We thank the reviewers for the thought-invoking questions and helpful comments on improving the manuscript. We will add more text to improve readability and, if necessary, remove some of the technical lemmas to make room. We also have extended our empirical proof of maximal informativeness to $k = 15$.

R1, R2, & R3: The LLW hinge loss is calibrated with respect to the 0-1 loss while the WW hinge loss is not. Why does the WW SVM still outperform the LLW SVM? In other words, how can calibration be used as a justification for performance?

A: The LLW SVM performs worse for a reason unrelated to calibration. Doğan et al. [2016] makes the distinction between *relative* and *absolute* margin losses. The WW and CS loss are both based on relative margins, while LLW is based on absolute margins. Doğan et al. [2016] on their page 20 gave an explanation for the worse performance of all losses based on absolute margin. Hence, the poor performance of LLW is a consequence of using absolute margin. Out of the nine SVM losses considered by Doğan et al. [2016], only the CS and WW losses are relative margin based. We will add this discussion to our manuscript.

R2, R3 & R4: Why is consistency with respect to the ordered partition loss desirable? What is the intuition?

A: Regarding the intuition behind the ordered partition loss, the basic idea is that we want to rank the labels, where ties are allowed and each S_i is a set of labels that are tied. We want the correct label to be as high up the ranking as possible. The lower the true class is ranked, the larger the loss. That is what the definition of the ordered partition loss says. We should have said this in the initial draft and will add this discussion to our manuscript.

R1 & R4: How to get the surrogate decision function ψ to recover the ordered partition/buckets? Is there an excess risk bound?

A: In line 125 of our manuscript, we cited Finocchiaro et al. [2019] who provided an explicit ψ given L, ℓ, and φ. We refer to [Finocchiaro et al., 2019, Definition 6] for the construction of ψ. The excess risk bound for their constructed ψ can be found in [Finocchiaro et al., 2019, Theorem 6]. We will make the theorem references explicit in the manuscript.

R2: What are the consequences of the maximally informative property?

A: Intuitively, a discrete loss $\ell : \mathcal{R} \to \mathbb{R}^k_+$ (where $\mathcal{R}$ is finite) with embedding φ (an injection into the domain of L) is maximally informative for a surrogate L if $\varphi(\mathcal{R})$ captures all the essential information contained in the surrogate L in the most compact way. To better convey this intuition, we replace "maximally informative" with the new terminology "minimally emblematic." Let us say that a set of vectors $E \subseteq \mathbb{R}^k$ is an *emblem* of L if for all $p \in \Delta^k$, the set $E \cap \operatorname{argmin}_v \langle p, L(v) \rangle$ is nonempty. Then we can equivalently define ℓ with φ to be *minimally emblematic* for L if $\varphi(\mathcal{R})$ is an emblem of L of minimal cardinality. In other words, $\varphi(\mathcal{R})$ is a minimal set of minimizers of all possible L-inner risks. We will update our manuscript with this discussion and the new terminology.

R3: How does performance over the ordered partition loss translate to the 0-1 loss?

A: Results from our section on the "argmax link" provide a partial answer to this. Namely, we show in two common regimes, the Bayes optimal ordered partition has a top bucket consisting of a single element. When this occurs, the argmax link recovers the most probable class, i.e., the unique element from the top bucket. We will modify the manuscript to clarify this point.

References

U. Doğan, T. Glasmachers, and C. Igel. A unified view on multi-class support vector classification. *The Journal of Machine Learning Research*, 17(1):1550–1831, 2016.

J. Finocchiaro, R. Frongillo, and B. Waggoner. An embedding framework for consistent polyhedral surrogates. In *Advances in neural information processing systems*, pages 10780–10790, 2019.