We thank all the reviewers for their comments.

Responses to Reviewer-2's comments:

"... would prefer if the authors mention more clearly that their results are significant only in the agnostic setting.... "

Indeed. We mentioned that our work focuses on the agnostic setting in several places (including the abstract and the introduction), but we will elaborate more on this point as suggested by Reviewer 2.

" Is there [..] intuitive way to explain why there is such a discontinuity at public sample size $1/\alpha$?"

Here is one way to think about this: this kind of sharp transition is a by-product of the fact that the definition of PAC learnability is a worst-case (min-max style) definition. Similar discontinuities are also exhibited by standard (non-private) PAC sample complexity bounds: for example, a class is either learnable with $O(VC(H)/\alpha^2)$ examples, or it is not learnable at all (if $VC(H) = \infty$).

"Is there any way the lower bound on the public sample size to become $VC(H)/\alpha$ instead of $1/\alpha$? ... I would suggest the authors to mention whether this is a hard next research step or not."

This is a very good question. Although it is natural to think that the upper bound should be tight, it is not immediately obvious, at least for general VC classes, how to involve this factor of $VC(H)$ in the lower bound. We believe this to be an interesting research question.

"what does the term $negl(n_{priv})$ mean in Definition 2.3?"

This means it is a negligible function of n_{priv}. The function $negl(.)$ is formally defined earlier in the first paragraph of Section 2.

"In Algorithm 1, step 5: By "add to $\tilde{H}$ arbitrary h.." do you mean "add to $\tilde{H}$ every h.." or "add to $\tilde{H}$ one h arbitrarily chosen.." ? I suspect the former but it is not clear."

It is the latter. To construct the α-cover, one only needs one representative hypothesis (chosen arbitrarily) for each dichotomy. We will rephrase this step to make it entirely clear.

Response to Reviewer-3's comments:

"For the lower bound, it seems not very complete. Authors show that if a concept can't be pure privately learned, then any semi-private learner must have $\Omega(1/\alpha)$ public samples. So there is a problem, does a non-trivial semi-private learner for this concept always exist? Non-trivial means that the learner doesn't learn only from the public data, otherwise, there is no privacy issue in this learning. If a concept can't be semi-privately learned nontrivially, then the lower bound has no sense. Recall that they show an algorithm for semi-privately learning a concept with finite VC dimension, then whether there is a semi-private algorithm for infinite VC dimension, this is not clear."

We have not been able to understand the comment. If the VC-dimension is infinite, then learning is impossible, even ignoring any privacy issues. On the other hand, if the Littlestone dimension is infinite, then private leaning is impossible. Thus the lower bound is interesting mainly when the VC-dimension is finite and the Littlestone dimension is infinite. In this case our positive result shows that a non-trivial semi-private learner always exists, indeed the learner needs only VC/α public examples and hence does not learn only from the public examples as altogether VC/α^2 examples are needed for learning in the general agnostic setting, which is the setting we focus on in this work. Our lower bound shows that the dependence on α in the number of public examples for this non-trivial semi-private learner is tight.

"There are some typos and expressions can be fixed: Line 82, it should be VC/α, rather than VC/α^2"

This is not a typo. What we are saying here is that constructing an α-cover using VC/α^2 examples is rather straightforward using standard uniform convergence arguments. Hence, a construction (like ours) that involves only VC/α public examples is non-trivial.

"Line 270: [This implies that the total variation between $\hat{S}$ and S is at most 0.01.] This sentence is confusing. The above inequality means that the probability of $\hat{S} \neq S$ is at most 0.01. *How does the total variation mean here?"*

This follows from the sequence of steps before that line. We are happy to elaborate and will include this clarification in the paper. First, note that the distribution of the examples in $\hat{S}_{pub}$ is a mixture of two distributions $b \cdot D + (1-b) \cdot D_0$, where D is the original distribution (realizable by H), and D_0 is the distribution of the examples in S_{pub}. Second, note that the probability that $\hat{S}_{pub} \neq S_{pub}$ is an upper bound on the measure attributed to the first component of the mixture distribution of $\hat{S}_{pub}$ (i.e., the component from D). Hence, it follows that the total variation between the distribution of $\hat{S}_{pub}$ (induced by the mixture) and the distribution of S_{pub} (induced by D_0) is upper bounded by the aforementioned quantity.