Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/37/20230497

Survey of dynamic pricing based on Multi-Armed Bandit algorithms

Jiaming Qu

Department of Mathematics, the Ohio State University, United States

qu.381@osu.edu

Abstract. Dynamic pricing seeks to determine the most optimal selling price for a product or service, taking into account factors like limited supply and uncertain demand. This study aims to provide a comprehensive exploration of dynamic pricing using the multi-armed bandit problem framework in various contexts. The investigation highlights the prevalence of Thompson sampling in dynamic pricing scenarios with a Bayesian backdrop, where the seller possesses prior knowledge of demand functions. On the other hand, in non-Bayesian situations, the Upper Confidence Bound (UCB) algorithm family gains traction due to their favorable regret bounds. As markets often exhibit temporal fluctuations, the domain of non-stationary multi-armed bandits within dynamic pricing emerges as crucial. Future research directions include enhancing traditional multi-armed bandit algorithms to suit online learning settings, especially those involving dynamic reward distributions. Additionally, merging prior insights into demand functions with contextual multi-armed bandit approaches holds promise for advancing dynamic pricing strategies. In conclusion, this study sheds light on dynamic pricing through the lens of multi-armed bandit problems, offering insights and pathways for further exploration.

Keywords: Dynamic Pricing, Multi-armed Bandit, Machine Learning, Reinforcement Learning, Operation Research.

1. Introduction

Dynamic pricing aims to find the optimal selling price of a good or service under the limited supply and the uncertain demands. As a fundamental problem in management science, dynamic pricing has various applications in electricity market, financial services, online real-time auctions and retailing. In this problem, a seller provides a sequence of prices to the potential buyers and observes if each sale attempt succeeds or fails at each price. Every customer is characterized by a uniform demand function, denoted as f(p), representing the likelihood of a successful sale at a given price p. This function remains concealed from the seller and must be acquired progressively through sequential observations. The overarching goal is to optimize the anticipated cumulative reward within a sales constraint of T. In the process of selecting prices at each step, the seller grapples with a dilemma that involves striking a balance between delving deeper into understanding the demand function (exploration) and capitalizing on the price that has demonstrated the most impressive sales history (exploitation). As the seller increases the knowledge towards the unknown demand function, the pricing strategy can be improved over time.

DOI: 10.54254/2755-2721/37/20230497

Such a sequential decision problem can be formulated as the multi-armed bandit (MAB) problem, which involves K actions(arms), each associated with an unknown independent and identically distributed random reward probability, and a player. At each round t the player selects an arm and receives the corresponding reward. The objective is to choose the actions so that after the horizon T rounds, the cumulative reward is maximized. Different solutions can be used for this problem, like Gittins Index, Upper Confidence Bound (UCB), Thompson Sampling and other heuristic algorithms [1-5].

Each potential price p in the sequence is the action(arm), and the mean reward at price p is $p \cdot f(p)$. Since the goal is to maximize the total expected reward $\sum_{t=1}^{T} p_t f(p_t)$, the seller needs to determine the prices(arms) with the possible highest expected reward within limited rounds and make use of them. However, different from traditional discrete arms, the price is usually a continuous variable, therefore dynamic pricing should be described by a continuum-armed bandit problem [6, 7].

The initial association between dynamic pricing and Multi-armed Bandit (MAB) concepts was established by Rothschild [8] in 1974. In this seminal work, Rothschild focused on a scenario involving two prices (resembling a two-armed bandit) and modeled the demand for each price using an unknown mean Bernoulli distribution. Since then, a lot of literature about this connection have been published. Below a more detailed overview of dynamic pricing based on multi-armed bandit problem will be provided, dividing by whether the problem is studied in a Bayesian or non-Bayesian framework, and whether in a stationary or non-stationary environment.

2. Bayesian vs non-Bayesian

Depending on the prior knowledge about the underlying demand function, MAB approaches can be considered for dynamic pricing in both Bayesian and non–Bayesian settings. Within the Bayesian context, policy performance is assessed through the accumulation of rewards, while in non-Bayesian scenarios, it is typically quantified by cumulative regret. In a study by Rothschild [8], a scenario was explored where a seller selects prices from a finite assortment, casting the problem as a Multi-armed Bandit (MAB) dilemma within the Bayesian framework, assuming a priori probabilistic understanding of the demand function. The pivotal finding indicates that, under the optimal Bayesian strategy, there exists a non-zero probability of the price sequence converging towards a suboptimal price. Furthermore, Rothschild and McLennan [9] demonstrated that the occurrence of incomplete learning persists even when the seller's pricing options extend across a continuous spectrum. Incomplete learning means that the seller may only focus on the reward of one of the demand functions, rather than the average over all possible demand functions. Harrison et al. [10] solved this by coming up with a modified version of myopic Bayesian policy which achieves complete learning (finite regret).

Thompson sampling finds extensive application in the realm of dynamic pricing, particularly within the Bayesian framework. The merits of employing Thompson sampling encompass its uncomplicated algorithmic implementation and its compatibility with intricate reward models, a characteristic often encountered in real-world scenarios like electricity pricing. In contrast, methodologies based on the Upper Confidence Bound (UCB) principle face limitations when attempting to expand beyond generalized linear models, which typically lack the comprehensive scope required for optimizing demand response.

In the study conducted by Moradipari et al. [11], they delve into the intricacies of the electricity pricing predicament confronting sellers engaged in real-time pricing strategies aimed at shaping customer demand. To address this, they devised an algorithm hinging on Thompson sampling, with the primary objective of minimizing the seller's regret despite lacking precise insight into customers' price reactions. The Thompson Sampling approach operates under the assumption of an existing prior distribution over unknown parameters, along with a non-zero probability linked to the authentic parameter value. Within each round or day, the algorithm draws a random sample from the prior distribution, subsequently making an optimal selection of the electricity price that minimizes projected costs. Upon observing the load response to the designated price, the algorithm undertakes a Bayesian update of the probability distribution based on this fresh data point. Simulation results unveil that the

DOI: 10.54254/2755-2721/37/20230497

dynamic pricing decisions are considerably enhanced as the certainty surrounding the accurate demand model grows through Bayesian updating. Consequently, the Thompson sampling algorithm's cost aligns more closely with the expense tied to knowledge of the true demand model.

Genalti [12] introduces a new dynamic pricing algorithm which uses Thompson sampling for the exploitation strategy in a Bayesian setting. Through its design, a Bayesian model furnishes a probability distribution of posterior estimates pertaining to the weights in Bayesian linear regression. The technique of Thompson sampling facilitates the generation of random samples derived from the posterior distribution of these weights. This methodology effectively translates into the incorporation of posterior-related characteristics—linked to both time and price—into the curve dictating sales volume patterns. So it can evaluate the volumes for selecting the best arm with respect to only price values. This solution is able to produce a pricing schedule accounting for seasonality and integrating a data-driven volume discounts policy. By performing an online A/B test on an Italian e-commerce for more than 4 months, the algorithm improved the net cash flow margin by 55% with respect to the set B.

However, when the distribution of bandit rewards is unknown to the seller, the MAB problem becomes non-Bayesian. Compared to the Bayesian setting, the non-Bayesian setting is more common in dynamic pricing since the seller usually don't have priori information about the corresponding rewards of different prices. Consider the scenario where a seller ventures into a new market. In such instances, the actual demand distribution might remain elusive, and the process of acquiring this knowledge often comes at a considerable expense. Similarly, when the market undergoes substantial transformations, deriving the new demand function from existing data might not be a straightforward endeavor.

Nevertheless, sellers have the option to make informed assumptions grounded in economic principles and other relevant applications. In the study by Misra et al. [13], a dynamic pricing strategy is presented, addressing the Multi-armed Bandit (MAB) challenge through a scalable distribution-free algorithm. To ensure its applicability across diverse product landscapes, the model deliberately relies on minimal assumptions regarding the underlying demand function for specific items. The researchers expand the conventional Upper Confidence Bound (UCB) algorithm to accommodate the intricacies of learning partially identified demand. Specifically, they introduce the premise of weakly downward-sloping customer demand curves. This modest yet influential assumption allows the manager to infer a consumer's underlying preference across various potential prices when exposed to any particular price point. By amalgamating this partial identification with demand learning, they enhance profit maximization through the extended UCB1 algorithm, creating the Upper Confidence Bound bandit algorithm that incorporates partially identified demand (UCB-PI). Ultimately, the authors provide analytical evidence that their algorithm ensures asymptotic optimality for any demand curve exhibiting a weakly downward-sloping pattern. Similarly, Trovo et al. [14] introduce modifications to the well-known Upper Confidence Bound (UCB) bandit algorithm, harnessing two specific aspects relevant to pricing scenarios: 1) a decrease in the probability of an item being sold as its selling price increases; 2) the common practice of consumers comparing prices from various sellers and monitoring price fluctuations before making purchases, especially in online shopping. Notably, the frequency of item purchases is only a fraction of the instances where potential buyers view its price. The authors integrate these factors into UCB-like algorithms by incorporating insights about inter-arm correlations and prior knowledge concerning the arms' maximum conversion probability.

Their primary approach for leveraging correlation entails utilizing the diminishing pattern of expected conversion probabilities to refine the UCB calculations. Furthermore, to encompass a priori data about the maximum conversion probability, they adopt a variant of the Chernoff bound that offers improved precision in cases of notably low conversion probabilities. In summary, the authors embed fundamental assumptions pertinent to pricing applications, such as the inter-arm correlations and prior knowledge regarding maximum conversion probabilities, into the formulation of a UCB-like arm selection strategy.

The results of the simulation show that the joint use of the two methodologies is very effective. While taking inventory into consideration, Babaioff et al. [15] also concentrate on strategies devoid of any

DOI: 10.54254/2755-2721/37/20230497

distribution information, termed as detail-free or prior-independent methods. A pivotal concept underlying their approach is the formulation of arm indices based on the estimated anticipated cumulative payoff from an arm, considering the known constraints. This stands in contrast to establishing indices according to the expected payoff in a single instance. Leveraging the foundation of UCB1, they devise a novel index that corresponds to an upper confidence bound (UCB) on the anticipated cumulative payoff derived from a specific price. This UCB pertains to the anticipated total payoff associated with that price within the context of a fixed-price strategy, taking into account the number of agents and the inventory size. This index amalgamates factors such as the average payoff from the arm (exploitation), the number of samples collected for that arm (exploration), and the supply constraint.

Just like in UCB, they pick the price which maximizes this index instead of the upper confidence bound value. They prove that this index-based bandit-style pricing algorithm, called cappedUCB, achieve near-optimal performance, and show that the dynamic pricing with limited supply beyond IID valuation can also be dealt as a MAB problem.

In scenarios where the demand model is constrained to a finite set of known demand functions, the dynamic pricing problem can be interpreted as a multi-armed bandit problem featuring interdependent arms. Building upon prior exploration of binary (two-armed) bandit issues, such as the studies by Rothschild [8] and Harrison et al. [10], Tehrani et al. [16] introduce a dynamic pricing policy grounded in the likelihood ratio test. Their work demonstrates that this policy achieves complete learning in a non-Bayesian context, thereby delivering bounded regret. Regret, in this context, refers to the revenue loss in relation to a situation where the demand model is known. This stands in stark contrast to the logarithmically growing regret observed in multi-armed bandit problems with independent arms.

Furthermore, Jain et al. [17] propose an exploration-separated multi-armed bandit mechanism (MAB-MDR) as a means to devise incentive offerings for consumers whose demand response characteristics remain unknown. This approach aims to minimize costs. While access to the Bayesian setting is often challenging, understanding the demand function in dynamic pricing remains highly valuable. The magnitude of this divergence hinges on the assumptions made regarding buyers' valuations. Kleinberg and Leighton [18] explore three such assumptions: 1) valuations are uniformly identical to an undisclosed constant; 2) valuations are independent samples from an undisclosed probability distribution; 3) valuations are determined by an oblivious adversary. In each of these cases, they derive upper and lower bounds on regret that align within a logarithmic factor of "n." In the instance of identical valuations, these bounds match up to a constant factor.

3. Stationary vs non-Stationary

The stationary multi-armed bandit (MAB) problem entails a consistent scenario where both the bandit and the reward distribution remain unchanging over time. Rewards linked to each bandit choice are drawn from distributions that remain constant throughout. The overarching objective is to maximize accumulated rewards over the long term. The solutions outlined above can be categorized as stationary. However, when it comes to dynamic pricing and learning in a fluctuating market, the scenario shifts to the realm of non-stationary multi-armed bandit problems.

Common algorithms employed in the context of stationary MAB comprise the Greedy algorithm, Epsilon-greedy algorithm, and Epsilon-greedy with upper confidence bound algorithm. In accordance with [19], the Greedy method underperforms in the long run due to its convergence to suboptimal policies. The Epsilon-greedy method fares better due to its ongoing exploration. The UCB algorithm outperforms other alternatives by constraining the pursuit of rewards to maximize profit. However, these algorithms resort to averaging methods, leading to convergence towards a singular value, rendering them unsuitable for non-stationary environments where the bandits change over time.

Qureshi et al. [19] tackle the dynamic pricing problem in the context of a non-stationary demand faced by a commercial charging station aiming to maximize long-term profit. They devise a non-stationary algorithm for their dynamic pricing predicament, focusing solely on recent information about uncertain demand rather than historical data. Their approach demonstrates that their strategy tends to

DOI: 10.54254/2755-2721/37/20230497

choose optimal prices for uncertain electric vehicle demand about 90% of the time over a 24-hour time frame.

Another instance of non-stationary MAB in dynamic pricing emerges in the domain of pricing privacy data. Xu et al. [20] investigate the pricing challenge within a scenario where a data collector sequentially procures data from multiple data owners, with privacy prices drawn randomly from an unknown distribution. To optimize overall payoff, the collector must dynamically adapt prices offered to owners. Considering that data volume fluctuates over time and hinges on past interactions between the collector and data owners, they model the pricing problem as a multi-armed bandit issue with time-varying reward distributions. They propose several UCB-based algorithms, essentially relying on the number of successful transactions in the past and the most recent data value estimation to compute the upper confidence bound of rewards. Combining the advantages of cumulative distribution estimation (referred to as VarUCB) and treating side information as context to execute a contextual MAB algorithm (referred to as LinUCB), they introduce a learning strategy named VarLinUCB. Simulation results indicate that these three UCB-based policies proficiently learn data owner type distributions and outperform alternative approaches.

In addition, Trovo et al. [21] take two properties of online pricing into account: the decreasing monotonicity of the demand curve, and online sellers' priori knowledge towards customers. Based on the properties above they introduced improved UCB family algorithms (UCB-LM, UCBV-M and the combinations with sliding-window to adapt to non-stationary setting) and apply these to both stationary and non-stationary settings. The experimental evaluation shows that their algorithms outperform other frequentist MAB algorithms.

4. Conclusion

Due to the general applicability to complex reward models, Thompson sampling is widely used in dynamic pricing with Bayesian setting, where the seller has prior knowledge about the demand function. UCB family algorithms are more popular in non-Bayesian settings for their good regret bounds. In addition, considering the time-varying property of most markets, non-stationary MAB in dynamic pricing deserves more focus. The promising directions for future studies include making improvements on traditional MAB algorithms for online learning settings (i.e., dynamic reward distribution) by exploiting the properties of different scenarios, and combining prior knowledge towards demand functions with contextual MAB methods.

References

[1] Lattimore T and Csaba S 2020 Bandit algorithms Cambridge University Press
[2] Auer P Nicolo C B and Paul F 2002 Finite-time analysis of the multiarmed bandit problem Machine learning 47 235-256
[3] Scott S L 2010 A modern Bayesian look at the multi-armed bandit Applied Stochastic Models in Business and Industry 26.6 639-658.
[4] Gittins J C 1979 Bandit processes and dynamic allocation indices Journal of the Royal Statistical Society Series B: Statistical Methodology 41.2 148-164
[5] Lai T L and Herbert R 1985 Asymptotically efficient adaptive allocation rules Advances in applied mathematics 6.1 4-22
[6] Agrawal R 1995 The continuum-armed bandit problem SIAM journal on control and optimization 33.6 1926-1951
[7] Kleinberg R 2004 Nearly tight bounds for the continuum-armed bandit problem Advances in Neural Information Processing Systems 17
[8] Rothschild M 1974 A two-armed bandit theory of market pricing Journal of Economic Theory 9.2 185-202
[9] McLennan A 1984 Price dispersion and incomplete learning in the long run Journal of Economic dynamics and control 7.3 331-347

DOI: 10.54254/2755-2721/37/20230497

[10] Harrison J Michael N et al 2012 Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution Management Science 58.3 570-586
[11] Moradipari A Cody S and Mahnoosh A 2018 Learning to dynamically price electricity demand based on multi-armed bandits 2018 IEEE global conference on signal and information processing (GlobalSIP) IEEE
[12] Genalti G 2021 A multi-armed bandit approach to dynamic pricing
[13] Misra K Eric M S and Jacob A 2019 Dynamic online pricing with incomplete information using multiarmed bandit experiments Marketing Science 38.2 226-252
[14] Trovò F et al 2015 Multi-armed bandit for pricing Proceedings of the 12th European Workshop on Reinforcement Learning
[15] Babaioff M et al 2015 Dynamic pricing with limited supply 1-26
[16] Tehrani P et al 2012 Dynamic pricing under finite space demand uncertainty: a multi-armed bandit with dependent arms arXiv preprint arXiv:1206.5345
[17] Jain S et al 2014 A multiarmed bandit incentive mechanism for crowdsourcing demand response in smart grids Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 28. No. 1
[18] Kleinberg R and Tom L 2003 The value of knowing a demand curve: Bounds on regret for online posted-price auctions 44th Annual IEEE Symposium on Foundations of Computer Science IEEE
[19] Qureshi U et al 2023 Dynamic Pricing for Electric Vehicle Charging at a Commercial Charging Station in Presence of Uncertainty: A Multi-armed Bandit Reinforcement Learning Approach Proceedings of International Conference on Data Science and Applications: ICDSA 2022 Volume 2. Singapore: Springer Nature Singapore
[20] Xu L et al. 2016 Dynamic privacy pricing: A multi-armed bandit approach with time-variant rewards IEEE Transactions on Information Forensics and Security 12.2 271-285
[21] Trovò F Paladino S Restelli M et al. 2018 Improving multi-armed bandit algorithms in online pricing settings International Journal of Approximate Reasoning 98: 196-235