We would like to thank all the reviewers for their feedback, it is very much appreciated. Below, we respond to the chief questions and comments raised, particularly from Reviewer 1.

Data-dependent h Regarding the point raised on whether h can be chosen in a data-dependent fashion, the answer is that this is possible. As the reviewer points out, the basic deviation bounds on the estimator use independence, but these guarantees can be made "uniform" over $h \in \mathcal{H}$ using an argument analogous to McAllester's classical pre-theorems. This is in fact what we are trying to show with our pre-theorem 7; the bound holds with high probability uniformly over the choice of h, meaning it can be selected by a data-dependent procedure. We can definitely adjust the wording around this pre-theorem to reinforce this point. In addition, we shall be more explicit in the text regarding why the classical approach fails with unbounded losses.

Tightness of bounds Regarding the point raised about when the bounds are actually tighter – this is definitely a valid point. Under weak moment assumptions on the loss distribution, while the dependence of deviation bounds on the confidence level is exponentially better than the case of using the empirical mean, this is certainly not the only factor in the bounds. As a simple example, when compared with variance-dependent bounds that depend on $1/\delta$ linearly, if one takes the confidence level low enough (δ large enough), and the second moment is larger than the variance, then one could force such bounds to be numerically smaller than ours. That said, our basic hope is that the proposed approach should really shine at high confidence levels, in particular limiting the risk variance across independent samples when the data may be heavy-tailed. Tightly controlled simulations and numerical tests of precisely when different methods realize superior guarantees are a point of interest that we are pursuing at present. To get things started, a direct comparison of our bound with the Alquier and Guedj bound under some restricted assumptions could be added easily. As well, we will revise the text to be more precise regarding computational cost, including the potential extra cost for setting the scale parameter.

Empirical study Regarding the point about computation, as reviewer 1 says, there is definitely "more to the story," and one of our major goals moving forward is to develop a general-purpose computational strategy to make this robust PAC-Bayes approach practical for a large class of learning problems. As the reviewer points out, there is a great line of recent literature on robust mean estimation in high dimensions. In the context of PAC-Bayes, we are currently seeking a more refined understanding of the tradeoffs between statistical performance guarantees, computational cost, and computational error (i.e., when the estimator cannot be computed exactly) that come up when introducing more sophisticated sub-routines, such as those borne of the recent line of work starting with Lugosi and Mendelson in 2017.

Algorithm parameter settings Regarding the scaling and shifting strategies, in the present paper we have left this mostly abstract, with the implication being that one could plug in empirical moment estimates for the former, and just split the sample in half for the latter, but certainly this strategy is by no means optimal, and for practical purposes, it must be handled very carefully. As we continue with numerical tests in this direction, we plan to flesh out the theory in this direction further, accompanied with some empirical insights to back up our approach. Some comments regarding possible strategies can be added to the manuscript to elucidate this point.

Finally, we would also like to thank all the reviewers for raising various minor points to be corrected. These will all be reflected as we revise the manuscript.