previous section, we find the following asymptotic expression for the magnitude of $\widehat{\boldsymbol{w}}_\ell + \widehat{\boldsymbol{w}}_c$:

$$\begin{aligned}\|\widehat{\boldsymbol{w}}_\ell + \widehat{\boldsymbol{w}}_c\|_{\ell_2}^2 \xrightarrow{P} &\left\|\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi}\right)\cdot\left(\boldsymbol{\nu}_\ell\cdot\left(\boldsymbol{e}_\ell-\boldsymbol{\pi}_\ell\boldsymbol{\nu}\right)+\boldsymbol{\nu}_c\cdot\left(\boldsymbol{e}_c-\boldsymbol{\pi}_c\boldsymbol{\nu}\right)\right)\right\|_{\ell_2}^2\\ &+\frac{1}{\left(1/\gamma-\widetilde{\boldsymbol{\pi}}^T\boldsymbol{\nu}\right)}\cdot\left\{\left(\boldsymbol{c}_\ell+\boldsymbol{c}_c\right)^T\boldsymbol{A}^{-1}\boldsymbol{A}'\boldsymbol{A}^{-1}\left(\boldsymbol{c}_\ell+\boldsymbol{c}_c\right)-2\left(\boldsymbol{\nu}_\ell\boldsymbol{c}_\ell+\boldsymbol{\nu}_c\boldsymbol{c}_c\right)^T\boldsymbol{A}^{-1}\left(\boldsymbol{\nu}_\ell\boldsymbol{c}_\ell+\boldsymbol{\nu}_c\boldsymbol{c}_c\right)+\widetilde{\boldsymbol{\pi}}_\ell\cdot\boldsymbol{\nu}_\ell+\widetilde{\boldsymbol{\pi}}_c\cdot\boldsymbol{\nu}_c\right\},\end{aligned} \tag{J.36}$$

We may now combine this with (J.32) to conclude with the following asymptotic limits for the cross-correlations for all $\ell \neq c \in [k]$:

$$\begin{aligned}\langle\widehat{\boldsymbol{w}}_\ell,\widehat{\boldsymbol{w}}_c\rangle \xrightarrow{P} &\boldsymbol{\nu}_c\left(\boldsymbol{e}_c-\boldsymbol{\pi}_c\boldsymbol{\nu}\right)^T\left(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi}\right)\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-2}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi}\right)\boldsymbol{\nu}_\ell\left(\boldsymbol{e}_\ell-\boldsymbol{\pi}_\ell\boldsymbol{\nu}\right)\\ &+\frac{1}{\left(1/\gamma-\widetilde{\boldsymbol{\pi}}^T\boldsymbol{\nu}\right)}\cdot\left\{\boldsymbol{c}_c^T\boldsymbol{A}^{-1}\boldsymbol{A}'\boldsymbol{A}^{-1}\boldsymbol{c}_\ell-\boldsymbol{\nu}_c\boldsymbol{\nu}_\ell\boldsymbol{c}_c^T\boldsymbol{A}^{-1}\boldsymbol{c}_\ell\right\}.\end{aligned} \tag{J.37}$$

$$\begin{aligned}=&\boldsymbol{\nu}_c\left(\boldsymbol{e}_c-\boldsymbol{\pi}_c\boldsymbol{\nu}\right)^T\left(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi}\right)\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-2}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi}\right)\boldsymbol{\nu}_\ell\left(\boldsymbol{e}_\ell-\boldsymbol{\pi}_\ell\boldsymbol{\nu}\right)\\ &+\frac{1}{\left(1/\gamma-\widetilde{\boldsymbol{\pi}}^T\boldsymbol{\nu}\right)}\cdot\left\{\boldsymbol{c}_c^T\left(\boldsymbol{A}^{-1}\boldsymbol{A}'\boldsymbol{A}^{-1}-\boldsymbol{\nu}_c\boldsymbol{\nu}_\ell\boldsymbol{A}^{-1}\right)\boldsymbol{c}_\ell\right\}.\end{aligned} \tag{J.38}$$

In matrix form, we have

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}} \xrightarrow{P} &\operatorname{diag}(\boldsymbol{\nu})\left(\boldsymbol{I}_k-\boldsymbol{\pi}\boldsymbol{\nu}^T\right)\left(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi}\right)\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-2}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi}\right)\left(\boldsymbol{I}_k-\boldsymbol{\pi}\boldsymbol{\nu}^T\right)\operatorname{diag}(\boldsymbol{\nu})\\ &+\frac{1}{\left(1/\gamma-\widetilde{\boldsymbol{\pi}}^T\boldsymbol{\nu}\right)}\Big\{\left[\operatorname{diag}(\boldsymbol{\nu})\left(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi}\right)\boldsymbol{V}\boldsymbol{\Sigma}\quad\widetilde{\boldsymbol{\pi}}\right]\boldsymbol{A}^{-1}\boldsymbol{A}'\boldsymbol{A}^{-1}\begin{bmatrix}\operatorname{diag}(\boldsymbol{\nu})\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi}\right)\\ \widetilde{\boldsymbol{\pi}}^T\end{bmatrix}\\ &+\operatorname{diag}(\boldsymbol{\nu})\left[\operatorname{diag}(\boldsymbol{\nu})\left(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi}\right)\boldsymbol{V}\boldsymbol{\Sigma}\quad\widetilde{\boldsymbol{\pi}}\right]\boldsymbol{A}^{-1}\begin{bmatrix}\operatorname{diag}(\boldsymbol{\nu})\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\operatorname{diag}(\boldsymbol{\pi})-\boldsymbol{\Pi}\right)\\ \widetilde{\boldsymbol{\pi}}^T\end{bmatrix}\operatorname{diag}(\boldsymbol{\nu})\Big\}\end{aligned} \tag{J.39}$$