We thank the reviewers for their feedback and the time spent on our submission.

First, let us elaborate on the concerns by **Reviewers 1 & 3** regarding restrictions to specific linear classifiers. To quote **Reviewer 4** – who we thank for the encouraging feedback – : "*The theoretical analysis of multiclass classification is an open problem at the core of machine learning/statistical modelling. While the specific setting considered seem limited, they are insightful and likely an important stepping stone in the full analysis of multi-class classification...Such results are a very good lead for future investigations of more general settings*". Below, we remark on the following regarding the motivation behind the studied classifiers and the impact of our results. We will expand upon these in the revision.

(I) On the averaging estimator: The averaging estimator is Bayes optimal for balanced GMM with Gaussian means (Prop. 3.4). As such, it serves well as a baseline and a "natural algorithm" for this data model. One then might wonder how the performance of this algorithm depends on the data model. Thus, we further analyze its performance for the second basic model considered here: the MLM. **(II)** On the LS classifier: There have been numerous empirical works that investigate the role of the loss function in classification tasks for various data models. Several of these find empirically that simple LS can have comparable performance compared to the (perhaps most commonly used) hinge/logistic losses, e.g. [53,74,75]. Quoting [p.105, 53]: "*Intuitively, it seems that the square loss may be less well suited to classification than the hinge loss (...) However, in practice, we have found that the accuracy of RLSC (regularized LS classification) is essentially equivalent to that of SVMs*." One of the long-term goals of our project is to provide theoretical evidence against/in-favor of such empirical findings and to characterize what loss is suitable for each setting. As a first step, we naturally ask whether these claims are already justified (or not) in simple linear models, and if so, under what conditions. Along these lines, there are several recent works that theoretically study the role of LS in high-dimensional *binary* linear classification. For example, under the same asymptotic regime as in our paper, [44] proves that LS is optimal for GMM within the family of convex un-regularized empirical-risk minimization, and, [60] proves that LS is approximately optimal (thus, comparable to the ML solution: logistic loss) for logistic data. We take the first steps towards extending these to the more challenging, but more versatile, *multiclass* setting. **(III)** On WLS: (a) We are motivated by recent findings [13] that "weighted" variations of LS can significantly boost the performance over simple LS. (b) Compared to LS, WLS offers the flexibility to adjust the algorithm to balance performance between majority vs minority classes (together with our ability to accurately predict class-wise errors).

While there is a lot to do further down the road, our results (model setup, analysis, sharp asymptotics) are the first step towards this direction and facing some of the new challenges in multiclass settings (see lines 45-52). Certain important additions such as *regularized* (W)LS and correlated Gaussian features – while requiring extra work – are almost direct extensions of our current framework. Others, such as the study of cross-entropy minimization or extreme classification will likely require combining elements of our work with new ideas. We believe that our paper sets the fundamentals and will inspire further investigations in this direction. Of course, extensions to non-asymptotic regimes and non-linear models (e.g., RFF, NTK) are highly desirable. Such results, only recently obtained for regression settings, are typically founded on long prior work on simpler regression models – linear, (isotropic) Gaussian, etc.. Our work, together with refs. in (lines 81-82) for binary settings, resemble these essential precursor works for the setting of classification.

Second, on **Reviewer's 1** question on the relative performance of the averaging and LS estimators for the two data models: This is discussed in Sec. 3.2 and 4.2. In Prop. 3.3, we prove that averaging outperforms LS for balanced GMM with orthogonal means. Intuitively, this is because "compared to the weight vectors $\mathbf{w}_i$ of the class averaging classifier that are also (asymptotically) orthogonal when means are orthogonal, this is not the case for LS" (line 222, pg.6). In fact in Sec. 3.3 we formally study the optimality of the averaging estimator in a Bayesian GMM setting. Similarly, in Prop. 4.3 and Sec. 4.2, we show that LS outperforms the averaging estimator in MLM for large data samples.

Third, we agree with **Reviewers 1 and 4** that sketching key proof ideas in the main body of the paper will benefit the reader. If accepted, we will use the extra space to move the corresponding discussions from App. F to the main body.

Fourth, in response to **Reviewer's 4** suggestion. Indeed, results for binary classification can be lifted to characterize the limit of $\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{\mu}}$ and diagonal entries of $\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}}$ for one-vs-all classifiers (including LS) (with some additional technical work to capture correlations $\boldsymbol{\Sigma}_{\boldsymbol{\mu},\boldsymbol{\mu}}$ for k>2). As mentioned in the paper, this alone does not give any information on the off-diagonals of $\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}}$, needed in the exact test-error formulas (2.3)/(2.4). It is possible to derive heuristic approximations and union bound arguments leading to error expressions that depend only on the diagonals of $\boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{w}}$. Indeed, Fig. 5 in App. A provides a result of this flavor and gives a sense of how our exact results improve upon such approximations. We will expand upon this comparison in Appendix A in the revision.

Reviewer 2: With respect to test error, our paper is precisely about characterizing the performance of the studied linear classifiers *in the sense of test error.* The formulas of Thms. 3.1,3.2,4.1,4.2 can be directly plugged in (2.3) and (2.4) to obtain test error. Regarding "calculations in the double asymptotic regime are not quite new": Of course, there are numerous works in this regime under numerous settings over the last decade (lines 77-79). However, ours is *the first such work in multiclass classification* This point is well-articulated in the introduction (lines 73-76, 82-91).