We warmly thank all four reviewers for their careful reading and evaluation of our work, and for their input which helped improve the manuscript. We very much appreciate the overall positive assessment from all reviewers. We provide now individual responses to some of the questions or comments in the reviews.

Reviewer #1. ***"Few comments on computational hardness of the CVaR minimization problem (do we need to consider surrogate risks"*** Minimizing the empirical CVaR is not much harder than minimizing the standard empirical expectation. In fact, the expression of the empirical CVaR in display (3) reveals that one can reduce the problem of minimizing CVaR to that of minimizing an empirical average with an extra real variable—this is the μ variable in display (3).

"build stronger motivation for investigating CVaR; make the derivation of the PAC-Bayesian bound more transparent. Bayesian (even PAC-Bayesian) are not always edible. Make readers life easier." With the extra space provided by the ninth page, we will add more on the motivation behind CVaR and some additional (proof) details in Subsection 4.3 (i.e. the last step in deriving the PAC-Bayesian bound).

Reviewer #2. ***" Starting with the concentration bound of section 5, isn't it possible to get a wider range of generalization bounds, i.e bounds from other frameworks other than the PAC-Bayes framework such as, eg, Rademacher bounds? This part is not discussed."***

"Also, in the same vein... 1) introduce the concentration result and 2) give the PAC Bayes bound."

Starting from concentration inequalities (such as the one in Theorem 11), it is certainly possible to recover generalization bounds either through Rademacher analysis or via the PAC-Bayesian analysis due to McAllester (we discuss the latter possibility in the two paragraphs between the lines 145 and 161). However, starting from Theorem 11, for example, and directly applying such techniques, will yield looser bounds than the one we present in our main Theorem 1 (in the best case, some terms will be off by a "Jensen gap"; an instance of this is described in lines 152 and 153). We avoid this gap in Theorem 1 because we use a bound on the moment generating function of the auxiliary random variable Y—this is Lemma 5. This lemma is stronger than Theorem 11 (in fact, Lemma 5 implies Theorem 11), and so the former leads to a tighter generalization bound. However, we did not want Lemma 5 to take center stage in the story since it is less interpretable; it involves the *implicit* auxiliary random variable Y. We will add a note on this matter in the final version.

"Coherent risk measures are introduced and a bit discussed, but it is not clear how the results provided here cannot carry over to those CRMs in general, the result for CVaR being a specific case."

Our technique relies on a dual property of CVaR, which is not necessarily shared by all CRMs. In particular, there exists a convex function φ such that $\mathrm{CVaR}[X] = \sup_{Q \in \mathcal{B}_\varphi} \mathbb{E}_Q[X]$, where $\mathcal{B}_\varphi$ is a φ-divergence ball. Varying the choice of the convex function φ leads to a rich class of CRMs called *entropic risk measures*. In Appendix B, we explain how our techniques may be transferred to this case, given the structural similarity. However, it is not clear to us how to obtain generalization bounds for all CRMs beyond entropic risk measures. We consider this an exciting direction for future research.

Reviewer #3. ***"Question 1: What is the optimal classifier in binary classification with the CVaR of the 0-1 loss? Is it a variant of the Bayes classifier with the auxiliary random variable Y (from the reduction to expectation)?"***

The optimal classifier in binary classification with CVaR of the 0-1 loss is the "new" SVM—see e.g. the paper "ν-support vector machine as conditional value-at-risk minimization" by Takeda and Masashi 2008.

"Question 2: More practically, could this reduction to expectation trick be used for optimization purpose, by e.g. using the Stochastic Gradient Descent algorithm?" The reduction to the expected risk is only useful in the analysis; note that the random variable Y (as in display (17)) that we introduce in the reduction depends on the "support point" $q_\star$ in the dual formulation of CVaR. This support point is implicit (it depends on the unknown data-generating distribution), and so it is not clear how it can be used for practical optimization purposes.

Reviewer #4. We thank the reviewer for their feedback. We will correct the typos which were identified.