We thank all reviewers for their time and feedback. Major comments are addressed below. Minor comments (typos/stylistic changes/adding expository text) will be addressed in the final version to make it more reader friendly.

R1: *"difficult to review the novelty and innovation of the paper"* We understand that the paper was outside of your main area of interest. Note that SPO is a new framework that includes multiclass classification as a special case. By developing combinatorial dimension based and margin-based generalization bounds for the SPO framework, we are rebuilding the two major pillars of standard generalization theory in a more challenging, novel setting.

R2: Thanks for your careful reading of our paper and an accurate understanding of our contributions. We are glad you highlighted the importance of: 1) construction of generalization theory for the SPO framework for the first time, 2) a generalized margin loss function and margin-theory in the strongly convex case, and 3) opening up a new area of investigation for researchers at the intersection of Operations Research (OR) and Machine Learning (ML).

R3: *"the authors claimed that their margin-based approach removes the dependency on the number of classes ... However, I can not see clearly this."* We agree that a bit of context is needed to interpret our statement in the introduction. We will revise the introduction accordingly. The main point is to see that one can generally construct a multiclass classification instance from an instance of an SPO problem by considering the "label" of each observed cost vector c_i to be the corresponding optimal solution $w^*(c_i)$, which is w.l.o.g. an extreme point of S. Thus, the "number of classes" is the number of extreme points of S. Note however that this reduction throws away potentially important information, namely the numerical values of the cost vectors c_i. (Note also that Example 1 presents a case where this reduction does not remove information, in which case S is the unit simplex in $\mathbb{R}^d$ with d extreme points.) Now, notice that the margin based approach of Section 4 makes an important assumption that S is strongly convex (which necessarily implies that the number of extreme points/classes is *infinite*) and also heavily uses the structure of the SPO loss via the construction of the γ-margin SPO loss. This refined analysis allows us to circumvent a naive bound that depends on the infinite number of classes, which would be vacuous. In fact, the dependence on the dimension d is often only mild – e.g. logarithmic or square root as you mention – and improves upon the linear dependency in the bound generated via the discretization argument presented in Corollary 2.

"It is not clear to me whether the dependency is optimal" In specific cases, such as multiclass classification, our bounds are comparable with the best available ones (see below). As R2 pointed out, this is the *first* work on generalization theory for SPO loss. Therefore, in the general case, optimality has not been investigated. We hope our work paves the way for deriving matching lower bounds (mentioned in open problems in Section 5).

"The authors should present comparison with existing work more clearly to show how this work advance the state of the art" Just to clarify, ours is the *first* work to develop generalization and margin theory for the SPO framework. At that level of generality, related work simply does not exist.

"how the results applied to multi-class classification can be compared with existing works" Since multiclass classification is a special case, it does make sense to compare bounds. Many multiclass approaches are theoretically compared in "Multiclass Learning Approaches: A Theoretical Comparison with Implications" by Daniely et al. Different approaches (like one-vs-all, all-pairs) use different hypothesis classes. The one which is most relevant to us is their MSVM approach that uses a single space of multiclass classifiers without reducing the problem to binary classification problems. In the linear hypothesis case with d classes and p features, they show that the bound of $\tilde{O}\left(\sqrt{pd/n}\right)$ is tight up to logarithmic factors. Since $d_N(w^*(\mathcal{H})) \leq dp$ in the case of linear classifiers, our result is also tight up to logarithmic factors. We will add this comparison to the final version.

R4: *"SPO loss has not attracted much attention in the machine learning community previously"* The general problem of understanding the impact of errors in machine learning predictions when they are fed to other decision analysis tools has been gathering attention (we cite 2 neurips papers [6, 13] and 2 OR papers [3,7]). Note that OR journals have long reviewing periods which explains why the references are to their arXiv versions. The problem will increase in importance as ML is integrated in real-world decision making pipelines.

"Theorem 1: it is not very clear how it is related to one in Bartlett and Mendelson, if these are exactly the same or not." Theorem 1 is indeed a minor rewriting of a landmark result of Bartlett and Mendelson in a form that is easy for us to use in our setting.

"Some connections between this work and multi-task representation learning as in Maurer et al JMLR 2016 as c_i could maybe be seen as a representation of the data". Representation learning in the SPO framework is a great idea! Actually c_i's are cost vectors and are more closely related to *labels (or outputs)* rather than *examples (or inputs)* in standard supervised learning. Representation learning in the sense of Maurer et al learns the representation of *examples* and would definitely make sense if we had several related SPO problems (e.g., shortest path for different cities) to solve. Your comment reinforces our conviction that our work has much to offer to the ML community in terms of fruitful follow-up directions!