Efficient Sequential UCB-based Hungarian Algorithm for Assignment Problems

Yuyang Shi and Yajun Mei
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
{yyshi, ymei3}@gatech.edu

***Abstract*—The assignment problem has many real-world applications such as allocations of agents and tasks for optimal utility gain. While it has been well-studied in the optimization literature when the underlying utility between every pair of agent and task is known, research is limited when the utilities are unknown and need to be learned from data on the fly. In this work, motivated by the mentor-mentee matching application in U.S. universities, we develop an efficient sequential assignment algorithm, with the objective of nearly maximizing the overall utility simultaneously for each time. Our proposed algorithm is to use stochastic binary bandit feedback to estimate the unknown utilities through the logistic regression, and then to combine the Upper Confidence Bound (UCB) method in the multi-armed bandit problem with the Hungarian algorithm in the assignment problem. We derive the theoretical bounds of our algorithm for both the estimation error and the total regret, and numerical studies are conducted to illustrate the usefulness of our algorithm.**

I. Introduction

The assignment problem is a classical problem in combinatoric optimization, with many classical real-world applications such as allocation of workers or resources for optimal utility gain. Under a general setup, it is assumed that we are given equal number of agents and tasks along with the utility associated with every possible agent-task pair, then we want to find the optimal assignment, i.e. a one-to-one mapping between the agents and tasks that yields maximal total utility. When the underlying utilities are known, the problem is well-studied in the combinatoric optimization literature. For instance, Kuhn (1955) first proposed the well-known Hungarian algorithm, which gives the optimal solution in polynomial time.

However, in many real-world applications, the underlying utility is often unknown and needs to be learned from data on the fly. The motivating example of our research is the mentor-mentee program of the Office of Alumni Relations (OAR) in many U.S. colleges and universities. In such a program, a student mentee is paired with an alumni mentor for regular meetings and career advice. Some universities might let the student mentees choose alumni mentors on their own, often first come first serve, but other universities might have a central pairing system, where the Alumni office decides how to pair between mentees and mentors. While the latter allows one to maximize the overall satisfaction constantly at the system-wise level, the pairing process can be time consuming and label-intensive if it is conducted manually. This inspires us to develop an efficient algorithm that not only speeds up the pairing process but also maximizes the overall satisfaction.

In this paper, we propose an efficient sequential assignment algorithm, with the objective of nearly maximizing the overall utility simultaneously for each round. A key assumption is that a survey is distributed to both mentors and mentees to collect their feedback of satisfaction at the end of the program, and a reasonable number of mentor-mentee pairs return their surveys. For the purpose of demonstration, we assume that the feedback is binary (i.e., satisfied/unsatisfied), which allows us to learn the unknown underlying utility on the fly. Our algorithm consists of two key components at each time: 1) Constructing the upper confidence bound of the utility for every agent-task pair based on previous outcomes through logistic regression; 2) Conducting the assignment that maximizes the overall upper confidence bound of utility by the Hungarian algorithm. This allows us to balance the trade-off between exploration and exploitation in the optimal assignment problem with unknown utilities.

Below it is useful to provide a brief literature review. First, our problem involves parameter estimation of the logistic regression model in a sequential manner, which is related to the field of sequential estimation (Anscombe, 1953a; Ghosh et al., 2011), as well as online learning and optimization (Anderson, 2008; Hazan et al., 2016; Shalev-Shwartz et al., 2012). Second, our research is also related to the so-called reciprocal recommendation systems in applications such as online friend recommendation (Pizzato et al., 2010; Xia et al., 2015), where the system recommends users potential partners based on their profiles, and learn the strategy to find good pairs. As different from our setting, their system gives a number of top recommendations for each user without conducting the assignments among users. Finally, our work is closely related to the area of combinatorial semi-bandit, where each time the player need to pull a collection of arms (called super-arm) subjected to certain constraints, see for example, Cesa-Bianchi and Lugosi (2012); Chen et al. (2013); Gai et al. (2012); Wen et al. (2015). In these studies, the agents and tasks are fixed at each round, and the player wants to learn the optimal assignment through bandit feedback of utility from matched pairs, which is different from our setting, where the

This research was supported in part by NSF grant DMS-2015405 and by the National Centerfor Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR002378. The content is solely the responsibility of the authors and doesnot necessarily represent the official views of the National Institutes of Health.

agents and tasks can change constantly.

The rest of the paper is organized as below. Section II introduces the problem formulation and relative background. Section III develops our proposed UCB-based sequential assignment algorithm, and Section IV presents the theoretical results of our algorithm. Section V presents the results of numerical studies. The concluding remarks are summarized in Section VI. The proofs of our main results are provided in the appendix.

Notations. For $n \in \mathbb{N}$, we denote $[n]$ the set $\{1, 2, \cdots, n\}$. For a d-dimensional vector $\boldsymbol{v} = (v_1, \cdots, v_d)$ and $d \times d$ positive semi-definite matrix M, we define the vector ℓ_2 norm $\|\boldsymbol{v}\|_2 = \sqrt{\sum_{i=1}^d v_i^2}$ and matrix norm $\|v\|_M = \sqrt{v^\top M v}$. We use $P(\cdot)$ to denote the probability of events, and $\mathbb{E}[\cdot]$ to denote the expectation of random variables. We use $I(\cdot)$ to denote the indicator function.

II. Problem Formulation and Background

In this section, we first formulate our sequential assignments with unknown utility, and then briefly present the classical Hungarian algorithm for the the optimal assignment problem with known utility.

Suppose that initially (denoted as time 0) a historic data set of size n_0 is available, denote by

$$\mathcal{D}_0 \coloneqq \{(\boldsymbol{x}_i^0, \boldsymbol{z}_i^0, U_i^0) : \ i \in [n_0]\},$$

where $\boldsymbol{x}_i^0$'s and $\boldsymbol{z}_i^0$'s are task and agent covariates, and U_i^0's are corresponding outcome for every pair of $(\boldsymbol{x}_i^0, \boldsymbol{z}_i^0)$. For each time $t = 1, 2, \cdots, T$, we are given n_t agents and tasks, where each agent or task is associated a vector of covariates, which is also referred to as side information or context. Let $\{\boldsymbol{x}_i^t : i \in [n_t]\} \subset \mathcal{X}$ and $\{\boldsymbol{z}_i^t : i \in [n_t]\} \subset \mathcal{Z}$ be the collection of the covariates of agents and tasks at time t, where $\mathcal{X}$ and $\mathcal{Z}$ are the spaces of the corresponding covariates.

Next, we need to conduct an assignment, denoted by δ_t between these agents and tasks. We assume that the utility associated to each and every matched pair $(\boldsymbol{x}_i^t, \boldsymbol{z}_{\delta_t(i)}^t)$, denoted by $U(\boldsymbol{x}_i^t, \boldsymbol{z}_{\delta_t(i)}^t)$, and we are able to observe the utilities of a subset of random pairs through the end-of-year surveys. Our goal is to maximize the total utility gained up to time T.

At the high-level, we summarize our problem in an online learning framework as below. For round $t = 1, 2, \cdots, T$:

(i) The player is given the agent and task covariates, $\{\boldsymbol{x}_i^t : i \in [n_t]\}$ and $\{\boldsymbol{z}_i^t : i \in [n_t]\}$.

(ii) The player decides an assignment, denoted by a one-to-one mapping $\delta_t : [n_t] \to [n_t]$.

(iii) The player observes the utility feedback for some assigned pair $U(\boldsymbol{x}_i^t, \boldsymbol{z}_{i,\delta_t(i)}^t)$.

Our goal is to decide the assignment δ_t at every round t such that the overall expected utility $\sum_{t=1}^T \sum_{i=1}^{n_t} \mathbb{E}\left[U(\boldsymbol{x}_i^t, \boldsymbol{z}_{\delta_t(i)}^t)\right]$ is maximized.

While the underlying utility of any agent-task pair is unknown by the time of assignment, we assume that the utility is related to the covariates of the agent and task through some probability models. To demonstrate our main ideas, for any pair of covariates $(\boldsymbol{x}_i^t, \boldsymbol{z}_j^t)$, we assume that the binary associated utility $U(\boldsymbol{x}_i^t, \boldsymbol{z}_j^t)$ can be modeled by the logistic regression model:

$$\log \frac{P(U(\boldsymbol{x}_i^t, \boldsymbol{z}_j^t) = 1)}{P(U(\boldsymbol{x}_i^t, \boldsymbol{z}_j^t) = 0)} = \sum_{k=1}^d x_{i,k}^t z_{j,k}^t \theta_k^*, \tag{1}$$

where θ^* is a d-dimensional unknown parameter, $x_{i,k}^t$ is the k-th entry of the vector $\boldsymbol{x}_i^t$, and likewise for $z_{j,k}^t$ and θ_k^*. By defining $\circ$ the entry-wise product, we can rewrite the right-hand side in (1) as $(\boldsymbol{x}_i^t \circ \boldsymbol{z}_j^t)^\top \theta^*$.

For notational simplicity, we define the expectation of the utility by

$$u(\boldsymbol{x}_i^t, \boldsymbol{z}_j^t) = \mathbb{E}[U(\boldsymbol{x}_i^t, \boldsymbol{z}_j^t)] = \frac{\exp\left((\boldsymbol{x}_i^t \circ \boldsymbol{z}_j^t)^\top \theta^*\right)}{1 + \exp\left((\boldsymbol{x}_i^t \circ \boldsymbol{z}_j^t)^\top \theta^*\right)}.$$

and use the shorthands $U_{i,j}^t = U(\boldsymbol{x}_i^t, \boldsymbol{z}_j^t)$ and $u_{i,j}^t = u(\boldsymbol{x}_i^t, \boldsymbol{z}_j^t)$ when there is no ambiguity. Also, we define the pair interaction covariate by $\phi_{i,j}^t = (\boldsymbol{x}_i^t \circ \boldsymbol{z}_j^t)$, which will be used in the rest of the paper.

Note that in our work, we assume that $\boldsymbol{x}$ and $\boldsymbol{z}$ has the same dimension, and the utility is related to these covariates through $\boldsymbol{x} \circ \boldsymbol{z}$. This setting is natural in scenarios (such as mentor-mentee matching) with specific data form, where the preference or attributes of the two parts are aligned well. For other applications, our method need to be extended to accommodate more general form of the utility functions.

To measure the loss of assignments conducted by certain algorithm $\mathcal{A}$, we define the oracle assignment δ_t^* at each round t as the assignment that maximizes the total expected utility:

$$\delta_t^* \in \arg\max_{\delta} \sum_{i=1}^{n_t} u(\boldsymbol{x}_i^t, \boldsymbol{z}_{\delta(i)}^t).$$

Note that since u depends on the unknown parameter θ^*, δ_t^* is also unknown in practice when one conducts the assignment. With the definition of δ_t^*, we further define the cumulative regret of an algorithm $\mathcal{A}$ to be

$$R_T(\mathcal{A}) = \sum_{t=1}^T \left\{ \frac{1}{n_t} \sum_{i=1}^{n_t} \left[u(\boldsymbol{x}_i^t, \boldsymbol{z}_{\delta_t^*(i)}^t) - u(\boldsymbol{x}_i^t, \boldsymbol{z}_{\delta_t(i)}^t) \right] \right\}, \tag{2}$$

where δ_t is the assignment conducted by $\mathcal{A}$, and δ_t^* is the oracle assignment at time t. By the definition of δ^*, $R_T(\mathcal{A})$ captures the performance gap between $\mathcal{A}$ and the oracle performance in expected utility. Ideally, one wish to design an algorithm with the total regret R_T as small as possible.

Note that in (2), we scale the utility gap by the number of pairs n_t, so that we eliminate the difference of the size across time and focus on the average utility gap. Alternatively, one might also consider the regret defined as the summation of the total utility gap directly, without taking the average at each time. The choice of the criteria for regret can depend on the specific problem of interest. In this paper we will discuss the regret bound with regret defined as in (2), while the result can be naturally adapted to the alternative definition.

A. Optimal assignment with known utilities

In this subsection, we briefly review the optimal assignment problem with known utilities and the Hungarian algorithm, which plays a crucial role in our proposed algorithm.

In the classical optimal assignment problem, for n agents and n tasks, there is a $n \times n$ utility matrix U, where $U_{i,j}$ is the utility gain when assigning agent i to task j. The objective is to find the optimal assignment plan (i.e., a one-to-one mapping between agents and tasks), such that the total utility gain is maximized. Mathematically, an assignment plan of δ can be represented either as a permutation or as a $n \times n$ 0-1 matrix, and one wants to solve the following optimization problem:

$$\max_{\delta_{i,j}} \sum_{i\in[n]} \sum_{j\in[n]} U_{i,j}\delta_{i,j} \tag{3}$$

subject to the one-to-one assignment constraints

$$\begin{aligned} &\sum_{i=1}^{n} \delta_{i,j} = 1,\ \forall j = 1, \cdots, n, \\ &\sum_{j=1}^{n} \delta_{i,j} = 1,\ \forall i = 1, \cdots, n, \\ &\delta_{i,j} \in \{0,1\},\ \forall i,j = 1, \cdots, n. \end{aligned} \tag{4}$$

Here $\delta_{i,j}$'s are binary decision variables, and $\delta_{i,j} = 1$ indicates that i is assigned to task j, and $\delta_{i,j} = 0$ otherwise.

When the utilities $U_{i,j}$'s are known, there are many exact or approximation algorithms for optimal assignment, see Avis (1983); Duan and Pettie (2014); Kurtzberg (1962). Here we focus on the classical Hungarian algorithm. It was originally developed by Kuhn (1955), and later improved by Anscombe (1953b) and Edmonds and Karp (1972) to achieve an $O(n^3)$ complexity.

III. Our Proposed Method

In this section, we present our proposed UCB-based algorithm for the sequential assignment problem with unknown utility. At the high level, our algorithm adaptively estimates the true model parameter θ^*, and consists of two components for every round: 1) Constructing the upper confidence bound for every agent-task pair based on past data, and 2) Deciding the assignment by maximizing the total upper bound of the expected utility by Hungarian algorithm.

For better understanding, we split this section into two subsections, each corresponding a component of our proposed algorithm.

A. Constructing upper confidence bound

In this subsection, we specify the construction of the confidence bound based on past data in our method. Loosely speaking, after time step $t_1 \geq 1$, we use the logistic regression to fit the past observations, obtain the estimate for model parameter θ^*, and construct the confidence bound accordingly. For logistic regression, we decide to adopt the ℓ_2-norm-penalized parameter estimation to avoid overfitting in early stage when the sample size is small. More specifically, for every time step $t \geq t_1$, we estimate the parameter θ^* by maximizing the ℓ_2-penalized log-likelihood based on past observations $\{(\boldsymbol{x}_i^\tau, \boldsymbol{z}_{\delta_\tau(i)}^\tau, U_{i,\delta_\tau(i)}^\tau) : \tau \in [t-1], i \in [n_\tau]\}$, or,

$$\begin{aligned} \widehat{\theta}^t \in \operatorname*{argmin}_{\theta} \sum_{\tau=0}^{t-1} \sum_{i\in[n_t]} \Big\{ &\log\Big(1 + \exp(\phi_{i,\delta_\tau(i)}^{\tau\top}\theta)\Big) \\ &- U_{i,\delta_\tau(i)}^\tau \phi_{i,\delta_\tau(i)}^{\tau\top}\theta \Big\} + \frac{r}{2}\|\theta\|_2^2, \end{aligned} \tag{5}$$

where $\phi_{i,\delta_\tau(i)}^\tau = \boldsymbol{x}_i^\tau \circ \boldsymbol{z}_{\delta_\tau(i)}^\tau$, r is a penalty coefficient, and $\delta_0(i) = i$ for consistency. It is equivalent to solving the following equation to obtain $\widehat{\theta}^t$:

$$\sum_{\tau=0}^{t-1} \sum_{i=1}^{n_\tau} \phi_{i,\delta_\tau(i)}^\tau \left(U_i^\tau - \frac{1}{1 + \exp(-\phi_{i,\delta_\tau(i)}^{\tau\top}\theta)} \right) = r\theta. \tag{6}$$

Let $\widehat{\theta}^t$ be the estimate of θ^* at the beginning of time t. Then for any new pair $(\boldsymbol{x}_i^t, \boldsymbol{z}_j^t)$, we construct the following upper confidence bound $b_{i,j}^t$ on the expected utility for the pair:

$$b_{i,j}^t = \frac{1}{1 + \exp\left\{-\phi_{i,j}^{\tau\top}\widehat{\theta}^t - \lambda\sqrt{\phi_{i,j}^{\tau\top}\bar{M}_{t-1}^{-1}\phi_{i,j}^\tau}\right\}}, \tag{7}$$

where λ is a parameter, and

$$\bar{M}_{t-1} = rI + \sum_{\tau=1}^{t-1} \sum_{i=1}^{n_\tau} \phi_{i,\delta_\tau(i)}^\tau \phi_{i,\delta_\tau(i)}^{\tau\top}.$$

Note that the term $\lambda\sqrt{\phi_{i,j}^{\tau\top}\bar{M}_{t-1}^{-1}\phi_{i,j}^\tau}$ scales with the standard error for the utility estimate and thus helps balance the exploration-exploitation trade-off through the parameter λ.

B. Assignment by maximizing upper confidence bound

After constructing the upper confidence bound $b_{i,j}^t$ in (7) for every possible agent-task pair $(\boldsymbol{x}_i^t, \boldsymbol{z}_j^t)$ at time t, we decide to adopt the classical Hungarian algorithm to solve the following optimization problem at each and every time step t:

$$\max_{\delta_{i,j}} \sum_{i\in[n_t]} \sum_{j\in[n_t]} b_{i,j}^t \delta_{i,j} \tag{8}$$

subject to the one-to-one assignment constraints in (4).

Note that here we maximize the objective of the total upper confidence bound of the utility. Our approach is inspired by the UCB-algorithm for classical multi-armed bandit and contextual bandit, as discussed in Lai and Robbins (1985), Chu et al. (2011) and Li et al. (2017) among others, where the player pull the arm with highest upper confidence bound of reward at each round. By considering the upper confidence bound instead of the utility estimate, one can balance the exploitation with exploration.

For better understanding, our proposed algorithm is summarized below in Algorithm 1.

Algorithm 1 UCB-based algorithm for sequential assignment with logistic model

Set $\mathcal{D}_0 = \{(\boldsymbol{x}_i^0, \boldsymbol{z}_i^0, U_i^0) : \ i \in [n_0]\}$. Specify the parameters $t_1 > 0$ and $\lambda > 0$.
for $t = 1$ to T **do**
  Observe covariates $\boldsymbol{x}_1^t, \cdots, \boldsymbol{x}_{n_t}^t$ and $\boldsymbol{z}_1^t, \cdots, \boldsymbol{z}_{n_t}^t$.
  if $t \le t_1$ **then**
    Conduct assignment through a random permutation δ_t.
  else
    Obtain $\widehat{\theta}^t$ by solving (6) based on $\mathcal{D}_{t-1}$.
    Let $\phi_{i,j}^t = (\boldsymbol{x}_i^t \circ \boldsymbol{z}_j^t)$ for every i and j in $[n_t]$. Construct the upper confidence bound $b_{i,j}^t$ for the utility as

$$b_{i,j}^t = \frac{1}{1 + \exp\left(-\phi_{i,j}^{t\top}\widehat{\theta}^t - \lambda \|\phi_{i,j}^t\|_{\bar{M}_{t-1}^{-1}}\right)},$$

    where $\bar{M}_{t-1} = rI + \sum_{\tau=1}^{t-1}\sum_{i=1}^{n_\tau} \phi_{i,j}^\tau \phi_{i,j}^{\tau\top}$.
    Finding the assignment δ_t by solving the optimization problem (8), using the Hungarian algorithm.
  end if
  Observe feedback $\{U_{i,\delta_t(i)}^t : i \in [n_t]\}$.
  Update

$$\mathcal{D}_t \leftarrow \mathcal{D}_{t-1} \cup \left\{(\boldsymbol{x}_i^t, \boldsymbol{z}_{\delta_t(i)}^t, U_{i,\delta_t(i)}^t) : \ i \in [n_t]\right\}.$$

end for

IV. Theoretical Results

In this section, we present the theoretical properties of our proposed algorithm on the estimation error and regret bound. The detailed technical proofs of our main results, Theorems 1 and 2, are postponed in the Appendix.

Let us begin with some necessary assumptions, which are standard in the statistics and bandit literature, see for example (Chu et al., 2011; Li et al., 2017).

Assumption 1. For every $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{z} \in \mathcal{Z}$, $\|\boldsymbol{x} \circ \boldsymbol{z}\|_2 \le 1$. Meanwhile, we assume $\|\theta^*\|_2 \le 1$.

Assumption 2. We assume that for every $t \ge 1$, $\boldsymbol{x}_i^t$'s and $\boldsymbol{z}_j^t$'s are i.i.d. sample drawn from distributions $\mathcal{P}_X$ and $\mathcal{P}_Z$, respectively. Furthermore, $\mathcal{P}_X$ and $\mathcal{P}_Z$ satisfies that

$$\lambda_{\min}\left(\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_X,\ \boldsymbol{z}\sim\mathcal{P}_Z}\left[(\boldsymbol{x}\circ\boldsymbol{z})(\boldsymbol{x}\circ\boldsymbol{z})^\top\right]\right) \ge \sigma^2 > 0.$$

We are now ready to present our main results. Our first main result is to bound the error of the parameter estimation. Theorem 1 below shows that with sufficient past observations, our estimates $\widehat{\theta}^t$ are close to the underlying true parameter θ^* with high probability.

Theorem 1. Suppose Assumptions 1 and 2 hold. Define

$$\kappa = \inf_{\|\phi\|_2 \le 1,\ \|\theta-\theta^*\|_2 \le 1} \frac{\exp\left(\phi^\top\theta\right)}{\left[1 + \exp\left(\phi^\top\theta\right)\right]^2}.$$

Then for fixed t, if M_{t1} satisfies that

$$\lambda_{\min}(M_{t_1}) \ge \frac{32(d + \log(T/\delta))}{\kappa^2} + \frac{2r}{\kappa}, \tag{9}$$

then with probability at least $1 - 2\delta$, we have

$$\|\widehat{\theta}^t - \theta^*\|_{\bar{M}_{t-1}}^2 \le \frac{1}{\kappa^2}\left[d\log\left(1 + \frac{\sum_{t=1}^T n_t}{rd}\right) + 2\log\frac{1}{\delta} + r\right]$$

for all $t \ge t_1 + 1$.

Our second main result is to bound the cumulative regret of our proposed algorithm, and is summarized as follows.

Theorem 2. Suppose that Assumptions 1 and 2 hold. Then for any $\delta \in (0,1)$, there exists universal constants $C_1,\ C_2$ and C_3, such that as long as

$$\sum_{\tau=1}^{t_1} n_\tau \ge C_1\frac{d + \log(1/\delta)}{\sigma^4} + C_2\frac{d + \log(T/\delta)}{\kappa^2\sigma^2} + C_3\frac{r}{\kappa\sigma^2}, \tag{10}$$

with the choice that

$$\lambda = \frac{1}{\kappa}\sqrt{d\log\left(1 + \frac{\sum_{t=1}^T n_t}{rd}\right) + 2\log(1/\delta) + r}, \tag{11}$$

with probability at least $1 - 3\delta$, the total regret of Algorithm 1 satisfies that

$$R_T \le t_1 + \frac{2d}{\kappa}\sqrt{T\log\left(1 + \frac{\sum_{t=1}^T n_t}{d}\right)} \sqrt{\log\left(1 + \frac{\sum_{t=1}^T n_t}{rd}\right) + 2\log(1/\delta) + r}. \tag{12}$$

Remark 1. Here the condition (10) is to guarantee that the minimum eigenvalue of M_{t_1} is sufficiently large under the worst case that there is no observation at time 0 (i.e. $\mathcal{D}_0 = \emptyset$), such that with high probability $\widehat{\theta}^t$ is close to θ^* for every $t > t_1$. Meanwhile, if we have sufficient observations at time 0 such that M_0 is well conditioned, this requirement on t_1 can be relaxed.

Remark 2. The parameter λ is important in our algorithm that trades off between exploration and exploitation. Specially, when $\lambda = 0$, the algorithm reduces to the greedy method, while when λ is very large, the algorithm tends to find the best design for θ^* estimate. For most desired performance, the choice of λ can depend on data in practice.

Remark 3. our high-probability regret bound in (12) is of the rate $\widetilde{O}(d\sqrt{T})$ when neglecting the logarithm factors. This rate is consistent with the result of regret for the logistic bandit, as presented in Li et al. (2017).

V. Numerical Studies

In this section, we report our simulation results to demonstrate the usefulness of our proposed UCB-based algorithm. For simplicity, we fix the total number of time steps $T = 20$ and the number of pairs $n_t \equiv n = 50$ for each time step. Our focus is to investigate the total regret and parameter estimation under several settings, with various data dimension d and choice of tuning parameter λ.

A. Settings

In this subsection, we specify the construction of our simulation examples. Specifically, we consider two cases of dimension d: one is $d = 10$ and the other is $d = 100$. For each dimension d, we consider the following two settings of θ^* when generating the true utility outcome:

(i) **Dense** $\theta^* = \frac{1}{\sqrt{d}}(1, 1, \cdots, 1, -1, -1 \cdots, -1)$, where the first $d/2$ entries are positive and the second $d/2$ entries are negative.

(ii) **Sparse** $\theta^* = \frac{1}{\sqrt{31}}(-1, 1, 2, 3, 4, 0, \cdots, 0)$, such that $\|\theta^*\|_2 = 1$.

At each round, we randomly sample $\boldsymbol{x}_i^t$'s and $\boldsymbol{z}_i^t$'s from the multivariate normal distribution $N(0, I_d)$. We also vary the tuning parameter λ in a grid within the range $[0, 1]$. We set the penalty parameter $r = 0.2$.

For every scenario, we have 20 replications of randomly sampled data, and we present the average performance. Note that we do not have history data at time 0, and we set $t_1 = 1$.

B. Performance

In this subsection, we compare the performance between our method and the ϵ-greedy method with various ϵ values. We first present the figures that characterize the growth rate of the cumulative regret in t, followed by a table with detailed performance for different d and λ. Figure 1 presents the cumulative regret of our algorithm with respect to t under settings (i) and (ii), with different choice of λ. As can be seen, the growth of the total regret is sublinear in t, which shows the statistical efficiency of our proposed algorithm.

Also, note that with the choice $\lambda = 0$, then the algorithm is greedy that does pure exploitation. From the figure, we can see that with a proper choice of λ, one can achieve a lower regret than the pure greedy method with $\lambda = 0$, by balancing the exploration and exploitation. Moreover, we also observe that the improvement of our proposed algorithm is more significantly better when the unknown parameter dimension d is smaller.

Furthermore, it is also worth noticing that while the choice of λ in (11) guarantees the theoretical properties, in practice one might prefer to tune λ for better empirical performance. Also under our simulation setting, while the choice of λ suggested by (11) can be much larger than 1, the empirical performance suggests that a chosen tuning parameter λ less than 1 will lead to good performance.

VI. Conclusions

In this work, we present a simple but useful algorithm for sequential assignment problem with unknown utility and stochastic feedback. Our main idea is to combine the UCB-based algorithm from the multi-armed bandit problem with the Hungarian Algorithm from the optimal assignment problems. We establish the theoretical properties on the parameter estimators and the regret bound that is consistent with research in the stochastic contextual bandit literature. In addition, we also conduct extensive numerical studies to demonstrate the usefulness and advantage of our proposed algorithm. There are a number of interesting issues that have not been addressed here. First, while the interaction of agents and tasks is essential from the optimization viewpoint, but they might be secondary when fitting the logistic regression or other utility model from the statistical or modeling viewpoint. Second, in practice, the underlying utility function might have complicated form, and thus we might need to investigate a more sophistical model such as deep neural networks or non-parametric models. Third, it is also interesting to investigate when the utility function is non-stationary, e.g., changing over time, by adapting the time-varying bandit algorithms in Vakili et al. (2014) and Zeng et al. (2016) to our context. Finally, it might be useful to combine our proposed algorithm with the divide and conquer algorithm to improve the computing efficiency, especially when we face the problem of large-scale assignments in real-world applications. Therefore, this work should be interpreted as a starting point for further investigation on optimal sequential assignment problems.

Appendix: Proofs of Theorems

This Appendix contains three sections. Sections A and B present the proof for Theorems 1 and 2, respectively. Section C presents the proof for Lemma 1, which is used in the proof for Theorem 2.

A. Proof of Theorem 1

Proof. We first define that

$$M_{t-1} = \sum_{\tau=1}^{t-1} \sum_{i \in \mathcal{I}_\tau} \phi_{i,\delta_\tau(i)}^{\tau} \phi_{i,\delta_\tau(i)}^{\tau\top}, \quad \bar{M}_{t-1} \quad = M_{t-1} + rI.$$

Our first step is to show that $\|\widehat{\theta}^t - \theta^*\|_2 \leq 1$ for all $t \geq t_1 + 1$ under the given condition with high probability.

Consider some η-neighborhood of θ^*, $\mathcal{B}_\eta(\theta^*) := \{\theta : \ \|\theta - \theta^*\|_2 \leq \eta\}$. Define

$$\kappa_\eta := \inf_{\|\phi\|_2 \leq 1, \ \theta \in \mathcal{B}_\eta(\theta^*)} \frac{\exp(\phi^\top \theta)}{1 + \exp(\phi^\top \theta)}.$$

Note that at time t, the estimator $\widehat{\theta}^t$ from penalized maximum-likelihood is the solution to the following equation:

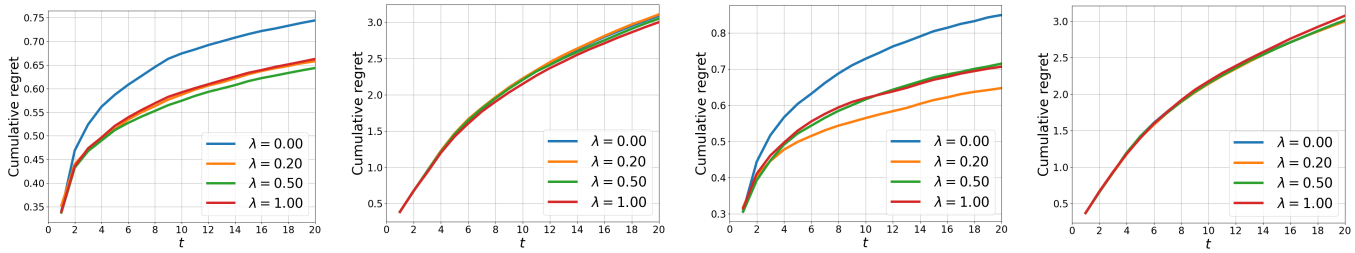
$$\sum_{\tau=1}^{t-1} \sum_{i \in \mathcal{I}_\tau} \left(U_{i,\delta_\tau(i)}^t - \frac{\exp(\phi_{i,\delta_\tau(i)}^{\tau\top} \theta)}{1 + \exp(\phi_{i,\delta_\tau(i)}^{\tau\top} \theta)} \right) \phi_{i,\delta_\tau(i)}^{\tau} - r\theta = 0.$$

Define

$$G(\theta) := \sum_{\tau=1}^{t-1} \sum_{i \in \mathcal{I}_\tau} \left(\frac{\exp(\phi_{i,\delta_\tau(i)}^{\tau\top} \theta)}{1 + \exp(\phi_{i,\delta_\tau(i)}^{\tau\top} \theta)} - \frac{\exp(\phi_{i,\delta_\tau(i)}^{\tau\top} \theta^*)}{1 + \exp(\phi_{i,\delta_\tau(i)}^{\tau\top} \theta^*)} \right) \phi_{i,\delta_\tau(i)}^{\tau} + r\theta.$$

Then we have

$$G(\theta^*) = r\theta^*, \quad G(\widehat{\theta}^t) = \sum_{\tau=1}^{t-1} \sum_{i \in [n_\tau]} \epsilon_{i,\delta_\tau(i)}^{\tau} \phi_{i,\delta_\tau(i)}^{\tau}.$$



(a) Setting (i), $d = 10$ (b) Setting (i), $d = 100$ (c) Setting (ii), $d = 10$ (d) Setting (ii), $d = 100$

Fig. 1. Cumulative regret with different d and λ under settings (i) and (ii). The curves of cumulative regret are sub-linear in t. With proper choice of λ, the performance is improved over the pure-greedy algorithm when $\lambda = 0$.

For any $\theta \in \mathcal{B}_\eta(\theta^*)$, since $0 < \kappa_\eta < 1$, we have

$$\begin{aligned}\|G(\theta) - G(\theta^*)\|^2_{M_{t-1}^{-1}} \geq & (\theta - \theta^*)^\top (\kappa_\eta M_{t-1} + rI)(M_{t-1})^{-1} \\ & (\kappa_\eta M_{t-1} + rI)(\theta - \theta^*) \\ \geq & \kappa_\eta^2 (\theta - \theta^*)^\top (M_{t-1} + \frac{r}{\kappa_\eta} I)(\theta - \theta^*) \\ \geq & \kappa_\eta^2 \lambda_{\min}(M_{t-1}) \|\theta - \theta^*\|_2^2.\end{aligned}$$

By Lemma A in Chen et al. (1999), we have that

$$\left\{\theta : \ \|G(\theta) - G(\theta^*)\|_{M_{t-1}^{-1}} \leq \kappa_\eta \eta \sqrt{\lambda_{\min}(M_{t-1})}\right\} \subseteq \mathcal{B}_\eta(\theta^*). \tag{13}$$

By Lemma 7 in Li et al. (2017), and using a union bound argument, we have that with probability at least $1 - \delta$,

$$\|G(\widehat{\theta}^t)\|_{M_{t-1}^{-1}} \leq 4\sqrt{d + \log(T/\delta)}, \text{ for } t \in [T]. \tag{14}$$

Also, note that

$$\begin{aligned}\|G(\theta)^*\|^2_{M_{t-1}^{-1}} = r^2 \theta^{*\top} M_{t-1}^{-1} \theta^* \leq & \frac{r^2}{\lambda_{\min}(M_{t-1})} \|\theta^*\|_2^2 \\ \leq & \frac{r^2}{\lambda_{\min}(M_{t-1})}.\end{aligned} \tag{15}$$

Combining (13), (14) and (15), we have that when

$$\lambda_{\min}(M_{t_1}) \geq \frac{16(d + \log \frac{T}{\delta})}{\kappa_\eta^2 \eta^2} + \frac{8\sqrt{rd + r \log \frac{T}{d}}}{(\kappa_\eta \eta)^{3/2}} + \frac{r}{\kappa_\eta \eta},$$

with probability at least $1 - \delta$, $\widehat{\theta}^t \in \mathcal{B}_\eta(\theta^*)$ for all $t \geq t_1 + 1$. Taking $\eta = 1$, the above condition is met in (9), so that with probability at least $1 - \delta$, we have $\|\widehat{\theta}^t - \theta^*\|_2 \leq 1$ for all $t \geq t_1 + 1$, which we assume to hold in the followings.

We next uniformly bound the terms $\|\widehat{\theta}^t - \theta^*\|^2_{\bar{M}_{t-1}}$ for all $t \geq t_1 + 1$. Let $N_t = \sum_{\tau=1}^t n_\tau$. By Theorem 1 in Abbasi-Yadkori et al. (2011), with probability at least $1 - \delta$, we have

$$\begin{aligned}\|G(\widehat{\theta}^t)\|^2_{\bar{M}_{t-1}^{-1}} \leq & 2 \log \left(\frac{\det(\bar{M}_{t-1})^{1/2} \det(rI)^{-1/2}}{\delta} \right) \\ \leq & 2 \log \left(\frac{r^{-d/2} (r + \frac{N_{t-1}}{d})^{d/2}}{\delta} \right) \\ \leq & d \log \left(1 + \frac{N_{t-1}}{rd} \right) + 2 \log(1/\delta).\end{aligned}$$

Meanwhile, we have $\|G(\theta^*)\|^2_{\bar{M}_{t-1}^{-1}} \leq r$ for all t. Note that when $\|\theta - \theta^*\|_2 \leq 1$,

$$\begin{aligned}\|G(\theta) - G(\theta^*)\|^2_{M_{t-1}^{-1}} \geq & \kappa^2 (\theta - \theta^*)^\top (M_{t-1} + \frac{r}{\kappa} I)(\theta - \theta^*) \\ \geq & \kappa^2 \|\theta - \theta^*\|^2_{\bar{M}_{t-1}}.\end{aligned}$$

Hence, we have

$$\begin{aligned}\|\widehat{\theta}^t - \theta^*\|^2_{\bar{M}_{t-1}} \leq & \frac{1}{\kappa^2} \|G(\theta) - G(\theta^*)\|^2_{M_{t-1}^{-1}} \\ \leq & \frac{1}{\kappa^2} \left[d \log \left(1 + \frac{N_{t-1}}{rd} \right) + 2 \log(1/\delta) + r \right] \\ \leq & \frac{1}{\kappa^2} \left[d \log \left(1 + \frac{N_T}{rd} \right) + 2 \log(1/\delta) + r \right]\end{aligned}$$

for all $t \geq t_1$. This completes our proof for Theorem 1. □

B. *Proof of Theorem 2*

In this section, we provide the proof for our Theorem 2. We first introduce a useful lemma.

Lemma 1. Let $\{\phi_i^t : t \in [T], i \in [n_t]\}$ be an arbitrary collection of d-dimensional vectors satisfying $\|\phi_i^t\|_2 \leq 1$ for every t and i. Suppose $M_0 \succeq I_d$, $M_t = M_{t-1} + \sum_{i=1}^{n_t} \phi_i^t \phi_i^{t\top}$. Denote $\|\phi\|_M = \sqrt{\phi^\top M \phi}$ the norm induced by a positive definite matrix M. Then we have

$$\sum_{t=t_1+1}^{T} \sum_{i=1}^{n_t} \frac{\|\phi_i^t\|^2_{M_{t-1}^{-1}}}{n_t} \leq 2d \log \left(1 + \frac{\sum_{t=1}^T n_t}{d} \right).$$

Lemma 1 will be used to upper bound the total uncertainty for our algorithm, which is closely related to the regret bound. The proof of The proof for Lemma 1 adapts the argument for UCB method in linear contextual bandits, see Dani et al. (2008) and Abbasi-Yadkori et al. (2011) for example. We defer the detailed proof for Lemma 1 to a later subsection.

Below we provide the proof to Theorem 2. In the high level, we first make use of Theorem 1 to bound the total regret in terms of the summation of $\|\phi_{i,\delta_t(i)}^t\|_{\bar{M}_{t-1}^{-1}}$'s. Then we use Lemma 1 to derive an upper bound for $\sum_t \sum_i \|\phi_{i,\delta_t(i)}^t\|_{\bar{M}_{t-1}^{-1}}$, which leads to an upper bound for the total regret.

Proof. For notational simplicity, we first define

$$V_t = \frac{1}{n_t} \sum_{i=1}^{n_t} u_{i,\delta_t(i)}^t, \quad V_t^* = \frac{1}{n_t} \sum_{i=1}^{n_t} u_{i,\delta_t^*(i)}^t.$$

where δ_t and δ_t^* denotes the assignment decided by the algorithm and the oracle assignment at t, respectively. Then we have

$$R_T \le t_1 + \sum_{t=t_1+1}^{T} (V_t^* - V_t).$$

We first upper bound the performance gap $(V_t^* - V_t)$ for each single round $t \ge t_1 + 1$, and then upper bound their summation as the total regret.

First, under Assumption 2, by Proposition 1 in Li et al. (2017), when t_1 satisfies (10), with probability at least $1-\delta$, we have

$$\lambda_{\min}(M_{t_1}) \ge \frac{32(d+\log(T/\delta))}{\kappa^2} + \frac{2r}{\kappa},$$

so that with probability at least $1-2\delta$, we have

$$\|\widehat{\theta}^t - \theta^*\|_{\bar{M}_{t-1}}^2 \le \frac{1}{\kappa^2}\left[d\log\left(1+\frac{\sum_{t=1}^T n_t}{rd}\right) + 2\log\frac{1}{\delta} + r\right] \tag{16}$$

for all $t \ge t_1 + 1$, according to Theorem 1.

Now we upper bound the difference between $\widehat{u}_{i,j}^t - u_{i,j}^t$ for $t \ge t_1+1$ and $i,j \in [n_t]$. Note that

$$\widehat{u}_{i,j}^t = \frac{1}{1+\exp(-\phi_{i,j}^{t\top}\widehat{\theta}^t)}, \quad u_{i,j}^t = \frac{1}{1+\exp(-\phi_{i,j}^{t\top}\theta^*)}.$$

Since the function $\frac{1}{1+\exp(-\phi_{i,j}^{t\top}\theta)}$ is convex in θ, we have

$$\begin{aligned}
|\widehat{u}_{i,j}^t - u_{i,j}^t| &\le \frac{\exp(-\phi_{i,j}^{t\top}\widehat{\theta}^t)}{[1+\exp(-\phi_{i,j}^{t\top}\widehat{\theta}^t)]^2}\left|\phi_{i,j}^{t\top}(\widehat{\theta}^t - \theta^*)\right| \\
&\le \frac{1}{2}\|\phi_{i,j}^t\|_{\bar{M}_{t-1}^{-1}}\|\widehat{\theta}^t - \theta^*\|_{\bar{M}_{t-1}} \\
&\le \frac{\lambda}{2}\|\phi_{i,j}^t\|_{\bar{M}_{t-1}^{-1}}.
\end{aligned} \tag{17}$$

Here in the last inequality, we use (11) and (16). Then for every $t \ge t_1+1$, we have that

$$\begin{aligned}
V_t^* - V_t &= \frac{1}{n_t}\sum_{i\in[n_t]} u_{i,\delta_t^*(i)}^t - u_{i,\delta_t(i)}^t \\
&\le \frac{1}{n_t}\sum_{i\in[n_t]}\left[\left(\widehat{u}_{i,\delta_t^*(i)}^t + \frac{\lambda}{2}\|\phi_{i,\delta_t^*(i)}^t\|_{\bar{M}_{t-1}^{-1}}\right) - u_{i,\delta_t(i)}^t\right] \\
&\le \frac{1}{n_t}\sum_{i\in[n_t]}\left[\left(\widehat{u}_{i,\delta_t(i)}^t + \frac{\lambda}{2}\|\phi_{i,\delta_t(i)}^t\|_{\bar{M}_{t-1}^{-1}}\right) - u_{i,\delta_t(i)}^t\right] \\
&\le \frac{\lambda}{n_t}\sum_{i\in[n_t]}\|\phi_{i,\delta_t(i)}^t\|_{\bar{M}_{t-1}^{-1}}.
\end{aligned} \tag{18}$$

Here the second inequality is by the construction of our assignment δ_t that maximizes the total upper confidence bound, and the first and third inequalities use (17). From (18) we have that

$$R_T \le t_1 + \lambda\sum_{t=t_1+1}^{T}\sum_{i\in[n_t]}\frac{\|\phi_{i,\delta_t(i)}^t\|_{\bar{M}_{t-1}^{-1}}}{n_t}.$$

To further bound the right-hand side, we use Lemma 1 to obtain that

$$\begin{aligned}
\sum_{t=t_1}^{T}\sum_{i=1}^{n_t}\frac{\|\phi_{i,\delta_t(i)}^t\|_{\bar{M}_{t-1}^{-1}}}{n_t} &\le \sqrt{T\sum_{t=1}^{T}\left(\sum_{i=1}^{n_t}\frac{\|\phi_{i,\delta_t(i)}^t\|_{\bar{M}_{t-1}^{-1}}}{n_t}\right)^2} \\
&\le \sqrt{T\sum_{t=1}^{T}\sum_{i=1}^{n_t}\frac{\|\phi_{i,\delta_t(i)}^t\|_{M_{t-1}^{-1}}^2}{n_t}} \\
&\le \sqrt{2dT\log\left(1+\frac{\sum_{t=1}^T n_t}{d}\right)}.
\end{aligned} \tag{19}$$

Combining (19) with our choice of λ in (11), we obtain that

$$\begin{aligned}
R_T \le& t_1 + \lambda\sum_{t=t_1+1}^{T}\sum_{i=1}^{n_t}\frac{s_{i,\delta_t(i)}^t}{n_t} \\
\le& t_1 + \frac{2d}{\kappa}\sqrt{T\log\left(1+\frac{\sum_{t=1}^T n_t}{d}\right)} \\
&\sqrt{\log\left(1+\frac{\sum_{t=1}^T n_t}{rd}\right) + 2\log(1/\delta) + r},
\end{aligned}$$

which completes our proof for Theorem 2. □

C. Proof of Lemma 1

Proof. We first introduce an inequality that will be used in the later proof of Lemma 1. Let M be a $d\times d$ positive definite matrix with the minimum eigenvalue $\lambda_{\min}(M) \ge 1$. Let ϕ be a d-dimensional vector with $\|\phi\|_2 \le 1$. Note that

$$\begin{aligned}
\det(M+\phi\phi^\top) &= \det(M)\det\left(I_d + M^{-1/2}\phi(M^{-1/2}\phi)^\top\right) \\
&= \det(M)\left(1+\|\phi\|_{M^{-1}}^2\right)
\end{aligned}$$

Since $\|\phi\|_2 \le 1$ and $\lambda_{\min}(M) \ge 1$, we have $\|\phi\|_{M^{-1}}^2 \le 1$. Using the fact that $x \le 2\log(1+x)$ for $x \in [0,1]$, we get

$$\|\phi\|_{M^{-1}}^2 \le 2\log(1+\|\phi\|_{M^{-1}}^2) = 2\log\frac{\det(M+\phi\phi^\top)}{\det(M)}.$$

We now continue to prove Lemma 1. Note that we assume $\|\phi_{i,j}^t\|_2 \le 1$ for all t,i,j. Also, by our construction of $\bar{M}_t$'s, we have $\lambda_{\min}(\bar{M}_t) \ge r$ for all $t \ge t_1+1$. By defining $\bar{M}_{t-1,i} = \bar{M}_{t-1} + \phi_{i,\delta_t(i)}^t\phi_{i,\delta_t(i)}^{t\top}$ and using the previous inequality, we have

$$\begin{aligned}
&\sum_{t=t_1+1}^{T}\sum_{i=1}^{n_t}\frac{\|\phi_{i,\delta_t(i)}^t\|_{\bar{M}_{t-1}^{-1}}^2}{n_t} \\
\le& \sum_{t=t_1+1}^{T}\sum_{i=1}^{n_t} 2\log\left[\frac{\det(\bar{M}_{t-1,i})}{\det(\bar{M}_{t-1})}\right]^{\frac{1}{n_t}} \\
=& 2\log\prod_{t=t_1+1}^{T}\frac{\left[\prod_{i=1}^{n_t}\det(\bar{M}_{t-1,i})\right]^{\frac{1}{n_t}}}{\det(\bar{M}_{t-1})} \\
\le& 2\log\frac{\left[\prod_{i=1}^{n_T}\det(\bar{M}_{T-1,i})\right]^{\frac{1}{n_T}}}{\det(\bar{M}_{t_1})}.
\end{aligned} \tag{20}$$

Since $M_{t_1} \succeq I_d$ and $\|\phi_{i,j}^t\|_2 \leq 1$, we have

$$\det(M_{T-1,i}) \leq \left(1 + \frac{\left(\sum_{t=1}^{T-1} n_t\right) + 1}{d}\right)^d.$$

Therefore, following (20), we have

$$\sum_{t=t_1}^{T} \sum_{i=1}^{n_t} \frac{\|\phi_{i,\delta_t(i)}^t\|_{M_{t-1}^{-1}}^2}{n_t} \leq 2d \log\left(1 + \frac{\sum_{t=1}^{T} n_t}{d}\right).$$

This completes our proof for Lemma 1. □

References

Abbasi-Yadkori, Y., D. Pál, and C. Szepesvári (2011). Improved algorithms for linear stochastic bandits. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, Red Hook, NY, USA, pp. 2312–2320. Curran Associates Inc.

Anderson, T. (2008). *The theory and practice of online learning*. Athabasca University Press.

Anscombe, F. J. (1953a). Sequential estimation. *Journal of the Royal Statistical Society: Series B (Methodological) 15*(1), 1–21.

Anscombe, F. J. (1953b). Sequential estimation. *Journal of the Royal Statistical Society: Series B (Methodological) 15*(1), 1–21.

Avis, D. (1983). A survey of heuristics for the weighted matching problem. *Networks 13*(4), 475–493.

Cesa-Bianchi, N. and G. Lugosi (2012). Combinatorial bandits. *Journal of Computer and System Sciences 78*(5), 1404–1422.

Chen, K., I. Hu, and Z. Ying (1999). Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *The Annals of Statistics 27*(4), 1155–1163.

Chen, W., Y. Wang, and Y. Yuan (2013, 17–19 Jun). Combinatorial multi-armed bandit: General framework and applications. In S. Dasgupta and D. McAllester (Eds.), *Proceedings of the 30th International Conference on Machine Learning*, Volume 28 of *Proceedings of Machine Learning Research*, pp. 151–159. PMLR.

Chu, W., L. Li, L. Reyzin, and R. Schapire (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214. JMLR Workshop and Conference Proceedings.

Dani, V., T. P. Hayes, and S. M. Kakade (2008, 01). Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*, pp. 355–366.

Duan, R. and S. Pettie (2014). Linear-time approximation for maximum weight matching. *Journal of the ACM (JACM) 61*(1), 1–23.

Edmonds, J. and R. M. Karp (1972). Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the ACM (JACM) 19*(2), 248–264.

Gai, Y., B. Krishnamachari, and R. Jain (2012). Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking 20*(5), 1466–1478.

Ghosh, M., N. Mukhopadhyay, and P. K. Sen (2011). *Sequential estimation*. John Wiley & Sons.

Hazan, E. et al. (2016). Introduction to online convex optimization. *Foundations and Trends® in Optimization 2*(3-4), 157–325.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics Quarterly 2*(1-2), 83–97.

Kurtzberg, J. M. (1962). On approximation methods for the assignment problem. *Journal of the ACM (JACM) 9*(4), 419–439.

Lai, T. L. and H. Robbins (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics 6*(1), 4–22.

Li, L., Y. Lu, and D. Zhou (2017). Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pp. 2071–2080. PMLR.

Pizzato, L., T. Rej, T. Chung, I. Koprinska, and J. Kay (2010). Recon: A reciprocal recommender for online dating. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, New York, NY, USA, pp. 207–214. Association for Computing Machinery.

Shalev-Shwartz, S. et al. (2012). Online learning and online convex optimization. *Foundations and Trends® in Machine Learning 4*(2), 107–194.

Vakili, S., Q. Zhao, and Y. Zhou (2014). Time-varying stochastic multi-armed bandit problems. In *2014 48th Asilomar Conference on Signals, Systems and Computers*, pp. 2103–2107. IEEE.

Wen, Z., B. Kveton, and A. Ashkan (2015). Efficient learning in large-scale combinatorial semi-bandits. In *International Conference on Machine Learning*, pp. 1113–1122. PMLR.

Xia, P., B. Liu, Y. Sun, and C. Chen (2015). Reciprocal recommendation system for online dating. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 234–241. IEEE.

Zeng, C., Q. Wang, S. Mokhtari, and T. Li (2016). Online context-aware recommendation with time varying multi-armed bandit. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, New York, NY, USA, pp. 2025–2034. Association for Computing Machinery.