Application of Unbalanced Optimal Transport in Healthcare

Qui Phu Pham[1]*, Nghia Thu Truong[2]*, Hoang-Hiep Nguyen-Mau[3], Cuong Nguyen[4],
Mai Ngoc Tran[5], Dung Luong[6]
University of California, Irvine, California, USA[1]
Pasadena City College, California, USA[2]
VinUniversity, Hanoi, Vietnam[3,4]
Binh Duong University, Ho Chi Minh City, Vietnam[5]
VietDynamic, Ho Chi Minh City, Vietnam[6]
Acuitas Education, Ho Chi Minh City, Vietnam[5]

***Abstract*—Optimal Transport (OT) is a powerful tool widely used in healthcare applications, but its high computational cost and sensitivity to data changes make it less practical for resource-constrained settings. These limitations also contribute to increased environmental impact due to higher CO2 emissions from computing. To address these challenges, we explore Unbalanced Optimal Transport (UOT), a variation of OT that is both computationally efficient and more robust to data variability. We apply UOT to two healthcare scenarios: independence testing on breast cancer data and modeling heart rate variability (HRV). Our experiments show that UOT not only reduces computational costs but also delivers reliable results, making it a practical alternative to OT for socially impactful applications.**

***Keywords*—Optimal transport; unbalanced optimal transport; healthcare**

I. Introduction

Optimal Transport (OT), first formulated by Gaspard Monge [23] and further developed by Kantorovic [15], addresses the fundamental question of finding the most efficient way to minimize the cost of transporting mass from one distribution to another. OT has evolved into many practical applications in fields such as healthcare [42], [39], [41], machine learning [12] and domain adaptation [7]. For healthcare applications such as breast cancer detection [39] or heart rate variability (HRV) modeling [42], which may be needed widely by also resource-constrained medical institutes and have a direct impact on human wellness, the computational efficiency and the robustness of the deployed models are paramount. Nevertheless, OT has been known to suffer from computational bottleneck [21] and sensitivity to problem structure or data perturbation [17]. Such limitations of vanilla OT can make solution models to these healthcare applications not accessible to medical institutes with limited budget [36], raise the CO2 output of computing resources thereby negatively impacting the environment [16], and become a less reliable tool in the realm of healthcare [8].

To alleviate the above limitations, Unbalanced Optimal Transport (UOT) is recently proposed variant of the classical OT formulation that penalizes the marginal constraints based on some given divergence. Among the various divergences used in the literature such as Kullback-Leiber (KL) divergence [6], squared ℓ_2 norm [3], ℓ_1 norm [4], and ℓ_p norm [18], UOT with KL divergence is the most prominent for its wide applicability, flexibility and efficient computation [30]. UOT has shown its prominence in various applications in statistics and machine learning [11]. Recent works [35], [17], [30] have facilitated the fast computation of UOT and provided guarantees on its statistical and approximation properties.

Recent advancements in UOT have significantly reduced its computational complexity and improved its scalability, enabling its application in large-scale machine learning tasks. Notable among these advancements are efficient gradient-based methods [30], which not only accelerate UOT computations but also provide theoretical guarantees for convergence and statistical properties. These methods are especially important in scenarios requiring real-time processing, such as medical diagnostics or dynamic resource allocation in healthcare. Furthermore, UOT has been shown to be more robust than traditional OT in handling outliers and noisy data, making it a valuable tool in applications where data quality is variable [17].

As computational efficiency and environmental concerns become more critical in the age of large-scale AI, UOT stands out as a method that balances performance with resource utilization. By relaxing the mass conservation constraint, UOT reduces the computational burden while maintaining the accuracy required in sensitive applications such as healthcare. This reduced computational cost also has positive implications for sustainability, as it lowers the energy consumption and CO2 emissions associated with large-scale computations [16]. As a result, UOT not only provides a more flexible and scalable approach to OT problems but also addresses key limitations that have historically hindered the adoption of OT in resource-constrained settings.

A. Contributions

In this paper, we benchmark and empirically validate the effectiveness of UOT on various healthcare applications, which are the statistical independence test on the breast cancer dataset following the setting in [39] and HRV modeling in [42]. The code is given in https://github.com/quipp12/UOT_Healthcare.git. Our contributions can be summarized as follows:

- For both healthcare applications, statistical independence test [39] and HRV modeling [42], the OT

*Equal Contribution

distance is used as a component in these pipelines, where the celebrated Sinkhorn algorithm with the costly computational cost of $\tilde{O}\left(n^2\varepsilon^{-2}\right)$ [19] is used. To alleviate the computational bottleneck, we propose the adoption of UOT as well as the Sinkhorn variant specifically designed for UOT distance with improved complexity of $\tilde{O}\left(n^2\varepsilon^{-1}\right)$ [35]. This facilitates not only seamless integration of UOT (in place of OT) in these applications but also an acceleration in computation, which is consistently demonstrated in Section III and Section IV

- For HRV modeling [42], in addition to the realization of UOT's computational benefit, our experimental investigation reviews the high training cost from the original model using Gradient Descent (GD). Consequently, we implement the GD with Momentum (GDM) into the model to significantly expedite the training process, while maintaining comparable or even better Mean Squared Error (MSE)- the main performance metric (Section IV).
- Our primary theoretical contribution focuses on establishing a bound that quantifies the difference between the costs of Unbalanced Optimal Transport (UOT) and regular Optimal Transport (OT). We provide a rigorous guarantee that as the parameter controlling the mass relaxation in UOT increases, the difference between the UOT cost and the OT cost becomes smaller. This result ensures that the approximation made by UOT closely mirrors the exact cost computed by OT when enough relaxation is allowed.

II. Approximating Optimal Transport via Unbalanced Optimal Transport

A. Notations

Denote by $\mathbb{R}_+^n$ the set of all vectors in $\mathbb{R}^n$ with nonnegative entries. Bold capital letters and lowercase letters respectively stand for matrices and vectors. For $p \in [1, \infty), \|.\|_p$ denotes the l_p norm. The Frobenius inner product of two matrices of the same size is defined as $\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i,j=1}^{n} A_{ij} B_{ij}$.

B. Optimal Transport

Consider two discrete distributions $\mathbf{a}, \mathbf{b} \in \mathbb{R}_+^n$, specifically $\mathbf{a} := (a_1, \ldots, a_n), \mathbf{b} := (b_1, \ldots, b_n)$ with equal masses, i.e., $\|\mathbf{a}\|_1 = \|\mathbf{b}\|_1$. Denote $a_{min} = \min_{1 \le i \le n}\{a_i\}$ and $b_{min} = \min_{1 \le i \le n}\{b_i\}$ as the minimum masses. The OT problem seeks to find a matrix $\mathbf{X} \in \mathbb{R}_+^{n \times n}$ represented a transport plan which maps $\mathbf{a}$ to $\mathbf{b}$ at a minimum cost, i.e.

$$\mathbf{OT}(\mathbf{a}, \mathbf{b}) := \min_{\mathbf{X} \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \mathbf{X} \rangle, \tag{1}$$

where $\mathbf{C} \in \mathbb{R}_+^{n \times n}$ is a cost matrix whose entries are distances between measures of these distributions and $\Pi(\mathbf{a}, \mathbf{b}) := \left\{\mathbf{X} \in \mathbb{R}_+^{n \times n} : \mathbf{X}\mathbf{1}_n = \mathbf{a}, \mathbf{X}^\top \mathbf{1}_n = \mathbf{b}\right\}$. Denote by $\mathbf{X}^{\text{OT}} = \operatorname{argmin}_{\mathbf{X} \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \mathbf{X} \rangle$ be the optimal solution to the OT problem (1).

C. Unbalanced Optimal Transport

First, we define the $\mathbf{KL}$ divergence function between two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^n$ as

$$\mathbf{KL}(\mathbf{x} \| \mathbf{y}) := \sum_{i=1}^{n} \left(\mathbf{x}_i \log\left(\frac{\mathbf{x}_i}{\mathbf{y}_i}\right) - \mathbf{x}_i + \mathbf{y}_i \right)$$

Assume two finite measures $\mathbf{a}, \mathbf{b} \in \mathbb{R}_+^n$, specifically $\mathbf{a} := (a_1, \ldots, a_n), \mathbf{b} := (b_1, \ldots, b_n)$ with possibly different total mass. The UOT problem seeks to find a matrix $\mathbf{X} \in \mathbb{R}_+^{n \times n}$ represented a transport plan, i.e.

$$\begin{aligned} \mathbf{UOT}_{\mathbf{KL}}(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{X} \in \mathbb{R}_+^{n \times n}} \big\{ f(\mathbf{X}) := & \langle \mathbf{C}, \mathbf{X} \rangle + \tau \mathbf{KL}\left(\mathbf{X}\mathbf{1}_n \| \mathbf{a}\right) \\ & + \tau \mathbf{KL}\left(\mathbf{X}^\top \mathbf{1}_n \| \mathbf{b}\right) \big\} \end{aligned} \tag{2}$$

where $\mathbf{C} \in \mathbb{R}_+^{n \times n}$ is a given cost matrix and $\tau > 0$ is a given regularization parameter. Denote by $\mathbf{X}^{\text{UOT}} = \operatorname{argmin}_{\mathbf{X} \in \mathbb{R}_+^{n \times n}} f(\mathbf{X})$ be the optimal solution to the UOT problem (2).

The parameter τ effectively acts as a regularization term. A larger τ means a stronger penalty on the mass divergence, making the solution more balanced (closer to the original distributions $\mathbf{a}$ and $\mathbf{b}$). A smaller τ reduces the effect of the regularization term, allowing for more flexibility in the transport plan. With a very small τ, the solution may deviate significantly from the target distributions in favor of minimizing the transport cost.

D. Approximation Error of UOT

When $\mathbf{a}^\top \mathbf{1}_n = \mathbf{b}^\top \mathbf{1}_n$ and $\tau \to \infty, \mathbf{UOT}_{\mathbf{KL}}(\mathbf{a}, \mathbf{b})$ turns to the regular $\mathbf{OT}(\mathbf{a}, \mathbf{b})$.

Formally, taking the limit as $\tau \to \infty$ in the UOT objective function:

$$\lim_{\tau \to \infty} \left(\langle \mathbf{C}, \mathbf{X} \rangle + \tau \mathbf{KL}(\mathbf{X}\mathbf{1}_n \,||\, \mathbf{a}) + \tau \mathbf{KL}(\mathbf{X}^\top \mathbf{1}_n \,||\, \mathbf{b}) \right)$$

enforces $\mathbf{X}\mathbf{1}_n = \mathbf{a}$ and $\mathbf{X}^\top \mathbf{1}_n = \mathbf{b}$ at the optimal solution, which recovers the OT problem (1). Moreover, [30, Theorem 26] provided the tight non-asymptotic characterization on the distance gap between $\mathbf{UOT}_{\mathbf{KL}}(\mathbf{a}, \mathbf{b})$ and $\mathbf{OT}(\mathbf{a}, \mathbf{b})$ to be $O\left(\frac{1}{\tau}\right)$, where the big-$O$ notation here neglects the terms other than τ. Nevertheless, such results as above do not fully capture how well the UOT solution $\mathbf{X}^{\text{UOT}}$ can approximate the OT solution $\mathbf{X}^{\text{OT}}$ in the pure sense of transportation cost $\langle \mathbf{C}, \mathbf{X} \rangle$, which may better represents the raw performance within the application of interest. To this end, we establish in the next theorem such approximation error in transportation cost of using the UOT solution instead of the OT solution.

Theorem 1. *Under the balanced setting of $\|\mathbf{a}\|_1 = \|\mathbf{b}\|_1 = 1$, i.e. $\mathbf{a}$ and $\mathbf{b}$ are distributions, we have the following bound on the difference between the transportation costs incurred by UOT and OT solutions:*

$$0 \le \langle \mathbf{C}, \mathbf{X}^{OT} \rangle - \langle \mathbf{C}, \mathbf{X}^{UOT} \rangle \le O\left(\frac{1}{\tau}\right). \tag{3}$$

Proof: Let $\kappa = \min\{a_{min}, b_{min}\}^{-1}$. From [10, Theorem 1] that provides an upper bound for KL divergence, we have:

$$\begin{aligned} 0 \leq \mathbf{KL}(\mathbf{X}^{\text{UOT}}\mathbf{1}_n \| \mathbf{a}) &\leq \sum_{i=1}^{n} \frac{\left[(\mathbf{X}^{\text{UOT}}\mathbf{1}_n)_i - a_i\right]^2}{a_i} \\ &\overset{(i)}{\leq} \kappa \sum_{i=1}^{n} \left[(\mathbf{X}^{\text{UOT}}\mathbf{1}_n)_i - a_i\right]^2 \\ &= \kappa \|\mathbf{X}^{\text{UOT}}\mathbf{1}_n - \mathbf{a}\|_2^2 \\ &\overset{(ii)}{\leq} \kappa \|\mathbf{X}^{\text{UOT}}\mathbf{1}_n - \mathbf{a}\|_1^2, \end{aligned} \tag{4}$$

where for (i), we use $a_i \geq \min\{a_{min}, b_{min}\} = \kappa^{-1}$, and for (ii), we use $\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1$. Similarly, we obtain that:

$$0 \leq \mathbf{KL}\big((\mathbf{X}^{\text{UOT}})^\top \mathbf{1}_n \| \mathbf{b}\big) \leq \kappa \|(\mathbf{X}^{\text{UOT}})^\top \mathbf{1}_n - \mathbf{b}\|_1^2. \tag{5}$$

Summing up (4) and (5), we have:

$$\begin{aligned} 0 &\leq \mathbf{KL}(\mathbf{X}^{\text{UOT}}\mathbf{1}_n \| \mathbf{a}) + \mathbf{KL}\big((\mathbf{X}^{\text{UOT}})^\top \mathbf{1}_n \| \mathbf{b}\big) \\ &\leq \kappa \Big[\|\mathbf{X}^{\text{UOT}}\mathbf{1}_n - \mathbf{a}\|_1^2 + \|(\mathbf{X}^{\text{UOT}})^\top \mathbf{1}_n - \mathbf{b}\|_1^2 \Big] \\ &\leq \kappa \Big[\|\mathbf{X}^{\text{UOT}}\mathbf{1}_n - \mathbf{a}\|_1 + \|(\mathbf{X}^{\text{UOT}})^\top \mathbf{1}_n - \mathbf{b}\|_1 \Big]^2 \\ &\leq \kappa \Big(\frac{2n\|\mathbf{C}\|_\infty}{\tau} \Big)^2 = \frac{4\kappa n^2 \|\mathbf{C}\|_\infty^2}{\tau^2}, \end{aligned} \tag{6}$$

where the last inequality follows directly from [30, Theorem 23]. Now, from [30, Theorem 26] and given $M = \log(2)\|C\|_\infty^2 \big(n + 3\kappa\big)^2 + 2n\|C\|_\infty^2$, we have:

$$\begin{aligned} 0 \leq \mathbf{OT}(\mathbf{a}, \mathbf{b}) - \mathbf{UOT}(\mathbf{a}, \mathbf{b}) &\leq \frac{M}{\tau} \\ \therefore 0 \leq \langle \mathbf{C}, \mathbf{X}^{\text{OT}} \rangle - \langle \mathbf{C}, \mathbf{X}^{\text{UOT}} \rangle &- \tau \mathbf{KL}(\mathbf{X}^{\text{UOT}}\mathbf{1}_n \| \mathbf{a}) \\ - \tau \mathbf{KL}\big((\mathbf{X}^{\text{UOT}})^\top \mathbf{1}_n \| \mathbf{b}\big) &\leq \frac{M}{\tau}, \end{aligned} \tag{7}$$

where the last line is by definitions of $\mathbf{OT}(\mathbf{a}, \mathbf{b})$ and $\mathbf{UOT}(\mathbf{a}, \mathbf{b})$. Finally, combining (6) and (7), we can conclude the statement of the theorem:

$$0 \leq \langle \mathbf{C}, \mathbf{X}^{\text{OT}} \rangle - \langle \mathbf{C}, \mathbf{X}^{\text{UOT}} \rangle \leq \frac{M + 4\kappa n^2 \|\mathbf{C}\|_\infty^2}{\tau} = O\left(\frac{1}{\tau}\right).$$

■

1) Remark: Thereom 1 provides a bound on the difference between the transportation costs of OT solution and UOT solution under a balanced setting. Specifically, when the marginal distributions $\mathbf{a}$ and $\mathbf{b}$ are probability distributions both summing up to 1, the theorem establishes that the cost difference between the OT and UOT solutions grows inversely with τ, and thus can be made arbitrarily negligible by tuning the hyper-parameter $\tau > 0$ to be sufficiently large. This motivates our usage of UOT in place of OT for computational acceleration while maintaining the original performance of OT via the proper choice of τ in the next sections.

III. Experimental Results of UOT in Breast Cancer Data

Breast cancer is one of the most prevalent cancers worldwide, and accurately distinguishing between benign and malignant cases is crucial for early diagnosis and treatment. Machine learning methods have advanced breast cancer data analysis, but traditional approaches often struggle with computational efficiency and imbalanced data sets.

Optimal transport (OT) is a powerful tool for comparing probability distributions, yet assumes balanced datasets, which is rarely the case in real-world scenarios like breast cancer data. Unbalanced Optimal Transport (UOT) addresses this issue by allowing for differences in data mass, making it well-suited for medical datasets with uneven class distributions.

In [39], the authors proposed the Hungarian algorithm to solve a special type of optimal transport. It showed that Hungarian outperforms Sinkhorn algorithm and network simplex algorithm in all cases.

In this study, we apply UOT to breast cancer data to test for statistical independence between features of benign and malignant cases. By leveraging the efficiency of UOT combined with the Sinkhorn algorithm, we aim to provide a scalable method that outperforms traditional OT in runtime while maintaining accuracy, offering valuable insights for large-scale healthcare data analysis.

A. Problem Setting of UOT in Breast Cancer Data

1) Wasserstein-distance-based independence test and UOT: One crucial application of OT distances such as Wasserstein-1 [34], [20] is the independence test [39]. To assess the independence between the variables $Y \sim \nu_1$ and $Z \sim \nu_2$, the Wasserstein-1 distance with ℓ_p-norm cost function, which belongs to the class of OT problem [39], [20], between the joint distribution π of Y, Z and the product distribution of Y, Z is used. Specifically, this process requires the evaluation of $\mathbf{OT}(\pi, \nu_1 \otimes \nu_2)$, where $\nu_1 \otimes \nu_2$ represents the product distribution of Y, Z. It is proven in [39] that Y and Z are independent if and only if $\mathbf{OT}(\pi, \nu_1 \otimes \nu_2) = 0$. In practice, given n i.i.d. samples $\{(y_1, z_1), ..., (y_n, z_n)\}$ generated from (Y, Z), one can construct the statistic $\mathbf{OT}(\hat{\pi}, \hat{\nu}_1 \otimes \hat{\nu}_2)$, where $\hat{\pi}$ and $\hat{\nu}$ represent the empirical distributions, to test for independence. UOT distance has been known to well approximate OT distance with vanishing approximation error [30, Theorem 26], while enjoying more favorable computational complexity through various solvers vastly used in the ML/AI literature [35], [30], [5]. In this study, we aim to use UOT as an alternative to OT for Wasserstein-distance-based independence testing, and utilizes the celebrated Sinkhorn algorithm [35] to solve for the UOT problem.

2) Breast cancer data: The dataset consists of 569 instances, each characterized by 30 features. The instances are classified into two categories: benign and malignant. Let $X \in \mathbb{R}^{30}$ represent the distribution generated uniformly from the benign class, and $Y \in \mathbb{R}^{30}$ represent the distribution generated uniformly from the malignant class. We compute the empirical OT/UOT distance in two scenarios: 1. Independent case: Between X_1 and Y_2, where X_1 comprises the first 5 coordinates of X, and Y_2 comprises the last 25 coordinates of Y. 2. Dependent case: Between X and Z, where $Z = X_1 * Y_1$, with X_1 being the first 5 coordinates of X, Y_1 being the first 5 coordinates of Y, and $*$ representing the coordinatewise product.

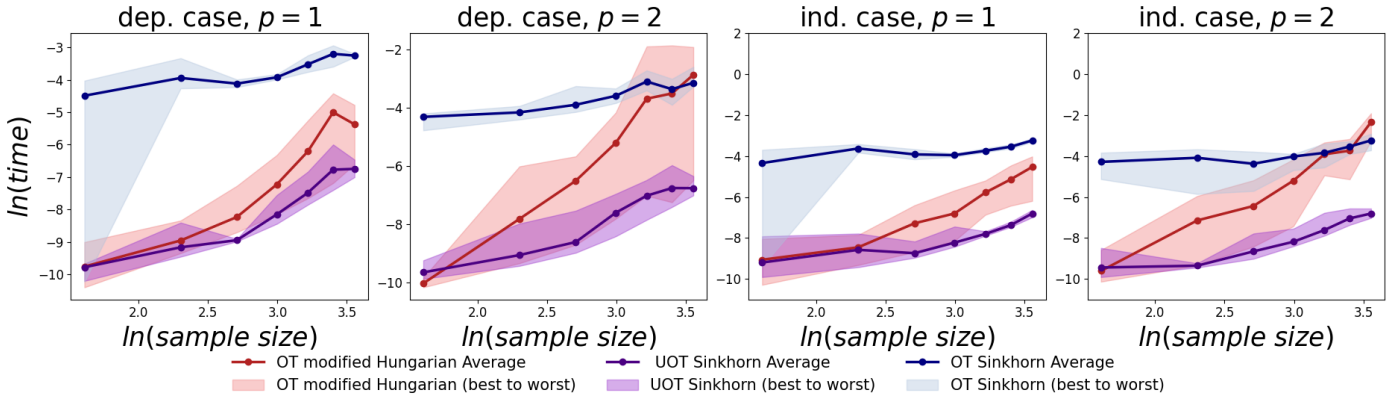

Fig. 1. Runtime evaluation with UOT sinkhorn, OT sinkhorn and modified hungarian algorithm on breast cancer data.

B. Application of UOT in Breast Cancer Data

1) Data preprocessing: In order to separate classes, the second column of the dataset contains the labels that classify each instance as either benign (B) or malignant (M). We use this column to separate the data into two subsets such that rows labeled 'B' and 'M' are extracted to form the X and Y dataset respectively.

From both X and Y, we focus on the feature values found in columns 2 to 32 (the 30 numerical features of each instance). These features are selected because they represent the main characteristics of the data relevant for classification.

The selected feature values are normalized by dividing each value by the maximum value in its respective column. This step rescales all feature components to the range [0, 1].

Normalization ensures that features with different scales do not disproportionately influence the analysis and that the data is suitable for OT/UOT computations.

2) Experimental setup: We follow the experimental setup outlined in [39], where the independence test based on Wasserstein distance is applied. In our experiments, we calculate the cost matrix $\mathbf{C}$ using the ℓ_p-norm as the cost function. Specifically, we evaluate the performance under two different norms: $p = 1$ and $p = 2$. These norms are chosen to assess the behavior of OT and UOT algorithms under different geometries of the data.

For both the dependent and independent cases, we vary the sample sizes of the breast cancer data to examine the effect of sample size ranging from small to large on the computational performance. In each experimental run, we generate data for both cases (dependent and independent) using uniformly distributed samples from the benign and malignant classes. The independence tests are then performed using UOT-based and OT-based methods.

Each experiment is repeated 10 times to account for random variations in the data and solvers. For each sample size and test, we report the following runtime statistics: the average, best, and worst runtimes over the 10 trials. This ensures a robust evaluation of the algorithm's performance and allows us to observe both the typical performance and the variability across runs.

Additionally, to evaluate the statistical significance of the results, we analyze the empirical distribution of the runtimes for each method and apply appropriate tests (e.g., t-tests) to confirm that the differences in runtimes between UOT Sinkhorn and the baseline methods are statistically significant.

3) Baselines: In our comparison, we include several state-of-the-art methods for solving OT problems as baselines. The first baseline is the exact OT solver based on the modified Hungarian algorithm [39]. The second baseline is the OT Sinkhorn algorithm [19], which approximates the OT distance by adding an entropic regularization term to the original OT problem.

Our proposed method uses the UOT Sinkhorn algorithm to approximate the OT distance under the UOT framework. UOT is particularly suitable for handling the unbalanced nature of distributions that may arise in real-world datasets like breast cancer data, where the number of benign and malignant cases might not be perfectly matched. The Sinkhorn solver adds entropic regularization, making it highly efficient for large-scale problems.

To ensure a fair comparison, we set the desired error tolerance for the approximate algorithms (UOT Sinkhorn and OT Sinkhorn) to 0.01. This tolerance provides a good balance between computational efficiency and approximation accuracy.

4) Experimental results: The results of our experiments, shown in Fig. 1, illustrate the runtime performance of the tested algorithms as a function of sample size. The x-axis represents the logarithm of the sample size, while the y-axis represents the logarithm of the total runtime in seconds. These log-log plots provide a clear visualization of the scalability of each algorithm.

For each tested sample size, UOT Sinkhorn consistently outperforms both OT-based baselines in terms of runtime. Specifically, UOT Sinkhorn achieves lower average, best, and worst runtimes across all sample sizes, with significant improvements as the sample size increases. The performance advantage of UOT Sinkhorn becomes especially pronounced for large sample sizes, where the modified Hungarian algorithm exhibits much slower runtimes due to its higher computational complexity.

In terms of robustness, UOT Sinkhorn demonstrates consis-

tent performance, where even its worst runtime remains faster than the average runtime of the second-best baseline, OT solved using the modified Hungarian algorithm. This robustness is a crucial factor for practical applications where runtime variability can impact the reliability of the method. Additionally, OT Sinkhorn performs better than the exact solver for moderate sample sizes but is still outperformed by UOT Sinkhorn in most cases.

Finally, we also observe that the choice of ℓ_p-norm ($p = 1$ or $p = 2$) has a relatively minor effect on the runtime but does influence the accuracy of the independence test, as the distance metrics capture different aspects of the data geometry.

IV. UOT in HRV Estimation for Physiological Research

In physiological research, Heart Rate Variability (HRV) is often used as a measure for its reliability and noninvasiveness [1]. However, assessing cardiovascular functioning using HRV in practice is challenging due to noise and irregularly sampled data.

Previously, [42] proposed a multitask-learning approach to address this issue. However, clinical and healthcare data in practice often have a high degree of heterogeneity (such as in demographics, treatments, devices, etc), which means domain generalization is an essential task. Thus, [42] proposes to use OT to estimate a mapping that is generalizable for unseen out-of-domain task distributions. The multitask model using Optimal Transport as a regularizer showed the lowest RMSE amongst other transport maps such as Group Lasso [40], Multi-level Lasso [22], Dirty models [13], Multitask Wasserstein and Reweighted Multi-task Wasserstein [14].

However, the stringent nature of OT maps may cause strict mapping, which would be problematic under noisy data regimes [2], [17]. Thus, we propose using UOT instead of OT for the domain generalization task.

A. OT/UOT Map Estimation for Physio Multitask-learning

Consider the task-wise feature vectors S_t and their underlying predictive functions W_t^T, each following the measures μ_S and ν_W respectively. Our goal is to find a push forward mapping $T_{\#}\mu_S = \nu_W$. Here, one can estimate μ_S, ν_W from the empirical distributions, and use them as the two input marginal vectors of the OT/UOT problem. The optimal mapping allows us to measure the similarity of model parameters to obtain a predictive transformation, which will eventually be used to perform domain generalization.

B. Multitask-learning (MTL)

We follow the MTL formulation in [42] and use the following optimization objective, where the first term represents the prediction loss and the second term represents the regularization that induces task similarities:

1) Dataset: We are given a set of T tasks, each represented by:

- $X_t \in \mathbb{R}^{d_x \times N_t}$: The feature matrix for task t, where N_t is the number of samples for task t, and each sample has d_x features.
- $Y_t \in \mathbb{R}^{1 \times N_t}$: The labels for the samples in task t.
- $s_t \in \mathbb{R}^{d_s \times 1}$: A task-specific feature vector, which may contain information unique to that task.

2) Objective: We are learning the parameters, represented by the matrix W_t, for each task t, where each W_t is part of a larger matrix $W \in \mathbb{R}^{T \times d_w}$ that contains the weight vectors for all T tasks. The goal is to minimize the loss across all tasks, represented as:

$$\min_{\mathbf{W},\mathbf{F}} \underbrace{\frac{1}{2}\sum_{t=1}^{T} \|W_t^T X_t - y_t\|_2^2}_{\text{Prediction loss}} + \underbrace{\alpha \sum_{i,j=1}^{T} \pi_{i,j}^* \|F(s_i) - W_j\|_2^2}_{\text{Regularization term}}. \quad (1)$$

Here, $\mathbf{W} = [W_1, W_2, \ldots, W_T]$ are the task-specific weights, and $\mathbf{F}$ is the transformation that models the similarities between different tasks, π^* is the OT/UOT coupling obtained from the Sinkhorn-OT/Sinkhorn-UOT algorithm and α is a weighting parameter. The loss (1) will learn $\mathbf{W}$ and $\mathbf{F}$ using Algorithm 1 in [42] where $\mathbf{W}$ and $\mathbf{F}$ are jointly updated using GD.

C. Linear and Non-linear Transformation for $\boldsymbol{F}$

First, we consider $\mathcal{F}$ being a linear transformation, where the set $\mathcal{F}$ is characterized by a matrix $\mathbf{F} \in \mathbb{R}^{d_s \times d_\theta}$, which captures all affine transformations. Mathematically, the set $\mathcal{F}$ is expressed as:

$$\mathcal{F} = \left\{ F : \mathbf{F} \in \mathbb{R}^{d_\theta \times d_s},\ s_t \in \Omega_S,\ F(s_t) = \mathbf{F} s_t \right\}.$$

However, linear transformations may not sufficiently approximate the transport map, particularly for modeling complex systems such as the human Autonomic Nervous System (ANS). To address this, non-linear transformations is considered.

Let ϕ be a non-linear function associated with a kernel function $k : \Omega_S \times \Omega_S \to \mathbb{R}$, where $k(s_i, s_j) = \langle \phi(s_i), \phi(s_j) \rangle$. For a given set of samples S, we define the set $\mathcal{F}$ as:

$$\mathcal{F} = \left\{ F : \mathbf{F} \in \mathbb{R}^{d_\theta \times T},\ s_t \in \Omega_S,\ F(s_t) = \mathbf{F}\mathbf{k}_{s_t}(s_t) \right\},$$

where $\mathbf{k}_{s_t}(\cdot)$ denotes the vector $k(s_1, \cdot), \ldots, k(s_T, \cdot)$.

The non-linear transformation allows the model to capture more intricate relationships between tasks, making it particularly useful when task dependencies are complex and cannot be adequately captured by a linear mapping. The challenge in this formulation lies in the increased complexity of learning $\mathbf{F}$, as the optimization problem becomes non-convex and may require advanced techniques such as back-propagation for training.

Despite the increased computational cost, non-linear transformations can significantly improve performance in multitask learning scenarios where tasks exhibit non-linear similarities, enabling the model to generalize better accros tasks.

D. Gradient Descent for MTL

To apply Gradient Descent (GD) to solve the MTL objective, we first differentiate the loss function with respect to the MTL parameters $\mathbf{W}$ and $\mathbf{F}$ is a non-linear transformation. Given the optimization objective from Eq. (1), the gradients with respect to the MTL parameters $\mathbf{W}$ and $\mathbf{F}$ are:

Gradient with respect to W:

$$\nabla_{\mathbf{W}}\mathcal{L} = (W_j^T X_j - y_j)X_j^T - 2\alpha \sum_i \pi^*_{i,j}(\mathbf{F}\mathbf{k}_{s_i} - W_j).$$

The first term corresponds to the gradient of the prediction loss, while the second term reflects the gradient of the regularization term, weighted by α and the optimal coupling matrix π^* obtained from OT/UOT.

Gradient with respect to F:

$$\nabla_{\mathbf{F}}\mathcal{L} = 2\alpha \sum_{i,j} \pi^*_{i,j}(\mathbf{F}\mathbf{k}_{s_i} - W_j)\mathbf{k}_{s_i}^T.$$

Here, the gradient is driven solely by the regularization term, as **F** does not appear in the prediction loss.

PhysioMTL using gradient descent has a disadvantage when processing complex data such as HRV where it confronts multiple local minimums which will affect learning rate. To address the problem, we introduce the gradient descent with momentum, a more robust version of gradient descent which can potentially handle local minimums and saddle points.

Using the gradients proposed above, we iteratively update the parameters **W** and **F** using the standard GD update rule:

$$\mathbf{W}^{(k+1)} = \mathbf{W}^{(k)} - \eta\nabla_{\mathbf{W}}\mathcal{L}, \quad \mathbf{F}^{(k+1)} = \mathbf{F}^{(k)} - \eta\nabla_{\mathbf{F}}\mathcal{L}$$

where η is the learning rate and k denotes the iteration index.

So, the update rules using gradient descent for $\mathbf{W}_j$ and **F** are:

For $\mathbf{W}_j$:

$$\mathbf{W}_j \leftarrow \mathbf{W}_j - \eta\left[(W_j^T X_j - y_j)X_j^T - 2\alpha \sum_i \pi^*_{i,j}(\mathbf{F}\mathbf{k}_{s_i} - W_j)\right].$$

For F:

$$\mathbf{F} \leftarrow \mathbf{F} - \eta\left[2\alpha \sum_{i,j} \pi^*_{i,j}(\mathbf{F}\mathbf{k}_{s_i} - W_j)\mathbf{k}_{s_i}^T\right].$$

Algorithm 1 Solving MTL: Gradient Descent

Input: η, α, $\{\pi^*_{i,j}\}$, $\{X_t, y_t\}_{t=1}^T$, $\{\mathbf{k}_{s_i}\}$, $\mathbf{W}_j$, **F**
for $n = 1$ to N **do**
  for $j = 1$ to T **do**
    $\mathbf{W}_j \leftarrow \mathbf{W}_j - \eta\nabla_{W_j}$
  end for
  $\nabla_F = 2\alpha \sum_{i,j} \pi^*_{i,j}(\mathbf{F}\mathbf{k}_{s_i} - W_j)\mathbf{k}_{s_i}^T$
  $\mathbf{F} \leftarrow \mathbf{F} - \eta\nabla_F$
end for
Output: $\mathbf{W}_j$, **F**

In practice, choosing an appropriate learning rate η is critical for convergence. A small learning rate can slow down the convergence, while a large one might cause the updates to overshoot the optimal solution. The iterative GD process continues until convergence, which is typically defined by a threshold on the change in the loss function or the norm of the gradient.

E. Gradient Descent with Momentum for MTL

In gradient descent with momentum, we introduce a velocity term to accelerate the optimization process. The update rules are modified to include momentum, where the gradient is accumulated over time.

Let $\mathbf{v}_W$ and $\mathbf{v}_F$ be the velocity terms for $\mathbf{W}_j$ and **F**, respectively. The momentum update is controlled by a parameter $\beta \in [0, 1)$. The update rules for the velocities and parameters are as follows:

Velocity update:

$$\mathbf{v}_W^{(k+1)} = \beta\mathbf{v}_W^{(k)} + (1-\beta)\nabla_{\mathbf{W}}\mathcal{L}, \quad \mathbf{v}_F^{(k+1)} = \beta\mathbf{v}_F^{(k)} + (1-\beta)\nabla_{\mathbf{F}}\mathcal{L}$$

Parameter update:

$$\mathbf{W}^{(k+1)} = \mathbf{W}^{(k)} - \eta\mathbf{v}_W^{(k+1)}, \quad \mathbf{F}^{(k+1)} = \mathbf{F}^{(k)} - \eta\mathbf{v}_F^{(k+1)}$$

Algorithm 2 Solving MTL: Gradient Descent with Momentum

Input: η, β, α, $\{\pi^*_{i,j}\}$, $\{X_t, y_t\}_{t=1}^T$, $\{\mathbf{k}_{s_i}\}$, $\mathbf{W}_j$, **F**
$\mathbf{v}_{W_j} = 0$, $\mathbf{v}_F = 0$
for $n = 1$ to N **do**
  for $j = 1$ to T **do**
    $\mathbf{v}_{W_j} \leftarrow \beta\mathbf{v}_{W_j} + (1-\beta)\nabla_{W_j}$
    $\mathbf{W}_j \leftarrow \mathbf{W}_j - \eta\mathbf{v}_{W_j}$
  end for
  $\nabla_F = 2\alpha \sum_{i,j} \pi^*_{i,j}(\mathbf{F}\mathbf{k}_{s_i} - W_j)\mathbf{k}_{s_i}^T$
  $\mathbf{v}_F \leftarrow \beta\mathbf{v}_F + (1-\beta)\nabla_F$
  $\mathbf{F} \leftarrow \mathbf{F} - \eta\mathbf{v}_F$
end for
Output: $\mathbf{W}_j$, **F**

F. Literature and Motivation for Gradient Descent with Momentum for MTL

Momentum helps to speed up convergence in scenarios where the optimization landscape has long, narrow valleys—common in deep learning and multitask-learning. The velocity terms accumulate the gradients in such valleys, allowing the optimization to move faster along the flat directions and more slowly in directions where the gradients change rapidly.

1) Escaping saddle points: One of the key advantages of GDM is its ability to help the optimization process escape saddle points. Saddle points are regions in the loss surface where the gradient is close to zero but the point is not a local minimum. By incorporating past gradients, GDM can provide sufficient momentum to push the optimization out of these regions, avoiding the problem of getting stuck at suboptimal points. This is particularly important for multitask-learning, where the complex interaction between tasks may introduce non-convexities in the loss surface.

2) Convergence considerations: While GDM generally converges faster than standard GD, careful tuning of both the learning rate η and the momentum parameter β is essential for achieving optimal performance. A typical heuristic is to start with $\beta = 0.9$ and adjust η based on empirical results, ensuring that the updates do not become too aggressive or oscillatory.

In the MTL context, GDM is particularly useful when dealing with heterogeneous tasks, as the added momentum helps

TABLE I. SUMMARY OF MODEL PERFORMANCES WITH $\alpha = 0.1$.UOT-GD AND UOT-GDM SHOW LOWER RMSE AND FASTER RUNTIMES COMPARED TO OT-BASED METHODS, WITH MOMENTUM FURTHER REDUCING RUNTIME

Method	20%		40%		60%		80%	
	RMSE	Runtime	RMSE	Runtime	RMSE	Runtime	RMSE	Runtime
OT-GD	29.992 ± 0.809	0.514	29.890 ± 1.198	1.219	29.504 ± 0.963	3.347	28.800 ± 2.432	6.765
UOT-GD	29.978 ± 0.843	0.287	29.911 ± 1.222	0.534	29.500 ± 0.942	1.032	28.749 ± 2.455	2.072
OT-GDM	30.070 ± 0.749	0.291	29.940 ± 1.206	0.641	29.584 ± 0.990	1.154	28.897 ± 2.394	3.190
UOT-GDM	30.013 ± 0.794	0.244	29.891 ± 1.203	0.372	29.521 ± 0.965	0.874	28.845 ± 2.420	1.603

TABLE II. SUMMARY OF MODEL PERFORMANCES WITH $\alpha = 0.5$. UOT MAINTAINS CONSISTENT RMSE AND FASTER RUNTIMES ACROSS SAMPLE SIZES, WHILE OT METHODS SHOW SLIGHTLY HIGHER RMSE AND SLOWER PERFORMANCE

Method	20%		40%		60%		80%	
	RMSE	Runtime	RMSE	Runtime	RMSE	Runtime	RMSE	Runtime
OT-GD	30.753 ± 1.536	0.290	29.841 ± 0.924	1.480	29.364 ± 1.692	4.448	30.282 ± 2.218	7.974
UOT-GD	30.779 ± 1.549	0.163	29.748 ± 0.967	0.684	29.298 ± 1.703	1.383	30.151 ± 2.259	2.550
OT-GDM	30.769 ± 1.574	0.286	29.934 ± 0.942	1.012	29.461 ± 1.671	2.197	30.459 ± 2.207	3.584
UOT-GDM	30.759 ± 1.529	0.154	29.883 ± 0.926	0.459	29.388 ± 1.696	0.888	30.336 ± 2.218	1.786

TABLE III. SUMMARY OF MODEL PERFORMANCES WITH $\alpha = 0.9$. UOT STILL PERFORMS BETTER THAN OT METHODS IN TERMS OF RMSE AND RUNTIME, WITH MOMENTUM (UOT-GDM) ENSURING THE FASTEST CONVERGENCE, EVEN WITH A LARGER α

Method	20%		40%		60%		80%	
	RMSE	Runtime	RMSE	Runtime	RMSE	Runtime	RMSE	Runtime
OT-GD	30.366 ± 0.812	0.338	30.025 ± 1.370	1.833	30.352 ± 1.240	4.106	30.783 ± 2.393	6.660
UOT-GD	30.291 ± 0.821	0.197	29.996 ± 1.377	0.657	30.232 ± 1.265	1.289	30.718 ± 2.394	2.314
OT-GDM	30.491 ± 0.830	0.274	30.144 ± 1.417	0.896	30.491 ± 1.236	2.049	30.881 ± 2.436	3.584
UOT-GDM	30.395 ± 0.812	0.156	30.046 ± 1.388	0.450	30.399 ± 1.237	0.831	30.823 ± 2.409	1.362

balance the convergence rates across tasks with varying levels of difficulty. The accumulated gradients guide the optimization process toward a more stable solution, reducing the likelihood of overfitting to specific tasks.

By applying GDM, we can achieve a more efficient and reliable solution to the MTL problem, especially in the context of large-scale data or noisy, heterogeneous domains such as HRV estimation.

G. Application of UOT in HRV

1) Data preprocessing: The MMASH dataset [37] contains 24-hour continuous data from 22 healthy male participants, including inter-beat intervals (IBI), wrist accelerometry, sleep duration and quality, physical activity levels, and psychological factors like stress, anxiety, and emotions. HRV is calculated using RMSSD, the root mean square of successive differences between normal heartbeats, over 5-minute intervals, which is the standard duration for short-term HRV analysis. We use key features - activity, sleep, stress, and anthropometric data (age, height, weight). Sleep is expressed as total hours in bed, while physical activity is represented by hours of moderate (e.g. walking, cycling) and intense (e.g., running, gym) exercise. Stress levels are measured via the Daily Stress Inventory (DSI) score.

We following the data preprocessing procedure in [42]: (1) removing RMSSD outliers (z-score greater than 2.5), (2) excluding subjects with abnormal data (e.g. subject 4 with an RMSSD average of 318), and (3) imputing missing values for sleep and age for subjects 11 and 18 using dataset-wide averages. After preprocessing, the final dataset includes 21 subjects.

2) Experimental setup: We applied our model to predict Heart Rate Variability (HRV) across various tasks from the MMASH dataset, which is publicly available through the PhysioNet repository [37]. The tasks used for testing were completely unseen during training to ensure a rigorous evaluation. The performance of the model was measured using Root Mean Square Error (RMSE) to quantify the prediction accuracy. To assess the model's performance under different data availability scenarios, we randomly selected varying proportions of tasks—specifically, 20%, 40%, 60%, and 80%—for training. The model was then evaluated on the remaining unseen tasks.

Additionally, we experimented with different values of $\alpha = 0.1, 0.5, 0.9$ to investigate the robustness of the Sinkhorn-Unbalanced Optimal Transport (Sinkhorn-UOT) model under different mass relaxation parameters. Varying α allows us to explore how the model behaves when placing more or less emphasis on balancing the transport plan, offering insights into the flexibility and adaptability of the method.

3) Baselines: To further improve computational efficiency, we implemented Gradient Descent with Momentum (GDM), which is known to help iterates quickly escape saddle points and accelerate convergence to a stationary point, as suggested by [38]. This technique is particularly valuable for large-scale datasets where faster convergence can significantly reduce runtime. We compared four experimental settings to evaluate both OT and UOT under different optimization strategies: OT with Gradient Descent (OT-GD), OT with Gradient Descent and Momentum (OT-GDM), UOT with Gradient Descent (UOT-GD), and UOT with Gradient Descent and Momentum (UOT-GDM). These baselines allowed us to assess both the impact of mass relaxation in UOT and the computational benefits of incorporating momentum into the optimization process.

4) Experimental results: The results of our experiments are summarized in Tables I, II, and III, corresponding to $\alpha = 0.1$, $\alpha = 0.5$, and $\alpha = 0.9$ respectively. Across all α values, our UOT-based approaches consistently demonstrated significantly lower runtime compared to the OT-based methods, while maintaining similar levels of accuracy as measured by RMSE. This highlights the computational advantage of UOT, particularly for large datasets with imbalanced distributions. Additionally, the inclusion of momentum in the optimization process (GDM) resulted in faster convergence and further reduced runtime compared to standard Gradient Descent (GD), confirming the effectiveness of GDM in accelerating training. These findings underline the practicality of using UOT with momentum for tasks requiring fast and accurate predictions in complex datasets.

5) Results and Discussion: The experimental results summarized in Tables I, II, and III demonstrate the consistent advantages of UOT over OT across all tested values of α (0.1, 0.5, 0.9). While runtime differences were significant—UOT required up to 40% less time than OT—the benefits of UOT extend beyond computational efficiency.

Additionally, the inclusion of momentum in UOT further accelerated convergence while maintaining similar accuracy. This enhancement is particularly valuable in iterative medical research tasks, where faster training enables rapid model updates based on new data.

H. Summary of UOT Algorithms as Compared to OT and Related Practical Healthcare Applications

By relaxing OT's strict constraints, UOT enables faster and more adaptable multitask learning algorithms. Its computational advantages $\tilde{O}\left(n^2\varepsilon^{-1}\right)$ and smoother optimization dynamics make it a competitive alternative to OT, especially in healthcare applications requiring efficient and reliable predictions. Coupled with techniques like GDM, UOT further enhances its utility for large-scale, real-world datasets, demonstrating its practicality in domains where computational resources are limited but accuracy remains paramount.

The advantages of UOT algorithms, particularly their computational efficiency and adaptability, make them highly suitable for resource-constrained healthcare applications:

1) Real-Time HRV Monitoring: Faster convergence of UOT allows real-time heart rate variability predictions in wearable devices, enabling timely interventions in critical scenarios.

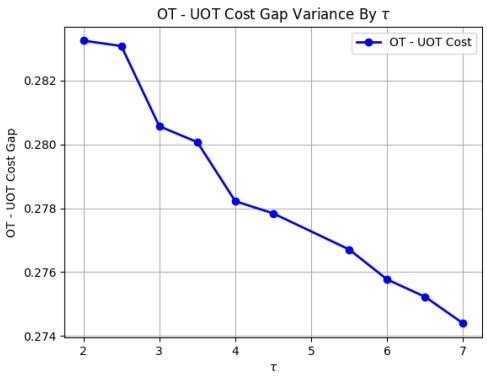

Fig. 2. Cost gap on breast cancer data.

2) Large-Scale Population Studies: UOT's scalability supports applications involving population-level diagnostics, such as analyzing longitudinal health data or predicting patient outcomes across diverse demographics.

3) Personalized Medicine: The flexibility of UOT in handling imbalanced distributions is crucial for personalized medicine, where data from some patient subgroups may be underrepresented. For example, drug response predictions can benefit from UOT's ability to align heterogeneous task distributions effectively. asks, as reflected in the lower Root Mean Square Error (RMSE) observed for UOT.

V. Aprroximation Error

In the context of the breast cancer data experiment, we aim to empirically validate the theorem 1. Specifically, we are interested in investigating how closely the UOT cost approximates the OT cost in real-world datasets, which will be done in the following set up:

1) Computing OT and UOT cost: Using the Sinkhorn algorithm, we compute the OT transport plan $\mathbf{X}^{OT}$ and its corresponding transport cost. Similarly, we compute the UOT transport plan $\mathbf{X}^{UOT}$ for varying values of τ (the regularization parameter) and calculate the associated UOT cost.

2) Comparison and Validation of Approximation Error: We empirically measure this difference to see how well UOT approximates OT in our breast cancer dataset. Specifically, we calculate $\langle \mathbf{C}, \mathbf{X}^{OT} \rangle - \langle \mathbf{C}, \mathbf{X}^{UOT} \rangle$ for different values of τ to observe whether the theoretical bound holds in practice.

3) Experimental result: The results shown in Fig. 2 as τ increases, the cost gap steadily decreases. This demonstrates that with higher regularization, the difference between the UOT and OT solutions becomes negligible. At this point, UOT provides a good approximation to the OT cost while retaining computational advantages. The graph exhibits a smooth, nonlinear decrease in the cost gap, implying that increasing τ provides diminishing returns in terms of cost difference.

VI. Conclusion and Future Works

In this work, we investigate the computational benefits of the Sinkhorn-UOT algorithm across different healthcare applications such as the independence test on breast cancer

data and HRV estimation in physiological research. We find that Sinkhorn-UOT consistently outperforms other popular computational OT methods such as Sinkhorn-OT and the modified Hungarian algorithm, which partially makes various healthcare applications more accessible to budget-constraint medical institutes by alleviating the prohibitive computational cost, and mitigates the CO2 emission from computing resources toward better environment.

Building on these results, future work should focus on further optimizing the Sinkhorn-UOT algorithm by reducing the computational complexity and enhancing scalability for even larger datasets. Another interesting direction is to investigate the effectiveness of Partial Optimal Transport (POT) [24] as an alternative to OT, besides the UOT metric considered in this paper. Furthermore, Stochastic or Constrained Decentralized Optimization techniques [25], [26] can be leveraged to create sample-efficient computational approaches for noisy, dynamic, and multi-agent scenarios [9], [28], [32], [31], [33] that commonly emerge in modern distributed systems [27], [29].

ACKNOWLEDGMENT

This project was supported by VietDynamic and Binh Duong University. In addition, the authors extend profound gratitude to Mr. Quang Minh Nguyen, Mr. Hoang Huy Nguyen, Ms. My Ngoc Tran Le, and Mr. Nhat Minh Phung for their invaluable guidance and supervision throughout this research endeavor.

REFERENCES

[1] U. Acharya, Paul Joseph, Natarajan Kannathal, Choo Lim, and Jasjit Suri. Heart rate variability: A review. *Medical & biological engineering & computing*, 44:1031–51, 01 2007.

[2] Yogesh Balaji, Rama Chellappa, and Soheil Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems*, 33:12934–12944, 2020.

[3] Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and sparse optimal transport, 2018.

[4] Luis A. Caffarelli and Robert J. McCann. Free boundaries in optimal transport and monge-ampère obstacle problems. *Annals of Mathematics*, 171(2):673–730, 2010.

[5] Laetitia Chapel, Rémi Flamary, Haoran Wu, Cédric Févotte, and Gilles Gasso. Unbalanced optimal transport through non-negative penalized linear regression. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[6] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

[7] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017.

[8] Alejandro Deniz-Garcia, Himar Fabelo, Antonio J Rodriguez-Almeida, Garlene Zamora-Zamorano, Maria Castro-Fernandez, Maria Del Pino Alberiche Ruano, Terje Solvoll, Conceição Granja, Thomas Roger Schopf, Gustavo M Callico, Cristina Soguero-Ruiz, Ana M Wägner, and WARIFA Consortium. Quality, usability, and effectiveness of mhealth apps and the role of artificial intelligence: Current scenario and challenges. *J Med Internet Res*, 25:e44030, May 2023.

[9] Minh Ngoc Dinh and Quang Minh Nguyen. Measurements of errors in large-scale computational simulations at runtime. In *2020 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 1–7, 2020.

[10] S.S. Dragomir, M.L. Scholz, and J. Sunde. Some upper bounds for relative entropy and applications. *Computers & Mathematics with Applications*, 39(9):91–100, 2000.

[11] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061, 2015.

[12] Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617, 2018.

[13] Ali Jalali, Sujay Sanghavi, Chao Ruan, and Pradeep Ravikumar. A dirty model for multi-task learning. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

[14] Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Wasserstein regularization for sparse multi-task regression. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1407–1416. PMLR, 16–18 Apr 2019.

[15] L. V. Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.

[16] Loïc Lannelongue, Jason Grealey, Alex Bateman, and Michael Inouye. Ten simple rules to make your computing more environmentally sustainable. *PLoS Comput Biol*, 17(9):e1009324, September 2021.

[17] Khang Le, Huy Nguyen, Quang Nguyen, Tung Pham, Hung Bui, and Nhat Ho. On robust optimal transport: Computational complexity and barycenter computation. arXiv, 2021.

[18] John Lee, Nicholas P Bertrand, and Christopher J Rozell. Parallel unbalanced optimal transport regularization for large scale imaging problems. *arXiv preprint arXiv:1909.00149*, 2019.

[19] Tianyi Lin, Nhat Ho, and Michael I. Jordan. On the efficiency of entropic regularized algorithms for optimal transport. *Journal of Machine Learning Research*, 23(137):1–42, 2022.

[20] Jialin Liu, Wotao Yin, Wuchen Li, and Yat Tin Chow. Multilevel optimal transport: a fast approximation of wasserstein-1 distances, 2019.

[21] Weijie Liu, Chao Zhang, Nenggan Zheng, and Hui Qian. Approximating optimal transport via low-rank and sparse factorization. *ArXiv*, abs/2111.06546, 2021.

[22] Aurélie C. Lozano and Grzegorz Swirszcz. Multi-level lasso for sparse multi-task regression. In *International Conference on Machine Learning*, 2012.

[23] G Monge. *Mémoire sur la théorie des déblais et des remblais*. 1781.

[24] Anh Duc Nguyen, Tuan Dung Nguyen, Quang Minh Nguyen, Hoang H. Nguyen, Lam M. Nguyen, and Kim-Chuan Toh. On partial optimal transport: Revising the infeasibility of sinkhorn and efficient gradient methods, 2023.

[25] Hoang Huy Nguyen, Yan Li, and Tuo Zhao. Stochastic constrained decentralized optimization for machine learning with fewer data oracles: a gradient sliding approach, 2024.

[26] Hoang Huy Nguyen and Siva Theja Maguluri. Stochastic Approximation for Nonlinear Discrete Stochastic Control: Finite-Sample Bounds. 2023.

[27] Minh Nguyen, Dumitrel Loghin, and Tien Tuan Anh Dinh. Understanding the scalability of hyperledger fabric. *ArXiv*, abs/2107.09886, 2021.

[28] Quang Minh Nguyen, Haewon Jeong, and Pulkit Grover. Coded qr decomposition. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 191–196, 2020.

[29] Quang Minh Nguyen, Nhan Khanh Le, and Lam M. Nguyen. Scalable and secure federated xgboost. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[30] Quang Minh Nguyen, Hoang H. Nguyen, Yi Zhou, and Lam M. Nguyen. On unbalanced optimal transport: Gradient methods, sparsity and approximation error. *Journal of Machine Learning Research*, 24(384):1–41, 2023.

[31] Quang Minh Nguyen, Lam M. Nguyen, and Subhro Das. Correlated attention in transformers for multivariate time series, 2023.

[32] Quang Minh Nguyen, Iain Weissburg, and Haewon Jeong. Coded computing for fault-tolerant parallel qr decomposition, 2023.

[33] Khuong Nguyen-Vinh, Quang-Nguyen Vo-Huynh, Khoa Nguyen-Minh, Minh Hoang, and Surender Rangaraju. *Case Study: Utilising of Deep Learning Models for Fault Detection and Diagnosis of Photovoltaic Modules to Improve Solar Energy Constructions' O&M Activities Quality*, pages 53–67. Springer Nature Singapore, Singapore, 2024.

[34] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[35] K. Pham, K. Le, N. Ho, T. Pham, and H. Bui. On unbalanced optimal transport: An analysis of sinkhorn algorithm. In *ICML*, 2020.

[36] Claudia A Rhoades, Brian E Whitacre, and Alison F Davis. Higher electronic health record functionality is associated with lower operating costs in urban-but not Rural-Hospitals. *Appl Clin Inform*, 13(3):665–676, July 2022.

[37] Alessio Rossi, Eleonora Da Pozzo, Dario Menicagli, Chiara Tremolanti, Corrado Priami, Alina Sîrbu, David A. Clifton, Claudia Martini, and Davide Morelli. A public dataset of 24-h multi-levels psycho-physiological responses in young healthy adults. *Data*, 5(4), 2020.

[38] Jun-Kun Wang, Chi-Heng Lin, and Jacob Abernethy. Escaping saddle points faster with stochastic momentum, 2021.

[39] Yiling Xie, Yiling Luo, and Xiaoming Huo. Solving a special type of optimal transport problem by a modified hungarian algorithm, 2023.

[40] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 12 2005.

[41] Kevin Zhang, Junhao Zhu, Dehan Kong, and Zhaolei Zhang. Modeling single cell trajectory using forward-backward stochastic differential equations. *PLoS Comput Biol*, 20(4):e1012015, April 2024.

[42] Jiacheng Zhu, Gregory Darnell, Agni Kumar, Ding Zhao, Bo Li, Xuanlong Nguyen, and Shirley You Ren. Physiomtl: Personalizing physiological patterns using optimal transport multi-task regression. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 354–374. PMLR, 07–08 Apr 2022.