We thank the reviewers for their thorough and positive reviews. Overall, we were glad to see the reviewers found the paper to be clearly written, technically sound, and the results to be of interest to the (fair) ML community.

We will of course incorporate all the suggested edits by the reviewers as well as more clarifications.

General comments:

Regarding our lower bound: Our lower bound holds both for finite predictor classes and infinite predictor classes with finite graph dimension. We will restate the theorem statement so it would state precisely what is proven. Indeed, it remains to be an open problem whether it can be improved, and in particular exhibit a dependency on graph dimension, which we assume is necessary.

As for other complexity measures such as fat-shattering and margin classifiers, it could be an interesting direction for future work. In this paper, we chose to derive the generalization bounds using Graph dimension and VC-dimension.

Reviewer 1: Thank you for your positive review. The reviewer expressed concern regarding the brevity of the discussion about the potential societal impacts of the results. We will address these important issues in detail in the final version.

Reviewer 2: Thanks for your careful reading and constructive questions. We address your concerns as follows:

1. The upper bounds of [Hebert-Johnson et al.] are implicitly guaranteed only for "interesting categories" with large enough mass.
2. *Dependence on ψ:* In the case of finite predictor classes, we applied union-bound directly on all the (large enough) "interesting categories", that have finite cardinality due to the finite number of hypotheses. In the case of infinite predictor classes, this technique is no longer possible. To derive the concentration bound of $c(h, U, v)$ we bound the deviations of $p_1(h, U, v)$ and $p_2(h, U, v)$ (see line 592, in the proof of Theorem 10 in the supplementary material). To achieve our desired accuracy, we need to bound the deviation by $\psi\epsilon/3$ (see Lemma 19, line 331). This increased accuracy translates to the ψ^{-2} in the sample size.
3. *Instance optimality:* Your intuition is correct, there are some problem instances with smaller VCs for all the values of v that determines the "intresting categories", thus their sample complexity is lower. On the other hand, the bound is tight in the worst case. Namely, for any graph dimension d, we can consider a predictor classes with $VC(\mathcal{H}_v) = VC(\Phi_{\mathcal{H}_V}) = d$ for any prediction value v.
4. *Applicability of the proposed techniques:* Our techniques are suitable to derive generalization bounds for complex measures such as F-score. Specifically for F-score, one can derive generalization bounds, assuming that at least one of TP, FP, or FN is "not negligible".
5. *Objectives which contain both a loss component as well as a calibration error component* Our generalization bounds extend to such settings. For example, for zero-one loss, we can guarantee simultaneously for each predictor in the class that all the empirical multicalibraion errors of "interesting categroies" and the empirical loss are close to their expectations. This does not even require an increase in the sample complexity.

Reviewer 3: We thank the reviewer for their positive and constructive review. As for multi-category pattern classification, it can be reduced to binary classification, which we address in this work.

Reviewer 4: Thank you for diving deeply into the paper and for your detailed comments and corrections. Sure, one can choose α to be the discretization scale, but this is a limitation we wanted to avoid in order to give more flexibility.