JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2025, VOL. 00, NO. 0, 1–14: Special Issue on Statistical Science in Artificial Intelligence
https://doi.org/10.1080/01621459.2025.2540653

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Optimal Transport based Cross-Domain Integration for Heterogeneous Data

Yubai Yuan[a], Yijiao Zhang[b], Babak Shahbaba[c], Norbert Fortin[d], Keiland Cooper[d], Qing Nie[e], and Annie Qu[f]

[a]Department of Statistics, The Pennsylvania State University, University Park, PA; [b]Department of Statistics and Data Science, Fudan University, Shanghai, China; [c]Department of Statistics, University of California Irvine, Irvine, CA; [d]Department of Neurobiology and Behavior, University of California Irvine, Irvine, CA; [e]Department of Mathematics, Department of Developmental and Cell Biology, Department of Biomedical Engineering, University of California Irvine, Irvine, CA; [f]Department of Statistics and Applied Probability, University of California Santa Barbara, Santa Barbara, CA

ABSTRACT

Detecting dynamic patterns shared across heterogeneous datasets is a critical yet challenging task in many scientific domains, particularly within the biomedical sciences. Systematic heterogeneity inherent in diverse data sources can significantly hinder the effectiveness of existing machine learning methods in uncovering shared underlying dynamics. Additionally, practical and technical constraints in real-world experimental designs often limit data collection to only a small number of subjects, even when rich, time-dependent measurements are available for each individual. These limited sample sizes further diminish the power to detect common dynamic patterns across subjects. In this article, we propose a novel heterogeneous data integration framework based on optimal transport to extract shared patterns in the conditional mean dynamics of target responses. The key advantage of the proposed method is its ability to enhance discriminative power by reducing heterogeneity unrelated to the signal. This is achieved through the alignment of extracted domain-shared temporal information across multiple datasets from different domains. Our approach is effective regardless of the number of datasets and does not require auxiliary matching information for alignment. Specifically, the method aligns longitudinal data from heterogeneous datasets within a common latent space, capturing shared dynamic patterns while leveraging temporal dependencies within subjects. Theoretically, we establish generalization error bounds for the proposed data integration approach in supervised learning tasks, highlighting a novel tradeoff between data alignment and pattern learning. Additionally, we derive convergence rates for the barycentric projection under Gromov-Wasserstein and fused Gromov-Wasserstein distances. Numerical studies on both simulated data and neuroscience applications demonstrate that the proposed data integration framework substantially improves prediction accuracy by effectively aggregating information across diverse data sources and subjects. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received December 2024
Accepted July 2025

KEYWORDS
Distribution alignment; Gromov-Wasserstein Barycenter; Latent embedding; Optimal transport; Semi-supervised learning

1. Introduction

In large-scale scientific studies, extracting generalizable knowledge from multimodal and heterogeneous data is crucial for achieving more accurate predictions, decisions, and interpretations (Halevy, Rajaraman, and Ordille 2006). These studies often involve collecting vast amounts of data from diverse modalities and subjects, making efficient information aggregation particularly challenging (Boufares and Salem 2012). Overcoming these challenges necessitates the development of advanced data integration strategies.

A significant and often intractable source of data heterogeneity arises from systematic distinctions between data sources in terms of representation and underlying distribution. This leads to misaligned patterns, as the same information may be inconsistently measured and collected across sources. For instance, medical institutions in different regions often use distinct diagnostic processes, complicating efforts to aggregate data from electronic health records or medical tests (Mikkelsen and Aasly 2001). Furthermore, intrinsic heterogeneity in data distributions across sources can obscure common patterns. In neuroscience applications, which are the focus of detailed discussion in this article, the distributions of neuronal activation frequency and intensity can vary significantly across subjects or species, even when responding to the same external stimuli or cognitive tasks. These variations reflect the complex individual differences inherent in neural systems (Akil et al. 2011; Sejnowski, Churchland, and Movshon 2014).

Many domain adaptation methods have been proposed in recent years for integrating complex multi-source datasets (Kouw and Loog 2019; Farahani et al. 2021). Existing techniques can be broadly classified into instance-based and feature-based approaches. Instance-based methods (Huang et al. 2006; Gretton et al. 2008) address data distribution shifts among sources by employing various sample re-weighting strategies. However, these methods are generally not feasible when the covariates collected from each source differ, or the measurement formats are not uniform, as the density ratios of data distributions

CONTACT Annie Qu aqu2@ucsb.edu Department of Statistics and Applied Probability, University of California Santa Barbara, Santa Barbara, CA 93106-3110.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.
© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

across datasets are not well-defined. Feature-based adaptation approaches (Gopalan, Li, and Chellappa 2011; Gong et al. 2012) aim to transform the original features into new feature representations that are shared across datasets from different domains, enabling the learning of cross-domain patterns based on the transformed features. Existing methods use a variety of parametric and nonparametric models to learn feature transformation mappings, including factor models (Lindén et al. 2022), Gaussian processes (Ebrahimi et al. 2020), kernel methods (Long et al. 2013), and neural networks (Long et al. 2015). The performance of domain adaptation heavily depends on whether the feature transformation model can accurately capture the mechanisms underlying data heterogeneity in each source. In real-world applications, this assumption is often violated due to the complex and significant within-source heterogeneity. Furthermore, in many fields, such as neuroscience, the number of data sources is often small, and samples collected from the same source are strongly dependent. These issues reduce the effective information extraction for modeling feature transformations and limit the power of over-parameterized models, such as deep learning. Another important line of work utilizes the idea of multi-task representation learning to extract transferable features from heterogeneous datasets (Gebru, Hoffman, and Fei-Fei 2017; Zhang and Chen 2022; Tian, Gu, and Feng 2023).

Optimal transport, as a powerful tool for comparing and learning from data distributions, has been widely used in various machine learning tasks, including supervised learning (Frogner et al. 2015), unsupervised learning (Yan, Wang, and Rigollet 2024), and representation learning (Cherian and Aeron 2020). In this article, we aim to develop a novel and general cross-domain heterogeneous data integration framework, called integrated latent alignment (ILA), building on recent advances in optimal transport (Agueh and Carlier 2011; Courty, Flamary, and Tuia 2014; Titouan et al. 2019) and domain adaptation (Courty et al. 2016, 2017). Specifically, we first align both the geometric structure of domain-specific covariates and the covariate-response patterns from different datasets. Cross-domain alignment enables filtering out domain-specific heterogeneity and aggregating data from each domain into a shared latent space. Then we can adopt supervised learning methods in the aligned latent space to capture the common conditional mean patterns of response. Unlike existing methods, the proposed method bypasses the data heterogeneity modeling. The key novelty of the proposed method lies in connecting the common patterns in the conditional distribution of responses with the geometric structure of the data distribution and integrating datasets from different sources through geometric alignment.

The new integration design introduces several significant advantages. It offers a principled strategy for integrating data from heterogeneous subjects and increasing statistical power in identifying common patterns. Additionally, the geometric alignment avoids requiring restrictive assumptions about the transformation mappings from heterogeneous domain-specific data to cross-domain features. As a result, the integration is more flexible and robust to intractable data heterogeneity and perturbations between responses and covariates across subjects. Furthermore, the integration remains effective even when the number of data sources is small. Notably, our alignment procedure does not require cross-source data to have uniform variable collections. Thus, the covariates across datasets can differ in terms of measurement, dimension, and domain. Finally, the proposed method can incorporate heterogeneous longitudinal datasets from different domains and extract the common temporal patterns of the target responses. Both the numerical studies on simulations and the neuroscience applications of the proposed data integration method demonstrate substantial improvements in supervised learning task performance compared to methods that use individual datasets alone.

From a theoretical perspective, we provide rigorous guarantees for the proposed data integration approach in supervised learning tasks. Specifically, we derive generalization error bounds for both regression and classification setting by integrating data from multiple sources. We establish a series of convergence rates for the discrepancy between the integrated data distribution and the true data-generating distribution. Our results show that generalization errors depend on both the bias introduced by data integration and the estimation errors of the supervised learning models. These findings apply to a class of widely used nonparametric models. Different to existing theoretical works, our analysis reveals a trade-off between data alignment and pattern learning in the context of data integration. Specifically, we demonstrate that data integration can improve generalization errors in supervised learning tasks if the complexity of regression or classification models exceeds that of integration mappings. Additionally, our theoretical results contribute to a deeper understanding of optimal transport, particularly with respect to the statistical properties of the Gromov-Wasserstein (GW) distance and the fused Gromov-Wasserstein (FGW) distance. First, we establish the existence of Monge maps for the FGW problem under squared distances. Second, we derive convergence rates for the barycentric projection under GW and FGW distances. To the best of our knowledge, these two results are the first in this area of research. Third, we provide an explicit representation of the FGW barycenter under the assumption of Gaussian covariate distributions. These results are nontrivial extensions of well-established findings within the Wasserstein distance framework and pave the way for future studies on optimal transport problems involving GW and FGW.

This article is organized as follows: Section 2 introduces the background of optimal transport and its notations. Section 3 describes the proposed data integration method for supervised learning, along with the corresponding optimization algorithms. Section 4 provides a theoretical analysis of the new integration approach. Section 5 presents simulation studies, while Section 6 illustrates the application of the proposed method to a challenging, yet common, neuroscience problem. The final section offers conclusions, a discussion, and future directions.

2. Background and Notations

Optimal transport (OT) has been widely applied in various machine learning fields (e.g., domain adaptation and transfer learning), to provide a principled way of aligning distributions while preserving the geometric structures. The classical OT based on the Wasserstein distance requires that data distribu-

tions be defined on the same space, which may not be satisfied in many real applications where data could arise from different domains of distributions. To overcome this limitation, the Gromov–Wasserstein distance (Mémoli 2014) and fused Gromov–Wasserstein distance (Vayer et al. 2020) have been introduced to extend the concept of OT to cross-domain distribution alignment. In what follows, we present the relevant concepts and notations of OT, particularly those pertinent to the proposed method in this article.

For a Polish space $\boldsymbol{X}$, let $\mathcal{P}(\boldsymbol{X})$ be a Borel probability measure on random vectors $X \in \boldsymbol{X}$. For a measurable map $T : \boldsymbol{X} \to \boldsymbol{Y}$, the *pushforward* of a measure $\mu \in \mathcal{P}(\boldsymbol{X})$, denoted by $T_{\#}\mu$, is defined as $T_{\#}\mu(A) := \mu\left(T^{-1}(A)\right)$ for any Borel set $A \in \boldsymbol{Y}$. For any joint distribution $\mu \in \mathcal{P}(\boldsymbol{X} \times \boldsymbol{Y})$, we denote μ_x as the corresponding marginal distribution on $\boldsymbol{X}$. For a map $T : \boldsymbol{X} \times \boldsymbol{Y} \to \boldsymbol{X} \times \boldsymbol{Y}$, define T_x (T_y) as the map that only transforms $X(Y)$ by marginalizing $Y(X)$ out. Consider S data sources where each source is defined on domain $\boldsymbol{V}^{(s)} := \boldsymbol{X}^{(s)} \times \boldsymbol{Y}$ with response space $\boldsymbol{Y} \in \mathbb{R}$ and covariate space $\boldsymbol{X}^{(s)} \in \mathbb{R}^{d_s}$, $s = 1, \ldots, S$. For two datasets $1 \le s \ne s' \le S$, let $\Pi(\mu^{(s)}, \mu^{(s')})$ denote the class of joint distribution $\mathcal{P}(\boldsymbol{V}^{(s)} \times \boldsymbol{V}^{(s')})$ with marginal distributions on $\boldsymbol{V}^{(s)}$ and $\boldsymbol{V}^{(s)'}$ being $\mu^{(s)}$ and $\mu^{(s')}$, respectively. An element $\pi \in \Pi(\mu^{(s)}, \mu^{(s')})$ is called a *coupling plan* between $\mu^{(s)}$ and $\mu^{(s')}$. We call $T : \boldsymbol{V}^{(s)} \to \boldsymbol{V}^{(s')}$ a measurable *map* between $\mu^{(s)}$ and $\mu^{(s')}$ and $\pi \in \Pi(\mu^{(s)}, \mu^{(s')})$ being induced by T if $T_{\#}\mu^{(s)} = \mu^{(s')}$ and $\pi = (\mathrm{id}, T)_{\#}\mu^{(s)}$.

Gromov-Wasserstein distance For $\mu_x \in \mathcal{P}(\boldsymbol{X})$, define a distance metric $D : \boldsymbol{X} \times \boldsymbol{X} \to \mathbb{R}$ as *gauge*, if D is symmetric and integrable (Beier and Beinert 2024). A *gauged measure space (gm-space)* is a triple $\mathcal{X} = (\boldsymbol{X}, D, \mu_x)$. Consider the covariate spaces from source s and s' as $\mathcal{X}_s$ and $\mathcal{X}_{s'}$. The Gromov-Wasserstein distance between $\mathcal{X}_s$ and $\mathcal{X}_{s'}$ is defined as

$$\begin{aligned}&\mathrm{GW}^2\left(\mathcal{X}_s, \mathcal{X}_{s'}\right)\\ &= \min_{\pi \in \Pi(\mu_x^{(s)}, \mu_x^{(s')})} \int_{\boldsymbol{X}^{(s)} \times \boldsymbol{X}^{(s')}} \left| D^{(s)}(x_1, x_2) - D^{(s')}(x_3, x_4) \right|^2 \\ &\quad \times \mathrm{d}\pi\left(x_1, x_3\right) \mathrm{d}\pi\left(x_2, x_4\right),\end{aligned}$$

where $x_1, x_2 \in \boldsymbol{X}^{(s)}$ and $x_3, x_4 \in \boldsymbol{X}^{(s')}$. Notice that $\boldsymbol{X}^{(s)}$ and $\boldsymbol{X}^{(s')}$ can be from different domains with different data types and dimensions of covariates. In practice, we do not have the distribution information $\mu_x^{(s)}$, $\mu_x^{(s')}$. Instead, we can collect samples in each data source as $X^{(s)} = \{X_i^{(s)}\}_{i=1}^{N_s}$ from $\mu_x^{(s)}$, $s = 1, \ldots, S$ where n_s is the sample size of the sth dataset. Therefore, we can construct the empirical distribution $\hat{\mu}_x^{(s)} = \frac{1}{n_s} \sum_{i=1}^{n_s} \delta_{X_i^{(s)}}$, $s = 1, \ldots, S$. Then the empirical Gromov-Wasserstein distance is

$$\begin{aligned}\mathrm{GW}^2\left(\mathcal{X}_s, \mathcal{X}_{s'}\right) = \min_{\Omega \in \boldsymbol{\Omega}} \sum_{i,j,k,l} &\left| D^{(s)}\left(X_i^{(s)}, X_k^{(s)}\right) - D^{(s')}\left(X_j^{(s')}, X_l^{(s')}\right) \right|^2 \\ &\cdot \Omega_{i,j} \cdot \Omega_{k,\ell},\end{aligned}$$

where $\boldsymbol{\Omega}$ is the class of transport coupling matrix between $\hat{\mu}_x^{(s)}$ and $\hat{\mu}_x^{(s')}$, with $\mathbb{1}_N$ being N-dimensional one vector, defined as

$$\boldsymbol{\Omega} = \left\{ \Omega \in (\mathbb{R}_+)^{N_s \times N_{s'}} : \Omega \mathbb{1}_{n_{s'}}^\top = \hat{\mu}_x^{(s)}, \Omega^\top \mathbb{1}_{n_s}^\top = \hat{\mu}_x^{(s')} \right\}.$$

The ith column in the coupling matrix Ω decides the proportions of probability mass in the ith sample point from the dataset s to be transported to each sample point in the dataset s'.

Fused Gromov-Wasserstein distance To further measure the discrepancy of joint distribution between two datasets, the fused Gromov-Wasserstein distance has been introduced as a generalization of the aforementioned Gromov-Wasserstein distance, where a *labeled gm-space* is defined as a triple $\mathfrak{X} := (\boldsymbol{X} \times \boldsymbol{Y}, D, d, \mu)$ with a joint distribution $\mu \in \mathcal{P}(\boldsymbol{X} \times \boldsymbol{Y})$ on a gm-space $(\boldsymbol{X}, D, \mu_x)$ and a compact metric space $(\boldsymbol{Y}, d)$ (Beier, Beinert, and Steidl 2023). The gm-space $(\boldsymbol{X}, D, \mu_x)$ contains the covariate structure information, and $(\boldsymbol{Y}, d)$ the label information. Denote the labeled gm-spaces for sth data source as $\mathfrak{X}_s := (\boldsymbol{V}^{(s)}, D^{(s)}, d, \mu^{(s)})$. The fused Gromov-Wasserstein distance between two labeled gm-spaces $\mathfrak{X}_s$ and $\mathfrak{X}_{s'}$ is defined as

$$\begin{aligned}\mathrm{FGW}_\alpha^2\left(\mathfrak{X}_s, \mathfrak{X}_{s'}\right) = \min_{\pi \in \Pi(\mu^{(s)}, \mu^{(s')})} \int_{\boldsymbol{V}^{(s)} \times \boldsymbol{V}^{(s')}} &\Big\{ (1-\alpha) \\ &\times \left| D^{(s)}\left(x_1, x_2\right) - D^{(s')}\left(x_3, x_4\right) \right|^2 \\ &+ \alpha d^2\left(y_1, y_3\right) \Big\} \mathrm{d}\pi\left(v_1, v_3\right) \mathrm{d}\pi\left(v_2, v_4\right),\end{aligned} \tag{1}$$

where $\alpha \in (0, 1)$ and $v_i = (x_i, y_i) \in \boldsymbol{V}^{(s)}$ for $i = 1, 2$ and $v_i \in \boldsymbol{V}^{(s')}$ for $i = 3, 4$. A plan $\pi \in \Pi(\mu^{(s)}, \mu^{(s')})$ solving (1) is the *optimal plan* between $\mu^{(s)}$ and $\mu^{(s')}$, and the map T corresponding to the optimal plan π is the *FGW map* between $\mu^{(s)}$ and $\mu^{(s')}$. By Theorem 3.1 in Vayer et al. (2020), the FGW distance in (1) always can be achieved at an infimum $\pi^* \in \Pi(\mu^{(s)}, \mu^{(s')})$. We can similarly define empirical FGW distance by replacing $\mu^{(s)}$ and $\mu^{(s')}$ by the empirical distributions. Notice that when $\alpha = 0$, the FGW distance degenerates to the GW distance.

3. Methodology

In this section, we propose a new data integration framework for supervised and semi-supervised learning tasks. In Section 3.1, we introduce a target integrated learning method that integrated samples from multiple source datasets into a target dataset. In Section 3.2, we introduce a temporal alignment method to learn the cross-dataset longitudinal patterns.

3.1. Target Integrated Learning

In this section, we consider multiple datasets $\{(X^{(s)}, Y^{(s)})\}_{s=1}^S$ defined in S different domains. The proposed data integration method can be used for the semi-supervised learning scenario, where each dataset consists of a small proportion of labeled samples and a large number of unlabeled samples. The objective is to predict outcomes for the unlabeled samples. To enable the cross-domain alignment, we assume that there exists a mapping $T^{(s,s')} : \boldsymbol{V}^{(s')} \to \boldsymbol{V}^{(s)}$ satisfying

$$Y^{(s')} \mid X^{(s')} \overset{d}{\approx} Y^{(s)} \mid T_x^{(s,s')}\left(X^{(s')}\right). \tag{2}$$

Notice that the measure-preserving mapping $T^{(s,s')}$ corresponds to the optimal plan π in (1). Motivated by real applications, we consider the scenario where the covariate marginal distributions $\mu_x^{(s)}, s = 1, 2, \ldots, S$, share the same geometric structure, which

can be preserved via $T^{(s,s')}$. Specifically, the geometry information of $\mu_x^{(s)}$ can be represented as a set of pairwise distances $\{D^{(s)}(x_i, x_j)\}$ where x_i, x_j are samples from $\mu_x^{(s)}$. Then the distribution of $\{D^{(s)}(x_i, x_j)\}$ represents the concentration of samples from $\boldsymbol{X}^{(s)}$ at different regions. Therefore, a common geometric structure between $\boldsymbol{X}^{(s)}$ and $\boldsymbol{X}^{(s')}$ implies that for $x_i, x_j \in \boldsymbol{X}^{(s')}$ $D^{(s')}\big(x_i, x_j\big) \approx D^{(s)}\big(T_x^{(s,s')}(x_i), T_x^{(s,s')}(x_j)\big), 1 \le s \ne s' \le S$. Intuitively, the above setup suggests that the response variable Y is determined by the intrinsic structure of the covariates X, which remains invariant across different representations of X in various data sources. The cross-domain relation (2) enables us to increase the effective training sample size in the s dataset via integrating labeled data from other datasets s' associated with s, thereby improving the performance of the machine learning model in the sth domain.

Denote the samples from the sth dataset as $\mathcal{N}^{(s)}$ with $n_s = |\mathcal{N}^{(s)}|$ where $1 \le s \le S$. Within each dataset, let $\mathcal{N}_l^{(s)}$ and $\mathcal{N}_u^{(s)}$ be the subset of labeled and unlabeled samples such that $\mathcal{N}^{(s)} = \mathcal{N}_l^{(s)} \cup \mathcal{N}_u^{(s)}$ and $n_l^{(s)} = |\mathcal{N}_l^{(s)}|, n_u^{(s)} = |\mathcal{N}_u^{(s)}|$ hence, $n_s = n_l^{(s)} + n_u^{(s)}$. We denote the labeled samples in the sth dataset as $(X_l^{(s)}, Y_l^{(s)})$ where $X_l^{(s)} = \{X_i^{(s)}\}_{i\in\mathcal{N}_l^{(s)}}$ and $Y_l^{(s)} = \{Y_i^{(s)}\}_{i\in\mathcal{N}_l^{(s)}}$. Denote the covariates of unlabeled samples as $X_u^{(s)} = \{X_i^{(s)}\}_{i\in\mathcal{N}_u^{(s)}}$. Without loss of generality, we select the sth dataset as target dataset, and integrate labeled samples from other source datasets $1 \le s' \ne s \le S$ to improve the prediction on target dataset. The data integration for other datasets $1 \le s' \ne s \le S$ follows the same procedure. Define the distance matrices $M^{s,s'} \in \mathbb{R}_+^{n_u^{(s)}\times n_l^{(s')}} = \big(d(\hat{Y}_i^{(s)}, Y_j^{(s')})\big)_{i\in\mathcal{N}_u^{(s)}, j\in\mathcal{N}_l^{(s')}}$, $D^{(s)} \in \mathbb{R}_+^{n_u^{(s)}\times n_u^{(s)}} = \big(d^{(s)}(X_i^{(s)}, X_j^{(s)})\big)_{i\in\mathcal{N}_u^{(s)}, j\in\mathcal{N}_u^{(s)}}$, and $D^{(s')} \in \mathbb{R}_+^{n_l^{(s')}\times n_l^{(s')}} = \big(d^{(s')}(X_i^{(s')}, X_j^{(s')})\big)_{i\in\mathcal{N}_l^{(s')}, j\in\mathcal{N}_l^{(s')}}$ where $d(\cdot), d^{(s)}(\cdot), d^{(s')}(\cdot)$ are the metrics on response, covariates on target dataset, and covariates on source dataset, respectively. The proposed method is summarized as follows:

Step 1: Individual supervised learning. Given a specific classifier $f(\cdot)$, we train the model on $(X_l^{(s)}, Y_l^{(s)})$ separately as $f^{(s)}(\cdot)$, and then use it to predict the unlabeled set $X_u^{(s)}$ in order to obtain predicted labels $\hat{Y}_u^{(s)} = f^{(s)}(X_u^{(s)})$.

Step 2: One-to-many cross-domain alignment. We consecutively align labeled samples from other $S-1$ datasets $(X_l^{(s')}, Y_l^{(s')})$, $s' \ne s, 1 \le s' \le S$ to the region of unlabeled data $X_u^{(s)}$ in the sth dataset. Specifically, we adopt the fused Gromov-Wasserstein optimal transport as follows:

$$\Omega^{(s,s')} = \underset{\boldsymbol{\Omega}}{\operatorname{argmin}} \sum_{i,j,k,l} \big\{(1-\alpha)M_{ij}^{s,s'} + \alpha(D_{i,k}^{(s)} - D_{j,\ell}^{(s')})^2\big\}\Omega_{i,j}\Omega_{k,\ell},$$

$$\boldsymbol{\Omega} = \Bigg\{\Omega \in (\mathbb{R}_+)^{n_u^{(s)}\times n_l^{(s')}} : \Omega\mathbb{1}_{n_l^{(s')}} = \frac{1}{n_u^{(s)}}\mathbb{1}_{n_u^{(s)}}, \times \Omega^\top \mathbb{1}_{n_u^{(s)}} = \frac{1}{n_l^{(s')}}\mathbb{1}_{n_l^{(s')}}\Bigg\}. \tag{3}$$

Given that $X_l^{(s')}$ and $X_u^{(s)}$ are not in the same domain, $\Omega^{(s,s')}$ intends to match the internal geometric structures between $X_l^{(s)}$ and $X_u^{(s')}$, which are measured by distance matrices $D^{(s)}$ and $D^{(s')}$, respectively. On the other hand, $\Omega^{(s,s')}$ also aims to align the label distribution between datasets s and s' though the term $(1-\alpha)M_{ij}^{s,s'}$. Minimizing (3) allows us to match the joint distribution $\mathbb{P}(X_l^{(s')}, Y_l^{(s')})$ to $\mathbb{P}(X_u^{(s)}, Y_u^{(s)})$ so that (2) can be held where $\Omega^{(s,s')}$ serves as the underlying cross-domain transformation $T^{(s,s')}$, and hence we can integrate labeled data from heterogeneous domains to capture the common label-predictor pattern. Because $X_u^{(s)}$ are not labeled, that is, the corresponding $Y_u^{(s)}$ are not observed, we approximate $\mathbb{P}(X_u^{(s)}, Y_u^{(s)})$ by $\mathbb{P}(X_u^{(s)}, \hat{Y}_u^{(s)})$.

Step 3: Labeled data mapping. We map the labeled set $(X_l^{(s')}, Y_l^{(s')})$ $(s' \ne s)$ from other $S-1$ datasets to the domain of the sth dataset. Specifically, we reconstruct the jth labeled sample in the s'th dataset $(X_j^{(s')}, Y_j^{(s')})$ on the sth domain as $(\hat{X}_j^{(s')}, Y_j^{(s')})$, where

$$\hat{X}_j^{(s')} = \underset{X}{\operatorname{argmin}} \sum_{i\in\mathcal{N}_u^{(s)}} \Omega_{i,j}^{(s,s')} D^{(s)}\left(X, X_i^{(s)}\right), \ j \in \mathcal{N}_l^{(s')}, \tag{4}$$

and $\Omega^{(s,s')}$ is obtained from Step 2. The barycentric projection (4) corresponds to a weighted average of unlabeled data $X_u^{(s)}$ in the sth domain. Intuitively, the barycentric mapping reconstructs the labeled sample from the $s$the dataset in the domain of the s'th dataset. We denote the reconstructed samples as $\hat{X}_l^{(s')} = \{\hat{X}_j^{(s')}\}_{j\in\mathcal{N}_l^{(s')}}$. Based on Step 2 and assumption (2), $(\hat{X}_l^{(s')}, Y_l^{(s')})$ follows the same distribution in the s domain as $(X_u^{(s)}, Y_u^{(s)})$.

Step 4: Retrain the model with augmented labeled data. We obtain reconstructed labeled data from other $S-1$ datasets as $(\tilde{X}_l^{(s)}, \tilde{Y}_l^{(s)}) = \bigcup_{s'\ne s} (\hat{X}_l^{(s')}, Y_l^{(s')})$, and train the classifier $f(\cdot)$ on the augmented labeled dataset $(X_l^{(s)} \cup \tilde{X}_l^{(s)}, Y_l^{(s)} \cup \tilde{Y}_l^{(s)})$ as $f_s^{aug}(\cdot)$. Then we update the prediction on unlabeled data $X_u^{(s)}$ as $\hat{Y}_u^{(s)} = f_s^{aug}(X_u^{(s)})$.

3.2. Temporal Alignment Integrated Learning

In this section, we introduce data integration for supervised learning on multiple longitudinal heterogeneous datasets. Specifically, we consider S longitudinal heterogeneous datasets $\{\big(X^{(s)}(t), Y^{(s)}(t)\big)\}_{s=1}^S$ measured at time points $t = 1, \ldots, T$, where covariates $X^{(s)}(t)$, $s = 1, \ldots, S$ change over time, but are defined on different domains. We assume that there exists alignment mappings $T^{(s)}$ from domains s to a cross-source latent space, that is,

$$\mathbb{P}\big(\boldsymbol{Y}(t) \mid \boldsymbol{Z}(t)\big) \approx \mu^{(s)}\Big(\boldsymbol{Y}(t) \mid T_x^{(s)}\big(\boldsymbol{X}^{(s)}(t)\big)\Big), \quad 1 \le s \le S, \ 1 \le t \le T. \tag{5}$$

where $\{\boldsymbol{Z}(t)\}_{t=1}^T$ and $\{\boldsymbol{Y}(t)\}_{t=1}^T$ denote the covariate-associated and response-associated cross-source dynamic on an aligned domains, respectively. In other words, the common temporal pattern on the conditional distribution of response given that covariates is captured by $\{(\boldsymbol{Z}(t), \boldsymbol{Y}(t)\}_{t=1}^T$. Then the sth mapping $\{T^{(s)}\}_{s=1}^S$ projects the common temporal patterns to domains of

the sth dataset. By inferring $\{(\boldsymbol{Z}(t), \boldsymbol{Y}(t)\}_{t=1}^{T}$ and $\{T^{(s)}\}_{s=1}^{S}$, we can integrate cross-domain temporal data into a common latent space to improve the performance of the supervised learning task as the effective training sample size increases.

We follow the notation system in Section 3.1 for longitudinal labeled and unlabeled samples from datasets $1 \le s \le S$ with additional time variable $1 \le t \le T$. The proposed method is summarized as follows:

Step 1: Individual supervised learning. Train a specific classifier $f(\cdot)$ on $(X_l^{(s)}(t), Y_l^{(s)}(t))$ separately as $f_t^{(s)}(\cdot)$ for each dataset s and time point t, $1 \le s \le S, 1 \le t \le T$. Then use the model to obtain predicted labels for unlabeled data $X_u^{(s)}(t)$ via $\hat{Y}_u^{(s)}(t) = f_t^{(s)}\big(X_u^{(s)}(t)\big)$ for each dataset and time point.

Step 2: Cross-domain temporal alignment. We align the time-evolving datasets from S different domains via the proposed temporal Fused Gromov-Wasserstains (FGW) barycenter. Denoting the augmented labels for the sth dataset as $Y^{(s)}(t) = \big(Y_l^{(s)}(t), \hat{Y}_u^{(s)}(t)\big), 1 \le t \le T$, we estimate the alignment couplings $\{\Omega^{(s)}\}_{s=1}^{S}$ and the common dynamic process $\{C(t)\}_{t=1}^{T} = \{D(t), Y(t)\}_{t=1}^{T}$ in the following aligned space:

$$\begin{aligned}\{C(t)\}_{t=1}^{T}, \{\Omega^{(s)}\}_{s=1}^{S} = &\underset{\substack{C(t),t=1,\ldots,T\\ \Omega^{(s)}\in\boldsymbol{\Omega}^{(s)},s=1,\ldots,S}}{\operatorname{argmin}} \sum_{t=1}^{T}\sum_{s=1}^{S} w_s \\ &\times \mathrm{FGW}_{\alpha}^{2}\Big(C(t), \Omega^{(s)}\Big),\\ \text{where } \mathrm{FGW}_{\alpha}^{2}\Big(C(t), \Omega^{(s)}\Big) = &\alpha\sum_{ij}\big(M^{(s)}(t)\big)_{ij}\Omega_{i,j}^{(s)} + (1-\alpha)\\ &\times \sum_{i,j,u,v}\|\big(D^{(s)}(t)\big)_{ij} - \big(D(t)\big)_{uv}\|^2\\ &\times \Omega_{i,u}^{(s)}\Omega_{j,v}^{(s)},\\ \boldsymbol{\Omega}^{(s)} = &\Big\{\Omega \in (\mathbb{R}_+)^{n_s\times n} : \Omega\mathbb{1}_n = \frac{1}{n_s}\mathbb{1}_{n_s},\\ &\times \Omega^{\top}\mathbb{1}_{n_s} = \frac{1}{n}\mathbb{1}_n\Big\}, \end{aligned} \tag{6}$$

where $D^{(s)}(t) \in \mathbb{R}_+^{n_s\times n_s} = \big(d^{(s)}\big(X^{(s)}(t)_i,\ X^{(s)}(t)_j\big)\big)_{i\in\mathcal{N}^{(s)}, j\in\mathcal{N}^{(s')}}$, $1\le t\le T$, $M^{(s)}(t)\in\mathbb{R}_+^{n_s\times n_s} = \big(d(Y^{(s)}(t)_i, Y^{(s)}(t)_j)\big)_{i\in\mathcal{N}^{(s)}, j\in\mathcal{N}^{(s')}}$, and n is the number of template samples in the common latent space. The similarity matrix $D(t) \in \mathbb{R}_+^{n\times n}$ represents the pairwise distances aligned with $D^{(s)}(t)$ from different datasets at time t, and consists of the common covariates dynamics $Z(t)$ in (5). Specifically, given the positions of n template samples $\{Z(t)_i\}_{i=1}^{n}$ in the aligned latent space at time t, we have $\big(D(t)\big)_{ij} = \tilde{d}\big(Z(t)_i, Z(t)_j\big)$, where $\tilde{d}(\cdot)$ is the distance between $Z(t)_i$ and $Z(t)$ in the latent space. In addition, the transport coupling $\{\Omega^{(s)} \in [0,1)^{n_s\times n}\}$ serves as the domain transformation $\{T^{(s)}\}$ in (5) which assigns the probability mass on $\{X^{(s)}(t)_i, Y^{(s)}(t)_i\}_{i=1}^{n_s}$ for template samples $\{Z(t)_i, Y(t)_i\}_{i=1}^{n}$ in the latent space with $1 \le t \le T$. Finally, $w_s \ge 0$ in (6) denotes the weights of the sth dataset in the alignment as $\sum_{s=1}^{S} w_s = 1$, and we set $w_s = \frac{1}{S}$.

Step 3: Data integration on aligned space. We embed the aligned similarity matrix $D(t)$ to a k-dimensional latent space at each time point t. Specifically, we infer $Z(t) = \{Z(t)_i \in \mathbb{R}^k\}_{i=1}^{n}$ via

$$Z(t) = \underset{\boldsymbol{Z}}{\operatorname{argmin}} \sum_{i,j}^{n}(D(t)_{ij} - \|\boldsymbol{Z}_i - \boldsymbol{Z}_j\|_2)^2.$$

Then we map all s datasets to latent space $\{Z(t)\}_{t=1}^{T}$, that is, the labeled sample $\big(X_l^{(s)}(t)_i, Y_l^{(s)}(t)_i\big)$ and the jth unlabeled sample $X_u^{(s)}(t)_j$ in the sth dataset are reconstructed in the latent space as $\big(\hat{X}_l^{(s)}(t)_i, Y_l^{(s)}(t)_i\big)$ and $\hat{X}_u^{(s)}(t)_j$ respectively, where

$$\hat{X}_l^{(s)}(t)_i = \underset{x}{\operatorname{argmin}} \sum_{k=1}^{n}\Omega_{i,k}^{(s)}\tilde{d}\big(x, Z(t)_k\big),$$

$$\hat{X}_u^{(s)}(t)_j = \underset{x}{\operatorname{argmin}} \sum_{k=1}^{n}\Omega_{j,k}^{(s)}\tilde{d}\big(x, Z(t)_k\big),$$

with $1 \le s \le S, 1 \le t \le T$, and $\Omega^{(s)}$ is obtained from Step 2. In conjunction with data integration, inferring $\{Z(t), Y(t)\}$ in a low-dimensional latent space also allows us to identify and model the governing temporal pattern on conditional distribution of the response after integrating heterogeneous data. This strategy can be useful for other downstream tasks such as clustering.

Step 4: Retrain the model with augmented labeled data. For the sth dataset at time t, we aggregate labeled data from all S datasets reconstructed in the common latent space as

$$\big(\tilde{X}_l^{(s)}(t), \tilde{Y}_l^{(s)}(t)\big) = \underset{1\le s\le S}{\cup}\big(\hat{X}_l^{(s)}(t), Y_l^{(s)}(t)\big),$$

and train the classifier $f(\cdot)$ on the augmented labeled dataset $\big(\tilde{X}_l^{(s)}(t), \tilde{Y}_l^{(s)}(t)\big)$ as $\tilde{f}_t^{(s)}(\cdot)$. Consequently, we update the prediction on unlabeled data $X_u^{(s)}(t)$ as $\hat{Y}_u^{(s)}(t) = \tilde{f}_t^{(s)}\big(\hat{X}_u^{(s)}(t)\big)$. The optimization for the proposed method and more discussion on the assumptions are provided in Sections 8 and 14 of the supplementary materials. **Remark**: approaches in Sections 3.1 and 3.2 are designed for different tasks and rely on different assumptions. Specifically, Section 3.1 aims to improve prediction performance on each individual dataset through data integration. This method assumes that the target dataset can be aligned with each source dataset, which, however, does not necessarily imply the existence of a shared data structure across all datasets. In contrast, Section 3.2 introduces integration method for longitudinal data to capture common latent dynamic patterns across multiple datasets. Identifying these common patterns does not necessarily enhance prediction performance on individual datasets due to cross-dataset heterogeneity. The method in Section 3.2 relies on the existence of an invariant data structure across different datasets.

4. Theoretical Analysis

In this section, we provide theoretical guarantees for supervise learning tasks using the proposed ILA framework. Specifically, we derive the estimation error bounds for regression and the excess risks for classification. The results for target integrated learning and alignment integrated learning are presented in Sections 4.1 and 4.2, respectively. Throughout the theoretical

analysis, we let $D^{(s)}(x_1^{(s)}, x_2^{(s')}) = \|x_1^{(s)} - x_2^{(s')}\|_2^2$, $\tilde{d}(Z(t)_i, Z(t)_j) = \|Z(t)_i - Z(t)_j\|_2^2$, and $d(y^{(s)}, y^{(s')}) = |y^{(s)} - y^{(s')}|^2$. Without loss of generality, we consider the binary classification problem. For a probability measure $\mu \in \mathcal{P}(\boldsymbol{X} \times \boldsymbol{Y})$ and two classifiers $h, h^* : X \mapsto \{0, 1\}$, the risk function and excess risk are $\mathcal{R}_\mu(h) = \mathbb{E}_\mu\{I\{Y \neq h(X)\}\}$ and $\mathcal{E}_\mu(h, h^*) = \mathcal{R}_\mu(h) - \mathcal{R}_\mu(h^*)$, respectively. In the following, we assume that the size of labeled samples is at the same order of size of unlabeled samples in each dataset, that is, $n_j^{(l)} = cn_j^{(u)}$, $1 \leq j \leq S$ for some constant $c \in (0, 1)$.

4.1. Theory for Target Integrated Learning

We first investigate the property of $\Omega^{(s,s')}$ obtained from (4) in Section 3.1. Let $\hat{\mu}^{(s)} = \frac{1}{n_s}\sum_{i=1}^{n_s} \delta_{v_i^{(s)}}$ denote the empirical version of $\mu^{(s)}$ and $\hat{\mathfrak{X}}_s$ as the empirical counterpart of $\mathfrak{X}_s$. Then $\Omega^{(s,s')}$ is the empirical optimal plans solving $\mathrm{FGW}_\alpha(\hat{\mathfrak{X}}^{(s')}, \hat{\mathfrak{X}}^{(s)})$. Notice that $\Omega^{(s,s')}$ corresponding to the FGW distance may not be unique (Deb, Ghosal, and Sen 2021). However, all results in this section hold for any $\Omega^{(s,s')}$ which minimizes $\mathrm{FGW}_\alpha(\hat{\mathfrak{X}}^{(s')}, \hat{\mathfrak{X}}^{(s)})$.

The barycentric projection of $\Omega^{(s,s')}$ (Ambrosio, Gigli, and Savaré 2008) is the conditional mean of $v^{(s)}$ given $v^{(s')}$ under $\Omega^{(s,s')}$, that is,

$$\hat{T}^{(s,s')}(v^{(s')}) := \frac{\int_{v^{(s)}} v^{(s)} d\Omega^{(s,s')}(v^{(s')}, v^{(s)})}{\int_{v^{(s)}} d\Omega^{(s,s')}(v^{(s')}, v^{(s)})}, \tag{7}$$

for $v^{(s')} \in \mathrm{supp}\left(\hat{\mu}^{(s')}\right)$. Following (4) in Section 3.1, the transformed data from source s' can be presented via the barycentric projection as $\hat{X}^{(s')} = \bigcup_{i=1}^{n_{s'}} \hat{X}_i^{(s')}$, where $\hat{X}_i^{(s')} = \hat{T}_x^{(s,s')}(X_i^{(s')})$.

Assumption 4.1. Given the labeled gm-spaces $\{\mathfrak{X}_s\}_{s=1}^S$, there exists a FGW map $T^{(s,s')}$ from $\mu^{(s')}$ to $\mu^{(s)}$ for all $s' \neq s$.

In Theorem 15.5 in Section 15.3 of the supplementary materials, we establish the existence of the FGW Monge map given several regularity conditions, which is of independent interest. The following theorem is essential for analyzing both target and alignment integrated learning. Due to space limitations, we defer the regularity conditions to Section 15 of the supplementary materials.

Theorem 4.1 (Rate of FGW Barycentric Projection). Under Assumptions 4.1 and regularity conditions, the empirical FGW Barycentric Projection $\hat{T}^{(s,s')}(v^{(s')})$ satisfies

$$\int \left\| \hat{T}^{(s,s')}(v^{(s')}) - T^{(s,s')}(v^{(s')}) \right\|^2 d\hat{\mu}^{(s')}(v^{(s')}) \lesssim \sum_{j=s,s'} n_j^{-\frac{2}{\max\{d_s,3\}+1}} (\log n_j)^{\mathbf{1}\{d_s==3\}}. \tag{8}$$

Theorem 4.1 also applies to the barycentric projection under GW distance. To the best of our knowledge, Theorem 4.1 is the first result to establish the convergence rate for barycentric projection under GW and FGW distance. Notably, the above rate of convergence aligns with the convergence rate of GW distance (Zhang et al. 2024) and also matches that of barycentric projection under Wasserstein distance Deb, Ghosal, and Sen (2021). The extension from Wasserstein distances to GW/FGW distance is non-trivial and requires to use new technical tools to analyze the quadratic form of GW distances and the pairwise similarity in the label space. We present the results by the following setup:

Assumption 4.2 (Conditional mean discrepancy). Given the target dataset s, for all $s' \neq s$, the conditional mean of label in source s' satisfies $\int_x \left| f^{(s)}(x) - \mathbb{E}[Y^{(s')}|T_x^{(s,s')}(X^{(s')}) = x] \right|^2 \mathrm{d}x \leq \delta_{s,s'}^2$, where $f^{(s)}(x) := \mathbb{E}[Y^{(s)}|X^{(s)} = x]$ denotes the conditional mean in data source s.

In Assumption 4.2, $\delta_{s,s'}$ quantifies the difference in the conditional label means between the target and transformed source data. Specifically, $\delta_{s,s'} = 0$ means $E[Y^{(s)}|X^{(s)}] = E[Y^{(s')}|T_x^{(s,s')}(X^{(s')})]$ almost surely, while a larger $\delta_{s,s'}$ reflects a greater discrepancy between two datasets.

Definition 4.1 (β-Hölder class). The β-Hölder class is defined as $\mathcal{H}_{\beta,C_\beta} = \{f : \mathbb{R}^{d_s} \mapsto \mathbb{R} \mid |f(x_1) - f(x_2)| \leq C_\beta \|x_1 - x_2\|^\beta\}$.

The β-Hölder class characterizes function smoothness. A larger β indicates that the function class is smoother and hence easier to estimate.

Definition 4.2 (η-sensitive nonparametric estimator). Let $\boldsymbol{W}_n := \{(X_i, Y_i)\}_{i=1}^n$ be an iid sample from $\mu \in \mathcal{P}(\boldsymbol{X} \times \boldsymbol{Y})$. Denote $\hat{g}(x \mid \boldsymbol{W}_n)$ as an estimator for $g = \mathbb{E}[Y|X]$. Let $\tilde{X}_i$ be a perturbed version of covariate X_i for $i = 1, \ldots, n$, and $\tilde{\boldsymbol{W}}_n := \{(\tilde{X}_i, Y_i)\}_{i=1}^n$. For some $\eta > 0$, the estimator $\hat{g}$ is an η-sensitive nonparametric estimator if

$$\sup_{x \in \boldsymbol{X}} \left| \hat{g}(x \mid \boldsymbol{W}_n) - \hat{g}(x \mid \tilde{\boldsymbol{W}}_n) \right|^2 \leq C_f \left(\sum_{i=1}^n \|X_i - \tilde{X}_i\|^2 \right)^\eta.$$

In Definition 4.2, η represents the sensitivity of estimator in terms of perturbation in training data (Koh and Liang 2017). The concept in Definition 4.2 is similar to the influence function from the literature of robust and semiparametric statistics (Cook and Weisberg 1980). In the supplementary materials, we provide examples to demonstrate that most popular nonparametric estimators satisfy $\eta = 1$.

Definition 4.3 (Classification noise). A measure $\mu \in \mathcal{P}(\boldsymbol{X} \times \boldsymbol{Y})$ has classification noise of level γ if $\mathbb{P}_{\mu^{(s)}}(|g - 1/2| < t) \leq C_\gamma t^\gamma$ for all $t > 0$, where $g = \mathbb{E}[Y|X]$.

Definition 4.3 is referred as the Tsybakov's noise condition (Audibert and Tsybakov 2007), and quantifies the noise level in labels. Similar to (4.1), γ characterizes the intrinsic difficulty of the classification. A larger γ implies an easier classification task.

Theorem 4.2 (Supervised learning via target integration). Denote the total sample size by $N = \sum_{s=1}^S n_s$ and the minimum sample size by $n_{\min} = \min_{s \in \{1,\ldots,S\}} n_s$.

Regression: assume that $f^{(s)}(\cdot)$ belongs to the β-Hölder class and $\hat{f}^{(s)}$ is an η-sensitive nonparametric estimator. Conditioning on

the integrated data $\tilde{X}^{(s)} = (\bigcup_{s'\neq s} \hat{X}^{(s')}) \bigcup X^{(s)}$, we have

$$\mathbb{E}\|\hat{f}^{(s)} - f^{(s)}\|^2 \lesssim r_{N,\beta} + \left[n_{\min}^{-\frac{2}{\max\{d_s,3\}+1}} (\log n_{\min})^{I\{d_s==3\}}\right]^{\eta} + \sum_{s'\neq s} \delta_{s,s'}^2, \tag{9}$$

where $r_{N,\beta}$ is the rate of estimation $\hat{f}^{(s)}$ via N iid samples integrated on the sth dataset.

Classification: let $h^{(s)} = I\{f^{(s)} \geq 1/2\}$ be the Bayes classifier, and $\hat{h}^{(s)} = I\{\hat{f}^{(s)} \geq 1/2\}$ be the plug-in estimator. Assume that $\mu^{(s)}$ has a classification noise of level γ. Let $k_\gamma = \frac{1+\gamma}{2+\gamma}$, we have

$$\mathbb{E}[\mathcal{E}_{\mu_x^{(s)}}(\hat{h}^{(s)}, h^{(s)})] \lesssim (r_{N,\beta})^{2k_\gamma} + \left[n_{\min}^{-\frac{2k_\gamma}{\max\{d_s,3\}+1}} (\log n_{\min})^{k_\gamma I\{d_s==3\}}\right]^{\eta} + \sum_{s'\neq s} (\delta_{s,s'})^{4k_\gamma}. \tag{10}$$

Theorem 4.2 establishes the convergence rate of the conditional mean estimator and the excess risk of the plug-in classifier for target integrated learning.

Remark 4.1. Intuitively, Theorem 4.2 indicates a trade-off between the benefits and costs of learning via data integration. On the one hand, for most popular nonparametric estimators, the convergence rate of using integrated data is $r_{N,\beta} = N^{-\frac{2\beta}{2\beta+d_s+1}}$, which is faster than the rate $r_{n_s,\beta} = n_s^{-\frac{2\beta}{2\beta+d_s+1}}$ of using target dataset only. On the other hand, the second and third terms in RHS correspond to the cost of estimating the FGW map and the data discrepancy, respectively. When $\beta < \frac{d_s+1}{\max\{d_s,3\}+1-2\eta}$ and $\sum_{s'\neq s} \delta_{s,s'}^2 \lesssim r_{n_s,\beta}$, that is, learning the conditional expectation in the target dataset is more difficult than data integration. Consequently, the proposed method in Section 3.1 improves on learning $f^{(s)}$ given $r_{N,\beta}$ dominating the total error. Conversely, when $\beta \geq \frac{d_s+1}{\max\{d_s,3\}+1-2\eta}$ or $\sum_{s'\neq s} \delta_{s,s'}^2 \lesssim r_{n,\beta}$, the integration is more difficult than individual learning $f^{(s)}$. In this case, individual learning achieves better performance.

Remark 4.2. Theorem 4.2 also captures the tradeoff in data integration for classification tasks and improves upon the existing theory (Redko, Habrard, and Sebban 2017; Courty et al. 2017), where the excess risk is bounded as

$$\mathcal{E}_{\mu_x^{(s)}}(h) \leq \mathcal{E}_{\mu_x^{(s')}}(h) + \min_{h'}\{\mathcal{E}_{\mu_x^{(s)}}(h') + \mathcal{E}_{\mu_x^{(s')}}(h')\} + \mathrm{W}_1(\hat{\mu}^{(s)}, \hat{\mu}^{(s')}) + o_p(1), \tag{11}$$

and W_1 denotes the L_1-Wasserstein distance. According to (11), data integration offers no advantage when the W_1 distance between the source and target is large. In contrast, our result (10) refines the theoretical result by showing that data integration can still improve classification accuracy with non-negligible discrepancies between source and target distributions, given data integration is easier than the individual classification task.

4.2. Theory for Alignment Integrated Learning

In this section, we discuss the property of the FGW barycenter. Given a set of weights $\boldsymbol{w} = (w_1, \ldots, w_S)$ with $\sum_{s=1}^S w_s = 1$, the FGW barycenter of labeled gm-space $\mathfrak{X}_1, \ldots, \mathfrak{X}_S$ is the labeled gm-space $\mathfrak{C} = (\boldsymbol{Z} \times \boldsymbol{Y}, D, d, \nu)$ solving $\mathfrak{C} = \mathrm{argmin}_{\mathfrak{C}} \sum_{s=1}^S w_s \mathrm{FGW}_\alpha^2(\mathfrak{X}_s, \mathfrak{C})$, where ν is the measure on $\boldsymbol{Z} \times \boldsymbol{Y}$. Without loss of generality, we let the first dataset $\boldsymbol{X}^{(1)} \times \boldsymbol{Y}$ be a reference space. Let $T^{(s,1)} : \boldsymbol{X}^{(1)} \times \boldsymbol{Y} \mapsto \boldsymbol{X}^{(s)} \times \boldsymbol{Y}$ be the FGW map from the distribution $\mu^{(1)}$ on reference space to distributions on datasets $\mu^{(s)}, s = 1, \ldots, S$. The following theorem is essential since it provides an explicit connection between the underlying FGW barycenter $\mathfrak{C}$ and pairwise FGW maps $\{T^{(s,1)}\}_{s=1}^S$.

Theorem 4.3 (FGW Barycenter). Let $\mathfrak{Q}^* = \left(\boldsymbol{X}^{(1)} \times \boldsymbol{Y}, D^{(1)}, d, T_{\#}^{(o)}\mu^{(1)}\right)$ be a labeled gm-space, where $T^o = \left((\sum_{s=1}^S w_s T_x^{(s,1)})^{1/2}, \sum_{s=1}^S w_s T_y^{(s,1)}\right)$ is a map from the reference measure $\mu^{(1)}$ to the measure ν on the barycenter space of $\mathfrak{C}$. Under Assumption 4.1 and regularity conditions, the FGW barycenter $\mathfrak{C}$ and $\mathfrak{Q}^*$ are homomorphic such that $\mathrm{FGW}(\mathfrak{C}, \mathfrak{Q}^*) = 0$.

Theorem 4.3 enables us to analyze the properties of the FGW barycenter on the reference space. Further details are provided in Section 15 of the supplementary materials. The counterparts of Theorem 4.3 for Wasserstein barycenter are studied in Agueh and Carlier (2011) and Bigot and Klein (2018), and the result for GW barycenters is provided in Beier and Beinert (2024) under Gaussian assumption. Theorem 4.3 is an non-trivial extension of the above existing results to the FGW barycenter. In the following, we use $\mathfrak{X}_s(t) = (\boldsymbol{X}^{(s)}(t) \times \boldsymbol{Y}(t), D^{(s)}, d, \mu^{(s)}(t))$ to denote the dynamic labeled gm-spaces, for $1 \leq t \leq L, 1 \leq s \leq S$. Let $\mathfrak{C}(t) = (\boldsymbol{Z}(t) \times \boldsymbol{Y}(t), D(t), d, \nu(t))$ be the barycenter space of $\{\mathfrak{X}_s(t)\}_{s=1}^S$ obtained according to Section 3.2.

Assumption 4.3 (Temporal Shared Coupling). The FGW map from $\mu^{(1)}(t)$ to $\mu^{(s)}(t)$ for all $1 \leq t \leq L$ is time-invariant, denoted by $T^{(s,1)}$. Similarly, the FGW map from $\mu^{(s)}(t)$ to $\nu(t)$ is time-invariant, denoted by $T^{(s)}$.

Intuitively, Assumption 4.3 require that the FGW maps from each dataset to the barycenter space are time-invariant, which is crucial for achieving temporal alignment in Section 3.2 across multiple time points. The time-invariant FGW maps reflects the intrinsic dependency within temporal data from same source, which is common in many longitudinal studies.

Assumption 4.4 (Conditional mean discrepancy). Let $f_t := \mathbb{E}[Y(t)|Z(t)]$ denote the conditional mean in the shared latent space at time t. For datasets $s = 1, \ldots, S$ and time points $t = 1, \ldots, T$, the conditional means satisfy $\int_x \| f_t(x) - \mathbb{E}[Y^{(s)}(t)|T_x^{(s)}(X^{(s)}(t)) = x]\|^2 \, \mathrm{d}x \leq (\delta_s^{(t)})^2$.

The quantity $\delta_s^{(t)}$ in Assumption 4.4 characterizes the discrepancy between the conditional means of response in the shared latent space and the transformed sth source space at time t. In particular, $\delta_s^{(t)} = 0$ when the FGW distance between the two spaces is zero. A larger $\delta_s^{(t)}$ implies a greater intrinsic difference at

time t. Let $\hat{T}^{(s)}$ be the barycenteric projection on the barycenter space obtained from $\Omega^{(s)}$. The reconstructed data in Step 3 can be denoted as $\hat{Z}^{(s)}(t) = \bigcup_{i=1}^{n_s} \hat{Z}_i^{(s)}(t)$, where $\hat{Z}_i^{(s)}(t) = \hat{T}_x^{(s)}(X_i^{(s)}(t))$.

Theorem 4.4 (Supervised learning via alignment integration). Regression: assume that $f_t(\cdot)$ belongs to the β-Hölder class and $\hat{f}_t$ is an η-sensitive nonparametric estimator.Conditioning on the reconstructed data $\tilde{Z}(t) = \bigcup_{s=1}^{S} \hat{Z}^{(s)}(t)$, we have

$$\mathbb{E}\|\hat{f}_t - f_t\|^2 \lesssim r_{N,\beta} + \left[(n_{\min}L)^{-\frac{2}{\max\{d_1+1,4\}}} (\log(n_{\min}L))^{I\{d_1==3\}}\right]^{\eta} + \sum_{s=1}^{S} (\delta_s^{(t)})^2,$$

where $r_{N,\beta}$ is the rate of $\hat{f}$ using N iid samples at time t from the barycenter space.

Classification: let $h_t = I\{f_t \geq 1/2\}$ be the Bayes classifier with $\hat{h}_t = I\{\hat{f}_t \geq 1/2\}$ being the plug-in estimator. Further assuming that $v(t)$ on the barycenter space is a measure with classification noise of level γ, we have

$$\mathbb{E}[\mathcal{E}_{v_x(t)}(\hat{h}_t, h_t)] \lesssim (r_{N,\beta})^{\frac{2(1+\gamma)}{(2+\gamma)}} + \left[(n_{\min}L)^{-\frac{2(1+\gamma)}{(\max\{d_1,3\}+1)(1+\gamma)}} \times (\log(n_{\min}L))^{\frac{(1+\gamma)}{(2+\gamma)} I\{d_1==3\}}\right]^{\eta} + \sum_{s=1}^{S} (\delta_s^{(t)})^{\frac{4(1+\gamma)}{(2+\gamma)}}.$$

Theorem 4.4 extends the results on target integrated learning from Section 4.1 to the alignment integrated setting with longitudinal data, focusing on estimands in the aligned space. Specifically, Theorem 4.4 shows that the proposed aligned integrated learning approach provides benefits when $\beta < \frac{d_s+1}{\max\{d_1,3\}+1-2\eta\log L}$ and $\sum_{s=1}^{S}(\delta_s^{(t)})^2 \lesssim r_{n_s,\beta}$. A notable distinction here is the presence of the $\log L$ term in $\frac{d_s+1}{\max\{d_1,3\}+1-2\eta\log L}$, which enhances a gain from data integration as discussed in Remark 4.2. This improvement arises from leveraging shared coupling to effectively incorporate temporal information.

5. Simulation Study

In this section, we investigate the performance of the proposed data integration approach and temporal integration method for several classification problems, and compare the performance to a baseline model that uses the data from individuals. Here, we select two classifiers for the numerical comparisons: quadratic discriminant analysis (QDA) and label propagation (LP); the former is a supervised learning classifier, whereas the latter is a semi-supervised learning method using the structure information on covariates.

5.1. Simulations of Data Integration for Classification

In this simulation setting, we consider six different datasets where the label of data points Y in each dataset can belong to one of three classes. The data generation process is as follows. For each class $Y = k, (k = 1, 2, 3)$, the two-dimensional covariates X are sampled from the mixture of Gaussian distributions as $X \mid (Y = k) \sim 0.5 \times N(\mu_k^{(1)}, I) + 0.5 \times N(\mu_k^{(2)}, I)$ where $I = \left((1,0)^T, (0,1)^T\right)$ is the covariance matrix, and $\mu_k^{(1)}, \mu_k^{(2)}$ denote the class-specific means. Specifically, we set

$$\begin{aligned} &\mu_1^{(1)} = (-a, a),\ \mu_1^{(2)} = (a, a),\\ &\mu_2^{(1)} = (2a, 0),\ \mu_2^{(2)} = (a, -a),\\ &\mu_3^{(1)} = (-a, -a),\ \mu_3^{(2)} = (-2a, 0), \end{aligned} \tag{12}$$

where a represents the separability of X among three classes. For each dataset, we randomly generate N samples, where Y follows a multinomial distribution with three classes of equal probability, and X is sampled from the aforementioned mixture of Gaussian distributions. Then we obtain data points $(\Gamma_l X, Y)$ as the lth dataset, where $\Gamma_l (l = 1 \cdots, 6)$ are the rotation matrices $\Gamma_l = \left((\cos\theta_l, -\sin\theta_l)^T, (\sin\theta_l, \cos\theta_l{}^T)\right)$, and $\theta_l = 20 \times (l-1)$. Notice that the covariates X from each dataset have six clusters corresponding to different labels Y; we index these clusters as $c = 1, \ldots, 6$. Next we split data points from each dataset into a training set with $N_{train} = \sum_{s=1}^{6} n_l^{(s)}$ data points, and a testing set with $N_{test} = \sum_{s=1}^{6} n_p^{(s)}$ data points, where $N_{train} + N_{test} = N$. For the first dataset, we sample 90% of the training data from clusters $c = \{1, 3, 6\}$ and 10% from the other three clusters. Then the remaining data points serve as testing data. The generation of the training and testing data for other five datasets follows the same procedure, while the training data are sampled from different clusters. Specifically, the cluster indexes of training data for the second to sixth dataset are $\{1, 4, 6\}, \{2, 3, 5\}, \{2, 3, 6\}, \{1, 4, 5\}, \{2, 4, 5\}$, respectively. The above data rotation and training data sampling scheme increase the heterogeneity in the distribution of X across different datasets.

In the following, we compare the classification performance of the data integration procedure in Section 3.1 with that of the individual learning scheme. For individual learning, we independently train a classifier on the training data, and then make predictions on the testing data for each dataset. To construct the node graph for label propagation classifier, we use a radial basis function kernel to measure the distance between samples. Whereas, for the data integration scheme, we train the classifier for the lth dataset by using the training data from both the lth dataset and M datasets. The number of datasets for integration varies as $M = 1, 3, 5$. We then calculate the classification accuracy averaged from the 6 datasets using two different schemes. The comparisons are conducted under different sizes of training set N_{train} and testing set N_{test}. In addition, we measure the separability intensity of the classification as pairwise distances among cluster centers as $\Delta := \sum_{1\leq k_1 \leq k_2 \leq 3} \sum_{1 \leq i \leq j \leq 2} \|\mu_{k_1}^i - \mu_{k_2}^j\|_2$, and denote the intensity as Δ_0, Δ_1, Δ_2 when a in (12) takes values 0.7, 1, 1.5. A larger a indicates a strong separation of classifications.

Table 1 shows that the proposed data integration method can improve the classification accuracies for both supervised learning classifier QDA and semi-supervised learning classifier LP by 5%–21% over different simulation schemes. Notice that the improvement from the data integration increases

Table 1. Classification accuracies over 30 repeated experiments from the individual learning scheme (Ind.) and data integration scheme (ILA), where the data integration uses M datasets.

Signal intensity	N_{train}/N_{test}	LP				QDA			
		Ind.	ILA			Ind.	ILA		
			$M=1$	$M=3$	$M=5$		$M=1$	$M=3$	$M=5$
Δ_0	90/120	0.64(0.02)	0.67(0.02)	0.67(0.02)	0.67(0.02)	0.67(0.02)	0.70(0.02)	0.71(0.02)	0.72(0.02)
	150/150	0.65(0.02)	0.68(0.02)	0.69(0.02)	0.69(0.02)	0.67(0.02)	0.71(0.02)	0.72(0.02)	0.73(0.02)
	240/180	0.62 (0.02)	0.66(0.02)	0.66(0.02)	0.67(0.02)	0.63(0.01)	0.69(0.01)	0.70(0.01)	0.71(0.01)
$\Delta_1 = 1.4\Delta_0$	90/120	0.73(0.02)	0.77(0.02)	0.78(0.02)	0.78(0.03)	0.75(0.02)	0.81(0.02)	0.81(0.02)	0.82(0.02)
	150/150	0.73(0.02)	0.77(0.02)	0.78(0.02)	0.78(0.02)	0.74(0.02)	0.80(0.02)	0.82(0.02)	0.82(0.02)
	240/180	0.66 (0.02)	0.72(0.04)	0.73(0.04)	0.73(0.04)	0.67(0.02)	0.77 (0.02)	0.78(0.02)	0.79(0.02)
$\Delta_2 = 2.2\Delta_0$	90/120	0.85(0.03)	0.89(0.02)	0.90(0.02)	0.90(0.02)	0.84(0.03)	0.91(0.02)	0.92(0.02)	0.92(0.02)
	150/150	0.80(0.02)	0.91(0.02)	0.92(0.01)	0.92(0.01)	0.86(0.02)	0.93(0.01)	0.94(0.02)	0.94(0.01)
	240/180	0.71 (0.02)	0.77(0.07)	0.79(0.06)	0.80(0.06)	0.71(0.02)	0.84 (0.03)	0.85(0.03)	0.86(0.03)

NOTE: The classification accuracies reported in this table is the average over six different datasets.

as the number of incorporated datasets and the size of the training set within N_{train} increase. This improvement pattern implies that the proposed alignment scheme captures the underlying common conditional distribution of label Y given covariate X, which boosts the effective sample size of training data on each dataset. In addition, the improvement of our method is more significant when the separability of X across different classes is strong, implying that the improvement also depends on the signal intensity of the shared response-covariate patterns within each individual dataset. In addition, we perform an additional simulation study where the labels are unbalanced in training sets. The results show that the proposed method achieves higher classification accuracies compared with the individual learning method. The detailed simulation setup and results are in Section 12 of the supplemental materials.

5.2. Simulations for Temporal Integration for Multiple Datasets

In this section, we consider simulation studies with the two-dimensional covariates dynamically changing over time. We compare the classification performance of integrating the temporal data procedure in Section 3.2 with that of the individual learning approach. For individual learning, we independently train a classifier on the training data, and then make a prediction on the testing data from each dataset at each time point. For our temporal integration scheme, we train the classifier on the common latent space integrating aligned training data from different datasets and time points. We then calculate the classification accuracy averaged from the six temporal datasets using these two approaches.

Specifically, we consider data points with three classes $Y = k(k = 1, 2, 3)$. For each dataset at time point $t = 1$, we generate the covariates in the same way as Sections 4.1, and use $X^{(1)}$ to denote the observed X at time $t = 1$. To mimic the temporal pattern in X over time points $t = 1, \ldots, T$, we propose a sequential update on $X^{(t)}$ based on $X^{(t-1)}$ as $X^{(t)} = \boldsymbol{b} + \boldsymbol{A}X^{(t-1)}$, where $\boldsymbol{b}$ introduces a global drift of X from time point $t-1$ to t, and $\boldsymbol{A}$ provides a transformation on $X^{(t-1)}$. The stochastic drift term is $\boldsymbol{b} = (b_1, b_2)$, where b_1, b_2 are sampled from uniform distribution Unif$(0, 0.4)$. The transformation matrix $\boldsymbol{A}$ is defined as follows:

$$\boldsymbol{A} = \begin{pmatrix} 1+\lambda & 0 \\ 0 & 1+\lambda \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$

where λ and θ are randomly sampled from Unif$(-0.2, 0.2)$ and Unif$(-0.2\pi, 0.2\pi)$, respectively. Notice that the labels of the data points do not change during the process. Therefore, we construct a data sequence $(X^{(1)}, X^{(2)}, \ldots, X^{(T)}, Y)$ for each dataset. Then we split data points from each dataset into 30 training samples and 90 testing samples following the same procedure discussed in Section 5.1. In the following, we perform numerical experiments under different lengths of time $T = 4, 8, 12, 20$, using different numbers of integrated datasets $M = 1, 3, 5$.

Table 2. Classification accuracies from the individual learning approach (Ind.) and the proposed temporal data integration scheme (ILA) under different lengths of time T, where the proposed method uses M datasets.

Length of time	Num of datasets	Classifier			
		LP		QDA	
		Ind.	ILA	Ind.	ILA
$T=4$	$M=1$	0.834(0.03)	0.848(0.04)	0.813(0.04)	0.832(0.04)
	$M=3$	0.832(0.02)	0.854(0.03)	0.810(0.03)	0.867(0.02)
	$M=5$	0.822 (0.02)	0.849(0.02)	0.802(0.03)	0.855(0.03)
$T=8$	$M=1$	0.830(0.03)	0.846(0.04)	0.813(0.04)	0.836(0.05)
	$M=3$	0.826(0.02)	0.850(0.03)	0.810(0.03)	0.866(0.02)
	$M=5$	0.816 (0.02)	0.847(0.02)	0.802(0.03)	0.857(0.02)
$T=12$	$M=1$	0.825(0.03)	0.849(0.04)	0.813(0.04)	0.831(0.04)
	$M=3$	0.821(0.02)	0.853(0.02)	0.810(0.03)	0.866(0.03)
	$M=5$	0.812 (0.02)	0.847(0.02)	0.802(0.03)	0.856(0.02)
$T=20$	$M=1$	0.820(0.03)	0.846(0.04)	0.813(0.04)	0.831(0.05)
	$M=3$	0.812(0.03)	0.849(0.03)	0.810(0.03)	0.866(0.02)
	$M=5$	0.803 (0.02)	0.845(0.02)	0.802(0.03)	0.855(0.02)

NOTE: QDA and LP stand for quadratic discriminative analysis and label propagation algorithm, respectively.

Table 2 shows that the proposed temporal data integration method can improve the classification accuracies for both classifiers QDA and LP compared with the individual learning even when the length of time is short ($T = 4$) or the number of incorporated datasets is small ($M = 1$). These results demonstrate the effectiveness of temporal alignment in reducing the systematic heterogeneity even with a small sample size. For both the QDA and LP classifiers, the improvement from temporal integration increases as T increases given that M is fixed. On the other hand, the improvement also increases as M increases given

Table 3. Classification accuracies from the individual learning approach (Ind.) and the proposed temporal data integration method (ILA) under different number of time points T, where the proposed method integrates M datasets.

Length of time	Num of datasets	Classifier			
		LP		QDA	
		Ind.	ILA	Ind.	ILA
$T = 4$	$M = 2$	0.611(0.03)	0.622(0.04)	0.572(0.04)	0.597(0.05)
	$M = 4$	0.604(0.03)	0.670(0.03)	0.562(0.03)	0.618(0.04)
	$M = 6$	0.612(0.02)	0.688(0.03)	0.566(0.03)	0.638(0.04)
$T = 10$	$M = 2$	0.656(0.03)	0.673(0.03)	0.615(0.03)	0.622(0.03)
	$M = 4$	0.645(0.03)	0.741(0.04)	0.608(0.02)	0.685(0.04)
	$M = 6$	0.651(0.02)	0.770(0.03)	0.608(0.02)	0.722(0.04)
$T = 16$	$M = 2$	0.708(0.03)	0.731(0.04)	0.655(0.03)	0.687(0.04)
	$M = 4$	0.698(0.03)	0.790(0.05)	0.645(0.02)	0.755(0.05)
	$M = 6$	0.699(0.02)	0.816(0.03)	0.645(0.02)	0.795(0.05)

NOTE: Here QDA and LP stand for quadratic discriminative analysis and label propagation algorithm, respectively. The spike data are simulated based on Poisson generalized linear modeling and latent variable modeling.

that T is fixed. This result implies that the proposed method can use common temporal patterns on the conditional distribution of response to further enhance the power of detecting the patterns shared across different datasets. In above experiments, we investigate the classification accuracies of each time point, and calculate the average classification accuracies on the trajectory. A higher average classification accuracies indicates that the aligned temporal pattern can better capture the overall dynamic of response-covariate association.

5.3. Simulations for Synthetic Neuronal Spike Data

In this section, we use simulation experiments that mimic the application of the proposed data integration method to real neuronal data. To this end, we simulate neuronal spike data by following the widely used Poisson generalized linear modeling and latent variable modeling (Keeley et al. 2020). Specifically, we consider a total of $M = 6$ datasets, each consisting of spike counts over $T = 20$ time points. The spiking patterns in the mth dataset at the tth time point are governed by dynamic low-dimensional latent status $X_t^{(m)} \in \mathbb{R}^k, m = 1, \ldots, 6, t = 1, \ldots, 20$, where k denotes the number of latent factors. The spike counts for the jth neuron at the ith trial are generated as follows: $S_j^{(i)}(t) \mid X_t^{(m)} \sim \text{Poisson}\big(\sigma(W_{ij} X_t^{(m)})\big), i = 1, \ldots, n, \ j = 1, \ldots, p$, where $W_{ij} \in \mathbb{R}^p$ denotes the loading vector, and $\sigma(\cdot) = \frac{\exp(\cdot)}{1+\exp(\cdot)}$ is the link function. Here we set the number of trials to $n = 60$ and the number of neurons to $p = 60$. The latent states $X_t^{(m)}$ are sampled from a mixture of Gaussian distributions with multiple clusters corresponding to different external stimuli. The detailed generation procedure for $\{X_t^{(m)}\}$ is similar as the setup in Section 5.2. We perform stimuli classification based on the labeled spike data $\{S^{(i)}(t), Y^{(i)}\}_{i=1}^n$. Then classifiers are trained on the integrated data points and are compared with the individual learning method. Table 3 illustrates the average classification accuracies across different numbers of datasets and time points. Table 3 shows that the proposed temporal data integration method can improve the classification accuracies for both classifiers compared to the independent training scheme. For both classifiers, the improvement from the proposed method increases as T or M increases, and the improvement maximizes at about 25% when $T = 16$ and $M = 6$. The simulation results show the effectiveness of the temporal alignment in reducing the systematic heterogeneity in subject-wise neural data.

6. Real Data Application

In this section, we apply the proposed ILA framework to a rodent electrophysiological study, where the neuronal spiking data were collected as rats performed a non-spatial sequence memory task Allen et al. (2014). Briefly, the task involves repeated presentations of a sequence of odors (A,B,C,and D) through the same port, and requires subjects (rats) to determine whether each odor is presented "in sequence" (InSeq; e.g., AB<u>C</u>…) or "out of sequence" (OutSeq; e.g., AB<u>D</u>…) to receive a reward.

Here we use spike data from five rats as they are presented different odors in the sequence memory task. The spike data from the lth rat is reformulated as a three-dimensional data array $\mathcal{A}^{(l)} = \{A_{ijk}\}, l = 1, 2, 3, 4, 5$ where the i, j, k represent the trials, neurons, and time points for each rat, respectively. The spike data $\mathcal{A}^{(l)}$ consists of binary elements, where $A_{ijk} = 1$ indicates that a specific neuron fired and $A_{ijk} = 0$ otherwise. In each trial, the subjects are presented with one of four odors labeled "A", "B", "C", and "D". We then extract 4 sec of spike data per trial (2 sec before and after odor onset) in bins of 10 ms, resulting in 400 time points for each trial. In the following, we investigate a subset of the original spike data from time interval between the 250 millisecond and 500 millisecond after the onset of odors. This time interval of the data was shown to exhibit the strongest differentiation of odor information, as well as significant differentiation of neuronal coding for different temporal order and trial outcome information (Shahbaba et al. 2022). In summary, the dimensions corresponding to trials, neurons, and time points of spike data for the five rats are $168 \times 46 \times 25$, $152 \times 49 \times 25$, $207 \times 104 \times 25$, $133 \times 92 \times 25$, and $174 \times 79 \times 25$, respectively. More detailed background on the electrophysiological study and neural spiking data are provided in Section 1 and 11 of the supplemental materials.

In the following, we first compress the original temporal spike data into a low-dimensional latent space for each rodent by using autoencoder with implementation details in the Section 10 of the supplemental materials. Through the above process, the spike data from each 100ms sub-window are compressed into two-dimensional latent features. We collect the preprocessed spike data from the 16 sub-windows for the sth rodent ($s = 1, 2, 3, 4, 5$) as $(X^{(s)}(1), X^{(s)}(2), \ldots, X^{(s)}(16), Y^{(s)})$, where $X^{(s)}(t) \in \mathbb{R}^{n_l \times 2}$ denotes the latent neural representation of the sth rat at the tth sub-windows ($t = 1, \ldots, 16$) and n_l is the number of trials from the lth rat. In addition, $Y^{(s)} \in \{1, 2, 3, 4\}^{n_s}$ denotes the odor labels "A", "B", "C", "D" corresponding to each trial of the sth rat. In the following, we first apply the proposed integrated supervised learning method in Section 3.1 to the preprocessed spike data, and compare the results with that of individual learning on the odor classification. Specifically, we sample 20% of the trials from each rat as training data, and the remaining 80% of trials serve as testing data. Then we train a classifier by our integration method and the individual learning method separately, and investigate the corresponding classifi-

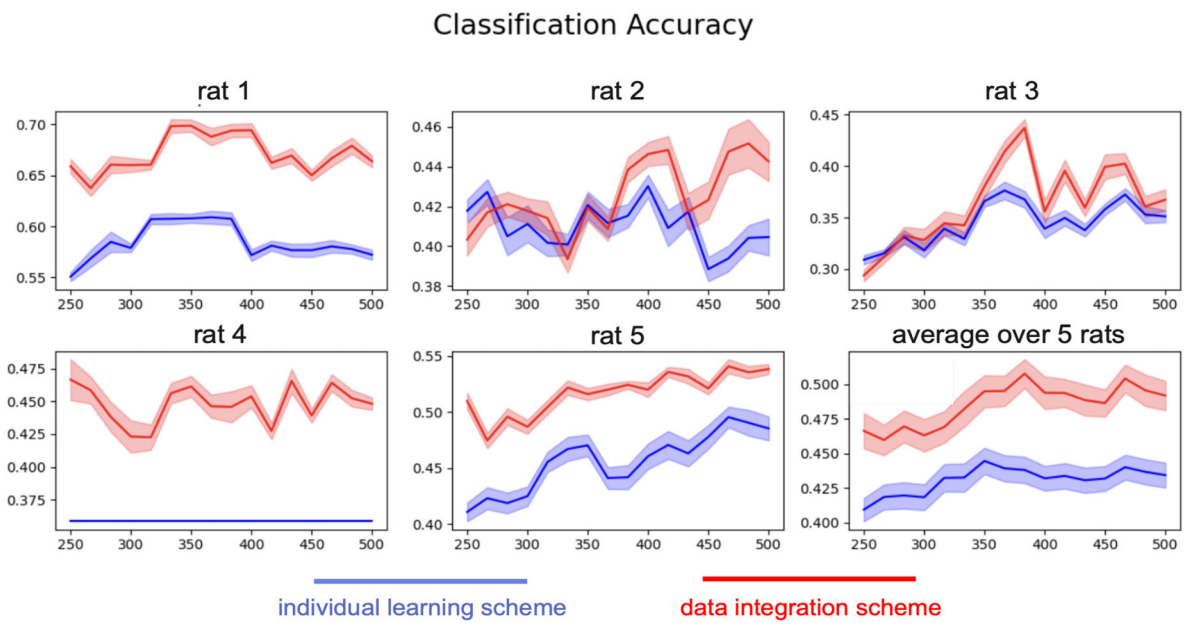

Figure 1. The four-odor classification accuracy based on the three rodents' latent neural features extracted by autoencoder within time interval 250 $\sim$ 500 milliseconds. The blue line denotes the classification accuracy from KNN trained by the individual learning method at each time point. The red line denotes KNN's classification accuracy trained by the proposed data integration method in Section 3.1.

cation accuracies for the two methods on each sub-window. Following (Shahbaba et al. 2022), we use the K-nearest neighbors algorithm (KNN) as our classifier to allow for nonlinear boundaries among the latent neural representations from different odors.

Figure 1 illustrates the classification accuracy of KNN trained at each 100*ms* sub-windows by the individual learning method (blue line) and the proposed integration learning method (red line), respectively. The results show that the proposed ILA method achieves higher odor classification accuracies for all five rodents over the sequential sub-windows. The improvement from ILA suggests that the neuronal ensembles from the five rodents share a common feature: they exhibit a differentiation of odor information within that time window. More specifically, it shows that our approach can detect the common neuronal coding among the different subjects, and uses the shared pattern to enhance the performance of odor classification. The result shows that the data integration method achieves about 10% higher average classification accuracy compared with the individual learning method. For this specific application, this is a substantial improvement.

In the following, we investigate odor classification performance of the temporal integration method introduced in Section 3.2, and compare the performance of the individual learning approach based on the five rodents spike samples within the time interval 250 $\sim$ 500 milliseconds. For each rodent, we sample the odor labels and the corresponding latent neuronal coding from 20% of the trials as training data, and the remaining 80% of the trials as testing data. Then we train the KNN classifier by our temporal integration method and individual learning method separately, and examine the classification accuracies of the two methods on each 16 sub-windows. We use the individual learning scheme as the baseline and provide the classification comparison Figure 2. The results show about 10% accuracy improvements from the proposed method for rodent 2 and rodent 4 across all time intervals. For rodent 1, rodent 3, and rodent 5, the accuracy of our method is higher than the individual learning method at most of the time points, and the improvements can be about 10% at some time windows. In addition, compared with the individual learning method, the accuracy of our method changes more smoothly over the time because it uses the time dependency among latent representations of spike data. This is another advantage of the proposed approach in terms of scientific interpretation in addition to improving classification accuracy.

In summary, the proposed alignment approach can integrate information from both different subjects and different time windows. The improvement from the temporal data integration on the five subjects implies the existence of a common structure in the neuronal coding associated with the processing of the different odors. More importantly, the dynamic of the neuronal coding is also shared across subjects to some extent. This finding is consistent with prior evidence that hippocampal neurons represent information about task events critical for accurate performance (Allen and Fortin 2013; Eichenbaum 2014).

We also use the temporal data integration in Section 3.2 to investigate the common dynamic pattern of neuronal coding among the five rodents. Specifically, we obtain the rodent-wise alignment mapping $\{\Omega^{(s)}\}_{s=1}^{5}$ and temporal similarity matrices $\{D(t)\}_{t=1}^{16}$ of 16 sub-windows in the temporal integration where we use the true odor labels to construct the label distance matrices $\{M^{(s)}(t)\}, s = 1, \ldots, 5, \quad t = 1, \ldots, 16$. Figure 3 shows neural activity trajectories for all five rodents throughout the 0–1500ms period. Our data integration approach also allows the visualization of each subject's neural activity in the aligned latent space (Fig 3c), which reveals comparable trajectories across the five subjects. The results show that the neural activity on odor A trials is very distinct than that of odors B, C, and D trials. This can be explained by the fact that the sequence always starts with odor A, and it is preceded by a running period (so that activity

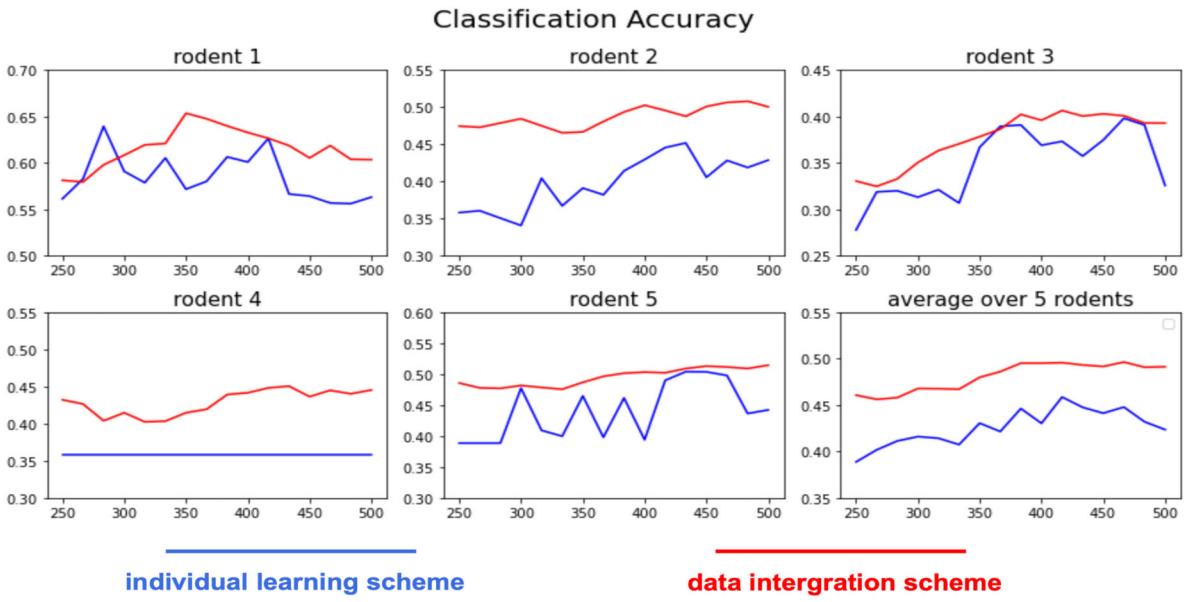

Figure 2. The four-odor classification accuracy within time interval 250 ∼ 500 milliseconds based on rodents' electrophysiological data. The blue line denotes the classification accuracy from KNN trained by the individual learning method at each time point. The red line denotes KNN's classification accuracy trained by the proposed temporal integration method in Section 3.1.

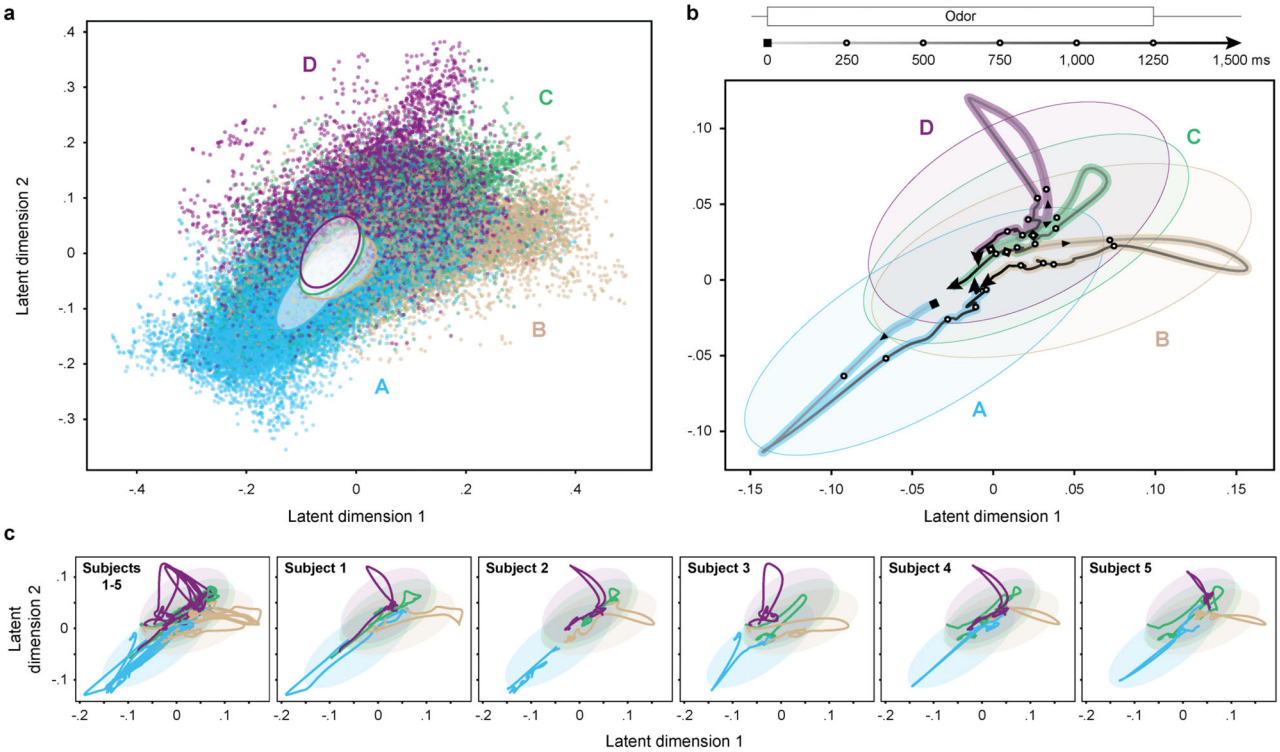

Figure 3. Odor-specific neuronal coding and dynamics in the aligned latent space. (a) Differentiation of the neural activity across odor presentations on the aligned latent dimensions (odors A-D; data aggregated across all five subjects). Each dot indicates the location of the neural activity for a 100-ms bin from a given subject, with the color denoting the odor label of the corresponding trial. (b) Neural activity trajectory (mean across subjects) on the latent dimensions during each odor trial type. (c) Neural activity trajectories for each subject showing comparable patterns across subjects. Ellipses are consistent across panels and represent the spread of the latent coding for each odor type, specifically the two largest square-rooted singular values from the covariance of latent coding within each cluster.

dynamics associated with running extend into the odor period). Therefore, we focus on comparing the coding properties and dynamics observed during odors B–D, in which the behavioral and cognitive demands are matched. Specifically, the activity on odors B, C, and D trials starts at a similar location in the latent space and heads toward the center of their respective clusters (0–250 ms), reaching maximum differentiation in the 250-500 ms period. Toward the end of the odor presentations (750–1250 ms), the activity for each odor then converges back toward the beginning of their trajectory (faster for odor D, than C, than B), before reaching a distinct region of the latent space in the post-trial activity (1500ms; indicated by arrowheads). These

results would have been difficult to interpret without this data integration approach, as the neural activity of each subject would otherwise reside in distinct latent spaces, making cross-subject comparisons impossible.

7. Discussion

In this article, we propose a novel and general data integration framework based on optimal transport for supervised learning tasks. Our method is effective when the data domains and dimensionalities of each dataset differ, and it remains effective even with a small number of datasets. We provide comprehensive theoretical analysis and performance guarantees for the proposed method. We apply our framework to analyze electrophysiological data from a rodent study investigating nonspatial memory. The method identifies common neuronal coding patterns across subjects during the processing of different odors. This finding aligns with previous research showing that hippocampal neurons represent information critical for task performance. Our results suggest that integrating heterogeneous datasets can enhance the discriminative power of predictive models, fostering advancements in scientific research and related fields.

Supplementary Materials

The supplementary material includes data descriptions, optimization implementation details, additional numerical experiments, and proofs of the theorems.

Acknowledgments

The authors thank the Associate Editor and anonymous reviewers for helpful suggestions which improved the article significantly.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

This work is supported by NSF Grants DMS 2210640, DMS 1763272, CDS&E-MSS 2401271, NCS 2319618, Simons foundation 594598, NIH R01 DC017687, NIH R01 MH115697, NSF BCS-1439267, and NSF CAREER IOS-1150292.

ORCID

Norbert Fortin http://orcid.org/0000-0002-6793-6984
Keiland Cooper http://orcid.org/0000-0002-0358-9645
Keiland Cooper http://orcid.org/0000-0002-8804-3368
Qing Nie http://orcid.org/0000-0002-8396-7828

References

Agueh, M., and Carlier, G. (2011), "Barycenters in the Wasserstein Space," *SIAM Journal on Mathematical Analysis*, 43, 904–924. [2,7]

Akil, H., Martone, M. E., and Van Essen, D. C. (2011), "Challenges and Opportunities in Mining Neuroscience Data," *Science*, 331, 708–712. [1]

Allen, T. A., and Fortin, N. J. (2013), "The Evolution of Episodic Memory," *Proceedings of the National Academy of Sciences*, 110, 10379–10386. [11]

Allen, T. A., Morris, A. M., Mattfeld, A. T., Stark, C. E., and Fortin, N. J. (2014), "A Sequence of Events Model of Episodic Memory Shows Parallels in Rats and Humans," *Hippocampus*, 24, 1178–1188. [10]

Ambrosio, L., Gigli, N., and Savaré, G. (2008), *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*, Basel: Springer. [6]

Audibert, J.-Y., and Tsybakov, A. B. (2007), "Fast Learning Rates for Plug-in Classifiers," *The Annals of Statistics*, 35, 608–633. [6]

Beier, F., and Beinert, R. (2024), "Tangential Fixpoint Iterations for Gromov-Wasserstein Barycenters," arXiv preprint arXiv:2403.08612. [3,7]

Beier, F., Beinert, R., and Steidl, G. (2023), "Multi-Marginal Gromov–Wasserstein Transport and Barycentres," *Information and Inference: A Journal of the IMA*, 12, 2753–2781. [3]

Bigot, J., and Klein, T. (2018), "Characterization of Barycenters in the Wasserstein Space by Averaging Optimal Transport Maps," *ESAIM: Probability and Statistics*, 22, 35–57. [7]

Boufares, F., and Salem, A. B. (2012), "Heterogeneous Data-Integration and Data Quality: Overview of Conflicts," in *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, pp. 867–874, IEEE. [1]

Cherian, A., and Aeron, S. (2020), "Representation Learning via Adversarially-Contrastive Optimal Transport," in *International Conference on Machine Learning*, pp. 1820–1830, PMLR. [2]

Cook, R. D., and Weisberg, S. (1980), "Characterizations of An Empirical Influence Function for Detecting Influential Cases in Regression," *Technometrics*, 22, 495–508. [6]

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. (2017), "Joint Distribution Optimal Transportation for Domain Adaptation," in *Advances in Neural Information Processing Systems* (Vol. 30). [2,7]

Courty, N., Flamary, R., and Tuia, D. (2014), "Domain Adaptation with Regularized Optimal Transport," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 274–289, Springer. [2]

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2016), "Optimal Transport for Domain Adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39, 1853–1865. [2]

Deb, N., Ghosal, P., and Sen, B. (2021), "Rates of Estimation of Optimal Transport Maps Using Plug-in Estimators via Barycentric Projections," *Advances in Neural Information Processing Systems*, 34, 29736–29753. [6]

Ebrahimi, M., Calarco, N., Campbell, K., Hawco, C., Voineskos, A., and Khisti, A. (2020), "Time-Resolved fMRI Shared Response Model Using Gaussian Process Factor Analysis," arXiv preprint arXiv:2006.05572 [2]

Eichenbaum, H. (2014), "Time Cells in the Hippocampus: A New Dimension for Mapping Memories," *Nature Reviews Neuroscience*, 15, 732–744. [11]

Farahani, A., Voghoei, S., Rasheed, K., and Arabnia, H. R. (2021), "A Brief Review of Domain Adaptation," in *Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020*, pp. 877–894. [1]

Frogner, C., Zhang, C., Mobahi, H., Araya, M., and Poggio, T. A. (2015), "Learning with a Wasserstein Loss," in *Advances in Neural Information Processing Systems* (Vol. 28). [2]

Gebru, T., Hoffman, J., and Fei-Fei, L. (2017), "Fine-Grained Recognition in the Wild: A Multi-Task Domain Adaptation Approach," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1349–1358. [2]

Gong, B., Shi, Y., Sha, F., and Grauman, K. (2012), "Geodesic Flow Kernel for Unsupervised Domain Adaptation," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2066–2073, IEEE. [2]

Gopalan, R., Li, R., and Chellappa, R. (2011), "Domain Adaptation for Object Recognition: An Unsupervised Approach," in *2011 International Conference on Computer Vision*, pp. 999–1006, IEEE. [2]

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. (2008), "Covariate Shift by Kernel Mean Matching," *Dataset Shift in Machine Learning*, 3, 5. [1]

Halevy, A., Rajaraman, A., and Ordille, J. (2006), "Data Integration: The Teenage Years," in *Proceedings of the 32nd International Conference on Very Large Data Bases*, pp. 9–16. [1]

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. (2006), "Correcting Sample Selection Bias by Unlabeled Data," in *Advances in Neural Information Processing Systems* (Vol. 19). [1]

Keeley, S. L., Zoltowski, D. M., Aoi, M. C., and Pillow, J. W. (2020), "Modeling Statistical Dependencies in Multi-Region Spike Train Data," *Current Opinion in Neurobiology*, 65, 194–202. [10]

Koh, P. W., and Liang, P. (2017), "Understanding Black-Box Predictions via Influence Functions," in *International Conference on Machine Learning*, pp. 1885–1894, PMLR. [6]

Kouw, W. M., and Loog, M. (2019), "A Review of Domain Adaptation Without Target Labels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43, 766–785. [1]

Lindén, H., Petersen, P. C., Vestergaard, M., and Berg, R. W. (2022), "Movement is Governed by Rotational Neural Dynamics in Spinal Motor Networks," *Nature*, 610, 526–531. [2]

Long, M., Cao, Y., Wang, J., and Jordan, M. (2015), "Learning Transferable Features with Deep Adaptation Networks," in *International Conference on Machine Learning*, pp. 97–105, PMLR. [2]

Long, M., Wang, J., Ding, G., Sun, J., and Yu, P. S. (2013), "Transfer Feature Learning with Joint Distribution Adaptation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2200–2207. [2]

Mémoli, F. (2014), "The Gromov–Wasserstein Distance: A Brief Overview," *Axioms*, 3, 335–341. [3]

Mikkelsen, G., and Aasly, J. (2001), "Concordance of Information in Parallel Electronic and Paper based Patient Records," *International Journal of Medical Informatics*, 63, 123–131. [1]

Redko, I., Habrard, A., and Sebban, M. (2017), "Theoretical Analysis of Domain Adaptation with Optimal Transport," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part II 10*, pp. 737–753, Springer. [7]

Sejnowski, T. J., Churchland, P. S., and Movshon, J. A. (2014), "Putting Big Data to Good Use in Neuroscience," *Nature Neuroscience*, 17, 1440–1441. [1]

Shahbaba, B., Li, L., Agostinelli, F., Saraf, M., Cooper, K. W., Haghverdian, D., Elias, G. A., Baldi, P., and Fortin, N. J. (2022), "Hippocampal Ensembles Represent Sequential Relationships Among An Extended Sequence of Nonspatial Events," *Nature Communications*, 13, 1–17. [10,11]

Tian, Y., Gu, Y., and Feng, Y. (2023), "Learning from Similar Linear Representations: Adaptivity, Minimaxity, and Robustness," arXiv preprint arXiv:2303.17765. [2]

Titouan, V., Courty, N., Tavenard, R., and Flamary, R. (2019), "Optimal Transport for Structured Data with Application on Graphs," in *International Conference on Machine Learning*, pp. 6275–6284, PMLR. [2]

Vayer, T., Chapel, L., Flamary, R., Tavenard, R., and Courty, N. (2020), "Fused Gromov-Wasserstein Distance for Structured Objects," *Algorithms*, 13, 212. [3]

Yan, Y., Wang, K., and Rigollet, P. (2024), "Learning Gaussian Mixtures Using the Wasserstein–Fisher–Rao Gradient Flow," *The Annals of Statistics*, 52, 1774–1795. [2]

Zhang, Q., and Chen, J. (2022), "Distributed Learning of Finite Gaussian Mixtures," *Journal of Machine Learning Research*, 23, 1–40. [2]

Zhang, Z., Goldfeld, Z., Mroueh, Y., and Sriperumbudur, B. K. (2024), "Gromov–Wasserstein Distances: Entropic Regularization, Duality and Sample Complexity," *The Annals of Statistics*, 52, 1616–1645. [6]