Journal of Machine Learning Research 20 (2019) 1-37 Submitted 10/18; Revised 3/19; Published 4/19

Learning to Match via Inverse Optimal Transport

Ruilin Li RUILIN.LI@GATECH.EDU
School of Mathematics
School of Computational Science and Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA

Xiaojing Ye XYE@GSU.EDU
Department of Mathematics and Statistics
Georgia State University
Atlanta, GA 30302, USA

Haomin Zhou HMZHOU@MATH.GATECH.EDU
School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332, USA

Hongyuan Zha ZHA@CC.GATECH.EDU
School of Computational Science and Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA

Editor: Xiaotong Shen

Abstract

We propose a unified data-driven framework based on inverse optimal transport that can learn adaptive, nonlinear interaction cost function from noisy and incomplete empirical matching matrix and predict new matching in various matching contexts. We emphasize that the discrete optimal transport plays the role of a variational principle which gives rise to an optimization based framework for modeling the observed empirical matching data. Our formulation leads to a non-convex optimization problem which can be solved efficiently by an alternating optimization method. A key novel aspect of our formulation is the incorporation of marginal relaxation via regularized Wasserstein distance, significantly improving the robustness of the method in the face of noisy or missing empirical matching data. Our model falls into the category of prescriptive models, which not only predict potential future matching, but is also able to explain what leads to empirical matching and quantifies the impact of changes in matching factors. The proposed approach has wide applicability including predicting matching in online dating, labor market, college application and crowdsourcing. We back up our claims with numerical experiments on both synthetic data and real world data sets.

Keywords: Matching, Inverse Problem, Optimal Transport, Robustification, Variational Inference

1. Introduction

Matching is a key problem at the heart of many real-world applications, including online dating (Hitsch et al., 2010), labor market (David, 2001), crowdsourcing (Yuen et al., 2011), marriage (Becker, 1973), paper-to-reviewer assignment (Charlin et al., 2011), kidney transplant donor matching (Dickerson and Sandholm, 2015) and ad allocation (Mehta et al., 2013). Owing to the wide applicability and great importance of matching, 2012 Nobel prize in economics were awarded to two economists Lloyd Shapley and Alvin Roth for their fundamental theoretic work (Gale and Shapley, 1962) and substantive empirical investigations, experiments and practical design (Roth and Sotomayor, 1989, 1992) on matching. A good matching of individuals from two sides (e.g., men vs. women, students vs. school, papers vs. reviewers) is essential to the overall health of the specific market/community. However, matching is a challenging problem due to two major complications: individuals from both sides exhibit various observable and latent features, which makes "suitability of a match" far more complex to assess; and the matching is implicitly, but significantly, influenced by the supply limitations of individuals from each side, so that the supply of an item can only satisfy a small number of users even though it is preferred by many. These two issues must be properly tackled in an optimal matching system.

In many matching problems, feature and preference data can be collected from individuals of either or both sides of the matching. Then a central planner may use such data sets to infer suitable matching or assignment. The feature and preference data collected in this way, however, can be incomplete, noisy, and biased for two reasons:

- an individual may not be aware of the competitors from her own side and/or limited quantity of her preferred match from the opposite side

- collection of a full spectrum of features is inherently difficult or even infeasible (e.g., a student's merit outside of her school curriculum in college admission, or religious belief of a person in a marriage, may not be included in the collected data)

The former factor prevents individuals from listing their orders of preferences and positioning themselves strategically in the market, and the latter results in feature data set that is incomplete and biased.

One possible approach is to use observed (and perhaps latent) features of individuals to generate rating matrix for user-item combinations as in many recomender systems (RS). However, this approach is not suitable given the bias and noise in collected feature or preference data and limited supply constraints in our matching problems. For example, in a standard movie RS problem, a movie can receive numerous high ratings and be watched by many people. In contrary, in a matching-based college admission problem, a student can enter only one college. Therefore, an optimal matching cannot be obtained solely based on personal ratings and preferences—the population of both sides also need to be taken into consideration in a matching problem. This significant difference between standard recommendation and matching demands for new theoretical and algorithmic developments.

Our approach to tackle the aforementioned challenges in matching inference is to consider a generalized framework based on inverse optimal transport, where the diversified population of each side of the matching is naturally modeled as a probability distribution, and the bilateral preference of two individuals in a potential match is captured by a matching reward (or equivalently, negative matching cost). More specifically, we obtain kernel representation of the cost by learning the feature interaction matrix from the matching data, under which the total social surplus is supposed to be maximal in a healthy matching market as suggested by economists (Carlier and Ekeland, 2010). Moreover, we employ a robust and flexible Wasserstein metric to learn feature-enriched marginal distributions, which proves to be very effective and robust in dealing with incomplete/noisy data in the matching problem.

From a broader perspective, our approach is in the framework of optimization based on variational principles—the observed data are results of some optimization with an unknown objective function (or a known objective function with unknown parameters) that models the problem, and the goal is to learn the objective function (or its parameters) from the data. This approach is a type of prescriptive analytics: it exploits the motivation and mechanism of the subject, and produces results that are interpretable and meaningful to human. The solution process is more instructive and can make use of the observed data more effectively. In this broader sense, our proposed approach based on inverse optimal transport is in a similar spirit as inverse reinforcement learning (Ng et al., 2000). Furthermore, the learned objective can be used to understand the effect of various factors in a matching and infer optimal matching strategy given new data. For instance, in online dating, riders allocation, and many other settings, the central planners (Tinder, OkCupid, Uber, Lyft, etc.) can use such prescriptive models to improve customer experience and align social good with their own profit goal.

Our work is the first to establish a systematic framework for optimal matching learning using incomplete, noisy data under *limited-supply constraints*. In particular, we advocate a nonlinear representation of cost/reward in a matching and view the matching strategy as a solution of (regularized) optimal transport. The equilibrium of certain matching markets, such as marriage, with simplifying assumptions, coincide with optimal transport plans (Becker, 1973). Even for matching markets with complex structure and factors, whose matching mechanism is not yet completely unveiled, the proposed model serves as a powerful modeling tool to study those matchings. In terms of algorithmic development, we derive a highly efficient learning method to estimate the parameters in the cost function representation in the presence of computationally complex Wasserstein metrics. Numerical results show that our method contrasts favorably to other matching approaches in terms of robustness and efficiency, and can be used to infer optimal matching for new data sets accurately.

The rest of this paper is organized as follows: we briefly summarize related work in Section 2 and review discrete optimal transport and its regularized version as well as their close connections in Section 3. Section 4 describes the setup of proposed model and introduces our robust formulation via regularized Wasserstein distance, which tries to capture matching mechanism by leveraging regularized optimal transport. The derivation of optimization algorithm is detailed in Section 5. We evaluate our model in section 6 on both synthetic data and real-world data sets. The last section concludes the paper and points to several directions for potential future research.

2. Related Work

In this section, we briefly summarize some related work, including matching, ecological inference, recommender systems, distance metric learning and reciprocal recommendation.

2.1. Matching

Matching has been widely studied in economics community since the seminal work of Koopmans and Beckmann (1957). Gale and Shapley (1962) studied optimal matching in college admission, marriage market and proposed the famous Gale-Shapley algorithm. Becker (1973) gave a theoretic analysis in marriage market matching. Roth and Sotomayor (1992) did a thorough study and analysis in two-sided matching. Chiappori et al. (2010); Carlier and Ekeland (2010) used optimal transport theory to study the equilibrium of certain matching markets. Galichon and Salanié (2010) theoretically justified the usage of entropy-regularized optimal transport plan to model empirical matching in the presence of unobserved characteristics. Another interesting work (Charlin et al., 2011) proposed to predict optimal matching from learning suitability score in paper-to-review context where they used well-known linear regression, collaborative filtering algorithms to learn suitability scores. There are also some work studying dynamic matching theory and applications such as kidney exchange (Dickerson et al., 2012; Dickerson and Sandholm, 2015) and barter exchange (Anderson et al., 2017; Ashlagi et al., 2017).

A recent work closely related to ours is (Dupuy et al., 2016), where they worked with regularized optimal transport plan and modeled the cost by a bilinear form using an affinity matrix learned from data. By contrast, our work models the cost using a nonlinear kernel representation and incorporate regularized Wasserstein distance to tackle the challenging issues due the incomplete and noisy data in real-world matching problems.

2.2. Ecological Inference

Ecological inference infers the nature of individual level behavior using aggregate (historically called “ecological”) data, and is of particular interest to political scientists, sociologists, historians and epidemiologists. Due to privacy or cost issue, individual level data are eluding from researchers, hence the inference made through aggregate data are often subject to ecological fallacy [1]. Previously, people proposed neighborhood model (Freedman et al., 1991), ecological regression (Goodman, 1953) and King's method (King, 2013). A recent progress (Flaxman et al., 2015) is made by using additional information and leverage kernel embedding of distributions, distribution regression to approach this problem.

Our work differs from classical ecological inference problem and methods in four ways. First, we assume access to empirical matching at individual-level granularity which is not available in standard ecological inference setting. Second, in our framework, we focus on learning the preference of two sides in the matching and propose a novel and efficient method to learn it, after which inference/prediction problem becomes trivial as preference is known. Third, different from previous statistical methods, we adopt a model-based approach, leverages optimal transport to model matching and draw a connection between these two fields. Lastly, thanks to the model-based approach, we are able to shed light on

[1] `https://en.wikipedia.org/wiki/Ecological_fallacy`

what factors lead to empirical matching and quantitatively estimate the influence caused by changes of those factors, which are beyond the reach of traditional statistical approaches.

2.3. Recommender Systems

Collaborative filtering (CF) type recommender systems share many similarities with optimal matching problem as both need to learn user preference from rating/matching data and predict rating/matching in a collaborative manner. Matrix-factorization based models (Mnih and Salakhutdinov, 2008; Salakhutdinov and Mnih, 2008) enjoyed great success in Netflix Prize Competition. Rendle (2010, 2012) proposed factorization machine model with strong sparse predictive power and ability to mimic several state-of-the-art, specific factorization methods. Recently there is trend of combining collaborative filtering with deep learning (He and Chua, 2017; He et al., 2017). Most items recommended by conventional recommender systems, however, are non-exclusive and can be consumed by many customers such as movies and music. They do not take supply limit of either or both sides into consideration hence may perform poorly in matching context.

2.4. Distance Metric Learning

Our model essentially aims to learn an adaptive, nonlinear representation of the matching cost. This is closely related to, but more general than, ground metric learning. Prior research on learning different distance metrics in various contexts are fruitful, such as learning cosine similarity for face verification (Nguyen and Bai, 2010), learning Mahalanobis distance for clustering (Xing et al., 2003) and face identification (Guillaumin et al., 2009). However, distance learning for optimal transport distance is largely unexplored. Cuturi and Avis (2014) proposed to learn the ground metric by minimizing the difference of two convex polyhedral functions. Wang and Guibas (2012) formulated a SVM-like minimization problem to learn Earth Mover's distance. Both approaches work with Wasserstein distance which involves solving linear programming as subroutine hence may be computationally too expensive. This paper works with regularized optimal transport distance, involving solving a matrix scaling problem as subroutine which is much lighter than linear programming.

2.5. Reciprocal Recommendation

Another line of related research is reciprocal recommendation (Brozovsky and Petricek, 2007; Pizzato et al., 2013), which also tries to model two-side preference by computing reciprocal score via a hand-craft score function. By a sharp contrast, our model learns how two sides interact with each other from observed noisy/incomplete matching in a data-driven fashion.

3. Background and Preliminaries

In this section, we present Kantorovich's formulation of optimal transportation problem (in discretized setting) and its regularized version.

3.1. Optimal Transport

Given two probability vectors $\boldsymbol{\mu} \in \Sigma_m$ and $\boldsymbol{\nu} \in \Sigma_n$, where $\Sigma_d := \{\boldsymbol{x} \in \mathbb{R}_+^d | \mathbf{1}^T \boldsymbol{x} = 1\}$ is the standard $(d-1)$-dimensional probability simplex, denote the transport polytope of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ by

$$U(\boldsymbol{\mu}, \boldsymbol{\nu}) := \{\pi \in \mathbb{R}_+^{m \times n} | \pi \mathbf{1} = \boldsymbol{\mu}, \pi^T \mathbf{1} = \boldsymbol{\nu}\}$$

namely the set of all $m \times n$ non-negative matrices satisfying marginal constraints specified by $\boldsymbol{\mu}, \boldsymbol{\nu}$. Note that $U(\boldsymbol{\mu}, \boldsymbol{\nu})$ is a convex, closed and bounded set containing joint probability distributions with $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ as marginals. Furthermore, if given a cost matrix $C = [C_{ij}] \in \mathbb{R}^{m \times n}$ where C_{ij} measures the cost of moving a unit mass from μ_i to ν_j, define

$$d(C, \boldsymbol{\mu}, \boldsymbol{\nu}) := \min_{\pi \in U(\boldsymbol{\mu}, \boldsymbol{\nu})} \langle \pi, C \rangle$$

where $\langle A, B \rangle = \mathrm{Tr}(A^T B)$ is the Frobenius inner product for matrices. This quantity describes how to optimally redistribute $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$ so that the total cost is minimized, hence providing a means to measure the similarity between the two distributions. In particular, when $C \in \mathcal{M}^d$, that is, C is in the cone of distance matrices (Brickell et al., 2008), defined as

$$\mathcal{M}^d := \{C \in \mathbb{R}_+^{d \times d} | C_{ii} = 0, C_{ij} = C_{ji}, C_{ij} \leq C_{ik} + C_{kj}, \forall i, j, k\},$$

then it is shown that $d(C)$ is a distance (or metric) on Σ_d (Villani, 2008), named the optimal transport distance (also known as the 1-Wasserstein distance or the earth mover distance). The minimizer π is called the optimal transport plan.

In discrete case, computing OT distance amounts to solving a linear programming problem, for which there exists dedicated algorithm with time complexity $\mathcal{O}(n^3 \log n)$ (Pele and Werman, 2009). Nevertheless, this is still too computationally expensive in large scale settings. In addition, OT plan π typically admits a sparse form which is not robust in data-driven applications. We refer readers to Villani (2008); Peyré et al. (2017) for a thorough theoretical and computational treatment of optimal transport.

3.2. Regularized Optimal Transport

To address the aforementioned computational difficulty, Cuturi (2013) proposed to use a computationally-friendly approximation of OT distance by introducing entropic regularization. This also mitigates the sparsity and improve the smoothness of OT plan. Concretely, consider

$$d_\lambda(C, \boldsymbol{\mu}, \boldsymbol{\nu}) := \min_{\pi \in U(\boldsymbol{\mu}, \boldsymbol{\nu})} \{\langle \pi, C \rangle - H(\pi)/\lambda\}$$

where $H(\pi)$ is the discrete entropy defined by

$$H(\pi) = -\sum_{i,j=1}^{m,n} \pi_{ij} (\log \pi_{ij} - 1),$$

and $\lambda > 0$ is the regularization parameter controlling the trade-off between sparsity and uniformity of π. We refer the above quantity as regularized optimal transport (ROT) distance (regularized Wasserstein distance) though it is *not* an actual distance measure.

Algorithm 1 Sinkhorn-Knopp Algorithm

```
Input: marginal distributions $\boldsymbol{\mu}, \boldsymbol{\nu}$, cost matrix $C$, regularization parameter $\lambda$
$K = \exp(-\lambda C)$
$\boldsymbol{a} = \mathbf{1}$
while not converge do
    $\boldsymbol{b} \leftarrow \frac{\boldsymbol{\nu}}{K^T \boldsymbol{a}}$
    $\boldsymbol{a} \leftarrow \frac{\boldsymbol{\mu}}{K \boldsymbol{b}}$
end while
$\pi = \mathbf{diag}(\boldsymbol{a}) K \, \mathbf{diag}(\boldsymbol{b})$
return $\pi, \boldsymbol{a}, \boldsymbol{b}$
```

Due to the strict convexity introduced by entropy, $d_\lambda(C, \boldsymbol{\mu}, \boldsymbol{\nu})$ admits a unique minimizer with full support $\pi^\lambda(C, \boldsymbol{\mu}, \boldsymbol{\nu})$, which we call regularized optimal transport plan in the sequel. The ROT plan π^λ has a semi-closed form solution

$$\pi^\lambda = \mathbf{diag}(\boldsymbol{a}) K \, \mathbf{diag}(\boldsymbol{b}) \tag{1}$$

where $\boldsymbol{a} \in \mathbb{R}^m, \boldsymbol{b} \in \mathbb{R}^n$ are positive vectors and are uniquely determined up to a multiplicative constant and $K := \exp(-\lambda C)$ is the component-wise exponential of $-\lambda C$. We can efficiently compute $\boldsymbol{a}$ and $\boldsymbol{b}$ by Sinkhorn-Knopp matrix scaling algorithm (Sinkhorn and Knopp, 1967), also known as iterative proportional fitting procedure (IPFP). The algorithm alternately scales rows and columns of K to fit the specified marginals. See Algorithm 1 for detailed description of the Sinkhorn-Knopp algorithm.

Not surprisingly, we have

$$\lim_{\lambda \to \infty} d_\lambda(C, \boldsymbol{\mu}, \boldsymbol{\nu}) = d(C, \boldsymbol{\mu}, \boldsymbol{\nu})$$

ROT distance converges to OT distance as λ tends to infinity, i.e., entropic regularization diminishes. Moreover, let

$$\Pi(C, \boldsymbol{\mu}, \boldsymbol{\nu}) = \{\pi | \langle \pi, C \rangle = \min_{\pi \in U(\boldsymbol{\mu}, \boldsymbol{\nu})} \langle \pi, C \rangle\}$$

be the set of all OT plan and

$$\pi^\star = \operatorname*{arg\,max}_{\pi \in \Pi(C, \boldsymbol{\mu}, \boldsymbol{\nu})} H(\pi)$$

be the joint distribution with highest entropy within $\Pi(C, \boldsymbol{\mu}, \boldsymbol{\nu})$, then

$$\lim_{\lambda \to \infty} \pi^\lambda = \pi^\star$$

in another word, ROT plan converges to the most uniform OT plan and the rate of convergence is exponential, as shown by Cominetti and San Martín (1994). The generalization of entropic regularization, Tsallis entropy regularized optimal transport also receives more and more attention and is studied by Muzellec et al. (2017).

ROT has more favorable computational properties than OT does, as it only involves component-wise operation and matrix-vector multiplication, all of which are of quadratic complexity, and can be parallelized (Cuturi, 2013). This fact makes ROT popular for measuring dissimilarity between potentially unnormalized distributions in many research fields: machine learning (Geneway et al., 2017; Rolet et al., 2016; Laclau et al., 2017), computer vision (Cuturi and Doucet, 2014) and image processing (Papadakis, 2015).

Besides computation efficiency, we argue in next section why it is more appropriate to use ROT in our setting from a modeling perspective.

4. Learning to Match

For the ease of exposition, we refer two sides of the matching market as users and items. The methodology is suitable in various applications where optimal matching is considered under supply limitations, such as marriage market, cab hailing, college admission, organ allocation, paper matching ans so on. Suppose we have m user profiles $\{\boldsymbol{u}_i\}_{i\in[m]} \subset \mathbb{R}^p$, n item profiles $\{\boldsymbol{v}_j\}_{j\in[n]} \subset \mathbb{R}^q$ and N_{ij}, the count of times $(\boldsymbol{u}_i, \boldsymbol{v}_j)$ appears in matching. Let $N = \sum_{i,j=1}^{m,n} N_{ij}$ be the number of all matchings, $[\hat{\pi}_{ij}] = [N_{ij}/N]$ be the observed matching matrix and $\hat{\boldsymbol{\mu}} = \hat{\pi}\mathbf{1}, \hat{\boldsymbol{\nu}} = \hat{\pi}^T\mathbf{1}$ be the sample marginals. Suppose we are also given two cost matrices C_u and C_v, measuring user-user dissimilarity and item-item dissimilarity respectively, we can then select two appropriate constants λ_u and λ_v and use $d_{\lambda_u}(C_u, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ and $d_{\lambda_v}(C_v, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2)$ to measure the dissimilarity of probability distributions $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ over user profile space and that of $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2$ over item profile space.

4.1. Modeling Observed Matching Matrix

Becker (1973) pointed out that equilibrium of some matching markets coincide with optimal transport plans which are often highly sparse. The implication of this theory is far from being realistic, though, as we observe heterogeneous matchings in real world. Galichon and Salanié (2015) argued that there are latent features having significant impact on matching but unfortunately unobservable to researchers. Hence they proposed to leverage a combination of pure optimal transport plan and mutual information of two sides of matching to model empirical matching data which is exactly entropy-regularized optimal transport.

Furthermore, the observed matching matrix $\hat{\pi}$ (hence the empirical marginals) often contains noisy, corrupted, and/or missing entries, consequently it is more robust to employ a regularized optimal transport plan rather than enforce an exact matching to empirical data in cost function learning.

To that end, we propose to use regularized optimal transport plan $\pi^{\lambda}(C, \boldsymbol{\mu}, \boldsymbol{\nu})$ in our learning task. This also has several important benefits that take the following aspects into modeling consideration in addition to unobserved latent features:

- **Enforced Diversity.** Diversity is enforced in certain matchings as is the case when admission committee making decisions on applicants, diversity is often an important criterion and underrepresented minorities may be preferred. Entropy term captures the uncertainty introduced by diversity. The idea of connecting entropy with matching to capture/promote diversity is also adopted, for example, by Agrawal et al. (2018) and Ahmed et al. (2017).

- **Aggregated Data.** Sometimes due to privacy issues or insufficient number of matched pairs, only grouped or aggregated data, rather than individual data are available. Accordingly, the aggregated matching is usually denser than individual level matching and is less likely to exhibit sparsity.

4.2. Cost Function via Kernel Representation

The cost function $C_{ij} = c(\boldsymbol{u}_i, \boldsymbol{v}_j)$ is of critical importance as it determines utility loss of user $\boldsymbol{u}_i$ and item $\boldsymbol{v}_j$. The lower the cost is, the more likely user $\boldsymbol{u}_i$ will match item $\boldsymbol{v}_j$, subject to supply limit of items. A main contribution of this work is to learn an adaptive, nonlinear representation of the cost function from empirical matching data. To that end, we present several properties of cost function in optimal matching that support the feasibility.

First of all, we show in the following proposition that the cost function C is not unique in general but can be uniquely determined in a special and important case.

Proposition 1 *Given two marginal probability vectors $\boldsymbol{\mu} \in \Sigma_m$, $\boldsymbol{\nu} \in \Sigma_n$, define $F : \mathbb{R}^{m\times n} \to U(\boldsymbol{\mu}, \boldsymbol{\nu})$, $F(C) = \pi^\lambda(C, \boldsymbol{\mu}, \boldsymbol{\nu})$ is the ROT plan of C. Then F is in general not injective, however, when $m = n$ and F is restricted on $\mathcal{M}^n$, $F_{|\mathcal{M}^n}(C)$ is injective.*

Proof One can easily verify that F is well-defined from the strict convexity of ROT. The optimality condition of ROT reads as

$$\pi^\lambda(C, \boldsymbol{\mu}, \boldsymbol{\nu}) = \exp(\lambda(-C + \boldsymbol{a}\mathbf{1}^T + \mathbf{1}\boldsymbol{b}^T))$$

where $\boldsymbol{a} \in \mathbb{R}^m$ and $\boldsymbol{b} \in \mathbb{R}^n$ are Lagrangian multipliers dependent on C and λ such that $\pi^\lambda(C, \boldsymbol{\mu}, \boldsymbol{\nu}) \in U(\boldsymbol{\mu}, \boldsymbol{\nu})$. Therefore, $\pi^\lambda(C + \epsilon\mathbf{1}\mathbf{1}^T, \boldsymbol{\mu}, \boldsymbol{\nu}) = \exp(\lambda(-C - \epsilon\mathbf{1}\mathbf{1}^T + (\boldsymbol{a} + \epsilon\mathbf{1})\mathbf{1}^T + \mathbf{1}\boldsymbol{b}^T)) = \pi^\lambda(C, \boldsymbol{\mu}, \boldsymbol{\nu})$ for any $\epsilon > 0$. Therefore F is in general not injective.

If $m = n$ and $C_1, C_2 \in \mathcal{M}^n$, by the semi-closed form (1) of ROT plan, there exist positive vectors $\boldsymbol{a}_1, \boldsymbol{b}_1$ and $\boldsymbol{a}_2, \boldsymbol{b}_2$ such that

$$\pi^\lambda(C_1, \boldsymbol{\mu}, \boldsymbol{\nu}) = \mathbf{diag}(\boldsymbol{a}_1)\exp(-\lambda C_1)\,\mathbf{diag}(\boldsymbol{b}_1)$$
$$\pi^\lambda(C_2, \boldsymbol{\mu}, \boldsymbol{\nu}) = \mathbf{diag}(\boldsymbol{a}_2)\exp(-\lambda C_2)\,\mathbf{diag}(\boldsymbol{b}_2)$$

If $\pi^\lambda(C_1, \boldsymbol{\mu}, \boldsymbol{\nu}) = \pi^\lambda(C_2, \boldsymbol{\mu}, \boldsymbol{\nu})$, we have

$$\exp(-\lambda C_1) = \mathbf{diag}(\boldsymbol{a})\exp(-\lambda C_2)\,\mathbf{diag}(\boldsymbol{b})$$

where $\exp(\cdot)$ is component-wise exponential, $\boldsymbol{a} = \log\frac{\boldsymbol{a}_2}{\boldsymbol{a}_1}$, $\boldsymbol{b} = \log\frac{\boldsymbol{b}_2}{\boldsymbol{b}_1}$.

Since C_1, C_2 are symmetric matrices, it follows that $\boldsymbol{a} = s\boldsymbol{b}$. By appropriately rescaling $\boldsymbol{a}$ and $\boldsymbol{b}$ to make them equal, we have

$$\exp(-\lambda C_1) = \mathbf{diag}(\boldsymbol{w})\exp(-\lambda C_2)\,\mathbf{diag}(\boldsymbol{w})$$

where $\boldsymbol{w} = \boldsymbol{a}/\sqrt{s}$. Inspecting (i, i) entry of both sides, we immediately conclude that $\boldsymbol{w} = \mathbf{1}$ and $C_1 = C_2$. ■

Actually, the general non-uniqueness or non-identifiability of cost C is quite natural. For instance, in an online auction setting, if all bidders raise their bids by the same amount, the result of the auction will not change because the rank of bidders remain the same and the original winner still wins the auction. Therefore, by observing empirical matching alone, we can not determine cost matrix definitively without further assumption. Proposition 1 guarantees the uniqueness of learned cost if we model it as a distance matrix, e.g. Mahalanobis distance ($C_{ij} = \sqrt{(\boldsymbol{u}_i - \boldsymbol{v}_j)^T M (\boldsymbol{u}_i - \boldsymbol{v}_j)}$, where M is a positive definite matrix). However, in many cases, cost may grow nonlinearly in the difference of features. An even more serious issue is that if the number of features of two sides of matching are inconsistent or two sides do not lie in the same feature space at all, it would be infeasible to use a distance metric to capture the cost between them due to such dimension incompatibility.

Therefore, as generalized distance functions (Schölkopf, 2001), kernel representation which is able to measure matching cost even when features of two sides do not lie in the same feature space can be leveraged to model the cost function, i.e.,

$$c(\boldsymbol{u}_i, \boldsymbol{v}_j) = k(G\boldsymbol{u}_i, D\boldsymbol{v}_j)$$

where $k(\boldsymbol{x}, \boldsymbol{y})$ is a specific (possibly nonlinear) kernel, $G \in \mathbb{R}^{r\times p}$ and $D \in \mathbb{R}^{r\times q}$ are two unknown linear transformations to be learned. $G\boldsymbol{u}, D\boldsymbol{v}$ can be interpreted as the latent profile associated with users and items and are studied by Agarwal and Chen (2009).

For a wide class of commonly used kernels including linear kernel, polynomial kernel and sigmoid kernel, they depend only on the inner product of two arguments through an activation function f, i.e. $k(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}^T\boldsymbol{y})$. For such kernels, we have

$$c(\boldsymbol{u}_i, \boldsymbol{v}_j) = f(\boldsymbol{u}_i^T G^T D \boldsymbol{v}_j)$$

and it suffices to learn $A = G^T D$. In this case, cost matrix

$$C(A) = f(U^T A V)$$

is parametrized by A and we refer A as interaction matrix. Here we apply f component-wise on $U^T AV$. For ease of presentation, we will work with kernels of this form in the sequel. With kernel function representation, it is still likely that a matching matrix corresponds to multiple cost matrices, and we will be contented with finding one of them that explains the observed empirical matching.

4.3. Kernel Inference with Wasserstein Marginal Regularization

A straight forward way to learn $C(A)$ in kernel representation is estimating parameter A through minimizing negative log likelihood

$$\min_A - \sum_{i=1}^{m} \sum_{j=1}^{n} \hat{\pi}_{ij} \log \pi_{ij} \tag{2}$$

where $\pi = \pi^\lambda(C(A), \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\nu}})$, i.e., one enforces the optimal plan π to satisfy $\pi\mathbf{1} = \hat{\boldsymbol{\mu}}$ and $\pi^T\mathbf{1} = \hat{\boldsymbol{\nu}}$. Note that (2) is equivalent to minimizing the reverse Kullback-Leibler divergence (Bishop, 2006) of ROT plan π with respect to empirical matching $\hat{\pi}$, i.e.,

$$\min_A \mathrm{KL}(\hat{\pi}\|\pi)$$

This is the formulation proposed in Dupuy et al. (2016) which we refer as inverse optimal transport formulation (IOT) in the sequel.

In this variation principle based framework, the ROT plan π has the same marginals as the empirical matching $\hat{\pi}$ does, which is reasonable if the marginal information of empirical matching is sufficiently accurate. In practice, however, the size of samples available is hardly enough, hence the empirical marginals inferred from samples can be incomplete and noisy. To see why this is the case, suppose the ground space $\mathcal{D} = \{1, -1\}^d$, $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{R}^{2^d}$ are two discrete probability distributions over $\mathcal{D}$ and $\pi^0 \in \mathbb{R}^{2^d \times 2^d}$ is the ground truth matching matrix between $\boldsymbol{\mu}, \boldsymbol{\nu}$. Let $\{X_1, X_2, \cdots, X_N\} \subset \mathbb{R}^{2^d \times 2^d}$ be i.i.d random matrices, where $X_1 = E_{ij}$ with probability π^0_{ij}. Then empirical matching matrix is given by $\hat{\pi} = \bar{X} = \frac{1}{N}\sum_{n=1}^N X_n$ and two empirical marginals are $\hat{\boldsymbol{\mu}} = \bar{X}\mathbf{1}, \hat{\boldsymbol{\nu}} = \bar{X}^T\mathbf{1}$. By Bretagnolle-Huber-Carol inequality (Bretagnolle and Huber, 1979; Wellner et al., 2013), for any $\epsilon > 0$, we have

$$\mathbb{P}(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_1 > \epsilon) < 2^{2^d} \exp(\frac{-N}{2}\epsilon^2) = \exp(2^d \ln 2 - \frac{\epsilon^2}{2}N)$$
$$\mathbb{P}(\|\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}\|_1 > \epsilon) < 2^{2^d} \exp(\frac{-N}{2}\epsilon^2) = \exp(2^d \ln 2 - \frac{\epsilon^2}{2}N)$$

Hence

$$\begin{aligned}\mathbb{P}(\max\{\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_1, \|\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}\|_1\} > \epsilon) &= \mathbb{P}(\{\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_1 > \epsilon\} \cup \{\|\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}\|_1 > \epsilon\}) \\ &\leq \mathbb{P}(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_1 > \epsilon) + \mathbb{P}(\|\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}\|_1 > \epsilon) \\ &< 2\exp(2^d \ln 2 - \frac{\epsilon^2}{2}N)\end{aligned}$$

To ensure with at least $1 - \delta$ probability, the maximum error in marginal distributions $\max\{\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_1, \|\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}\|_1\} \leq \epsilon$, we might need $N(\epsilon, \delta) = \frac{2}{\epsilon^2}(2^d \ln 2 + \ln \frac{2}{\delta})$ samples. Note that $N(\epsilon, \delta)$ is quadratic in $\frac{1}{\epsilon}$ and exponential in d, indicating that practically there are hardly enough samples for us to obtain accurate estimate of marginal distributions $\boldsymbol{\mu}, \boldsymbol{\nu}$.

If using IOT formulation with noisy estimate of marginal distributions, it causes a *systematic error* no smaller than $\max\{\|\Delta\boldsymbol{\mu}\|_1, \|\Delta\boldsymbol{\nu}\|_1\}$ as shown in proposition 2.

Proposition 2 *If empirical $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\nu}}$ are off from true $\boldsymbol{\mu}, \boldsymbol{\nu}$ by $\Delta\boldsymbol{\mu}, \Delta\boldsymbol{\nu}$, then the matching matrix π_{IOT} recovered by solving equation* (2) *has error lower bounded by*

$$\|\pi_0 - \pi_{IOT}\|_1 \geq \max\{\|\Delta\boldsymbol{\mu}\|_1, \|\Delta\boldsymbol{\nu}\|_1\}$$

where $\|\pi\|_1 = \sum_{i,j=1}^{m,n} |\pi_{ij}|$, $\hat{\boldsymbol{\mu}}, \boldsymbol{\mu} \in \mathbb{R}^m$ and $\hat{\boldsymbol{\nu}}, \boldsymbol{\nu} \in \mathbb{R}^n$, π_0 is the ground truth matching matrix, $\pi_{IOT} = \pi^\lambda(C(A^\star), \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\nu}})$ and $A^\star$ is the solution of equation (2).

Proof We know $\pi_0 \in U(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $\pi_{\text{IOT}} \in U(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\nu}})$. By triangle inequality, we have

$$\sum_{i=1}^m \sum_{j=1}^n |(\pi_0)_{ij} - (\pi_{IOT})_{ij}| \geq \sum_{i=1}^m |\sum_{j=1}^n (\pi_0)_{ij} - (\pi_{IOT})_{ij}| = \sum_{i=1}^m |\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i| = \|\Delta\boldsymbol{\mu}\|_1$$

Interchanging the summation and applying triangle inequality again, we obtain

$$\sum_{j=1}^{n}\sum_{i=1}^{m} |(\pi_0)_{ij} - (\pi_{IOT})_{ij}| \geq \|\Delta\boldsymbol{\nu}\|_1$$

Therefore, we conclude

$$\|\pi_0 - \pi_{\mathrm{IOT}}\|_1 \geq \max\{\|\Delta\boldsymbol{\mu}\|_1, \|\Delta\boldsymbol{\nu}\|_1\}$$

■

We have seen that inaccurate marginal information can serious harm the recovery performance of ground truth matching matrix. Not unexpectedly, it could mislead us to learn an inaccurate cost matrix as well, as stated in proposition 4.

Lemma 3 *Suppose $M \in \mathbb{R}^{m\times n}$ and $f(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{a}\mathbf{1}^T + \mathbf{1}\boldsymbol{b}^T - M\|_F^2$. Then we have*

$$f(\boldsymbol{a}, \boldsymbol{b}) \geq \|M\|_F^2 - \boldsymbol{f}^T A^+ \boldsymbol{f}$$

where $\boldsymbol{f} = [(M\mathbf{1})^T, \mathbf{1}^T M]^T$, $A = \begin{bmatrix} nI_{m\times m} & \mathbf{1}_m\mathbf{1}_n^T \\ \mathbf{1}_n\mathbf{1}_m^T & mI_{n\times n}\end{bmatrix}$, A^+ is the Moore-Penrose inverse of matrix A and $\|M\|_F = \sqrt{\sum_{i=1}^m\sum_{j=1}^n M_{ij}^2}$ is Frobenius norm. In particular, if M can not be written as $M = \boldsymbol{a}\mathbf{1}^T + \mathbf{1}\boldsymbol{b}^T$, the lower bound is strictly positive, i.e.,

$$f(\boldsymbol{a}, \boldsymbol{b}) \geq \|M\|_F^2 - \boldsymbol{f}^T A^+ \boldsymbol{f} > 0$$

Proof See Appendix A. ■

Proposition 4 *Suppose $\pi_0 \in \mathbb{R}^{m\times n}$ is the ground truth matching matrix, $\hat{\pi} \in \mathbb{R}^{m\times n}$ is an empirical matching matrix. Let C_0 be the ground truth cost matrix giving rise to π_0 and $C_{IOT} = \arg\min_{C\in\mathbb{R}^{m\times n}} KL(\hat{\pi}||C, \hat{\pi}\mathbf{1}, \hat{\pi}^T\mathbf{1})$ be the learned cost matrix via IOT formulation that gives rise to $\hat{\pi}$, i.e. $\pi_0 = \pi^\lambda(C_0, \pi_0\mathbf{1}, \pi_0^T\mathbf{1})$ and $\hat{\pi} = \pi^\lambda(C_{IOT}, \hat{\pi}\mathbf{1}, \hat{\pi}^T\mathbf{1})$. Denote $\Delta C = C_0 - C_{IOT}$ and $\Delta\log\pi = \log\pi_0 - \log\hat{\pi}$ and further assume $(\Delta\log\pi)_{ij}$ are independent (absolutely) continuous random variables (w.r.t. Lebesgue measure), we have*

$$\|\Delta C\|_F^2 \geq \frac{1}{\lambda^2}(\|\Delta\log\pi\|_F^2 - \boldsymbol{f}^T A^+ \boldsymbol{f}) > 0 \qquad a.e. \tag{3}$$

where $\boldsymbol{f} = [(\Delta\log\pi\mathbf{1})^T, \mathbf{1}^T\Delta\log\pi]^T$, $A = \begin{bmatrix} nI_{m\times m} & \mathbf{1}_m\mathbf{1}_n^T \\ \mathbf{1}_n\mathbf{1}_m^T & mI_{n\times n}\end{bmatrix}$, A^+ is the Moore-Penrose inverse of matrix A and $\|M\|_F = \sqrt{\sum_{i=1}^m\sum_{j=1}^n M_{ij}^2}$ is Frobenius norm.

Proof First we show that $\min\limits_{C\in\mathbb{R}^{m\times n}} KL(\hat{\pi}||C, \hat{\pi}\mathbf{1}, \hat{\pi}^T\mathbf{1}) = 0$, i.e., $\hat{\pi} = \pi^\lambda(C_{IOT}, \hat{\pi}\mathbf{1}^T, \hat{\pi}^T\mathbf{1})$, any empirical matching can be realized as regularized optimal transport plan for some cost matrix.

Given arbitrary $\boldsymbol{\alpha} \in \mathbb{R}^m, \boldsymbol{\beta} \in \mathbb{R}^n$, let $\tilde{C} = \boldsymbol{\alpha}\mathbf{1}^\top + \mathbf{1}\boldsymbol{\beta}^\top - \frac{1}{\lambda}\log\hat{\pi}$, we see easily verify that $\hat{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}$ solve the KKT condition of ROT

$$\min_{\pi \in U(\hat{\pi}\mathbf{1}, \hat{\pi}^T\mathbf{1})} \langle \tilde{C}, \pi \rangle - \frac{H(\pi)}{\lambda}$$

Hence $\hat{\pi} = \pi^\lambda(\tilde{C}, \hat{\pi}\mathbf{1}^T, \hat{\pi}^T\mathbf{1})$ and $\min\limits_{C \in \mathbb{R}^{m\times n}} KL(\hat{\pi}||C, \hat{\pi}\mathbf{1}, \hat{\pi}^T\mathbf{1}) = 0$. Therefore, any minimizer C_{IOT} (need not be $\tilde{C}$) of the problem $\min\limits_{C \in \mathbb{R}^{m\times n}} KL(\hat{\pi}||C, \hat{\pi}\mathbf{1}, \hat{\pi}^T\mathbf{1})$ must satisfy

$$\hat{\pi} = \pi^\lambda(C_{IOT}, \hat{\pi}\mathbf{1}^T, \hat{\pi}^T\mathbf{1})$$

By the optimality condition of ROT, we know that there exist $\boldsymbol{a}, \boldsymbol{b}$ such that

$$\pi^\lambda(C, \boldsymbol{\mu}, \boldsymbol{\nu}) = \exp(\lambda(-C + \boldsymbol{a}^T\mathbf{1} + \mathbf{1}\boldsymbol{b}^T))$$

hence there exist $\boldsymbol{a}_0, \boldsymbol{b}_0, \hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}$ such that

$$C_0 = \boldsymbol{a}_0\mathbf{1}^T + \mathbf{1}\boldsymbol{b}_0^T - \frac{1}{\lambda}\log\pi_0$$
$$C_{IOT} = \hat{\boldsymbol{a}}\mathbf{1}^T + \mathbf{1}\hat{\boldsymbol{b}}^T - \frac{1}{\lambda}\log\hat{\pi}$$

Take difference and denote $\boldsymbol{a}_0 - \hat{\boldsymbol{a}}, \boldsymbol{b}_0 - \hat{\boldsymbol{b}}$ by $\Delta\boldsymbol{a}, \Delta\boldsymbol{b}$ respectively, we have

$$\Delta C = \Delta\boldsymbol{a}\mathbf{1}^T + \mathbf{1}\Delta\boldsymbol{b}^T - \frac{1}{\lambda}\Delta\log\pi$$

Since singular matrices have zero Lebesgue measure and $(\Delta\log\pi)_{ij}$ are independent continuous random variables, we have

$$\mathbb{P}(\Delta\log\pi = \boldsymbol{a}\mathbf{1}^T + \mathbf{1}\boldsymbol{b}^T) \leq \mathbb{P}(\det(\Delta\log\pi) = 0) = 0$$

By lemma 3, we obtain

$$\|\Delta C\|_F^2 \geq \frac{1}{\lambda^2}(\|\Delta\log\pi\|_F^2 - \boldsymbol{f}^T A^+ \boldsymbol{f}) > 0 \qquad a.e.$$

■

If we use the inaccurate cost matrix learned via IOT approach, it could negatively affect the quality of future matching prediction, as justified in proposition 5.

Proposition 5 *Let C_0 be any ground truth cost matrix, C_{IOT} be any learned cost matrix via IOT formulation and assume $C_{IOT} \notin \{C | C = C_0 + \boldsymbol{a}\mathbf{1}^T + \mathbf{1}\boldsymbol{b}^T$ for some $\boldsymbol{a}, \boldsymbol{b}\}$. Suppose the ground truth matching matrix is $\pi_0 = \pi^\lambda(C_0, \boldsymbol{\mu}, \boldsymbol{\nu})$ and the predicted matching matrix is $\pi_{predict} = \pi^\lambda(C_{IOT}, \boldsymbol{\mu}, \boldsymbol{\nu})$. Denote $\Delta C = C_0 - C_{IOT}$ and $\Delta\log\pi = \log\pi_0 - \log\pi_{predict}$, we have*

$$\|\Delta\log\pi\|_F^2 \geq \lambda^2(\|\Delta C\|_F^2 - \boldsymbol{f}^T A^+ \boldsymbol{f}) > 0 \tag{4}$$

where $\boldsymbol{f} = [(\Delta C\mathbf{1})^T, \mathbf{1}^T\Delta C]^T$, $A = \begin{bmatrix} nI_{m\times m} & \mathbf{1}_m\mathbf{1}_n^T \\ \mathbf{1}_n\mathbf{1}_m^T & mI_{n\times n} \end{bmatrix}$, A^+ is the Moore-Penrose inverse of matrix A and $\|M\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n M_{ij}^2}$ is Frobenius norm.

Proof The proof is almost identical to that of proposition 4 except for interchanging the role of ΔC and $\Delta \log \pi$. We hence omit the details here. ■

To address aforementioned issues, we hence propose a more robust formulation with Wasserstein marginal relaxation, dropping the hard marginal constraint. Concretely, we consider the following optimization problem.

$$\min_{A,\boldsymbol{\mu}\in\Sigma_m,\boldsymbol{\nu}\in\Sigma_n} -\sum_{i=1}^{m}\sum_{j=1}^{n}\hat{\pi}_{ij}\log\pi_{ij} + \delta\big(d_{\lambda_u}(C_u,\boldsymbol{\mu},\hat{\boldsymbol{\mu}}) + d_{\lambda_v}(C_v,\boldsymbol{\nu},\hat{\boldsymbol{\nu}})\big) \tag{5}$$

where $\pi = \pi^{\lambda}(C(A),\boldsymbol{\mu},\boldsymbol{\nu})$ is the regularized optimal transport plan, δ is the relaxation parameter controlling the fitness of marginals, $\lambda, \lambda_u, \lambda_v$ are hyper-parameters controlling the regularity of regularized Wasserstein distance. We refer this formulation as robust inverse optimal transport (RIOT) in the sequel. Interestingly, we note that Chizat et al. (2016) proposed a similar but different formulation in solving unbalanced optimal transport problem.

The intuition of this RIOT formulation is that instead of enforcing noisy empirical marginals as hard constraints, we incorporate them as soft constraints in objective function. We use regularized Wasserstein distance as regularization because of the following reasons:

- as approximated Wasserstein distance, it drives $\boldsymbol{\mu},\boldsymbol{\nu}$ to $\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\nu}}$, but at the same time it also allows some uncertainty hence is able to robustify the result;
- in presence of missing entries in marginals, Wassertein distance is still well defined while other measures such as KL are not;
- Wasserstein distance can be applied to continuous, discrete, or even mixed distributions;
- computation of regularized Wasserstein distance (Cuturi, 2013) is efficient and hence potentially more scalable for large scale problem (5) in practice.

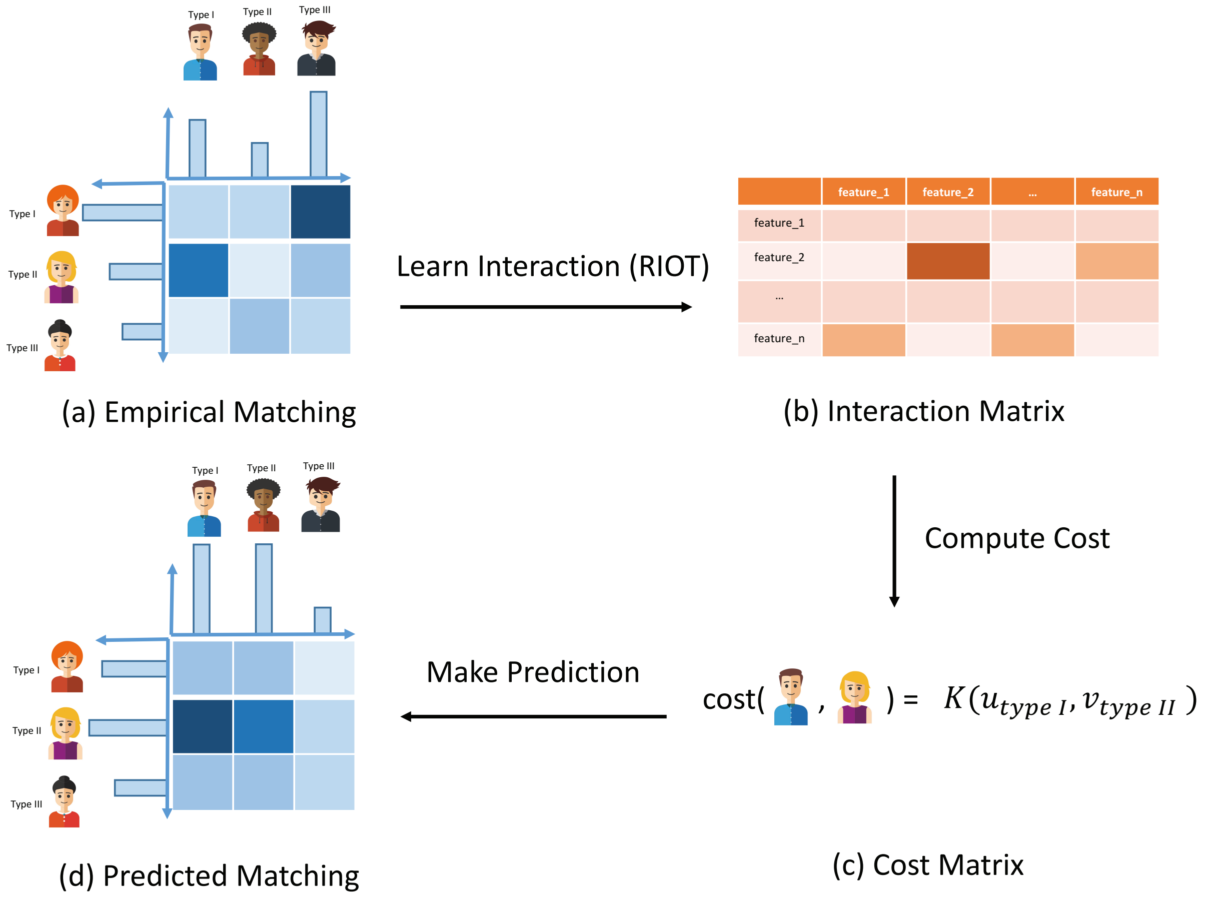
To be precise, the robustness in RIOT specifically means that RIOT outperforms IOT, in terms of the quality of learned matching matrix and cost matrix, when the observed joint distributions $\hat{\pi}$ and marginal distributions $\hat{\mu}$ and $\hat{\nu}$ are inaccurate due to insufficient and noisy samples, which is generally the case in real-world applications. In this situation, enforcing hard constraints to the inaccurate distributions is shown to cause severe bias in the estimation of the ground cost C, which will further induce errors in the matching when applied to testing data. The robustness of RIOT over IOT is verified by empirical experiment results in Section 6.1.

We assume access to C_u and C_v in our model because learning user-user/item-item similarity is relatively easier than our task, there are many existing work dedicated to that end (Cheung and Tian, 2004; Agarwal and Bharadwaj, 2013) and we want to single out and highlight our main contribution—learning the cost matrix that gives rise to observed matching and leverage it to infer matching for new data sets. In fact, our framework can also be extended to learn C_u and C_v jointly if needed, the optimization algorithm of which tends to be much more complex, though. See Appendix B for the extension. We postpone the detailed algorithmic derivation of the solution to (5) to next section.

4.4. Predict New Matching

After obtaining interaction matrix A from solving RIOT, we may then leverage it to predict new matching. Concretely, for a group of new users $\{\tilde{\boldsymbol{u}}_i\}_{i\in[m]}$ and items $\{\tilde{\boldsymbol{v}}_j\}_{j\in[n]}$, two marginal distributions, i.e., users profile distribution $\tilde{\boldsymbol{\mu}}$ and item profile distribution $\tilde{\boldsymbol{\nu}}$ can be easily obtained. First compute the cost matrix $\tilde{C}_{ij} = f(\tilde{\boldsymbol{u}}_i^T A \tilde{\boldsymbol{v}}_j)$ using kernel representation and apply Sinkhorn-Knopp algorithm to computing $\tilde{\pi}^{\lambda}(\tilde{C}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\nu}})$, which gives us the predicted matching of the given groups of users and items.

See Figure 1 for illustration of the complete pipeline or proposed learning-to-match framework.



(a) Empirical Matching

(b) Interaction Matrix

(c) Cost Matrix

(d) Predicted Matching

Figure 1: From noisy (a) empirical matching matrix, we learn (b) the interaction matrix via our proposed RIOT formulation. We then use kernel representation to compute (c) cost matrix and predict (d) matching matrix for new data. $\boldsymbol{u}_{type\ I}, \boldsymbol{v}_{type\ II}$ are feature vectors of type I men and type II women.

5. Derivation of Optimization Algorithm

Since the constraint set of ROT problem satisfies Slater's condition (Boyd and Vandenberghe, 2004), we have by strong duality that

$$d_\lambda(C, \boldsymbol{\mu}, \boldsymbol{\nu}) = \max_{\boldsymbol{z}} \langle \boldsymbol{z}, \boldsymbol{\mu} \rangle + \langle \boldsymbol{z}^C, \boldsymbol{\nu} \rangle - \frac{1}{\lambda}$$

where $z_j^C = \frac{1}{\lambda}\log c_j - \frac{1}{\lambda}\log(\sum_{i=1}^m e^{\lambda(z_i - C_{ij})})$. $\boldsymbol{z}, \boldsymbol{z}^C$ are essentially the Lagrangian multipliers corresponding to constraints $\pi\mathbf{1} = \boldsymbol{\mu}$ and $\pi^T\mathbf{1} = \boldsymbol{\nu}$. See also Genevay et al. (2016). Hence we have

$$d_{\lambda_u}(C_u, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \max_{\boldsymbol{z}} \langle \boldsymbol{z}, \boldsymbol{\mu} \rangle + \langle \boldsymbol{z}^{C_u}, \hat{\boldsymbol{\mu}} \rangle - \frac{1}{\lambda_u}$$

$$d_{\lambda_v}(C_v, \boldsymbol{\nu}, \hat{\boldsymbol{\nu}}) = \max_{\boldsymbol{w}} \langle \boldsymbol{w}, \boldsymbol{\nu} \rangle + \langle \boldsymbol{w}^{C_v}, \hat{\boldsymbol{\nu}} \rangle - \frac{1}{\lambda_v}$$

where $z_j^{C_u} = \frac{1}{\lambda_u}\log \hat{r}_j - \frac{1}{\lambda_u}\log(\sum_{i=1}^m e^{\lambda_u(z_i - C_{u ij})})$ and $w_j^{C_v} = \frac{1}{\lambda_v}\log \hat{c}_j - \frac{1}{\lambda_v}\log(\sum_{i=1}^n e^{\lambda_v(w_i - C_{v ij})})$. Given sample marginals, once $\boldsymbol{z}, \boldsymbol{w}$ are fixed, $\boldsymbol{z}^{C_u}, \boldsymbol{w}^{C_v}$ are also fixed. We can then convert (5) into a min-max problem

$$\min_{A,\boldsymbol{\mu},\boldsymbol{\nu}} \max_{\boldsymbol{z},\boldsymbol{w}} - \sum_{i=1}^m \sum_{j=1}^n \hat{\pi}_{ij} \log \pi_{ij} + \delta\big(\langle \boldsymbol{z}, \boldsymbol{\mu} \rangle + \langle \boldsymbol{z}^{C_u}, \hat{\boldsymbol{\mu}} \rangle + \langle \boldsymbol{w}, \boldsymbol{\nu} \rangle + \langle \boldsymbol{w}^{C_v}, \hat{\boldsymbol{\nu}} \rangle\big) \tag{6}$$

where constants are omitted. The optimal solution is a saddle-point of the objective in (6). To solve this min-max problem, we alternately update the primal variable $(A, \boldsymbol{\mu}, \boldsymbol{\nu})$ and dual variable $(\boldsymbol{z}, \boldsymbol{w})$, each time with the other ones fixed.

5.1. Update $(A, \boldsymbol{\mu}, \boldsymbol{\nu})$ for fixed $(\boldsymbol{z}, \boldsymbol{w})$

Now $\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{z}^{C_u}, \boldsymbol{w}^{C_v}$ are all fixed. Note that

$$\pi_{ij} = e^{\lambda(a_i + b_j - C_{ij})}$$

for some positive vectors $\boldsymbol{a}, \boldsymbol{b}$, such that $\pi\mathbf{1} = \boldsymbol{\mu}$, $\pi^T\mathbf{1} = \boldsymbol{\nu}$, and $\mathbf{1}^T\pi\mathbf{1} = 1$. Thus we may rewrite the minimization in this stage as

$$\begin{aligned} \min_{A,\boldsymbol{a},\boldsymbol{b}} \quad & \sum_{j=1}^n \hat{\pi}_{ij} \log \pi_{ij} + \delta(\langle z, \pi\mathbf{1} \rangle + \langle w, \pi^T\mathbf{1} \rangle) \\ \text{s.t.} \quad & \sum_{i=1}^m \sum_{j=1}^n e^{\lambda(a_i + b_j - C_{ij})} = 1 \end{aligned} \tag{7}$$

For fixed A, denote the optimum of objective function in equation (7) subject to the constraint by $E(C(A))$. Recall that the ultimate goal in this step is to find the interaction matrix A that cost C depends on, such that the minimum above can be attained. For any A, we have kernel representation $C(A)$ parameterized by interaction matrix A. Therefore the minimization above is equivalent to

$$\min_A E(C(A))$$

To minimize $E(C(A))$, the critical step is to evaluate gradient $\nabla_A E(C(A))$ and by envelope theorem (Milgrom and Segal, 2002) we have

Proposition 6 *The gradient $\nabla_A E(C(A))$ is*

$$\nabla_A E = \sum_{i=1}^{m}\sum_{j=1}^{n} \lambda[\hat{\pi}_{ij} + (\theta - \delta(z_i + w_j))\pi_{ij}]C'_{ij}(A)$$

where θ is the Lagrangian multiplier of the constrained minimization problem in equation (7)*.*

Proof By chain rule, we have that

$$\nabla_A E = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\partial E}{\partial C_{ij}} \frac{\partial C_{ij}}{A}$$

With the kernel representation, $C'_{ij}(A)$ is easily available. For fixed $C = C(A)$, by envelop theorem (Milgrom and Segal, 2002), we have

$$\begin{aligned}
\nabla_{C_{ij}} E(C) &= \frac{\partial}{\partial C} - \sum_{i=1}^{m}\sum_{j=1}^{n} \hat{\pi}_{ij}\log \pi_{ij} + \delta\langle \boldsymbol{z}, \pi\mathbf{1}\rangle + \delta\langle \boldsymbol{w}, \pi^T\mathbf{1}\rangle - \theta(\sum_{i,j=1}^{m,n} e^{\lambda(a_i + b_j - C_{ij})}) \\
&= (-\frac{\hat{\pi}_{ij}}{\pi_{ij}} + \delta(z_i + w_j) - \theta)\frac{\partial \pi_{ij}}{\partial C_{ij}} \\
&= \lambda[\hat{\pi}_{ij} + (\theta - \delta(z_i + w_j))\pi_{ij}]
\end{aligned}$$

■

Hence in each evaluation of $\nabla_C E$, we need to solve $E(C(A))$ once. If we denote $\xi_i = e^{\lambda a_i}$, $\eta_j = e^{\lambda b_j}$, $Z_{ij} = e^{-\lambda C_{ij}}$ and $M_{ij} = \delta(z_i + w_j)Z_{ij}$, then computing $E(C(A))$ is equivalent to solving

$$\begin{aligned}
\min_{\boldsymbol{\xi},\boldsymbol{\eta}} \quad & -\langle \hat{\boldsymbol{\mu}}, \log\boldsymbol{\xi}\rangle - \langle \hat{\boldsymbol{\nu}}, \log\boldsymbol{\eta}\rangle + \boldsymbol{\xi}^T M\boldsymbol{\eta} \\
\text{s.t.} \quad & \boldsymbol{\xi}^T Z\boldsymbol{\eta} = 1
\end{aligned} \tag{8}$$

Note that this is a non-convex optimization problem, both the objective function and constraints are non-convex which is difficult to solve in general. However, once we fix $\boldsymbol{\eta}$, the problem with respect to $\boldsymbol{\xi}$ alone is a convex problem and vice versa. We can solve this problem efficiently by alternately updating $\boldsymbol{\xi}, \boldsymbol{\eta}$.

Proposition 7 *Denote the objective in equation* (8) *by*

$$h(\boldsymbol{\xi}, \boldsymbol{\eta}) = -\langle \hat{\boldsymbol{\mu}}, \log\boldsymbol{\xi}\rangle - \langle \hat{\boldsymbol{\nu}}, \log\boldsymbol{\eta}\rangle + \boldsymbol{\xi}^T M\boldsymbol{\eta}$$

Initialize $\boldsymbol{\xi}^{(0)}, \boldsymbol{\eta}^{(0)}$ and alternately update $\boldsymbol{\xi}^{(k)}, \boldsymbol{\eta}^{(k)}$ in the following fashion

$$\boldsymbol{\xi}^{(k)} = \operatorname*{arg\,min}_{\boldsymbol{\xi}^T Z\boldsymbol{\eta}^{(k-1)}=1} h(\boldsymbol{\xi}, \boldsymbol{\eta}^{(k-1)}) \tag{9}$$

$$\boldsymbol{\eta}^{(k)} = \underset{{\boldsymbol{\xi}^{(k)}}^T Z\boldsymbol{\eta}=1}{\arg\min}\; h(\boldsymbol{\xi}^{(k)}, \boldsymbol{\eta}) \tag{10}$$

If $\lim_{k\to\infty}(\boldsymbol{\xi}^{(k)}, \boldsymbol{\eta}^{(k)}, \theta_1^{(k)}, \theta_2^{(k)}) = (\boldsymbol{\xi}^\star, \boldsymbol{\eta}^\star, \theta_1^\star, \theta_2^\star)$*, then* $\theta_1^\star = \theta_2^\star$ *and* $(\boldsymbol{\xi}^\star, \boldsymbol{\eta}^\star)$ *is a local minimizer of* h*, where* $\theta_1^{(k)}$ *and* $\theta_2^{(k)}$ *are Lagrangian multipliers corresponding to problem* (9) *and* (10) *respectively.*

Proof From the definition of $E(C(A))$ in equation (7) it is easily seen that $h(\boldsymbol{\xi}, \boldsymbol{\eta})$ is lower bounded. Moreover, since

$$h(\boldsymbol{\xi}^{(k)}, \boldsymbol{\eta}^{(k)}) \le h(\boldsymbol{\xi}^{(k)}, \boldsymbol{\eta}^{(k-1)}) \le h(\boldsymbol{\xi}^{(k-1)}, \boldsymbol{\eta}^{(k-1)})$$

there exists a convergent subsequence of $\{h(\boldsymbol{\xi}^{(k)}, \boldsymbol{\eta}^{(k)})\}$ and we denote the limit by $h^\star$.

The KKT condition of equation (9) and (10) are

$$-\frac{\hat{\boldsymbol{\mu}}}{\boldsymbol{\xi}^{(k)}} + M\boldsymbol{\eta}^{(k-1)} - \theta_1^{(k)} Z\boldsymbol{\eta}^{(k-1)} = 0 \tag{11}$$

$${\boldsymbol{\xi}^{(k)}}^T Z\boldsymbol{\eta}^{(k-1)} = 1$$

$$-\frac{\hat{\boldsymbol{\nu}}}{\boldsymbol{\eta}^{(k)}} + M^T\boldsymbol{\xi}^{(k)} - \theta_2^{(k)} Z^T\boldsymbol{\xi}^{(k)} = 0 \tag{12}$$

$${\boldsymbol{\xi}^{(k)}}^T Z\boldsymbol{\eta}^{(k)} = 1$$

Let k tend to infinity and take inner product with $\boldsymbol{\xi}^\star$ for equation (11) and take inner product with $\boldsymbol{\eta}^\star$ for equation (12), compare two equations and use the fact that both $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\nu}}$ are probability vectors, we find that $\theta_1^\star = \theta_2^\star$ and $(\boldsymbol{\xi}^\star, \boldsymbol{\eta}^\star)$ solves the KKT condition of constrained problem (9) and (10). Therefore, $(\boldsymbol{\xi}^\star, \boldsymbol{\eta}^\star)$ is a local minimizer and $h^\star = h(\boldsymbol{\xi}^\star, \boldsymbol{\eta}^\star)$ is a local minimum. ■

Once we obtain $(\boldsymbol{\xi}^\star, \boldsymbol{\eta}^\star, \theta^\star)$, we can then get $(\boldsymbol{a}, \boldsymbol{b}, \theta)$ by setting $\boldsymbol{a} = \frac{1}{\lambda}\log\boldsymbol{\xi}^\star$ and $\boldsymbol{b} = \frac{1}{\lambda}\log\boldsymbol{\eta}^\star$ and $\theta = \theta^\star$. Then plug in $(\boldsymbol{a}, \boldsymbol{b}, \theta)$ to evaluate $\nabla_C E$ for the current $C = C(A)$.

A careful analysis of the KKT condition of equations (9) and (10) shows that $\theta_1^{(k)}$ and $\theta_2^{(k)}$ are roots of

$$p(\theta) = \left\langle \frac{\hat{\boldsymbol{\mu}} \odot (Z\boldsymbol{\eta}^{(k-1)})}{(M - \theta Z)\boldsymbol{\eta}^{(k-1)}}, \mathbf{1} \right\rangle, \quad q(\theta) = \left\langle \frac{\hat{\boldsymbol{\nu}} \odot (Z^T\boldsymbol{\xi}^{(k)})}{(M - \theta Z)^T\boldsymbol{\xi}^{(k)}}, \mathbf{1} \right\rangle$$

respectively. The univariate root finding problem can be solved efficiently by off-the-shelf package. After obtaining $\theta_1^{(k)}, \theta_2^{(k)}$, we can update

$$\boldsymbol{\xi}^{(k)} = \frac{\hat{\boldsymbol{\mu}}}{(M - \theta_1^{(k)} Z)\boldsymbol{\eta}^{(k-1)}}, \quad \boldsymbol{\eta}^{(k)} = \frac{\hat{\boldsymbol{\nu}}}{(M - \theta_2^{(k)} Z)^T\boldsymbol{\xi}^{(k)}}$$

directly. Computationally, this approach to solving problem (9) and (10) is much cheaper than gradient-type iterative methods when m and/or n are large.

5.2. Update $(\boldsymbol{z}, \boldsymbol{w})$ for fixed $(A, \boldsymbol{\mu}, \boldsymbol{\nu})$

When $(A, \boldsymbol{\mu}, \boldsymbol{\nu})$ are fixed, π is also fixed, we then only need to solve

$$\max_{\boldsymbol{z},\boldsymbol{w}} \langle \boldsymbol{z}, \pi \mathbf{1} \rangle + \langle \boldsymbol{z}^{C_u}, \hat{\boldsymbol{\mu}} \rangle + \langle \boldsymbol{w}, \pi^T \mathbf{1} \rangle + \langle \boldsymbol{w}^{C_v}, \hat{\boldsymbol{\nu}} \rangle$$

and one immediately recognizes this is equivalent to applying Sinkhorn-Knopp algorithm to compute $d_{\lambda_u}(C_u, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$ and $d_{\lambda_v}(C_v, \boldsymbol{\nu}, \hat{\boldsymbol{\nu}})$.

To summarize, in each iteration, we perform a gradient-type update for A, followed by two calls of Sinkhorn-Knopp algorithm to compute $d_{\lambda_u}(C_u, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$ and $d_{\lambda_v}(C_v, \boldsymbol{\nu}, \hat{\boldsymbol{\nu}})$. Algorithm 2 details the algorithm. Note that this algorithm only finds a local minimum of equation 5.

Algorithm 2 Solve RIOT

```
Input: observed matching matrix $\hat{\pi}$, cost matrices $C_u, C_v$, regularization parameter $\lambda, \lambda_u, \lambda_v$
for $l = 1, 2, \cdots, L$ do
    $Z \leftarrow \exp(-\lambda C)$
    $M \leftarrow \delta(\boldsymbol{z}\mathbf{1}^T + \mathbf{1}\boldsymbol{w}^T) \odot Z$
    Initialize $\boldsymbol{\xi}^{(0)}, \boldsymbol{\eta}^{(0)}$
    for $k = 1, 2, \cdots, K$ do
        $\theta_1^{(k)} \leftarrow$ root of $p(\theta)$
        $\theta_2^{(k)} \leftarrow$ root of $q(\theta)$
        $\boldsymbol{\xi}^{(k)} \leftarrow \frac{\hat{\boldsymbol{\mu}}}{(M-\theta_1^{(k)} Z)\boldsymbol{\eta}^{(k-1)}}$
        $\boldsymbol{\eta}^{(k)} \leftarrow \frac{\hat{\boldsymbol{\nu}}}{(M-\theta_2^{(k)} Z)^T\boldsymbol{\xi}^{(k)}}$
    end for
    $\boldsymbol{a} \leftarrow \frac{1}{\lambda}\log \boldsymbol{\xi}^{(k)}, \quad \boldsymbol{b} \leftarrow \frac{1}{\lambda}\boldsymbol{\eta}^{(k)}, \quad \theta = \theta_1^{(k)}$
    $\pi \leftarrow \exp(\lambda(\boldsymbol{a}\mathbf{1}^T + \mathbf{1}\boldsymbol{b}^T - C))$
    $\nabla_A \leftarrow \sum_{i,j=1}^{m,n} \lambda[\hat{\pi}_{ij} + (\theta - \delta(z_i + w_j)\pi_{ij}]C'_{ij}(A)$
    $A \leftarrow A - s\nabla_A$
    $\boldsymbol{a}_1 \leftarrow$ Sinkhorn-Knopp$(C_u, \pi\mathbf{1}, \hat{\boldsymbol{\mu}}, \lambda_u)[1]$
    $\boldsymbol{a}_2 \leftarrow$ Sinkhorn-Knopp$(C_v, \pi^T\mathbf{1}, \hat{\boldsymbol{\nu}}, \lambda_v)[1]$
    $\boldsymbol{z} \leftarrow \frac{1}{\lambda_u}\log \boldsymbol{a}_1, \quad \boldsymbol{w} \leftarrow \frac{1}{\lambda_v}\log \boldsymbol{a}_2$
end for
```

6. Experiments

In this section, we evaluate our proposed RIOT model on both synthetic data and real world data sets. For synthetic data set, we illustrate its robustness against IOT and show our model can achieve better performance in learning cost matrix C than IOT could. For election data set, we show our method can effectively learn meaningful preference of voters based on their demographics. For taxi trip data set, we demonstrate that the proposed model is able to predict matching of taxi drivers and passengers fairly accurate. For marriage data set, we demonstrate the applicability of RIOT in predicting new matching and make recommendation accordingly by comparing it with baseline and state-of-art recommender systems.

6.1. Synthetic Data

We set $\lambda = \lambda_u = \lambda_v = 1$ and simulate $m = 10$ user profiles $\{\boldsymbol{u}_i\} \subset \mathbb{R}^{10}$, $n = 10$ item profiles $\{\boldsymbol{v}_j\} \subset \mathbb{R}^8$, two probability vectors $\boldsymbol{\mu}_0, \boldsymbol{\nu}_0 \in \mathbb{R}^{10}$, an interaction matrix A_0 of size 10×8 and pick polynomial kernel $k(\boldsymbol{x}, \boldsymbol{y}) = (\gamma \boldsymbol{x}^T \boldsymbol{y} + c_0)^d$ where $\gamma = 0.05, c_0 = 1, d = 2$, hence $C_{0ij} = (0.05\boldsymbol{u}_i^T A \boldsymbol{v}_j + 1)^2$. For C_u, C_v, we randomly generate m and n points from $\mathcal{N}(\boldsymbol{0}, 5I_2)$ on plane and use their Euclidean distance matrix as C_u and C_v. The ground truth entropy-regularized optimal transport plan is given by $\pi_0 = \pi^\lambda(C_0, \boldsymbol{\mu}_0, \boldsymbol{\nu}_0)$. We independently sample N samples from π_0 and then compute empirical matching matrix $\hat{\pi}$ accordingly. In algorithm 2, we set the number of iterations of inner loop $K = 20$.

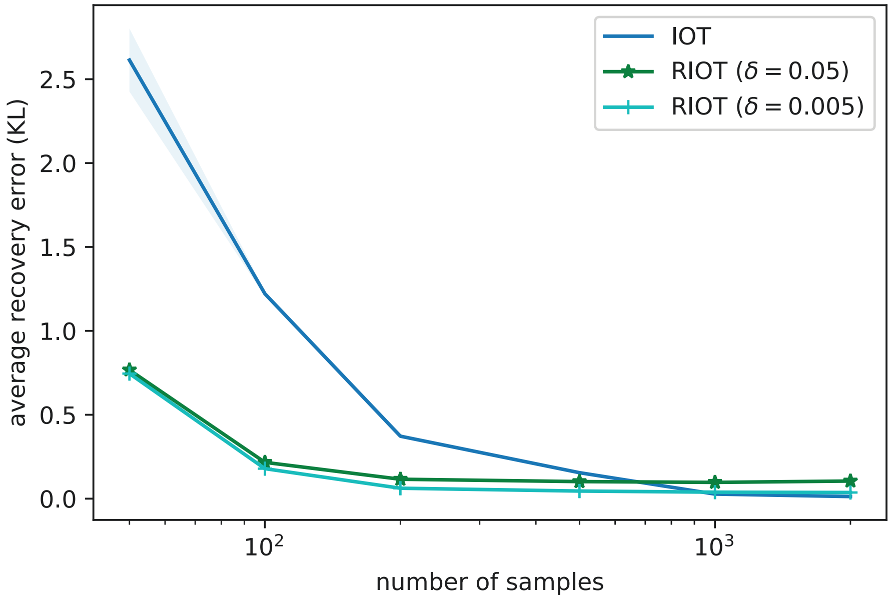
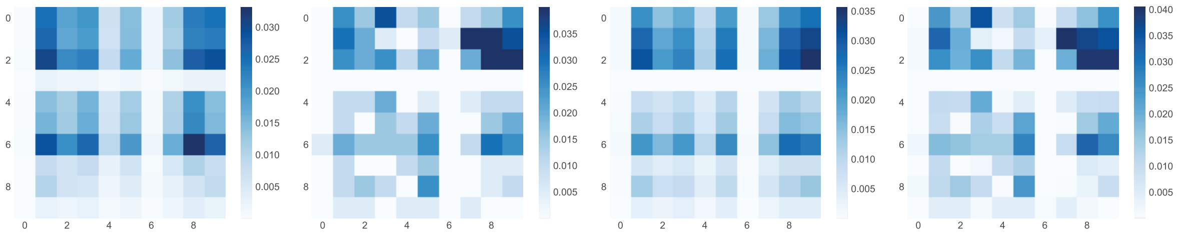


Figure 2: Comparison of recovery performance of fixed marginal approach (IOT) and marginal relaxation approach (RIOT). For sample size N, we sample from π_0, compute empiricla matching $\hat{\pi}$ and run algorithm 2 for 50 times. The shaded region is one standard deviation.

6.1.1. Improved Robustness

To produce Figure 2, we set the number of iterations in outer loop $L = 50$, learning rate $s = 10$. For each $N \in \{50, 100, 200, 500, 1000, 2000\}$ we run algorithm 2 and record Kullback-Leibler divergence between learned matching matrix π_{IOT}, π_{RIOT} and ground truth matching matrix π_0. Figure 2 shows that RIOT with different relaxation parameters



(a) π_0 (b) $\hat{\pi}$ (c) π_{RIOT} (d) π_{IOT}

Figure 3: Comparison of (a) actual matching matrix π_0 (b) noised matching matrix $\hat{\pi}$ ($\text{KL}(\pi_0\|\hat{\pi}) = 1.269$) (c) matching matrix π_{RIOT} learned by marginal relaxation approach ($\text{KL}(\pi_0\|\pi_{\text{RIOT}}) = 0.219$) (d) matching matrix π_{IOT} learned with fixed marginal approach ($\text{KL}(\pi_0\|\pi_{\text{IOT}}) = 0.464$) (sample size $N = 200$)

demonstrate improved robustness than IOT does when noise size is large. If δ is set too large (e.g. $\delta = 0.05$), however, the entropy term tends to dominate and negatively affect recovery performance when noise size is modestly small. If δ is tuned carefully (e.g. $\delta = 0.005$), RIOT can achieve comparable performance even when noise size is quite small. Moreover, we observe that curves corresponding to different δ intersect with the curve of fixed marginal at different noise size. Therefore, when prior knowledge or statistical estimate of noise size is available, we may tune δ accordingly to achieve best practical performance. In addition, we observe notable variation of KL divergence between π_0 and π_{IOT} when sample size is small (e.g., $N = 50$), in contrast, one standard deviation of $KL(\pi_0||\pi_{RIOT})$ is negligible, supporting our argument of RIOT being more robust than IOT.

To produce Figure 3, we set sample size $N = 200$ and relaxation parameter $\delta = 0.001$ with other parameters same as those for producing Figure 2. Figure 3 visually illustrates π_0, $\hat{\pi}$, π_{RIOT} and π_{IOT} and we see that when sample size is small, sampling noise can significantly corrupts= the ground truth matching matrix. π_{RIOT} exhibits less distortion compared to π_{IOT}, which demonstrates improved robustness again. Numerical results also back up our observation.

$$\text{KL}(\pi_0\|\hat{\pi}) = 1.269,\ \text{KL}(\pi_0\|\pi_{\text{RIOT}}) = 0.219,\ \text{KL}(\pi_0\|\pi_{\text{IOT}}) = 0.464$$

Compared to IOT, marginal relaxation via regularized Wasserstein distance does help improve the robustness of solution.

6.1.2. Superior Learning Performance

To produce Figure 4, we set sample size $N = 200$, relaxation parameter $\delta = 0.001$, the number of iterations in outer loop $L = 100$ and learning rate $s = 1$, we then run algorithm 2 to compare the performance of learning cost matrix C_0. To avoid non-uniqueness/non-identifiability issue, we use

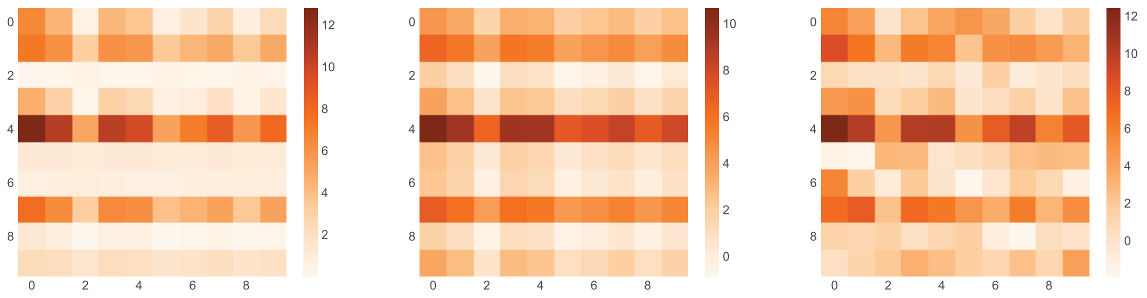
$$d(C_1, C_2) = \min_{D=\boldsymbol{a}\mathbf{1}^T+\mathbf{1}\boldsymbol{b}^T+C_1} \|D - C_2)\|_F$$

to measure the closeness of cost matrices C_1 and C_2 and denote the minimizer of $d(C, C_0)$ by $\tilde{C}$. The results are shown below,

$$d(C_{RIOT}, C_0) = 7.831, \quad d(C_{IOT}, C_0) = 12.439$$

where C_{RIOT}, C_{IOT} are cost matrices learned by RIOT formulation and IOT formulation respectively. Compared to C_{IOT}, C_{RIOT} learned via our proposed method almost halves the distance to ground truth cost matrix. Figure 4 also illustrates that our model can learn the structure of cost matrix better than IOT does. Our approach improves the learning performance and is able to reveal the structure of ground truth cost matrix.

To sum up, we show that with appropriately tuned relaxation parameter, RIOT is superior to IOT in terms of both robustness and learning performance.



(a) C_0 (b) $\tilde{C}_{\text{RIOT}}$ (c) $\tilde{C}_{\text{IOT}}$

Figure 4: Comparison of (a) ground truth cost matrix C_0 (b) $\tilde{C}_{\text{RIOT}}$, the minimizer of $d(C_{RIOT}, C_0)$ and (c) $\tilde{C}_{\text{IOT}}$, the minimizer of $d(C_{IOT}, C_0)$

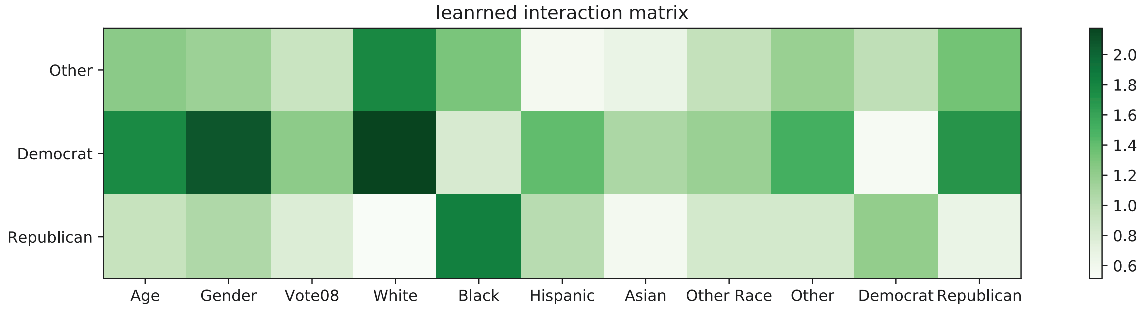


Figure 5: Interaction matrix learned by RIOT for election data set

6.2. Election data set

We show in this subsection that RIOT can effectively learn the user-item interaction by applying it to 2012 presidential election data of Florida. The experiment setup is similar to that of Muzellec et al. (2017)[2].

The data set contains more than 9.23×10^6 voters profile, each voter has features like gender, age, race, party and whether the voter voted in 2008 election. 0-1 encoding are used for gender (M:1, F:0) and voting in 2008 (Yes:1, No:0), age is linearly mapped onto $[0, 1]$ and we use one-hot encoding for both race and party. We obtain the empiricl matching data from exit poll provided by GALLUP[3]. The empirical matching matrix $\hat{\pi}$ is 3-by-5 matrix, candidates are either Democratic or Republican, or from a third party and voters are categorized into five races (White, Black, Hispanic, Asian and Other). We use mean profile as features for each race and one-hot encoding as features for candidates. We set $\lambda = \lambda_u = \lambda_v = 1, \delta = 5 \times 10^{-3}, K = 20$, for C_u, C_v, we randomly generate m and n points from $\mathcal{N}(\mathbf{0}, 5I_2)$ on plane and use their Euclidean distance matrix as C_u and C_v. Polynomial kernel $C_{ij} = (0.2\boldsymbol{u}_i^T A \boldsymbol{v}_j + 1)^2$ is used for this experiment. We run RIOT for this data set and the learned interaction matrix is shown in Figure 5.

In Figure 5, the lighter the color of a cell is, the lower the cost caused by that feature combination is. Take 'Age' column as an example, ('Democratic', 'Age') is the darkest and ('Republican', 'Age') is the lightest among the column, it means that elder voters are more likely to favor Republican candidate as it has lower cost compared to supporting Democratic candidate. Other cells can be interpreted in a similar manner.

From Figure 5, we see that most white voters tend to support Republican candidate Romney while black voters tend to support Democratic candidate Obama, Democratic and Republican voters tend to support candidate from their own party, elder voters tend to support Romney while female voters tend to support Obama. All above observations are consistent with CNN's[4] exit polls. This demonstrates that RIOT can learn meaningful interaction preference from empirical matching effectively.

6.3. New York Taxi data set

We demonstrate in this subsection that the proposed RIOT framework is able to predict fairly accurate matching on New York Taxi data set [5]. This data set contains 1458644 taxi trip records from January to June in 2016 in New York city. Each trip record is associated with one of the two data vendors (Creative Mobile Technologies, LLC and VeriFone Inc.) and contains detailed trip information such as pickup/drop-off time, longitude, latitude and so on. As no unique identifiers of taxis are provided, we can not predict new matching on individual level. Instead, we predict matching between data vendors and passengers (a passenger is matched with one of the data vendors if a taxi associated with that data vendor rides with the passenger).

[2] part of the experiment in this subsection is based on the code kindly shared by Boris Muzellec(https://github.com/BorisMuzellec/TROT)

[3] http://news.gallup.com/poll/160373/democrats-racially-diverse-republicans-mostly-white.aspx

[4] http://www.cnn.com/election/2012/results/state/FL/president/

[5] https://www.kaggle.com/c/nyc-taxi-trip-duration/data

To reflect the proximity of passengers and taxis, we cluster all trip records into 50 regions and plot them in Figure 6. If a passenger and a taxi are in the same region, it indicates they are close to each other and it is desirable to match them up. Further, since we do not have real-time location of taxis, we use the last known drop-off location as taxis' current location. This assumption is usually not true in large time scale as taxis are likely to leave the region and search for next passenger. To alleviate this issue, we only use trip records within a short time period, 6:00-6:30pm on Friday, June 3rd to predict matching of 6:00-6:30pm on Friday, June 10. Moreover, this is typically the rush hour in New York city and location of taxis are not likely to change dramatically during the period. Vendors' features are the distribution of associated taxis across 50 regions, i.e., $U \in \mathbb{R}^{50\times 2}$, passengers' features are simply the one-hot encoding of their current location, i.e, $V \in \mathbb{R}^{50\times 50}$. So the interaction matrix $A \in \mathbb{R}^{50\times 50}$. We set $\lambda = \lambda_u = \lambda_v = 1, \delta = 1 \times 10^{-3}, K = 20$, for C_u, C_v, we randomly generate m and n points from $\mathcal{N}(\mathbf{0}, 5I_2)$ on plane and use their Euclidean distance matrix as C_u and C_v. Linear kernel $C_{ij} = 0.2\boldsymbol{u}_i^T A\boldsymbol{v}_j + 1$ is used for this experiment.

The comparison of the actual matching π_{new} and the predicted matching $\pi_{predicted}$ is shown in Figure 7. Visually speaking, we see that the predicted matching is able to capture the pattern of actual empirical matching and the prediction is fairly accurate. Quantitative result is also reported, measured in Kullback-Leibler divergence

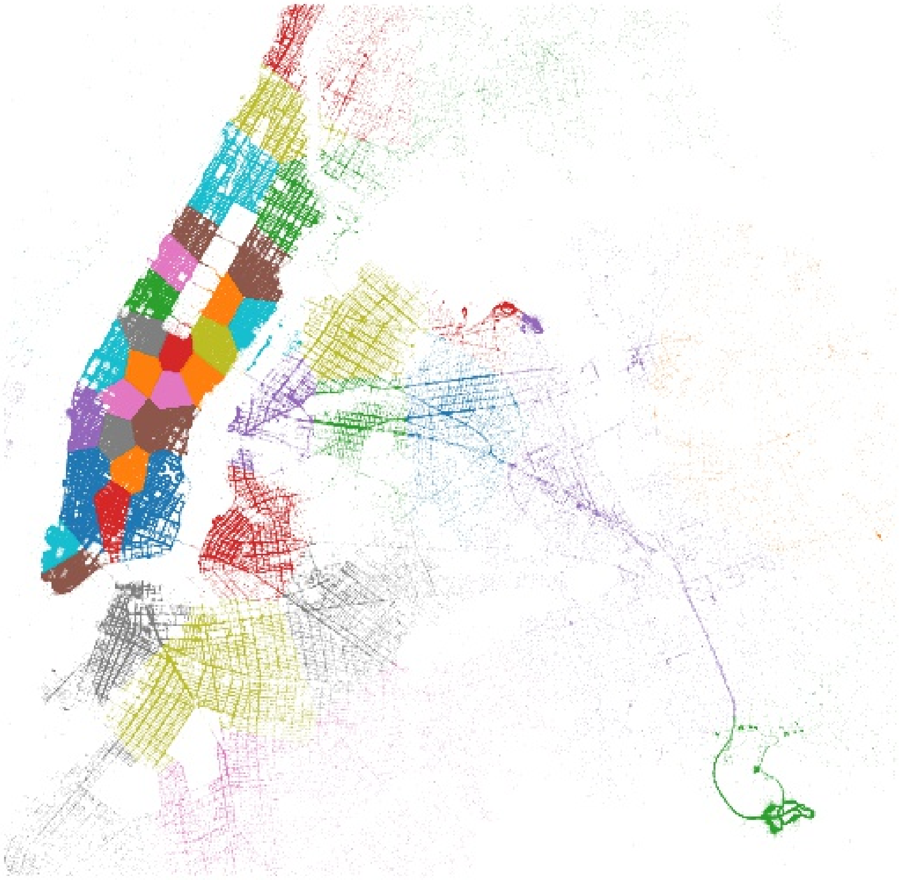
$$\mathrm{KL}(\pi_{new}||\pi_{predicted}) = 0.1659.$$

6.4. Marriage data set

In this subsection, we illustrate the applicability of our model in suggesting new matching and it can make more accurate and realistic recommendations than conventional recommender systems do. Once the interaction between two sides of matching market is learned, one may use that to predict matching for new groups and make recommendations accordingly. We compare our RIOT with baseline random predictor model (Random), classical SVD model (Koren et al., 2009) and item-based collaborative filtering model (itemKNN) (Cremonesi et al., 2010), probabilistic matrix factorization model (PMF) (Mnih and Salakhutdinov, 2008) and the state-of-art factorization machine model (FM) (Rendle, 2012). To fit conventional recommender systems in our setting, one possible approach is simply treating each cell of matching matrix as rating and ignoring the underlying matching mechanism. In RIOT, we set $\lambda = \lambda_u = \lambda_v = 1$, relaxation parameter $\delta = 0.001$, inner iteration $K = 20$ and use polynomial kernel $k(\boldsymbol{x}, \boldsymbol{y}) = (0.2\boldsymbol{x}^T\boldsymbol{y} + 0.8)^2$. For C_u, C_v, we randomly generate m and n points from $\mathcal{N}(\mathbf{0}, 5I_2)$ on plane and use their Euclidean distance matrix as C_u and C_v.

We evaluate all models on Dutch Household Survey (DHS) data set [6] from 2005 to 2014 excluding 2008 (due to data field inconsistency). After data cleaning, the data set consists of 2475 pairs of couple. For each person we extract 11 features including education level, height, weight, health and 6 characteristic traits, namely irresponsible, accurate, ever-ready, disciplined, ordered, clumsy and detail-oriented. Education levels are first categorized into elementary, middle and high and then mapped linearly onto $[0, 1]$. Height and weight are normalized by dividing the largest height/weight. Health and characteristic

[6] https://www.dhsdata.nl/site/users/login



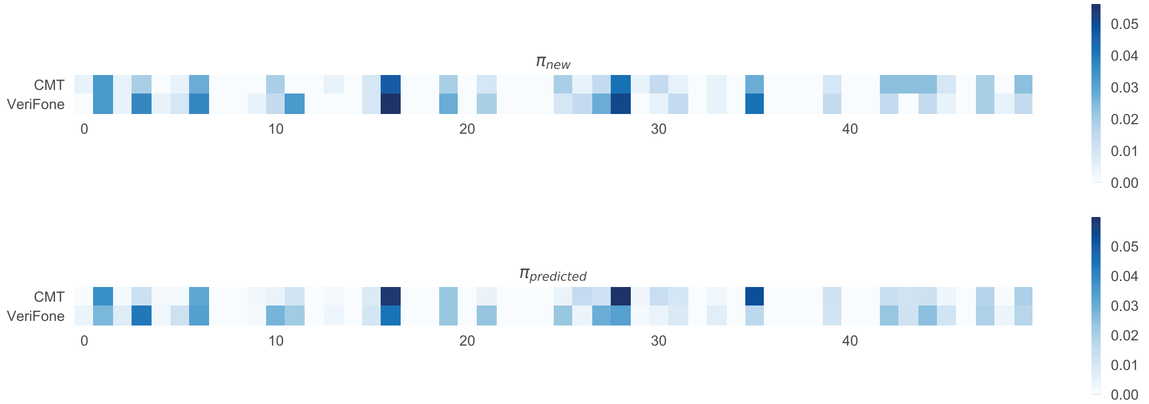


Figure 7: Comparison between actual matching π_{new} (top) and predicted matching $\pi_{predicted}$ (bottom) between 6:00-6:30pm on Friday, June 10th in New York city. Ticks of x-axis are labels of regions.

square error (RMSE) and mean absolute error (MAE) using 5-fold cross-validation. The result is shown in Table 1.

	Random	PMF	SVD	itemKNN	RIOT	FM
RMSE	54.5	8.4	29.9	2.4	**2.3**	3.6
MAE	36.6	2.0	16.8	1.6	**1.5**	2.8

Table 1: Average error of 5-fold cross-validation measured in RMSE and MAE ($\times 10^{-4}$)

In both measures, RIOT beats other conventional RS competitors. The comparison clearly shows that being able to take supply limitation into consideration and capture matching mechanism is of critical importance in suggest matching in such context and our proposed RIOT model can do a better job than conventional recommender systems do.

7. Conclusion

In this paper, we develop a novel, unified, data-driven inverse-optimal-transport-based matching framework RIOT which can learn adaptive, nonlinear interaction preference from noisy/incomplete empirical matching matrix in various matching contexts. The proposed RIOT is shown to be more robust than the state of the art IOT formulation and exhibits better performance in learning cost. Moreover, our framework can be extended to make recommendations based on predicted matching and outperforms conventional recommender systems in matching context.

In the future, our work can be continued in multiple ways. First, our model does batch prediction for a group of users and items and we would like to develop online algorithm to deal with streaming data and make matching suggestion for a previous unseen user/item in an online fashion. A recent method proposed by Perrot et al. (2016) that allows to update the plan using out-of-sample data without recomputing might be useful. From business standpoint, we may study optimal pricing within our framework, i.e., how to set a reasonable price and adjust item distribution in a most profitable way (Azaria et al., 2013). In addition, we hope to combine impressive expressiveness of deep neural networks to further boost the performance of our proposed model.

Acknowledgments

We would like to acknowledge support for this project from the National Science Foundation grants CMMI-1745382, DMS-1620342, DMS-1818886, DMS-1830225 and DMS-1620345, the Office of Naval Research grant N00014-18-1-2852. The authors are grateful to anonymous reviewers and the action editor for their insightful comments and suggestions which helped to improve the manuscript significantly.

Appendix A. Proof of Lemma 3

Proof

$$f(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{x}^T A \boldsymbol{x} - 2\boldsymbol{f}^T \boldsymbol{x} + \|M\|_F^2$$

where $\boldsymbol{x} = \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix}$, $\boldsymbol{f} = [(M\mathbf{1})^T, \mathbf{1}^T M]^T$ and $A = \begin{bmatrix} nI_{m\times m} & \mathbf{1}_m \mathbf{1}_n^T \\ \mathbf{1}_n \mathbf{1}_m^T & mI_{n\times n} \end{bmatrix}$. Note that A is a positive semi-definite matrix, the algebraic multiplicity of its 0 eigenvalue is 1 and $\text{null}(A) = \text{span}\{[\mathbf{1}_m^T, -\mathbf{1}_n^T]^T\}$. Moreover, $\boldsymbol{f} \perp \text{null}(A)$, hence the quadratic form $f(\boldsymbol{a}, \boldsymbol{b})$ admits a minimum in $\text{null}(A)^\perp$ and it is straightforward to obtain

$$\begin{aligned} \min_{\boldsymbol{a},\boldsymbol{b}} f(\boldsymbol{a}, \boldsymbol{b}) &= \min_{\boldsymbol{x}} \boldsymbol{x}^T A \boldsymbol{x} - 2\boldsymbol{f}^T \boldsymbol{x} + \|M\|_F^2 \\ &= \|M\|_F^2 - \boldsymbol{f}^T A^+ \boldsymbol{f} \end{aligned}$$

where A^+ is the Moore-Penrose inverse of matrix A. If M can not be written as $M = \boldsymbol{a}\mathbf{1}^T + \mathbf{1}\boldsymbol{b}^T$, then the minimum of $f(\boldsymbol{a}, \boldsymbol{b})$ must be positive, hence

$$f(\boldsymbol{a}, \boldsymbol{b}) \geq \min_{\boldsymbol{a},\boldsymbol{b}} f(\boldsymbol{a}, \boldsymbol{b}) = \|M\|_F^2 - \boldsymbol{f}^T A^+ \boldsymbol{f} > 0$$

.

■

Appendix B. Extension of RIOT model to learn C_u and C_v jointly

In this section, we extend proposed RIOT model to settings where C_u and C_v are unknown and need to be learned jointly with the main cost matrix $C(A)$. Following the same derivation, we end up with an optimization problem almost identical to the one in equation (5), i.e.,

$$\min_{A,\boldsymbol{\mu}\in\Sigma_m,\boldsymbol{\nu}\in\Sigma_n,C_u,C_v} -\sum_{i=1}^{m}\sum_{j=1}^{n}\hat{\pi}_{ij}\log\pi_{ij} + \delta\big(d_{\lambda_u}(C_u,\boldsymbol{\mu},\hat{\boldsymbol{\mu}}) + d_{\lambda_v}(C_v,\boldsymbol{\nu},\hat{\boldsymbol{\nu}})\big) \tag{13}$$

except for the fact that now we need to optimize two additional variables C_u and C_v. Genericly, inverse problems are usually not well-posed, in our cases, if no constraints imposed on C_u, C_v, one could trivially let, say $C_u = 0_{m\times m}, C_v = 0_{n\times n}$. To avoid such ill-posedness, we assume that $C_u \in \mathcal{M}^m \cap \Sigma^{m\times m}, C_v \in \mathcal{M}^n \cap \Sigma^{n\times n}$, where $\mathcal{M}^d$ is the cone of $d\times d$ distance matrix and Σ^d is $d-1$ simplex (see Section 3 for definition). Other regularization can also be explored.

By strong duality, we may convert equation (13) to its dual problem in a similar fashion as equation (6),

$$\min_{A,\boldsymbol{\mu},\boldsymbol{\nu},C_u,C_v}\max_{\boldsymbol{z},\boldsymbol{w}} -\sum_{i=1}^{m}\sum_{j=1}^{n}\hat{\pi}_{ij}\log\pi_{ij} + \delta\big(\langle\boldsymbol{z},\boldsymbol{\mu}\rangle + \langle\boldsymbol{z}^{C_u},\hat{\boldsymbol{\mu}}\rangle + \langle\boldsymbol{w},\boldsymbol{\nu}\rangle + \langle\boldsymbol{w}^{C_v},\hat{\boldsymbol{\nu}}\rangle\big)$$

where $z_j^{C_u} = \frac{1}{\lambda_u}\log\hat{r}_j - \frac{1}{\lambda_u}\log(\sum_{i=1}^m e^{\lambda_u(z_i - C_{uij})})$ and $w_j^{C_v} = \frac{1}{\lambda_v}\log\hat{c}_j - \frac{1}{\lambda_v}\log(\sum_{i=1}^n e^{\lambda_v(w_i - C_{vij})})$.

One way to solve equation (14), without too many changes of proposed algorithm in Section 5, is to rewrite it as

$$\min_{A,\boldsymbol{\mu},\boldsymbol{\nu}}\min_{C_u,C_v}\max_{\boldsymbol{z},\boldsymbol{w}} -\sum_{i=1}^{m}\sum_{j=1}^{n}\hat{\pi}_{ij}\log\pi_{ij} + \delta\big(\langle\boldsymbol{z},\boldsymbol{\mu}\rangle + \langle\boldsymbol{z}^{C_u},\hat{\boldsymbol{\mu}}\rangle + \langle\boldsymbol{w},\boldsymbol{\nu}\rangle + \langle\boldsymbol{w}^{C_v},\hat{\boldsymbol{\nu}}\rangle\big) \tag{14}$$

and alternatively update three groups of variables $(A,\boldsymbol{\mu},\boldsymbol{\nu})$, (C_u, C_v) and $(\boldsymbol{z},\boldsymbol{w})$.

B.1. Update $(A,\boldsymbol{\mu},\boldsymbol{\nu})$, with (C_u, C_v) and $(\boldsymbol{z},\boldsymbol{w})$ fixed

Once (C_u, C_v) and $(\boldsymbol{z},\boldsymbol{w})$ fixed, $\boldsymbol{z}^{C_u}, \boldsymbol{w}^{C_v}$ are fixed as well, hence the optimization problem at this stage becomes

$$\min_{A,\boldsymbol{\mu},\boldsymbol{\nu}} -\sum_{i=1}^{m}\sum_{j=1}^{n}\hat{\pi}_{ij}\log\pi_{ij} + \delta\big(\langle\boldsymbol{z},\boldsymbol{\mu}\rangle + \langle\boldsymbol{w},\boldsymbol{\nu}\rangle\big)$$

where constants are omitted. This minimization problem is identical to that in subsection 5.1. Please see detailed update scheme for $(A,\boldsymbol{\mu},\boldsymbol{\nu})$ there.

B.2. Update (C_u, C_v), with $(A,\boldsymbol{\mu},\boldsymbol{\nu})$ and $(\boldsymbol{z},\boldsymbol{w})$ fixed

If $(A,\boldsymbol{\mu},\boldsymbol{\nu})$ and $(\boldsymbol{z},\boldsymbol{w})$ are fixed, π is also fixed as it is the regularized OT plan determined by parameters $(A,\boldsymbol{\mu},\boldsymbol{\nu})$, hence the optimization problem at this stage becomes

$$\min_{C_u\in\mathcal{M}^m\cap\Sigma^{m\times m},C_v\in\mathcal{M}^n\cap\Sigma^{n\times n}} \delta\big(\langle\boldsymbol{z}^{C_u},\hat{\boldsymbol{\mu}}\rangle + \langle\boldsymbol{w}^{C_v},\hat{\boldsymbol{\nu}}\rangle\big)$$

and can be further splitted into two independent optimization problems

$$\min_{C_u \in \mathcal{M}^m \cap \Sigma^{m\times m}} \delta\langle \boldsymbol{z}^{C_u}, \hat{\boldsymbol{\mu}} \rangle, \qquad \min_{C_v \in \mathcal{M}^n \cap \Sigma^{n\times n}} \delta\langle \boldsymbol{w}^{C_v}, \hat{\boldsymbol{\nu}} \rangle \tag{15}$$

which can be solved simultaneously.

Both $\mathcal{M}^d$ and $\Sigma^{d\times d}$ are convex sets, so is their intersection. Therefore we can perform projected gradient method to solve two separate minimization problems in equation (15).

B.3. Update $(\boldsymbol{z}, \boldsymbol{w})$, with $(A, \boldsymbol{\mu}, \boldsymbol{\nu})$ and (C_u, C_v) fixed

When $(A, \boldsymbol{\mu}, \boldsymbol{\nu})$ and (C_u, C_v) are fixed, π is also fixed, we then only need to solve

$$\max_{\boldsymbol{z},\boldsymbol{w}} \langle \boldsymbol{z}, \pi \mathbf{1} \rangle + \langle \boldsymbol{z}^{C_u}, \hat{\boldsymbol{\mu}} \rangle + \langle \boldsymbol{w}, \pi^T \mathbf{1} \rangle + \langle \boldsymbol{w}^{C_v}, \hat{\boldsymbol{\nu}} \rangle$$

and one immediately recognizes this is equivalent to applying Sinkhorn-Knopp algorithm to compute $d_{\lambda_u}(C_u, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$ and $d_{\lambda_v}(C_v, \boldsymbol{\nu}, \hat{\boldsymbol{\nu}})$.

To summarize, to jointly learn $C(A)$, C_u and C_v, we formulate an optimization problem similar to that in equation (5) and propose an alternating algorithm to solve it by alternately update $(A, \boldsymbol{\mu}, \boldsymbol{\nu})$, (C_u, C_v) and $(\boldsymbol{z}, \boldsymbol{w})$. Practically, the update of (C_u, C_v) requires expensive projection onto $\mathcal{M}^d \cap \Sigma^{d\times d}$, therefore we suggest learning C_u, C_v first and then using RIOT formulation to learn the main cost matrix $C(A)$, rather than learning three cost matrices simultaneously.

Appendix C. Hyper-parameter Tuning

There are many hyper-parameters in RIOT formulation, including regularization parameter $\lambda, \lambda_u, \lambda_v$, relaxation parameter δ, kernel function $k(\cdot, \cdot)$, cost matrix for two sides C_u, C_v and the number of iteration in inner loop K. With so many hyper-parameters, it would be very expensive to perform grid-search type tuning. Two parameters, relaxation parameter δ and kernel $k(\cdot, \cdot)$, turn out to be critical to the performance of RIOT. Therefore, we fixed other hyper-parameters (based on coarse-tuning) and (fine) tuned these two hyper-parameters.

- $\lambda, \lambda_u, \lambda_v$: as mentioned in main paper, one main motivation of introducing entropy regularization to optimal transport is computation. Too large λ usually causes numerical instability in Sinkhorn-Knopp algorithm, hence we set $\lambda = \lambda_u = \lambda_v = 1$ across all experiments.

- C_u, C_v: Randomly generate m and n points from $\mathcal{N}(\mathbf{0}, 5I_2)$ on plane and use their Euclidean distance matrix as C_u and C_v. This setting is consistent across all experiments.

- K: the number of iterations in inner loop should be large enough to ensure the iterative algorithms described in proposition 7 to converge. In all experiments, we observed the inner loop usually converges within 20 steps, hence we set $K = 20$.

- δ: Both small and large δ can regularize the problem quite well compared to IOT when sample size is small. When sample size is large, empirical matching matrix is usually accurate enough. In this case, large δ can introduce large bias whereas small δ has comparable performance of IOT. In practice, we suggest users to select small relaxation parameter. In our experiments, we found 10^{-3} is an appropriate order to work with.

- $k(\cdot, \cdot)$: we primarily work with polynomial kernels (including linear kernels) in the paper. Kernel function was chosen from $\{k(\boldsymbol{x}, \boldsymbol{y}) = (\gamma \boldsymbol{x}^T \boldsymbol{y} + c_0)^d | \gamma \in \{0.05, 0.1, 0.2\}, c_0 \in \{0.5, 0.8, 1, 2\}, d \in \{1, 2, 3\}\}$

Appendix D. Table of Notations

$\boldsymbol{\mu}$ $\boldsymbol{\nu}$

d $-$

$\boldsymbol{\mu}$ $\boldsymbol{\nu}$

IOT

$RIOT$

λ $\boldsymbol{\mu}$ $\boldsymbol{\nu}$

λ $\boldsymbol{\mu}$ $\boldsymbol{\nu}$

$\boldsymbol{\mu}$ $\boldsymbol{\nu}$

References

Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining

Social Network Analysis and Mining

International Conference on Machine Learning

arXiv preprint arXiv:1702.07134

Operations Research

Proceedings of the 7th ACM conference on Recommender systems

Journal of Political economy

Pattern Recognition and Machine Learning

Convex optimization

Probability Theory and Related Fields

SIAM Journal on Matrix Analysis and Applications

arXiv preprint cs/0703042

Economic theory

Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence

Information Retrieval

Economic Theory

arXiv preprint arXiv:1607.05816

Mathematical Programming

Proceedings of the fourth ACM conference on Recommender systems

Advances in neural information processing systems

Journal of Machine Learning Research

International Conference on Machine Learning

Journal of Economic Perspectives

AAAI

AAAI

Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining

Evaluation Review

The American Mathematical Monthly

Advances in Neural Information Processing Systems

American sociological review

Computer Vision, 2009 IEEE 12th international conference on

Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval

Proceedings of the 26th International Conference on World Wide Web

American Economic Review

A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data

Econometrica: journal of the Econometric Society

Computer

arXiv preprint arXiv:1705.06189

Foundations and Trends○ in Theoretical Computer Science

Econometrica

Advances in neural information processing systems

AAAI

Icml

Asian conference on computer vision

Optimal Transport for Image Processing

Computer vision, 2009 IEEE 12th international conference on

Advances in Neural Information Processing Systems

User Modeling and User-Adapted Interaction

Data Mining (ICDM), 2010 IEEE 10th International Conference on

ACM Transactions on Intelligent Systems and Technology (TIST)

Artificial Intelligence and Statistics

Econometrica: Journal of the Econometric Society

Two-sided matching: A study in game-theoretic modeling and analysis

Proceedings of the 25th international conference on Machine learning

Advances in neural information processing systems

Pacific Journal of Mathematics

Optimal transport: old and new

European Conference on Computer Vision

Weak convergence and empirical processes: with applications to statistics

Advances in neural information processing systems

Internet of Things (iThings/CPSCom), 2011 International Conference on and 4th International Conference on Cyber, Physical and Social Computing