Foreword

We would like to thank the reviewers for their valuable feedback and suggested references. Two of the reviewers raised the question of whether the use of the cross-entropy loss is justified, and we would like to address this point. First, Rankmax can be thought of as an adaptive projection method that is not tied to a particular loss function, and it can be used with other losses such as Fenchel-Young losses or differentiable approximations of top-k losses – we thank the reviewers for these references. On the other hand, we focused on the cross-entropy loss because it is commonly used with the closely related Softmax projection. Moreover, we have shown that in the cross-entropy setting, there are useful connections between Rankmax and other rank losses such as OWA and pairwise losses (see Section 3.3 and Section 5). Note that our purpose is not to make a statement about which loss is better suited to multi-label classification, but rather to show that adaptivity of the Rankmax projection can accelerate training and improve model quality.

Detailed answers

(R1–R4) We will correct all typos, missing citations, spelling mistakes, and minor technical errors in the final revision.

(R1) ***Mathematical notation and writing.*** Section 2 and 3 follow the notational conventions of the field of optimization. Nevertheless, we tried to make the core findings and claims of the paper accessible to a wider audience.

(R2) ***Connections to top-k classification.*** Thank you for the suggestions. We will expand our discussion of the connection between Rankmax and top-k classification, both in Section 3 and when discussing the gradient properties. In particular, it can be shown that the magnitude of the gradient under OWA only depends on the number of negatives above the margin, while it is adaptive under Rankmax (it depends on the distribution of negative scores). This gives another interpretation of adaptivity. We will give specific examples in which the magnitudes can vary significantly.

(R2) ***Problems with overfitting.*** Based on our experimental results, Rankmax was prone to overfitting on the smaller dataset (as discussed line 269), but this can be remedied using early stopping on a cross-validation set. With early stopping, it outperformed non-adaptive projections. On the larger datasets, we have not observed any overfitting.

(R2) ***Compute time vs. epochs.*** We will add wall-time plots to the supplement. As briefly discussed Line 257, Rankmax and Softmax had the same computational cost per epoch, but Sparsemax was slower.

(R3) ***Projection on the (n, k)-simplex (or capped simplex).*** Thank you for these additional references, we will correct this omission in the revision. For comparison, [1,2] consider special cases of projections on the capped simplex (respectively for Euclidian and entropy regularizers) and [3] considers the projection on the standard simplex. Our result (Theorem 2) can be viewed as a generalization of both. The permutahedron projection in [4] can indeed be applied to the capped simplex, though the result in Theorem 2 is much more direct to obtain, and easier to interpret and to implement (as it is more specialized). We will add a detailed discussion in the related work section.

(R3) ***Fenchel-Young (FY) losses.*** Thank you for bringing this important work to our attention. We will add a discussion of FY losses in Section 3. As discussed above, though we focused the presentation on cross-entropy, the Rankmax projection can be used with other losses. Combining the adaptive projection of Rankmax with FY losses is an interesting direction for future work. Specifically, the FY loss with regularizer αg, label y and score vector z, can be written as $L(z, y) = f_z(y) - f_z(p_\alpha(z))$ where $p_\alpha(z) = \arg\min_p -\langle z, p\rangle + \alpha g(p)$ is the projection defined in Eq. (1) of our paper. It is therefore possible to apply the same loss with adaptive α.

(R4) ***Relevance of projecting onto the (n,k)-simplex.*** The adaptivity of Rankmax projection is the central contribution in the paper. We developed our framework in the more general context of projecting onto the (n, k)-simplex because it was found useful in other studies (see [1,5,6]). However, Rankmax provides benefits even when $k = 1$, as shown in our numerical experiments.

(R4) ***Differentiable approximations of Precision@k.*** Thank you for the suggestion, we will add in Section 3 a discussion of differentiable approximations to top-k metrics. As discussed above, Rankmax is an adaptive projection that can be used with any loss function, including such approximations. We focused on the cross-entropy loss as it enjoys additional properties and connections with other losses.

[1] Amos, Koltun, and Kolter. The Limited Multi-Label Projection Layer, arXiv:1906.08707, 2019.
[2] Wang and Lu. Projection onto the Capped Simplex, arXiv:1503.01002, 2015.
[3] Blondel, Martins, and Niculae. Learning Classifiers with Fenchel-Young Losses: Generalized Entropies, Margins, and Algorithms. AISTATS, 2019.
[4] Lim and Wright. Efficient Bregman Projections onto the Permutahedron and Related Polytopes. AISTATS, 2016.
[5] Lapin, Hein, and Schiele. Top-k multiclass SVM. Advances in Neural Information Processing Systems, 2015.
[6] Lapin, Hein, and Schiele. Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification. IEEE transactions on pattern analysis and machine intelligence, 2017.