We thank the reviewers for their detailed comments and thoughtful feedback.

Following your feedback, we will work on making the paper easier to read and especially making it more self-contained: we will include a full and self-contained proof of Theorem 3.1 (instead of relying on [MHS19]), add a description of α-Boost and lemmas stating its properties, and a lemma for the sparsification technique due to Moran-Yehudayoff.

Below we address other key concerns that were raised:

Reviewer 1: This paper studies what guarantees can we (theoretically) hope for in adversarially robust learning when all we have access to is a black-box non-robust learner. This question helps us formulate two settings (Section 3 and Section 4). Formally defining these settings can be helpful for practice, in the sense that they progress us towards the non-trivial task of formulating the "right" access we need to allow/assume/use on perturbation sets. Such progress can help us handle robust learning beyond ℓ_p perturbations, and obtain generic wrappers for non-robust learning for generic perturbation sets. For instance, the mistake oracle model presented in Section 4 captures what's referred to in practice as adversarial training, and can be applied for generic perturbation sets beyond ℓ_p perturbations.

In the statement of Theorem 4.3 (lines 317-320), we do explicitly specify the sample/oracle complexity of the resulting robust learner, in terms of the mistake bound of the black-box learner we are reducing to (as in any reduction, the guarantee on the wrapper is in terms of the guarantee of the method being wrapped–we hope this is clear). But we agree it will be useful to include a concrete example of how this can be instantiated, and will do so.

Reviewer 2: Thank you for pointing us to the related work papers, we will add these citations to the paper. We briefly discuss the agnostic setting in lines 335-342, and mention that the techniques in [MHS19] give us a reduction with an exponential runtime dependence on VC dimension. Our results can also be extended to the Massart noise model. This can be done with roughly similar sample/oracle complexity as in the realizable setting.

Sampling Oracle Over Perturbations: We need an oracle that takes as input a point x and a function $E : \mathcal{X} \to \mathbb{R}$, and does the following:

1. Samples a perturbation z from a distribution given by $p_x(z) \propto \exp(E(z)) * \mathbb{1}[z \in \mathcal{U}(x)]$. That is, the oracle samples from the set $\mathcal{U}(x)$ based on the weighting encoded in E.
2. Calculates $\Pr[z \in \mathcal{U}(x)]$ for the distribution given by $p(z) \propto \exp(E(z))$.

Using such an oracle suffices to sample from distributions over the inflated set $S_{\mathcal{U}}$ that are constructed by α-Boost in Algorithm 1 and its subprocedure `ZeroRobustLoss`. We will update the sampling paragraph (lines 287-295) in the paper with a more elaborate/formal explanation of these details, and we will also address your other comments/corrections.

Reviewer 3: First, we think that reductions with a $\log|\mathcal{U}|$ dependence could, sometimes, be reasonable, as pointed out by Reviewer 2. Second, since we agree that in many cases we want to avoid the $\log|\mathcal{U}|$ dependence, our work does indeed indicate that this means moving away from reductions based on an explicit $\mathcal{U}$ oracle, which is exactly what we do in Section 4 (using the mistake oracle model). We view this as a significant contribution of our work, since this isn't obvious (at least to us) a-priori, or from [MHS19], and it progresses us towards the non-trivial task of formulating the "right" access we need to allow/assume/use. Although the lower bound holds for aggregation reductions, a cheating reduction algorithm that ignores the non-robust learners can't generalize in general, and we will make this formal in the paper. Regarding the proof of the lower bound, your intuition and understanding is correct. We will add an informal description of the strategy before the formal proof as suggested. In the definition of M_1, there is a missing multiplicative factor of $\Pr_{(z,y)\sim D_t}[z \notin P^t \wedge y = 0]$, and M_2 should be $(1-\varepsilon) - M_1$. Thank you for pointing out these typos. Basically, M_1 is the fraction of examples correctly classified under distribution D_t by predictor h_{t-1}. If M_1 is high enough (at least $1-\varepsilon$), then it suffices to return h_{t-1}, if not, we shift the threshold as little as possible based on M_2.

Reviewer 4: While true that the oracle complexity can be exponential in the VC dimension, in lines 107–109 we discuss how for many natural classes, the dual VC dimension is polynomially related to the VC dimension. This holds beyond just linear predictors, and is true also, e.g., for neural networks with threshold activation functions. This can be shown using Lemma 3.2 which is a novel contribution in this paper. We briefly discuss limitations/challenges of extending our results to the agnostic setting in lines 335-342. Our results can also be extended to the Massart noise model. This can be done with roughly similar sample/oracle complexity as in the realizable setting. Regarding line 233, yes, it does NOT mean that there is no hope for robust learning in general. We are just showing that $\log|\mathcal{U}|$ oracle calls are unavoidable without extra assumptions on the non-robust learner.