We first would like to thank all reviewers for their reviews and constructive comments. We updated the paper to take into account the suggestions and corrections that were proposed: we

- (obviously) corrected typos and other minor formulation errors
- expanded on the related work and discussions, including adding references suggested by R1, R2, R4
- clarified section 2.1 following our answer to R2's questions
- added graphical representations of connected/disconnected sublevel sets to complement Figure 1, with a table of formula of different ranking losses and associated utilities if any, following R3's suggestion

We give more details on some discussion points below.

R1: "2. [...] it would imply that the CCDim of the loss function is equal to the rank of the loss matrix." Yes. In fact, the symmetry assumption in our definition of a ranking loss (" Items are equivalent a priori" in Definition 3) implies that ranking losses satisfy the assumptions of Theorem 18 in Ramaswamy & Agarwal [22]. So for a convex calibrated loss, we do have CCdim $\geq$ affdim(L) - 1, which is what Theorem 7 of Agarwal & Agarwal would give. Note that these two theorems use a definition of "calibration" that does not use argsort as the inference procedure. In their case, inference may be intractable. Thanks for this remark, we will add it.

R1: "does a convex calibrated surrogate in a given dimension exist if and only if there is a squared loss that is consistent for that dimension?" Indeed, it would be interesting to extend our analysis to higher dimensions, for fixed "interesting" inference procedures other than argsort. Note that we need to focus on fixed inference schemes: If we accept possibly intractable inference procedures, the approach of Ramaswamy & Agarwal (2012), based on decomposing the loss matrix, works to define calibrated square losses in any dimension.

R2: "the loss L takes a tuple (Y,pi) as input, where pi is a predicted ranking. Normally, a loss compares a prediction with the corresponding ground truth, but it seems a supervision signal is not a ground truth ranking." We agree with the reviewer that in most supervised learning tasks, the supervision is a ground truth in prediction space (a ranking in our case). Yet, in many practical ranking tasks, the supervision is not a complete ranking. For instance, in search engines, the task is to rank documents in response for a query. A typical setup is when annotators give binary relevance judgments to each document given a query. The set of relevance judgments does not define a full ranking, because it does not specify the relative order of two documents with the same relevance. By decoupling the supervision space $\mathcal{Y}$ from the prediction space $\mathfrak{S}_n$, our framework is more general than a standard supervised learning framework since it allows for $\mathcal{Y} = \mathfrak{S}_n$, but also for other supervisions such as relevance judgements.

R2: ""In recommender systems or search engines, this means that the score of an item depends on the other available items" –> is this consistent with defining a utility function on individual items?" For a utility function, we can say *the input of the utility function is the entire supervision (e.g., all relevance judgments for all items to rank), and it computes jointly the utility values for all items*. The sentence quoted by the reviewer makes the analogous statement for scoring functions: *the input of the scoring function are the features of all items to rank, and it jointly computes the scores of all items*. These are consistent: there is a utility value per item on one side, and one score per item on the other side.