The thorough reviews asked numerous questions and suggested great clarifications/improvements. Due to space limitations, we are not able to respond to all of them, but we will incorporate all the feedback into the next version.

Reviewer 1. You make an excellent point that the skin-tone face example is a poor motivating example since one should collect a diverse set of labeled faces for training. We will give this as an example of how *not* to use our algorithm, and we are so grateful that you identified this issue. We will also note that research on PQ algorithms may help raise awareness that P and Q may differ. We will switch to the following motivating natural example:

> One wishes to provide a service classifying medical scans as normal or abnormal. Training data consists of a set of volunteered examples hand-labeled by multiple radiologists over a period of time. Test examples are to be classified in large daily batches. A concerning distribution shift may occur due a new disease, e.g. COVID-19 scans show "the presence of bilateral nodular and peripheral ground glass opacities" which may not occur in labeled training data [Sawani, 2020]. If there aren't enough unlabeled test examples at classification time, periodic rejections may be useful *in hindsight*, e.g., it could be immensely helpful if a machine can recognize a new disease (or a problematic change in the scanning pipeline) at the end of a week.

About the question of when one cannot simply label a sample of test examples: note that large-scale classification services may have millions of test examples per hour, but it may be impractical or even unethical for human annotators to label them. For instance, privacy policies may prohibit random private Twitter or SnapChat messages from being classified by humans, though one may have curated messages (e.g., public tweets) at train time. One may simply not know how representative the training examples are of the private messages, and one would prefer to abstain then misclassify at test (deploy) time.

About the concern of whether or not the adversary can choose examples based on S, in fact, in our model S is a given deterministic function of the training data and unlabeled test examples provided by the adversary. Since the adversary is all-knowing and all-powerful (e.g., knows the training data), they can determine exactly what S will be based on their chosen test set. Hence S is effectively known. Forcing the adversary to provide a batch of test examples avoids the need for an "arms race" as the classifier always wins the game.

In our empirical evaluation, the simple "baseline" worked really well on our toy example, illustrating our fundamental contribution: *PQ learning is possible but provably requires unlabeled test data*. That is why we did not implement any of our more advanced algorithms, thus their value remains theoretical at this point. Future work evaluating and designing new PQ algorithms on real-world datasets is indeed of interest.

Finally, thank you for referring us to the work on "unrestricted adversarial examples", we will add this to the related work in the updated version.

Reviewer 2. Thank you for the enthusiasm and presentation suggestions. We will try to clarify the transductive setting. As you suggest, one may think of a one-to-one map between examples in $\tilde{x}$ and z: one way to think about it is that every example in $\tilde{x}$ is the result of a manipulation of some example in z. To see why our transductive guarantees are about z and not P, consider a case where $P = Q$ is uniform on a huge set X. It would be unreasonable to reject the entire test set $\tilde{x}$ (and thus achieve zero test error), though if this is all you rejected you would have a negligible rejection rate on P. Regarding L159, It is true that the notion of arbitrary test examples is more general than the notion of adversarial examples with some specific perturbation set (e.g. ℓ_p perturbations). However, in this paper we do selective classification, which is unlike work on adversarial examples where a prediction is usually always required (even on perturbations). Also, thank you for referring us to the line of work on adversarial examples that uses unlabeled data, we will cite these papers in the updated version. We will replace the ? symbol with a different symbol, and we will address your many other questions though space prohibits us from commenting on them here.

Reviewer 3. Thank you as well for your enthusiasm and suggestions. As you suggest, the rejection rate bounds can be improved beyond statistical distance. In Appendix B, we give an example where a tighter bound of 0.1 can be shown while the statistical distance is 0.91, and Equation (2) also gives a refinement. We agree that it would be especially interesting to consider better bounds specific to the given concept class C.

As the reviewers suggest, this work is clearly not the final word on learning with arbitrary adversarial examples but we hope the theoretical results on the possibility of PQ learning inspire people to work in this area.

References

Jina Sawani. How does covid 19 appear in the lungs? *U of M Health Blog*, March 2 2020.