We thank all reviewers and we will modify the paper to clarify each of the points raised, as discussed/clarified below.

**R1**. (***Q3***) The primary contribution of our paper is a general theoretical framework for the analysis of Wasserstein distance between an empirical measure and a probability measure when the hypothesis class is infinite dimensional. Our goal is to provide insights to explain the empirical success of Wasserstein distance in practice, despite the well-known curse of dimensionality. Two-sample testing is a specific application that can be addressed by our general framework, but not the only one. For this specific two-sample testing case, we seek to verify the hypothesis test that an empirical set $\{X_1, X_2, \ldots, X_n\}$ has been sampled from a given probability distribution $P_*$. We compute $\hat{R}_n$ and reject the hypothesis if $\hat{R}_n$ is larger than the 95% quantile. The quantile is obtained by sampling $\hat{R}_n$ between $P_*$ and $P_n$, where $P_n$ is the empirical distribution sampled from $P_*$. We will revise the presentation to make these points more clearly.

(***Q5***) The 4th paragraph of the introduction provides a concrete example of the high-level concepts discussed in the 3rd paragraph, specifically describing how our theoretical results can be used in the discrimination step for constructing a Wasserstein GAN. The questions concerning Section 4 and our theoretical results for hypothesis testing are addressed in (*Q8*). We will refine the introduction and Section 4 to clarify how our theoretical results can be used by the community.

(***Q6***) While $R_n$ for infinite-dimensional function classes can scale as slowly as $O_p(n^{-1/d})$ (as discussed in the paragraphs starting on L71 and L166), we establish that the rate of convergence for $R_n$ does not depend on the dimension $d$ for the general class of infinite-dimensional functions considered in the paper, as formally presented in Theorems 2 and 4 for the compact and non-compact settings, respectively. With respect to the literature you mentioned, the work of Tameling et al. (2019) does recover parametric rates of convergence, but under the assumption that the underlying measures are atomic. The entirety of Section 2.4 in Tameling et al. (2019) is dedicated to explaining (via the development of a lower bound) why their method does not apply to continuous measures. The paper of Genevay et al. (2017) is not a theoretical statistical analysis and it does not provide a rigorous rate of convergence for statistical learning. We will include more discussion to clarify these points and comparisons with previous works.

(***Q8***) As discussed in our responses to (*Q3*) and (*Q5*), for the specific two-sample testing case, we seek to verify the hypothesis test that an empirical set has been sampled from a given probability distribution. This includes, in both the compact and non-compact settings, that we can exploit our strong duality results to compute an estimate for the quantile of $R_n$ required to accept or reject a given hypothesis; refer also to (*Q3*) response. Theorem 2 provides a characterization of the limiting distribution for $nR_n$ for the compact setting, which renders additional flexibility by providing an alternate way to compute the required quantile. For the more complex non-compact setting, Theorem 4 establishes the convergence rate of $R_n$ to be $O_p(n^{-1/2})$. The primary goal of Theorem 4 is to establish a parametric rate that provides insights into why Wasserstein distance can beat the curse of dimensionality in practice. Hence, we do not use Theorem 4 to compute the specific quantile. We will refine the presentation to address all of these points more clearly throughout the paper, including to clarify how to apply our theoretical results for hypothesis testing.

**R2 & R4**. Thank you both for your positive comments on our paper. As **R2** noted, we do plan to further extend the practical impact of our theoretical analysis as part of ongoing research. To address the point raised by **R2**, we will briefly expand the discussion on rates of convergence; and to address the points raised by **R4**, we will conclude with a discussion of our theoretical results and planned future work, and we will refine the appendix to improve clarity and the formatting of some of the equations. We also thank you both for the identified typos, which we will readily address.

**R3**. Thank you for your positive comments on our paper. ***(1)*** You raise a good point about our statistical convergence results, which is well taken, with the caveat that we consider finitely many linear projections, not only a single one. This, actually, makes the result very difficult to prove. To address your point, we will add a discussion about what hints at the general condition to ensure the parametric rate, which, we believe, is closely related to the tightness of the formal limiting object. This generalization is a topic of our current research. Note that Assumption 2 is a sufficient condition given in terms of the problem primitives for this particular class. For more general function classes, we conjecture that the convergence rate should be $O(n^{-1/d'})$, where $d'$ is the "effective dimension" (suitably defined) of the function class. (***2***) You are indeed correct that IPM is similar to our formulation in terms of the duality representation. While there are a few important technical differences, we note that it is not our primary intention to define a new metric. Rather we seek to provide a thorough analysis of the Wasserstein distance, which has been the focus of a great deal of attention in the statistical learning research literature. In particular, we add a new modeling feature, which is the hypothesis class or the actor critic class. This induces a class of dual functions; and we note that our expression for the strong duality (generalizing the celebrated Kantorovich-Rubinstein duality) uses the combination of both the function $f$ and its dual $f^c$ in contrast with IPM. We thank you for raising this point, and we will include a discussion of this comparison in the final version. (***3***) The primary contributions of our paper are theoretical in nature. However, as noted in our response to **R2**, we plan to extend the practical impact of our theoretical work as part of ongoing research. This includes applying our theoretical results to real-world dataset.