Submodular Framework for Structured-Sparse Optimal Transport

Piyushi Manupriya [1] **Pratik Jawanpuria** [2] **Karthik S. Gurumoorthy** [3] **SakethaNath Jagarlapudi** [1]
Bamdev Mishra [2]

Abstract

Unbalanced optimal transport (UOT) has recently gained much attention due to its flexible framework for handling un-normalized measures and its robustness properties. In this work, we explore learning (structured) sparse transport plans in the UOT setting, i.e., transport plans have an upper bound on the number of non-sparse entries in each column (structured sparse pattern) or in the whole plan (general sparse pattern). We propose novel sparsity-constrained UOT formulations building on the recently explored maximum mean discrepancy based UOT. We show that the proposed optimization problem is equivalent to the maximization of a weakly submodular function over a uniform matroid or a partition matroid. We develop efficient gradient-based discrete greedy algorithms and provide the corresponding theoretical guarantees. Empirically, we observe that our proposed greedy algorithms select a diverse support set and we illustrate the efficacy of the proposed approach in various applications.

1. Introduction

Optimal transport (OT) has emerged as a popular tool in machine learning applications for comparing probability distributions (Peyré et al., 2019). OT computes the minimal cost to transform one distribution into another and generates a transport plan, offering a deeper understanding of the underlying geometry. The obtained transport plan may be used for aligning the support of the distributions (Alvarez-Melis & Jaakkola, 2018; Jawanpuria et al., 2020), domain adaptation (Courty et al., 2017; Nath & Jawanpuria, 2020), ecological inference (Muzellec et al., 2017), etc. Furthermore, OT has been explored in diverse applications such as generative modeling (Arjovsky et al., 2017), shape interpolation (Solomon et al., 2015; Han et al., 2022), prototype selection (Gurumoorthy et al., 2021), multi-label classification (Frogner et al., 2015; Jawanpuria et al., 2021), single-cell RNA sequencing (Schiebinger et al., 2019), and hypothesis testing (Manupriya et al., 2024b), to name a few.

The seminal work of Cuturi (2013) popularized the entropic regularized variants of OT for their computational and generalization benefits. However, a notable drawback of entropic regularized OT approaches is that they usually learn dense transport plans, where sparse (zero) entries are almost nonexistent (Blondel et al., 2018; Liu et al., 2023). Sparser transport plans are often preferred as they offer more interpretability in alignments (Muzellec et al., 2017; Swanson et al., 2020). In this regard, existing works (Blondel et al., 2018; Essid & Solomon, 2018) have shown that the squared 2-norm regularizer for OT leads to a sparse OT plan. More recently, Liu et al. (2023) introduced an explicit cardinality constraint to control the sparsity level. It should be noted that the above works explore sparsity in the balanced OT setups, i.e., when the marginals of the transport plan are enforced to match the given distributions.

While balanced OT is suitable for many applications, the need for robustness in the case of noisy measures motivates relaxing the marginal matching constraints (Frogner et al., 2015; Fatras et al., 2021). This has led to several unbalanced OT (UOT) methods (Liero et al., 2018; Chizat et al., 2017) where a KL-divergence based regularization is employed for (softly) enforcing marginal constraints. Recently, Manupriya et al. (2024b) proposed a maximum mean discrepancy (MMD) regularized UOT approach, termed as MMD-UOT, as an alternative to KL-regularized UOT. However, to the best of our knowledge, the existing UOT works do not focus on learning (structured) sparse transport plans with explicit cardinality constraints.

Contributions. In this work, we propose novel sparsity-constrained UOT formulations. In particular, we learn UOT plans with a general sparsity constraint or a column-wise sparsity constraint. While the corresponding search space is non-convex and non-smooth, we identify them with well-studied matroid structures such as uniform matroid or partition matroid. Our contributions are as follows.

[1] Department of Computer Science and Engineering, IIT Hyderabad, India. [2] Microsoft, India. [3] Walmart Global Tech, India. Correspondence to: Pratik Jawanpuria <pratik.jawanpuria@microsoft.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

- We show that the MMD-UOT problem (Manupriya et al., 2024b), when viewed as a function of the support set of transport plan, is equivalent to maximizing a weakly submodular function. This allows us to view our proposed sparsity-constrained UOT problems as maximizing a weakly submodular function over a matroid constraint.
- We propose novel efficient gradient-based greedy algorithms (Algorithms 1 and 2) with attractive theoretical guarantees for maximizing a weakly submodular function over a (uniform or partition) matroid constraint. While the algorithms can be readily applied to solve our proposed constrained UOT formulations, they are also of independent interest for maximizing a general weakly submodular function.
- A salient feature of our investigation is the dual analysis of (non-convex) weakly submodular problems. The usual approximation results corresponding to greedy maximization of (weakly) submodular functions are on lower bounds. While these lower bounds capture the worst case performance, often in practice, they do not explain the good performance of the greedy algorithms. In this context, the duality gap analysis provides a more optimistic bound on the performance.
- Finally, we empirically demonstrate the effectiveness of the proposed approach in several applications.

The proofs of our theoretical results and additional experimental details are provided in the appendix sections.

2. Preliminaries

We begin with a few notations. Let $X := \{\mathbf{x}_i\}_{i=1}^m$ and $Y := \{\mathbf{y}_j\}_{j=1}^n$ be the source and target datasets, respectively, where $\mathbf{x}_i \in \mathcal{X}$ and $\mathbf{y}_i \in \mathcal{Y}$. The corresponding empirical distributions may be written as $\mu := \sum_{i=1}^m \boldsymbol{\mu}_i \delta_{\mathbf{x}_i}$ and $\nu := \sum_{j=1}^n \boldsymbol{\nu}_i \delta_{\mathbf{y}_j}$, where $\boldsymbol{\mu}_i$ and $\boldsymbol{\nu}_j$ denote the mass associated with samples $\mathbf{x}_i$ and $\mathbf{y}_j$, respectively, and $\delta_{\mathbf{z}}$ represents the Dirac measure centered on $\mathbf{z}$. Let $\mathbf{1}$ and $\mathbf{0}$ denote a vector/matrix of ones and zeros, respectively, whose size could be understood from the context. Then, $\boldsymbol{\mu} \in \Delta_m$ and $\boldsymbol{\nu} \in \Delta_n$, where $\Delta_d = \{\mathbf{z} \in \mathbb{R}_+^d : \mathbf{z}^\top \mathbf{1} = 1\}$. For $m \in \mathbb{N}$, let $[m] = \{1, 2, \ldots, m\}$. Let $V \equiv \{(i,j) : i \in [m]; j \in [n]\}$ represent the index set of an $m \times n$ matrix. Let $\mathrm{vec}(\mathbf{M})$ denote the vectorization of the matrix $\mathbf{M}$, and for an index $u \equiv (i,j)$, $\mathbf{g}_u$ denotes the element $\mathbf{g}[i,j]$. For a non-negative vector $\mathbf{z} \in \mathbb{R}_+^d$, the indices of non-zero entries in $\mathbf{z}$ (its support) are denoted by the set $\mathrm{supp}(\mathbf{z}) = \{i \in [d] : \mathbf{z}_i > 0\}$.

2.1. Optimal Transport

Optimal transport (OT) quantifies the distance between two distributions μ and ν while incorporating the geometry over their supports. Let $\mathbf{C} \in \mathbb{R}_+^{m \times n}$ be a cost matrix induced by a cost metric $c : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}_+$ such that $\mathbf{C}_{ij} = c(\mathbf{x}_i, \mathbf{y}_j)$. Kantorovich (1942) proposed the OT problem between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ as $\min_{\boldsymbol{\gamma} \in \Gamma(\boldsymbol{\mu},\boldsymbol{\nu})} \langle \mathbf{C}, \boldsymbol{\gamma} \rangle$, where $\Gamma(\boldsymbol{\mu},\boldsymbol{\nu}) := \{\boldsymbol{\gamma} \in \mathbb{R}_+^{m \times n} : \boldsymbol{\gamma}\mathbf{1} = \boldsymbol{\mu}; \boldsymbol{\gamma}^\top \mathbf{1} = \boldsymbol{\nu}\}$. This is a *balanced* OT problem due to the presence of mass preservation constraints $\boldsymbol{\gamma}\mathbf{1} = \boldsymbol{\mu}$ and $\boldsymbol{\gamma}^\top \mathbf{1} = \boldsymbol{\nu}$. The transport plan $\boldsymbol{\gamma}$ is a joint distribution with marginals $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ and supported over the (index) set V.

Recent works have explored relaxing the mass-preservation constraint of classical OT for settings where measures are noisy (Balaji et al., 2020) or un-normalized (Chizat et al., 2017). Unbalanced optimal transport (UOT) replaces the constraint $\boldsymbol{\gamma} \in \Gamma(\boldsymbol{\mu},\boldsymbol{\nu})$ with regularizers $\mathcal{D}(\boldsymbol{\gamma}\mathbf{1}, \boldsymbol{\mu})$ and $\mathcal{D}(\boldsymbol{\gamma}^\top \mathbf{1}, \boldsymbol{\nu})$, which promote the marginals of $\boldsymbol{\gamma}$ to be close to the given $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ distributions. Here, $\mathcal{D}$ is a divergence or distance between distributions such as KL-divergence (Fatras et al., 2021), maximum mean discrepancy (MMD) (Gretton et al., 2012), etc. A recent work (Manupriya et al., 2024b) has studied MMD regularization for the UOT problem. Given a cost matrix $\mathbf{C}$ and a universal kernel k (Sriperumbudur et al., 2011), MMD-UOT (Manupriya et al., 2024b) is the following convex problem:

$$\begin{aligned} &\min_{\boldsymbol{\gamma} \geq \mathbf{0}} \ \mathcal{U}(\boldsymbol{\gamma}), \text{ where} \\ \mathcal{U}(\boldsymbol{\gamma}) := \langle \mathbf{C}, \boldsymbol{\gamma} \rangle &+ \lambda_1 \mathrm{MMD}_k^2(\boldsymbol{\gamma}\mathbf{1}, \boldsymbol{\mu}) \\ &+ \lambda_1 \mathrm{MMD}_k^2(\boldsymbol{\gamma}^\top \mathbf{1}, \boldsymbol{\nu}) + \tfrac{\lambda_2}{2}\|\boldsymbol{\gamma}\|^2. \end{aligned} \tag{1}$$

Here, $\mathrm{MMD}_k(\boldsymbol{\gamma}\mathbf{1}, \boldsymbol{\mu}) = \|\boldsymbol{\gamma}\mathbf{1} - \boldsymbol{\mu}\|_{\mathbf{G}_1}$, $\mathrm{MMD}_k(\boldsymbol{\gamma}^\top \mathbf{1}, \boldsymbol{\nu}) = \|\boldsymbol{\gamma}^\top \mathbf{1} - \boldsymbol{\nu}\|_{\mathbf{G}_2}$, $\mathbf{G}_1$ and $\mathbf{G}_2$ are the Gram matrices corresponding to kernel k over the source and target points, respectively, and $\|\mathbf{z}\|_{\mathbf{G}} = \sqrt{\mathbf{z}^\top \mathbf{G} \mathbf{z}}$. We may additionally employ a squared ℓ_2-norm regularization ($\lambda_2 \geq 0$) for computational benefits (Blondel et al., 2018).

2.2. Submodularity

Submodularity is a property of set functions that exhibit diminishing returns. Given two sets A and B such that $A \subseteq B \subseteq V$, a set function is submodular if and only if for any $u \notin B$, $F(A \cup \{u\}) - F(A) \geq F(B \cup \{u\}) - F(B)$. The term, $F(A \cup \{u\}) - F(A)$, is the marginal gain on adding an element u to the set the A and is popularly denoted as $F(u|A)$. Likewise $F(S|B)$ denotes $F(B \cup S) - F(B)$. The set function is monotone increasing iff $F(A) \leq F(B)$ when $A \subseteq B \subseteq V$. For non-negative monotone submodular maximization problem, $\max_{S \subseteq V, |S| \leq K} F(S)$, Nemhauser et al. (1978) showed that the classical greedy algorithm obtains a $(1 - e^{-1})$ approximation to the optimal objective.

Another naturally occurring structure is that of a matroid defined as follows. Given a non-empty collection $\mathcal{I} \subseteq 2^V$, the pair $\mathcal{M} = (V, \mathcal{I})$ is a matroid if for every two sets $A, B \subseteq V$, the following are satisfied: (i) $\emptyset \in \mathcal{I}$; (ii) If $A \subseteq B$ and $B \in \mathcal{I}$, then $A \in \mathcal{I}$; and (iii) If $|A| < |B|$

and $A, B \in \mathcal{I}$, then $\exists u \in B \setminus A$ such that $A \cup \{u\} \in \mathcal{I}$. The elements of set $\mathcal{I}$ are called the independent sets of matroid $\mathcal{M}$. A set $X \subseteq V$ such that $X \notin \mathcal{I}$ is called a dependent set of $\mathcal{M}$. A maximal independent set that becomes dependent upon adding any element of V is called a base for the matroid. Given a matroid $\mathcal{M} = (V, \mathcal{I})$, the associated matroid constraint is $S \in \mathcal{I}(\mathcal{M})$, which implies that set S is an independent set of $\mathcal{M}$.

A function is said to exhibit a weaker notion of submodularity, characterized by $\boldsymbol{\alpha}$-weakly submodular (Das & Kempe, 2018) for some $\alpha \in (0, 1]$, if $\sum_{u \in S} F(u|B) \geq \alpha.F(S|B)$ for all $S, B \subseteq V$. Similar to submodular functions, constant-factor approximation guarantees also exist for maximizing a weakly submodular set function under cardinality and matroid constraints (Das & Kempe, 2018; Chen et al., 2018).

2.3. Restricted Strong Concavity and Restricted Strong Smoothness

On a domain $\Omega \subset \mathbb{R}^N \times \mathbb{R}^N$, a function $l : \mathbb{R}^N \mapsto \mathbb{R}$ is said to be restricted strong concave (RSC) with parameter u_Ω and restricted smooth (RSM) with parameter U_Ω if for all $(\mathbf{x}, \mathbf{y}) \in \Omega$, the following holds (Elenberg et al., 2018):

$$\frac{u_\Omega}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 \leq l(\mathbf{x})-l(\mathbf{y})+\langle \nabla l(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \leq \frac{U_\Omega}{2}\|\mathbf{y}-\mathbf{x}\|_2^2.$$

We denote the RSC and RSM parameters on the domain $\Omega_K = \{(\mathbf{x}, \mathbf{y}) : \mathbf{x} \geq 0; \|\mathbf{x}\|_0 \leq K; \mathbf{y} \geq 0; \|\mathbf{y}\|_0 \leq K\}$ of all K-sparse non-negative vectors by u_K and U_K, respectively. This set is of interest as we aim to learn non-negative transport plans with at most K non-zero entries. Also, let $\tilde{\Omega} = \{(\mathbf{x}, \mathbf{y}) : \|\mathbf{x} - \mathbf{y}\|_0 \leq 1\}$ with the corresponding smoothness parameter $\tilde{U}_1$. It can be easily verified that if $\hat{K} \leq K$, then $u_{\hat{K}} \geq u_K$ and $U_{\hat{K}} \leq U_K$ as $\Omega_{\hat{K}} \subseteq \Omega_K$.

3. Proposed Method

Given a source μ and a target ν distributions, we now propose a novel submodular framework for structured-sparse UOT. In this regard, we generalize the MMD-UOT formulation (1) by introducing additional (structured) sparsity constraints on the transport plan as follows:

$$\min_{\boldsymbol{\gamma} \in \mathcal{C}} \mathcal{U}(\boldsymbol{\gamma}), \tag{2}$$

where $\mathcal{U} : \mathbb{R}_+^{m \times n} \mapsto \mathbb{R}_+$ is the function defined in (1) and $\mathcal{C}$ denotes a set of sparsity constraints. In this work, we focus on two different sparsity constraints: (a) $\mathcal{C} \equiv \mathcal{C}_1 := \{\boldsymbol{\gamma} \in \mathbb{R}_+^{m \times n} : \|\text{vec}(\boldsymbol{\gamma})\|_0 \leq K_1\}$ or (b) $\mathcal{C} \equiv \mathcal{C}_2 := \{\boldsymbol{\gamma} \in \mathbb{R}_+^{m \times n} : \|\boldsymbol{\gamma}_j\|_0 \leq K_2 \, \forall \, j \in [n]\}$, where $\|\cdot\|_0$ denotes the ℓ_0-norm and $\boldsymbol{\gamma}_j$ denotes the j-th column of matrix $\boldsymbol{\gamma}$. While $\mathcal{C}_1$ imposes a cardinality constraint on the entire transport plan $\boldsymbol{\gamma}$, $\mathcal{C}_2$ imposes the cardinality constraint on each column of $\boldsymbol{\gamma}$. Note that MMD-UOT formulation (1) is a special case of Problem (2), e.g., when $K_1 = mn$ or $K_2 = m$. Problem (2) is non-convex over a non-smooth search space $\mathcal{C}$, and hence tricky to optimize even though the objective $\mathcal{U}$ is a convex function. However, we note that the constraint sets $\mathcal{C}_1$ or $\mathcal{C}_2$ essentially restrict the support of the transport plan $\boldsymbol{\gamma}$ to certain patterns which may be modeled using a matroid structure. For instance, the set $\mathcal{C}_1$ may equivalently be represented as a uniform matroid $\mathcal{M}_1 = (V, \mathcal{I}_1)$ where $\mathcal{I}_1 = \{S \subseteq V : |S| \leq K_1\}$. Similarly, the set $\mathcal{C}_2$ may be equivalently modeled using a partition matroid $\mathcal{M}_2 = (V, \mathcal{I}_2)$ where $\mathcal{I}_2 = \{S : S \subseteq V; |S \cap P_j| \leq K_2 \, \forall \, j \in [n]\}$ with $P_j = \{(i, j) : i \in [m]\}$.

Due to this interesting correspondence between the sparsity constraints $\mathcal{C}_1$ or $\mathcal{C}_2$ and the matroids, we equivalently pose the continuous Problem (2) as the following maximization problem over discrete sets representing the support of $\boldsymbol{\gamma}$:

$$\max_{S \in \mathcal{I}(\mathcal{M})} F(S)(:= \mathcal{U}(\mathbf{0}) - \min_{\boldsymbol{\gamma}:\text{supp}(\boldsymbol{\gamma}) \subseteq S, \boldsymbol{\gamma} \geq \mathbf{0}} \mathcal{U}(\boldsymbol{\gamma})), \tag{3}$$

where the matroid $\mathcal{M}$ corresponds to either the uniform matroid ($\mathcal{M} = \mathcal{M}_1$) or the partition matroid ($\mathcal{M} = \mathcal{M}_2$). Hence, we decouple the non-convex non-smooth problem (2) into a discrete optimization problem (3) whose objective evaluation requires solving a convex problem. For a candidate set $S \in \mathcal{I}(\mathcal{M})$, computing $F(S)$ essentially requires solving the MMD-UOT problem (1) with the support of $\boldsymbol{\gamma}$ restricted to set S. Since the objective $\mathcal{U}(\boldsymbol{\gamma})$ is L-smooth, (1) can solved using the accelerated projected gradient descent (APGD) method with a fixed step size of $1/L$ and has a linear convergence rate (Manupriya et al., 2024b).

A key outcome of the above reformulation is our next result, which proves that the set function $F(\cdot)$ is weakly submodular under mild assumptions on the kernel employed in (3). Please refer to Appendix A1.2 for more details.

Lemma 3.1. *$F(.)$ is a monotone, non-negative, and α-weakly submodular function with the submodularity ratio $\alpha \geq \frac{u_{2K}}{\tilde{U}_1} > 0$, where K denotes the sparsity level of the transport plan $\boldsymbol{\gamma}$. Here, $K = K_1$ for $\mathcal{M} = \mathcal{M}_1$ and $K = nK_2$ for $\mathcal{M} = \mathcal{M}_2$.*

The proof of Lemma 3.1 is discussed in Appendix A2.2. In the following sections, we propose efficient greedy algorithms with attractive approximation guarantees for maximizing our weakly submodular problem (3).

3.1. Learning (General) Sparse Transport Plan

As discussed, sparse transport plans are more interpretable and are useful in applications such as designing topology (Luo et al., 2023), word alignment (Arase et al., 2023), etc. To this end, we consider solving (3) with $\mathcal{M} = \mathcal{M}_1$, i.e.,

$$\max_{S \in \mathcal{I}_1(\mathcal{M}_1)} F(S). \tag{4}$$

This problem learns a sparse transport plan with a maximum of $K = K_1$ non-sparse and we term it as **GenSparseUOT**.

Algorithm 1 Stochastic OMP algorithm for maximizing weakly submodular problems with cardinality constraint

```
Input: $\lambda_1, \lambda_2, \boldsymbol{\mu}, \boldsymbol{\nu}, \mathbf{C}, \mathbf{G}_1, \mathbf{G}_2$, sparsity level $K$, $\epsilon$.
$i = 1, S_0 = \emptyset, \boldsymbol{\gamma}_{S_0} = \mathbf{0}$ and $\mathbf{g} = -\nabla\mathcal{U}(\boldsymbol{\gamma}_{S_0})$.
while $i \leq K$ do
    1. Set $R_i$ as a random subset of $V \setminus S_{i-1}$ with $mnK^{-1}\log(1/\epsilon)$ elements
    2. $u = \arg\max_{e \in R_i} \mathbf{g}_e$
    3. $S_i = S_{i-1} \cup \{u\}$
    4. $\boldsymbol{\gamma}_{S_i} = \arg\min_{\boldsymbol{\gamma}:\text{supp}(\boldsymbol{\gamma}) \subseteq S_i, \boldsymbol{\gamma} \geq \mathbf{0}} \mathcal{U}(\boldsymbol{\gamma})$
    5. $\mathbf{g} = -\nabla\mathcal{U}(\boldsymbol{\gamma}_{S_i})$
    6. $i = i + 1$
end while
return $S_K, \boldsymbol{\gamma}_{S_K}$
```

Since (4) is a monotone, non-negative, and α-weakly submodular maximization problem with cardinality constraint, the classical greedy method gives a constant-factor approximation guarantee of $F(S_K) \geq (1 - e^{-\alpha})\text{OPT}$ (Das & Kempe, 2018). Here, S_K is the solution returned by the greedy algorithm and OPT is the optimal objective of (4). The classical greedy algorithm begins with an empty set $S_0 = \emptyset$ and at each iteration i, it finds an element $u \in V \setminus S_{i-1}$ such that the marginal gain $F(u|S_{i-1})$ is maximized. Hence, in the context of solving (4), the classical greedy algorithm requires solving various instances of MMD-UOT $mnK - K(K-1)/2$ times. The classical greedy algorithm is detailed in Algorithm A3.

Since the function $-\mathcal{U}(\cdot)$ in the definition of $F(S)$ has RSC and RSM properties (Lemma A2.1), we propose to employ a computationally efficient orthogonal matching pursuit (OMP) based greedy algorithm (Elenberg et al., 2018; Gurumoorthy et al., 2019) for solving (4). A key feature of such strategies is that they greedily select the next element which maximally correlates with the residual of what has already been selected, i.e., choosing the element corresponding to the largest gradient value. In our case, this implies solving the MMD-UOT problem (1) for a given support set S (Appendix A3) and using its solution $\boldsymbol{\gamma}$ to compute the gradient $-\nabla\mathcal{U}(\boldsymbol{\gamma}_S)$ (6) for a candidate set of elements R.

In Algorithm 1, we propose a stochastic greedy algorithm for maximizing weakly submodular problems with cardinality constraint. It employs the above discussed OMP technique for greedy selection. We observe that Algorithm 1 requires solving the MMD-UOT problem (1) of size $|i|$ only once in each iteration i. Step 1 in Algorithm 1 corresponds to stochastic selection of the candidate set R_i for every iteration i (Mirzasoleiman et al., 2015). The vanilla non-stochastic OMP algorithm for maximizing weakly submodular problems with cardinality constraint (Gurumoorthy et al., 2019) is presented in Algorithm A4. Compared to its (non-stochastic) counterpart, Algorithm 1 is more efficient as the gradient (step 6) is computed only for a subset of the remaining elements. The approximation guarantee provided by Algorithm 1 is as follows.

Lemma 3.2. *Let $\{S_K, \boldsymbol{\gamma}_{S_K}\}$ be a solution returned by the proposed Algorithm 1, where S_K is the support of the transport plan $\boldsymbol{\gamma}_{S_K}$. Let S^* be an optimal solution of Problem (4). Then, $\mathbb{E}[F(S_K)] \geq (1 - e^{-u_{2K}/\tilde{U}_1} - \epsilon)F(S^*)$.*

The proof of Lemma 3.2 is discussed in Appendix A2.3.

3.2. Learning Column-wise Sparse Transport Plan

We now consider learning the transport plan $\boldsymbol{\gamma}$ with column-wise sparsity constraint, i.e., every column of $\boldsymbol{\gamma}$ has at most K_2 non-sparse entries. Such an OT approach is useful in learning a sparse mixture of experts (Liu et al., 2023). To this end, we consider solving (3) with $\mathcal{M} = \mathcal{M}_2$, i.e.,

$$\max_{S \in \mathcal{I}_2(\mathcal{M}_2)} F(S). \tag{5}$$

The partition matroid constraint ensures that (5) learns a transport plan in which each column has at most K_2 non-sparse entries. We term the proposed problem (5) as **ColSparseUOT**. The total number of non-sparse entries in the learned transport plan $\boldsymbol{\gamma}$ is $K = nK_2$. We note that (5) can alternatively learn row-sparse transport plans as well.

Algorithm 1 cannot be directly employed for solving (5) as its greedy selection does not respect the partition matroid constraint. Hence, we consider the residual randomized greedy approach for matroids (Chen et al., 2018), which provides a $(1 + 1/\alpha)^{-2}$ approximation guarantee for α-weakly submodular maximization subject to a general matroid constraint. However, it has a high computational cost as it requires solving multiple MMD-UOT instances in each iteration. We propose a novel OMP-based greedy algorithm, Algorithm 2, for efficiently maximizing weakly submodular problems with a general matroid constraint.

In each iteration i, Algorithm 2 selects a uniformly random element from the best maximal independent set (base) of $\mathcal{M}_2/S_{i-1}$. Here, $\mathcal{M}_2/S = \big(V \setminus S, \mathcal{I}_{\mathcal{M}_2/S}\big)$ denotes the contraction of $\mathcal{M}_2$ by S, which is a matroid on $V \setminus S$ consisting of independent sets $\mathcal{I}_{\mathcal{M}_2/S} := \{I \subseteq V \setminus S : I \cup S \in \mathcal{I}\}$. The gradient $\nabla\mathcal{U}(\boldsymbol{\gamma}_{S_i})$ is computed via by (6). It should be noted that in every iteration, the gradient needs to be computed only for the elements in $R = \{u : u \in I, I \in \mathcal{I}_{\mathcal{M}_2/S}\}$. The solution $\boldsymbol{\gamma}_S$ in step 4 is obtained efficiently using the APGD algorithm. It should be noted that step 1 of Algorithm 2 may not require a search over all possible maximal independent sets of $\mathcal{M}_2/S$. For partition matroids, step 1 essentially involves selecting the top-$(K_2 - |S \cap P_j|)$ elements with the largest (thresholded) gradient values from

Algorithm 2 OMP algorithm for maximizing weakly submodular problems with matroid constraint

```
**Input:** $\lambda_1, \lambda_2, \boldsymbol{\mu}, \boldsymbol{\nu}, \mathbf{C}, \mathbf{G}_1, \mathbf{G}_2$, per column sparsity level $K_2$.
$S_0 = \emptyset, \boldsymbol{\gamma}_{S_0} = \mathbf{0}, K = nK_2, \mathbf{g} = -\nabla\mathcal{U}(\boldsymbol{\gamma}_{S_0})$.
**for** $i = 1, \cdots, K$ **do**
    **1.** Let $M_i$ be a maximal independent set of $\mathcal{M}_2/S_{i-1}$ maximizing the sum $\sum_{u\in M_i}\max(0, \mathbf{g}_u)$.
    **2.** Let $u$ be a uniformly random element from $M_i$.
    **3.** $S_i = S_{i-1} \cup \{u\}$
    **4.** $\boldsymbol{\gamma}_{S_i} = \underset{\boldsymbol{\gamma}:\mathrm{supp}(\boldsymbol{\gamma})\in S_i, \boldsymbol{\gamma}\geq\mathbf{0}}{\arg\min} \mathcal{U}(\boldsymbol{\gamma})$
    **5.** $\mathbf{g} = -\nabla\mathcal{U}(\boldsymbol{\gamma}_{S_i})$
**end for**
**return** $S_K, \boldsymbol{\gamma}_{S_K}$
```

the set $P_j \setminus S$. The approximation guarantee provided by our proposed Algorithm 2 is as follows.

Lemma 3.3. *Let $\{S_K, \boldsymbol{\gamma}_{S_K}\}$ be the solution returned by our Algorithm 2, where S_K is the support of the transport plan $\boldsymbol{\gamma}_{S_K}$. Let S^* be an optimal solution of (5). Then,*

$$\mathbb{E}[F(S_K)] \geq F(S^*)\left(1 + \tilde{U}_1/u_{2K}\right)^{-2}.$$

Appendix A2.4 discusses the proof for Lemma 3.3.

3.3. Gradient Computation & Computational Cost

Gradient computation: The gradient $\nabla\mathcal{U}(\boldsymbol{\gamma})$ is employed in steps 4 and 5 of both the proposed Algorithms 1 & 2. The partial gradient expression is as follows:

$$\begin{aligned}\frac{\partial\mathcal{U}(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}_{ij}} = \mathbf{C}_{ij} &+ 2\lambda_1\left((\mathbf{G}_1)_i^\top(\boldsymbol{\gamma}\mathbf{1}) + (\mathbf{1}^\top\boldsymbol{\gamma})(\mathbf{G}_2)_j\right) \\ &- 2\lambda_1\left((\mathbf{G}_1)_i^\top\boldsymbol{\mu} + \boldsymbol{\nu}^\top(\mathbf{G}_2)_j\right) + \lambda_2\boldsymbol{\gamma}_{ij}.\end{aligned} \tag{6}$$

In (6), we observe that (a) the last term $(\mathbf{G}_1)_i^\top\boldsymbol{\mu} + \boldsymbol{\nu}^\top(\mathbf{G}_2)_j$ is independent of $\boldsymbol{\gamma}$ and can be precomputed, and (b) the terms involving the full matrix $\boldsymbol{\gamma}$ decouple in i and j. We leverage this structure for computing $\nabla\mathcal{U}(\boldsymbol{\gamma})$ efficiently.

Computational cost: We now discuss the per-iteration computational cost of both the proposed algorithms. For a given support set S, both Algorithms 1 & 2 involve solving the corresponding MMD-UOT problem to obtain the solution $\boldsymbol{\gamma}_S$. Let $R \subseteq V \setminus S$ be the set on which the gradient needs to be computed. The set S is updated via greedy selection (step 2 in Algorithm 1 or steps 1 & 2 in Algorithm 2) as $S \leftarrow S \cup \{u\}$, where $u \in R$ is the chosen element in the current iteration. The per-iteration cost of both the algorithms is $O(N + t \cdot M)$, where N is the cost of computing the gradient of candidate elements, t is the maximum iterations used for solving MMD-UOT using APGD, and M is the gradient cost in every APGD iteration. The above expression does not include the one-time cost of computing matrices $\mathbf{C}, \mathbf{G}_1, \mathbf{G}_2$ and vectors $\mathbf{G}_1\boldsymbol{\mu}, \mathbf{G}_2\boldsymbol{\nu}$.

Let $I_S = \{i \in [m] : (i,j) \in S\}$, $J_S = \{j \in [n] : (i,j) \in S\}$, $I_R = \{i \in [m] : (i,j) \in R\}$, and $J_R = \{j \in [n] : (i,j) \in R\}$. Then, $M = \mathcal{O}(|I_S|^2 + |J_S|^2 + |S|)$ and $N = \mathcal{O}(|I_S||I_R| + |J_S||J_R| + |S| + |R|)$. For both the algorithms, $1 \leq |I_S|, |I_R| \leq m$ and $1 \leq |J_S|, |J_R| \leq n$, where the value of these terms depend on S and R. For Algorithm 1, $|R| = mnK^{-1}\log(1/\epsilon)$ and for Algorithm 2 with partition matroid constraint, $2 \leq |R| \leq mn$.

3.4. Dual Analysis of (2) and (3)

In the previous sections, we analyzed (2) with $\mathcal{C} = \mathcal{C}_1$ or $\mathcal{C} = \mathcal{C}_2$ using discrete submodular maximization framework, developed Algorithms 1 & 2, and obtained corresponding approximation guarantees (Lemma 3.2 and Lemma 3.3). However, (2) may also be viewed in a continuous optimization setting. It has a convex objective but a non-convex and non-smooth constraint set. From this perspective, we now analyze a dual of the non-convex (2). While only weak duality holds in our setting, the duality gap analysis may still provide insights on the closeness to optimality.

Our next result details the primal-dual formulations corresponding to the proposed structured sparse optimal transport problem (2) with $\mathcal{C} = \mathcal{C}_2$. The expressions for (2) with $\mathcal{C} = \mathcal{C}_1$ can be derived likewise.

Lemma 3.4. *Problem (2) with $\mathcal{C} = \mathcal{C}_2$ and $\lambda_2 > 0$ may equivalently be written as:*

$$\begin{aligned}\min_{\boldsymbol{\gamma}\geq\mathbf{0}} P(\boldsymbol{\gamma})\big(&:= \langle\mathbf{C}, \boldsymbol{\gamma}\rangle + \textstyle\sum_{j=1}^n \Theta(\boldsymbol{\gamma}_j) \\ &+ \lambda_1(\|\boldsymbol{\gamma}\mathbf{1} - \boldsymbol{\mu}\|_{\mathbf{G}_1}^2 + \|\boldsymbol{\gamma}^\top\mathbf{1} - \boldsymbol{\nu}\|_{\mathbf{G}_2}^2)\big),\end{aligned} \tag{7}$$

where $\boldsymbol{\gamma}_j$ denotes the j^{th} column of $\boldsymbol{\gamma}$, $\Theta(\boldsymbol{\gamma}_j) = \frac{\lambda_2}{2}\|\boldsymbol{\gamma}_j\|^2 + \delta_{B_K}(\boldsymbol{\gamma}_j)$ and $B_K = \{\mathbf{z} \in \mathbb{R}_+^m : \|\mathbf{z}\|_0 \leq K\}$. Here, δ_B is the indicator function of a set B such that $\delta_B(\mathbf{z}) = 0$ if $\mathbf{z} \in B$, and $\delta_B(\mathbf{z}) = \infty$ otherwise. The following is a convex (weak) dual of the primal (7):

$$\begin{aligned}\max_{\boldsymbol{\alpha}\in\mathbb{R}^m, \boldsymbol{\beta}\in\mathbb{R}^n} D(\boldsymbol{\alpha}, \boldsymbol{\beta})\big(&:= \langle\boldsymbol{\alpha}, \boldsymbol{\mu}\rangle + \langle\boldsymbol{\beta}, \boldsymbol{\nu}\rangle - \tfrac{1}{4\lambda_1}\boldsymbol{\alpha}^\top\mathbf{G}_1^{-1}\boldsymbol{\alpha} \\ &- \tfrac{1}{4\lambda_1}\boldsymbol{\beta}^\top\mathbf{G}_2^{-1}\boldsymbol{\beta} - \textstyle\sum_{j=1}^n \Theta^*(\boldsymbol{\alpha} + \boldsymbol{\beta}_j\mathbf{1} - \mathbf{C}_j)\big),\end{aligned} \tag{8}$$

where $\mathbf{C}_j$ denote the j^{th} column of $\mathbf{C}$ and

$$\Theta^*(\mathbf{w}) = \max_{\mathbf{z}\in B_K}\langle\mathbf{w}, \mathbf{z}\rangle - \frac{\lambda_2}{2}\|\mathbf{z}\|^2. \tag{9}$$

The above result can be obtained using Lagrangian duality, and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the Lagrangian parameters corresponding to $\boldsymbol{\gamma}\mathbf{1} - \boldsymbol{\mu} = \mathbf{p}$ and $\boldsymbol{\gamma}^\top\mathbf{1} - \boldsymbol{\nu} = \mathbf{q}$ constraints, respectively, where $\mathbf{p}$ and $\mathbf{q}$ are auxiliary variables. To compute $\Theta^*(\mathbf{w})$, consider the permutation π on $[m]$ such that $\mathbf{w}_{\pi(i)} \geq \mathbf{w}_{\pi(i+1)}$ for $1 \leq i < m$. The solution is given by: $\mathbf{z}_{\pi(i)} = \max\left(0, \frac{\mathbf{w}_{\pi(i)}}{\lambda_2}\right)$ for $i \in [K]$, 0 otherwise, and $\Theta^*(\mathbf{w}) = \frac{1}{2\lambda_2}\sum_{i=1}^{K}(\max(0, \mathbf{w}_{\pi(i)}))^2$ (Liu et al., 2023).

For a feasible primal-dual pair $\{\boldsymbol{\gamma}_S, (\boldsymbol{\alpha}_S, \boldsymbol{\beta}_S)\}$ corresponding to (7) and (8), $\Delta(\boldsymbol{\gamma}_S, \boldsymbol{\alpha}_S, \boldsymbol{\beta}_S) = P(\boldsymbol{\gamma}_S) - D(\boldsymbol{\alpha}_S, \boldsymbol{\beta}_S)$ is the associated duality gap. However, $\Delta(\boldsymbol{\gamma}_S, \boldsymbol{\alpha}_S, \boldsymbol{\beta}_S)$ requires computing the dual candidate $(\boldsymbol{\alpha}_S, \boldsymbol{\beta}_S)$ for the given primal candidate $\{\boldsymbol{\gamma}_S\}$, which leads to our next result.

Proposition 3.5. *Let $\boldsymbol{\gamma}_S$ be a feasible primal candidate for (7), e.g., obtained from Algorithm 2 as (7) and (5) are equivalent problems. Then, the dual candidate corresponding to $\boldsymbol{\gamma}_S$ is $\boldsymbol{\alpha}_S = 2\lambda_1\mathbf{G}_1(\boldsymbol{\mu} - \boldsymbol{\gamma}\mathbf{1})$ and $\boldsymbol{\beta}_S = 2\lambda_1\mathbf{G}_2(\boldsymbol{\nu} - \boldsymbol{\gamma}^\top\mathbf{1})$.*

Proposition 3.5 provides concrete expressions for computing the duality gap Δ. While weak duality only guarantees $\Delta \geq 0$, computing Δ may still provide an estimate of how far a candidate solution could be from optimality. We present the proofs of Lemma 3.4 and Proposition 3.5 in Appendix A2.5.

4. Related Works

Since entropic-regularized OT (Cuturi, 2013) usually learns dense transport plan, Blondel et al. (2018) proposed an alternative ℓ_2-norm regularization for balanced OT and showed that it learns a sparse transport plan. While the degree of sparsity in ℓ_2-norm regularized OT depends on the magnitude of the regularization parameter, it cannot be explicitly controlled as desired in several applications. Hence, Liu et al. (2023) impose explicit column-wise sparsity constraints on the transport plan in the balanced ℓ_2-regularized OT problem. To solve their ℓ_2-regularized sparsity constrained balanced OT problem, henceforth termed as SCOT, Liu et al. (2023) propose a (semi-)dual relaxation of their primal formulation in the continuous optimization setting. SCOT uses gradient updates (LBFGS or ADAM solver) to solve the (semi-)dual and requires solving (9) at each iteration. We note that Alvarez-Melis et al. (2018) also leverages submodularity in the OT framework. In particular, they employ a submodular cost function.

In contrast, we propose to learn a general or column-wise sparse transport plan in the unbalanced optimal transport (UOT) setting. We pose these as equivalent (weakly) submodular maximization problems under matroid (uniform or partition) constraints. Overall, we develop efficient discrete greedy algorithms to solve the primal formulation (3) and present corresponding approximation guarantees (Lemmas 3.2 & 3.3). The equivalence between the discrete (3) and the continuous (2) problems allows us to derive a convex (weak) dual (8) of (3). While this dual analysis requires $\lambda_2 > 0$, Algorithms 1 & 2 (and Lemmas 3.2 & 3.3) also work with $\lambda_2 = 0$, i.e., no additional ℓ_2-norm regularization in (2). On the other hand, the presence of ℓ_2-norm regularizer is essential for SCOT (Liu et al., 2023).

5. Experimental Results

We evaluate the proposed approach in various applications. Experiments related to general sparse transport plans are discussed in Section 5.1, while those related to column-wise sparse transport plans are discussed in Section 5.2. Additional experimental results and details are presented in Appendix A5. Code can be downloaded from https://github.com/Piyushi-0/Sparse-UOT.

5.1. General Sparsity

We begin with experiments where learning a sparse transport plan is desired.

5.1.1. Designing Topology

Sparse process flexibility design (SPFD) aims to design a network topology that handles unpredictable demands of n products by matching them to the supplies from m plants. Designing a network topology requires adding edges between the nodes that facilitate the flow of goods. A recent work (Luo et al., 2023) models SPFD as an OT problem. While the supplies are predefined and can be modeled as $\boldsymbol{\mu} \in [0,\infty)^m$, the demands follow a given distribution ν. Hence, a set of demands $\{\boldsymbol{\nu}_i\}_{i=1}^{z}$ can be sampled from ν, i.e., $\boldsymbol{\nu}_i \in [0,\infty)^n \sim \nu$. Then, the SPFD problem may be defined as (Luo et al., 2023)

$$\max_{\{\boldsymbol{\gamma}_i \in \Gamma(\boldsymbol{\mu},\boldsymbol{\nu}_i)\}_{i=1}^{z}} \frac{1}{z}\sum_{i=1}^{z}\langle \mathbf{P}, \boldsymbol{\gamma}_i\rangle, \text{ s.t. } \left\|\sum_{i=1}^{z}\boldsymbol{\gamma}_i\right\|_0 \leq l, \quad (10)$$

where $\mathbf{P} \in \mathbb{R}^{m\times n}$ denotes the matrix of profits (negative of the cost matrix in the OT setting) and l is the total number of edges allowed in the network.

GSOT: Luo et al. (2023) propose a convex relaxation of the ℓ_0-norm constraint in (10) with a ℓ_1-norm regularizer and solve the resulting OT problem, termed as group sparse OT (GSOT), using an ADMM algorithm. Given a solution $\{\boldsymbol{\gamma}_{i,\mathrm{GSOT}}\}_{i=1}^{z}$ of the relaxed GSOT problem, the network topology may be obtained from the aggregate solution $\boldsymbol{\gamma}_{\mathrm{GSOT}} = \frac{1}{z}\sum_{i=1}^{z}\boldsymbol{\gamma}_{i,\mathrm{GSOT}}$. The profit created by the network is approximated as $\langle \mathbf{P}, \boldsymbol{\gamma}_{\mathrm{GSOT}}\rangle$. It should be noted that since the aggregate $\boldsymbol{\gamma}_{\mathrm{GSOT}}$ may not satisfy $\|\boldsymbol{\gamma}_{\mathrm{GSOT}}\|_0 \leq l$, the top-$l$ edges in $\boldsymbol{\gamma}_{\mathrm{GSOT}}$ which maximize the profit $\langle \mathbf{P}, \boldsymbol{\gamma}_{\mathrm{GSOT}}\rangle$ are selected as the network topology.

Proposed: We propose to model the SPFD problem (10) as the following UOT problem:

$$\frac{1}{z}\sum_{i=1}^{z}\max_{\boldsymbol{\gamma}_i \in \Gamma(\boldsymbol{\mu},\boldsymbol{\nu}_i)}\langle \mathbf{P}, \boldsymbol{\gamma}_i\rangle, \text{ s.t. } \|\boldsymbol{\gamma}_i\|_0 \leq \frac{l}{z}, \quad (11)$$

where we solve z independent instances of our GenSparseUOT problem (4). Thus, we employ the proposed

Table 1. Expected profit (higher is better) for SPFD experiment with varying network size constraint l. Proposed refers to our GenSparseUOT formulation (4) solved via Algorithm 1. The result is averaged over five random trials. We observe that our approach outperforms the GSOT baseline.

Method	$l = 100$	$l = 175$	$l = 250$
GSOT	0.014	0.031	0.044
Proposed ($\epsilon = 10^{-2}$)	0.166	0.224	$\mathbf{0.293}$
Proposed ($\epsilon = 10^{-3}$)	$\mathbf{0.167}$	0.238	0.286
Proposed ($\epsilon = 10^{-4}$)	0.147	$\mathbf{0.240}$	0.274

Algorithm 1 to solve (11). Let $\{\gamma_i^*\}_{i=1}^{z}$ be the obtained solution. The final network topology is obtained by selecting the top-l significant edges of $\mathbf{P} \odot \sum_i \gamma_i^*$, as discussed in the case of GSOT.

Experimental setup and results: Using the data generation process described in Luo et al. (2023), we generate the source and target datasets with $m = n = 100$ and $z = 20$. We compare the proposed GenSparseUOT approach only against GSOT, as other baselines such as SSOT (Blondel et al., 2018) and MMD-UOT (Manupriya et al., 2024b) do not incorporate sparsity constraint over transport plan. The hyperparameters of both GSOT and the proposed approach are tuned. Please refer to Appendix A5.2 for more details. In Table 1, we report the expected profit (averaged across five random trials) obtained by both the approaches with $l = \{100, 175, 250\}$, i.e., $\max\{m, n\} \leq l \leq \mathrm{round}(2.5 \max\{\mathrm{m}, \mathrm{n}\})$ (Luo et al., 2023). Our solution is obtained via Algorithm 1 and we report our performance with different stochastic greedy parameter ϵ. We observe that the proposed approach is significantly better than GSOT across varying network size constraints.

5.1.2. Monolingual Word Alignment

Aligning words in a (monolingual) sentence pair is an important sub-task in various natural language processing applications such as question answering, paraphrase identification, sentence fusion, and textual entailment recognition to name a few (MacCartney et al., 2008; Yao et al., 2013; Feldman & El-Yaniv, 2019; Brook Weiss et al., 2021). Recently, Arase et al. (2023) employed the OT machinery to align words between given two sentences. The sentences are represented as a histogram over words and the OT cost matrix is computed using contextualized word embeddings using a (pretrained) BERT-base-uncased model (Devlin et al., 2019). The learned OT plan represents the alignment. Arase et al. (2023) showed that existing OT variants perform at par with tailor-made word alignment techniques (Sabet et al., 2020).

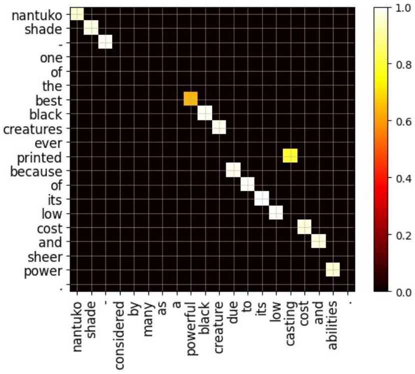
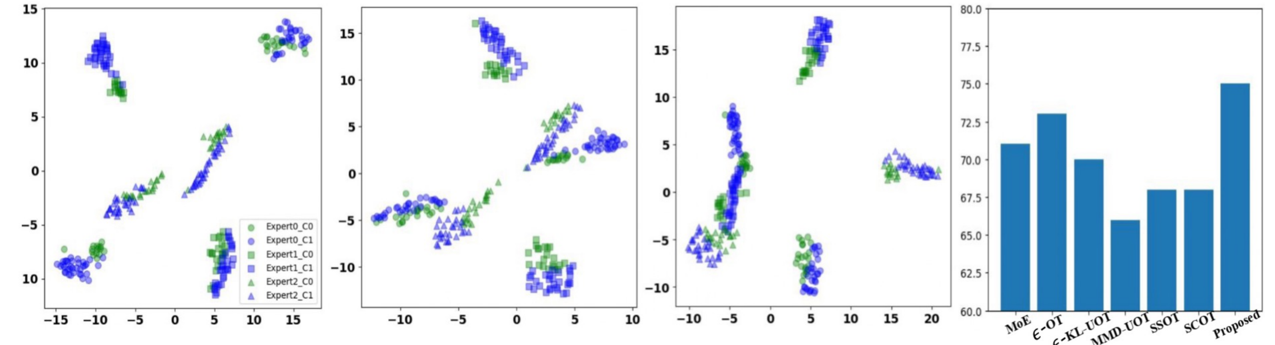


Figure 1. Example of a word alignment matrix obtained by our GenSparseUOT approach. Since the sentences convey similar information, most words in either sentences have a semantic counterpart, and our approach aligns them (almost) correctly. E.g., it correctly aligns 'powerful'↔'best' and 'abilities'↔'power' and (correctly) does not map 'powerful'↔'power' even though this pair is semantically close. Words without a semantic counterpart are left unaligned (null alignment).

It should be noted that words in one sentence may lack semantic counterparts in the other sentence, especially when the sentences convey different meanings. Such words correspond to *null* alignments. Identifying null alignments is essential because it helps us reason about the semantic similarity between sentences by highlighting information inequality. This motivates the need of learning sparsity constrained unbalanced transport plan for such a task and we evaluate the suitability of our GenSpareUOT approach (4) for this problem. Figure 1 illustrates a word alignment matrix learned by our approach for a given pair of (semantically similar) sentences.

Experimental setup and results: We follow the experimental setup described in (Arase et al., 2023). The evaluation is performed on the aligned Wikipedia sentences in an unsupervised setting with the 'sure' alignments, i.e., with the alignments agreed upon by multiple annotators (Arase et al., 2023). Since the number of words in the input sentences is usually small, we solve GenSpareUOT (4) via Algorithm A4 (which is the non-stochastic variant of Algorithm 1) and compare it against the OT baselines BOT, POT, KL-UOT studied by Arase et al. (2023), SSOT (Blondel et al., 2018), and MMD-UOT (Manupriya et al., 2024b). The hyperparameters of all methods are tuned. Please refer to Appendix A5.3 for more details.

Table 2 reports the accuracy and the F1 scores corresponding to matching the null and the total (null + non-null) assignments. We see that the proposed approach is at par or better than the OT baselines studied by Arase et al. (2023). On the other hand, our approach outperforms MMD-UOT and the sparse OT approach SSOT. The corresponding precision and recall scores are detailed in Table A7.



(a) Proposed (b) SCOT (c) Vanilla MoE (d) Accuracy

Figure 2. (a)-(c) t-SNE mappings of the experts learned by different approaches. ‘Expert$i$_Cj’ denotes the embeddings learnt by expert i for samples belonging to class j. The embeddings learned with the proposed approach not only distinguish the instances from the two classes but also exhibit more diversity in the knowledge acquired by every expert. (d) The accuracy obtained on the test set.

Table 2. F1 and accuracy (Acc.) scores on the test split of the Wiki dataset. The scores are reported for both null and total alignments. Higher scores are better. The proposed approach is at par with the best performing baseline.

	Null		Total	
Method	Acc.	F1	Acc.	F1
BOT	48.95	**80.05**	47.05	**94.96**
POT	37.07	72.48	34.32	94.15
KL-UOT	44.68	78.71	42.02	94.63
MMD-UOT	41.35	75.92	37.74	93.14
SSOT	16.54	29.40	12.74	64.13
Proposed	**49.14**	79.92	**48.00**	94.79

5.2. Column-wise Sparse Transport Plan

Mixture-of-Expert (MoE) (Jacobs et al., 1991; Jordan & Jacobs, 1993; Eigen et al., 2014) is a popular architecture that helps scale up model capacity with relatively small computational overhead. MoE consists of m experts, which are neural networks with identical architecture, trained with a gating function (often, a shallow neural network) that routes inputs to a chosen subset of the experts. Shazeer et al. (2017) demonstrated the utility of a sparsely-gated mixture of experts (SMoE) that selects only the top-K_2 experts for processing the input, where $1 \leq K_2 < m$. A key motivation behind MoE/SMoE is that a complex problem may be solved by a combination of experts, each specializing on different sub-problem(s).

Given an input $\mathbf{x}$, the output of SMoE is given by $\text{SMoE}(\mathbf{x}) = \sum_{r=1}^{m} \text{Gate}_r(\mathbf{x}) E_r(\mathbf{x})$, where $\text{Gate}_r : \mathbb{R}^d \mapsto \mathbb{R}_+$ is the sparse gating function and $\{E_r\}_{r=1}^m$ are the experts. Clark et al. (2022) proposed an entropy-regularized OT based gating function with the aim of achieving a more balanced assignment across experts. For instance, load balancing becomes crucial in distributed systems. Recently, Liu et al. (2023) employed their SCOT method in the SMoE application, where the goal is to map each input in a batch of size n to top-K_2 (out of m) experts. In the following, we illustrate the utility of the proposed ColSparseUOT approach (5), solved via Algorithm 2, in the SMoE setting.

Toy dataset. We begin with the classification task on a toy binary dataset (with random train/test split). We train an SMoE with three (shallow) experts and a top-2 gating function using various approaches. The architectural and training details are provided in Appendix A5.4. We first qualitatively assess the latent representations learned by (vanilla) MoE (Shazeer et al., 2017), SCOT (Liu et al., 2023), and the proposed ColSparseUOT approaches. In Figure 2 we show their 2-D t-SNE visualizations (van der Maaten & Hinton, 2008) on the toy dataset. Figure 2(a) reveals that the proposed approach's experts not only distinguish the two classes effectively but also demonstrate variety in the knowledge acquired by each expert. This is because the t-SNE maps the proposed approach's experts to well-separated locations on the 2-D plane. On the other hand, t-SNE maps the experts learned by SCOT and (vanilla) MoE approaches to overlapping/nearby regions in their respective plots shown in Figures 2(b) & 2(c). We also compare the performance of the learned SMoE models on the test split. In Figure 2(d), we report the accuracy of our proposed approach, (vanilla) MoE, SCOT, and other SMoE baselines in which the (top-2) gating function is based on entropy-regularized OT (ϵOT), entropy-regularized KL-UOT (ϵKL-UOT), SSOT, and MMD-UOT. We observe that our method obtains the highest accuracy.

CIFAR dataset. We next focus on the binary classification problem of identifying whether a given image belongs to the CIFAR-10 dataset or the CIFAR-10-rotate dataset (Chen et al., 2022). CIFAR-10-rotate consists of CIFAR-10 images, rotated by 30 degrees. For SMoE, we consider four ResNet18 experts (He et al., 2016) and train SMoEs with the

Table 3. Accuracy obtained on SMoE experiment along with the number of inputs allocated to each expert. We observe that the proposed approach obtains the best generalization performance with balanced allocation across experts.

Method	Acc.	Exp 1	Exp 2	Exp 3	Exp 4
Top-1 MoE$_{\text{default}}$	95.91	0	10962	9038	0
Top-1 MoE$_{\text{balanced}}$	93.74	4953	5018	4779	5250
SCOT$_{\lambda=0.1}$	77.96	3341	1895	11119	3645
SCOT$_{\lambda=10}$	90.56	1929	6112	6113	5846
SCOT$_{\lambda=1000}$	56.48	0	7678	7788	4534
Proposed$_{\lambda_1=0.1}$	95.56	5435	4854	4977	4734
Proposed$_{\lambda_1=10}$	85.18	5000	5002	4998	5000
Proposed$_{\lambda_1=1000}$	90.54	5000	5000	5000	5000

Table 4. Duality gap (Δ) comparison for solving (7) with various hyperparameters. Lower duality gap is better. We observe that our approach obtains significantly lower duality gap than SCOT.

λ_1	λ_2	Proposed solver		SCOT solver	
		Primal obj.	Δ	Primal obj.	Δ
0.1	0.1	**0.02993**	$< \mathbf{10^{-10}}$	0.03169	0.00232
1	0.1	**0.09183**	**0.01911**	0.27172	0.19111
10	0.1	**0.11682**	**0.64896**	2.30889	2.21029
0.1	1	**0.03036**	$< \mathbf{10^{-10}}$	0.03116	0.00114
1	1	**0.09409**	**0.00286**	0.10371	0.01216
10	1	**0.11897**	**0.05468**	0.32334	0.21289

gating network based on: (a) top-1 linear activation (Chen et al., 2022), (b) SCOT, and (c) the proposed ColSparseUOT (5). In both SCOT and ColSparseUOT, a sparse transport plan is learned between the $m = 4$ experts and a given batch of n inputs with the goal of mapping each input with only one expert ($K_2 = 1$).

In Table 3, we report the performance of all the three approaches. Since load balancing is an important aspect in MoE setup, we also report the corresponding number of inputs assigned to every expert during the inference stage. However, balanced allocation may not be achieved by the SMoEs by default. Hence, we report their results with different hyperparameters values. Please refer to Appendix A5.4 for more details on the experimental setup. From the results we observe that our approach obtains a good generalization performance with balanced assignments across hyperparameters. While SCOT obtains a reasonable accuracy in one case (with $\lambda = 10$), its load balancing is skewed. For the non-OT based Top-1 MoE basline, its default setting obtains a heavily skewed allocation with two experts never getting used. In a more balanced configuration, it suffers a small accuracy drop. Overall, we see that our method is well suited for SMoE setting from both the generalization and load balancing points of view.

5.3. Duality Gap Comparison

In this section, we compare the optimization quality of the proposed Algorithm 2 and the SCOT algorithm (Liu et al., 2023) in solving our sparse UOT problem with the column-wise sparsity constraint (7). In Section 3.2, we propose to solve (7) in discrete optimization setting, via an equivalent reformulation (3). It should be noted that Liu et al. (2023) study a ℓ_2-regularized *balanced* OT problem with column-wise sparsity constraint. They propose a gradient descent based algorithm (e.g., LBFGS) with sparse projections to optimize a dual relaxation of their primal problem. We use their algorithm to solve (8), which is a dual of (7). While Algorithm 2 learns a primal solution $\boldsymbol{\gamma}_1$ of (7), the corresponding dual solution $\{\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1\}$ can be obtained via Proposition 3.5. SCOT, on the other hand, obtains a dual solution $\{\boldsymbol{\alpha}_2, \boldsymbol{\beta}_2\}$ of (8) and then obtains the corresponding primal solution $\boldsymbol{\gamma}_2$ by solving the sparse projection problem (9). Hence, we can compute and compare the duality gap $\Delta(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = P(\boldsymbol{\gamma}) - D(\boldsymbol{\alpha}, \boldsymbol{\beta})$ associated with the solutions obtained by both the algorithms.

Experimental setup and results: The source and target measures are taken to be the empirical measures over two randomly chosen 100-sized batches of CIFAR-10. We compare the duality gap over a range of hyperparameter (λ_1, λ_2) values and different kernels employed for the MMD computation in (2). The kernel hyperparameters are fixed according to the median heuristics (Gretton et al., 2012).

Table 4 reports the results with a inverse multiquadratic kernel (Sriperumbudur et al., 2011). We observe that our approach outperforms SCOT by obtaining at least three times lower duality gap. In a couple of cases, the duality gap associated with our Algorithm 2 is $< 10^{-10}$, signifying that it has converged at (or very close to) a global optimum. Additional results are discussed in Appendix A5.5.

6. Conclusion

In this work we proposed sparsity-constrained unbalanced OT formulations and presented an interesting viewpoint of the problem as that of maximization of a weakly submodular function over a uniform or partition matroid. To this end, we propose novel greedy algorithms having attractive approximation guarantees. A duality gap analysis further provides an empirical way of validating the optimality of our greedy solution. Experiments across different applications shows the efficacy of the proposed approach. At a conceptual level, our work shows a novel connection between OT and submodularity. A future work could be to expand on the variants of structured sparsity patterns in the OT plan.

Acknowledgements

PM acknowledges the support of the Google PhD Fellowship.

Impact Statement

This paper tries to advance the field of optimal transport and its applications to machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Alvarez-Melis, D. and Jaakkola, T. Gromov-Wasserstein alignment of word embedding spaces. In *EMNLP*, 2018.

Alvarez-Melis, D., Jaakkola, T., and Jegelka, S. Structured optimal transport. In *AISTATS*, 2018.

Arase, Y., Bao, H., and Yokoi, S. Unbalanced optimal transport for unbalanced word alignment. In *ACL*, 2023.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *ICML*, 2017.

Balaji, Y., Chellappa, R., and Feizi, S. Robust optimal transport with applications in generative modeling and domain adaptation. In *NeurIPS*, 2020.

Blondel, M., Seguy, V., and Rolet, A. Smooth and sparse optimal transport. In *AISTATS*, 2018.

Brook Weiss, D., Roit, P., Klein, A., Ernst, O., and Dagan, I. QA-align: Representing cross-text content overlap by aligning question-answer propositions. In *EMNLP*, 2021.

Chen, L., Feldman, M., and Karbasi, A. Weakly submodular maximization beyond cardinality constraints: Does randomization help greedy? In *ICML*, 2018.

Chen, Z., Deng, Y., Wu, Y., Gu, Q., and Li, Y. Towards understanding the mixture-of-experts layer in deep learning. In *NeurIPS*, 2022.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. Scaling algorithms for unbalanced optimal transport problems. *Math. Comput.*, 87:2563–2609, 2017.

Clark, A., Casas, D. d. l., Guy, A., Mensch, A., Paganini, M., Hoffmann, J., Damoc, B., Hechtman, B., Cai, T., Borgeaud, S., Driessche, G. v. d., Rutherford, E., Hennigan, T., Johnson, M., Millican, K., Cassirer, A., Jones, C., Buchatskaya, E., Budden, D., Sifre, L., Osindero, S., Vinyals, O., Rae, J., Elsen, E., Kavukcuoglu, K., and Simonyan, K. Unified scaling laws for routed language models. In *ICML*, 2022.

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

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013.

Das, A. and Kempe, D. Approximate submodularity and its applications: subset selection, sparse approximation and dictionary selection. *Journal of Machine Learning Research*, 19(1):74–107, jan 2018. ISSN 1532-4435.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

Dwivedi, R. and Mackey, L. Generalized kernel thinning. In *ICLR*, 2022.

Eigen, D., Ranzato, M., and Sutskever, I. Learning factored representations in a deep mixture of experts. In *ICLR, Workshop Track Proceedings*, 2014.

Elenberg, E. R., Khanna, R., Dimakis, A. G., and Negahban, S. Restricted strong convexity implies weak submodularity. *The Annals of Statistics*, 46(6B):3539 – 3568, 2018. doi: 10.1214/17-AOS1679.

Essid, M. and Solomon, J. Quadratically regularized optimal transport on graphs. *SIAM Journal on Scientific Computing*, 40(4):A1961–A1986, 2018.

Fatras, K., Zine, Y., Flamary, R., Gribonval, R., and Courty, N. Learning with minibatch wasserstein: asymptotic and gradient properties. In *AISTATS*, 2020.

Fatras, K., Séjourné, T., Courty, N., and Flamary, R. Unbalanced minibatch optimal transport; applications to domain adaptation. In *ICML*, 2021.

Feldman, Y. and El-Yaniv, R. Multi-hop paragraph retrieval for open-domain question answering. In *ACL*, 2019.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouve, A., and Peyré, G. Interpolating between optimal transport and mmd using sinkhorn divergences. In *AISTATS*, 2019.

Frogner, C., Zhang, C., Mobahi, H., Araya-Polo, M., and Poggio, T. Learning with a wasserstein loss. In *NIPS*, 2015.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Gurumoorthy, K., Jawanpuria, P., and Mishra, B. SPOT: A framework for selection of prototypes using optimal transport. In *ECML*, 2021.

Gurumoorthy, K. S., Dhurandhar, A., Cecchi, G., and Aggarwal, C. Efficient data representation by selecting prototypes with importance weights. In *IEEE ICDM*, 2019.

Han, A., Mishra, B., Jawanpuria, P., and Gao, J. Riemannian block SPD coupling manifold and its application to optimal transport. *Machine Learning Journal*, 113(4): 1595–1622, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.

Jawanpuria, P., Meghwanshi, M., and Mishra, B. Geometry-aware domain adaptation for unsupervised alignment of word embeddings. In *Annual Meeting of the Association for Computational Linguistics*, 2020.

Jawanpuria, P., Satya Dev, N. T. V., and Mishra, B. Efficient robust optimal transport: formulations and algorithms. In *IEEE Conference on Decision and Control*, 2021.

Jitkrittum, W., Sangkloy, P., Gondal, M. W., Raj, A., Hays, J., and Schölkopf, B. Kernel mean matching for content addressability of GANs. In *ICML*, 2019.

Jordan, M. and Jacobs, R. Hierarchical mixtures of experts and the em algorithm. In *Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan)*, volume 2, pp. 1339–1344 vol.2, 1993.

Kantorovich, L. On the transfer of masses (in russian). *Doklady Akademii Nauk*, 37(2):227–229, 1942.

Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Poczos, B. Mmd gan: Towards deeper understanding of moment matching network. In *NIPS*, 2017.

Liero, M., Mielke, A., and Savaré, G. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.

Liu, T., Puigcerver, J., and Blondel, M. Sparsity-constrained optimal transport. In *ICLR*, 2023.

Luo, D., Yu, T., and Xu, H. Group sparse optimal transport for sparse process flexibility design. In *IJCAI*, 2023. AI for Good.

MacCartney, B., Galley, M., and Manning, C. D. A phrase-based alignment model for natural language inference. In *EMNLP*, 2008.

Manupriya, P., Das, R. K., Biswas, S., and Jagarlapudi, S. N. Consistent optimal transport with empirical conditional measures. In *AISTATS*, 2024a.

Manupriya, P., Nath, J. S., and Jawanpuria, P. Mmd-regularized unbalanced optimal transport. *Transactions on Machine Learning Research*, 2024b.

Mirzasoleiman, B., Badanidiyuru, A., Karbasi, A., Vondrák, J., and Krause, A. Lazier than lazy greedy. In *AAAI*, 2015.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1–2):1–141, 2017.

Muzellec, B., Nock, R., Patrini, G., and Nielsen, F. Tsallis regularized optimal transport and ecological inference. In *AAAI*, 2017.

Nath, J. S. and Jawanpuria, P. K. Statistical optimal transport posed as learning kernel embedding. In *NeurIPS*, 2020.

Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. An analysis of approximations for maximizing submodular set functions - I. *Math. Program.*, 14(1), 1978.

Nguyen, K., Nguyen, D., Nguyen, Q., Pham, T., Bui, H., Phung, D., Le, T., and Ho, N. On transportation of mini-batches: A hierarchical approach. In *ICML*, 2022.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Sabet, M. J., Dufter, P., and Schütze, H. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. In *EMNLP*, 2020.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., Lee, L., Chen, J., Brumbaugh, J., Rigollet, P., Hochedlinger, K., Jaenisch, R., Regev, A., and Lander, E. S. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943.e22, 2019.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017.

Solomon, J., de Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics*, 34(4):66:1–66:11, 2015.

Song, L. Learning via hilbert space embedding of distributions, 2008.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. G. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12: 2389–2410, 2011.

Swanson, K., Yu, L., and Lei, T. Rationalizing text matching: Learning sparse alignments via optimal transport. In *ACL*, 2020.

van der Maaten, L. and Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605, 2008.

Yao, X., Van Durme, B., Callison-Burch, C., and Clark, P. A lightweight and high performance monolingual word aligner. In *ACL*, 2013.

A1. Background

A1.1. Weak Submodularity

Das & Kempe (2018) defined the notion of approximate submodularity governed by a submodularity ratio. For a monotone function F the submodularity ratio, w.r.t. a set S and a parameter $K \geq 1$, is defined as follows. $\alpha_{L,K}(F) = \min_{S \subseteq L, A:|A| \leq K, A \cap S = \phi} \frac{\sum_{u \in A} F(S \cup \{u\}) - F(S)}{F(S \cup A) - F(S)}$, with $0/0 := 1$. F is submodular iff $\alpha_{S,K} \geq 1$. If the ratio $\alpha \equiv \frac{\sum_{u \in A} F(S \cup \{u\}) - F(S)}{F(S \cup A) - F(S)}$ is greater than 0 and not necessarily greater than 1, then F is α-weakly submodular.

A1.2. Characteristic Kernel, Universal Kernel, and Maximum Mean Discrepancy (MMD)

We have the following assumption on the kernel corresponding to the MMD regularization in (3):

Assumption A1.1. The kernel k corresponding to the MMD regularizations in (3) is bounded and universal.

In the following, we briefly discuss the above concepts.

Boundedness: A kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is said to be bounded if $k(\mathbf{x}, \mathbf{y}) < \infty, \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$. In the continuous domain, examples of bounded kernels include the RBF (Gaussian) kernel or the IMQ (inverse multiquadratic) kernel.

Kernel mean embeddings: Let $\phi(\cdot)$ and $\mathcal{H}$ be the canonical feature map and the canonical reproducing kernel Hilbert space (RKHS) corresponding to the kernel k. The kernel mean embedding (Muandet et al., 2017) of a random variable $X \sim \mathcal{P}$ is defined as $\psi_{\mathcal{P}} := \mathbb{E}_{X \sim \mathcal{P}}[\phi(X)]$. If the kernel k is bounded, then $\psi_X \in \mathcal{H}$ and is well defined.

Characteristic and universal kernels: Characteristic kernels (Sriperumbudur et al., 2011) are those for which the map $\mathcal{P} \mapsto \psi_{\mathcal{P}}$ is injective (one-to-one). A kernel defined over a domain $\mathcal{X}$ is universal if and only if its RKHS is dense in the set of all continuous functions over $\mathcal{X}$. All universal kernels are also characteristic kernels (over their respective domains). Examples of universal kernels include the Kronecker delta kernel for discrete measures, the Gaussian kernel for continuous measures, the IMQ kernel for continuous measures, etc.

Maximum mean discrepancy (MMD): Given a characteristic kernel k, and distributions μ and ν, the MMD metric between μ and ν is defined as (Gretton et al., 2012)

$$\mathrm{MMD}_k(p, q) = \|\psi_\mu - \psi_\nu\|_{\mathcal{H}} = \max_{f : \|f\|_{\mathcal{H}} \leq 1} \langle f, \psi_\mu \rangle_{\mathcal{H}} - \langle f, \psi_\nu \rangle_{\mathcal{H}}, \tag{A12}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\| \cdot \|_{\mathcal{H}}$ denote the RKHS inner product and RKHS norm corresponding to the kernel k, respectively.

A2. Proofs on the theoretical results presented in the main paper

A2.1. A few useful properties of $F(S)$ as defined in (3)

Let the function $F(S)$ be as defined in (3), i.e., $F(S) := \mathcal{U}(\mathbf{0}) - \min_{\boldsymbol{\gamma}:\text{supp}(\boldsymbol{\gamma}) \subseteq S, \boldsymbol{\gamma} \geq \mathbf{0}} \mathcal{U}(\boldsymbol{\gamma}) = \mathcal{U}(\mathbf{0}) - \mathcal{U}(\boldsymbol{\gamma}_S)$. Here, $\mathcal{U} : \mathbb{R}_+^{m \times n} \mapsto \mathbb{R}_+$ and $\boldsymbol{\gamma}_S$ denotes an optimal solution of the convex MMD-UOT problem (1) with a given fixed support S. In the following, we first prove that the function $-\mathcal{U}(\cdot)$ has a finite RSC and RSM parameters (Section 2.3) and then use this property to prove a couple of results corresponding to $F(S)$.

Lemma A2.1. *$-\mathcal{U}(\cdot)$ has a finite restricted strong concavity (RSC) parameter (u_Ω) and a finite restricted smoothness (RSM) parameter (U_Ω) whenever the employed kernel function k is universal.*

Proof. We first prove that $-\mathcal{U}(.)$ has a finite RSC, RSM parameters for the case $\lambda_2 = 0$. Given $\boldsymbol{\gamma}, \boldsymbol{\gamma}' \in \mathbb{R}^{m \times n}$, we have the following result

$$\begin{aligned} & -\left(\mathcal{U}(\boldsymbol{\gamma}) - \mathcal{U}(\boldsymbol{\gamma}') - \langle \nabla \mathcal{U}(\boldsymbol{\gamma}), \boldsymbol{\gamma} - \boldsymbol{\gamma}' \rangle\right) \\ & \qquad = \lambda_1 \left((\boldsymbol{\gamma}\mathbf{1}_n - \boldsymbol{\gamma}'\mathbf{1}_n)^\top \mathbf{G}_1 (\boldsymbol{\gamma}\mathbf{1}_n - \boldsymbol{\gamma}'\mathbf{1}_n) + (\boldsymbol{\gamma}^\top \mathbf{1}_m - \boldsymbol{\gamma}'^\top \mathbf{1}_m)^\top \mathbf{G}_2 (\boldsymbol{\gamma}^\top \mathbf{1}_m - \boldsymbol{\gamma}'^\top \mathbf{1}_m) \right) \\ & \qquad = \lambda_1 \left(\mathrm{Tr}\left((\boldsymbol{\gamma} - \boldsymbol{\gamma}')^\top \mathbf{G}_1 (\boldsymbol{\gamma} - \boldsymbol{\gamma}') \mathbf{1}_\mathrm{n} \mathbf{1}_\mathrm{n}^\top \right) + \mathrm{Tr}\left((\boldsymbol{\gamma}^\top - \boldsymbol{\gamma}'^\top)^\top \mathbf{G}_2 (\boldsymbol{\gamma}^\top - \boldsymbol{\gamma}'^\top) \mathbf{1}_\mathrm{m} \mathbf{1}_\mathrm{m}^\top \right) \right). \end{aligned} \tag{A13}$$

where the function $\mathcal{U}(\boldsymbol{\gamma})$ and its gradient are defined in (1) and (6). $\mathrm{Tr}(\cdot)$ denotes the trace operator.

Let $e_0^1,\ e_1^1$ denote the minimum and maximum eigenvalues of $\mathbf{G}_1$. Let $e_0^2,\ e_1^2$ denote the minimum and maximum eigenvalues of $\mathbf{G}_2$. Then (A13) implies, $\lambda_1(e_0^1 n + e_0^2 m) \leq\ -\left(\mathcal{U}(\boldsymbol{\gamma}) - \mathcal{U}(\boldsymbol{\gamma}') - \langle \nabla\mathcal{U}(\boldsymbol{\gamma}), \boldsymbol{\gamma}' - \boldsymbol{\gamma}\rangle\right)\ \leq \lambda_1(e_1^1 n + e_1^2 m)$. Thus, the RSC constant becomes $u_\Omega = \lambda_1(e_0^1 n + e_0^2 m)$ and the RSM constant becomes $U_\Omega = \lambda_1(e_1^1 n + e_1^2 m)$. We recall that all characteristic kernels are universal (A1.2). We use that the gram matrices of universal kernels are full-rank (Song, 2008, Corollary 32). Hence, with a characteristic kernel (like the Gaussian kernel or the inverse multi-quadratic kernel), u_Ω and $U_\Omega > 0$. Invoking the Gershgorin circle theorem, the maximum eigenvalue of the gram matrices can be upper-bounded by the maximum row sum, which is finite for bounded kernels (like the Gaussian kernel or the inverse multi-quadratic kernel). We have that $0 < u_\Omega \leq U_\Omega < \infty$.

It can be easily seen that, when $\lambda_2 > 0$, the RSM constant becomes $\lambda_1(e_1^1 n + e_1^2 m) + \lambda_2/2$ and the RSC constant becomes $\lambda_1(e_0^1 n + e_0^2 m) + \lambda_2/2$. □

Lemma A2.2. *$F(S \cup \{u\}) - F(S) \geq \frac{1}{2\tilde{U}_1}\left(\mathbf{g}_u^+(\boldsymbol{\gamma}_S)\right)^2$, where $\mathbf{g}_u^+(.) \equiv \max\{-\nabla\mathcal{U}_u(.),\ 0\}$.*

Proof. Let $\mathbf{1}^{\{u\}} \in \mathbb{R}^{m\times n}$ denote a matrix of zeros with 1 at the index given by $\{u\}$. Let $\mathbf{y}^{\{u\}} \equiv \boldsymbol{\gamma}_S + \eta\mathbf{1}^{\{u\}}$ for some $\eta \geq 0$.

$$\begin{aligned} F(S \cup \{u\}) - F(S) &= -\mathcal{U}(\boldsymbol{\gamma}_{S\cup\{u\}}) + \mathcal{U}(\boldsymbol{\gamma}_S) \\ &\geq -\mathcal{U}(\mathbf{y}^{\{u\}}) + \mathcal{U}(\boldsymbol{\gamma}_S) \\ &\geq \langle -\nabla\mathcal{U}(\boldsymbol{\gamma}_S), \eta\mathbf{1}^{\{u\}}\rangle - \frac{\tilde{U}_1}{2}\eta^2. \end{aligned}$$

On maximizing wrt $\eta \geq 0$, we get $F(S\cup\{u\}) - F(S) \geq \frac{1}{2\tilde{U}_1}\left(\mathbf{g}_u^+(\boldsymbol{\gamma}_S)\right)^2$ $\left(\text{when } \eta = \frac{\mathbf{g}_u^+(\boldsymbol{\gamma}_S)}{\tilde{U}_1}\right)$. □

Lemma A2.3. *$F(S \cup A) - F(S) \leq \frac{1}{2u_{\bar{m}}}\sum_{u\in A}\left(\mathbf{g}_u^+(\boldsymbol{\gamma}_S)\right)^2$, where $\mathbf{g}_u^+(.) \equiv \max\{-\nabla\mathcal{U}_u(.),\ 0\}$ and $\bar{m} = |S| + |A|$.*

Proof. As $\boldsymbol{\gamma}_S,\ \boldsymbol{\gamma}_{S\cup A}$ are the minimizes, $\mathcal{U}(\mathbf{0}_{m\times n}) - \mathcal{U}(\boldsymbol{\gamma}_S) = F(S)$ and $\mathcal{U}(\mathbf{0}_{m\times n}) - \mathcal{U}(\boldsymbol{\gamma}_{S\cup A}) = F(A\cup S)$. We now upper-bound, $F(S\cup A) - F(S) = \mathcal{U}(\boldsymbol{\gamma}_S) - \mathcal{U}(\boldsymbol{\gamma}_{S\cup A})$.

Using the RSC and RSM constants of $-\mathcal{U}$ (A2.1), we have the following.

$$\begin{aligned} &\frac{u_{\bar{m}}}{2}\|\boldsymbol{\gamma}_{S\cup A} - \boldsymbol{\gamma}_S\|^2 \leq -\mathcal{U}(\boldsymbol{\gamma}_S) + \mathcal{U}(\boldsymbol{\gamma}_{S\cup A}) + \langle -\nabla\mathcal{U}(\boldsymbol{\gamma}_S), \boldsymbol{\gamma}_{S\cup A} - \boldsymbol{\gamma}_S\rangle \\ \implies &0 \leq -\mathcal{U}(\boldsymbol{\gamma}_{S\cup A}) + \mathcal{U}(\boldsymbol{\gamma}_S) \leq \langle -\nabla\mathcal{U}(\boldsymbol{\gamma}_S), \boldsymbol{\gamma}_{S\cup A} - \boldsymbol{\gamma}_S\rangle - \frac{u_{\bar{m}}}{2}\|\boldsymbol{\gamma}_{S\cup A} - \boldsymbol{\gamma}_S\|^2 \\ &\leq \max_{\mathbf{W}:\mathbf{W}\geq\mathbf{0}_{m\times n}, \mathbf{W}_{V\setminus(S\cup A)}=\mathbf{0}} \langle -\nabla\mathcal{U}(\boldsymbol{\gamma}_S), \mathbf{W} - \boldsymbol{\gamma}_S\rangle - \frac{u_{\bar{m}}}{2}\|\mathbf{W} - \boldsymbol{\gamma}_S\|^2. \end{aligned} \tag{A14}$$

The matrix $\mathbf{W}^*$ that attains the maximum is described as follows. $\mathbf{W}^*_{S\cup A} = \max\left\{\frac{1}{u_{\bar{m}}}(-\nabla\mathcal{U}_{S\cup A}(\boldsymbol{\gamma}_S)) + (\boldsymbol{\gamma}_S)_{S\cup A}, \mathbf{0}\right\}$. Now, from the KKT conditions, we have that $\forall j \in S$,

$$(\boldsymbol{\gamma}_S)_j > 0 \implies -\nabla\mathcal{U}_j(\boldsymbol{\gamma}_S) = 0 \text{ and } (\boldsymbol{\gamma}_S)_j = 0 \implies -\nabla\mathcal{U}_j(\boldsymbol{\gamma}_S) \leq 0.$$

Hence, $(\mathbf{W}^* - \boldsymbol{\gamma}_S)_j = 0\ \forall j \in S$. Also, $(\boldsymbol{\gamma}_S)_j = 0\ \forall j \in A$. Thus, $\forall j \in A, (\mathbf{W}^* - \boldsymbol{\gamma}_S)_j = \max\left\{\frac{1}{u_{\bar{m}}}\left(-\nabla\mathcal{U}_j(\boldsymbol{\gamma}_S)\right), 0\right\}$. Using this in (A14), we get

$$F(S\cup A) - F(S) \leq \frac{1}{2u_{\bar{m}}}\sum_{u\in A}\left(\mathbf{g}_u^+(\boldsymbol{\gamma}_S)\right)^2,$$

where $\mathbf{g}_u^+(.) \equiv \max\{-\nabla\mathcal{U}_u(.),\ 0\}$ and $\bar{m} = |S| + |A|$. □

A2.2. Proof of Lemma 3.1

Proof. We have that, $F(S) \equiv \mathcal{U}(\mathbf{0}_{m\times n}) - \min_{\boldsymbol{\gamma}:\text{supp}(\boldsymbol{\gamma})\in S,\ \boldsymbol{\gamma}\geq\mathbf{0}_{m\times n}} \mathcal{U}(\boldsymbol{\gamma})$. From the definition of min, it follows that $F(.)$ is a monotonically increasing function of S, i.e., if $S_1 \subseteq S_2 \subseteq V$, then $F(S_1) \leq F(S_2)$. As $F(.)$ is monotonically increasing, $F(S) \geq F(\phi) = \mathcal{U}(\mathbf{0}_{m\times n}) - \mathcal{U}(\mathbf{0}_{m\times n}) = 0$. This shows the non-negativity of $F(.)$. From Lemma A2.1, we know that

$\mathcal{U}(.)$ has a finite RSC and RSM constants: u_Ω and U_Ω respectively. Now, the proof of α-weak submodularity of $F(.)$ follows the proof technique used in Gurumoorthy et al. (2019, Theorem $IV.3$).

For weak-submodularity (Appendix A1.1), we need to lower bound $F(S\cup\{u\}) - F(S)$ and upper bound $F(S\cup A) - F(S)$. Let $\bar{m} = |S| + |A|$. From Lemma A2.2, we have that, $F(S\cup\{u\}) - F(S) \geq \frac{1}{2\tilde{U}_1}\left(\mathbf{g}_u^+(\boldsymbol{\gamma}_S)\right)^2$. From Lemma A2.3, we have that, $0 \leq F(S\cup A) - F(S) \leq \frac{1}{2u_{\bar{m}}}\sum_{u\in A}\left(\mathbf{g}_u^+(\boldsymbol{\gamma}_S)\right)^2$. Using these, the ratio $\frac{\sum_{u\in A} F(S\cup\{u\})-F(S)}{F(S\cup A)-F(S)} \geq \frac{u_{\bar{m}}}{\tilde{U}_1}$. We now lower bound $u_{\bar{m}}$. We recall that $\bar{m} := |S| + |A|$ for $S, A \subseteq V$. With the general sparsity constraints on the support (Section 3.1), $\bar{m} \leq 2K_1$, which makes $u_{\bar{m}} \geq u_{2K_1}$ (Section 2.3). With the column-wise sparsity constraints on the support, (Section 3.2), $\bar{m} \leq 2nK_2$, which makes $u_{\bar{m}} \geq u_{2nK_2}$ (Section 2.3). Hence, we proved that $F(\cdot)$ is α-weakly submodularity with $\alpha \geq \frac{u_{2K_1}}{\tilde{U}_1}$ for the general sparsity case and $\alpha \geq \frac{u_{2nK_2}}{\tilde{U}_1}$ for the column-wise sparsity constraints. Combining the two cases, we have that $\alpha \geq \frac{u_{2K}}{\tilde{U}_1}$ where K denotes the sparsity level of the transport plan. □

A2.3. Proof of Lemma 3.2

Proof. Let S^* be the optimal support set. Let V denote the ground set of cardinality $N \equiv m \times n$ and $K = K_1$ as the general sparsity cardinality constraint. Let S_i be the subset chosen by Algorithm 1 up to iteration i. We define $\mathbf{g}_j^+(\boldsymbol{\gamma}_{S_i}) \equiv \max\{-\nabla\mathcal{U}_j(\boldsymbol{\gamma}_{S_i}),\ 0\}$.

Let a randomly chosen set R consist of $s = \frac{N}{K}\log\frac{1}{\epsilon}$ elements from $V\setminus S_i$. We first estimate the probability that $R\cap(S^*\setminus S_i)$ is non-empty.

$$
\begin{aligned}
\Pr\left[R\cap(S^*\setminus S_i) \neq \phi\right] &= 1 - \Pr\left[R\cap(S^*\setminus S_i) = \phi\right] \\
&= 1 - \left(1 - \frac{|S^*\setminus S_i|}{|V\setminus S_i|}\right)^s \\
&\geq 1 - e^{-s\frac{|S^*\setminus S_i|}{|V\setminus S_i|}} \ (\because 1 - x \leq e^{-x}) \\
&\geq 1 - e^{-s\frac{|S^*\setminus S_i|}{N}} \ (\because |V\setminus S_i| \leq N) \\
&\overset{(1)}{\geq} \left(1 - e^{\frac{sK}{N}}\right)\frac{|S^*\setminus S_i|}{K} \ \left(\text{Using concavity of } f(x) = 1 - e^{-\frac{s}{N}x}.\right) \\
&= (1-\epsilon)\frac{|S^*\setminus S_i|}{K} \ (\text{Substituting the value of } s.)
\end{aligned} \tag{A15}
$$

The inequality (1) is detailed as follows. As $f(x) = 1 - e^{-\frac{s}{N}x}$ is a concave function for $x\in\mathbb{R}$ and as $\frac{|S^*\setminus S_i|}{K} \in [0,1]$, we have that $f\left(\frac{|S^*\setminus S_i|}{K}K + \left(1 - \frac{|S^*\setminus S_i|}{K}\right).0\right) \geq \frac{|S^*\setminus S_i|}{K}f(K) + \left(1 - \frac{|S^*\setminus S_i|}{K}\right)f(0)$.

Now, we observe that, for an element u to be picked by Algorithm 1, $\mathbf{g}_u^+(\boldsymbol{\gamma}_{S_i}) \geq \mathbf{g}_b^+(\boldsymbol{\gamma}_{S_i}),\ \forall b \in R\cap(S^*\setminus S_i)$ (if non-empty). We have that,

$$
\begin{aligned}
\mathbf{g}_u^+(\boldsymbol{\gamma}_{S_i}) \geq \mathbf{g}_b^+(\boldsymbol{\gamma}_{S_i}) &\implies \left(\mathbf{g}_u^+(\boldsymbol{\gamma}_{S_i})\right)^2 \geq \left(\mathbf{g}_b^+(\boldsymbol{\gamma}_{S_i})\right)^2 \\
&\implies \mathbb{E}\left[\left(\mathbf{g}_u^+(\boldsymbol{\gamma}_{S_i})\right)^2 |S_i\right] \geq \mathbb{E}\left[\left(\mathbf{g}_b^+(\boldsymbol{\gamma}_{S_i})\right)^2 |S_i\right]\Pr\left[R\cap(S^*\setminus S_i)\neq\phi\right]
\end{aligned} \tag{A16}
$$

for any element $b\in R\cap(S^*\setminus S_i)$ (if non-empty).

We then use that R is equally likely to contain each element of $S^*\setminus S_i$, so a uniformly random element of $R\cap(S^*\setminus S_i)$ is a uniformly random element of $S^*\setminus S_i$. From (A16),

$$
\begin{aligned}
\mathbb{E}\left[\left(\mathbf{g}_u^+(\boldsymbol{\gamma}_{S_i})\right)^2 |S_i\right] &\geq \Pr\left[R\cap(S^*\setminus S_i)\neq\phi\right]\frac{1}{|S^*\setminus S_i|}\sum_{b\in S^*\setminus S_i}\left(\mathbf{g}_b^+(\boldsymbol{\gamma}_{S_i})\right)^2 \\
&\geq \frac{1-\epsilon}{K}\sum_{b\in S^*\setminus S_i}\left(\mathbf{g}_b^+(\boldsymbol{\gamma}_{S_i})\right)^2 \ (\text{From A15.})
\end{aligned} \tag{A17}
$$

With $S_{i+1} = S_i \cup \{u\}$, Lemma A2.2 gives us,

$$\mathbb{E}\left[2\tilde{U}_1\left(F(S_{i+1}) - F(S_i)\right)|S_i\right] \geq \mathbb{E}\left[\left(\mathbf{g}_u^+(\boldsymbol{\gamma}_{S_i})\right)^2|S_i\right]. \tag{A18}$$

Using Lemma A2.3, we have that,

$$\frac{1-\epsilon}{K}\sum_{b\in S^*\setminus S_i}\left(\mathbf{g}_b^+(\boldsymbol{\gamma}_{S_i})\right)^2 \geq \frac{2u_{\bar{m}}(1-\epsilon)}{K}\left(F(S^*) - F(S_i)\right) \geq \frac{2u_{2K}(1-\epsilon)}{K}\left(F(S^*) - F(S_i)\right). \tag{A19}$$

The last inequality uses that $\bar{m} = |S_i| + |S^* \setminus S_i| \leq 2K$ and hence $u_{\bar{m}} \geq u_{2K}$.

From inequalities (A17), (A18), and (A19), we get the following.

$$\begin{aligned}
&\mathbb{E}\left[2\tilde{U}_1\left(F(S_{i+1}) - F(S_i)\right)|S_i\right] \geq \frac{2u_{2K}(1-\epsilon)}{K}\left(F(S^*) - F(S_i)\right)\\
\implies &\mathbb{E}\left[\left(F(S_{i+1}) - F(S_i)\right)|S_i\right] \geq \frac{u_{2K}(1-\epsilon)}{K\tilde{U}_1}\left(F(S^*) - F(S_i)\right)\ \left(\because \frac{u_{2K}}{\tilde{U}_1} \in (0,1]\right)\\
\implies &\mathbb{E}\left[F(S_{i+1}) - F(S_i)\right] \geq \frac{u_{2K}(1-\epsilon)}{K\tilde{U}_1}\left(F(S^*) - \mathbb{E}\left[F(S_i)\right]\right)\ \text{(Taking an expectation over } A_i\text{.)}
\end{aligned}$$

On re-arranging and using induction, we get

$$\begin{aligned}
\mathbb{E}\left[F(S_K)\right] &\geq \frac{u_{2K}(1-\epsilon)}{\tilde{U}_1 K}F(S^*)\left(\sum_{i=0}^{K-1}\left(1 - \frac{u_{2K}(1-\epsilon)}{K\tilde{U}_1}\right)^i\right)\\
&\geq \left(1 - \left(1 - \frac{u_{2K}(1-\epsilon)}{K\tilde{U}_1}\right)^K\right)F(S^*)\ \left(\text{We use } \frac{u_{2K}}{\tilde{U}_1} \in (0,1] \text{ and sum the Geometric series.}\right)\\
&\geq \left(1 - e^{-\frac{u_{2K}(1-\epsilon)}{\tilde{U}_1}}\right)F(S^*)\ \text{(Using } e^{-x} \geq 1-x\text{)}\\
&= \left(1 - e^{-r(1-\epsilon)}\right)F(S^*)\ \left(\text{where } r = \frac{u_{2K}}{\tilde{U}_1}\right)\\
&\geq \left(1 - e^{-r} - \epsilon\right)F(S^*).
\end{aligned}$$

The last inequality is detailed as follows. Let us first consider a function f over the domain $[0,1]$ defined as $f(x) = z^x - xz$ for some $z \geq 0$. This is a convex function with $f(0) = 1, f(1) = 0$. Thus, $z^x - xz \leq 1$. Taking $z = e^r$ proofs the result. The proof technique is inspired by the proof for the stochastic-greedy algorithm (Mirzasoleiman et al., 2015). □

A2.4. Proof of Lemma 3.3

Our proof is inspired by the approximation ratio proof in Chen et al. (2018). We first discuss the following lemma where our proof differs from that in Chen et al. (2018). In this subsection, we use K to denote the overall cardinality constraint of the column-wise sparse transport plan, i.e., $K = nK_2$.

Lemma A2.4. *For every* $1 \leq i \leq K$, $\mathbb{E}[F(S_i)] \geq \mathbb{E}[F(S_{i-1})] + \alpha\frac{\mathbb{E}[F(OPT_{i-1}\cup S_{i-1})] - \mathbb{E}[F(S_{i-1})]}{K-i+1}$, *where* $\alpha = \frac{u_{2K}}{\tilde{U}_1}$.

Proof. The base OPT_{i-1} is a possible candidate to be the maximizing base, M_i. Now, with the criteria used in Algorithm 2 to pick the next element, we have the following.

$$\sum_{u\in M_i}\mathbf{g}^+(u|S_{i-1}) \geq \sum_{u\in OPT_{i-1}}\mathbf{g}^+(u|S_{i-1}) \implies \sum_{u\in M_i}\left(\mathbf{g}^+(u|S_{i-1})\right)^2 \geq \sum_{u\in OPT_{i-1}}\left(\mathbf{g}^+(u|S_{i-1})\right)^2.$$

Using Lemma A2.2 and Lemma A2.3, we have that,

$$\sum_{u\in M_i}2\tilde{U}_1F(u|S_{i-1}) \geq 2u_{\bar{m}}F(OPT_{i-1}|S_{i-1}) \implies \sum_{u\in M_i}F(u|S_{i-1}) \geq \frac{u_{\bar{m}}}{\tilde{U}_1}F(OPT_{i-1}|S_{i-1}),$$

where $\bar{m} = |OPT_{i-1}| + |S_{i-1}|$.

We denote $\frac{u_{2K}}{\tilde{U}_1} (\leq \frac{u_{\bar{m}}}{\tilde{U}_1})$ by α. Algorithm 2 adds a uniformly random element $u_i \in M_i$ to the set S_{i-1} to obtain the set S_i. As M_i is of the size $K - i + 1$,

$$\begin{aligned}
\mathbb{E}[F(S_i)] = F(S_{i-1}) + \mathbb{E}[F(u_i|S_{i-1})] &= F(S_{i-1}) + \frac{1}{K-i+1} \sum_{u \in M_i} F(u|S_{i-1}) \\
&\geq F(S_{i-1}) + \frac{\alpha}{K-i+1} F(OPT_{i-1}|S_{i-1}) \\
&= F(S_{i-1}) + \alpha \frac{F(OPT_{i-1} \cup S_{i-1}) - F(S_{i-1})}{K-i+1}.
\end{aligned}$$

□

We now discuss the proof of Lemma 3.3.

Proof. Let OPT be an arbitrary optimal solution. As $F(\cdot)$ is monotone, we may assume OPT is a base of $\mathcal{M}$. Chen et al. (2018, Lemma 2.2) describes constructing a random set OPT_i for which $S_i \cup OPT_i$ is a base, for every $0 \leq i \leq K$. From Chen et al. (2018, Lemma 2.3), we have that for every $0 \leq i \leq K$, $\mathbb{E}[F(OPT_i)] \geq \left[1 - \left(\frac{i+1}{K+1}\right)^{\alpha}\right] F(OPT)$. This result uses the non-negativity of F and the property that OPT_i is a uniformly random subset (of size $K - i$) of OPT.

Combining this result with Lemma A2.4, Chen et al. (2018, Corollary 2.5) gives us that for every $1 \leq i \leq K$, $\mathbb{E}[F(S_i)] \geq \mathbb{E}[F(S_{i-1})] + \alpha \frac{\{1-[i/(K+1)]^{\alpha}\}F(OPT) - \mathbb{E}[F(S_{i-1})]}{K-i+1}$. Now, the proof for the approximation ratio of Algorithm 2 follows from Chen et al. (2018, Theorem 2.6). The proof is based on induction. □

In the above proof, we refer the results of (Chen et al., 2018) as given in their arXiv version (`https://arxiv.org/pdf/1707.04347`).

A2.5. Proofs of Lemma 3.4 and Proposition 3.5

Proof. We begin by re-stating the primal problem.

$$\min_{\boldsymbol{\gamma} \geq \mathbf{0}} P(\boldsymbol{\gamma}) \big(:= \langle \mathbf{C}, \boldsymbol{\gamma} \rangle + \textstyle\sum_{j=1}^{n} \Theta(\boldsymbol{\gamma}_j) + \lambda_1 (\|\boldsymbol{\gamma}\mathbf{1} - \boldsymbol{\mu}\|^2_{\mathbf{G}_1} + \|\boldsymbol{\gamma}^\top \mathbf{1} - \boldsymbol{\nu}\|^2_{\mathbf{G}_2}) \big), \tag{A20}$$

where $\Theta(\boldsymbol{\gamma}_j) = \frac{\lambda_2}{2} \|\boldsymbol{\gamma}_j\|^2 + \delta_{B_K}(\boldsymbol{\gamma}_j)$ and $B_K = \{\mathbf{z} \in \mathbb{R}^m_+ : \|\mathbf{z}\|_0 \leq K\}$. We use auxiliary variables $\mathbf{p} \in \mathbb{R}^m$ and $\mathbf{q} \in \mathbb{R}^n$ to set $\boldsymbol{\gamma}\mathbf{1} - \boldsymbol{\mu} = \mathbf{p}$ and $\boldsymbol{\gamma}^\top \mathbf{1} - \boldsymbol{\nu} = \mathbf{q}$. The Lagrangian becomes

$$\min_{\boldsymbol{\gamma} \geq \mathbf{0}; \mathbf{p} \in \mathbb{R}^m; \mathbf{q} \in \mathbb{R}^n} \langle \mathbf{C}, \boldsymbol{\gamma} \rangle + \textstyle\sum_{j=1}^{n} \Theta(\boldsymbol{\gamma}_j) + \lambda_1 (\|\mathbf{p}\|^2_{\mathbf{G}_1} + \|\mathbf{q}\|^2_{\mathbf{G}_2}) + \boldsymbol{\alpha}^\top (\mathbf{p} - \boldsymbol{\gamma}\mathbf{1} + \boldsymbol{\mu}) + \boldsymbol{\beta}^\top (\mathbf{q} - \boldsymbol{\gamma}^\top \mathbf{1} + \boldsymbol{\nu}),$$

where $\boldsymbol{\alpha} \in \mathbb{R}^m$ and $\boldsymbol{\beta} \in \mathbb{R}^n$. We simplify the Lagrangian as follows.

$$\begin{aligned}
&\min_{\boldsymbol{\gamma} \geq \mathbf{0}; \mathbf{p} \in \mathbb{R}^m; \mathbf{q} \in \mathbb{R}^n} \langle \mathbf{C}, \boldsymbol{\gamma} \rangle + \sum_{j=1}^{n} \Theta(\boldsymbol{\gamma}_j) + \lambda_1 (\|\mathbf{p}\|^2_{\mathbf{G}_1} + \|\mathbf{q}\|^2_{\mathbf{G}_2}) + \boldsymbol{\alpha}^\top (\mathbf{p} - \boldsymbol{\gamma}\mathbf{1} + \boldsymbol{\mu}) + \boldsymbol{\beta}^\top (\mathbf{q} - \boldsymbol{\gamma}^\top \mathbf{1} + \boldsymbol{\nu}) \\
&= \sum_{j=1}^{n} \min_{\boldsymbol{\gamma}_j \geq \mathbf{0}} \Big(\langle \mathbf{C}_j - \boldsymbol{\alpha} - \boldsymbol{\beta}_j \mathbf{1}, \boldsymbol{\gamma}_j \rangle + \Theta(\boldsymbol{\gamma}_j) \Big) + \min_{\mathbf{p} \in \mathbb{R}^m; \mathbf{q} \in \mathbb{R}^n} \Big(\lambda_1 (\|\mathbf{p}\|^2_{\mathbf{G}_1} + \|\mathbf{q}\|^2_{\mathbf{G}_2}) + \boldsymbol{\alpha}^\top (\mathbf{p} + \boldsymbol{\mu}) + \boldsymbol{\beta}^\top (\mathbf{q} + \boldsymbol{\nu}) \Big) \\
&= \sum_{j=1}^{n} -\Theta^*(\boldsymbol{\alpha} + \boldsymbol{\beta}_j \mathbf{1} - \mathbf{C}_j) + \min_{\mathbf{p} \in \mathbb{R}^m; \mathbf{q} \in \mathbb{R}^n} \Big(\lambda_1 (\|\mathbf{p}\|^2_{\mathbf{G}_1} + \|\mathbf{q}\|^2_{\mathbf{G}_2}) + \boldsymbol{\alpha}^\top (\mathbf{p} + \boldsymbol{\mu}) + \boldsymbol{\beta}^\top (\mathbf{q} + \boldsymbol{\nu}) \Big).
\end{aligned} \tag{A21}$$

From the optimality conditions, we have that $2\lambda_1 \mathbf{G}_1 \mathbf{p} + \boldsymbol{\alpha} = 0$ and $2\lambda_1 \mathbf{G}_2 + \boldsymbol{\beta} = 0$. On substituting the values of $\mathbf{p}$ as $\boldsymbol{\gamma}\mathbf{1} - \boldsymbol{\mu}$ and $\mathbf{q}$ as $\boldsymbol{\gamma}^\top \mathbf{1} - \boldsymbol{\nu}$, we prove Proposition 3.5. Using this relationship in equation (A21), the simplified Lagrangian is denoted by $D(\boldsymbol{\alpha}, \boldsymbol{\beta})$. The dual problem of (A20) then becomes the following.

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^m, \boldsymbol{\beta} \in \mathbb{R}^n} D(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \max_{\boldsymbol{\alpha} \in \mathbb{R}^m, \boldsymbol{\beta} \in \mathbb{R}^n} \langle \boldsymbol{\alpha}, \boldsymbol{\mu} \rangle + \langle \boldsymbol{\beta}, \boldsymbol{\nu} \rangle - \frac{1}{4\lambda_1} \boldsymbol{\alpha}^\top \mathbf{G}_1^{-1} \boldsymbol{\alpha} - \frac{1}{4\lambda_1} \boldsymbol{\beta}^\top \mathbf{G}_2^{-1} \boldsymbol{\beta} - \sum_{j=1}^{n} \Theta^*(\boldsymbol{\alpha} + \boldsymbol{\beta}_j \mathbf{1} - \mathbf{C}_j).$$

This proves Lemma 3.4. □

Algorithm A3 Classical greedy algorithm for maximizing (weakly) submodular problems with cardinality constraint

```
Input: $\lambda_1, \lambda_2, \boldsymbol{\mu}, \boldsymbol{\nu}, \mathbf{C}, \mathbf{G}_1, \mathbf{G}_2$, sparsity level $K$.
$i = 1, S_0 = \phi, \boldsymbol{\gamma}_{S_0} = \mathbf{0}$.
while $i \leq K$ do
    1. $u = \operatorname*{arg\,max}_{e \in V \setminus S_{i-1}} F(S_{i-1} \cup \{e\}) - F(S_{i-1})$
    2. $S_i = S_{i-1} \cup \{u\}$
    3. $\boldsymbol{\gamma}_{S_i} = \operatorname*{arg\,min}_{\boldsymbol{\gamma}:\text{supp}(\boldsymbol{\gamma}) \in S_i,\ \boldsymbol{\gamma} \geq \mathbf{0}} \mathcal{U}(\boldsymbol{\gamma})$
    4. $i = i + 1$
end while
return $S_K, \boldsymbol{\gamma}_{S_K}$
```

A3. MMD-UOT problem with the support set of the variable $\boldsymbol{\gamma}$ restricted to a given set S

We first present the MMD-UOT formulation in which the support set of $\boldsymbol{\gamma}$ is $T = \{(i,j) | i \in [m], j \in [n]\}$, i.e., no sparsity constraints:

$$\min_{\boldsymbol{\gamma} \geq 0} \mathcal{U}(\boldsymbol{\gamma}), \quad \text{where}$$

$$\begin{aligned}\mathcal{U}(\boldsymbol{\gamma}) \coloneqq \sum_{(i,j)\in T} \Big[&\mathbf{C}_{ij}\boldsymbol{\gamma}_{ij} + \tfrac{\lambda_2}{2}\boldsymbol{\gamma}_{ij}^2 + \lambda_1 \boldsymbol{\gamma}_{ij}\Big(\sum_{(p,q)\in T} \boldsymbol{\gamma}_{pq}(\mathbf{G}_1)_{ip} - 2(\mathbf{G}_1\boldsymbol{\mu}1_n^\top)_{ij}\Big) \\ &+ \lambda_1 \boldsymbol{\gamma}_{ij}\Big(\sum_{(p,q)\in T} \boldsymbol{\gamma}_{pq}(\mathbf{G}_2)_{qj} - 2(1_m\boldsymbol{\nu}^\top\mathbf{G}_2)_{ij}\Big)\Big] + \lambda_1(\|\boldsymbol{\mu}\|^2_{\mathbf{G}_1} + \|\boldsymbol{\nu}\|^2_{\mathbf{G}_2}).\end{aligned} \tag{A22}$$

In the proposed formulation (3), the optimization is solved with the support set of $\boldsymbol{\gamma}$ being restricted to a given set $S \subseteq T$. This equivalently implies that $\boldsymbol{\gamma}_{ij}$ (and $\boldsymbol{\gamma}_{pq}$ terms) can be set to zero for all $(i,j) \in T \setminus S$ (and $(p,q) \in T \setminus S$). Consequently, the optimization problem is only for $\boldsymbol{\gamma}_{ij}$ for $(i,j) \in S$. We denote this variable as $\boldsymbol{\gamma}_S$ in the following:

$$\min_{\boldsymbol{\gamma}:\text{supp}(\boldsymbol{\gamma}) \subseteq S, \boldsymbol{\gamma} \geq 0} \mathcal{U}(\boldsymbol{\gamma}) \equiv \min_{\boldsymbol{\gamma}_S \geq 0} \mathcal{U}(\boldsymbol{\gamma}_S), \quad \text{where}$$

$$\begin{aligned}\mathcal{U}(\boldsymbol{\gamma}_S) \coloneqq \sum_{(i,j)\in S} \Big[&\mathbf{C}_{ij}\boldsymbol{\gamma}_{ij} + \tfrac{\lambda_2}{2}\boldsymbol{\gamma}_{ij}^2 + \lambda_1 \boldsymbol{\gamma}_{ij}\Big(\sum_{(p,q)\in S} \boldsymbol{\gamma}_{pq}(\mathbf{G}_1)_{ip} - 2(\mathbf{G}_1\boldsymbol{\mu}1_n^\top)_{ij}\Big) \\ &+ \lambda_1 \boldsymbol{\gamma}_{ij}\Big(\sum_{(p,q)\in S} \boldsymbol{\gamma}_{pq}(\mathbf{G}_2)_{qj} - 2(1_m\boldsymbol{\nu}^\top\mathbf{G}_2)_{ij}\Big)\Big] + \lambda_1(\|\boldsymbol{\mu}\|^2_{\mathbf{G}_1} + \|\boldsymbol{\nu}\|^2_{\mathbf{G}_2}).\end{aligned} \tag{A23}$$

We note that the above problem has a smooth convex quadratic objective with non-negativity constraint and, therefore, can be solved using the APGD solver (Manupriya et al., 2024b).

A4. Classical greedy and non-stochastic OMP algorithms for solving our GenSparseUOT (4)

Algorithm A3 is the classical greedy algorithm for solving the proposed GenSparseUOT formulation (4).

Algorithm A4 is a non-stochastic OMP algorithm for for maximizing weakly sub- modular problems with cardinality constraint. It is used for solving GenSparseUOT sub-problems in the SPFD experiments (Section 5.1.1, Problem (11)).

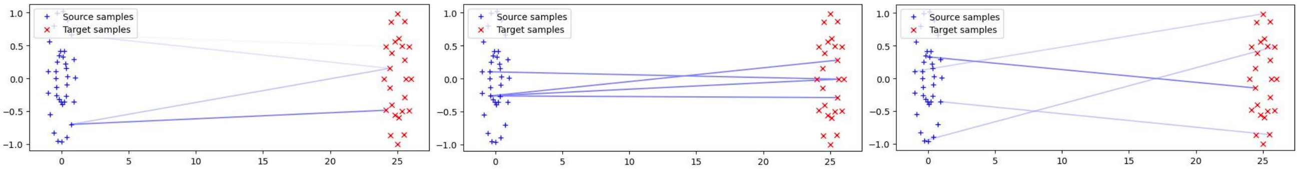
A5. More on Experiments

We present details of experiments discussed in the main paper along with some additional results.

Common Experimental Details: In the proposed approach, we either use the RBF kernel $k(x,y) = \exp\frac{-\|x-y\|^2}{2\sigma^2}$ or kernels from the inverse multiquadratic (IMQ) family: $k(x,y) = (\sigma^2 + \|x-y\|^2)^{-0.5}$ (referred to as IMQ) and $k(x,y) = \left(\frac{1+\|x-y\|^2}{\sigma^2}\right)^{-0.5}$ (referred to as IMQ-v2). These are the universal kernels (Sriperumbudur et al., 2011; Li et al., 2017; Jitkrittum et al., 2019; Dwivedi & Mackey, 2022; Manupriya et al., 2024b;a). The cost function is squared-Euclidean unless otherwise mentioned. We also normalize the cost matrix such that all entries are upper-bounded by 1. The coefficient of quadratic regularization λ_2 is 0 unless otherwise mentioned.

Algorithm A4 Vanilla OMP algorithm for maximizing weakly submodular problems with cardinality constraint

```
Input: $\lambda_1, \lambda_2, \boldsymbol{\mu}, \boldsymbol{\nu}, \mathbf{C}, \mathbf{G}_1, \mathbf{G}_2$, sparsity level $K$.
$i = 1, S_0 = \phi, \boldsymbol{\gamma}_{S_0} = \mathbf{0}$ and $\mathbf{g} = -\nabla\mathcal{U}(\boldsymbol{\gamma}_{S_0})$.
while $i \leq K$ do
    1. $u = \underset{e \in V \setminus S_{i-1}}{\arg\max}\, \mathbf{g}_e$
    2. $S_i = S_{i-1} \cup \{u\}$
    3. $\boldsymbol{\gamma}_{S_i} = \underset{\boldsymbol{\gamma}:\text{supp}(\boldsymbol{\gamma}) \in S_i,\ \boldsymbol{\gamma} \geq \mathbf{0}}{\arg\min} \mathcal{U}(\boldsymbol{\gamma})$
    4. $\mathbf{g} = -\nabla\mathcal{U}(\boldsymbol{\gamma}_{S_i})$
    5. $i = i + 1$
end while
return $S_K, \boldsymbol{\gamma}_{S_K}$
```



(a)

(b)

(c)

Figure A3. (Best viewed in color.) The source samples are shown in blue and the target samples are shown in red. We show an edge between source point i and target point j if $\boldsymbol{\gamma}_{i,j} > 0$. The intensity of the color represents the magnitude of $\boldsymbol{\gamma}_{i,j}$. (a) SSOT (Blondel et al., 2018) results in 4 non-zero entries in $\boldsymbol{\gamma}$. (b) The top-4 entries of the MMD-UOT transport plan. (c) Proposed GenSparseUOT transport plan obtained with sparsity constraint $K = 4$. We can see that the support points of the transport plan obtained by GenSparseUOT are the most diverse, resulting in one-to-one mapping between the source and the target.

The experiments in Section 5.2 are done on an NVIDIA A100-SXM4-40GB GPU, and the remaining experiments are done on an NVIDIA GeForce RTX 4090 GPU.

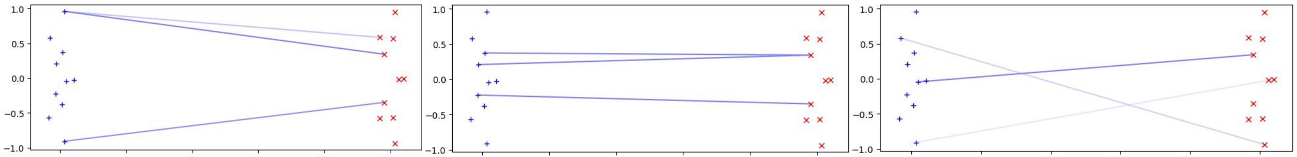
A5.1. Synthetic Experiments

Diversity in the support set. A key feature of submodular maximization is obtaining a diverse solution set (Das & Kempe, 2018) since the selection of the next element essentially involves incremental gain maximization. Hence, we expect the support set of the transport plan learned by our Algorithm 1 for the GenSparseUOT problem (4) to exhibit diversity. Diversity in the support set of a transport plan implies primarily learning one-to-one mappings between the source and the target points rather than one-to-many or many-to-one mappings.

In Figure A3, we observe the diversity in the learned transport plan with $K = 4$ on the two-dimensional source and target sets. Figure A3(a) shows the transport plan obtained using SSOT (Blondel et al., 2018). We observe that SSOT learns many-to-many mappings. For Figure A3(b), we observe that MMD-UOT (Manupriya et al., 2024b) also has similar issues. It should be noted that both SSOT and MMD-UOT do not provide a direct control over the size of support of the transport plan. Hence, one may require a top-K selection heuristic to learn a transport plan with K non-sparse entries (Arase et al., 2023). However, as discussed, proposed Algorithm 1 directly learns a transport plan $\boldsymbol{\gamma}$ with K non-sparse entries. In addition, as observed in Figure A3(c), Algorithm 1 learns several one-to-one mappings, highlighting the diversity in the support set of $\boldsymbol{\gamma}$.

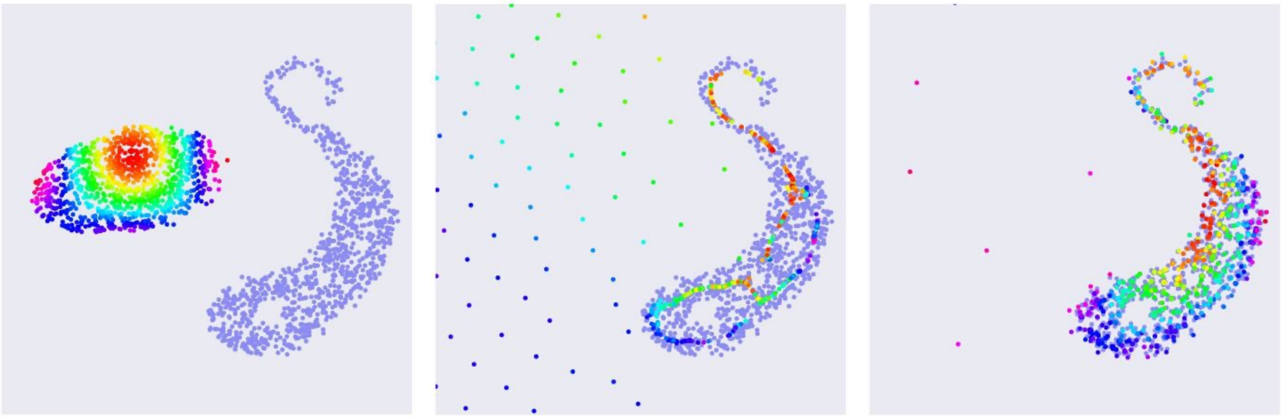
Gradient flow. Gradient flow constructs the trajectory of a source distribution $\bar{\boldsymbol{\mu}}$ being transformed to a given target distribution $\bar{\boldsymbol{\nu}}$. The underlying problem in gradient flow is of solving $\partial_t \bar{\boldsymbol{\mu}}_t = -\nabla_{\bar{\boldsymbol{\mu}}_t} D(\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\nu}})$ for different timesteps $t > 0$, where D is a divergence over measures. Prior works have employed an OT-based divergence (Fatras et al., 2020; Nguyen et al., 2022) and used the Euler scheme for solving this problem (Feydy et al., 2019). Often, in practice, the gradient updates are performed only over the support of the distribution, keeping the mass values of the distribution fixed to uniform (Nguyen et al., 2022). We compare our approach (4) with MMD-UOT (Manupriya et al., 2024b).

In our experiment, the initial source distribution and the target distributions are shown in Figure A5(a). Both the source and the target sets have 1000 data points each. The learning rate for gradient updates is fixed to 0.01 and the number of iterations



(a) (b) (c)

Figure A4. (Best viewed in color.) The source samples are shown in blue and the target samples are shown in red. We show an edge between source point i and target point j if $\boldsymbol{\gamma}_{i,j} > 0$. The intensity of the color represents the magnitude of $\boldsymbol{\gamma}_{i,j}$. (a) SSOT (Blondel et al., 2018) results in 3 non-zero entries in $\boldsymbol{\gamma}$. (b) The top-3 entries of the MMD-UOT transport plan. (c) Proposed GenSparseUOT transport plan obtained with sparsity constraint $K = 3$. We can see that the support points of the transport plan obtained by GenSparseUOT are the most diverse, resulting in one-to-one mapping between the source and the target.



(a) (b) (c)

Figure A5. (Best viewed in color.) (a) Initial source points (rainbow color) on the left and target points (in blue) on the right. (b) Gradient Flow results of MMD-UOT (c) Gradient Flow results of proposed GenSparseUOT solved with Algorithm A4.

is set to 2450. Figures *A*5(b) & *A*5(c) plot the results for MMD-UOT and the proposed GenSparseUOT formulations. We observe that the solution $\boldsymbol{\mu}_t$ obtained by GenSparseUOT closely mimics the target distribution, while the solution obtained with MMD-UOT performs poorly. This is interesting because while GenSparseUOT employs MMD-UOT based objective, the additional sparsity constraint and the submodular maximization approach (Algorithm 1) ensures diversely selected support for GenSparseUOT. This makes the gradients with the proposed approach more informative.

The details of the hyperparameters used for Fig. *A*3 are as follows. We consider empirical measures over the two-dimensional source and target samples with no. of source samples as 35 and no. of target samples as 25. The coefficient of regularization hyperparameter for SSOT is chosen from $\{0.1, 0.5, 1\}$. The result in Fig. *A*3(a) has a coefficient of 0.5, which resulted in the OT plan with the most diverse support points. The results obtained with the proposed method and with MMD-UOT use RBF kernel with σ^2 as 1 and λ_1 as 10. Fig. *A*4 shows results with empirical measures over the two-dimensional source and target samples with no. of source and target samples as 10 each. The coefficient of regularization for SSOT is 0.5, which resulted in the OT plan with the most diverse support points. The results obtained with the proposed method and with MMD-UOT use IMQ kernel with σ^2 as 10 and λ_1 as 10.

The details of hyperparameters used for Fig. *A*5 are as follows. For the proposed method, we use IMQ kernel with σ^2 as 10^{-4} and λ_1 as 10^{-1}. For MMD-UOT, we also use the IMQ kernel but additionally validated over a range of hyperparameters: $\sigma^2 \in \{10^{-4}, 10^{-3}, 10^{-2}\},\ \lambda_1 \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$. The best σ^2 is 10^{-3} and the best $\lambda_1 = 10^{-2}$.

A5.2. Sparse Process Flexibility Design Experiment Details (Section 5.1.1 of the main paper)

Table *A*5 shows the detailed result where we present the mean and the standard deviation of expected profit when run for different seeds. We also show the result with our non-stochastic variant Algorithm A4.

Table A5. Expected profit (higher is better) for SPFD experiment with varying network size constraint l. Proposed refers to our GenSparseUOT formulation (4). We report the mean and std. deviation with 5 different seed values. We observe that our approach outperforms GSOT.

Method	$l = 100$	$l = 175$	$l = 250$
GSOT	0.014 ± 0.001	0.031 ± 0.001	0.044 ± 0.002
Proposed (Algorithm A4)	0.152 ± 0.006	0.212 ± 0.011	0.252 ± 0.008
Proposed (Algorithm 1, $\epsilon = 10^{-2}$)	0.166 ± 0.013	0.224 ± 0.029	$\mathbf{0.293 \pm 0.023}$
Proposed (Algorithm 1, $\epsilon = 10^{-3}$)	$\mathbf{0.167 \pm 0.017}$	0.238 ± 0.021	0.286 ± 0.017
Proposed (Algorithm 1, $\epsilon = 10^{-4}$)	0.147 ± 0.018	$\mathbf{0.240 \pm 0.015}$	0.274 ± 0.008

Table A6. Computation time (s) corresponding to Table 1 results.

Method	$l = 100$	$l = 175$	$l = 250$
GSOT	17.69	20.09	23.00
Proposed ($\epsilon = 10^{-2}$)	6.74	11.99	17.59
Proposed ($\epsilon = 10^{-3}$)	6.33	12.03	17.93
Proposed ($\epsilon = 10^{-4}$)	6.44	11.92	17.68

Validation details: The validation data split is generated following the procedure given by Luo et al. (2023). Both the GSOT's hyperparameters (α and ρ) are independently chosen from the set $\{10^{-2}, 10^{-1}, 1, 10\}$. For the proposed approach, the regularization parameter λ_1 and the kernel parameter σ^2 are chosen from the sets $\{1, 10, 100\}$ and $\{10^{-3}, 10^{-2}, 10^{-1}, 1\}$, respectively.

The hyperparameters chosen after validation are as follows. The proposed approach uses RBF kernel with $\sigma^2 = 1$ and $\lambda_1 = 100$. The hyperparameters (α, ρ) in GSOT are set as $(10, 1)$.

Timing results: Table *A6* compares the time taken for the SPFD experiment by the GSOT method and the proposed approach. This computation time includes the time taken to compute the OT plans and the time taken to compute the overall profit as described in Section 5.1.1. It should be noted that the algorithm to compute the overall profit is the same for both the methods.

A5.3. Word Alignment Experiment Details (Section 5.1.2 of the main paper)

Dataset: The Wiki dataset consists of 2514 training instances, 533 validation instances, and 1052 test instances.

Experimental setup: For baseline methods BOT, POT, and KL-UOT, the results were obtained with the code and optimal hyperparameters shared by Arase et al. (2023). To evaluate the proposed method, we use the same experimental setup as in (Arase et al., 2023) and only tune the hyperparameters. Cosine-distance is employed as the cost function.

Validation Details: The validation data split is the same provided by Arase et al. (2023). The grid for regularization hyperparameters for BOT, POT, KL-UOT are the same as in their code, i.e., BOT, POT have 50 equally-spaced values between 0 and 1 and KL-UOT has 200 equally spaced values in the log space between -3 and 3. For SSOT, the regularization hyperparameter λ is chosen from $\{10^{-7}, 10^{-6}, \ldots, 1\}$. For both MMD-UOT and the proposed approach: (a) the kernel function is validated between RBF and IMQ, (b) the kernel hyperparameters are tuned from the set $\{m/8, m/4, m, 4m, 8m\}$, where m denotes the median used in median heuristics (Gretton et al., 2012), and (c) λ_1 is tuned from the set $\{0.1, 1, 10\}$. For the proposed method, λ_2 is validated on the set $\{0.1, 0\}$.

The chosen hyperparameters for SSOT, MMD-UOT, and our approach are as follows. The coefficient of ℓ_2-norm regularization λ for SSOT is 10^{-4}. For MMD-UOT, IMQ kernel is chosen with $\sigma^2 = 4m$ and λ_1 as 10. For our approach, the chosen kernel is RBF with $\sigma^2 = m/8$, $\lambda_1 = 1$, and $\lambda_2 = 0.1$.

Results: Table 2 reports the F1 and accuracy scores while Table *A7* reports the corresponding precision and recall scores.

Timing results: The average time (in seconds) to compute OT plans: (a) BOT, POT, and KL-UOT baselines of Arase et al. (2023) require 0.01 seconds, (b) SSOT requires 0.08 seconds, (c) MMD-UOT requires 0.40 seconds, and (d) our approach

Table A7. Precision and Recall values on the test split of the Wiki dataset. Higher values are better.

	Null		Total	
Method	Precision	Recall	Precision	Recall
BOT (Arase et al., 2023)	**79.80**	80.29	95.11	**94.81**
POT (Arase et al., 2023)	66.96	79.01	93.64	94.67
KL-UOT (Arase et al., 2023)	77.31	80.16	94.50	94.75
MMD-UOT (Manupriya et al., 2024b)	76.41	75.42	92.73	93.57
SSOT (Blondel et al., 2018)	23.02	40.64	57.71	72.15
Proposed	79.14	**80.72**	**95.13**	94.45

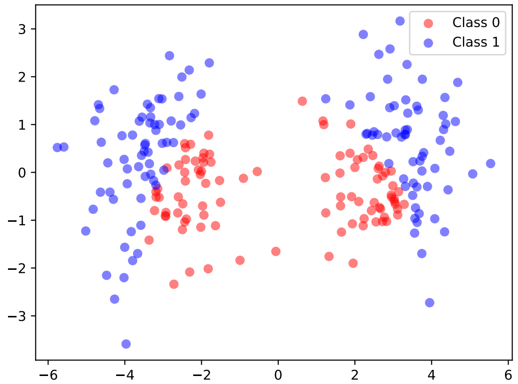


Figure A6. Synthetic data (best viewed in color) for the toy experiment in Section 5.2.

requires 1.06 seconds.

A5.4. Column-wise Sparse Transport Plan Experiment Details (Section 5.2 of the main paper)

A5.4.1. Toy Experiment

Figure A6 shows the data generated for this experiment. Each of the experts is a 1-hidden-layer neural network with GELU as the activation function. The gating function is a single-layer neural network. We employ the Adam optimizer with a constant learning rate scheduler and train for 100 epochs.

Validation Details: The randomly sampled validation split consists of 15% instances. The regularization hyperparameter for SCOT is from $\{10^{-2}, 10^{-1}, 1, 10, 100\}$. For the proposed approach, λ_1 is chosen from $\{10^2, 10, 1\}$ and λ_2 is chosen from $\{10, 1\}$. The validation set for other baselines are as follows: (a) regularization hyperparameter for SSOT: $\{10^{-2}, 10^{-1}, 1, 10\}$, (b) KL-UOT: marginal's regularization $\{10^{-1}, 1, 10\}$, entropic regularization $\{10^{-2}, 10^{-1}, 1, 10\}$, and (c) regularization hyperparameter for entropic OT $\{10^{-2}, 10^{-1}, 1, 10\}$.

The hyperparameter chosen for the proposed approach: IMQ-v2 kernel with $\sigma^2 = 100$ and regularization parameters $\lambda_1 = 100, \lambda_2 = 10$. The coefficient of regularization chosen for SCOT is $\lambda = 10$.

A5.4.2. CIFAR-10 vs CIFAR-10-ROT

We follow the default setting of the code provided by Chen et al. (2022).

For the proposed method, we fix $\lambda_2 = 10$, the kernel function as IMQ-v2, and we set the kernel hyperparameter following the median-heuristics (Gretton et al., 2012). The ℓ_2-norm regularization hyperparameter λ of SCOT and the marginal regularization hyperparameter λ_1 of the proposed approach are chosen $\{0.1, 10, 1000\}$. The default coefficient of the load-balancing loss taken from the code by Chen et al. (2022) is n^2, where n is the number of experts. The default setting results in a skewed allocation across the experts. On increasing the coefficient of the load-balancing loss to n^8, we get a more balanced split of inputs across the experts. Other hyperparameters are set to the default value in the code by Chen et al.

Table A8. Duality gap comparison for solving Problem (7) on varying the regularization hyperparameters. All values are rounded to 6 decimal places. The kernel used is IMQ. A lower duality gap is better.

λ_1	λ_2	Proposed solver		SCOT solver	
		Primal obj.	Duality Gap	Primal obj.	Duality Gap
0.1	0.1	0.006073	$< \mathbf{10^{-10}}$	0.006073	0.000020
1	0.1	0.040079	$\mathbf{0.000014}$	0.060187	0.021549
10	0.1	0.090064	$\mathbf{0.015801}$	0.502088	0.418079
0.1	1	0.006073	$< \mathbf{10^{-10}}$	0.006073	0.000417
1	1	0.042633	$\mathbf{0.000012}$	0.043374	0.001185
10	1	0.092715	$\mathbf{0.001890}$	0.095961	0.005033

Table A9. Duality gap comparison for solving Problem (7) on varying the regularization hyperparameters. All values are rounded to 6 decimal places. The kernel used is RBF. A lower duality gap is better.

λ_1	λ_2	Proposed solver		SCOT solver	
		Primal obj.	Duality Gap	Primal obj.	Duality Gap
0.1	0.1	0.002000	$< \mathbf{10^{-10}}$	0.002000	$< \mathbf{10^{-10}}$
1	0.1	0.019944	$< \mathbf{10^{-10}}$	0.020003	0.000317
10	0.1	0.094129	$\mathbf{0.057683}$	0.174627	0.102349
0.1	1	0.002000	$< \mathbf{10^{-10}}$	0.002000	$< \mathbf{10^{-10}}$
1	1	0.019953	$< \mathbf{10^{-10}}$	0.020218	0.000881
10	1	0.094866	$\mathbf{0.008412}$	0.155751	0.064417

(2022). The test data consists of 20000 examples, with 10000 examples each from CIFAR-10 and the CIFAR-10 rotated.

Timing results: The per-epoch computation time in seconds corresponding to Table 3 are: 58.75s for (vanilla) MoE, 59.92s for SCOT, and 617.90s for our approach.

A5.5. Duality Gap Comparison Experiment Details (Section 5.3 of the main paper)

We present a comparison of the duality gaps obtained using the proposed solver (Algorithm 2) and the SCOT-based solver for optimizing (7).

Adapting the SCOT solver for solving the dual problem (8) corresponding to the primal problem (7): Following Liu et al. (2023), we use the LBFGS optimizer from `scipy.optimize` and initialize the dual variables $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ as zero vectors of appropriate dimensions. Using the LBFGS optimizer requires one to pass a module that takes in inputs as the optimization variable ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ in our case) and returns the objective value in (8) along with the expression for the gradient of the objective w.r.t. the optimization variables. The gradient of the dual objective w.r.t. $\boldsymbol{\alpha}$ is $\boldsymbol{\mu} - \frac{\mathbf{G}_1^{-1}\boldsymbol{\alpha}}{2\lambda_1} - \mathbf{z}\mathbf{1}$ and the gradient w.r.t. $\boldsymbol{\beta}$ is $\boldsymbol{\nu} - \frac{\mathbf{G}_2^{-1}\boldsymbol{\beta}}{2\lambda_1} - \mathbf{z}^\top\mathbf{1}$, where $\mathbf{z}$ is the solution of the sparse projection problem (9) (Liu et al., 2023).

We set max-iter for the APGD algorithm used in the proposed solver (Algorithm 2) as 1000. The results with the SCOT solver are also reported with 1000 as the maximum iteration (after confirming that a higher max-iter does not change the duality gap).

Results: Tables *A8* and *A9* show the duality gaps associated with the proposed solver and the SCOT solver with RBF and IMQ kernels, respectively, and over different hyperparameter values. Table 4 in the main paper shows duality gaps with the solvers employing IMQ-v2 kernel. The kernel hyperparameter is fixed according to the median heuristics (Gretton et al., 2012), and the column-wise sparsity level is fixed as $K_2 = 4$. We observe that the proposed solver obtains better duality gaps across regularization hyperparameters and kernels.

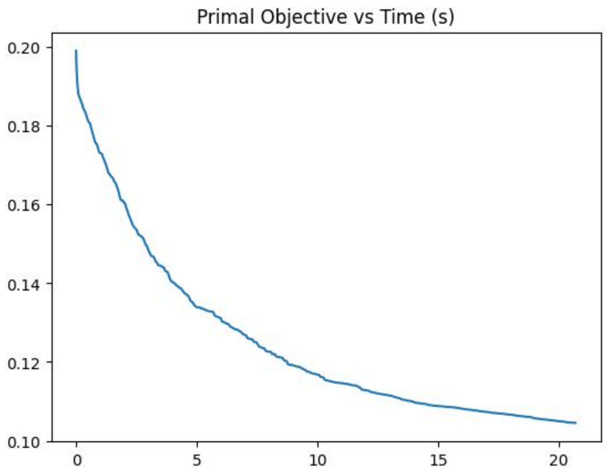


Figure A7. Primal obj vs Time (s) plot for solving the ColSparseUOT formulation using Algorithm 2. The time is computed on an Intel-i9 CPU.

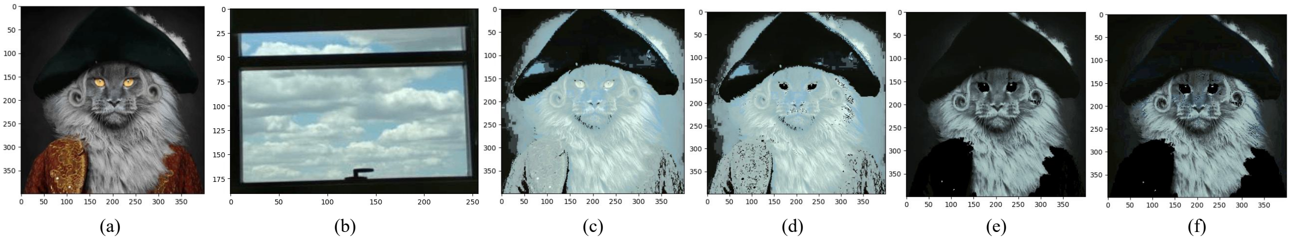


Figure A8. (Best viewed in color.) (a) Image 1 (b) Image 2. (c)-(f) show results (along with the level of sparsity) obtained by transferring the colors from Image 2 to Image 1: (c) OT (99.22%) (Kantorovich, 1942) (d) SSOT (97.86%) (Blondel et al., 2018) (e) MMD-UOT (96.12%) (Manupriya et al., 2024b) (f) Proposed GenSparseUOT (99.61%).

A5.6. Computation Time

Figure A7 shows the objective over time (s) plot while computing column-wise sparse transport plan using Algorithm 2. The source and target measures are empirical measures over two randomly chosen 100-sized batches of CIFAR-10. The kernel used is RBF with median heuristics (Gretton et al., 2012). The sparsity level K_2 is 4 and $\lambda_1 = \lambda_2 = 10$. The computation is done on an Intel-i9 CPU.

A5.7. Color Transfer Experiment

Following Blondel et al. (2018), we perform an experiment of OT-based color transfer. Figure A8 shows the results with various methods and the sparsity level in the obtained transport map. The coefficient of ℓ_2-norm regularization hyperparameter for SSOT is 1. For the proposed method and for MMD-UOT, we use RBF kernel with $\sigma^2 = 10^{-2}$ and $\lambda_1 = 0.1$.