
Initialization and Coordinate Optimization for Multi-way Matching

Da Tang

Columbia University

Tony Jebara

Columbia University & Netflix Inc.

Abstract

We consider the problem of consistently matching multiple sets of elements to each other, which is a common task in fields such as computer vision. To solve the underlying NP-hard objective, existing methods often relax or approximate it, but end up with unsatisfying empirical performance due to a misaligned objective. We propose a coordinate update algorithm that directly optimizes the target objective. By using pairwise alignment information to build an undirected graph and initializing the permutation matrices along the edges of its Maximum Spanning Tree, our algorithm successfully avoids bad local optima. Theoretically, with high probability our algorithm guarantees an optimal solution under reasonable noise assumptions. Empirically, our algorithm consistently and significantly outperforms existing methods on several benchmark tasks on real datasets.

1 INTRODUCTION

Given element sets X_1, \dots, X_n ($n \geq 2$), the problem of finding consistent pairwise bijections between all pairs of sets is known as multi-way matching. As a critical problem in computer science, it is widely used in many computer vision tasks, such as object recognition (Demirci et al., 2006; Yang et al., 2011), shape analysis (Pettersson et al., 2009), and structure from motion (Dai et al., 2014; Roberts et al., 2011). It can also be applied to other fields (e.g. multiple graph matching (Lacoste-Julien et al., 2006; Yan et al., 2013) and data source integration (Zhang et al., 2015)).

In most cases, the multi-way matching problem is ap-

proached as a weighted multi-dimensional matching optimization or relaxation (e.g. the works (Pachauri et al., 2013; Yan et al., 2015b,a)). This objective is easy to solve when $n = 2$, since no consistency between matchings is required. However, as $n \geq 3$, the problem becomes hard due to the combinatorial constraints induced by consistency. The underlying optimization is NP-hard to solve even when $n = 3$ for the unweighted case (Ausiello et al., 2012; Kann, 1991). Therefore, approximate methods, such as convex relaxation (Pachauri et al., 2013) and matrix decomposition (Yan et al., 2015b) have been proposed. These methods work well on datasets with little noise but may become unreliable when more realistic noise levels are present in the data.

In this paper, we aim to find algorithms that directly optimize the true objective of the weighted multi-dimensional matching problem. One intuitive approach is to iteratively update the matching between pairs of sets (X_i, X_{i+1}) 's for $i = 1, \dots, n-1$, but this may produce significant errors once one erroneous pairwise matching is found in the iterative process (Le et al., 2007; Tsochantaridis et al., 2005; Volkovs and Zemel, 2012). Alternatively, one can simply perform coordinate updates on the objective since each coordinate update subproblem is a weighted bipartite matching which can be efficiently solved optimally. However, coordinate update approaches depends heavily on good initialization and may produce bad performance due to local optima.

In this paper, we combine the above ideas and design an effective method for the multi-way matching problem. We build an undirected graph with edge weights from all pairwise matching similarity values, and use its Maximum Spanning Tree (MST) to find a good order for computing $n - 1$ pairwise matchings. This helps avoid bad local optima since it focuses initially on more reliable matchings in the coordinate updates. This seemingly simple idea yields good performance in practice while also enjoying theoretical guarantees. Similar ideas have been discussed in previous works (e.g. (Yan et al., 2016)), but lacked a comprehensive theoretical analysis. In real experiments, we ob-

tain surprisingly strong results on many well-known datasets. For instance, we reliably get 0% error on the famous datasets **CMU House** and **CMU Hotel** for the task of stereo landmark alignments with $m = 30$ points (which has not been easy for previous algorithms to achieve). Theoretically, we not only guarantee that our algorithm solves the problem optimally when pairwise alignment methods work but we *also* guarantee optimality with high probability when a spanning tree on the noise parameter graph has small bottleneck weight (the largest weight in a spanning tree) after imposing some other mild assumptions.

2 CONSISTENT MATCHING FOR SETS OF ELEMENTS

We frame the multiway matching problem as described by Pachauri et al. (2013). Assume we have n element sets X_1, X_2, \dots, X_n where each X_i contains m elements $X_i = \{x_1^i, x_2^i, \dots, x_m^i\}$. For any pair of element sets (X_i, X_j) , we assume that there exists a bijection between their elements such that element x_p^i is mapped to element x_q^j if they are similar to each other. For example, X_1, \dots, X_n could be n images of an everyday object (say a chair) and each image X_i therein contains m pixels $x_1^i, x_2^i, \dots, x_m^i$. Since these images describe the same object type, we expect a bijection to exist between the parts (or pixels) within the pairs of images.

Clearly, such bijections should be consistent with each other. In other words, if element x_p^i is mapped to element x_q^j and element x_q^j is mapped to element x_r^k , then element x_p^i should be mapped to element x_r^k . More specifically, given the element sets X_1, \dots, X_n , we are interested in finding a consistent bijection $\tau_{ij} : \{1, \dots, m\} \rightarrow \{1, \dots, m\}$ between each pair of element sets (X_i, X_j) such that: $x_{\tau_{ij}(p)}^j$ is mapped to x_p^i , τ_{ii} is the identity transform, $\tau_{ij} = \tau_{ji}^{-1}$ and $\tau_{jk} \circ \tau_{ij} = \tau_{ik}$, for any element sets X_i, X_j, X_k and any element x_p^i .

Achieving the above is equivalent to reordering the elements in each element set X_i such that the elements with the same index correspond to each other. Mathematically, finding a consistent bijection τ_{ij} for each pair of element sets (X_i, X_j) is equivalent to finding a bijection $\sigma_i : \{1, \dots, m\} \rightarrow \{1, \dots, m\}$ for each element set X_i , such that element x_p^i is mapped to element x_q^j if and only if $\sigma_i(p) = \sigma_j(q)$. We easily see that these mappings satisfy $\tau_{ij} = \sigma_j^{-1} \circ \sigma_i$ for any τ_{ij} , σ_i and σ_j .

In order to find the mappings σ_i 's, Pachauri et al. (2013) proposed an alternative objective function.

They assume that we are given a *similarity matrix* $T_{ij} \in \mathbb{R}^{m \times m}$ for each pair of sets (X_i, X_j) . The entry $[T_{ij}]_{p,q}$ in the p^{th} row and q^{th} column of T_{ij} represents the similarity level between elements x_p^i and x_q^j . The closer two elements are each other, the larger this similarity level is. By symmetry, we also require that $T_{ij} = T_{ji}^\top$ for any pair of (T_{ij}, T_{ji}) . Without loss of generality, we will assume that $[T_{ij}]_{p,q}$ are constrained to the range $[0, 1]$. Ideally, elements x_p^i and elements x_q^j can be perfectly matched to each other if $[T_{ij}]_{p,q} = 1$ and is maximal. We also hope to avoid matching pairs of elements that not related to each other, e.g. when $[T_{ij}]_{p,q} = 0$ or is minimal. Pachauri et al. (2013) recovered the mappings $\sigma_1, \dots, \sigma_n$ by solving the following optimization problem:

$$\max_{\sigma_1, \dots, \sigma_n} \mathcal{L}(\sigma_1, \dots, \sigma_n) := \sum_{i=1}^n \sum_{j=1}^n \langle P(\sigma_j^{-1} \circ \sigma_i), T_{ij} \rangle \quad (1)$$

where $P(\sigma) \in \mathbb{R}^{m \times m}$ is a permutation matrix satisfying

$$[P]_{p,q} = \begin{cases} 1 & \text{if } \sigma(p) = q \\ 0 & \text{otherwise.} \end{cases}$$

Notice that all permutation matrices are orthogonal matrices and $P(\sigma_j^{-1} \circ \sigma_i) = P(\sigma_i)^{-1} P(\sigma_j)$ for any mappings σ_i and σ_j . Denote the set of all $m \times m$ permutation matrices as \mathcal{P}_m . We rewrite the objective function in Equation (1) as

$$\max_{A_1, \dots, A_n \in \mathcal{P}_m} \mathcal{L}(A_1, \dots, A_n) := \sum_{i=1}^n \sum_{j=1}^n \text{tr}(A_i T_{ij} A_j^\top) \quad (2)$$

since $A_i^\top = A_i^{-1}$, where $A_i = P(\sigma_i)$ is a permutation matrix for the element set X_i . Note that the solution for this optimization problem is not unique (in fact, it has at least $m!$ different tuples of solutions) since $(A_1, \dots, A_n) = (P\hat{A}_1, \dots, P\hat{A}_n)$ is an optimal solution if $(A_1, \dots, A_n) = (\hat{A}_1, \dots, \hat{A}_n)$ is, for any permutation matrix $P \in \mathcal{P}_m$.

A naive method for solving this problem is to recover A_1, \dots, A_n from equations $P_{ij} = A_i^\top A_j$, where $P_{ij} = \arg\max_{P \in \mathcal{P}_m} \text{tr}(P^\top T_{ij})$, for all $i, j \in \{1, \dots, n\}$. We call this method *Pairwise Alignment*. It clearly does not always work since the matrices P_{ij} 's may not be consistent with each other (i.e. they don't always satisfy $P_{ij} P_{jk} = P_{ik}$) and may not correspond to a solution for (A_1, \dots, A_n) . In the next section, we will propose novel algorithms that solve this optimization problem.

3 COORDINATE OPTIMIZATION WITH SMART INITIALIZATION

The optimization problem in Equation (1) is essentially a maximum weighted n -dimensional matching problem. Even for $n = 3$ with binary similarity matrices T_{ij} , (i.e. the unweighted 3-dimensional matching problem), this is an NP-hard problem. Moreover, approximation of the solution within ratio of $\frac{3}{2}$ is also NP-hard (Ausiello et al., 2012; Kann, 1991). Therefore, we cannot find solutions to this problem for arbitrary input values of T_{ij} . Instead, we will constrain the similarity matrices that are used as inputs to the problem. To approximate the problem, Pachauri et al. (2013) proposed an eigenvalue decomposition-based method by first relaxing the combinatorial optimization into a continuous one and then rounding the solution using the Kuhn-Munkres algorithm (Kuhn, 2010). However, Pachauri et al. (2013) could only guarantee their solution when every similarity matrix T_{ij} was close to the ground truth permutation matrix $P(\tilde{\sigma}_j^{-1} \circ \tilde{\sigma}_i)$ (denote $\tilde{\sigma}_1, \dots, \tilde{\sigma}_n$ to be the ground truth mappings we want to find). Unfortunately, this is rarely the case in practice. In the next section, we will present a more general method for solving this problem via coordinate ascent.

3.1 Coordinate ascent over permutations

Consider the objective function in Equation (2). For each permutation matrix A_i , since $\text{tr}(A_i T_{ii} A_i^\top) = \text{tr}(A_i^\top A_i T_{ii}) = \text{tr}(T_{ii})$ is a constant, and $\text{tr}(A_i T_{ij} A_j^\top) = \text{tr}(A_j T_{ij}^\top A_i^\top) = \text{tr}(A_j T_{ji} A_i^\top)$ for any permutation matrix A_j , we know that, if we fix all of the other permutation matrices $A_1, \dots, A_{i-1}, A_{i+1}, \dots, A_n$, then the maximization problem becomes

$$\arg\max_{A_i \in \mathcal{P}_m} \mathcal{L}(A_1, \dots, A_n) = \arg\max_{A_i \in \mathcal{P}_m} \text{tr}(A_i^\top \sum_{1 \leq j \leq n, i \neq j} A_j T_{ij}). \quad (3)$$

The optimization in Equation (3) can be solved in polynomial time (for example, through the $O(m^3)$ Kuhn-Munkres algorithm (Kuhn, 2010)). Hence, a naive coordinate algorithm is easy to derive: initialize the permutation matrices A_1, \dots, A_n (either randomly or deterministically). Then, for each iteration, randomly pick $i \in \{1, \dots, n\}$, update A_i according to Equation (3), and repeat until convergence. Unfortunately, standard ways of initializing such an algorithm lead to poor local optima (see §4.2). Better performance can be achieved, however, if we use pairwise alignment information to construct a good initialization. This approach is discussed in §3.2.

3.2 MST-based initialization

We seek a good initialization for the coordinate update approach summarized in Equation (1). Consider a single term $\text{tr}(A_i T_{ij} A_j^\top)$ in the objective function in Equation (2). As we maximize that objective, we say that we are *confident* in the values chosen for A_i and A_j if the corresponding term $\text{tr}(A_i T_{ij} A_j^\top)$ is large. Define

$$f(T_{ij}) := \max_{P \in \mathcal{P}_m} \text{tr}(P^\top T_{ij}) = \max_{A_i, A_j \in \mathcal{P}_m} \text{tr}(A_i T_{ij} A_j^\top) \quad (4)$$

for each T_{ij} . We call an initialization of our algorithm *convincing* if $A_i = \hat{A}_i$ and $A_j = \hat{A}_j$ and $\text{tr}(\hat{A}_i T_{ij} \hat{A}_j^\top)$ is close or equal to $f(T_{ij})$ for some permutation matrices $\hat{A}_i, \hat{A}_j \in \mathcal{P}_m$.

The above intuition encourages us to first initialize the matrices A_i, A_j that correspond to values of $f(T_{ij})$ that are large. To achieve this, we build an undirected graph $G = (V, E)$ where each element set X_i corresponds to one vertex $v_i \in V$ and each pair of element sets (X_i, X_j) ($i \neq j$) corresponds to an edge $(v_i, v_j) \in E$ with weight $f(T_{ij}) = f(T_{ji})$. We then find a Maximum Spanning Tree $T = (V, E')$ of G . Then, we initialize the matrices A_1, \dots, A_n along the edges in E' as follows. Initially, we have n sets S_1, \dots, S_n of vertices, each containing one vertex in V . Then, for every edge in each edge $(v_i, v_j) \in E'$, we try to *combine* them together and use the similarity T_{ij} to find the permutation matrices corresponding to vertices in the sets contain v_i and v_j . Details are shown in Algorithm 1.

Algorithm 1 MST-based coordinate updates for multi-way matching

Input:

The similarity matrices T_{ij} 's ($i, j \in \{1, \dots, n\}$).

Output:

The permutation matrices A_i 's ($i \in \{1, \dots, n\}$).

- 1: Construct graph $G = (V, E)$ as in §3.2, compute a Maximum Spanning Tree $T = (V, E')$ of G ;
 - 2: Initialize $S_i \leftarrow \{v_i\}$ for each $v_i \in V$; For each A_i , initialize it to be any permutation matrix in \mathcal{P}_m ;
 - 3: **for** each edge $(v_i, v_j) \in E'$ **do**
 - 4: Compute $\hat{P} = \arg\max_P \text{tr}(P^\top A_i T_{ij} A_j^\top)$;
 - 5: Update $A_{j'} \leftarrow \hat{P} A_{j'}$ for each $v_{j'} \in S_j$;
 - 6: Let $S' = S_i \cup S_j$;
 - 7: Update $S_k \leftarrow S'$ for each $v_k \in S'$;
 - 8: **end for**
 - 9: **while** Not converged **do**
 - 10: Randomly pick $i \in \{1, \dots, n\}$;
 - 11: Update A_i according to Equation (3);
 - 12: **end while**
-

The above algorithm uses a Maximum Spanning Tree to initialize the permutation matrices A_1, \dots, A_n . To iteratively *combine* the vertices in V to find an initialization, we need each edge (v_i, v_j) that is selected to have a relatively large $f(T_{ij})$ value. The Maximum Spanning Tree of G achieves this. Subsequently, the algorithm above simply iterates the usual coordinate update process. We will next analyze how this initialization provides a reliable starting point for the coordinate updates that will ultimately produce a good final set of permutation matrices (A_1, \dots, A_n) .

3.3 Analysis of the coordinate update

3.3.1 Analysis without noise

We now analyze the behavior of Algorithm 1. First, consider a simple case where we guarantee through the Pairwise Alignment method that there are *consistent* permutation matrices $P_{ij} = f(T_{ij})$ for each pair of element sets (X_i, X_j) , i.e. $P_{ij}P_{jk} = P_{ik}$ for all $i, j, k \in \{1, \dots, n\}$ (here by guarantee we mean that the maximum value of $\text{tr}(P^\top T_{ij})$ will be achieved for some unique permutation matrix P , for each similarity matrix T_{ij}). If we have consistency, then we can easily recover the optimal (A_1, \dots, A_n) from the P_{ij} matrices by setting $A_i = P_{1i}$ for each A_i since each single term $\text{tr}(A_i T_{ij} A_j^\top) = \text{tr}(A_j^\top A_i T_{ij}) = \text{tr}(P_{ij}^\top T_{ij})$ in the objective function in the Equation (2) is maximized.

What is Algorithm 1's behavior under this constraint? Can we guarantee that it recovers all A_i 's matrices optimally? The answer is YES. We leverage the following theorem:

Theorem 1. *If we recover consistent permutation matrices $P_{ij} = f(T_{ij})$ for all pairs of element sets (X_i, X_j) using the Pairwise Alignment method, then we can guarantee that Algorithm 1 solves the optimization problem in Equation 2 optimally.*

The proof is in Section A of the supplementary material. From Theorem 1, we know that Algorithm 1 is at least as good as the Pairwise Alignment method. Moreover, the optimality cases in Theorem 1 subsume all cases that Pachauri et al. (2013) claimed they could solve optimally. Next, we go even further and guarantee optimality in much more general settings.

3.3.2 Analysis with noise

A more interesting setting is when the matrices T_{ij} 's are not perfect and consistent permutation matrices but rather have been corrupted by noise. If we denote the optimal solution for the optimization problem in Equation (2) as $(A_1, \dots, A_n) = (\hat{A}_1, \dots, \hat{A}_n)$, then ideally the best input data we could have for each T_{ij} would be $T_{ij} = \hat{T}_{ij} := \hat{A}_i^\top \hat{A}_j$. In the case where

$T_{ij} = \hat{T}_{ij}$ for each $i, j \in \{1, \dots, n\}$, it is obvious from Theorem 1 that Algorithm 1 solves this optimization problem optimally.

What if the similarity matrices have noise and T_{ij} is not perfectly equal to \hat{T}_{ij} for some (or all) of the T_{ij} 's? To analyze Algorithm 1, we will assume that the T_{ij} inputs are random perturbations near \hat{T}_{ij} . We only need to consider matrices T_{ij} where $i \neq j$ since the Algorithm does not depend on T_{ii} in any way. Recall that we assumed that the entries of T_{ij} ranged from $[0, 1]$. We propose the following model of the noise that generates the entries of T_{ij} as perturbations of the ground-truth \hat{T}_{ij} :

$$[T_{ij}]_{p,q} = \begin{cases} 1 - Z_{ijpq}^2 & \text{if } i < j \text{ and } [\hat{T}_{ij}]_{p,q} = 1 \\ Z_{ijpq}^2 & \text{if } i < j \text{ and } [\hat{T}_{ij}]_{p,q} = 0 \\ [T_{ji}]_{q,p} & \text{if } i > j. \end{cases} \quad (5)$$

Here $Z_{ijpq} \sim \mathcal{N}(0, \eta_{ij})$'s are independent Gaussian random variables for any $1 \leq i < j \leq n$ and $p, q \in \{1, \dots, m\}$. We assume that different T_{ij} matrices may have different variance parameters η_{ij} 's since we may have different noise levels for different pairs of element sets. Also, we require $\eta_{ij} \leq O(1)$ for each η_{ij} since we want the similarity matrices to only have entries in $[0, 1]$. Notice that we still maintain $T_{ij} = T_{ji}^\top$ for all $i \neq j$ under the model in Equation (5). We now have the following more general theorem:

Theorem 2. *With probability $1 - o(1)$ and for sufficiently large n and m , Algorithm 1 finds an optimal solution for the optimization problem in Equation (2) under the following conditions:*

- $n \geq 20 \ln m$, and $\exists \gamma > 0$ such that $n \leq m^\gamma$,
- the bottleneck length of the **Minimum Bottleneck Spanning Tree** of G is at most $\frac{1}{4(3+\gamma) \ln m + 4}$ where $G = (V, E)$ is a complete undirected weighted graph, with a vertex $v_i \in V$ for each set X_i and with edges $(v_i, v_j) \in E$ with weight η_{ij} ,
- and $\max_{1 \leq i < j \leq n} \eta_{ij} \leq \frac{1}{3}$.

The proof is in Section B of the supplementary material. Here the Minimum Bottleneck Spanning Tree of a graph G means a spanning tree of G which has minimal edge weight on its heaviest edge. It seems that our algorithm could work well if n and m are both large and there exists a spanning tree of graph G with all edge weights no more than $O(\frac{1}{\log m})$. In the proof for this theorem we will show that we can use the Pairwise Alignment method to solve the optimization problem optimally with high probability if all edges of G have weight no more than $O(\frac{1}{\log m})$. This is the same guarantee asymptotically as our bottleneck weight bound

but the latter applies for all edges in E . So, our algorithm remains optimal (with high probability) for a much broader set of inputs.

3.4 Practical improvements

In §3.2 and 3.3, we introduced our algorithm and discussed its theoretical guarantees. However, to make Algorithm 1 better in practice, we also suggest some minor improvements that tend to provide slightly better empirical performance.

3.4.1 Combining initialization with coordinate optimization

In Algorithm 1, we propose a coordinate update process after an initialization step. However, there is a possibility that we may find bad solutions under this initialization as well. Therefore, it is helpful to add a coordinate update process right after each iteration of initialization that may potentially fix some errors the algorithm made during that iteration. During each iteration, after we have processed the vertices in the set S' (line 6 of Algorithm 1), we can do a coordinate update on the corresponding permutation matrices A_k 's where $v_k \in S'$ as:

$$A_k = \operatorname{argmax}_P \operatorname{tr}(P^\top \sum_{v_{k'} \in S', k' \neq k} A_{k'} T_{kk'}). \quad (6)$$

By adding these intermediate update steps, we no longer need to have a final coordinate update step since the additional coordinate updates after the last iteration of initialization have already played that role.

3.4.2 Using a good MST edge ordering

In Algorithm 1, we performed initialization by enumerating the edges of the Maximum Spanning Tree T . It is reasonable that running the updates in a good order along the edges may be beneficial. In this section, we propose two kinds of ordering that we have found work well in practice: Prim's order and Kruskal's order.

As in line 5 of Algorithm 1, we need to update $|S_j|$ different permutation matrices in one step. Even though we have proved that this algorithm works well in many cases, it can be improved if we are more *cautious* and update fewer permutation matrices at each iteration. One way is to use Prim's algorithm (Ahuja, 1988) to compute the Maximum Spanning Tree and then process the edges in the order that we get them through the execution of Prim's algorithm. Since there is only one vertex in the set $|S_j|$ each time, we only need to update one permutation matrix at each iteration. We call this ordering *Prim's order*.

Alternatively, the edge weights themselves are potentially important for initialization. As discussed in §3.3.1, we are more confident in edges (v_i, v_j) whose weights $f(T_{ij})$'s are large. Therefore, we update according to edges that we trust more first. To achieve that goal, we can process the edges in the descending weight order. This is exactly the edge order that we get from running Kruskal's algorithm (Ahuja, 1988). We call this ordering *Kruskal's order*.

3.4.3 The overall algorithm

Algorithm 2 Improved MST-based coordinate updates for multi-way matching

Input:

The similarity matrices T_{ij} 's ($i, j \in \{1, \dots, n\}$).

Output:

The permutation matrices A_i 's ($i \in \{1, \dots, n\}$).

- 1: Construct the Graph $G = (V, E)$ as in the §3.2, Compute a Maximum Spanning Tree $T = (V, E')$ of G ;
 - 2: Sort the edges in E' with Prim's order or Kruskal's order as discussed in §3.4;
 - 3: Initialize $S_i \leftarrow \{v_i\}$ for each $v_i \in V$; For each A_i , initialize it to be any permutation matrix in \mathcal{P}_m ;
 - 4: **for** each edge $(v_i, v_j) \in E'$ **do**
 - 5: Compute $\hat{P} = \operatorname{argmax}_P \operatorname{tr}(P^\top A_i T_{ij} A_j^\top)$;
 - 6: Update $A_{j'} \leftarrow \hat{P} A_{j'}$ for each $v_{j'} \in S_j$;
 - 7: Let $S' = S_i \cup S_j$;
 - 8: Update $S_k \leftarrow S'$ for each $v_k \in S'$;
 - 9: **while** Not converged **do**
 - 10: Randomly pick $v_k \in S'$;
 - 11: Update A_k according to Equation (6);
 - 12: **end while**
 - 13: **end for**
-

By adding the heuristics mentioned above, we obtain a slight modification of our algorithm as shown in Algorithm Box 2. This algorithm works slightly better and we will explore how these heuristics fare in the experiments section. Using techniques similar to those in the proof of Theorem 1, it is easy to show that Algorithm 2 is at least as good as the Pairwise Alignment method:

Theorem 3. *If we can guarantee the recovery of pairwise-consistent permutation matrices $P_{ij} = f(T_{ij})$ for each pair of element sets (X_i, X_j) using the Pairwise Alignment method, then we can guarantee that Algorithm 2 solves the optimization problem in Equation 2 optimally.*

4 EXPERIMENTS

In this section, we will show how our algorithms behave in practice. We focus primarily on computer vision datasets. For each dataset, we compare our algorithm with the Permutation Synchronization algorithm Pachauri et al. (2013), which is a state-of-the-art method for multi-way matching.

4.1 PCA reconstruction of MNIST digits

Our first experiment is on image compression and recovery. We use the **MNIST** dataset (LeCun et al., 1998), which contains 70,000 images of individual handwritten digits from $\{0, \dots, 9\}$.

In one experiment, we randomly selected $n = 100$ images I_1, \dots, I_n from the dataset, where each digit has roughly $\frac{n}{10}$ images. Note that we do not use a larger number of images because of scalability limitations of Pachauri et al. (2013) which we need as our baseline in the evaluation. Our algorithms, however, easily scale to much larger data-sets. The goal of this experiment is to compress the MNIST images with low dimensionality. We represent each image I_i as an element set by randomly selecting $m = 30$ white pixels (the MNIST digits are white drawings on a black background). This forms the element set $X_i = \{x_j^i = (a_j^i, b_j^i) : j \in \{1, \dots, m\}\}$, where (a_j^i, b_j^i) is the coordinate of the j^{th} selected pixel of image I_i .

We will use Principal Components Analysis (PCA) as our compression technique and apply it to the element sets X_1, \dots, X_n . We can view each X_i as a matrix $Y_i \in \mathbb{R}^{m \times 2}$ where the j^{th} row of Y_i is merely (a_j^i, b_j^i) . It is straightforward to vectorize these Y_i matrices and apply PCA to compressively store the vectors $\text{vec}(Y_1), \dots, \text{vec}(Y_n) \in \mathbb{R}^{2m}$. We explore various levels of compression by keeping only the $k \in \{1, \dots, \min(n, 2m)\}$ leading eigenvectors from PCA as well as a mean vector in \mathbb{R}^{2m} . However, since X_1, \dots, X_n are element sets, we are free to permute the order of the element sets prior to the application of PCA. Since each of the MNIST images is a digit, it is reasonable to believe that for many pairs of images, there should be a bijection or correspondence relationship between the pixels that compose each digit. Intuitively, permutations that discover that bijection will help improve PCA performance. We will evaluate how various permutation algorithms that precede PCA help its ability to reconstruct the original data at various compression levels k . Our goal is to reduce the reconstruction error by using our algorithms to reorder the elements in each element set X_i prior to PCA compression.

Results are shown in Figure 1. Here we use the ra-

dial basis function (RBF) kernel to compute the similarity matrices T_{ij} 's. More specifically, $[T_{ij}]_{p,q} = \exp(-\frac{\|x_p^i - x_q^j\|^2}{2\sigma^2})$ where σ is a parameter. This function is suitable for our settings since we expect $[T_{ij}]_{p,q}$ to be close to 1 when x_p^i and x_q^j are close to each other and to be close to 0 otherwise. We compare the two versions of our algorithms with two baseline cases. In the first baseline, we do not reorder the pixels (the no permutation case). In the second baseline, we use the Permutation Synchronization algorithm to reorder the pixels before compression. For all methods in our experiment, we selected the best σ value that achieves the best performance. Figure 1 shows the reconstruction errors and the recovered digits obtained from all five methods. All versions of our algorithm outperform the two baseline methods. Note that both versions of Algorithm 2 slightly outperform Algorithm 1 (although they often produce similar results) which is due to the additional heuristics we discussed in §3.4.

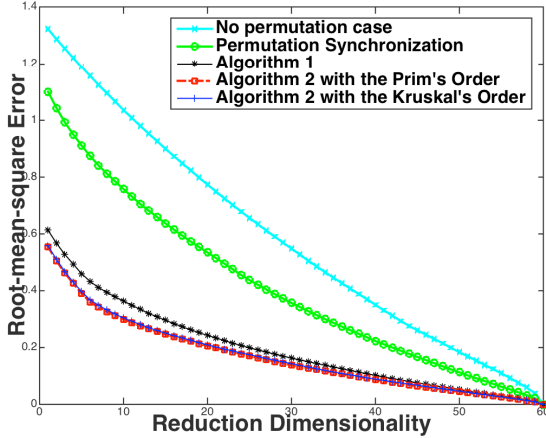
4.2 Stereo landmark alignment

The second computer vision task we focus on is stereo matching. The goal is to align pixels from multiple images of a single object where the images are taken from various vantage points. For this experiment, we have 2 datasets, the **CMU House** dataset and the **CMU Hotel** dataset¹. Each of them contains n images of the same toy house ($n = 111$ for the CMU House dataset and $n = 101$ for the CMU Hotel dataset). For each of the two toy houses, we have $m = 30$ landmark points selected, and each of the n figures contains a different view of these m landmark points. Our goal is to find a consistent mapping that aligns points that correspond to the same landmarks together.

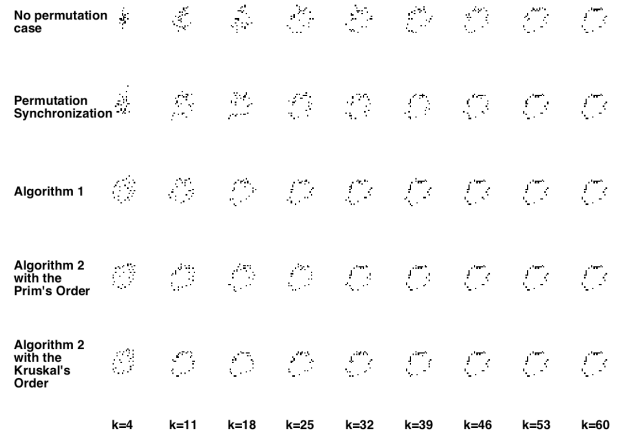
The element sets are constructed by extracting visual features. Once again, we use the RBF kernel to compute similarity matrices. We denote the tasks in this experiment as *CMU House RBF* and *CMU Hotel RBF*. Since we have the ground truth alignments for these two datasets, it is also reasonable to use some local alignments between pairs of figures to construct the similarity matrices. Hence, we also use the outputs of the Pairwise Alignment method as our similarity matrices. We call these corresponding tasks *CMU House Alignment* and *CMU Hotel Alignment*. To evaluate performance, we compute the average error rates of all pairs of element sets.

Results are shown in the first four lines of Table 1. For this experiment, we only report performance from Algorithm 2 though our other algorithm performs almost as well. We compare the two versions of our proposed algorithm with the Permutation Synchroniza-

¹<http://vasc.ri.cmu.edu/idb/html/motion/>.



(a) The PCA reconstruction errors of our methods compared to two baseline methods. The horizontal axis represents the reduction dimensionality k . The performance of the two versions of Algorithm 2 are almost the same, and both clearly outperform previous approaches.



(b) The reconstructed results of an image of digit 0 by all methods. For each figure, we show the reconstructed images with $k = 4, 11, 18, \dots, 60$ eigenvectors, respectively. We see that our methods can reconstruct the image well even at $k = 4$. Meanwhile, the other two methods require many more eigenvectors to reconstruct a recognizable digit 0 image.

Figure 1: Results for the PCA reconstruction experiment

Table 1: Average Error Rates of alignments for the datasets House, Hotel, Building and Sentence

TASK	PERM-SYNC	PRIM'S ORDER	KRUSKAL'S ORDER
CMU House RBF	$(22.61 \pm 7.49)\%$	$(0.00 \pm 0.00)\%$	$(0.00 \pm 0.00)\%$
CMU House Alignment	$(4.71 \pm 1.98)\%$	$(0.81 \pm 0.96)\%$	$(1.89 \pm 2.21)\%$
CMU Hotel RBF	$(18.63 \pm 2.90)\%$	$(0.00 \pm 0.00)\%$	$(0.00 \pm 0.00)\%$
CMU Hotel Alignment	$(5.22 \pm 2.55)\%$	$(3.59 \pm 0.75)\%$	$(4.20 \pm 0.71)\%$
Building RBF	$(86.71 \pm 3.36)\%$	$(49.87 \pm 0.24)\%$	$(50.39 \pm 0.25)\%$
Building Alignment	$(50.49 \pm 1.09)\%$	$(48.52 \pm 0.50)\%$	$(48.61 \pm 0.52)\%$
Sentence RBF	$(62.65 \pm 2.90)\%$	$(55.26 \pm 0.82)\%$	$(55.85 \pm 0.95)\%$
Sentence Alignment	$(58.69 \pm 2.99)\%$	$(56.06 \pm 1.19)\%$	$(55.59 \pm 1.58)\%$

tion method. Here we show results over 10 trials of the experiments since performance is stochastic due to the random re-ordering of the images and the pixels prior to input to the algorithms. We can see that both our methods reliably solve this problem with **100%** accuracy in the RBF setting. Meanwhile the baseline method behaves much worse. For the Alignment setting, the baseline method's behaves better since it tends to excel with those types of inputs. Nevertheless, our methods still outperform it, especially the Prim's Order version.

Notice that the initialization component of our algorithms is very important. Without the smart initialization technique, coordinate updates behave unreliably. For instance, in the CMU House RBF task, we obtain an average error rate of $(3.90 \pm 2.63)\%$ if we randomly initialize all permutation matrices. Mean-

while we always get 0% error rate when we use our MST-based initialization technique.

4.3 Repetitive structures of key points

Beyond stereo matching, we are also interested in matching images with complicated geometric ambiguities. For instance, images with frequent repetitive structures are extremely hard to handle if we use high-dimensional features such as SIFT (Pachauri et al., 2013). In this experiment, we use the **Building** dataset (Roberts et al., 2011) which has such kind of structures. Pachauri et al. (2013) used this dataset and hand-annotated 25 similar-looking landmark points in the scene across the dataset. We use their hand-annotated data and select $n = 14$ images for evaluation. However, in each image, we have $m = 28$ key points, and we do not always have all of the 25 land-



(a) The Permutation Synchronization method under the RBF setting



(b) Algorithm 2 with Prim's Order under the RBF setting



(c) Algorithm 2 with Kruskal's Order under the RBF setting



(d) The Permutation Synchronization method under the Alignment setting



(e) Algorithm 2 with Prim's Order under the Alignment setting



(f) Algorithm 2 with Kruskal's Order under the Alignment setting

Figure 2: Key Points Alignment for the Building dataset. Green lines are the ground truth alignments while red lines are the computed alignments. Less green lines being exposed means a better performance.

mark points in each scene. Hence, there are many useless key points in each image, which makes the multi-way matching task even harder.

As with the Stereo Landmark Alignments test in §4.2, we explore the RBF setting and the Alignment setting for this experiment. The performance of our Algorithm 2 is compared against the baseline Permutation Synchronization method as shown in Table 1. This task is much more difficult than stereo matching and we do not expect a very high accuracy. Nevertheless, both our methods outperform the baseline algorithm in both settings (more notably in the RBF setting). Furthermore, the alignments we get for each setting and each algorithm are shown in Figure 2. All of those 6 groups of alignments are between a same pair of images. We can see that, even when we have highly repetitive structures and significant noise in the datasets, our Algorithm 2 is still stable and gets satisfactory consistent matchings.

4.4 Experiments in domains beyond computer vision

Our work aligns multiple objects that are composed of sets of consistent parts. This setting is not limited to images and computer vision tasks. For example, we can apply our algorithms on natural language processing datasets such as the **Sentence**² dataset. This data-set contains human-labeled sentences from 30 re-

search papers. We tried to consistently align the word frequency vectors of each label from different papers. Table 1 shows that our methods outperform the baseline method, which leads us to believe that our methods could extend to fields beyond computer vision.

5 CONCLUSION

Straightforward coordinate updates for the multi-way matching objective are not reliable and produce poor performance. However, if seeded with a good initialization, coordinate updates will (with high probability) not get stuck at bad local optima. By combining a traditional coordinate ascent algorithm with iterative initializations along the edges of a Maximum Spanning Tree (of the graph with edge weights given by pairwise matching similarity values), we obtain stronger empirical results and stronger theoretical guarantees. We outperform leading baseline methods on various problems in computer vision as well as other domains. In addition, our theoretical analysis shows that we do not require all of the noise parameters to be small to ensure a perfect alignment with high probability. Rather, we only require that the spanning tree (on the noise parameter graph) has a small bottleneck edge weight (along with other mild conditions).

Acknowledgements

Work supported in part by DARPA N66001-15-2-4026, N66001-15-C-4032 and NSF III-1526914, IIS-1451500, CCF-1302269.

²<https://archive.ics.uci.edu/ml/datasets/Sentence+Classification>

References

- Ravindra K Ahuja. *Network flows*. PhD thesis, TECHNISCHE HOCHSCHULE DARMSTADT, 1988.
- Giorgio Ausiello, Pierluigi Crescenzi, Giorgio Gambosi, Viggo Kann, Alberto Marchetti-Spaccamela, and Marco Protasi. *Complexity and approximation: Combinatorial optimization problems and their approximability properties*. Springer Science & Business Media, 2012.
- Yuchao Dai, Hongdong Li, and Mingyi He. A simple prior-free method for non-rigid structure-from-motion factorization. *International Journal of Computer Vision*, 107(2):101–122, 2014.
- M Fatih Demirci, Ali Shokoufandeh, Yakov Keselman, Lars Bretzner, and Sven Dickinson. Object recognition as many-to-many feature matching. *International Journal of Computer Vision*, 69(2):203–222, 2006.
- Heinrich Dörrie. *100 great problems of elementary mathematics*. Courier Corporation, 2013.
- Viggo Kann. Maximum bounded 3-dimensional matching is max snp-complete. *Information Processing Letters*, 37(1):27–35, 1991.
- Harold W Kuhn. The hungarian method for the assignment problem. In *50 Years of Integer Programming 1958-2008*, pages 29–47. Springer, 2010.
- Simon Lacoste-Julien, Ben Taskar, Dan Klein, and Michael I Jordan. Word alignment via quadratic assignment. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 112–119. Association for Computational Linguistics, 2006.
- Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.
- Quoc V Le, Alex J Smola, and Svn Vishwanathan. Bundle methods for machine learning. In *Advances in neural information processing systems*, pages 1377–1384, 2007.
- Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- Deepti Pachauri, Risi Kondor, and Vikas Singh. Solving the multi-way matching problem by permutation synchronization. In *Advances in neural information processing systems*, pages 1860–1868, 2013.
- James Petterson, Jin Yu, Julian J McAuley, and Tibério S Caetano. Exponential family graph matching and ranking. In *Advances in Neural Information Processing Systems*, pages 1455–1463, 2009.
- Tobias Polzin and Siavash Vahdati Daneshmand. On steiner trees and minimum spanning trees in hypergraphs. *Operations Research Letters*, 31(1):12–20, 2003.
- Richard Roberts, Sudipta N Sinha, Richard Szeliski, and Drew Steedly. Structure from motion for scenes with large duplicate structures. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3137–3144. IEEE, 2011.
- Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6(Sep):1453–1484, 2005.
- Maksims Volkovs and Richard S Zemel. Efficient sampling for bipartite matching problems. In *Advances in Neural Information Processing Systems*, pages 1313–1321, 2012.
- Junchi Yan, Yu Tian, Hongyuan Zha, Xiaokang Yang, Ya Zhang, and Stephen M Chu. Joint optimization for consistent multiple graph matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1649–1656, 2013.
- Junchi Yan, Jun Wang, Hongyuan Zha, Xiaokang Yang, and Stephen Chu. Consistency-driven alternating optimization for multigraph matching: A unified approach. *IEEE Transactions on Image Processing*, 24(3):994–1009, 2015a.
- Junchi Yan, Hongteng Xu, Hongyuan Zha, Xiaokang Yang, Huanxi Liu, and Stephen Chu. A matrix decomposition perspective to multiple graph matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 199–207, 2015b.
- Junchi Yan, Minsu Cho, Hongyuan Zha, Xiaokang Yang, and Stephen M Chu. Multi-graph matching via affinity optimization with graduated consistency regularization. *IEEE transactions on pattern analysis and machine intelligence*, 38(6):1228–1242, 2016.
- Allen Y Yang, Subhransu Maji, C Mario Christoudias, Trevor Darrell, Jitendra Malik, and S Shankar Sstry. Multiple-view object recognition in smart camera networks. In *Distributed Video Sensor Networks*, pages 55–68. Springer, 2011.
- Duo Zhang, Benjamin IP Rubinstein, and Jim Gemmell. Principled graph matching algorithms for integrating multiple data sources. *IEEE Transactions on Knowledge and Data Engineering*, 27(10):2784–2796, 2015.

SUPPLEMENTARY MATERIAL

In this supplementary material, we prove Theorems 1 and 2 in the main paper. Theorem 3 can be proved using almost the same techniques as the proof for Theorem 1 and hence those details are omitted.

A. Proof of Theorem 1

Proof. Since we have mentioned that, under the case of this theorem the matrices P_{ij} satisfy the sum $\sum_{i=1}^n \sum_{j=1}^n \text{tr}(P_{ij}^\top T_{ij})$ reaches the optimal value for the objective in Equation (2) in the main paper, it is sufficient to show that the matrices A_1, \dots, A_n returned by Algorithm 1 satisfy $P_{ij} = A_i^\top A_j$ for each P_{ij} . We first show that, before the coordinate update part (line 9 to 12) of Algorithm 1, we have already ensured that the matrices A_1, \dots, A_n satisfy the property $P_{ij} = A_i^\top A_j$ for each P_{ij} . We will use induction to prove that, after each iteration during the initialization part of the Algorithm 1, for any set S_k and any $v_i, v_j \in S_k$, we have $P_{ij} = A_i^\top A_j$.

1. Initially (after the 0^{th} iteration), each set S_k only contains one vertex v_k . Since $P_{ii} = I = A_k^\top A_k$, the induction assumption is correct.
2. Assume the induction assumption is correct after the t^{th} iteration ($t \geq 0$). For the $(t+1)^{th}$ iteration, denote the edge we use in this iteration as (v_i, v_j) . Then, from the algorithm we know that the matrix $\hat{P} = \arg\max_P \text{tr}(P^\top A_i T_{ij} A_j^\top) = A_i P_{ij} A_j^\top$ for the old values of A_i and A_j . Therefore, after the update on line 5, we will get $P_{ij} = A_i^\top A_j$ for the new values of A_i and A_j . Since we are multiplying on the lefthand side the matrices $A_{j'}$'s on line 5 by the same matrix \hat{P} , this does not break the induction assumption inside the set S_j . After the update on line 5, for each $v_{i'} \in S_i$ and each $v_{j'} \in S_j$, we have $A_{i'}^\top A_{j'} = A_{i'}^\top A_i A_i^\top A_j A_j^\top A_{j'} = P_{i'i} P_{ij} P_{jj'} = P_{i'j'}$. Hence, after line 6 and 7, we know that for each $v_k \in S'$ (the set defined on the line 6) and each $v_{i'}, v_{j'} \in S_k$, we have $P_{i'j'} = A_{i'}^\top A_{j'}$. Since the permutation matrices that are changed during this iteration have their corresponding vertices in the set S' , we know that the induction assumption is correct after this iteration.

From 1, 2 we know that we have $P_{ij} = A_i^\top A_j$ for each P_{ij} after initialization. Since we have shown in the main paper that the Pairwise Alignment method can solve the problem optimally on this case, we know that our algorithm has also solved the problem optimally after initialization, and hence we do not have

any updates in the coordinate update part. Therefore, Algorithm 1 guarantees an optimal solution in this case. \square

B. Proof of Theorem 2

Proof. Ideally, we want to recover (A_1, \dots, A_n) such that $A_i^\top A_j = \hat{T}_{ij}$ for each pair (A_i, A_j) . Let us analyze the probability that we recover such a tuple of (A_1, \dots, A_n) under the model in Equation (5).

First, let us consider the probability that we recover the correct permutation matrices \hat{T}_{ij} from the optimization problem $\max_P \text{tr}(P^\top T_{ij})$ for any $i \neq j$. For any permutation matrix $P' \in \mathcal{P}_m$, $P' \neq T_{ij}$, if we denote k to be the number of entries where T_{ij} equals 1 but P' does not equal 1, then $k = \text{tr}((\hat{T}_{ij} - P')^\top \hat{T}_{ij})$. Therefore, $U := \frac{\text{tr}(P'^\top T_{ij}) - \text{tr}(\hat{T}_{ij}^\top T_{ij}) + k}{\eta_{ij}}$ follows the Chi-Square distribution $\chi^2(2k)$. Hence, the probability that P' is a better permutation matrix compared to \hat{T}_{ij} is

$$\begin{aligned} & \Pr[\text{tr}(P'^\top T_{ij}) \geq \text{tr}(\hat{T}_{ij}^\top T_{ij})] \\ &= \Pr[\eta_{ij}U - k \geq 0] = \Pr\left[\frac{U}{\mathbb{E}[U]} - 1 \geq \frac{1}{2}\left(\frac{1}{\eta_{ij}} - 2\right)\right]. \end{aligned} \quad (7)$$

For $\eta_{ij} \leq \frac{1}{10}$, by the Chi-Square tail bounds that Laurent and Massart (2000) proposed,

$$\begin{aligned} & \Pr\left[\frac{U_{ij}}{\mathbb{E}[U_{ij}]} - 1 \geq \frac{1}{2}\left(\frac{1}{\eta_{ij}} - 2\right)\right] \\ & \leq \Pr\left[\frac{U_{ij}}{\mathbb{E}[U_{ij}]} - 1 \geq \frac{1}{4}\left(\frac{1}{\eta_{ij}} - 2\right) + \sqrt{\frac{1}{2}\left(\frac{1}{\eta_{ij}} - 2\right)}\right] \quad (8) \\ & \leq \exp\left(-\frac{k}{4}\left(\frac{1}{\eta_{ij}} - 2\right)\right). \end{aligned}$$

Denote the probability of misaddressing k letters to k envelopes (The Bernoulli-Euler Problem of the Misaddressed Letters (Dörrie, 2013)) as $p_k = \sum_{i=0}^{\infty} \frac{(-1)^i}{i!} \leq \frac{1}{2}$ (for $k \geq 2$). Then, by union bound on Equation (8) for $k = 2, 3, \dots, n$, we know the probability that some $P' \neq \hat{T}_{ij}$ is better than \hat{T}_{ij} is at most

$$\begin{aligned} & \sum_{k=2}^m p_k \cdot \frac{m!}{(m-k)!} \cdot \exp\left(-\frac{k}{4}\left(\frac{1}{\eta_{ij}} - 2\right)\right) \\ & \leq \frac{1}{2} \sum_{k=2}^m m^k \cdot \exp\left(-\frac{k}{4}\left(\frac{1}{\eta_{ij}} - 2\right)\right). \end{aligned} \quad (9)$$

If we have $\eta_{ij} \leq \frac{1}{4(1+\varepsilon)\ln m+2}$ for some $\varepsilon > 0$, then,

$$\begin{aligned} & \frac{1}{2} \sum_{k=2}^m m^k \cdot \exp\left(-\frac{k}{4} \left(\frac{1}{\eta_{ij}} - 2\right)\right) \\ & \leq \frac{1}{2} \sum_{k=2}^m m^{-\varepsilon k} = \frac{m^{-2\varepsilon}}{2(1-m^{-\varepsilon})}. \end{aligned} \quad (10)$$

Hence, if we choose the variance parameter $\eta_{ij} \leq \min(\frac{1}{10}, \frac{1}{4(1+\varepsilon)\ln m+2})$ for some $\varepsilon > 0$, then for $m \geq 2^{\frac{1}{\varepsilon}}$ we have probability at least $1 - m^{-2\varepsilon}$ to guarantee that we recover \hat{T}_{ij} from the optimization problem $\max_P \text{tr}(P^\top T_{ij})$.

Therefore, if we assume that the number of element sets n is not too large as there exists some constant $\gamma > 0$ such that $n \leq m^\gamma$, then by union bound we know that with probability at least $1 - m^{-\delta}$ for any $\delta > 0$ that we can guarantee that using the Pairwise Alignment method recovers a correct solution if $m \geq 2^{\frac{2}{4\gamma+\delta}}$ and if we set each $\eta_{ij} \leq \min(\frac{1}{10}, \frac{1}{2(2+4\gamma+\delta)\ln m+2}) = O(\frac{1}{\log m})$.

Next let us consider the probability that our Algorithm 1 recovers the correct permutation matrices. We would only make some errors on the updates on line 5 and 11. Basically, if we don't make any error at any iteration at the step of computing \hat{P} on line 4 and don't make any updates on line 11, then we are sure that our algorithm solves the problem optimally.

Here we consider $m \geq 8$ such that $\frac{1}{10} > \frac{1}{4(1+\varepsilon)\ln m+2}$ for any $\varepsilon > 0$. Let us first bound the probability that we might make a mistake when computing the matrix \hat{P} on line 4. At each iteration when we are considering edge (v_i, v_j) , if we have $\eta_{ij} \leq \frac{1}{4(1+\varepsilon)\ln m+2}$ for any $\varepsilon > 0$, then from the above analysis we know that with probability at least $1 - m^{-2\varepsilon}$ we do not make mistakes on this step.

Otherwise, let us take $(i^*, j^*) = \underset{i' \in S_i, j' \in S_j}{\operatorname{argmin}} \eta_{i'j'}$ (S_i and S_j are the sets before being updated on line 7). If we have $\eta_{i^*j^*} \leq \frac{1}{8(1+\varepsilon)\ln m+4} \leq \frac{1}{4(1+\varepsilon)\ln m+2}$, then from the above analysis we know that with probability at least $1 - m^{-2\varepsilon}$ we get $\hat{T}_{i^*j^*}$ from the optimization problem $\max_P \text{tr}(P^\top T_{i^*j^*})$, and we also know that $\eta_{ij} - \eta_{i^*j^*} \geq \frac{1}{8(1+\varepsilon)\ln m+4}$.

Notice that (v_i, v_j) is an edge of the Maximum Spanning Tree of G . It must be the edge with largest edge weight between vertices in S_i and S_j . Therefore, we have $f(T_{ij}) \geq f(T_{i^*j^*})$. Conditioned on the cases that we recover $\hat{T}_{i^*j^*}$ from $\max_P \text{tr}(P^\top T_{i^*j^*})$ (we will omit some conditional probability notation from now on for brevity), and let $U \sim \chi^2(m)$ be a Chi-Square ran-

dom variable with free degree m , then by the Chi-Square tail bounds that Laurent and Massart (2000) proposed,

$$\begin{aligned} & \Pr[f(T_{i^*j^*}) \leq m(1 - \eta_{i^*j^*} - \frac{1}{16(1+\varepsilon)\ln m+8})] \\ & = \Pr[U - m \geq \frac{m}{\eta_{i^*j^*}(16(1+\varepsilon)\ln m+8)}] \\ & \leq \Pr[U - m \geq \frac{m}{2}] \leq \Pr[U - m \geq 0.48m] \\ & \leq \exp(-\frac{m}{25}) \leq m^{-2\varepsilon} \end{aligned} \quad (11)$$

for sufficiently large m . On the other hand, consider the value of $f(T_{ij})$, denote $P' = \underset{P}{\operatorname{argmax}} \text{tr}(P^\top T_{ij})$ and

k to be the number of entries where \hat{T}_{ij} equals 1 while P' does not equal 1 ($0 \leq k \leq m$). Since we require all $\eta_{ij} \leq O(1)$, let us assume that we have $\eta_{ij} \leq \frac{1}{3}$. Let $V_1 \sim \chi^2(k)$, $V_2 \sim \chi^2(m-k)$ be two independent Chi-Square random variables (we use $\chi^2(0)$ to be the random variable that only has support on a single point 0). Conditioned on k , the distribution of $f(T_{ij})$ is the same with $\eta_{ij}(V_1 - V_2) + m - k$. If $k > 0$, we know that

$$\begin{aligned} & \Pr[V_1 \geq k + \frac{m}{\eta_{ij}(32(1+\varepsilon)\ln m+16)}] \\ & \leq \Pr[V_1 - k \geq \frac{3m}{32(1+\varepsilon)\ln m+16}] \\ & \leq \Pr[V_1 - k \geq 2\sqrt{2k\varepsilon\ln m} + 4\varepsilon\ln m] \leq m^{-2\varepsilon} \end{aligned} \quad (12)$$

for sufficiently large m . Symmetrically, if $k < m$,

$$\begin{aligned} & \Pr[V_2 \leq (m-k) - \frac{n}{\eta_{ij}(32(1+\varepsilon)\ln m+16)}] \\ & \leq \Pr[(m-k) - V_2 \geq \frac{3m}{32(1+\varepsilon)\ln m+16}] \\ & \leq \Pr[(m-k) - V_2 \geq 2\sqrt{2k\varepsilon\ln m}] \leq m^{-2\varepsilon}. \end{aligned} \quad (13)$$

for sufficiently large m . Therefore, conditioned on k , if we have $\eta_{ij} \leq \frac{1}{3}$, we always have

$$\begin{aligned} & \Pr[\eta_{ij}(V_1 - V_2) + m - k \geq m(1 - \eta_{ij} + \frac{1}{16(1+\varepsilon)\ln m+8})] \\ & \leq \Pr[\eta_{ij}(V_1 - V_2) - (2k - m)\eta_{ij} \geq \frac{1}{16(1+\varepsilon)\ln m+8}] \\ & \leq \Pr[V_1 \geq k + \frac{m}{\eta_{ij}(32(1+\varepsilon)\ln m+16)}] \\ & + \Pr[V_2 \leq (m-k) - \frac{m}{\eta_{ij}(32(1+\varepsilon)\ln m+16)}] \leq 2m^{-2\varepsilon}. \end{aligned} \quad (14)$$

This is true for all k . Hence, without conditioning on k , we know that

$$\Pr[f(T_{ij}) \geq m(1 - \eta_{ij} + \frac{1}{16(1+\varepsilon)\ln m+8})] \leq 2m^{-2\varepsilon} \quad (15)$$

for sufficiently large m and if we have $\eta_{ij} \leq \frac{1}{3}$.

By union bound on Equations (11) and (15), we know that, conditioned on the cases where we recover $\hat{T}_{i^*j^*}$ from $\max_P \text{tr}(P^\top T_{i^*j^*})$, since $\eta_{ij} - \eta_{i^*j^*} \geq \frac{1}{8(1+\varepsilon)\ln m+4}$, we have

$$\begin{aligned} & \Pr[f(T_{ij}) \geq f(T_{i^*j^*})] \\ & \leq \Pr[f(T_{i^*j^*}) \leq m(1 - \eta_{i^*j^*} - \frac{1}{16(1+\varepsilon)\ln m+8})] \\ & \quad + \Pr[f(T_{ij}) \geq m(1 - \eta_{ij} + \frac{1}{16(1+\varepsilon)\ln m+8})] \\ & \leq 3m^{-2\varepsilon}. \end{aligned} \tag{16}$$

Since we know that, if we have $\eta_{i^*j^*} \leq \frac{1}{8(1+\varepsilon)\ln m+4}$, then with probability at least $1 - m^{-2\varepsilon}$ we would recover $\hat{T}_{i^*j^*}$ from $\max_P \text{tr}(P^\top T_{i^*j^*})$. Hence, conditioned on the case that $\eta_{ij} > \frac{1}{4(1+\varepsilon)\ln m+2}$, we know that the probability $\Pr[f(T_{ij}) \geq f(T_{i^*j^*})] \leq 3m^{-2\varepsilon} + m^{-2\varepsilon} = 4m^{-2\varepsilon}$. Plus the opposite case where $\eta_{ij} \leq \frac{1}{4(1+\varepsilon)\ln m+2}$, by union bound we know that the probability that we make an error during each iteration of the initialization part of Algorithm 1 is at most $5m^{-2\varepsilon}$. This is true under the condition that $\eta_{ij} \leq \frac{1}{3}$ and $\min_{i' \in S_i, j' \in S_j} \eta_{ij} \leq \frac{1}{8(1+\varepsilon)\ln m+4}$. To make these two conditions true, we impose the following two requirements:

- Consider an undirected weighted graph $G' = (V', E'')$, where there is a vertex v'_i for each element set X_i and there is an edge $(v'_i, v'_j) \in E''$ with edge weight η_{ij} . Then the bottleneck weight of the minimum bottleneck spanning tree of G' should be at most $\frac{1}{8(1+\varepsilon)\ln m+4}$.
- $\max_{i < j} \eta_{ij} \leq \frac{1}{3}$.

Therefore, under the above two conditions, assume the number of element sets m satisfy $n \leq m^\gamma$ for some constant $\gamma > 0$. Then, by union bound we know that, for sufficiently large n , the probability that we recover the correct solution for $(\hat{A}_1, \dots, \hat{A}_n)$ during the initialization part of the Algorithm 1 is $1 - 5m^{-2\varepsilon+\gamma}$.

For the coordinate update part of Algorithm 1 (line 9 to 12), let us consider the probability that we do not perform any updates conditioned on the case that we already have an optimal solution in the initialization part. For each step, denote the matrix we are optimizing as A_i . The update rule is Equation (3). Using the same approach as before, assume that there is some matrix $P' \neq A_i$ such that $\text{tr}(P'^\top \sum_{1 \leq j \leq n, i \neq j} A_j T_{ij}) \geq \text{tr}(A_i^\top \sum_{1 \leq j \leq n, i \neq j} A_j T_{ij})$. Denote k as the number of

entries where A_i equals 1 but P' does not. Also, denote $U_1, \dots, U_{i-1}, U_{i+1}, \dots, U_n$ to be independent random variables following the distribution $\chi^2(2k)$, and let U be a random variable following distribution $\chi^2(2k(n-1))$. Then, by the Chi-Square tail bounds that Laurent and Massart (2000) proposed

$$\begin{aligned} & \Pr[\text{tr}(P'^\top \sum_{1 \leq j \leq n, i \neq j} A_j T_{ij}) \geq \text{tr}(A_i^\top \sum_{1 \leq j \leq n, i \neq j} A_j T_{ij})] \\ & = \Pr[\sum_{1 \leq j \leq n, i \neq j} \eta_{ij} U_j \geq k(n-1)] \\ & \leq \Pr[\frac{1}{3}U \geq k(n-1)] \leq \exp(-\frac{2k(n-1)}{25}). \end{aligned} \tag{17}$$

Then, again by union bound on all values for k , we know that the probability that we might get a wrong answer for A_i in a single step is at most

$$\begin{aligned} & \sum_{k=2}^m p_k \cdot \frac{m!}{(m-k)!} \cdot \exp(-\frac{2k(n-1)}{25}) \\ & \leq \frac{1}{2} \sum_{k=2}^m m^k \cdot \exp(-\frac{2k(n-1)}{25}). \end{aligned} \tag{18}$$

If we have $n \geq 20 \ln m$, then for sufficient large value of m we have

$$\frac{1}{2} \sum_{k=2}^m m^k \cdot \exp(-\frac{2k(n-1)}{25}) \leq m^{-2}. \tag{19}$$

By union bound on all m matrices A_i 's, we know that the probability at least one of them needs updates is at most m^{-1} . Hence, we can solve the optimization problem with probability at least $1 - 5m^{-2\varepsilon+\gamma} + m^{-1}$ under all of the above constraints. If we set $\varepsilon = \frac{\gamma+1}{2}$, then the probability becomes $1 - 6m^{-1} = 1 - o(1)$ for sufficiently large m . Under that setting, we require the bottleneck weight of the minimum bottleneck spanning tree of G' to be at most $\frac{1}{8(1+\varepsilon)\ln m+4} = \frac{1}{4(3+\gamma)\ln m+4}$. \square