We thank the reviewers for their encouraging comments and constructive feedback.

Reviewer comment: The empirical distribution estimator is often suboptimal. *Author response:* We thank the reviewer for their comment. We stress that the references Orlitsky et al. and Valiant and Valiant all assume an upper bound on the support size, and all their bounds break down for general distribution supported on countably many bins. Removing this assumption while still recovering proper convergence rates is our primary contribution. The empirical distribution still achieves the accurate minimax rate up to a universal constant for the learning problem with respect to total variation. We leave the investigations on improving the constant in our bounds and devising estimators that would be be first-order optimal as an interesting open question.

Reviewer comment: The results are similar to those in some intermediary steps of the prior work. Example of "On learning distributions from their samples" by Kamath et al. 2015, with an $O(\sqrt{(d-1)/m})$ upper bound on page 20, and an $O(\sum_i \sqrt{\mu(i)/m})$ upper bound just before. *Author response:* Indeed, the Kamath et al. 2015 paper is quite relevant to ours (it is already discusssed in the manuscript but we will expand the discussion in the revision). We note that their work is concerned with the finite support setting. Regarding the upper bound of the form $O(\sum_i \sqrt{\mu(i)/m})$, our response is three-fold. (i) Bounds of this form are well-known, appearing, e.g., in the Berend-Kontorovich 2013 paper we cite. (ii) Such bounds cannot handle general discrete distributions, as the sum diverges e.g., for $\mu(i) \propto 1/i^2$. (iii) Such bounds are not fully empirical (or adaptive) in the sense that they depend on the unknown μ. As discussed in our Introduction, the main motivation behind our paper was to address the limitations of (i,ii,iii).

Reviewer comment: The submission lacks comparison with prior works, in terms of both results and proof techniques. *Author response:* We will expand the discussion within the space constraints, and add more detail about some key references. We stress that in most cases the results are not directly comparable, in that prior work has focused on the finite support case, and all the bounds (e.g. 2015 "On learning distributions") depend on the support size.

Reviewer comment: The nature of the bound does not allow a user to obtain risk bounds a priori before seeing the samples. *Author response:* As the reviewer points out, such a priori bound is provably impossible to obtain; we therefore disagree with this being considered a weakness of the paper.

Reviewer comment: Combine with some underlying structure / shape-constrained technology, when this structure is still itself insufficient to obtain convergent rates. *Author response:* We thank reviewers #1 and #2 for this constructive comment. We agree that it will be interesting to investigate the influence on the complexity measure of shape-constraints such as log-concavity, monotonicity, or unimodality of the distribution. It is known that distribution learning and testing complexities can change drastically under proper structure assumptions, and we leave the question of this influence on the empirical complexity measure as an exciting research direction.

Reviewer comment: Why is the main result presented in terms of the 1/2-norm instead of the empirical Rademacher complexity? *Author response:* The first reason is that the empirical half-norm is computationally inexpensive, whereas the empirical Rademacher complexity is harder to compute. The half-norm is also conceptually simpler to visualize than the Rademacher complexity, and can be instructively compared to Valiant & Valiant's two-third-norm for the identity testing problem.

Reviewer comment: Is there something that can be thought of as a crude analogue of the measure for continuous measures? *Author response:* We thank for the reviewer for this question. Investigating extensions of this methodology to continuous measures (e.g., via kernel density estimates) is an active research direction of ours.

Reviewer comment: Minor typos and clarifications. *Author response:* We thank the reviewers for their careful reading. We will clarify L69 and add the missing expectation symbol at (13). Additionally, we remove our o_p claim (convergence in probability) in the remark at L160 and now simply assert that

$$\frac{\|\widehat{\boldsymbol{\mu}}_m\|_{1/2}^{1/2}}{\sqrt{2\pi m}} - \frac{3}{2}\sqrt{\frac{1}{2\pi}}\frac{1}{m^{3/2}}\left\|\widehat{\boldsymbol{\mu}}_m^+\right\|_{-1/2}^{-1/2} \le \hat{\mathfrak{R}}_m(\boldsymbol{X}) \le \frac{\|\widehat{\boldsymbol{\mu}}_m\|_{1/2}^{1/2}}{\sqrt{2\pi m}} + \sqrt{\frac{1}{2\pi}}\frac{1}{m^{3/2}}\left\|\widehat{\boldsymbol{\mu}}_m^+\right\|_{-1/2}^{-1/2}.$$

Reviewer comment: The discussion of "instance optimality" is a bit confusing, I didn't really see how this is analogous to the Valiant and Valiant result. It seems more similar to a standard minimax result. *Author response:* Indeed, the notion is distinct from that of Valiant and Valiant; we will clarify in the revision. Our intent was to draw attention to the fact that our empirical bounds are, in a sense, the best possible for *any* distribution, but perhaps *instance optimality* is not the best term for this.