We thank the reviewers for their careful reading of the paper. Please find the answers to the questions below.

Review #1

Q: The techniques in this paper are somewhat specific to halfspaces which are very simple hypotheses (adversarial robustness does not correspond to a standard L_p margin for other concept classes). But that being said, it's important to understand these basic hypothesis classes first.

A: As the reviewer states, halfspaces form one of the most basic and fundamental hypothesis class in machine learning, and we believe that obtaining a complete understanding of adversarial robustness for halfspaces is an important contribution. Furthermore, the hardness results for such a basic concept class might also be a good indication that the problem is hard for other, more complicated, classes too.

Q: I wonder if you could use these arguments to show guarantees based on the perceptron algorithm itself. The perceptron analysis shows that there is a combination of $O(1/\gamma^2)$ of the examples that gives a good classifier. I wonder if one can also show a similar statement with the margin error, and guess the examples in the combination.

A: The reviewer's suggestion is correct; a slight modification of the perceptron algorithm, where an update is performed whenever a sample is not correctly classified with margin $(1-\nu)\gamma$, is known to be an L_2 online learner with margin gap $(\gamma, (1-\nu)\gamma)$ and mistake bound $O\left(\frac{1}{\nu^2\gamma^2}\right)$. When plugging this so-called "margin perceptron" algorithm to our reduction (Proposition 5), this recovers our theorem in the case $p = 2$. Indeed, the more general algorithm of [Gen01a] that we invoke (Theorem 7) can be viewed as a slight modification of margin perceptron in the case $p = 2$.

Review #2

Q: I find the hardness result a bit limited in that it only show that the dependence in terms of d, γ is tight in the case L_∞.

A: As we briefly mentioned in the paper, the (essentially) tight running time lower bound of the form $2^{\gamma^{1-o(1)}}$ for L_p perturbations for any constant $p \geq 2$ follows already from [DKM19]. Specifically, [DKM19] proved such a result for $p = 2$. To get a similar lower bound for other constants $p \geq 2$, we simply take a random rotation of every sample $\mathbf{x}$ (of the hard instance for L_2 perturbation) and rescale so that it has unit L_p norm while keeping the label the same as before. (The optimal halfspace is also rotated and scaled so that it has unit L_q norm.) It is not hard to see that this preserves the margin for most of the samples up to a constant factor. We will add more detail about this in the revised version.

Q: The result is stated for a "small constant" $\nu > 0$. Perhaps it might help to say how small ν needs to be for the result to hold. For example, would $\nu = 0.1$ work?

A: We agree that obtaining a concrete value of ν is interesting; in fact, this is included in our "Additional Open Questions" in the supplementary material. For our current proof, ν is selected to be very tiny ($\approx 10^{-16}$) for simplicity of presentation. Per rough estimates, we can have $\nu \approx 10^{-4}$ but the bounds in the proof become more delicate.

Review #3

Q: The algorithmic upper bound seems to be a fairly straightforward application of existing halfspace learning algorithms; the main contribution here is realizing that the existing online learners are enough.

A: We view simplicity as an advantage of our work. Further, given that many works have studied the problem and that the online learning results have existed for a couple of decades, we believe that it is not straightforward.

Q: There are a number of other papers giving adversarially robust learning guarantees for halfspaces, including (non-agnostic) learning for L_p perturbations with random classification noise, and semi-agnostic learning for L_2.

A: As the reviewer points out, this is an active research area. Our results complement the existing works mentioned, which only apply to more restricted noise models. As such, we do not view this point as a weakness of our work.

Review #4

Q: Nevertheless, given the amount of work on closely related problems, this work is in some ways a little incremental. Also, one might argue that a stochastic noise model is often of more relevance in practice than the highly pessimistic agnostic noise model, where much stronger guarantees are generally possible for stochastic noise. (I feel that the agnostic model is still worthy of study and more relevant for some scenarios, though.)

A: While the stochastic noise model might be more realistic, the existing works often assume random classification noise where each label is flipped independently with probably *exactly* $\eta < 0.5$. Known algorithms in this model do not extend naturally even to the case where the flip probability is *at most* η (aka Massart Noise); such a limitation calls their practicality into question. On the other hand, our algorithms work in the most general agnostic noise model, which can be applied without strong assumptions about the specific random process creating the noise.