Online learning for two-sided sequential transport matching markets with temporal effects

A. Giudici[a,*], J. van Dalen[a], T. Lu[b] and R. Zuidwijk[a]

[a] Rotterdam School of Management, Rotterdam, The Netherlands
giudici@rsm.nl, jdalen@rsm.nl, rzuidwijk@rsm.nl

[b] School of Business, University of Connecticut, Storrs, Connecticut
tao.lu@uconn.edu

[*] Corresponding author

Extended abstract submitted for presentation at the 11th Triennial Symposium on Transportation Analysis conference (TRISTAN XI)
June 19-25, 2022, Mauritius Island

January 15, 2022

Keywords: two-sided platform, inland transport, spot-market, multi-armed bandits

1 Introduction

Two-sided digital platforms are ubiquitous in many industries. In many notable cases, these online markets have been disrupting industries; examples are ride-hailing platforms or home-stay platforms. In logistics, several platform initiatives have been taken, and despite the expected disruptive potential, these platforms have not been disrupting the industry yet. We consider the case of a two-sided digital marketplace for spot transport, where shippers - the companies in need of transport for available cargo - search for carriers - transport companies - that can transport their goods. In our case, we consider the inland waterway transport market where bulk goods are loaded onto and transported by river vessels (barges). Motivated by our collaboration with a digital transport company, we take the perspective of the company that operates such a two-sided platform where shippers and carriers match to execute transport. Transport requests are chiefly characterized by temporal, spatial, and commodity-related features. On the marketplace, decentralized matching decisions take place between the two sides: carriers screen transport requests and place offers, which are then evaluated by the shippers. When an offer is accepted by a shipper, a match is formed and the platform owner earns a revenue. After matching, agents might return to the marketplace to post or execute new requests, partially depending on the experienced service.

The platform owner manages how the stream of transport requests generated by the shippers is presented to each carrier. In other words, the platform owner decides what assortment of transport requests will each carrier choose from. As multiple carriers might signal their interest to the same shipper, the shipper will choose a single carrier between an *induced* assortment of interested ones. As two assortments are created in a sequence, our problem is similar to the two-sided sequential assortment problem of Ashlagi et al. (2019). The choices of either side are determined by the preferences of the individual users who rationally chose between the presented alternatives.

In this problem setting, information is scarce, partially available, and of transactional nature. The platform possesses only beliefs on users' characteristics, preferences, behavior, etc., which needs to be actively learned. A traditional approach of estimating beliefs once and then deciding on the optimal assortment is not practical in our setting where information is accumulated over time and past choices affect the exogenous information process. The platform owner has the opportunity to actively learn users' characteristics by experimenting with the assortment generated. This general problem has been formalized by the multi-armed bandit models which we adopt for our setting.

2 Methodology

Our model is largely inspired by *multi-armed bandits*, which has been a flourishing concept in recent years (Lattimore and Szepesvári, 2019). Essentially, a multi-armed bandit model captures the setting where a decision-maker aims at maximizing a reward over a time horizon $\mathcal{T} = \{0, \ldots, T\}$ ($T \in \mathbb{N}_{>0}$) while learning to choose between alternative actions $\mathbf{a} \in \mathcal{A}$ of unknown reward distribution. As the action is taken, the environment responds with a, possibly stochastic, reward which provides information on the reward distribution for that action.

In our setting, an action corresponds to choosing what assortment of transport requests each carrier will evaluate. The reward at time t for an assortment $\mathbf{a}^t$, i.e., the expected number of matches $Q^t_{\mathcal{H}^t}(\mathbf{a}^t)$, depends on the observed history $\mathcal{H}^t$ of information accumulated from past evaluations of the carriers and shippers up until time t.

The performance of an algorithm for multi-armed bandits is evaluated by computing the cumulative regret $R^T := \sum_{t=0}^{T} \left(\hat{Q}^t(\mathbf{a}^*) - Q^t_{\mathcal{H}^t}(\mathbf{a}^t)\right)$ between the expected reward $\hat{Q}^t(\mathbf{a}^*)$ obtained when knowing the true reward distribution ($\mathbf{a}^* := \arg\max_{\mathbf{a} \in \mathcal{A}} \hat{Q}^t(\mathbf{a})$) and the history-dependent reward $Q^t_{\mathcal{H}^t}(\mathbf{a}^t)$.

We solve the problem of finding a suitable approach to minimize the bandit regret in a data-driven setting by adapting the upper confidence bound approach to our case (Sutton and Barto, 2018, Agrawal et al., 2019a). The underlying idea is that of moderating action choice by both considering an estimated value of the reward and an upper bound on its performance. Because our industry partner needs to manage a large volume of requests quickly, we compute the next action rapidly by developing a tailored reduction of the objective value and a simulated annealing heuristic to maximize the upper confidence bound. Simulated annealing is a well-established random meta-heuristic framework where the solution space is explored by accepting both improved and worsened solutions with a certain probability (van Laarhoven and Aarts, 1987, Bertsimas and Tsitsiklis, 1993). An overview of the algorithm we implement is given in Algorithm 1 where the upper confidence bound $\mathrm{UCB}^t(\mathbf{a})$ is defined up to a given confidence parameter $c \in \mathbb{R}$, as follows

$$\mathrm{UCB}^t(\mathbf{a}) := Q^t(\mathbf{a}^t) + c\sqrt{\frac{\ln(t)}{N_t(\mathbf{a})}} \tag{1}$$

where $N_t(\mathbf{a})$ counts the number of times an action has been taken and, therefore, provides a measure of the confidence of our estimates. Note that if $N_t(\mathbf{a}) = 0$, then $\mathbf{a}$ is considered a maximizing action. The idea of the upper confidence bound selection is that the square root term is a measure of the uncertainty related to a certain action. Indeed, the more a given action has been taken, the more confident our estimates of the expected number of deals are. If instead, a certain action has never been taken, then our estimate of Q^t is poor and should not be trusted. The confidence parameter c tunes the choice of the next action between the beliefs or the measure of confidence.

3 Results

We test the performance of our bandit algorithm through simulations. For brevity, we report only the main result. In our experiments, we consider the case where a set I of 5 carriers interact with a set J of 5 shippers for $T = 300$ epochs. Carriers and shippers are characterized each

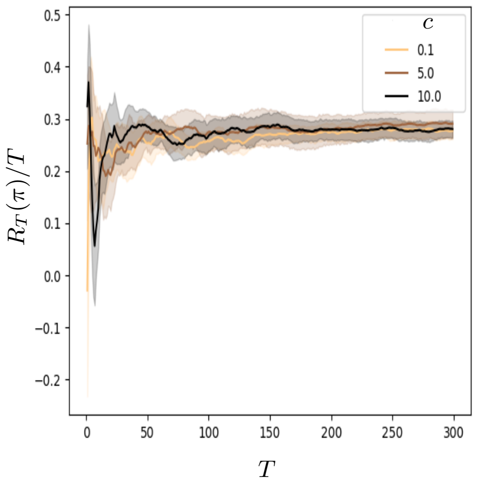

Figure 1 – *Performance of our method*

by a randomly chosen 3-dimensional vector $\mathbf{x}_i \in [0,1]^3$ and $\mathbf{y}_j \in [0,1]^3$ (for $i \in I$ and $j \in J$), respectively. Transport requests are characterized by randomly generated vectors $\mathbf{z}_j^t \in \mathbb{R}^5$. Choices are made following a multinomial choice model as typical in the assortment optimization literature (van Ryzin and Mahajan, 1999). Characteristics of users and transport requests affect their individual choices, i.e., the chance that a carrier reacts to a transport request among the assorted ones as well as the chance of a shipper accepting a carrier among the ones that reacted to the transport requests.

We evaluate the performance of our approach by computing the cumulative regret between the expected optimal decision $\mathbf{a}^*$ and the realized number of deals by the choice of the action. Given our parameters, we execute the same simulation 100 times to compute reliable estimates of the expected performance and present our results for different values of the confidence parameter c (cf. Eq. (1)).

We show in Figure 1 the ratio between the cumulative regret and the time horizon T, together with a 95% confidence band around it. We observe that, with minimal long-term dependency on the confidence parameter c, we obtain a limiting constant regret increase. Our result is in line with that of the model in Sauré and Zeevi (2013) where an asymptotically-constant regret is found in the application of their model on realistic data in Agrawal et al. (2019b). As in our case, Sauré and Zeevi (2013) achieves a constant cumulative regret because of the complex learning task where almost two thousand different items are considered. For us, instead, the complexity of the learning task is due to the exponential number of dual-sided combinations.

Algorithm 1 Multi-armed bandit algorithm overview

```
Initialize all parameters
for $t = 1, \dots, T$ do
    update all beliefs using the history $\mathcal{H}^t$
    update the assortment menu maintaining only open cargo offers
    take an action $\mathbf{a}'$ that maximizes the upper confidence bound $\mathrm{UCB}^t(\mathbf{a})$ using the simulated annealing heuristic
    update the history $\mathcal{H}^t$
end for
```

4 Conclusion

Our work provides a three-fold contribution to the literature. First, we propose a dynamic model of matching and learning for dual-sided digital transport marketplaces. Second, we enrich the assortment optimization model by capturing the timing of user logins and expiring shipments. Third, we achieve performances of dynamic assortment optimization models whilst capturing more complex dynamics.

Learning while matching in dual-sided digital transport marketplaces. To the best of our knowledge, we are the first to consider two-sided assortment optimization in the transport spot market. The main characteristic of this market is the inherent time pressure resulting from transport deadlines. We provide an algorithm that combines learning and exploration in dual-sided assortment optimization for real two-sided online marketplaces.

Stochastic evaluations on expiring assortments. Because the evaluation of an assortment requires carriers to log in while shipments expire over time, we include in our model the effect of uncertain carrier log-in and shipment deadlines. This means that beliefs on the next user login and shipment deadline affect the assortment decision.

Computational experiments. Our algorithm and model have been tested on a simulated online marketplace we constructed for testing purposes. We show that our method achieves asymptotic performances comparable to that of traditional, i.e., not two-sided, dynamic assortment optimization models with learning when exposed to similar learning challenges (Sauré and Zeevi, 2013).

Future research can expand this work in different directions by considering enriched solution methods or modeled features. For instance, we focus on the development of a single strategy that does not consider temporal changes of user behavior (e.g., due to market trends). Moreover, we do not consider the opportunity for optimizing matches from the transport point of view. Both directions could be part of future research.

5 References

References

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. (2019a). MNL-bandit: A dynamic learning approach to assortment selection. *Operations Research*, 67(5):1453–1485.

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. (2019b). MNL-Bandit: A Dynamic Learning Approach to Assortment Selection. *Operations Research*, 67(5):1453–1485.

Ashlagi, I., Krishnaswamy, A. K., Makhijani, R., Saban, D., and Shiragur, K. (2019). Assortment planning for two-sided sequential matching markets. *arXiv*.

Bertsimas, D. and Tsitsiklis, J. (1993). Simulated Annealing. *https://doi.org/10.1214/ss/1177011077*, 8(1):10–15.

Lattimore, T. and Szepesvári, C. (2019). Bandit Algorithms. Technical report.

Sauré, D. and Zeevi, A. (2013). Optimal Dynamic Assortment Planning with Demand Learning. *Manufacturing & Service Operations Management*, 15(3):387–404.

Sutton, R. and Barto, A. (2018). Reinforcement learning: An introduction.

van Laarhoven, P. J. M. and Aarts, E. H. L. (1987). Simulated annealing. In *Simulated Annealing: Theory and Applications*, pages 7–15. Springer Netherlands, Dordrecht.

van Ryzin, G. and Mahajan, S. (1999). On the Relationship Between Inventory Costs and Variety Benefits in Retail Assortments. *Management Science*, 45(11):1496–1509.