We thank the reviewers for their helpful comments and enthusiasm.

Reviewer #2: Thank you for your comments. Regarding your points:

1. Motivation: We would like to clarify our aims: our main goal is to study the connections and properties of MMD DRO, to reveal its potential benefits, and aid a better general understanding of the DRO landscape with different divergence measures.

We do not claim that Wasserstein or phi-divergence uncertainty are bad or yield poor out-of-sample bounds; we merely highlight differences between them and MMD DRO. For example, in discussing phi-divergences, we state in line 88 that one cannot obtain a generalization bound via Principle 2.1 alone (to be fair, we will add a reminder to the reader that e.g. Namkoong and Duchi (2017) achieve generalization bounds via other means). For Wasserstein, we focus in lines 34-36 and lines 89-99 on other complications, such as difficulty of optimization, and that many of the upper bounds are only asymptotic. Indeed, the convergence results of Blanchet, Kang, and Murthy (2016) (in Section 3 of their paper) are also asymptotic in nature.

2. Upper bounds instead of exact reformulation, and assuming ℓ_f is in an RKHS: We again emphasize that there are tradeoffs between all three DRO approaches. While discarding the non-negativity constraint in MMD DRO may weaken the bounds, in return we obtain a simple non-asymptotic upper bound (unlike Wasserstein).

Moreover, we contend that the assumption that ℓ_f is in an RKHS $\mathcal{H}$ is not so restrictive after all. If the kernel k is universal, as is the case for many kernels used in practice such as Gaussian and Laplace kernels, we can readily extend our results to all bounded continuous functions as described below. We will add this clarification to the paper:

Suppose the loss ℓ_f of our predictor f is any bounded continuous function on a compact metric space $\mathcal{X}$. By definition [Muandet et al., Definition 3.3] if k is a universal kernel on $\mathcal{X}$ (associated with the RKHS $\mathcal{H}$), then for any $\epsilon > 0$, there is some $\ell' \in \mathcal{H}$ with $\sup_{x \in \mathcal{X}} |\ell_f(x) - \ell'(x)| < \epsilon$. It follows that for any measure $\mathbb{P}$, we can bound the expectation of $\ell_f(x)$ by that of ℓ': $\mathbb{E}_{x\sim\mathbb{P}}[\ell_f(x)] < \mathbb{E}_{x\sim\mathbb{P}}[\ell'(x)] + \epsilon$. Then, we can apply our results to $\ell' \in \mathcal{H}$.

3. MMD DRO is a more conservative upper bound (Theorem 5.1): Separate from the task of producing a valid (and hopefully tight) upper bound is the task of designing a regularizer that is practically useful. And stronger regularizers are often better. One drawback of the previously considered chi-squared DRO/variance regularization is that the regularization ceases to have any effect when the training data can be fit perfectly e.g. in deep learning (since then the loss for each datapoint is zero, and so the variance of the loss on the dataset is also zero). In such a regime, stronger penalties such as the RKHS norm continue to be meaningful.

Reviewer #3: Thank you for your support.

Re: discrete approximation of MMD DRO uncertain set may not contain the population: Yes, any such discrete approximation can have similar issues. We present it mainly to link variance regularization and MMD DRO, e.g. as in Theorem 5.1.

Reviewer #4: Thank you for your feedback and support.

Before addressing your main comments in detail, we emphasize that at a high level we hope to present MMD DRO as an alternative worth studying, with complementary properties to existing techniques and rich connections. In that context,

(a) Wasserstein convergence with fewer assumptions: Thank you for pointing us to the references on non-Euclidean Wasserstein convergence; we are happy to mention them in the camera ready. Regarding your comment about assuming ℓ_f is in an RKHS, please see point 2 of our response to Reviewer #2.

(b) Faster convergence: The point you make about norms cancelling with rates is fair. We mention Wasserstein's $O(n^{-1/d})$ rate in the paper because it is relevant to the application of Principle 2.1. However, as discussed in point 1 of our response to Reviewer #2, we don't mean the remark about $O(n^{-1/2})$ vs $O(n^{-1/d})$ to claim the MMD results were always better. Instead, the different convergence properties of different distances motivates studying different DRO formulations. We will edit the paper to make this clearer.