We would like to thank the reviewers for their efforts on reading and evaluating our paper. We appreciate that they pointed out the importance of the problem and of our analysis given the increasing popularity of computational OT. All the minor comments will be addressed in the paper draft, and we will release the source code if the paper is accepted. In what follows, we provide specific responses to each reviewer.

Reviewer 1. We thank R1 for a very detailed evaluation and truly helpful feedback on our empirical analysis. We will use the sentence "PRW provides a computational advantage over SRW (provided that ϵ is not too small)" as you suggest to conclude our empirical analysis in the updated version.

♦ *Comparison between PRW and SRW:* We present Figures 1-3 and Tables 1-3 to show that it is reasonable to compute the PRW distance by our algorithm. To be more specific, while the SRW distance can be globally optimized, our algorithms only return an approximate stationary point which needs to be evaluated in practice. We agree that SRW has more discriminative power than PRW since it is equivalent to the Wasserstein distance [Prop.2, 64]. However, it also suffers from the curse of dimensionality in theory despite of practical performance. Finally, we highlight that the ideal sample complexity of PRW is partially demonstrated by its robustness to input data. As shown by Figure 4, PRW is more robust than SRW and Wasserstein distance, when the noise has the moderate to high variance.

♦ *The hyperparameter tuning [...] and initialization steps [...] are not explained clearly enough* We will move necessary details of hyperparameter tuning in Appendix H to the main paper to make the experiments clear. Furthermore, we will add the experiments with varying ϵ values with the same setting as in Figure 4 in the updated version.

♦ *Can't we apply the same strategy in the Frank-Wolfe algorithm of SRW [Alg.2, 64]:* Yes, we can apply the same strategy as in [Alg.2, 64] and it improves the computational efficiency and affects the quality of the approximate solution. Indeed, we made good use of the open source code they have provided to inspire our implementation.

♦ *Is it possible to extend this result to PRW?* We demonstrate that it is difficult to extend Prop. 3 from [64] to PRW. Indeed, the characterization of SRW as a sum of eigenvalues is crucial to the analysis but PRW does not have such property. Furthermore, as k becomes larger, the nonconvex max-min optimization problem has more stationary points. It is possible that RGAS and RAGAS converge to different stationary points with different PRW values.

♦ *is there an explanation on why RGAS is above RAGAS for $k < k^*$, and then under for $k \geq k^*$?* We provide one possible reason for the behavior of PRW value in Figure 4. Indeed, when $\sigma = 4$, our algorithm can converge to a stationary point where the PRW value is closer to the PRW value with $\sigma = 0$.

Reviewer 2. We thank R2 for the positive evaluation and very helpful feedback on the theoretical part.

♦ *Convergence to a stationary point:* We will fix this confusing point in the updated version.

♦ *Comparing a local optimal solution to the global optimal solution of a convex relaxation problem in [64]:* Investigating this relationship is difficult in general due to multiple stationary points of non-convex max-min problem but seems possible if data has certain structure. We will add a remark to elaborate our insights in the updated version.

♦ *High-dimensional data:* We agree and will present some results on high-dimensional text in the updated version.

Review 3. We thank R3 for an informative and thought provoking review.

♦ *I admit this work has relevance to optimal transport in terms of numerical optimisation method, and it is interesting to exploit more efficient technique for optimising WPP.* Indeed, we have observed in the field of computational OT that the discovery of efficient computational approaches often precede other types of advances (applied or statistical)

♦ *However, the main problem is this work only focus on Riemannian optimatzation part. So the idea is not very attractive to me.* Although we believe the statistical properties of entropic regularized WPP/PRW distance are definitively worth investigating, *we do not understand why our focus here on computational aspects should appear to you as a problem*. These are two distinct and complementary subjects. We argue that studying computational aspects for WPP/PRW distance with theoretical guarantee is needed for these tools to take off. Our methodology is new, and uses a nontrivial combination of techniques from OT and Riemannian optimization. Luckily, we are aware of recent work [`https://arxiv.org/abs/2006.12301`] that might answer some of your questions. Taken together, these two contributions provide new directions for computational OT.

Reviewer 4. We thank R4 for a positive evaluation and very helpful feedback on the paper organization.

♦ *The experimental section is not self sufficient:* As suggested, we will move some parts of implementation details back to the main context and make the experimental section self-contained.