Thanks for your comments and helpful suggestions! We begin with responses that might interest multiple reviewers.

REVIEWER 4: Framework requires knowledge of P, but in many practical settings it is difficult/impossible to know P, or control/design P.

We believe that the difficulty of knowing P in some practical settings makes our work *more* applicable, not less, in the following sense. Suppose someone conducts a poll or survey of students and then makes a claim like “I estimate the overall student body mean to be 74 based on the empirical mean of the respondents". No one truly knows the underlying sampling distribution, P (different students have different response probabilities, there might be some independence, but also likely correlations in participation between friends, etc.). Using our framework and algorithm, you could design a variety of plausible sampling distributions, P, and then evaluate the worst-case expected error of the empirical mean with respect to these P's, which would provide compelling evidence for whether or not to believe the claim about the student body mean being 74. [Recall Prop 1 in Sec 2.1 that argues that we can evaluate the worst-case-expected-error of a given/fixed (semi-linear) estimator such as the empirical mean, with respect to these “plausible" P's.]

In this sense, our framework and algorithm can be used to rigorously evaluate the stability of an estimator (in the above example, the empirical mean) with respect to various “plausible" P's. We think this is a useful alternative to more standard approaches that evaluate the estimator with respect to other kinds of assumptions on the data. We plan to add a discussion of this to the paper and thank the reviewer for eliciting this alternate use case for our framework.

REVIEWER 2: Extension to ℓ_1 or ℓ_2 constraints instead of ℓ_∞.

Good question. Our original motivation was for settings such as surveys/polling (e.g. COVID testing) where data values are binary or otherwise bounded in magnitude, and where ℓ_∞ is the natural constraint. We agree that investigating our model with respect to other constraints seems natural and worthwhile. We have done some preliminary work investigating the ℓ_2 case (not included in our submission): since the geometry of the ℓ_2 norm is so well behaved, there is potential for efficient algorithms that might even improve upon the $\pi/2$ approximation factor and surmount several of the more structural hurdles we encountered when investigating fully general (non semi-linear) estimators. We haven't fleshed out any details, though we agree this is a very interesting direction.

REVIEWER 4: Underlying domain is assumed to be finite (ie size n).

Our framework and definition of worst-case expected error apply, without modification, to infinite sets of underlying datapoints. Additionally, it seems like an extension of our algorithm might be adapted to such settings. To briefly sketch this, suppose P corresponds to a joint distribution over an infinite (or continuous) domain indexing potential elements, and that each sample/target set in the support of P has size at most k. Very roughly, one can (1) draw many (i.e. $poly(k)$) sample/target sets from P, let Z denote the union of these sampled sets (together with the $\leq 2k$ elements of A, B–the actual sample/target sets) and then (2) let P' denote the restriction of P to those sample/target sets that have non-empty intersections with Z, and (3) run our algorithm on this (now finite) set Z and distribution P'. The one delicate step (that would take some space to fully describe) is that one must adjust P' to account for the possibility that the measure in P that intersects Z in 1 point, might be (infinitely) larger than the measure intersecting in 2 points, etc.

This extension of our algorithm to infinite/continuously indexed data is certainly not trivial, and we haven't fleshed out a formal proof of this. It does seem quite interesting—thanks for pointing out this direction.

REVIEWER 4: “Results mainly apply to scalar sets"

We mainly focus on scalar sets, as that seems like the natural starting point for explaining/exploring this framework. Still, as our linear regression results (Thm 2) illustrate, the framework naturally extends to non-scalar settings, and one can obtain interesting results in these non-scalar settings. We imagine that future work will likely explore a number of different non-scalar settings within the framework we propose.

REVIEWER 3: "try to compare this model with the common case where we have a distribution D on the population and aim to minimize the statistic over D. Are there cases where a bound on the benchmark you propose implies a bound on the error of the estimator of expected value of the statistic over D?"

We aren't sure we understand this question. Our interpretation of your question is the following: Suppose we have a distribution D (ie over all images) and the ultimate goal is to train a model to classify, say, cat images, that has small expected loss wrt D. Given a model, how do we evaluate the expected loss over D? If we can sample from D, then great. If we can't sample from D, but instead can generate some (possibly dependent) set of samples S, drawn from a joint distribution D' over sets of $k = |S|$ samples, then we *could* use our framework to figure out how to estimate the expected loss over D based on the values of the loss on set S. [One natural example of such a D' might correspond to taking a sequence of k samples from a Markov Process whose stationary distribution is D...] We not sure if there are any ‘standard' instances of such settings for which there are clean results to which we could compare with. [If this doesn't address your question, please do clarify your question in your review...]