We deeply appreciate constructive and insightful comments from the reviewers.

Reviewer #1

Computing OCE minimizers. On the one hand, we clarify that Theorems 3 and 6 apply to any approximate OCE minimizer, including those given by [40], and Lemma 2 applies to an arbitrary algorithm, even including those not minimizing OCE risks. Our results provide generalization guarantees with the same or tighter dependency for algorithms given by [40]. We will clarify this point in the revised version. On the other hand, we agree to the reviewer's point that clearly stating the optimization-relevant properties of (inverted) OCE risks may facilitate future works in this direction. To this end, we will survey existing results (in the appendix) that are relevant to the optimization of standard risk-sensitive measures, e.g., smoothness/convexity properties of CVaR, entropic risk, and mean-variance.

Prior work on stability & convergence. We believe that R1 refers to (Line 42), where we point to [30]. The work [30] considers an MDP setting that where the standard Bellman's optimality principle cannot handle risk-sensitive cases. By contrast, we focus on nonsequential scenarios where such concerns do not arise. We will clarify this point in a separated "Related Works" section, along with detailed discussions on relevant literature.

Reviewer #2

Apparent lack of tailored analysis for risk-sensitivity. We begin by noting that the risk-sensitivity is indeed distilled into the smoothness characteristics of the disutility $\phi(\cdot)$ in our results, instead of completely disappearing from the analysis. The proposed OCE framework enables a deceptively simple treatment of risk-sensitivity by relating the problem to risk-neutral learning via contraction principles. Such strategy (similar to what [D$^+$19] did for adversarially robust learning) allows us to fully utilize accumulated insights on risk-neutral learning to understand risk-sensitive learning, without having to resort to overly complicated machineries (e.g., adapting Dvoretzky-Kiefer-Wolfowitz inequality to consider data-dependent weights). Given the significance of the considered problem, we view this as an *advantage* brought by our framework, instead of a weakness. Nevertheless, we also provide several "tailored" proof tools, including *product hypothesis space analysis* (Lines 430–438) and *two-sided variance-based characterization of OCE* (Appendix A.5), may be of readers' technical interest. We will highlight these points clearly in the main text.

Vacuity of Rademacher-style analysis for DNNs, and realizability. We thank R2 for pointing this out. While recent progress [S$^+$20, N$^+$20] shed new light on the power of Rademacher-style uniform convergence analysis for DNNs, we agree to the point that the current tone of the manuscript may be too bold. We will revise the manuscript to avoid exemplifying DNNs to make sense of realizability conditions.

Related works and connection to fair ML. Thank you for this valuable suggestion. We will add a standalone "Related Works" section to provide in-depth comparisons with existing statistical learning literature. Also, we will establish explicit connections to the fairness risk measures axiomatically defined in [49].

Reviewer #3

Experiments other than CVaR (points 1&2). Although our main scope is on a theoretical side than an algorithmic side, we agree to the reviewer's point that clarifying optimization properties of OCE risks may help readers grasp the potentials of our framework. For this purpose, we will make the following two revisions: (1) We will give a pointer to the concurrent work of Li et al. [L$^+$20] (which appeared on arXiv after the submission deadline). Via large-scale experiments [L$^+$20], the authors empirically observe that *both entropic risk and its inverse* can be optimized efficiently via standard mini-batch gradient descent, as R3 correctly anticipated. (2) For completeness, we will additionally survey existing theoretical results (in the appendix) that are relevant to the optimization for standard risk-sensitive measures.

Vacuity of Rademacher complexities. We resonate with your concern that any argument regarding the generalization properties of deep neural networks requires a delicate care. We will revise the manuscript to avoid exemplifying neural networks to support the applicability of realizability conditions.

Technical benefits. As R3 keenly points out, our primary focus is to establish an effective theoretical framework to formalize risk-sensitive learning, instead of pursuing technical exquisiteness. Nevertheless, we believe that several proof techniques may be of theoreticians' interest; for details, see Lines 22–24 of this response.

Reviewer #4

Suggestions on discussions and proofs. We express our deepest gratitude for the detailed comments, especially on the ideas to refine our result and the pointer to L-statistics/estimators. All these valuable comments will be incorporated.

Additional References

[D$^+$19] Y. Dong et al. Rademacher complexity for adversarially robust generalization. In *ICML*, 2019.

[L$^+$20] T. Li et al. Tilted empirical risk minimization. arXiv preprint 2007.01162, 2020.

[N$^+$20] J. Negrea et al. In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. In *ICML*, 2020.

[S$^+$20] T. Suzuki et al. Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network. In *ICLR*, 2020.