We thank all the reviewers for their insightful comments and detailed reviews. We share their enthusiasm that our work provides "a rigorous framework for dealing with label switching in mixture models" (R1) and "brings a new viewpoint on the problem, as well as new tools" (R3).

Below we discuss reviewer comments in detail. We are confident that we can address any requested revisions in time for publication to NeurIPS 2019 and that our work will be of interest to the optimal transport, Bayesian, and statistical audiences attending the conference.

**Theoretical contributions.** **R2** asserts that the Wasserstein barycenter is no better than the original posterior distribution as a summary statistic. This is an inaccurate assessment of our work: **In all practical scenarios, the Wasserstein barycenter is a *point estimate* for the true (non-degenerate) posterior mean (Theorem 2).** The posterior alone does not easily give this information due to the inherent *label switching* phenomenon; this is the key issue addressed by our work.

**The choice of sampler is orthogonal to the problem we tackle.** For our method to succeed, it is not necessary that the sampler visits all modes of the posterior, nor does it depend on the sampler not departing from the neighborhood of a single mode. Regardless of the coverage of modes in the posterior distribution, our approach provides a principled notion of correspondence between samples from different modes, resulting in a sensible and well-posed mean estimate on the quotient space.

We will detail the choice of MCMC sampler for the multi-reference alignment experiments. We used a Gibbs sampler and then applied our SGD algorithm in these experiments.

We thank **R3** for suggesting a Bayesian interpretation of our algorithm. While nonuniqueness of the barycenter is possible (see §1.4 and §2.2 of the supplementary), this problem never occurred for us empirically. The supplementary sections and the referenced results of Arnaudon et al. 2013 suggest that uniqueness is almost surely true; nonuniqueness occurs only under an extremely high (and unlikely) degree of symmetry in posterior samples.

**Experiments.** We are happy to provide additional experiments in our final revision, as suggested by **R1** and **R3**, and welcome any suggestions for additional experiments. We emphasize that mixture models are widely-used, effective probabilistic models in machine learning and statistics; our goal is not to improve them but rather to alleviate a common issue in Bayesian mixture modeling, which generalizes to problems with symmetry groups other than the permutation group.

As **R2** suggested, we will clarify characterizations of the baselines in the text. Next we offer a brief summary. The Stephens and Pivot methods relabel samples. Stephens minimizes the Kullback–Leibler divergence between average classification distribution and classification distribution of each MCMC sample. Pivot aligns every sample to a pre-specified sample (i.e. pivot) by solving a series of linear sum assignment problems. Pivot method requires pre-selecting a single sample for alignment — poor choice of the pivot sample leads to bad estimation quality, while making a "good" pivot choice may be highly non-trivial in practice. The default pivot choice is the MAP, however it may fail as discussed in lines 282-287 and illustrated in Figure 2. Stephens method is more accurate, however it is expensive computationally and has large memory requirement to store a tensor of size [data size $\times$ number of MCMC samples to be aligned $\times$ number of mixture components $K$].

**Clarity.** We agree with **R3's** suggestion that a simple running example of a mixture of Gaussians would improve clarity, and we will include one. Code will be made available via a Jupyter notebook.

- (**R1**) Line 125 : $S_K$ is the group of permutations of a finite set of $K$ points
- (**R1**) Line 131: An invariant transport plan $\pi : X \times X \to \mathbb{R}$ is invariant to the diagonal action of $G$ on $X \times X$. The invariance relation is one of equivariance if the coupling $\pi$ specifies a map, but this is not true in general. The proof strategy is correct; we will add a complete proof to the supplementary.
- (**R1**) Line 140: We were following the notation in the reference, but will change to match with the rest of the paper.
- (**R3**) Line 196: Our algorithm deals with samples as they come in, rather than collecting multiple samples and processing them together.
- (**R1**) Line 229: $\sigma$ refers to the map minimizing eq. (5).
- (**R1**) Line 230: Thanks for pointing this out. We'll fix it.
- (**R1**) Line 266: We followed the strategy of Muzellec & Cuturi and used a parameterization by factors to allow for more efficient computation of gradient steps. We will fix these inconsistencies in the final version.
- (**R2**) Page 3, eqn (1): We will use $\mu^*$ on the left hand side.
- (**R3**) Section 3, $\Omega$: interpretation of $\Omega$ suggested by **R3** is correct. Including a running example with Gaussian mixtures will help us to make the meaning of $\Omega$ more transparent.