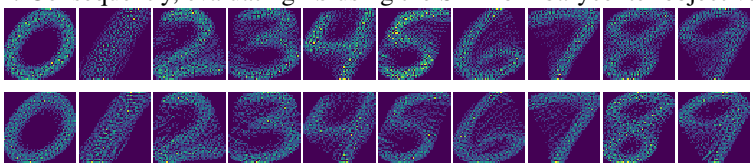
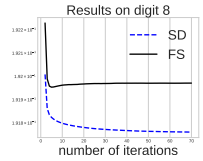
We thank the reviewers for their careful consideration and constructive feedback. Please find our responses below.

General response to all reviewers on empirical study of SD. A common suggestion of all the reviewers for improving the paper is to provide more empirical study of SD. We hence provide the results on the MNIST dataset. Specifically, we consider 10 target distributions (one for each digit) and our goal is to compute their Sinkhorn barycenter (we randomly select 100 images for each digit to construct the t.). We compare SD ($\gamma = 1e^{-4}$) with the free-support (FS) method, i.e. Algorithm 2 in [Cuturi and Doucet, 2014]. We use the implementation from the PythonOT library, which does not optimize over the weights (hence all particles have uniform weights like SD), nor it does local line searching. Note that this is the same as the fix point iteration (section 5.3) in [Claici et al, 2018]. Both methods use 2500 particles. We do not include the iterative Bregman projection (IBP) algorithm here since its implementation in PythonOT has numerical issues for the small γ case (it involves the division of $\exp(-1/\gamma)$). We will implement a numerical-stable version of IBP and add the result in our revision. Note that IBP does not handle the debiasing term in the Sinkhorn barycenter problem. [Chewi et al. 2020] is excluded here as it only applies to the barycenter problem of Gaussian distributions. In figure (i), the first row is the results of SD and the second row is the results of FS. To interpret these plots, brighter pixels mean more particles in a region. We can observe that the particles of SD are more concentrated on the digits compared to the ones of FS. We present in figure (ii) the comparison of convergence rate between SD and FS (only for digit 8–similar for others). Note that FS aims to solve the original Wasserstein barycenter problem without entropy regularization. Consequently, evaluating FS using the Sinkhorn barycenter objective value may not be very precise.



(i) Top row is from SD and the bottom row is from FS



(ii) loss vs # iter

To Reviewer #1. Q1. About Remark 3.1. A1. A deeper understanding of the RKHS restriction on the push-forward mapping is a very interesting direction and we are working on this right now. Q2. What if the variable measure α is discrete? A2. In this case, the result of Theorem 4.1 still stands and SD converges to a stable point. However, there is no guarantee for quality of such stable point: SD will *not* converge to the global optimal since α is not fully supported and the assumption in Theorem 4.3 does not hold. Q3. More experiments. A3. Please see our general response above. Q4. The limitation of discrete initialization α_0. A4. This is an excellent point raised by the reviewer. In practice, we start with a sufficient number of particles. Besides, we observe that increasing the number of particles reduces the Sinkhorn divergence at convergence, which, however, has diminishing returns. Q5. The gradients correspond to the extension of the potentials. A5. We will mention this and correct other typos in our revision. Thanks.

To Reviewer #2. Q1. Dependence on $\exp(1/\gamma)$. A1. Indeed, this problem is believed to be hard in the literature. The term $\exp(1/\gamma)$ appears in bounding the derivatives of Sinkhorn potentials, which also appears in bounding the sample complexity of the Sinkhorn divergence (see Theorem 2 and Lemma 3 of [Genevay et al. 2019]). The sample complexity in [Genevay et al. 2019] can be improved if one manages to remove this factor. However, this would potentially violate the lower bound on the sample complexity of the hard-to-compute Wasserstein distance, since it is the limit of the Sinkhorn divergence at $\gamma \to 0$. We will elaborate on this in our revision. Q2. How does $\exp(1/\gamma)$ impact practice. A2. Surprisingly, the empirical performance of SD does not suffer much from this factor: In our experiments, to produce good visual results, we pick $\gamma = 10^{-4}$ and we still observe that SD quickly converges (even in the high dimensional Gaussian barycenter task). We observe that the problem can be solved to high accuracy with different configurations of γ. Besides, a larger γ results in a more blurred barycenter. We will elaborate more on the impact of γ in our revision. Q3. Comparison with iterative Bregman projection. A3. Please see our general response above. Q4. logsumexp in SD. A4. Yes. Q5. y-axis of Figures 1a and 1c. A5. Both y-axes are the Barycenter function values (the latter).

To Reviewer #3. Q1. Implicit exponential dependence on the problem dimension. A1. Indeed, as shown in Lemma 4.1 and Theorem 4.1 (see line 209 and 214), our results depend on $\exp(M_c/\gamma)$ where M_c is the upper bound on ground cost on the domain $\mathcal{X}$ which contains an implicit dependence on the problem dimension. We will elaborate on this in our revision. We will also discuss that Sinkhorn divergence interpolates Wasserstein distance and MMD. Q2. cost of [Genevay et al. 2016]. A2. Since the entropy regularized optimal transport problem is strongly convex, SGD converges at the rate $\mathcal{O}(1/k)$ for such problem, where k is the number of SGD steps. Besides, the per-iteration complexity is $O(n^2)$ where n is the support size of input measures. We will discuss about it in the revision. Q3. Comparison with classical tools for computing barycenter. A3. Please see our general response above. Q4. Missing citation. A4. Thanks. We will properly cite all the works mentioned by the reviewer in our revision. Q5. Regularization parameter. A5. We set $\gamma = 10^{-4}$ in all of our experiments to produce results of good visual quality.

To Reviewer #4. Q1. Empirical result on MNIST. A1. Please see our general response above. Q2. Running time comparison of FW and SD. A2. In our experiment, we directly use the implementation of FW from the original paper and we observe that SD is much more efficient than FW. This is because each FW step requires to globally solve a nonconvex subproblem via grid search as discussed in lines 258-263 of our paper. We will highlight this in our revision.