We thank the reviewers for their careful consideration of our paper and their uniformly positive feedback. We will incorporate all minor comments and typos in the final version of our paper. Below we address specific questions and comments by the reviewers.

Reviewer 1

Our paper establishes SQ lower bounds against weakly learning an unknown function in the given class. As correctly pointed out by the reviewer, our lower bound construction produces instances in which OPT is close to $1/2$. Proving SQ lower bounds for the case that OPT is a small constant is left as an interesting open question that may require additional ideas.

The reviewer is correct that our lower bound works for the (broader) class of SQ algorithms – as opposed to only CSQ – even for ReLUs, essentially because the class of hard functions are boolean-valued. We will clarify this point in the final version.

We will add the definition of the unsupervised SQ dimension in our revised version, as well as the definition of χ^2-divergence.

We will add intuition regarding the notion of correlation. Intuitively, one can always think correlation as a metric of the closeness of two functions. For boolean-valued functions, correlation is closely related to the probability the functions disagree/agree. So, finding a function with a high correlation is the same as finding one with small error.

Reviewer 2

As pointed out by the reviewer and explained in the discussion section of our paper, our SQ lower bounds are *qualitatively* optimal, up to a degree of the polynomial in the exponent. In particular, we prove an SQ lower bound of $d^{\Omega(1/\epsilon)}$ for agnostically learning LTFs (under Gaussian marginals). In comparison, the best known algorithm for the problem has runtime $d^{O(1/\epsilon^2)}$; and the best previous lower bound was $d^{\Omega(\log 1/\epsilon)}$.

It remains an interesting open question for future work whether $d^{\Omega(1/\epsilon^2)}$ is an SQ lower bound for the problem.

Typos/Definition: We will revise line 42 and add the definition of a k-decision list in the final version.

Regarding the potential existence of faster PTAS for the problem: Suppose there was an algorithm for agnostically learning LTFs under Gaussian marginals that achieves error $(1+\gamma)\mathrm{OPT}+\epsilon$ and runs in time $\mathrm{poly}(d^{\log(1/\gamma)}, 1/\epsilon)$. Then, by setting $\gamma = \epsilon$ and using the fact that OPT is at most 1, we would obtain an algorithm with error $\mathrm{OPT}+2\epsilon$ that runs in quasi-polynomial time (as a function of $1/\epsilon$).

Proof of Proposition 3.3: We will add a clarification to line 274. In the proof, we need the breakpoints to be distinct, (as is stated in the beginning of the proof). In the proof of Proposition 3.3., we work with functions that have at most $k+1$ breakpoints, which means that for some ϵ it is possible that for some values of i, b_i and b_{i+1} may be equal. Of course, the compactness argument shows the existence of such a function and indeed it may have less than $k+1$ breakpoints, which will in fact yield a function with higher correlation. (Note for example that, if we knew that the function has $\sqrt{k}$ breakpoints, we could improve the lower bound to $d^{\Omega(1/\epsilon^2)}$.)

Reviewer 3

Thank you for pointing out this concurrent related work. We will add a paragraph with a comparison to our results and techniques in the revised version of our paper.

Thank you for the detailed technical comments. We will address them in the final version. We will also add prose providing the intuition behind Lemma 3.8.

Reviewer 4

Thank you for pointing out these typos. We will fix them in the final version and add the definition of χ^2-divergence.