We thank all the reviewers for their feedback and pointers to relevant papers. We first address some general comments raised by multiple reviewers and next respond to individual questions.

Extended discussion of related work: We will add a detailed discussion of all the relevant papers mentioned by reviewers to the final version. This includes (Kendall et al., 2018), where they learn $1/\sigma$ of a Gaussian kernel as weights for base objectives, and (Chen et al., 2018), where they dynamically update gradient norms for each base loss. The number of empirical papers, such as those mentioned above, that explore training with multiple objectives/tasks is quite large. While we are happy to discuss as many of the relevant references as we can, the main focus of our paper, as concisely summarized by REV1, is to provide a theoretical analysis and motivation for the agnostic learning scenario.

The goals and structure of our experiments: As suggested by REV2, we will add a detailed formulation of the goals of the experiments and will connect them to the evaluation metrics we use. Experiments serve to 1) illustrate the application of our novel agnostic learning formulation; and 2) support a claim that since our algorithm optimizes for the worst case mixture, it provides more robust results than training with a fixed mixture. Thus, comparing ALMO with a uniform mixture baseline is the most direct and clean experiment to demonstrate 1) and 2). We use AUC, which is not included in the training loss mixture, as a robustness measure and show that ALMO has better AUC than the baselines.

Connection to multi-task literature and baselines: We recognize in the paper that multi-task learning is relevant, it is nevertheless distinct from our framework. That being said, the ALMO algorithm could indeed benefit from some of the multi-task literature; and as pointed out, in reverse, our algorithm could be relevant to that setting. The key distinction lies in the way we work with base loss functions and output spaces. While multi-tasking typically considers the same loss function on different domains where the distributions and output spaces may differ (and many methods derive from that assumption, e.g. Kendall et al. 2018), we consider different loss functions on the same output space. This difference is why we have limited the discussion of multi-task literature, although we did include some references. The topic of our paper is "Agnostic Multi-Objective Learning" which is a bit different from "Agnostic Multi-Task Learning".

There are specific reasons we did not use several multi-task learning algorithms mentioned by REV4 as baselines. First, Kendall et al. (2018) assumes that all base losses are applications of the same function (max likelihood in this case) on different output. This is where the scaling factor $1/\sigma$ comes from, which is treated as weights for base Gaussian likelihoods. We don't see how this method can be extended to our scenario where base losses do not necessarily represent Gaussian likelihoods and in fact, they come from different functional families, such as the hinge loss and cross-entropy loss. Moreover, our regularization admits a very different nature. Second, the GRADNORM of Chen et al. (2018) also applies the same loss to different outputs. Since losses have different magnitudes for different output domains, GRADNORM is a useful multi-task normalization method. It can be used as a helpful subroutine in ALMO to balance the gradients. However, directly normalizing the base losses was sufficient for our experiments. Finally, for several of the suggested multi-task methods, we are not sure they provide Pareto-optimal solutions.

REV1: The reviewer is correct about the type of the Pareto-optimal solution achieved by ALMO, we will include relevant details. We have verified that lambda weights (as in Figure 2) are not significantly different for different values of initial weights. As for the AUC objective, since we are working with multi-class classification (e.g., 10 classes for MNIST) the ranking is based on the relative probabilities of each class. As recommended by REV1, we will fix the caption in Table 1 (the figures are for the testing set), and will add a detailed discussion of single-loss results.

REV2: In the final version, we will add details of how to extend Equation 11 to the regularized problem: when a convex regularization term is added to the loss ($\mathcal{W}$, Λ unchanged) the problem is still convex and the math is similar even for the projections. The minimax can indeed be viewed as an equilibrium in a two-player adversarial game, we will add more explanation for that. We provide more detail in lines 35-46 about why we don't apply Pareto-frontier search methods, but as suggested we will strengthen the argument and also illustrate with examples that in large-scale data settings, such as when the number of base losses is large, Pareto frontier estimation can often be computationally infeasible. On the other hand, our solution is theoretically proven to be Pareto-optimal, thus we don't need to verify that empirically.

REV3: The sensitivity to Lipschitz constants is directly captured by the M_k-weighted lambda norm and related terms in the generalization bound of Theorem 3. The zero-one loss that we report is 1–accuracy. As suggested, we will add more experimental details such as the training time (which took several hours at most). Please note that we have made sure all nontrivial and essential details for reproducibility are included in the NeurIPS submission.

REV4: Comments about related work and baselines are addressed above. Please note that Malkiel et al. was published only after the NeurIPS-2020 deadline. However, their optimization method can possibly help speed up the ALMO algorithm and we are interested in trying it. The auto-encoder experiment is definitely worth exploring in future work, though as we discussed, it is out of the scope of our learning scenario and the claims we put forward. Particularly, because there is no clear ground to assert that the goal of the learner is to optimize for the worst-case mixture of reconstruction, classification and regression losses. As hinted by REV4, the real goal of the learner might be to evaluate whether adding the reconstruction loss improves the generalization of the underlying model on multiple tasks.