Response to reviewers comments for the paper: “Faster Wasserstein Distance Estimation with the Sinkhorn Divergence”

We thank the reviewers for their comments and suggestions. Hereafter, we list reviewers’ (paraphrased) comments (**C**) followed by our responses (**R**). These answers will translate into clarifications in the final version of the paper.

Response to reviewer #1’s comments

- **(C)** *It’s not clear to me that the paper leads to any real practical gain.* **(R)** Our goal is mostly to deepen our theoretical understanding of known objects, such as the Sinkhorn divergence.
- **(C)** *It might have been interesting to see the plots for the error in squared Wasserstein distance.* **(R)** We will move such plots (currently in the supplementary material) to the main paper.
- **(C)** *The authors do cite Mena & Weed, and Genevay et al., but it might be helpful to discuss the relation of the current sample complexity results with these previous bounds* . **(R)** Those works were concerned with estimating the *regularized* optimal transport cost, while our goal is to estimate the *unregularized* cost. These are very different problems: the latter is cursed by the dimensionality while the former is not.

Response to reviewer #2’s comments

- **(C)** *The submission should be self-contained.* **(R)** Our submission is self-contained. The main paper contains all our results, with a self-contained narrative, and all the proofs are provided in the supplementary material, as referenced in the main text. We have often included proof ideas in the main text. This is consistent with NeurIPS guidelines which state that "Authors may submit up to 100MB of supplementary material, such as appendices, proofs, derivations [...]".
- **(C)** *I would try to focus on one or two results, and give a full(er) presentation..* **(R)** We think that the set of results that we present form a coherent story. We will use an additional page in the final version to give more details and ease the reading.

Response to reviewer #3’s comments

- **(C)** *How to choose the best regularization parameter λ^*?* **(R)** The optimal value λ^* depends on quantities which are typically not known so our theory is not yet able to answer this question. In practice, many machine learning tasks involving the Wasserstein distance come with a performance criterion, in which case cross-validation can be used.
- **(C)** *Can we verify the convergence rates with respect to the number of samples in Figure 1 by adding a plot for the theoretical rates?* **(R)** As mentioned in the text, we do not know the theoretical rates for the quantity shown in Figure 1 (L^1 error on the potential), so we cannot plot it. Note that the observed rate is different from $n^{-2/d}$.
- **(C)** *Be more specific about the required smoothness of the problem, in terms of the bound for the derivatives.* **(R)** This is a question that Proposition 1 partially addresses. We will add details on how to guarantee smoothness of the Brenier potential assuming smoothness of the densities (a typical and difficult problem in optimal transport theory).

Response to reviewer #5’s comments

- **(C)** *It would be good to see empirical errors as a function of dimensionality [also asked by **reviewer #6**].* **(R)** Such plots would not be very interpretable since we did not track the full dependency in the dimension d in the theory. Our focus is on the rates in the sample size n. Plots in higher dimension are qualitatively similar, but with different slopes.

Response to reviewer #6’s comments

- **(C)** *It is hard to determine whether the experiments verify the claims.* **(R)** As justified in the main text, Figure 1 and 2 are not intended to verify our theorems but rather to exhibit related phenomena and suggest future research directions. We will add in the main paper a plot for the error in cost and a plot to illustrate Theorem 2, which verify our claims.
- **(C)** *Computational time is not evaluated enough in the experimental section.* **(R)** We will expand a bit and add a plot.
- **(C)** *In the experimental results, the effectiveness of Richardson is not so significant. Why?* **(R)** We observed a significant debiasing effect (see Figure 1-(a)), but indeed no clear statistical or computational gain. We believe that this is mainly due to "constants" in the bounds: compared to S_λ, the estimation error for R_λ is up to 3 times larger and the computational time is up to 3 times larger.