Feature-Robust Optimal Transport for High-Dimensional Data

Mathis Petrovich[1,⋆], Chao Liang[2,⋆], Ryoma Sato[3], Yanbin Liu[4],
Yao-Hung Hubert Tsai[5], Linchao Zhu[4], Yi Yang[2],
Ruslan Salakhutdinov[5], and Makoto Yamada[3,Γ]

[1] École normale supérieure Paris-Saclay
[2] Zhejiang University
[3] Kyoto University, RIKEN AIP
[4] University of Technology Sydney
[5] Carnegie Mellon University

Abstract. Optimal transport is a machine learning problem with applications including distribution comparison, feature selection, and generative adversarial networks. In this paper, we propose feature-robust optimal transport (FROT) for high-dimensional data, which solves high-dimensional OT problems using feature selection to avoid the curse of dimensionality. Specifically, we find a transport plan with discriminative features. To this end, we formulate the FROT problem as a min–max optimization problem. We then propose a convex formulation of the FROT problem and solve it using a Frank–Wolfe-based optimization algorithm, whereby the subproblem can be efficiently solved using the Sinkhorn algorithm. Since FROT finds the transport plan from selected features, it is robust to noise features. To show the effectiveness of FROT, we propose using the FROT algorithm for the layer selection problem in deep neural networks for semantic correspondence. By conducting synthetic and benchmark experiments, we demonstrate that the proposed method can find a strong correspondence by determining important layers. We show that the FROT algorithm achieves state-of-the-art performance in real-world semantic correspondence datasets. Code can be found at `https://github.com/Mathux/FROT`

Keywords: Optimal transport · Feature selection

1 Introduction

Optimal transport (OT) is a machine learning problem with several applications in the computer vision and natural language processing communities. The applications include Wasserstein distance estimation (Peyré et al., 2019), domain adaptation (Yan et al., 2018), multitask learning (Janati et al., 2019), barycenter estimation (Cuturi and Doucet, 2014), semantic correspondence (Liu et al., 2020), feature matching (Sarlin et al., 2020), and photo album summarization

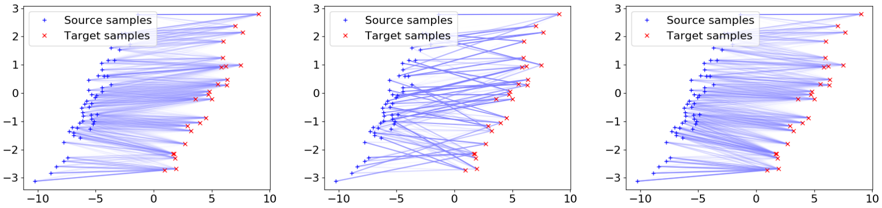
[⋆] The first two authors contributed equally

(Liu et al., 2021). The OT problem is extensively studied in the computer vision community as the earth mover's distance (EMD) (Rubner et al., 2000). However, the computational cost of EMD is cubic and highly expensive. Recently, the entropic regularized EMD problem was proposed; this problem can be solved using the Sinkhorn algorithm with a quadratic cost (Cuturi, 2013). Owing to the development of the Sinkhorn algorithm, researchers have replaced the EMD computation with its regularized counterparts. However, the optimal transport problem for high-dimensional data has remained unsolved for many years.

Recently, a robust variant of the OT was proposed for high-dimensional OT problems and used for divergence estimation (Paty and Cuturi, 2019, 2020). In the robust OT framework, the transport plan is computed with the discriminative subspace of the two data matrices $\boldsymbol{X} \in \mathbb{R}^{d \times n}$ and $\boldsymbol{Y} \in \mathbb{R}^{d \times m}$. The subspace can be obtained using dimensionality reduction. An advantage of the subspace robust approach is that it does not require prior information about the subspace. However, given prior information such as feature groups, we can consider a computationally efficient formulation. The computation of the subspace can be expensive if the dimensionality of data is high (e.g., 10^4).

One of the most common prior information items is a feature group. The use of group features is popular in feature selection problems in the biomedical domain and has been extensively studied in Group Lasso (Yuan and Lin, 2006). The key idea of Group Lasso is to prespecify the group variables and select the set of group variables using the group norm (also known as the sum of ℓ_2 norms). For example, if we use a pretrained neural network as a feature extractor and compute OT using the features, then we require careful selection of important layers to compute OT. Specifically, each layer output is regarded as a grouped input. Therefore, using a feature group as prior information is a natural setup and is important for considering OT for deep neural networks (DNNs).

In this paper, we propose a high-dimensional optimal transport method by utilizing prior information in the form of grouped features. Specifically, we propose a feature-robust optimal transport (FROT) problem, for which we select distinct group feature sets to estimate a transport plan instead of determining its distinct subsets, as proposed in (Paty and Cuturi, 2019, 2020). To this end, we formulate the FROT problem as a min–max optimization problem and transform it into a convex optimization problem, which can be accurately solved using the Frank–Wolfe algorithm (Frank and Wolfe, 1956; Jaggi, 2013). The FROT's subproblem can be efficiently solved using the Sinkhorn algorithm (Cuturi, 2013). An advantage of FROT is that it can yield a transport plan from high-dimensional data using feature selection, using which the significance of the features is obtained without any additional cost. Therefore, the FROT formulation is highly suited for high-dimensional OT problems. Moreover, we show the connection between FROT and the L1 regularized OT problem; this result supports the ability of FROT to select features and robustness of FROT. Through synthetic experiments, we initially demonstrate that the proposed FROT is robust to noise dimensions (See Figure 1). Furthermore, we apply FROT to a



(a) OT on clean data. (b) OT on noisy data. (c) FROT on noisy data ($\eta = 1$).

Fig. 1: Transport plans between two synthetic distributions with 10-dimensional vectors $\widetilde{\boldsymbol{x}} = (\boldsymbol{x}^\top, \boldsymbol{z}_x^\top)^\top$, $\widetilde{\boldsymbol{y}} = (\boldsymbol{y}^\top, \boldsymbol{z}_y^\top)^\top$, where two-dimensional vectors $\boldsymbol{x} \sim N(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ and $\boldsymbol{y} \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$ are true features; and $\boldsymbol{z}_x \sim N(\boldsymbol{0}_8, \boldsymbol{I}_8)$ and $\boldsymbol{z}_y \sim N(\boldsymbol{0}_8, \boldsymbol{I}_8)$ are noisy features. (a) OT between distribution $\boldsymbol{x}$ and $\boldsymbol{y}$ is a reference. (b) OT between distribution $\widetilde{\boldsymbol{x}}$ and $\widetilde{\boldsymbol{y}}$. (c) FROT transport plan between distribution $\widetilde{\boldsymbol{x}}$ and $\widetilde{\boldsymbol{y}}$ where true features and noisy features are grouped, respectively.

semantic correspondence problem (Liu et al., 2020) and show that the proposed algorithm achieves SOTA performance.
Contribution:

- We propose a feature robust optimal transport (FROT) problem and derive a simple and efficient Frank–Wolfe based algorithm. Furthermore, we propose a feature-robust Wasserstein distance (FRWD).
- We show the connection between FROT and the L1 regularized OT problem; this result supports the ability of FROT to select features and robustness of FROT.
- We apply FROT to a high-dimensional feature selection problem and show that FROT is consistent with the Wasserstein distance-based feature selection algorithm with less computational cost than the original algorithm.
- We used FROT for the layer selection problem in a semantic correspondence problem and showed that the proposed algorithm outperforms existing baseline algorithms.

2 Background

In this section, we briefly introduce the OT problem.
Optimal transport (OT): The following are given: independent and identically distributed (i.i.d.) samples $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n \in \mathbb{R}^{d \times n}$ from a d-dimensional distribution p, and i.i.d. samples $\boldsymbol{Y} = \{\boldsymbol{y}_j\}_{j=1}^m \in \mathbb{R}^{d \times m}$ from the d-dimensional distribution q. In the Kantorovich relaxation of OT, admissible couplings are defined by the set of the transport plan:

$$\boldsymbol{U}(\mu, \nu) = \{\boldsymbol{\Pi} \in \mathbb{R}_+^{n \times m} : \boldsymbol{\Pi}\boldsymbol{1}_m = \boldsymbol{a}, \boldsymbol{\Pi}^\top \boldsymbol{1}_n = \boldsymbol{b}\},$$

where $\boldsymbol{\Pi} \in \mathbb{R}_+^{n \times m}$ is called the transport plan, $\mathbf{1}_n$ is the n-dimensional vector whose elements are ones, and $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)^\top \in \mathbb{R}_+^n$ and $\boldsymbol{b} = (b_1, b_2, \ldots, b_m)^\top \in \mathbb{R}_+^m$ are the weights. The OT problem between two discrete measures $\mu = \sum_{i=1}^n a_i \delta_{\boldsymbol{x}_i}$ and $\nu = \sum_{j=1}^m b_j \delta_{\boldsymbol{y}_j}$ determines the optimal transport plan of the following problem:

$$\min_{\boldsymbol{\Pi} \in \boldsymbol{U}(\mu,\nu)} \quad \sum_{i=1}^n \sum_{j=1}^m \pi_{ij} c(\boldsymbol{x}_i, \boldsymbol{y}_j), \tag{1}$$

where $c(\boldsymbol{x}, \boldsymbol{y})$ is a cost function. For example, the squared Euclidean distance is used, that is, $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$. To solve the OT problem, Eq. (1) (also known as the earth mover's distance) using linear programming requires $O(n^3)$, $(n = m)$ computation, which is computationally expensive. To address this, an entropic-regularized optimal transport is used (Cuturi, 2013).

$$\min_{\boldsymbol{\Pi} \in \boldsymbol{U}(\mu,\nu)} \quad \sum_{i=1}^n \sum_{j=1}^m \pi_{ij} c(\boldsymbol{x}_i, \boldsymbol{y}_j) - \epsilon H(\boldsymbol{\Pi}),$$

where $\epsilon \geq 0$ is the regularization parameter, and $H(\boldsymbol{\Pi}) = -\sum_{i=1}^n \sum_{j=1}^m \pi_{ij}(\log(\pi_{ij}) - 1)$ is the entropic regularization. If $\epsilon = 0$, then the regularized OT problem reduces to the EMD problem. Owing to entropic regularization, the entropic regularized OT problem can be accurately solved using Sinkhorn iteration (Cuturi, 2013) with a $O(nm)$ computational cost (See Algorithm 2 in the supplementary material.).

Wasserstein distance: If the cost function is defined as $c(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{x}, \boldsymbol{y})$ with $d(\boldsymbol{x}, \boldsymbol{y})$ as a distance function and $p \geq 1$, then we define the p-Wasserstein distance of two discrete measures $\mu = \sum_{i=1}^n a_i \delta_{\boldsymbol{x}_i}$ and $\nu = \sum_{j=1}^m b_j \delta_{\boldsymbol{y}_j}$ as

$$W_p(\mu, \nu) = \left(\min_{\boldsymbol{\Pi} \in \boldsymbol{U}(\mu,\nu)} \sum_{i=1}^n \sum_{j=1}^m \pi_{ij} d(\boldsymbol{x}_i, \boldsymbol{y}_j)^p \right)^{1/p}.$$

Recently, a robust variant of the Wasserstein distance, called the subspace robust Wasserstein distance (SRW), was proposed (Paty and Cuturi, 2019). The SRW computes the OT problem in the discriminative subspace. This can be determined by solving dimensionality-reduction problems. Owing to the robustness, it can compute the Wasserstein from noisy data. The SRW is given as

$$\mathrm{SRW}(\mu, \nu) = \left(\min_{\boldsymbol{\Pi} \in \boldsymbol{U}(\mu,\nu)} \max_{\boldsymbol{U} \in \mathbb{R}^{d \times k}} \sum_{i=1}^n \sum_{j=1}^m \pi_{ij} \|\boldsymbol{U}^\top \boldsymbol{x}_i - \boldsymbol{U}^\top \boldsymbol{y}_j\|_2^2 \right)^{\frac{1}{2}},$$

where $\boldsymbol{U}$ is the orthonormal matrix with $k \leq d$, and $\boldsymbol{I}_k \in \mathbb{R}^{k \times k}$ is the identity matrix. The SRW or its relaxed problem can be efficiently estimated using either eigenvalue decomposition or the Frank–Wolfe algorithm.

3 Proposed Method

This paper proposes FROT. We assume that the vectors are grouped as $\boldsymbol{x} = (\boldsymbol{x}^{(1)^\top}, \ldots, \boldsymbol{x}^{(L)^\top})^\top$ and $\boldsymbol{y} = (\boldsymbol{y}^{(1)^\top}, \ldots, \boldsymbol{y}^{(L)^\top})^\top$. Here, $\boldsymbol{x}^{(\ell)} \in \mathbb{R}^{d_\ell}$ and $\boldsymbol{y}^{(\ell)} \in \mathbb{R}^{d_\ell}$ are the d_ℓ dimensional vectors, where $\sum_{\ell=1}^{L} d_\ell = d$. This setting is useful if we know the explicit group structure for the feature vectors a priori. In an application in L-layer neural networks, we consider $\boldsymbol{x}^{(\ell)}$ and $\boldsymbol{y}^{(\ell)}$ as outputs of the ℓth layer of the network. If we do not have a priori information, we can consider each feature independently (i.e., $d_1 = d_2 = \ldots = d_L = 1$ and $L = d$). All proofs in this section are provided in the the supplementary material.

3.1 Feature-Robust Optimal Transport (FROT)

The FROT formulation is given by

$$\mathrm{FROT}(\mu, \nu) = \min_{\boldsymbol{\Pi} \in \boldsymbol{U}(\mu,\nu)} \max_{\boldsymbol{\alpha} \in \boldsymbol{\Sigma}^L} \sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{ij} \sum_{\ell=1}^{L} \alpha_\ell c(\boldsymbol{x}_i^{(\ell)}, \boldsymbol{y}_j^{(\ell)}), \tag{2}$$

where $\boldsymbol{\Sigma}^L = \{\boldsymbol{\alpha} \in \mathbb{R}_+^L : \boldsymbol{\alpha}^\top \mathbf{1}_L = 1\}$ is the probability simplex. The underlying concept of FROT is to estimate the transport plan $\boldsymbol{\Pi}$ using distinct groups with large distances between $\{\boldsymbol{x}_i^{(\ell)}\}_{i=1}^n$ and $\{\boldsymbol{y}_j^{(\ell)}\}_{j=1}^m$. We note that determining the transport plan in nondistinct groups is difficult because the data samples in $\{\boldsymbol{x}_i^{(\ell)}\}_{i=1}^n$ and $\{\boldsymbol{y}_j^{(\ell)}\}_{j=1}^m$ overlap. By contrast, in distinct groups, $\{\boldsymbol{x}_i^{(\ell)}\}_{i=1}^n$ and $\{\boldsymbol{y}_j^{(\ell)}\}_{j=1}^m$ are different, and this aids in determining an optimal transport plan. This is an intrinsically similar idea to the subspace robust Wasserstein distance (Paty and Cuturi, 2019), which estimates the transport plan in the discriminative subspace, while our approach selects important groups. Therefore, FROT can be regarded as a feature selection variant of the vanilla OT problem in Eq. (1), whereas the subspace robust version uses dimensionality-reduction counterparts.

Using FROT, we can define a p-feature robust Wasserstein distance (p-FRWD).

Proposition 1 *For the distance function $d(\boldsymbol{x}, \boldsymbol{y})$,*

$$\mathrm{FRWD}_p(\mu, \nu) = \left(\min_{\boldsymbol{\Pi} \in \boldsymbol{U}(\mu,\nu)} \max_{\boldsymbol{\alpha} \in \boldsymbol{\Sigma}^L} \sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{ij} \sum_{\ell=1}^{L} \alpha_\ell d(\boldsymbol{x}_i^{(\ell)}, \boldsymbol{y}_j^{(\ell)})^p \right)^{1/p}, \tag{3}$$

is a distance for $p \geq 1$.

Note that we can show that FRWD_2 is a special case of SRW with $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2$ (See the supplementary material). Another difference between SRW and FRWD is that FRWD can use any distance, while SRW can only use $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2$. Moreover, FRWD_p can be regarded as a special case of the min-max optimal transport (Dhouib et al., 2020). A contribution of this paper is first to introduce feature selection using min-max optimal transport.

3.2 FROT Optimization

Here, we propose two FROT algorithms based on the Frank–Wolfe algorithm.
Frank–Wolfe: We propose a continuous variant of the FROT algorithm using the Frank–Wolfe algorithm, which is fully differentiable. To this end, we introduce entropic regularization for $\boldsymbol{\alpha}$ and rewrite the FROT as a function of $\boldsymbol{\Pi}$:

$$\min_{\boldsymbol{\Pi} \in \boldsymbol{U}(\mu,\nu)} \max_{\boldsymbol{\alpha} \in \boldsymbol{\Sigma}^L} \quad J_\eta(\boldsymbol{\Pi}, \boldsymbol{\alpha}),$$

$$\text{with } J_\eta(\boldsymbol{\Pi}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{ij} \sum_{\ell=1}^{L} \alpha_\ell c(\boldsymbol{x}_i^{(\ell)}, \boldsymbol{y}_j^{(\ell)}) - \eta H(\boldsymbol{\alpha}),$$

where $\eta \geq 0$ is the regularization parameter, and $H(\boldsymbol{\alpha}) = \sum_{\ell=1}^{L} \alpha_\ell(\log(\alpha_\ell) - 1)$ is the entropic regularization for $\boldsymbol{\alpha}$. An advantage of entropic regularization is that the nonnegative constraint is naturally satisfied, and the entropic regularizer is a strong convex function.

Lemma 2 *The optimal solution of the optimization problem*

$$\boldsymbol{\alpha}^* = \operatorname*{argmax}_{\boldsymbol{\alpha} \in \boldsymbol{\Sigma}^L} \quad J_\eta(\boldsymbol{\Pi}, \boldsymbol{\alpha}), \text{with } J_\eta(\boldsymbol{\Pi}, \boldsymbol{\alpha}) = \sum_{\ell=1}^{L} \alpha_\ell \phi_\ell - \eta H(\boldsymbol{\alpha})$$

with a fixed admissible transport plan $\boldsymbol{\Pi} \in \boldsymbol{U}(\mu, \nu)$*, is given by*

$$\alpha_\ell^* = \frac{\exp\left(\frac{1}{\eta}\phi_\ell\right)}{\sum_{\ell'=1}^{L} \exp\left(\frac{1}{\eta}\phi_{\ell'}\right)}, \text{with } J_\eta(\boldsymbol{\Pi}, \boldsymbol{\alpha}^*) = \eta \log\left(\sum_{\ell=1}^{L} \exp\left(\frac{1}{\eta}\phi_\ell\right)\right) + \eta.$$

Using Lemma 2 (or Lemma 4 in Nesterov (2005)) with the setting $\phi_\ell = \sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{ij} c(\boldsymbol{x}_i^{(\ell)}, \boldsymbol{y}_i^{(\ell)}) = \langle \boldsymbol{\Pi}, \boldsymbol{C}_\ell \rangle$, $[\boldsymbol{C}_\ell]_{ij} = c(\boldsymbol{x}_i^{(\ell)}, \boldsymbol{y}_i^{(\ell)})$, the global problem is equivalent to

$$\min_{\boldsymbol{\Pi} \in \boldsymbol{U}(\mu,\nu)} G_\eta(\boldsymbol{\Pi}) = \eta \log\left(\sum_{\ell=1}^{L} \exp\left(\frac{1}{\eta}\langle \boldsymbol{\Pi}, \boldsymbol{C}_\ell \rangle\right)\right). \tag{4}$$

Note that this is known as a smoothed max-operator (Nesterov, 2005; Blondel et al., 2018). The regularization parameter η controls the "smoothness" of the maximum. Moreover, α_ℓ^* becomes an one-hot vector if η is small; we select only one feature if we set $\eta = 0$. In contrast, thanks to the entropic regularization, α_ℓ^* takes non-zero values and we can select multiple features using α_ℓ^*.

Proposition 3 $G_\eta(\boldsymbol{\Pi})$ *is a convex function relative to* $\boldsymbol{\Pi}$*.*

The derived optimization problem of FROT is convex. Therefore, we can determine globally optimal solutions. Note that the SRW optimization problem is not jointly convex (Paty and Cuturi, 2019) for the projection matrix and the

Algorithm 1 FROT with the Frank–Wolfe.

```
Input: $\{\boldsymbol{x}_i\}_{i=1}^n$, $\{\boldsymbol{y}_j\}_{j=1}^m$, $\eta$, and $\epsilon$.
Initialize $\boldsymbol{\Pi}$, compute $\{\boldsymbol{C}_\ell\}_{\ell=1}^L$.
for $t = 0 \ldots T$ do
    $\widehat{\boldsymbol{\Pi}} = \text{argmin}_{\boldsymbol{\Pi}\in\boldsymbol{U}(\mu,\nu)}\langle\boldsymbol{\Pi}, \boldsymbol{M}_{\boldsymbol{\Pi}^{(t)}}\rangle + \epsilon H(\boldsymbol{\Pi})$
    $\boldsymbol{\Pi}^{(t+1)} = (1-\gamma)\boldsymbol{\Pi}^{(t)} + \gamma\widehat{\boldsymbol{\Pi}}$
    with $\gamma = \frac{2}{2+t}$.
end for
return $\boldsymbol{\Pi}^{(T)}$
```

transport plan. In this study, we employ the Frank–Wolfe algorithm (Frank and Wolfe, 1956; Jaggi, 2013), using which we approximate $G_\eta(\boldsymbol{\Pi})$ with linear functions at $\boldsymbol{\Pi}^{(t)}$ and move $\boldsymbol{\Pi}$ toward the optimal solution in the convex set (See Algorithm 1).

The derivative of $G_\eta(\boldsymbol{\Pi})$ at $\boldsymbol{\Pi}^{(t)}$ is given by

$$\frac{\partial G_\eta(\boldsymbol{\Pi})}{\partial \boldsymbol{\Pi}}\bigg|_{\boldsymbol{\Pi}=\boldsymbol{\Pi}^{(t)}} = \sum_{\ell=1}^{L} \alpha_\ell^{(t)} \boldsymbol{C}_\ell = \boldsymbol{M}_{\boldsymbol{\Pi}^{(t)}}, \text{with } \alpha_\ell^{(t)} = \frac{\exp\left(\frac{1}{\eta}\langle\boldsymbol{\Pi}^{(t)}, \boldsymbol{C}_\ell\rangle\right)}{\sum_{\ell'=1}^{L}\exp\left(\frac{1}{\eta}\langle\boldsymbol{\Pi}^{(t)}, \boldsymbol{C}_{\ell'}\rangle\right)}.$$

Then, we update the transport plan by solving the EMD problem:

$$\boldsymbol{\Pi}^{(t+1)} = (1-\gamma)\boldsymbol{\Pi}^{(t)} + \gamma\widehat{\boldsymbol{\Pi}}, \text{ with } \widehat{\boldsymbol{\Pi}} = \underset{\boldsymbol{\Pi}\in\boldsymbol{U}(\mu,\nu)}{\text{argmin}} \langle\boldsymbol{\Pi}, \boldsymbol{M}_{\boldsymbol{\Pi}^{(t)}}\rangle,$$

where $\gamma = 2/(2+k)$. Note that $\boldsymbol{M}_{\boldsymbol{\Pi}^{(t)}}$ is given by the weighted sum of the cost matrices. Thus, we can utilize multiple features to estimate the transport plan $\boldsymbol{\Pi}$ for the relaxed problem in Eq. (4).

Using the Frank–Wolfe algorithm, we can obtain the optimal solution. However, solving the EMD problem requires a cubic computational cost that can be expensive if n and m are large. To address this, we can solve the regularized OT problem, which requires $O(nm)$. We denote the Frank–Wolfe algorithm with EMD as FW-EMD and the Frank–Wolfe algorithm with Sinkhorn as FW-Sinkhorn.

Computational complexity: The proposed method depends on the Sinkhorn algorithm, which requires an $O(nm)$ operation. The computation of the cost matrix in each subproblem needs an $O(Lnm)$ operation, where L is the number of groups. Therefore, the entire complexity is $O(TLnm)$, where T is the number of Frank–Wolfe iterations (in general, $T = 10$ is sufficient).

Proposition 4 *For each $t \geq 1$, the iteration $\boldsymbol{\Pi}^{(t)}$ of Algorithm 1 satisfies*

$$G_\eta(\boldsymbol{\Pi}^{(t)}) - G_\eta(\boldsymbol{\Pi}^*) \leq \frac{4\sigma_{max}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})}{\eta(t+2)}(1+\delta),$$

where $\sigma_{max}(\boldsymbol{\Phi}^\top \boldsymbol{\Phi})$ *is the largest eigenvalue of the matrix* $\boldsymbol{\Phi}^\top \boldsymbol{\Phi}$ *and* $\boldsymbol{\Phi} = (\mathrm{vec}(\boldsymbol{C}_1), \mathrm{vec}(\boldsymbol{C}_2), \ldots, \mathrm{vec}(\boldsymbol{C}_L))^\top$*; and* $\delta \geq 0$ *is the accuracy to which internal linear subproblems are solved.*

Based on Proposition 4, the number of iterations depends on η, ϵ, and the number of groups. If we set a small η, convergence requires more time. In addition, if we use entropic regularization with a large ϵ, the δ in Proposition 4 can be large. Finally, if we use more groups, the largest eigenvalue of the matrix $\boldsymbol{\Phi}^\top \boldsymbol{\Phi}$ can be larger. Note that the constant term of the upper bound is large; however, the Frank–Wolfe algorithm converges quickly in practice.

3.3 Connection to L1 regularization

A natural way to select features is to introduce an L1 regularization term for the feature coefficient $\boldsymbol{\alpha}$. We prove the set of features selected by L1-regularized optimal transport is the same as that of FROT. Let the standard optimal transport with feature coefficient $\boldsymbol{\alpha}$ be:

$$\mathrm{OT}(\mu, \nu, \boldsymbol{\alpha}) = \min_{\boldsymbol{\Pi} \in \boldsymbol{U}(\mu,\nu)} \sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{ij} \sum_{\ell=1}^{L} \alpha_\ell c(\boldsymbol{x}_i^{(\ell)}, \boldsymbol{y}_j^{(\ell)}).$$

Then, the L1-regularized optimal transport is defined as follows:

$$\mathrm{L1OT}(\mu, \nu) = \max_{\boldsymbol{\alpha} \in \mathbb{R}^L_{\geq 0}} \mathrm{OT}(\mu, \nu, \boldsymbol{\alpha}) - \lambda \|\boldsymbol{\alpha}\|_1. \quad (5)$$

Note that the regularization is negative because this is a maximization problem. We assume that $\lambda \geq \mathrm{FROT}(\mu, \nu)$ because otherwise L1OT diverges. Let $\mathcal{F}_{\mathrm{L1}}$ be the set of features that the L1 regularization selects. We consider a feature is selected if the corresponding coefficient can take a positive value in the optimal solution. Specifically, $\mathcal{F}_{\mathrm{L1}}$ is the set of indices $f \in \{1, \cdots, L\}$ such that there exists $\boldsymbol{\alpha}$ such that $\alpha_f > 0$ and $\boldsymbol{\alpha}$ takes the optimum value in Eq. (5). Similarly, let $\mathcal{F}_{\mathrm{FROT}}$ be the set of selected features by FROT. To be precise, $\mathcal{F}_{\mathrm{FROT}}$ is the set of indices $f \in \{1, \cdots, L\}$ such that there exists $\boldsymbol{\Pi}$ and α such that $\alpha_f > 0$ and $(\boldsymbol{\Pi}, \alpha)$ takes the optimum value in Eq. (2).

Theorem 5 $\mathcal{F}_{\mathrm{FROT}} = \mathcal{F}_{L1}$ *when* $\lambda = \mathrm{FROT}(\mu, \nu)$*.*

In other words, FROT and L1 regularization select the same set of features. This result supports the ability of FROT to select features and robustness of FROT.

3.4 Connection to Subspace Robust Wasserstein

Here, we show that 2-FRWD with $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2$ is a special case of SRW. Let us define $\boldsymbol{U} = (\sqrt{\alpha_1}\boldsymbol{e}_1, \sqrt{\alpha_2}\boldsymbol{e}_2, \ldots, \sqrt{\alpha_d}\boldsymbol{e}_d)^\top \in \mathbb{R}^{d \times d}$, where $\boldsymbol{e}_\ell \in \mathbb{R}^d$ is the one-hot vector whose ℓth element is 1 and $\boldsymbol{\alpha}^\top \boldsymbol{1} = 1, \alpha_\ell \geq 0$. Then, the objective function of SRW can be written as

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{m}\pi_{ij}\|\boldsymbol{U}^\top \boldsymbol{x}_i - \boldsymbol{U}^\top \boldsymbol{y}_j\|_2^2 &= \sum_{i=1}^{n}\sum_{j=1}^{m}\pi_{ij}(\boldsymbol{x}_i - \boldsymbol{y}_j)^\top \mathrm{diag}(\boldsymbol{\alpha})(\boldsymbol{x}_i - \boldsymbol{y}_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m}\pi_{ij}\sum_{\ell=1}^{d}\alpha_\ell(x_i^{(\ell)} - y_j^{(\ell)})^2.\end{aligned}$$

Therefore, SRW and 2-FRWD are equivalent if we set $\boldsymbol{U} = (\sqrt{\alpha_1}\boldsymbol{e}_1, \sqrt{\alpha_2}\boldsymbol{e}_2, \ldots, \sqrt{\alpha_d}\boldsymbol{e}_d)^\top$ and $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2$.

3.5 Application: Semantic Correspondence

We applied our proposed FROT algorithm to semantic correspondence. The semantic correspondence is a problem that determines the matching of objects in two images. That is, given input image pairs (A, B), with common objects, we formulated the semantic correspondence problem to estimate the transport plan from the key points in A to those in B; this framework was proposed in (Liu et al., 2020). In Figure 2, we show an overview of our proposed framework.
Cost matrix computation $\boldsymbol{C}_\ell$: We employed a pretrained convolutional neural network to extract dense feature maps for each convolutional layer. The dense feature map of the ℓth layer output of the sth image is given by

$$\boldsymbol{f}_{s,q+(r-1)h_s}^{(\ell,s)} \in \mathbb{R}^{d_\ell},\ \ q \in [\![h_s]\!], r \in [\![w_s]\!], \ell \in [\![L]\!],$$

where $[\![L]\!] = \{1, 2, \ldots, L\}$, w_s and h_s are the width and height of the sth image, respectively, and d_ℓ is the dimension of the ℓth layer's feature map. Note that because the dimension of the dense feature map is different for each layer, we sample feature maps to the size of the 1st layer's feature map size (i.e., $h_s \times w_s$).

The ℓth layer's cost matrix for images s and s' is given by

$$[\boldsymbol{C}_\ell]_{ij} = \|\boldsymbol{f}_i^{(\ell,s)} - \boldsymbol{f}_j^{(\ell,s')}\|_2^2, \ \ i \in [\![w_s h_s]\!], j \in [\![w_{s'} h_{s'}]\!].$$

A potential problem with FROT is that the estimation depends significantly on the magnitude of the cost of each layer (also known as a group). Hence, normalizing each cost matrix is important. Therefore, we normalized each feature vector by $\boldsymbol{f}_i^{(\ell,s)} \leftarrow \boldsymbol{f}_i^{(\ell,s)}/\|\boldsymbol{f}_i^{(\ell,s)}\|_2$. Consequently, the cost matrix is given by $[\boldsymbol{C}_\ell]_{ij} = 2 - 2\boldsymbol{f}_i^{(\ell,s)^\top}\boldsymbol{f}_j^{(\ell,s')}$. We can use distances such as the $L1$ distance.
Computation of $\boldsymbol{a}$ and $\boldsymbol{b}$ with staircase re-weighting: Setting $\boldsymbol{a} \in \mathbb{R}^{h_s w_s}$ and $\boldsymbol{b} \in \mathbb{R}^{h_{s'} w_{s'}}$ is important because semantic correspondence can be affected by background clutter. Therefore, we generated the class activation maps (Zhou et al., 2016) for the source and target images and used them as $\boldsymbol{a}$ and $\boldsymbol{b}$, respectively. For CAM, we chose the class with the highest classification probability and normalized it to the range $[0, 1]$.

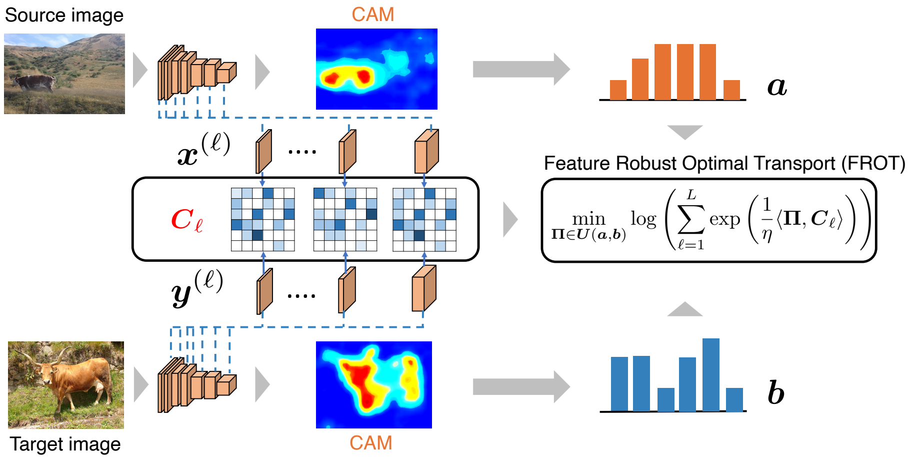

Source image
CAM
$\boldsymbol{a}$
$\boldsymbol{x}^{(\ell)}$
....
$\boldsymbol{C}_\ell$
Feature Robust Optimal Transport (FROT)
$\min_{\boldsymbol{\Pi}\in\boldsymbol{U}(\boldsymbol{a},\boldsymbol{b})} \log\left(\sum_{\ell=1}^{L}\exp\left(\frac{1}{\eta}\langle\boldsymbol{\Pi},\boldsymbol{C}_\ell\rangle\right)\right)$
$\boldsymbol{y}^{(\ell)}$
....
Target image
CAM
$\boldsymbol{b}$

Fig. 2: Semantic correspondence framework based on FROT.

3.6 Application: Feature Selection

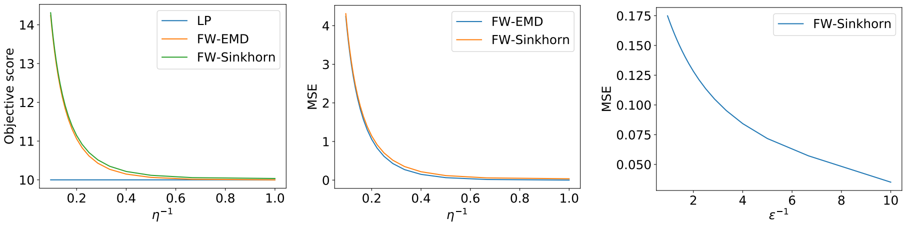
Since FROT finds the transport plan and discriminative features between $\boldsymbol{X}$ and $\boldsymbol{Y}$, we can use FROT as a feature-selection method. We considered $\boldsymbol{X} \in \mathbb{R}^{d\times n}$ and $\boldsymbol{Y} \in \mathbb{R}^{d\times m}$ as sets of samples from classes 1 and 2, respectively. The optimal feature importance is given by

$$\widehat{\alpha}_\ell = \frac{\exp\left(\frac{1}{\eta}\langle\widehat{\boldsymbol{\Pi}},\boldsymbol{C}_\ell\rangle\right)}{\sum_{\ell'=1}^{d}\exp\left(\frac{1}{\eta}\langle\widehat{\boldsymbol{\Pi}},\boldsymbol{C}_{\ell'}\rangle\right)}, \text{ with } \widehat{\boldsymbol{\Pi}} = \mathop{\mathrm{argmin}}_{\boldsymbol{\Pi}\in\boldsymbol{U}(\mu,\nu)} \eta\log\left(\sum_{\ell=1}^{d}\exp\left(\frac{1}{\eta}\langle\boldsymbol{\Pi},\boldsymbol{C}_\ell\rangle\right)\right),$$

where $[\boldsymbol{C}_\ell]_{ij} = (x_i^{(\ell)} - y_j^{(\ell)})^2$. Finally, we selected the top K features by the ranking $\widehat{\boldsymbol{\alpha}}$. Hence, $\boldsymbol{\alpha}$ changes to a one-hot vector for a small η and to $\alpha_k \approx \frac{1}{L}$ for a large η.

4 Related Work

The Wasserstein distance can be determined by solving the OT problem, and has many applications in NLP and CV such as measuring document similarity (Kusner et al., 2015; Sato et al., 2022) and finding local feature matching between images (Sarlin et al., 2020; Liu et al., 2020). An advantage of the Wasserstein distance is its robustness to noise; moreover, we can obtain the transport plan, which is useful for many applications. To reduce the computation cost for the Wasserstein distance, the sliced Wasserstein distance is useful (Kolouri et al., 2016). Recently, a tree variant of the Wasserstein distance was proposed (Evans and Matsen, 2012; Le et al., 2019; Sato et al., 2020; Takezawa et al., 2021, 2022); the sliced Wasserstein distance is a special case of this algorithm.



(a) Objective score. (b) MSE (η). (c) MSE (ϵ).

Fig. 3: (a) Objective scores for LP, FW-EMD, and FW-Sinkhorn. (b) MSE between transport plan of LP and FW-EMD and that with LP and FW-Sinkhorn with different η. (c) MSE between transport plan of LP and FW-Sinkhorn with different ϵ.

The approach most closely related to FROT is a robust variant of the Wasserstein distance, called the subspace robust Wasserstein distance (SRW) (Paty and Cuturi, 2019). SRW computes the OT problem in a discriminative subspace; this is possible by solving dimensionality-reduction problems. Owing to the robustness, SRW can successfully compute the Wasserstein distance from noisy data. The max–sliced Wasserstein distance (Deshpande et al., 2019) and its generalized counterpart (Kolouri et al., 2019) can also be regarded as subspace-robust Wasserstein methods. Note that SRW (Paty and Cuturi, 2019) is a *min–max* based approach, while the max–sliced Wasserstein distances (Deshpande et al., 2019; Kolouri et al., 2019) are *max–min* approaches. The FROT is a feature selection variant of the Wasserstein distance, whereas the subspace approaches are used for dimensionality reduction.

As a parallel work, a general minimax optimal transport problem called the robust Kantorovich problem (RKP) was recently proposed (Dhouib et al., 2020). RKP involves using a cutting-set method for a general minmax optimal transport problem that includes the FROT problem as a special case. The approaches are technically similar. However, we aim to solve a high-dimensional OT problem using feature selection and apply it to semantic correspondence problems, while the RKP approach focuses on providing a general framework and uses it for color transformation problems. As a technical difference, the cutting-set method may not converge to an optimal solution if we use the regularized OT (Dhouib et al., 2020). By contrast, because we use a Frank–Wolfe algorithm, our algorithm converges to a true objective function with regularized OT solvers. The multiobjective optimal transport (MOT) is an approach (Scetbon et al., 2021) parallel to ours. The key difference between FROT and MOT is that MOT tries to use the weighted sum of cost functions, while FROT considers the worst case. Moreover, we focus on the cost matrices computed from subsets of features, while MOT considers cost matrices with different distance functions.

5 Experiments

In this section, we evaluate the FROT algorithm using synthetic and real-world datasets.

5.1 Synthetic Data

We compare FROT with a standard OT using synthetic datasets. In these experiments, we initially generate two-dimensional vectors $\boldsymbol{x} \sim N(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ and $\boldsymbol{y} \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$. Here, we set $\boldsymbol{\mu}_x = (-5, 0)^\top$, $\boldsymbol{\mu}_y = (5, 0)^\top$, $\boldsymbol{\Sigma}_x = \boldsymbol{\Sigma}_y = ((5, 1)^\top, (4, 1)^\top)$. Then, we concatenate $\boldsymbol{z}_x \sim N(\boldsymbol{0}_8, \boldsymbol{I}_8)$ and $\boldsymbol{z}_y \sim N(\boldsymbol{0}_8, \boldsymbol{I}_8)$ to $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, to give $\widetilde{\boldsymbol{x}} = (\boldsymbol{x}^\top, \boldsymbol{z}_x^\top)^\top$, $\widetilde{\boldsymbol{y}} = (\boldsymbol{y}^\top, \boldsymbol{z}_y^\top)^\top$.

For FROT, we set $\eta = 1.0$, $T = 10$, and $\epsilon = 0.02$, respectively. To show the proof-of-concept, we set the true features as a group and the remaining noise features as another group.

Fig. 1a shows the correspondence from $\boldsymbol{x}$ and $\boldsymbol{y}$ with the vanilla OT algorithm. Figs. 1b and 1c show the correspondence of FROT and OT with $\widetilde{\boldsymbol{x}}$ and $\widetilde{\boldsymbol{y}}$, respectively. Although FROT can identify a suitable matching, the OT fails to obtain a significant correspondence. We observed that the $\boldsymbol{\alpha}$ parameter corresponding to a true group is $\alpha_1 = 0.9999$. Moreover, we compared the objective scores of the FROT with LP, FW-EMD, and FW-Sinkhorn ($\epsilon = 0.1$). Figure 3a shows the objective scores of FROTs with the different solvers, and both FW-EMD and FW-Sinkhorn can achieve almost the same objective score with a relatively small η. Moreover, Figure 3b shows the mean squared error between the LP method and the FW counterparts. Similar to the objective score cases, it can yield a similar transport plan with a relatively small η. Finally, we evaluated the FW-Sinkhorn by changing the regularization parameter η. In this experiment, we set $\eta = 1$ and varied the ϵ values. The result shows that we can obtain an accurate transport plan with a relatively small ϵ.

5.2 Semantic correspondence

We evaluated our FROT algorithm for semantic correspondence. In this study, we used the SPair-71k (Min et al., 2019b). The SPair-71k dataset consists of $70,958$ image pairs. For evaluation, we employed a percentage of accurate key points (PCK), which counts the number of accurately predicted key points given a fixed threshold (Min et al., 2019b). All semantic correspondence experiments were run on a Linux server with NVIDIA P100.

For the optimal transport based frameworks, we employed ResNet101 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009) for feature and activation map extraction. The ResNet101 consists of 34 convolutional layers and the entire number of features is $d = 32,576$. Note that we did not fine-tune the network. We compared the proposed method with several baselines (Min et al., 2019b) and the SRW Paty and Cuturi (2019). Owing to the computational cost and the required memory size for SRW, we used the first and the last few convolutional layers of ResNet101 as the input of SRW. In our experiments, we empirically set $T = 3$

Table 1: Per-class PCK ($\alpha_{bbox} = 0.1$) results using SPair-71k. All models use ResNet101. The numbers in the bracket of SRW are the input layer indicies.

Methods		aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	dog	horse	moto	person	plant	sheep	train	tv	all
SPair-71k finetuned models	CNNGeo (Rocco et al., 2017)	23.4	16.7	40.2	14.3	36.4	27.7	26.0	32.7	12.7	27.4	22.8	13.7	20.9	21.0	17.5	10.2	30.8	34.1	20.6
	A2Net (Hongsuck Seo et al., 2018)	22.6	18.5	42.0	16.4	37.9	**30.8**	26.5	35.6	13.3	29.6	24.3	16.0	21.6	22.8	20.5	13.5	31.4	36.5	22.3
	WeakAlign (Rocco et al., 2018a)	22.2	17.6	41.9	15.1	38.1	27.4	**27.2**	31.8	12.8	26.8	22.6	14.2	20.0	22.2	17.9	10.4	32.2	35.1	20.9
	NC-Net (Rocco et al., 2018b)	17.9	12.2	32.1	11.7	29.0	19.9	16.1	39.2	9.9	23.9	18.8	15.7	17.4	15.9	14.8	9.6	24.2	31.1	20.1
SPair-71k validation	HPF (Min et al., 2019a)	25.2	18.9	52.1	15.7	38.0	22.8	19.1	52.9	17.9	33.0	32.8	20.6	24.4	27.9	21.1	15.9	31.5	35.6	28.2
	OT-HPF (Liu et al., 2020)	32.6	18.9	**62.5**	20.7	42.0	26.1	20.4	61.4	**19.7**	**41.3**	**41.7**	**29.8**	**29.6**	**31.8**	25.0	23.5	44.7	37.0	33.9
	FROT($\eta = 0.2, \epsilon = 0.4$)	35.1	20.3	59.8	21.1	**42.9**	27.7	21.2	**63.5**	18.8	39.7	37.9	29.2	28.8	29.9	**28.2**	**24.3**	**52.1**	**39.5**	**34.7**
Without SPair-71k validation	OT	30.1	16.5	50.4	17.3	38.0	22.9	19.7	54.3	17.0	28.4	31.3	22.1	28.0	19.5	21.0	17.8	42.6	28.8	28.3
	FROT ($\eta = 0.3, T = 3$)	35.0	**20.9**	56.3	**23.4**	40.7	27.2	21.9	62.0	17.5	38.8	36.2	27.9	28.0	30.4	26.9	23.1	49.7	38.4	33.7
	FROT ($\eta = 0.3, T = 10$)	34.9	**20.9**	56.4	**23.4**	40.7	27.2	22.0	62.0	17.5	38.8	36.2	27.8	28.2	30.2	26.9	22.9	49.7	38.5	33.7
	FROT ($\eta = 0.5, T = 3$)	34.1	18.8	56.9	19.9	40.0	25.6	19.2	61.9	17.4	38.7	36.5	25.6	26.9	27.2	26.3	22.1	50.3	38.6	32.8
	FROT ($\eta = 0.5, T = 10$)	34.0	18.9	57.0	19.9	40.0	25.6	19.2	61.9	17.3	38.8	36.5	25.6	26.8	27.4	26.4	22.1	50.3	38.8	32.8
	FROT ($\eta = 0.7, T = 3$)	33.4	19.4	56.6	20.0	39.6	26.1	19.1	62.4	17.9	38.0	36.5	26.0	27.5	26.5	25.5	21.6	49.7	38.9	32.7
	FROT ($\eta = 0.7, T = 10$)	33.3	19.5	56.6	19.9	39.5	26.0	19.1	62.4	17.9	38.0	36.5	26.0	27.4	26.5	25.6	21.6	49.6	38.9	32.7
	SRW (layers = {1, 32–34})	29.4	14.0	43.7	15.6	33.8	21.0	17.6	48.0	12.9	23.3	26.5	19.8	25.5	17.6	16.7	15.2	37.1	20.5	24.5
	SRW (layers = {1, 31–34})	29.7	14.3	44.3	15.7	34.2	21.3	17.8	48.5	13.1	23.6	27.1	20.0	25.8	18.1	16.9	15.2	37.3	21.0	24.8
	SRW (layers = {1, 30–34})	29.8	14.7	45.6	15.9	34.8	21.5	18.0	49.3	13.3	24.0	27.7	20.6	25.7	18.7	17.2	15.3	37.7	21.5	25.2
	FROT (layers = {1, 32–34})	32.3	15.7	43.1	18.4	30.7	22.5	20.6	44.5	10.3	23.1	23.9	19.5	23.6	22.0	14.7	15.3	37.4	18.0	24.3
	FROT (layers = {1, 31–34})	35.3	16.5	45.3	20.5	33.0	25.0	21.6	48.1	11.4	25.9	26.9	22.4	25.2	25.0	16.5	17.1	40.5	21.2	26.6
	FROT (layers = {1, 30–34})	**36.7**	18.1	48.8	22.3	34.5	27.5	23.0	51.3	12.9	28.4	30.3	24.2	26.4	27.3	19.5	18.1	43.4	24.9	28.8

and $\epsilon = 0.1$ for FROT and SRW, respectively. For SRW, we set the number of latent dimension as $k = 50$ for all experiments. HPF (Min et al., 2019a) and OT-HPF (Liu et al., 2020) are state-of-the-art methods for semantic correspondence. HPF and OT-HPF required the validation dataset to select important layers, whereas SRW and FROT did not require the validation dataset. OT is a simple Sinkhorn-based method that does not select layers.

Table 1 lists the per-class PCK results obtained using the SPair-71k dataset. FROT ($\eta = 0.3$) outperforms most existing baselines, including HPF and OT. Moreover, FROT ($\eta = 0.3$) is consistent with OT-HPF (Liu et al., 2020), which requires the validation dataset to select important layers. In this experiment, setting $\eta < 1$ results in favorable performance (See Table 2 in the supplementary material). The computational costs of FROT is 0.29, while SRWs are 8.73, 11.73, 15.76, respectively. Surprisingly, FROT outperformed SRWs. However, this is mainly due to the used input layers.

We further evaluated FROT by tuning hyperparameters η and ϵ using validation sets, where the maximum search ranges for η and ϵ are set to 0.2 to 2.0 and 0.1 to 0.6 with intervals of 0.1, respectively. By using hyperparameter search, we selected $(\eta = 0.2, \epsilon = 0.4)$ as an optimal parameter. The FROT with optimal parameters outperforms the state-of-the-art method (Liu et al., 2020).

5.3 Feature Selection Experiments

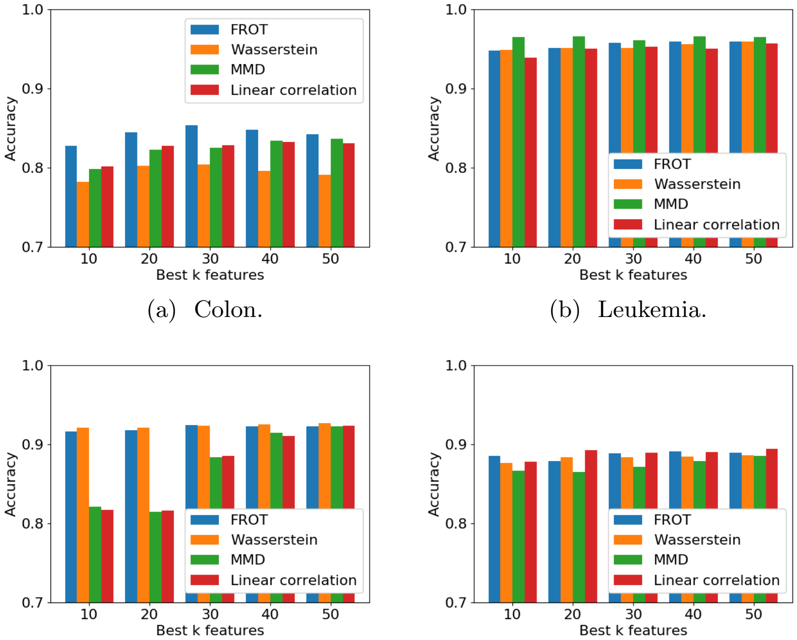
Here, we compared FROT with several baseline algorithms in terms of solving feature-selection problems. In this study, we employed a high-dimensional and a few sample datasets with two class classification tasks (see Table 3 in the supplementary material). All feature selection experiments were run on a Linux server with an Intel Xeon CPU E7-8890 v4 with 2.20 GHz and 2 TB RAM.

In our experiments, we initially randomly split the data into two sets (75% for training and 25% for testing) and used the training set for feature selection and building a classifier. Note that we standardized each feature using the training set. Then, we used the remaining set for the test. The trial was repeated 50 times, and we considered the averaged classification accuracy for all trials. Considered as baseline methods, we computed the Wasserstein distance, maximum mean discrepancy (MMD) (Gretton et al., 2007), and linear correlation[6] for each dimension and sorted them in descending order. Note that the Wasserstein distance is computed via sorting, which is computationally more efficient than the Sinkhorn algorithm when $d = 1$. Then, we selected the top K features as important features. For FROT, we computed the feature importance and selected the features that had significant importance scores. In our experiments, we set $\eta = 1.0$ and $T = 10$. Then, we trained a two-class SVM[7] with the selected features.

Fig. 4 shows the average classification accuracy relative to the number of selected features. From Figure 4, FROT is consistent with the Wasserstein

[6] `https://scikit-learn.org/stable/modules/feature_selection.html`

[7] `https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html`



(a) Colon.

(b) Leukemia.

(c) Prostate_ge.

(d) GLI_85.

Fig. 4: Feature selection results. We average over 50 runs of accuracy (on test set) of SVM trained with top k features selected by several methods.

distance-based feature selection and outperforms the linear correlation method and the MMD for two datasets. Table 3 in the supplementary file shows the computational time(s) of the methods. FROT is about two orders of magnitude faster than the Wasserstein distance and is also faster than MMD. Note that although MMD is as fast as the proposed method, it cannot determine the correspondence between samples.

6 Conclusion

In this paper, we proposed FROT for high-dimensional data. This approach jointly solves feature selection and OT problems. An advantage of FROT is that it is a convex optimization problem and can determine an accurate globally optimal solution using the Frank–Wolfe algorithm. We used FROT for feature selection and semantic correspondence problems. Through experiments, we demonstrated that the proposed algorithm is consistent with state-of-the-art algorithms in both feature selection and semantic correspondence.

7 Acknowledgement

M.Y. was supported by MEXT KAKENHI Grant Number 20H04243. R.S. was supported by JSPS KAKENHI GrantNumber 21J22490.

Bibliography

Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and sparse optimal transport. In *AISTATS*, 2018.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NIPS*, 2013.

Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. *ICML*, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. Max-sliced Wasserstein distance and its use for GANs. In *CVPR*, 2019.

Sofien Dhouib, Ievgen Redko, Tanguy Kerdoncuff, Rémi Emonet, and Marc Sebban. A swiss army knife for minimax optimal transport. In *ICML*, 2020.

Steven N Evans and Frederick A Matsen. The phylogenetic kantorovich–rubinstein metric for environmental sequence samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):569–592, 2012.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

Arthur. Gretton, Kenji. Fukumizu, C. Hui. Teo, Le. Song, Bernhard. Schölkopf, and Alex Smola. A kernel statistical test of independence. In *NIPS*, 2007.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Paul Hongsuck Seo, Jongmin Lee, Deunsol Jung, Bohyung Han, and Minsu Cho. Attentive semantic alignment with offset-aware correlation kernels. In *ECCV*, 2018.

Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML*, 2013.

Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Wasserstein regularization for sparse multi-task regression. In *AISTATS*, 2019.

Soheil Kolouri, Yang Zou, and Gustavo K Rohde. Sliced wasserstein kernels for probability distributions. In *CVPR*, 2016.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced wasserstein distances. In *NeurIPS*, 2019.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *ICML*, 2015.

Tam Le, Makoto Yamada, Kenji Fukumizu, and Marco Cuturi. Tree-sliced approximation of wasserstein distances. *NeurIPS*, 2019.

Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *CVPR*, 2020.

Yanbin Liu, Makoto Yamada, Yao-Hung Hubert Tsai, Tam Le, Ruslan Salakhutdinov, and Yi Yang. Lsmi-sinkhorn: Semi-supervised squared-loss mutual information estimation with optimal transport. *ECML*, 2021.

Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *ICCV*, 2019a.

Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv preprint arXiv:1908.10543*, 2019b.

Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

François-Pierre Paty and Marco Cuturi. Subspace robust wasserstein distances. In *ICML*, 2019.

François-Pierre Paty and Marco Cuturi. Regularized optimal transport is ground cost adversarial. *ICML*, 2020.

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

Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *CVPR*, 2017.

Ignacio Rocco, Relja Arandjelović, and Josef Sivic. End-to-end weakly-supervised semantic alignment. In *CVPR*, 2018a.

Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *NeurIPS*, 2018b.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020.

Ryoma Sato, Makoto Yamada, and Hisashi Kashima. Fast unbalanced optimal transport on tree. In *NeurIPS*, 2020.

Ryoma Sato, Makoto Yamada, and Hisashi Kashima. Re-evaluating word mover's distance. *ICML*, 2022.

Meyer Scetbon, Laurent Meunier, Jamal Atif, and Marco Cuturi. Equitable and optimal transport with multiple agents. In *AISTATS*, 2021.

Yuki Takezawa, Ryoma Sato, and Makoto Yamada. Supervised tree-wasserstein distance. In *ICML*, 2021.

Yuki Takezawa, Ryoma Sato, Zornitsa Kozareva, Sujith Ravi, and Makoto Yamada. Fixed support tree-sliced wasserstein barycenter. *AISTATS*, 2022.

Yuguang Yan, Wen Li, Hanrui Wu, Huaqing Min, Mingkui Tan, and Qingyao Wu. Semi-supervised optimal transport for heterogeneous domain adaptation. In *IJCAI*, 2018.

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, 2006.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.