Hierarchical Optimal Transport for Multimodal Distribution Alignment

John Lee†, Max Dabagia†, Eva L. Dyer†‡§, Christopher J. Rozell†§
†School of Electrical and Computer Engineering.
‡Coulter Department of Biomedical Engineering
Georgia Institute of Technology, Atlanta, GA, 30332 USA
{john.lee, maxdabagia, evadyer, crozell}@gatech.edu

Abstract

In many machine learning applications, it is necessary to meaningfully aggregate, through alignment, different but related datasets. Optimal transport (OT)-based approaches pose alignment as a divergence minimization problem: the aim is to transform a source dataset to match a target dataset using the Wasserstein distance as a divergence measure under alignment constraints. We introduce a hierarchical formulation of OT which leverages clustered structure in data to improve alignment in noisy, ambiguous, or multimodal settings. To solve this numerically, we propose a distributed ADMM algorithm that exploits the Sinkhorn distance, thus it has an efficient computational complexity that scales quadratically with the size of the largest cluster. When the transformation between two datasets is unitary, we provide performance guarantees that describe when and how well cluster correspondences can be recovered with our formulation, and then describe the worst-case dataset geometry for such a strategy. We apply this method to synthetic datasets that model data as mixtures of low-rank Gaussians and study the impact that different geometric properties of the data have on alignment. Next, we applied our approach to a neural decoding application where the goal is to predict movement directions and instantaneous velocities from populations of neurons in the macaque primary motor cortex. Our results demonstrate that when clustered structure exists in datasets, and is consistent across trials or time points, a hierarchical alignment strategy that leverages such structure can provide significant improvements in cross-domain alignment.

1 Introduction

In many machine learning applications, it is necessary to meaningfully aggregate, through alignment, different but related datasets (e.g., data across time points or under different conditions or contexts). Alignment is an important problem at the heart of transfer learning [1, 2], point set registration [3, 4, 5], and shape analysis [6, 7, 8], but is generally NP hard. In recent years, distribution alignment methods that use optimal transport (OT) to quantify similarity between two distributions have increased in popularity due to their attractive mathematical properties and impressive performance in a variety of tasks [9, 10]. However, using OT to solve unsupervised distribution alignment problems that must simultaneously match two datasets’ distributions (using OT) while also learning a transformation between their latent spaces, is extremely challenging, especially when the data has complicated multi-modal structure. Leveraging additional structure in the problem is thus necessary to regularize OT and constrain the solution space.

Here, we leverage the fact that heterogeneous datasets often admit clustered or multi-subspace structure to improve OT-based distribution alignment. Our solution to this problem is to simultaneously

†JL is currently with DSO National Laboratories of Singapore.
§Equal contributing senior authors.

estimate the cluster alignment across two datasets using their local geometry, while also solving a
global alignment problem to meld these local estimates. While it is advantageous to regularize the OT
problem with known cluster pairings [10, 11], we are instead concerned with the substantially harder
unsupervised setting where such information is missing. We introduce a hierarchical formulation of
OT for clustered and multi-subspace datasets called Hierarchical Wasserstein Alignment (HiWA).3

We empirically show that when data are well approximated with Gaussian mixture models (GMMs)
or lie on a union of subspaces, we may leverage existing clustering pipelines (e.g., sparse subspace
clustering [12, 13]) to improve alignment. When the transformation between datasets is unitary,
we provide analyses that reveal key geometric and sampling insights, as well as perturbation and
failure mode analyses. To solve the problem numerically, we propose an efficient distributed ADMM
algorithm that also exploits the Sinkhorn distance, thus benefitting from efficient computational
complexity that scales quadratically with the size of the largest cluster.

To test and benchmark our approach, we applied it to synthetic data generated from mixtures of
low-rank Gaussians and studied the impact of different geometric properties of the data on alignment
to confirm the predictions of our theoretical analysis. Next, we applied our approach to a neural
decoding application where the goal is to predict movement directions from populations of neurons
in the macaque primary motor cortex. Our results demonstrate that when clustered structure exists
in neural datasets and is consistent across trials or time points, a hierarchical alignment strategy
that leverages such structure can provide significant improvements in unsupervised decoding from
ambiguous (symmetric) movement patterns. This suggests OT can be applied to a wider range of
neural datasets, and shows that a hierarchical strategy avoids local minima encountered by a global
alignment strategy that ignores clustered structure.

2 Background and related work

Transfer learning and distribution alignment. A fundamental goal in transfer learning is to
aggregate related datasets by learning a mapping between them. We wish to learn a transformation
\( T \in \mathcal{T} \), where \( \mathcal{T} \) refers to some class of transformations that aligns distributions under a notion of
probability divergence \( D(\cdot | \cdot) \) between a target distribution \( \mu \) and a reference (source) distribution \( \nu \):

\[
\min_{T \in \mathcal{T}} D(T(\mu) | \nu). \tag{1}
\]

Various probability divergences have been proposed in the literature, such as Euclidean least-squares
(when data ordering is known) [14, 15, 16], Kullback-Leibler (KL) [17], maximum mean discrepancy
(MMD) [18, 19, 20, 21], and the Wasserstein distance [10], where trade-offs are often statistical
(e.g., consistency, sample complexity) versus computational. Alignment problems are ill-posed since
the space of \( \mathcal{T} \) is large, so a priori structure is often necessary to constrain \( \mathcal{T} \) based on geometric
assumptions. Compact manifolds like the Grassmann or Stiefel [22, 23] are primary choices when
little information is present, as they preserve isometry. Non-isometric transformations, though richer,
demand much more structure (e.g., manifold or graph structure) [24, 25, 26, 27, 10].

Low-rank and union of subspaces models. Principal components analysis (PCA), one of the most
popular methods in data science, assumes a low-rank model where the top-\( k \) principal components of
a dataset provide the optimal rank-\( k \) approximation under an Euclidean loss. This has been extended
to robust (sparse errors) settings [12], and multi- (union of) subspace settings where data can be
partitioned into disjoint subsets where each subset of data is locally low-rank [28]. Transfer learning
methods based on subspace alignment [29, 30, 31] work well with zero-mean unimodal datasets, but
struggle on more complicated modalities (e.g., Gaussian mixtures or union of subspaces) due to a
mixing of covariances. Related to our work, [32] performs multi-subspace alignment by greedily
assigning correspondences between subspaces using chordal distances; this however discards valuable
information about a distribution’s shape.

Optimal transport. Optimal transport (OT) [33] is a natural type of divergence for registration
problems because it accounts for the underlying geometry of the space. In Euclidean settings, OT
gives rise to a metric known as the Wasserstein distance \( W(\mu, \nu) \) which measures the minimum effort
required to “displace” points across measures \( \mu \) and \( \nu \) (understood here as empirical point clouds).
Therefore, OT relieves the need for kernel estimation to create an overlapping support of the measures

3MATLAB code can be found at https://github.com/siplab-gt/hiwa-matlab. Neural datasets and Python code
are provided at http://nerdslab.github.io/neuralign
\[ \mu, \nu. \] Despite this attractive property, it has both a poor numerical complexity of \( O(n^3 \log n) \) (where \( n \) is the sample size) and a dimension-dependent sample complexity of \( O(n^{-1/d}) \), where the data dimension is \( d \) [34, 35]. Recently, an entropically regularized version of OT known as the Sinkhorn distance [36] has emerged as a compelling divergence measure; it not only inherits OT’s geometric properties but also has superior computational and sample complexities of \( O(n^2) \) and \( O(n^{-1/2}) \)\(^4\), respectively. It has also become a versatile building block in domain adaptation [10, 38]. Prior art [10] has largely exploited the OT’s push-forward as the alignment map since this map minimizes the OT cost between the source and target distributions while allowing a priori structure to be easily incorporated (e.g., to preserve label/graphical integrity). Such an approach, however, is fundamentally expensive when \( d \ll n \) since the primary optimization variable is a large transport coupling (i.e., \( \mathbb{R}^{n \times n} \)), while in reality the alignment mapping is merely \( \mathbb{R}^d \rightarrow \mathbb{R}^d \). Moreover, it assumes that the source and target distributions are close in terms of their squared Euclidean distance (i.e., an identity transformation), but this does not generally hold between arbitrary latent spaces.

**Hierarchical OT and related work.** The idea of learning an affine or unitary transformation to align datasets with an OT-based divergence has previously been studied in [39, 40, 41], a problem known as OT Procrustes. However, these methods don’t use problem-specific or clustered structure in data. Hierarchical OT is a recent generalization of OT [42, 43, 44] that is an effective and efficient way of injecting structure into OT but it has never been used to jointly solve alignment problems – our work represents a first attempt at doing so. Thus, a key contribution of this paper is putting both of these two ingredients together to develop a scalable strategy that leverages multimodal structure in data solve the OT Procrustes problem.

### 3 Hierarchical Wasserstein alignment

#### Preliminaries and notation.

Consider clustered datasets \( \{X_i \in \mathbb{R}^{D \times n_{x,i}}\}_{i=1}^{S} \) and \( \{Y_j \in \mathbb{R}^{D \times n_{y,j}}\}_{j=1}^{S} \) whose clusters are denoted with the indices \( i, j \) and whose columns are treated as \( \mathbb{R}^D \) embedding coordinates. The number of samples in the \( i \)-th (\( j \)-th) cluster of dataset \( X \) (dataset \( Y \)) is given by \( n_{x,i} \) (\( n_{y,j} \)). We express the empirical measures of clusters \( X_i \) and \( Y_j \) as \( \mu_i := \frac{1}{n_{x,i}} \sum_{k=1}^{n_{x,i}} \delta x_i(k) \) and \( \nu_j := \frac{1}{n_{y,j}} \sum_{l=1}^{n_{y,j}} \delta y_j(l) \), respectively, where \( \delta_x \) refers to a point mass located at coordinate \( x \in \mathbb{R}^D \). The squared 2-Wasserstein distance between \( \mu_i \) and \( \nu_j \) is defined as

\[
W_2^2(\mu_i, \nu_j) := \min_{Q \in \mathcal{U}(n_{x,i}, n_{y,j})} \sum_{k=1}^{n_{x,i}} \sum_{l=1}^{n_{y,j}} Q(k, l) \| X_i(k) - Y_j(l) \|^2
\]

where \( Q \) is a doubly stochastic matrix that encodes point-wise correspondences (i.e., the \( (k, l) \)-th entry describes the flow of mass between \( \delta x_i(k) \) and \( \delta y_j(l) \)). \( X_i(k) \) is the \( k \)-th column of matrix \( X_i \), and the constraint \( \mathcal{U}(m, n) := \{ Q \in \mathbb{R}^{m \times n} : Q 1_n = 1_m/n, Q^T 1_m = 1_n/m \} \) refers to the uniform transport polytope (with \( 1_m \) a length \( m \) vector containing ones). We will use \( \| \cdot \| \) to denote the operator norm, \( X^\dagger \) to denote the pseudo-inverse of \( X \), and \( I_d \) to denote the \( d \times d \) identity matrix.

#### Overview.

Although unsupervised alignment is challenging due to the presence of local minima, the imposition of additional structure will help to prune them away. Our key insight is that hierarchical structure decomposes a complicated optimization surface into simpler ones that are less prone to local minima. We formulate a hierarchical Wasserstein approach to align datasets with known (or estimated) clusters \( \{ \mu_i \}_{i=1}^{S}, \{ \nu_j \}_{j=1}^{S} \) but whose correspondences are unknown. The task therefore is to jointly learn the alignment \( T \) and the cluster-correspondences:

\[
\min_{P \in \mathcal{B}_S, T \in \mathcal{T}} \sum_{i=1}^{S} \sum_{j=1}^{S} P_{i,j} W_2^2(T(\mu_i), \nu_j),
\]

where the matrix \( P \) encodes the strength of correspondences between clusters, with a large \( P_{i,j} \) value indicating a correspondence between clusters \( i, j \), and a small value indicating a lack thereof. We note that \( \mathcal{B}_S := \mathcal{U}(S, S) \) is a special type of transport polytope known as the \( S \)-th Birkhoff polytope. Interestingly, this becomes a nested (or block) OT formulation, where correspondences are resolved at two levels: the outer level resolves cluster-correspondences (via \( P \)) while the inner level resolves point-wise correspondences between cluster points (via the Wasserstein distance).

**Alignment over the Stiefel manifold.** Assuming clusters lie on subspaces and principal angles between subspaces are “well preserved” across \( X \) and \( Y \) (make this precise in Theorem 4.2), an

\[ \text{Dependent on a regularization parameter [37].} \]
isometric transformation suffices. Hence, we solve (2) with \( T \leftarrow \mathcal{V}_{D,D} \), the Stiefel manifold which is defined as \( \mathcal{V}_{k,d} := \{ \mathbf{R} \in \mathbb{R}^{k \times d} : \mathbf{R}^\top \mathbf{R} = \mathbf{I}_d \} \). Explicitly, we can re-formulate equation (2) as:

\[
\min_{\mathbf{P}, \mathbf{R}, \{Q_{ij}\}} \sum_{i,j} P_{ij} C_{ij}(\mathbf{R}, Q_{ij}) \quad \text{s.t.} \quad \mathbf{P} \in \mathcal{B}_S, \quad \mathbf{R} \in \mathcal{V}_{D,D}, \quad Q_{ij} \in \mathcal{U}(n_{x,i}, n_{y,j}),
\]

(3)

where \( C_{ij}(\mathbf{R}, Q_{ij}) := \frac{1}{D} \sum_{k,l} Q_{ij}(k,l) \|RX_i(k) - Y_j(l)\|_2^2 \).

measures pairwise cluster divergences using the squared 2-Wasserstein distance under a Stiefel transformation \( \mathbf{R} \) acting on the \( i^{th} \) cluster.

Finally, we include entropic regularization over transportation couplings \( \mathbf{P} \) and all \( Q_{ij} \)’s to modify the Wasserstein distances to Sinkhorn distances, so as to take advantage of its superior computational and sample complexities. Omitting constraints for brevity, our final problem is given as

\[
\min_{\mathbf{P}, \mathbf{R}, \{Q_{ij}\}} \sum_{i,j} (P_{ij} C_{ij}(\mathbf{R}, Q_{ij}) + H_{\gamma_2}(Q_{ij})) + H_{\gamma_1}(\mathbf{P}),
\]

(5)

where \( \gamma_1, \gamma_2 > 0 \) are the entropic regularization parameters and the negative entropy function is defined as \( H_{\gamma}(\mathbf{P}) := \gamma \sum_{i,j} P_{ij} \log P_{ij} \). Parameters \( \gamma_1, \gamma_2 \) control the correspondence entropy, therefore (5) approximates (3) when \( \gamma_1, \gamma_2 > 0 \), but reverts to the original problem (3) as \( \gamma_1, \gamma_2 \to 0 \).

**Distributed ADMM approach.** Problem (5) is non-convex due to multilinearity in the objective and its Stiefel manifold domain. Although alternating directions method of multipliers (ADMM) is a convergent convex solver framework [45, 46], it is being applied in increasingly many non-convex settings [47]. Since (5) readily admits a splitting structure that separates the individual \( C_{ij} \) blocks, we develop a distributed ADMM approach. We proceed to split (5) as follows:

\[
\min_{\mathbf{P}, \mathbf{R}, \{Q_{ij}\}} \sum_{i,j} \left( P_{ij} C_{ij}(\mathbf{R}_{ij}, Q_{ij}) + H_{\gamma_2}(Q_{ij}) \right) + H_{\gamma_1}(\mathbf{P}) \quad \text{s.t.} \quad \mathbf{R}_{ij} = \tilde{\mathbf{R}}, \ \forall i,j,
\]

noting that the set constraints are omitted for brevity. The augmented Lagrangian is given by

\[
L_{\mu} = \sum_{i,j} \left( P_{ij} C_{ij}(\mathbf{R}_{ij}, Q_{ij}) + \left( \frac{\mu}{D} A_{ij}, \mathbf{R}_{ij} - \tilde{\mathbf{R}} \right) + \frac{\mu}{2D} \|\mathbf{R}_{ij} - \tilde{\mathbf{R}}\|_F^2 + H_{\gamma_2}(Q_{ij}) \right) + H_{\gamma_1}(\mathbf{P}),
\]

where \( \mu > 0 \) is the ADMM parameter and \( \{A_{ij} \} \) are Lagrange multipliers. Full details of the update steps are included in the Supplementary Material. The algorithm may be summarized in two steps (Alg. 1): (i) a distributed step that asks all cluster pairs to individually find their optimal transformations \( \mathbf{R}_{ij} \) in parallel, and (ii) a consensus step that aggregates all the locally estimated transformations according to a weighting that is proportional to correspondence strengths \( P_{ij} \).

**Parameters.** Entropic parameters \( \gamma_1, \gamma_2 \) relax the one-to-one cluster correspondence assumption, balancing a trade off between alignment precision (small \( \gamma \)) and sample complexity (large \( \gamma \)). Numerically, negative entropy adds strong convexity to the program, reducing sensitivity towards perturbations at the cost of a slower convergence rate. The ADMM parameter \( \mu \) controls the ‘strength’ of the consensus, or from an algorithmic viewpoint, the gradient step size.

**Distributed consensus.** Update steps for \( Q_{ij}, \mathbf{R}_{ij}, \mathbf{L}_{ij} \) can be performed in parallel over all cluster pairs (\( S^2 \) in total), making it amenable for a distributed implementation. The runtime complexity of this algorithm is presented in the supplementary Materials.

**Robustness against initial conditions.** We intentionally build robustness against initial conditions by ordering updates for \( \mathbf{R}_{ij} \) and \( Q_{ij} \) before \( \mathbf{P} \) such that when \( \mu \) is sufficiently small, the ADMM sequence is influenced more by the data than by initial conditions.

4 **Theoretical guarantees for cluster-based alignment**

While the previous section explains *how* to align clustered datasets, in this section, we aim to answer the question of *when* and *how well* they can be aligned. We provide necessary conditions for cluster-based alignability as well as *alignment perturbation bounds* according to equation (3)’s formulation. To simplify our analysis, we make the following assumptions: (i) each of the clusters contain the same number of datapoints \( n \), (ii) the ground truth cluster correspondences are \( \mathbf{P}^* = \mathbf{I}_S/S \).
Algorithm 1 Hierarchical Wasserstein Alignment (HiWA) Algorithm

1: procedure HIERARCHICALWASSERSTEINALIGNMENT($\gamma_1, \gamma_2, \mu, \{X_i \}_{i=1}^S, \{Y_j \}_{j=1}^S$)
2: $R \leftarrow \text{random } \mathcal{V}_{D,D}, \ P \leftarrow \mathbb{I}_S \mathbb{T}/S^2, \ \Lambda_{ij} \leftarrow 0, \ \forall i, j$ \hspace{1cm} $\triangleright$ Initialization
3: while not converged do
4: for all $i, j$ in parallel do
5: \hspace{1cm} $Q_{ij} \leftarrow \mathbb{I}_{n_x}, \mathbb{I}_{n_y}/n_x,n_y$
6: \hspace{1cm} while not converged do
7: \hspace{1cm} $R_{ij} \leftarrow \text{STIEFELALIGNMENT}(2P_{ij}Y_jQ_{ij}^T X_i^T + \mu(R - \Lambda_{ij}))$
8: \hspace{1cm} $Q_{ij} \leftarrow \text{SINKHORN}(\gamma_2/P_{ij}, C(k,l) \leftarrow \frac{1}{D}\|R_{ij}X_i(k) - Y_j(l)\|_2^2)$
9: \hspace{1cm} end while
10: end for
11: $P \leftarrow \text{SINKHORN}(\gamma_1, C(i,j) \leftarrow C_{ij}(R_{ij}, Q_{ij}))$
12: $R \leftarrow \text{STIEFELALIGNMENT}(\sum_{i,j} R_{ij} + \Lambda_{ij})$
13: $\Lambda_{ij} \leftarrow \Lambda_{ij} + R_{ij} - R, \ \forall i, j$
14: end while
15: end procedure

Algorithm 1: Hierarchical Wasserstein Alignment (HiWA) Algorithm

1: procedure SINKHORN($\gamma, C \in \mathbb{R}^{m \times n}$)
2: $K \leftarrow \exp(-C/\gamma), \ \nu \leftarrow \frac{\nu}{m}$
3: while not converged do
4: \hspace{1cm} $u \leftarrow \frac{u}{m} \odot KV$
5: \hspace{1cm} $v \leftarrow \frac{v}{n} \odot K^T u$
6: end while
7: $P \leftarrow \text{diag}(u)K\text{diag}(v)$
8: end procedure

1: procedure STIEFELALIGNMENT($A$)
2: $(U, \Sigma, V) \leftarrow \text{SVD}(A)$
3: $R \leftarrow UV^T$
4: end procedure

Notation:
\(\odot\): elementwise division
\(\exp(\cdot)\): elementwise exponential
\(\text{diag}(\cdot)\): diagonal matrix of argument

Algorithm 1: Hierarchical Wasserstein Alignment (HiWA) Algorithm

diagonal containing 1/S). However, this analysis can be extended to the case where the number of points is unequal without loss of generality. Detailed proofs are given in the Supp. Material.

The following result is a criterion that, if met, ensures the existence of a global minimizer of the cluster correspondence $P^*$. This criterion requires that matched clusters must be closer in Wasserstein distance than mismatched clusters, according to a threshold determined by Wasserstein’s sample complexity (i.e., an asymptotic rate dependent on the clusters’ sample sizes and intrinsic dimensions). Since these sample complexity results are based on the Wasserstein distance, we expect a less stringent criterion when using the Sinkhorn distance in (5) (due to superior sample complexity [37]).

Theorem 4.1 (Correspondence disambiguity criterion). Let all clusters be strictly low-rank where the dimension of the $i$-th cluster in the $x$-th dataset is $d_{x,i}$. Let $d_{x,i}, d_{y,j} > 4, \ \forall i, j \in [S]$. Define $\hat{C}_{ij} \equiv \min_{R \in \mathcal{V}_{D,D}, Q_{ij} \in \mathcal{B}_C} C_{ij}(R, Q_{ij})$. Problem (3) yields the solution $P^* = I_S/S$ with probability at least $1 - \delta$ if, $\forall i, j : i \neq j$, the following criterion is satisfied:

$$C_{ij}^x + C_{ji}^y - \hat{C}_{ii}^x - \hat{C}_{jj}^y > B_{x,i}(\delta) + B_{y,j}(\delta) + B_{x,j}(\delta) + B_{y,i}(\delta)$$

where $B_{z,\delta}(\delta) := c_{z, kn} - \frac{\delta}{2} + \frac{\sqrt{\log(1/\delta)}/2n}{c_{z,k} = 1458(2 + \frac{1}{3d_{z,k}/2^2 - 1})}$. 

Proof sketch. The proof contains two parts. In the first part, we consider perturbation conditions of the cost matrix $C$ in a (non-variational) optimal transport program over the Birkhoff polytope. To be unperturbed from $P^* = I_S/S$, we require that $C_{ij} + C_{ji} - \hat{C}_{ii} - \hat{C}_{jj} > 0, \ \forall i, j : i \neq j$. In the second part, we extend this condition to the the finite-sample regime by utilizing recently developed concentration bounds [35] for the p-Wasserstein distance, which essentially raises the disambiguity lower bound due to finite-sample uncertainty. (Supp. Material, Section 2)

Now, even if we know the global correspondence $P^*$, we still do not have the full picture about the alignment’s quality. For example, all matching clusters may have very similar covariances, but principal angles between the clusters are “distorted” across the datasets. Our next theorem gives us an upper bound on the alignment error (for unitary transformations), and makes precise the notion of global structure distortion.
Theorem 4.2 (Cluster-based alignment perturbation bounds). Consider data matrices \( \{ X_i, Y_i \in \mathbb{R}^{D \times n} \}_{i=1}^c \) with known point-wise correspondence matrices \( Q_{ii} \in \mathbb{B}_n \). Define matrices
\[
X := [X_1 Q_{11}, X_2 Q_{22}, \ldots, X_c Q_{cc}], \quad Y := [Y_1, Y_2, \ldots, Y_c].
\]
Set \( \varepsilon^2 := \| Y^\top Y - X^\top X \|_F \). If the criterion stated in theorem 4.1 is satisfied, \( X \) is full row rank, and \( \varepsilon \| X^\top \| \leq \frac{1}{\sqrt{2}} (\| X \| \| X^\top \|)^{-1/2} \), then
\[
\min_{P \in \mathbb{R}, R \in \mathbb{R}^{D, D}} \sum_{i,j} P_{ij} C_{ij}(R) \leq (\| X \| \| X^\top \| + 2) \| X^\top \|^2 \varepsilon^2 + D,
\]
where \( D = \sum_{i=1}^c \text{tr}(X_i (I/n - Q_{ii} Q_{ii}^\top) X_i^\top) + (1/n - 1) Y_i Y_i^\top) \) is a data-dependent constant.

Proof sketch. We utilize a recent perturbation result on the Procrustes problem (on a Frobenius norm objective) by Arias-Castro et al. [48] and adapt it to our squared 2-Wasserstein objective. (Supp. Material, Section 3)

Note that \( \varepsilon \) plays a major role in the alignment error bound and quantifies the notion of global structure distortion, which allows us to understand on how phenomena like covariate shift or misclustering impacts alignment. To shed some light in this regard, we consider a simple analysis on a cluster-pair’s error contribution to \( \varepsilon \), denoted as \( \varepsilon_{ij} \). Consider the decomposition of the \((i, j)\)-th block of the Gramians related to clusters \( i \) and \( j \), where their respective singular value decompositions are \( X_i Q_{ii} = A_i \Sigma_{x,i} V^\top \) and \( Y_j = B_j \Sigma_{y,j} V^\top \). Defining the blockwise error between clusters \( i, j \)
\[
\varepsilon_{ij} := \| Y_j^\top Y_j - Q_{ij} X_i^\top X_j Q_{ij} \|_F = \| \Sigma_{y,j} B_j^\top B_j \Sigma_{y,j} - \Sigma_{x,i} A_i^\top A_j \Sigma_{x,i} \|_F,
\]
two components stand out: (i) angular shift, which is characterized by differences in principal angles between \( B_j^\top B_j \) and \( A_i^\top A_j \), and (ii) spectral shift, which is characterized by differences in spectra.

Finally, we show that the subspace configuration of a dataset’s clusters can also affect alignment. Pretend for a moment that external alignment information were present to aid in the disambiguation between two clusters. The following lemma tells us when such information is useless (Proof in Supp. Material, Section 4).

Lemma 4.3 (Uninformative alignment). Consider clusters \( X_i, Y_j \in \mathbb{R}^{D \times n} \) and known point-wise correspondences \( Q_{ij} \in \mathcal{U}(n,n) \). Denote the left and right singular vectors of \( Y_j Q_{ij}^\top X_i^\top \) associated with the non-zero singular values as \( \hat{U}, \hat{V} \in \mathbb{R}^{D \times r} \) with \( r \leq D \). Define the set of orthogonal transformations that are constrained to agree with known angular directions as
\[
T(U', V') := \{ R \in \mathbb{R}^{D \times D}_+ : R^\top R = I, RV' = U' \},
\]
where \( U', V' \in \mathbb{V}_{D,r} \) with \( r \leq D \). Given \( U', V' \in \mathbb{R}^{D \times r'} \) with \( r' \leq D \), we have
\[
\min_{R \in T(U', V')} C_{ij}(R) \geq \min_{R \in \mathbb{V}_{D,D}} C_{ij}(R), \tag{6}
\]
with equality holding when \( (\hat{U}, \hat{U}') = (V, V') \).

Direct consequences of this lemma are the following: When a dataset has equally-spaced subspaces, it has a maximally uninformative geometric configuration since angular information from other clusters (i.e., \( U', V' \)) can never increase the inter-cluster distance \( C_{ij} \) (i.e., equality in (6) always holds); it is hence a worst-case scenario for alignment. This also explains why alignment in very high-dimensional space is harder: All subspaces may be orthogonal to each other, and hence offer no “geometric” advantage.

5 Numerical experiments

5.1 Synthetic low-rank Gaussian mixture dataset

In this section, we validate our method as well as demonstrate its limiting characteristics under symmetric-subspace and finite-sample regimes. To generate our synthetic data, we repeat the following procedure for each of the \( S \) clusters. We first randomly generate Gaussian distribution parameters \( \mu_i \in \mathbb{R}^d, \Sigma_i \in \mathbb{R}^{d} \); \( \Sigma_i \succeq 0 \) (positive semi-definite), then randomly sample \( n \) data-points from these parameters, and finally project them into a random subspace \( \mathcal{V}_i \in \mathbb{R}^{D \times d} \) in a \( D > d \) dimensional embedding. In these experiments, we assume that the clusters are known, but the
To understand how global geometry impacts alignment, we applied HiWA in two different settings (average-case, solid) versus equally-spaced (worst-case, dashed) for electrophysiology measurements/readouts [31, 39, 31] for \( n = 50, d = 2 \), and (e) \( S = 5, D = 6 \), and (f) \( S = 2, D = 2 \).

Next, we studied the effect of dimensions \( d \) and sample size \( n \) on the accuracy of alignment (Figure 1 (c-d)). We tested HiWA across various dataset conditions by varying parameters \( d = \{2, 3, 4, 5\} \) and \( n = \{12, 25, 50, 100, 200\} \) while approximately maintaining the average subspace correlations (i.e., \( \mathbb{E}\|V_i^T V_j\| \)) by fixing the cluster size \( S = 5 \) and tuning \( D \) to control the subspace spacing. In both cases, sample complexities are better than the theoretical rate of \( O(n^{-1/d}) \), which is likely due to the Sinkhorn distance’s superior sample complexity. In Figure 1e-f, we conduct an ablation study and evaluate our algorithm against benchmark methods in transfer learning and point set registration in two settings: a simple one in low- \( d \) (e) and a harder one in higher- \( d \) (f). Specifically, we compare HiWA when clusters are known (but pairwise correspondences are unknown), HiWA with clustering via sparse subspace clustering [12] (HiWA-SSC) to represent completely unsupervised alignment, a Wasserstein alignment variant with no cluster-structure (WA) which is akin to OT Procrustes [50, 39, 40, 41], subspace alignment [29], correlation alignment [31], and iterative closest point (ICP) [49]. HiWA exhibits strongest performance, with HiWA-SSC trailing closely behind (since clusters are independently resolved), followed by WA, then other algorithms. Subspace alignment methods have remarkably poor performance in higher dimensions due to their inability to resolve subspace sign ambiguities, while ICP demonstrates its notorious dependence on good initial conditions. These results indicates HiWA’s strong robustness against initial conditions and good scaling properties.

5.2 Neural population decoding example
Decoding intent (e.g., where you want to move your arm) or evoked responses (e.g., what you are looking at or listening to) directly from neural activity is a widely studied problem in neuroscience, and the first step in the design of a brain machine interface (BMI). A critical challenge with BMIs is that neural decoders need to be recalibrated (or re-trained) due to drift in neural responses or electrophysiology measurements/readouts [51]. A recent method for semi-supervised brain decoding finds a transformation between projected neural responses and movements by solving a KL-divergence
Figure 2: Results on neural decoding dataset: How distribution alignment is used to translate neural activity into movement – low-dimensional embeddings of neural data are aligned with target movement patterns (a). In (b), we compare the performance (cluster correspondence) of HiWA, WA, and DAD as the number of points in the source dataset decreases. Next, we compared the performance of HiWA with known and estimated clusters (via GMM). Movement patterns in which cluster separability is high and the geometry is preserved across datasets, can be aligned in both cases (green stars). Patterns where separability is low but geometry is useful can be aligned when the cluster arrangements are known are denoted with yellow stars.

minimization problem [52]. Using this approach, one could build robust decoders that work across days and shifts in neural responses through alignment.

We test the utility of hierarchical alignment for neural decoding on datasets collected from the arm region of primary motor cortex of a non-human primate (NHP) during a center out reaching task [52]. After spike sorting and binning the data, we applied factor analysis to reduce the data dimensionality to 3D (source distribution) and applied HiWA to align the neural data to a 3D movement distribution (target distribution) (Figure 2). We compared its performance to (procrustes) Wasserstein alignment (WA) without hierarchical structure, and a baseline brute force search method called distribution alignment decoding (DAD) [52]. We examined the prediction accuracy of the target reach direction for the motor decoding task (i.e., the cluster classification accuracy).

Next, we examined the impact of the sampling density (Figure 2b) on alignment performance. Our results demonstrate that HiWA continues to produce consistent cluster correspondences (> 70% accuracy), even as the number of samples per cluster drops to 8. In comparison, DAD is competitive at larger sample sizes but its performance rapidly drops off as sampling density decreases because it requires estimating a distribution from samples. WA suffers from the presence of many local minima and fails to find the correct cluster correspondences. Our results suggest that HiWA consistently provides stable solutions, outperforming competitor methods for this application.

Finally, to study the impact of local and global geometry on whether an unlabeled source and target can be aligned, we applied HiWA to permutations of eight subsets of reach directions (movement patterns). When just two reach directions are considered (Figure 2c, Columns 1–4), global geometry becomes useless in determining the correct rotation. In this case, we observe that HiWA is only capable of consistent alignment when cluster asymmetries are sufficiently extreme in both the source and target. When three reach directions are considered (Figure 2c, Columns 5–8), the global geometry can be used, yet there still exist symmetrical cases where recovering the correct rotation is difficult without adequate local asymmetries or some supervised (labeled) data to match clusters. These results suggest that hierarchical structure can be critical in resolving ambiguities in alignment of globally symmetric movement distributions.

6 Conclusion

This paper introduces a new method for hierarchical alignment with Wasserstein distances, provided an efficient numerical solution with analytical guarantees. We tested our method and compared its performance against other methods on a synthetic mixture model dataset and on a real neural decoding dataset. Future directions include extensions to non-rigid transformations, and applications to higher dimensional neural datasets that do not rely on external measured behavioral covariates.
Acknowledgments
JL was supported by DSO National Laboratories of Singapore, ED and MD were supported by NSF grant IIS-1755871, and CR was supported by NSF grant CCF-1409422 and CAREER award CCF-1350954.

References


1 ADMM algorithmic details

1.1 Derivation of ADMM update steps

ADMM admits the following sequence of updates:

\[
\begin{align*}
(R_{ij}^{(t+1)}, Q_{ij}^{(t+1)}) & \leftarrow \min_{R_{ij} \in S(D,D)} \min_{Q_{ij} \in U(n_x, n_y)} P_{ij}^{(t)} C_{ij}(R_{ij}, Q_{ij}) + \frac{\mu}{2D} \|R_{ij} - R_{ij}^{(t)} + \Lambda_{ij}^{(t)}\|_F^2 + H_{\gamma_2}(Q_{ij}), \\
P^{(t+1)} & \leftarrow \min_{P \in B} \sum_{i,j} P_{ij} C_{ij}(R_{ij}^{(t+1)}, Q_{ij}^{(t+1)}) + H_{\gamma_1}(P), \\
R^{(t+1)} & \leftarrow \min_{R \in S(D,D)} \sum_{i,j} \|R_{ij}^{(t+1)} - R + \Lambda_{ij}^{(t)}\|_F^2, \\
\Lambda_{ij}^{(t+1)} & \leftarrow \Lambda_{ij}^{(t)} + R_{ij}^{(t+1)} - R.
\end{align*}
\]

Update (1) involves an alternating minimization over \(Q_{ij}\) and \(R_{ij}\) whereby the first variable is fixed while the second is minimized, followed by the second fixed and the first minimized, and the procedure is repeated until convergence is achieved. When solving for \(R_{ij}\) we have the following Stiefel manifold optimization:

\[
R_{ij}^{(t+1)} \leftarrow \min_{R_{ij} \in S(D,D)} P_{ij}^{(t)} C_{ij}(R_{ij}, Q_{ij}) + \frac{\mu}{2D} \|R_{ij} - R_{ij}^{(t)} + \Lambda_{ij}^{(t)}\|_F^2
= \max_{R_{ij} \in S(D,D)} \text{tr} \left( (2P_{ij}^{(t)} Y_j Q_{ij}^T X_i^T + \mu(R_{ij} - \Lambda_{ij}^{(t)})) R_{ij}^T \right) = UDV^T,
\]

where \(2P_{ij}^{(t)} Y_j Q_{ij}^T X_i^T + \mu(R_{ij} - \Lambda_{ij}^{(t)}) = UDV^T\) is its SVD. We employ the Sinkhorn algorithm (Algorithm 1 of [1]) to solve for \(Q_{ij}\), using an entropic parameter of \(\gamma_2/P_{ij}^{(t)}\) and uniform marginals. \(R_{ij}^{(t+1)}\) and \(Q_{ij}^{(t+1)}\) are retrieved once the alternating minimization converges.

Update (2) also employs the Sinkhorn algorithm over the cost matrix generated by \(C_{ij}(R_{ij}^{(t+1)}, Q_{ij}^{(t+1)})\) using variables found in update (1), along with an entropic parameter of \(\gamma_1\) and uniform marginals.

Update (3) is a consensus update over a Stiefel manifold optimization:

\[
R^{(t+1)} \leftarrow \min_{R \in S(D,D)} \sum_{i,j} \|R_{ij}^{(t+1)} - R + \Lambda_{ij}^{(t)}\|_F^2 = UDV^T,
\]

where \(\sum_{i,j} R_{ij}^{(t+1)} + \Lambda_{ij}^{(t)} = UDV^T\) is its SVD.

Preprint. Under review.
1.2 Computational complexity of distributed ADMM algorithm

The main computational complexity of the ADMM algorithm comes from line 8, which demands the solving of $S^2$ Sinkhorn problems per ADMM iteration. Fortunately, these Sinkhorn problems (i.e., update steps for $Q_{ij}, R_{ij}, L_{ij}$) may be conducted in parallel, making it amenable for a distributed implementation. When fully parallelized, the algorithm has a per-iteration computational complexity of $O(n_i n_j)$, where $n_i, n_j$ refers to the number of points in the largest clusters of $X, Y$ respectively (compared to vanilla Sinkhorn’s $O(n_x n_y)$ complexity where $n_x, n_y$ refers to the total number of points in respective datasets, assuming $D \ll \max(n_i, n_j)$). In Figure 1, we record run times of two versions of the same algorithm, with and without parallelism, operating on a similar test dataset, at varying cluster sizes $S$ (50 data points per cluster). 10 random trials were conducted and the means and standard deviations were reported (error lines). As expected, the parallel implementation scales better. These results were computed using an i7 intel processor with 6 cores clocked at 2.6 GHz.

Figure 1: Run time comparison between a parallelized and a non-parallelized implementation: We varied the number of clusters – the main bottleneck to run time - to compare the runtimes of the parallelized and non-parallelized implementations.

1.3 Stopping criteria.

In lines 3 and 6 of Algorithm 1, possible stopping criteria are (i) $\| R^{(t+1)} - R^{(t)} \|_F \leq \tau$ where the difference is between the current and previous iteration’s transformation and $\tau$ is the tolerance, and (ii) $t \leq T$ where $T$ is the maximum number of iterations.

2 Proof of Theorem 4.1

2.1 Part 1: Geometric perturbations conditions over the Birkhoff polytope

First, we provide the following lemma that illuminates some basic geometrical insights of the general OT cost matrix, whose coupling is in the Birkhoff polytope $B_c := \{ P \in \mathbb{R}^{c \times c} : P 1_c = P^1 1_c = 1_c \}$. We define the OT program with respect to a cost matrix $C$ as

$$L(C) := \arg \min_{P \in B_c} \langle P, C \rangle.$$ (5)

The following lemma describes the conditions on $C$ for $P$ to remain unperturbed at $P^*$.

**Lemma 2.1.** Define the set of vertices on the $c$-th Birkhoff polytope $B_c$ that are within a $\delta$-ball from $P^*$ as

$$P_\delta(P^*) = \{ P \in B_c \setminus \{ P^* \} : \| P - P^* \|_F \leq \delta \}. \quad (6)$$

Define the set of matrices that denote directions from $P^*$ to each neighboring vertex $P_i$ as

$$V_\delta(P^*) = \{ V \in \mathbb{R}^{c \times c} : V = P - P^*, \ P \in P_\delta(P^*) \}. \quad (7)$$

For the linear program’s solution $P^* = L(C)$ to remain unchanged, $\forall V \in V_\delta(P^*)$ for $\delta = 2$, the cost matrix $C$ should satisfy

$$\langle C, V \rangle > 0. \quad (8)$$
With this parameterized form, we may reexpress the optimization for
\[ P^* \] remains unchanged, \( C \) should not cause \( P^* \) to move to an adjacent edge of the Birkhoff polytope, nor should it cause it to extend beyond its adjacent edge because then, the neighboring vertex would assume the new solution. For the rest of this proof, we shall let \( P^* = I \) without any loss of generality. We define the set of nearest neighbors to \( P^* \), which are simply permutation matrices that can be described as taking \( P^* \) and exchanging any two columns. For a \( P \) with any two columns of \( P^* \) exchanged, notice that the difference matrix \( V = P - P^* \) is a symmetric matrix of mostly zeros except for two off-diagonal \( +1 \) entries and two diagonal \( -1 \) entries, hence \( \|P - P^*\|_F = 2 \). Formally, we describe the set of nearest neighbors with \( \mathcal{P}_\delta(P^*) \) with \( \delta = 2 \), defined by (6). Next, we define edges that are adjacent to \( P^* \) using the set \( \mathcal{V}_\delta(P^*) \) also with \( \delta = 2 \), defined by (7). Note that there are \( K := \binom{\delta}{2} - 1 = \frac{(c-1)c}{2} - 1 \) neighboring vertices and adjacent edges.

We will now show how perturbing \( C \) in just one direction \( V \) changes \( P^* \). First, consider a cost matrix that produce \( P^* \), which is defined as \( C^* := \sum_{V \in \mathcal{V}_\delta(P^*)} V_i \), meaning that it is equi-angle from all \( V \in \mathcal{V}_\delta(P^*) \), for \( \delta = 2 \). By enumerating over all \( V \), we may derive \( C^* = \mathbb{1}^\top - cI \), along with the fact that \( \frac{\langle C^*, V \rangle}{\|C^*\|_F \|V\|_F} = \frac{1}{\sqrt{\delta}} \). Define \( P_1 = L(C_1) \) with \( C_1 \) such that \( \langle V_1, C_1 \rangle \leq 0 \) and \( \|V_1, C_1\|_F = \frac{1}{\sqrt{\delta}} \), \( \forall i \in 2, \ldots, K \). Define \( P_2 = L(C_2) \) with \( C_2 \) such that \( \frac{\langle V_1, C_2 \rangle}{\|V_1, C_2\|_F} = \frac{1}{\sqrt{\delta}} \), \( \forall i \in [K] \). We make the claim that \( P_1 = P_2 \) and proceed with a proof by contradiction. As mentioned before, \( \forall V \in \mathcal{V}_\delta(P^*) \) for \( P^* = I \), \( \delta = 2 \) has exactly four non-zero entries, i.e., \( V_{ii} = V_{jj} = -1 \) and \( V_{ij} = V_{ji} = +1 \), where \( i, j \in [c], i \neq j \). Writing out \( \langle V, C \rangle \) explicitly, we have \( \langle V, C \rangle = - (C_{ij} + C_{ji}) + (C_{ii} + C_{jj}) \). To ensure that \( C_1 \) and \( C_2 \) does not interact with other edges of the polytope, we fix \( \langle V_i, C \rangle = \langle V_i, C^* \rangle = c \) for \( i = 2, \ldots, K \). Since we constructed \( C_1 \) and \( C_2 \) to differ only by the condition \( \langle V_1, C_1 \rangle \leq 0 \) or \( \langle V_1, C_2 \rangle > 0 \), and any \( V_i \) affects only four entries of the cost matrix, we may greatly simplify our analysis of \( C \) and \( P \) to only four entries. As such, we extract these four entries of \( P \), represent it using \( \hat{P} \in B_2 \), and parameterize using \( t \in [0, 1] \) as
\[
\hat{P}(t) = t \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + (1-t) \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.
\]

With this parameterized form, we may reexpress the optimization for \( L(C) \) as
\[
\min_{t \in [0,1]} t(C_{ii} + C_{jj}) + (1-t)(C_{ij} + C_{ji}) = \min_{t \in [0,1]} \left( \frac{C_{ii} + C_{jj}}{C_{ij} + C_{ji}} - 1 \right) t.
\]
The above minimization has three cases. If \( C_{ii} + C_{jj} = C_{ij} + C_{ji} \), then there exists no unique solution \( t \). If \( C_{ii} + C_{jj} > C_{ij} + C_{ji} \), then \( t = 0 \). If \( C_{ii} + C_{jj} < C_{ij} + C_{ji} \), then \( t = 1 \). The first two cases directly corresponds to \( \langle V_1, C_1 \rangle \leq 0 \), while the third case corresponds to \( \langle V_1, C_2 \rangle > 0 \). The fact that \( t \) is not consistent between all cases demonstrates a contradiction. Moreover, \( t = 1 \) produces the solution \( \hat{P} = I \), and if this holds for all \( V \in \mathcal{V}_\delta(P^*) \), then all off-diagonal entries must be zero and therefore \( P^* = I \) must be the minimizer for \( L(C) \).

A direct consequence of lemma 2.1 is the following.

**Corollary 2.2.** The solution to the linear program defined by (5) is \( P^* = I \) if the linear cost matrix \( C \) satisfies the following property
\[
C_{ij} + C_{ji} - C_{ii} - C_{jj} > 0, \ \forall i, j \in [c], i \neq j.
\]

**Proof.** Analyzing (8), we observe that any \( V \in \mathcal{V}_\delta(P^*) \) for \( \delta = 2 \) has only four symmetric non-negative entries, which we condense \( V \) and \( C \) into \( \mathbb{R}^{2 \times 2} \) matrices at these four support locations respectively as \( \bar{V} = \begin{bmatrix} -1 & +1 \\ +1 & -1 \end{bmatrix} \) and \( \bar{C} = \begin{bmatrix} C_{ii} & C_{ij} \\ C_{ij} & C_{jj} \end{bmatrix} \). It thus follows that an explicit computation produces:
\[
\langle P, C \rangle \geq 0 \Rightarrow \langle \hat{P}, \bar{C} \rangle = (C_{ij} + C_{ji}) - (C_{ii} + C_{jj}) \geq 0.
\]

Since the set \( \mathcal{V}_\delta(P^*) \) spans all permutations between \( i, j \in [c], i \neq j \), we conclude with (10).
In the variational setting, $C_{ij}(R, Q_{ij})$’s are themselves linearly coupled with $R$ and $Q_{ij}$. The following proposition introduces a trivial criterion on $C_{ij}(R, Q_{ij})$’s to guarantee that $P$ remains unperturbed from $P^*$. For pedagogical reasons, we shall assume that $n \to \infty$ for this proposition but subsequently relax this.

**Proposition 2.3** (Rotationally invariant disambiguity criterion). Problem (3) yields the solution $P^*$ if, $\forall i, j : i \neq j$, the following criterion is satisfied:

$$
\min_{R, Q_{ij}} C_{ij}(R, Q_{ij}) + \min_{R, Q_{ji}} C_{ji}(R, Q_{ji}) - \min_{R, Q_{ij}} C_{ii}(R, Q_{ii}) - \min_{R, Q_{ji}} C_{jj}(R, Q_{jj}) > 0.
$$

(11)

**Proof.** Consider a set of $c$ clusters where clusters $i, j$ satisfy

$$
C_{ij}(R_{ij}', Q_{ij}') + C_{ji}(R_{ji}', Q_{ji}') - C_{ii}(R', Q_{ii}') - C_{jj}(R', Q_{jj}') \leq 0,
$$

(12)

where

$$
(R_{ij}', Q_{ij}') := \arg \min_{R \in S(d,d), Q_{ij} \in B_n} C_{ij}(R), \quad (R_{ji}', Q_{ji}') := \arg \min_{R \in S(d,d), Q_{ji} \in B_n} C_{ji}(R),
$$

$$
(R', Q_{ii}', Q_{jj}') := \arg \min_{R \in S(d,d), Q_{ij} \in B_n} C_{ii}(R, Q_{ii}) + C_{jj}(R, Q_{jj}).
$$

Since the 2-Wasserstein is a valid metric, its distance between any two clusters must satisfy $C_{ij}(R) \geq 0$ with equality holding if and only if the clusters are exactly similar. If $P^* = I$, similar clusters are denoted with matching indices, and it must follow that $C_{ii}(R') + C_{jj}(R') = 0$. This implies that (12) must be false since $C_{ij}(R_{ij}'), C_{ji}(R_{ji}') > 0$ for mismatched clusters. Due to this contradiction, the disambiguity criterion (11) must hold for all cluster pairs $i, j : i \neq j$. 

This proposition provides a disambiguity criterion, requiring that matched clusters (i.e., $C_{ii}, C_{jj}$) should be more similar than mismatched clusters (i.e., $C_{ij}, C_{ji}$) up to some disambiguity threshold (in the case of $n \to \infty$, the threshold is 0). To extend this proposition to the finite-sample regime, we require a higher disambiguity-threshold to offset uncertainty due to finite samples.

### 2.2 Part 2: Disambiguity criterion in the finite-sample regime

We utilize a recent $p$-Wasserstein concentration bound by Weed and Bach [3] that describes finite sample behavior on the Wasserstein distance for data embedded in high-dimensional space, but whose clusters are themselves approximately low-dimensional. We will proceed our analysis with the language of probability measures $\mu$ to make our analysis consistent with [3]. We thus define the equivalent measure analogs as follows.

**Definition 2.4.** Let clusters $X_i \in \mathbb{R}^{D \times n_{x,i}}$ and $Y_j \in \mathbb{R}^{D \times n_{y,j}}$ be respectively denoted by empirical measures as

$$
\hat{\mu}_{x,i} := \frac{1}{n_{x,i}} \sum_{k=1}^{n_{x,i}} \delta_{X_i(k)}, \quad \hat{\mu}_{y,j} := \frac{1}{n_{y,j}} \sum_{k=1}^{n_{y,j}} \delta_{Y_j(k)},
$$

where $\delta_{X_i(k)}$ refers to a discrete point located at $X_i(k)$. At the limit, we denote the measures as

$$
\mu_{x,i} := \lim_{n_{x,i} \to \infty} \frac{1}{n_{x,i}} \sum_{k=1}^{n_{x,i}} \delta_{X_i(k)}, \quad \mu_{y,j} := \lim_{n_{y,j} \to \infty} \frac{1}{n_{y,j}} \sum_{k=1}^{n_{y,j}} \delta_{Y_j(k)}.
$$

**Definition 2.5.** Denote a linear transformation $R$ applied on the measure as

$$
R \circ \hat{\mu}_{x,i} := \frac{1}{n_{x,i}} \sum_{k=1}^{n_{x,i}} \delta_{RX_i(k)}, \quad R \circ \mu_{x,i} := \lim_{n_{x,i} \to \infty} \frac{1}{n_{x,i}} \sum_{k=1}^{n_{x,i}} \delta_{RX_i(k)}.
$$

The transformed inter-cluster distance between clusters may thus be denoted as

$$
\hat{C}_{ij}(R) := W^2_{p}(R \circ \hat{\mu}_{x,i}, \hat{\mu}_{y,i}), \quad \hat{C}_{ij}(R) := W^2_{p}(R \circ \mu_{x,i}, \mu_{y,i}).
$$
Now, we may proceed to state results from [3]. The following result pertains to the sample complexity of measures $\mu$ in $\mathbb{R}^d$ that are supported on an approximately low-dimensional set in $\mathbb{R}^d$, where $d \ll D$. First we require some definitions.

**Definition 2.6.** Given a set $S \subseteq \mathbb{R}^d$, let $N_{\varepsilon}(S)$ denote the $\varepsilon$-covering number of set $S$, which is defined as the minimum number of closed balls $B_1, \ldots, B_m$ of diameter $\varepsilon$ such that $S \subseteq \bigcup_{1 \leq i \leq m} B_i$.

**Definition 2.7.** For any set $S \subseteq \mathbb{R}^d$, the $\varepsilon$-fattening of $S$ is $S_\varepsilon := \{ y : D(y, S) \leq \varepsilon \}$.

**Proposition 2.8**. (Weed and Bach [3], Proposition 16). Let $S$ be a set that satisfies $N_{\varepsilon}(S) \leq (3\varepsilon')^{-d}$ for all $\varepsilon' \leq 1/27$ and for some $d > 2p$. Suppose there exists a positive constant $c$ such that $\mu$ satisfies $\mu(S_\varepsilon) \geq 1 - e^{-c\varepsilon^2/2\sigma^2}$ for all $\varepsilon > 0$. If $p \log \frac{1}{\varepsilon} \geq 1/18$, then for all $n \leq (18p\sigma^2 \log \frac{1}{\varepsilon})^{-d/2}$, $$\mathbb{E}[W_p^p(\mu, \hat{\mu}_n)] \leq cn^{-p/d},$$ where $c = 27p(2 + \frac{1}{3d_{z,k}/2-2-1})$.

This proposition states that the degree to which $\mu$ is concentrated (as parameterized by $\sigma$) around set $S$ (approximately supported in low-dimensions) affects how “long” (in terms of $n$) we can enjoy the fast convergence rate of $n^{-p/d}$. We will leverage this result to obtain the following theorem on cluster correspondence ambiguity with respect to sample complexity.

**Corollary 2.9.** Let sets $S_{z,i}$ and $S_{y,j}$ satisfy the conditions for $S$ in proposition 2.8 for some $\sigma_{z,i}, \sigma_{y,j} > 0$ and $d_{z,i}, d_{y,j} > 4$. If $\log \frac{1}{\sigma_{z,i}} \geq \frac{1}{36}$ and $\log \frac{1}{\sigma_{y,j}} \geq \frac{1}{36}$, then for all $n_{z,i} \leq (36\sigma_{z,i}^2 \log \frac{1}{\sigma_{z,i}})^{-d_{z,i}/2}$ and $n_{y,j} \leq (36\sigma_{y,j}^2 \log \frac{1}{\sigma_{y,j}})^{-d_{y,j}/2}$, $$\mathbb{E}[\|\hat{C}_{ij}(R) - \hat{C}_{ij}(R)\|] \leq c_{z,i}n_{z,i}^{-2/d_{z,i}} + c_{y,j}n_{y,j}^{-2/d_{y,j}},$$ where $$c_{z,k} = 729(2 + \frac{1}{3d_{z,k}/2-2-1}).$$

**Proof.** Denote $\mu_i, \mu_j$ as measures and $\hat{\mu}_i, \hat{\mu}_j$ as their empirical estimates. By the triangle inequality, $$\mathbb{E}[W_p^2(\hat{\mu}_i, \hat{\mu}_j)] \leq \mathbb{E}[W_2^2(\hat{\mu}_i, \mu_i) + W_2^2(\mu_i, \mu_j)] \leq \mathbb{E}[W_2^2(\hat{\mu}_i, \mu_i) + W_2^2(\mu_i, \hat{\mu}_j) + W_2^2(\hat{\mu}_j, \mu_j)]$$ $$\Rightarrow \mathbb{E}[\|W_2^2(\hat{\mu}_i, \mu_j) - W_2^2(\hat{\mu}_i, \hat{\mu}_j)\|] \leq \mathbb{E}[W_2^2(\hat{\mu}_i, \mu_i)] + \mathbb{E}[W_2^2(\mu_i, \hat{\mu}_j)]$$ $$\Rightarrow \mathbb{E}[\|\hat{C}_{ij}(R) - \hat{C}_{ij}(R)\|] \leq c_{z,i}n_{z,i}^{-2/d_{z,i}} + c_{y,j}n_{y,j}^{-2/d_{y,j}},$$ where the last line is a direct application of definition 2.5 and proposition 2.8.

**Lemma 2.10.** Let $S_{x,i}, S_{y,i}, S_{x,j}, S_{y,j}$ be sets that satisfy the conditions for $S$ in proposition 2.8 for some $\sigma_{x,i}, \sigma_{y,i}, \sigma_{x,j}, \sigma_{y,j} > 0$ and $d_{x,i}, d_{y,i}, d_{x,j}, d_{y,j} > 4$. If $\log \frac{1}{\sigma_{x,i}}, \log \frac{1}{\sigma_{y,i}}, \log \frac{1}{\sigma_{x,j}}, \log \frac{1}{\sigma_{y,j}} \geq \frac{1}{36}$, then for all $n_{x,i} \leq (36\sigma_{x,i}^2 \log \frac{1}{\sigma_{x,i}})^{-d_{x,i}/2}$, $n_{y,i} \leq (36\sigma_{y,i}^2 \log \frac{1}{\sigma_{y,i}})^{-d_{y,i}/2}$, $n_{x,j} \leq (36\sigma_{x,j}^2 \log \frac{1}{\sigma_{x,j}})^{-d_{x,j}/2}$, and $n_{y,j} \leq (36\sigma_{y,j}^2 \log \frac{1}{\sigma_{y,j}})^{-d_{y,j}/2}$, cluster correspondences in problem (3) may be disambiguated to achieve $P^* = I$ when the following criterion is fulfilled for all $i, j \in \{i, j\}$: $$\mathbb{E}[\min_{R} \hat{C}_{ij}(R) + \min_{R} \hat{C}_{ji}(R) - \min_{R} \hat{C}_{ii}(R) - \min_{R} \hat{C}_{jj}(R)] > 2(B_{x,i} + B_{y,i} + B_{x,j} + B_{y,j}),$$ where the constants are defined as $$B_{z,k} := c_{z,k}n_{z,k}^{-2/d_{z,k}}, \quad c_{z,k} := 729(2 + \frac{1}{3d_{z,k}/2-2-1}).$$

**Proof.** This follows a direct application of criterion (11) and corollary 2.9.

**Proposition 2.11** (Weed and Bach [3], Proposition 20). For all $n \geq 0$ and $0 \leq p < \infty$, $$\mathbb{P}[W_p^p(\mu, \hat{\mu}_n) \geq \mathbb{E}[W_p^p(\mu, \hat{\mu}_n)] + t] \leq \exp(-2nt^2).$$
Theorem 2.12. Define \( \hat{C}_{ij}^* := \min_{R \in S(D, D)} \hat{C}_{ij}(R) \). If the conditions in Lemma 2.10 are satisfied then problem (3) yields the solution \( P^* = I \) with probability at least \( 1 - \delta \) if, \( \forall i, j : i \neq j \), the following criterion is satisfied:

\[
\hat{C}_{ij}^* + \hat{C}_{ji}^* - \hat{C}_{ii}^* - \hat{C}_{jj}^* > 2(B_{x,i}(\delta) + B_{y,i}(\delta) + B_{x,j}(\delta) + B_{y,j}(\delta))
\]

where

\[
B_{z,k}(\delta) := c_{z,k}n_{z,k}\sqrt{\log(1/\delta)/2n_{z,k}}, \quad c_{z,k} := 729\left(2 + \frac{1}{3d_{z,k}^2 - 2 - 1}\right), \quad (13)
\]

where \( d_{z,k} \) refers to the intrinsic dimension of the \( k \)-th cluster from the \( z \)-th dataset.

Proof. For some measure \( \mu \) and its empirical finite-sample estimate \( \hat{\mu}_n \), proposition 2.11 may be equivalently stated with the choice of \( t = \sqrt{\log \frac{1}{\delta} / 2n} \) as:

\[
|W_p^f(\mu, \hat{\mu}_n) - E_W^f(\mu, \hat{\mu}_n)| \leq \sqrt{\log(1/\delta)/2n},
\]

holds with at least probability \( 1 - \delta \). Under the conditions stated in proposition (2.8), and combining its result with the above relation, we have

\[
W_p^f(\mu, \hat{\mu}_n) \leq cn^{-p/d} + \sqrt{\log(1/\delta)/2n},
\]

where \( c = 276^2(2 + \frac{1}{3d_{z,k}^2 - 2 - 1}) \). Combining this for all terms in the left-hand side of (13) yields the stated result. \( \square \)

2.3 Putting everything together

The final proof of Theorem 4.1 is a simplified version of Theorem 2.12’s since we assert a stronger (but cleaner) exact low-rank assumption to streamline communication. When the data is exactly supported in low-dimensions (as opposed to approximately), the \( \epsilon \)-fattening disappears (i.e., \( \epsilon \to 0 \)) thus any positive \( \sigma < \epsilon \) will send \( n \to \infty \), implying that the rapid convergence in dimensions \( d < D \) holds for \( n \to \infty \). Hence an identical result holds, with the sole condition that \( d > 4 \). \( \square \)

3 Proof of Theorem 4.2

We apply a very recent perturbation bound for the Procrustes problem developed by Arias-Castro et al. [4] to subsequently state a cluster-based alignment bound. First, we outline the perturbation bound for the classical Procrustes problem below.

Theorem 3.1 (Procrustes perturbation bounds. Theorem 1 [4]). Consider short matrices \( X, Y \in \mathbb{R}^{d \times n} \) with \( d < n \) and \( X \) having full rank. Set \( \epsilon^2 = \|Y^\top Y - X^\top X\|_p^p \), where \( \|\cdot\|_p \) denotes the Schatten \( p \)-norm. Denote the singular value decomposition of \( X = U \Sigma V^\top \), where \( \Sigma \) contains diagonal elements \( \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d > 0 = \ldots = 0 \), and let \( X^\dagger \) be the pseudo-inverse of \( X \), i.e., \( X^\dagger = U \Sigma^\dagger V^\top \), where \( \Sigma^\dagger = \text{diag}(\sigma_1^{-1}, \sigma_2^{-1}, \ldots, \sigma_d^{-1}, 0, \ldots, 0) \). If

\[
\|X^\dagger\| \leq \frac{\epsilon}{\sqrt{2}(\|X\| \|X^\dagger\|)^{1/2}}
\]

Then

\[
\min_{R \in S(D, d)} \|RX - Y\|_p \leq (\|X\| \|X^\dagger\| + 2) \|X^\dagger\| \epsilon^2.
\]

Directly applying Theorem 3.1 using the Schatten 2-norm (i.e., the Frobenius norm) yields

\[
\min_{R \in S(D, d)} \|RX - Y\|_F^2 = \min_{R \in S(D, d)} \sum_{i=1}^c \text{tr}(Q_i^\top X_i^\top X_i Q_i + Y_i^\top Y_i - 2Y_i Q_i^\top X_i^\top R) 
\]

\[
\leq ((\|X\| \|X^\dagger\| + 2) \|X^\dagger\| \epsilon^2)^2 := B^2;
\]

\[
\Rightarrow \max_{R \in S(D, d)} \sum_{i=1}^c \text{tr}(2Y_i Q_i^\top X_i^\top R) \geq \text{tr}(XX^\top + YY^\top) - B^2,
\]

where \( \epsilon^2 := \|Y^\top Y - X^\top X\|_F^2 \) and under the conditions that \( X \) is full rank and \( \|X^\dagger\| \epsilon \leq \frac{1}{\sqrt{2}}(\|X\| \|X^\dagger\|)^{-1/2} \). When the criterion given by corollary 2.3 is satisfied, we are guaranteed
What remains is for us to show the condition for equality. From (14), we have that
\[
\sum_{i=1}^c \frac{\text{tr}(X_i \text{diag}(Q_{ii} 1)X_i^T + Y_i \text{diag}(Q_{ii} 1)Y_j^T - 2Y_iQ_{ii}^T X_i^T R^T)}{n}
\]
\[
= \sum_{i=1}^c \text{tr}(X_i \text{diag}(Q_{ii} 1)X_i^T + Y_i \text{diag}(Q_{ii} 1)Y_j^T) - \max_{R \in S(d,d)} \sum_{i=1}^c \text{tr}(2Y_iQ_{ii}^T X_i^T R^T)
\]
\[
\leq B^2 + \sum_{i=1}^c \text{tr}(X_i \text{diag}(Q_{ii} 1) - Q_{ii} Q_{ii}^T)X_i^T + Y_j \text{diag}(Q_{ii} 1) - I)Y_j^T
\]
\[
= B^2 + \sum_{i=1}^c \text{tr}(X_i(1/n - Q_{ii} Q_{ii}^T)X_i^T + Y_j(1/n - I)Y_j^T).
\]

\[
\square
\]

4 Proof of Lemma 4.3

To ease notation, let \( A = Y_j Q_{ij}^T X_i^T \), and \( C = \frac{1}{n} X_i X_i^T + \frac{1}{n} Y_j Y_j^T \). Let \( A \) be decomposed by the singular-value decomposition as \( A = \tilde{U} \Sigma \tilde{V}^T \). Lastly, let \( U = [U', U''] \) and \( V = [V', V''] \), where \( \langle U', U' \rangle = \langle V', V' \rangle = 0 \). Then it follows that

\[
\text{tr}(U^T \tilde{U} \Sigma \tilde{V}^T V) \leq \text{tr}(\Sigma) = \text{tr}(\tilde{U}^T \tilde{U} \Sigma \tilde{V}^T \tilde{V}) \quad (14)
\]

\[
\text{tr}(U^T AV) \leq \text{tr}(\tilde{U}^T A \tilde{V})
\]

\[
\text{tr} \left( C - 2U'^T AV' - 2U''^T AV'' \right) \geq \min_{U''', V''' \in S(d,d-r); \langle U', U'' \rangle = \langle V', V'' \rangle = 0} \text{tr} \left( C - 2\tilde{U}^T A \tilde{V} \right)
\]

\[
\min_{U''', V''' \in S(d,d-r); \langle U', U'' \rangle = \langle V', V'' \rangle = 0} \text{tr} \left( C - 2AV'U'^T - 2AV''U''^T \right) \geq \min_{R \in S(d,d)} \text{tr} \left( C - 2AV\tilde{U}^T \right)
\]

What remains is for us to show the condition for equality. From (14), we have that

\[
\text{tr}(U^T \tilde{U} \Sigma \tilde{V}^T V) = \text{tr}(\tilde{V}^T \tilde{V} U^T \tilde{U} \Sigma) \leq \text{tr}(\Sigma),
\]

with equality holding if \( \tilde{V}^T \tilde{V} U^T \tilde{U} = I \), implying that \( \tilde{U}^T U = \tilde{V}^T V \), which imply that

\[
\langle \tilde{U}, U' \rangle = \langle \tilde{V}, V' \rangle, \quad \langle \tilde{U}, U'' \rangle = \langle \tilde{V}, V'' \rangle,
\]

which are obtained via the substitutions \( U = [U', U''] \) and \( V = [V', V''] \). \( \square \)

References


