Alleviating Label Switching with Optimal Transport

{Pierre Monteiller[1], Sebastian Claici[2,4], Edward Chien[2,4], Farzaneh Mirzazadeh[3,4], Justin Solomon[2,4] and Mikhail Yurochkin[3,4]} [1]ENS Ulm, [2]MIT CSAIL, [3]IBM Research, [4]MIT-IBM Watson AI Lab

Label Switching

Invariance of prior and likelihood under group action $\to K!$ **symmetric regions** in the posterior landscape

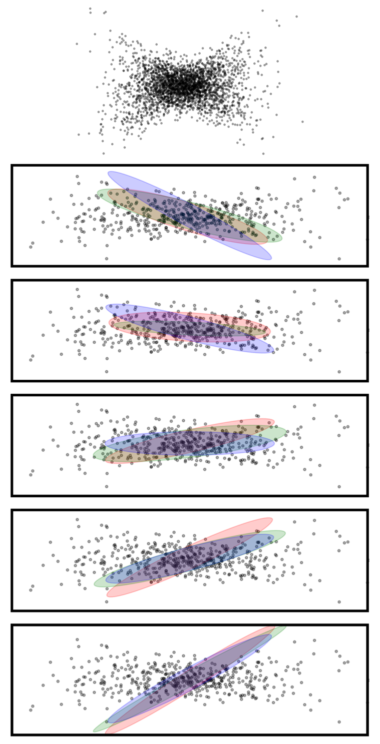
Example : Gaussian mixture

$$p(x|\Theta) = \sum_{k=1}^{K} \pi_k f(x; \mu_k, \Sigma_k) = p(x|\sigma(\Theta))$$

Contributions

- An algorithm to address the label switching problem.
- Theory relating Wasserstein barycenters to estimates of the symmetrized posterior statistics.
- A simple stochastic gradient descent algorithm.

Pivot Methods Fail



Setting :

- Mixture of five Gaussians
- Mean 0 and Covariances $\begin{pmatrix} 1 & 0 \\ 0 & 0.1 \end{pmatrix}$ rotated by angle $\theta \in \{-\pi/12, -\pi/24, 0, \pi/12, \pi/24\}$
- True covariances blue, SGD in green and pivot in red

Failure of fast Pivot method [3]

Optimal Transport with Group Actions

p-Wasserstein distance on $\mathbf{P}(\mathbf{X})$: for (X, d) *complete* and *separable* metric space, μ and ν measures, and $\Pi(\mu, \nu)$ the set of probability measures on the product space with marginals μ and ν :

$$W_p^p(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times X} d(x, y)^p \, \mathrm{d}\pi(x, y)$$

Set of measures invariant under group action: for G a *finite group* acting by *isometries* on X, and $P_2(X)$ finite second moments measures,

$$P_2(X)^G = \{\mu \in P_2(X) \mid g_{\#}\mu = \mu, \forall g \in G\}$$

Relation between the space $P_2(X)^G$ and $P_2(X/G)$: [2]

Let p be the quotient map, $p_* : P_2(X) \to P_2(X/G)$ restricts to an isometric isomorphism between the set of $P_2(X)^G$ of G-invariant elements in $P_2(X)$ and $P_2(X/G)$.

Wasserstein Barycenter

Generalization of Wasserstein barycenter: [1] Let $\Omega \in P_2(P_2(X))$

$$B(\mu) = \int_{P_2(X)} W_2^2(\mu, \nu) \, \mathrm{d}\Omega(\nu) = \mathbb{E}_{\nu \sim \Omega}\left[W_2^2(\mu, \nu)\right]. \quad (1)$$

Theorem 1. $B(\mu)$ has at least one minimizer in $P_2(X)$ if $\operatorname{supp}(\Omega)$ is tight.

Assumption on Ω: $\nu \sim \Omega$ has the following form: $\boxed{\nu = \frac{1}{|G|}\sum_{g \in G} \delta_{g \cdot x}}$ for some $x \in X$.

Barycenters under Group Action: Under this assumption, minimization of $B(\mu)$ is equivalent (with $\Omega_* := p_{*\#}\Omega$) to

$$\boxed{\arg\min_{\mu \in P_2(X/G)} \mathbb{E}_{\delta_x \sim \Omega_*}\left[W_2^2(\mu, \delta_x)\right].}$$

Main Theoretical Result:

Theorem: Single Orbit Barycenters

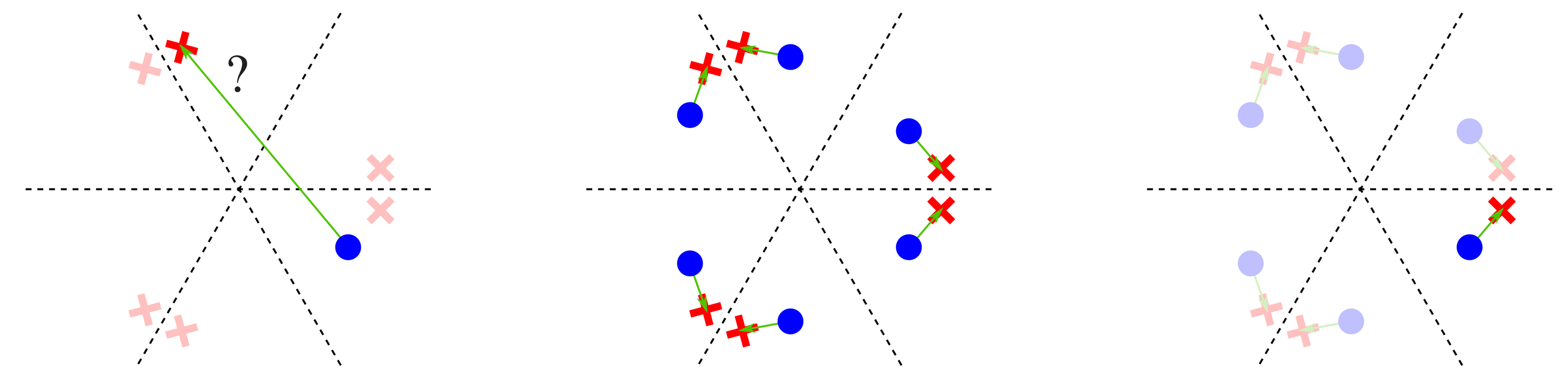
There is a barycenter solution of (1) that can be written as

$$\mu = \frac{1}{|G|} \sum_{g \in G} \delta_{g \cdot z^*} \quad \text{for a point } z^* \in X/G.$$

Principled method for extracting point estimates: take a quotient, find a mean in X/G, and then pull the result back to X.

Barycenter of Ω on quotient space

Input: sampler from Ω over a manifold $\mathcal{M}$ **Output:** a barycenter of the form $\frac{1}{|G|}\sum_{g \in G} \delta_{g \cdot x}$ for some $x \in \mathcal{M}$, using Riemannian SGD (i.e. taking the log then the exponential) on (1).



(a) Ambiguity (b) Orbit empirical distribution (c) Quotient update

Gradient descent on quotient space: for parameters $(p_1, \ldots, p_K) \in \mathcal{M}^K$, let's consider

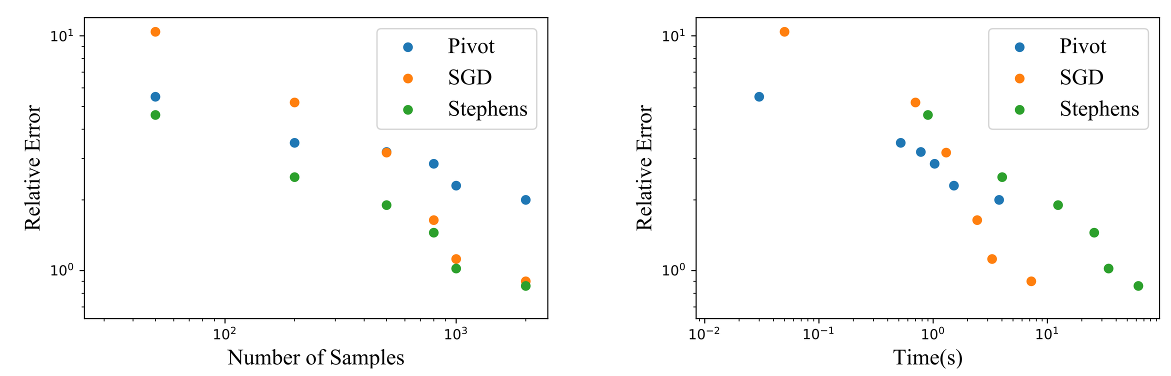
$$\mathrm{Conf}_K(\mathcal{M}) := \mathcal{M}^K \backslash \{(p_1, \ldots, p_K) \mid p_i = p_j \text{ for some } i \neq j\}$$

As $\Omega \in P(\mathrm{Conf}_K(M))$, we quotient $\mathrm{Conf}_K(\mathcal{M})$ by G, the obtained manifold $\mathrm{UConf}_K(M)$ has a structure inherited from the product metric,

$$d_{\mathrm{UConf}_K(M)}([(p_1, \ldots, p_K)], [(q_1, \ldots, q_K)]) = \min_{\sigma \in S_K} d_{\mathcal{M}^K}((p_1, \ldots, p_K), (q_{\sigma(1)}, \ldots, q_{\sigma(K)})). \quad (2)$$

At each iteration we draw q, compute σ, and apply a gradient step.

Estimating Gaussian mixture



Setting: Mixture of 5 Gaussians over $\mathbb{R}^5$ with means $0.5e_i$ and covariances $0.4I_{5\times5}$. **Results:** Pivoting obtains a suboptimal solution quickly, but if a more accurate solution is desired, our algorithm performs better.

Gradient step for Gaussian Mixtures

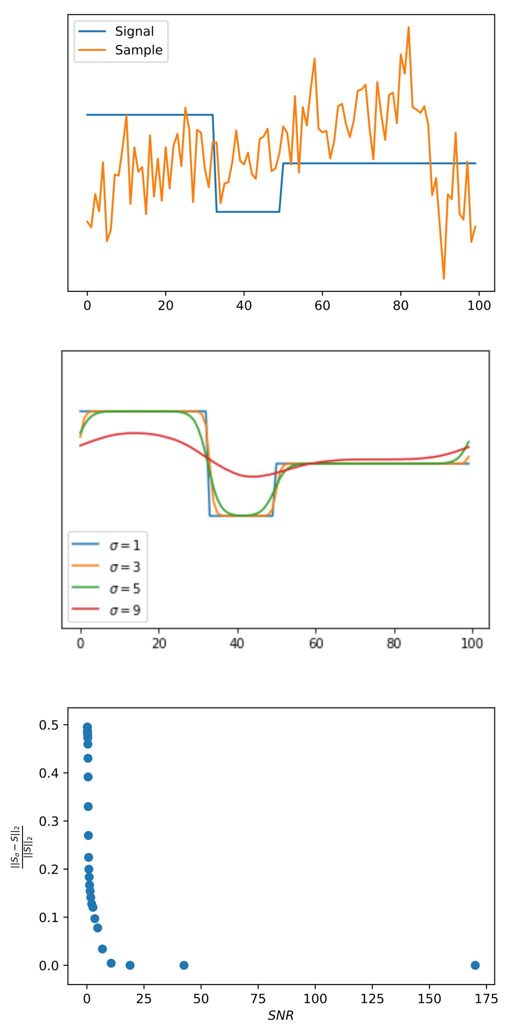
- **Means:** $\boxed{\mu^* = \mu^* - \eta(\mu^* - \mu)}$
- **Covariances:** with a Cholesky decomposition $\Sigma_i = L_i L_i^\intercal$ for every component in the mixture,

$$\boxed{L_i^* = L_i^* - \eta(I - T^{\Sigma_i^* \Sigma_i}) L_i^*} \quad T^{\Sigma_i^* \Sigma_i} = \Sigma_i^{*-\frac{1}{2}} (\Sigma_i^{*\frac{1}{2}} \Sigma_i \Sigma_i^{*\frac{1}{2}})^{\frac{1}{2}} \Sigma_i^{*-\frac{1}{2}}.$$

Algorithm

Input: Distribution Ω
Output: Barycenter $(p_1, \ldots, p_K)$
  $(p_1, \ldots, p_K) \sim \Omega$
  for $t = 1, \ldots$ **do**
    Draw $(q_1, \ldots, q_K) \sim \Omega$
    Compute σ in (2)
    for $i = 1, \ldots, K$ **do**
      $-D_{p_i} c(p_i, q_{\sigma(i)}) := \log_{p_i}(q_{\sigma(i)})$
      $p_i \leftarrow \exp_{p_i}\left(-\frac{1}{t} D_{p_i} c(p_i, q_{\sigma(i)})\right)$
    end for
  end for

Alignment



Multi-reference alignment: Reconstruction of a template signal $x \in \mathbb{R}^K$ given noisy and cyclically shifted samples $y \sim g \cdot x + \mathcal{N}(0, \sigma^2 I)$.

References

[1] Young-Heon Kim and Brendan Pass. Wasserstein barycenters over Riemannian manifolds. *Advances in Mathematics*, 307:640–683, February 2017.

[2] John Lott and Cédric Villani. Ricci curvature for metric-measure spaces via optimal transport. *Annals of Mathematics*, pages 903–991, 2009.

[3] Panagiotis Papastamoulis. label.switching: An R package for dealing with the label switching problem in MCMC outputs. *arXiv preprint arXiv:1503.02271*, 2015.