Reinforcement Learning Techniques for Innovation Diffusion and Optimal Pricing

Sindhu Padakandla
Perspectives Seminar Report
Advisor: Prof. Shalabh Bhatnagar
Dept. of Computer Science and Automation
Indian Institute of Science, Bangalore

***Abstract*—Optimal pricing of products and services is an important problem in marketing. In the presence of diffusion and saturation effects in the market as well as supply-side dynamics, it is imperative for a firm to adopt a dynamic optimal pricing scheme so as to maximize its profit.**

Innovation diffusion is the study of acceptance of new products throughout their life cycle and diffusion models can be used to predict future demand of a product. In this report, we first review innovation diffusion modelling. We then provide a comprehensive survey of the state-of-art in optimal pricing models, some of which employ innovation diffusion modelling. However, in many cases, it is not possible to find this dynamic optimal pricing scheme analytically due to the nature of underlying dynamics. In order to tackle this problem, we propose to formulate the dynamic optimal pricing problem in the Markov Decision Process framework and obtain a reinforcement learning based solution that enhances upon the existing solutions.

***Index Terms*—Diffusion modelling, dynamic pricing, reinforcement learning**

I. Introduction

Innovation diffusion is the *market penetration* of new product and services. A product or a service is always targeted at a particular segment of consumers. For e.g Tata Nano cars are targeted at low-income population groups while Mercedes Benz cars are targeted at high-income groups. Success of a product is indicated by its adoption levels. Moreover a firm selling a product must have some estimate of the future demand for the product in order to plan its short-term production capacities and long-term investments. Innovation diffusion models serve this purpose. They analyze the growth pattern of adoption level and forecast the future adoption levels.

Diffusion models can be temporal or spatial. Temporal models analyze the success of the product over time, while spatial models investigate the acceptance of a product across different geographical regions. Temporal diffusion models were first developed by Fourt [1] and Mansfield [2]. Other prominent temporal diffusion models in the literature are the Bass Model [3] and Rogers model [4]. The typical characteristic feature of these models is the emergence of S-shaped growth pattern. Spatial diffusion modelling has not received much focus in the literature.

Despite the fact that innovation diffusion process is intrinsically stochastic, much of the modeling effort has been restricted to the deterministic framework. In the deterministic case, one speaks of variation of a unique adoption level with time. However, in the stochastic case, one considers a probability distribution of all possible adoption levels. This distribution itself will vary with time.

An application of diffusion models is to develop a pricing scheme for a product. This can be done if some particular adoption level needs to be achieved. However realistically, the demand for a product cannot be predicted accurately and is subject to uncertainties like grabbing of market share by competing products, major breakthroughs in technology which may cause some products to become outdated and so on. Firms also need to account for inventory levels while pricing a product. Further current pricing decisions will influence future outcomes (e.g., sales). Because of these dynamics, optimally pricing a product becomes complicated. We propose to employ the tools of reinforcement learning to model this problem.

The report is organized as follows. Section II provides background on new product diffusion models. Section III briefly introduces the basic tools required to understand the problem setting. In Section IV we survey optimal pricing models and throw light on some future research directions.

II. Innovation Diffusion Models

Innovation diffusion models study how well a product is accepted in the market and what adoption levels a product achieves. The general idea in such prediction models is to formulate the evolution of sales as a ordinary differential equation with some constant parameters. The parameters are estimated from available(training) sales data, which are then used to predict the future sales. The main objective of these models is to study demand fluctuations as a consequence of changes in price and cumulative demand.

A. Bass model

Bass model is a single-purchase deterministic diffusion model. It considers the spread of a product introduced into a market with target population of size m. The target population is assumed to be fully connected. At any point of time potential adopters adopt(or purchase) the product as a result of two types of influences: word-of-mouth (or internal) influence or mass-mediated (or external) influence. The word-of-mouth influence is modeled by a coefficient q and represents the contact a consumer has with a prior adopter. Potential adopters who are influenced by prior adopters are “imitators”. Similarly,

the mass-mediated influence is modeled by coefficient p. It captures an individual's intrinsic need for a product. Potential adopters who adopt the product based on need are termed "innovators".

Considering the probability of an individual adopting a product, given that the individual has not yet adopted, the Bass model proposes that this probability is linear with respect to the number of previous adopters. Let $f(t)$ represent the density function of time to adoption and $F(t)$ represent the cumulative distribution function. If $r(t)$ is the conditional probability density that an individual will adopt a product, then $r(t)$ is a hazard rate function. So according to the Bass model, $r(t)$ is given by the following equation

$$r(t) = \frac{f(t)}{1 - F(t)} = p + qF(t). \tag{1}$$

Suppose $n(t)$ represents the number of adopters at time t and $N(t)$ represents the cumulative number of adopters by time t, then (1) can be used to derive the following equation

$$n(t) = \frac{dN(t)}{dt} = p[m - N(t)] + \frac{q}{m}N(t)[m - N(t)]. \tag{2}$$

The first term in the RHS of (2) represents the adopters who are not influenced by the number of previous adopters while the second term represents the adopters who are influenced by the number of previous buyers.

B. Extensions of Bass model

Bass model extensions enhance upon some of its simplifying assumptions. Two such extensions are described below.

1) Kalish model: The Kalish model [6] specifies the market potential as a time-varying function of product's price. Assuming full product awareness in the population, it specifies the market potential in the following manner-

$$m(t) = m_0 \exp\left[-dP(t)\frac{a+1}{a + \frac{N(t)}{m_0}}\right] \tag{3}$$

where a and d are constants, m_0 is the size of the market potential at the time of product introduction, $m(t)$ is the market potential at time t, $P(t)$ is the product price, $N(t)$ is the cumulative number of adopters at time t. The term $\frac{a+1}{a+\frac{N(t)}{m_0}}$ represents the effect of market penetration in increasing the size of market potential due to the word-of-mouth effect.

2) Generalized Bass model: The Generalized Bass Model [5] introduces a time-varying function $x(t)$ into the Bass model. This function is the "current marketing effort". It reflects the effect of product price, advertising cost and carry-over effects of advertising on the likelihood of adoption. The model describes the hazard rate $r(t)$ in the following manner-

$$r(t) = \frac{f(t)}{1 - F(t)} = [p + qF(t)]x(t). \tag{4}$$

C. Stochastic models

Need for stochastic diffusion models was highlighted by Eliashberg [8]. Piecewise diffusion model was proposed in [13]. It treats the cumulative number of adoptions $A_m(t)$ as a continuous-time pure-birth process, where m is the target population size. The number of states is $m + 1$ and for $0 \leq j \leq m - 1$, the transition rate from state j to state $j + 1$ is given by

$$\lambda_{mj} = (m - j)(\alpha + \frac{\beta}{m - 1}j) \tag{5}$$

where α and β are two non-negative parameters. Here α represents the intrinsic adoption rate while β represents the induction rate, similar to the Bass model parameters p and q respectively.

Diffusion model based on a nonlinear random differential equation was introduced by Karmeshu and Goswami [23]. The adoption level is considered to be random. The analytical investigation of the model in this paper establishes the existence of transient bimodality in the evolution of the adoption level.

D. Uses of innovation diffusion models

Innovation diffusion models provide a mechanism to generate an innovation life cycle which, in the marketing context, is referred to as product life cycle (PLC). Various empirical studies reveal uni-modal life-cycle patterns in the diffusion of new products. Such a uni-modal PLC pattern can easily be derived from the Bass model. The Generalized Bass model can be used to derive the price at a particular time t, if the adoption level at t is known. This way a pricing strategy can be devised.

III. Basic Framework

This section introduces the basic tools required which enables us to model optimal pricing of products as a learning problem.

A. Stochastic Differential Equations

A stochastic differential equation (SDE) is a differential equation in which one or more of the terms is a stochastic process, resulting in a solution which is itself a stochastic process. The general form of an SDE is

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t, \quad X_0 = 0$$

where X_t is a stochastic process, $b(t, X_t)$ is the *drift* term, $\sigma(t, X_t)$ is the *diffusion* term and W_t is the *Wiener Process*. W_t is normally distributed with mean 0 and variance $\sigma^2 t$. The derivative dW_t is referred to as white-noise. SDEs are used to model phenomena such as fluctuating stock prices and many physical phenomena.

B. Markov Decision Process

A stochastic process $\{X_n\}$ that takes values in a set S is called an MDP if its evolution is governed by a control- valued sequence Z_n so that the following controlled Markov property is satisfied:

$$\begin{aligned} Pr(X_{n+1} = j | X_n = i, Z_n = a, X_{n-1} = i_{n-1}, \\ Z_{n-1} = a_{n-1}, \ldots, X_0 = i_0, Z_0 = a_0) = p(i, j, a) \end{aligned} \tag{6}$$

for any $i_0, \ldots, i_{n-1}, i, j, a_0, \ldots, a_{n-1}, a$ in appropriate sets. $X_n = i$ represents the state and for any n, the set of

feasible actions or controls is $A(i)$. Thus, in (6), $a \in A(i)$, $a_{n-1} \in A(i_{n-1})$ and so on. Let $A = \cup_{i \in S} A(i)$ denote the control space (i.e., the set of all controls). Moreover (6) implies that when the state is i and action a is chosen, the next state is j with a probability of $p(i, j, a)$. These probabilities satisfy $p(i, j, a) \in [0, 1]$, $\forall i,\ j \in S,\ a \in A(i)$ and that $\sum_{j \in S} p(i, j, a) = 1$, for any given $i \in S$ and $a \in A(i)$. An action a chosen yields a scalar value $k(i, a)$ which can be interpreted either as a reward or a cost.

A controller has to pick an action at every decision epoch in a way to maximize(minimize) a long-term reward(cost). This is modeled as sequence of functions $\pi = \mu_1, \mu_2, \ldots$, with each $\mu_n : S \to A$, $n \geq 1$. The sequence π is referred to as a policy and is called admissible if $\mu_n(i) \in A(i)$, $\forall i \in S$. This condition corresponds to the choice of control $Z_n = \mu_n(X_n)$, $\forall n$. An admissible policy π with each $\mu_n = \mu$, $n \geq 1$ is said to be a stationary deterministic policy (SDP).

C. Discounted Reward Criterion

In the discounted reward criterion, the long-run reward to be maximized is the expected sum of discounted rewards. Let $\gamma \in (0, 1)$ denote the discount factor and Π the set of all admissible policies. For a given admissible policy $\pi \in \Pi$, the value function $V^\pi : S \to R$ is defined by

$$V^\pi(i) = \mathsf{E}\left[\sum_{m=0}^{\infty} \gamma^m k(X_m, \mu_m(X_m)) \big| X_0 = i\right] \qquad (7)$$

for all $i \in S$. Then, the aim is to find an optimal policy π^* that gives the optimal value function $V^* : S \to R$, which is defined by

$$V^*(i) = \max_{\pi \in \Pi} V^\pi(i). \qquad (8)$$

It is well known that an SDP achieves the optimal policy, i.e., the policy that corresponds to the optimal value $V^*(i), \forall i \in S$. Furthermore, the optimal value function $V^*(\cdot)$ satisfies the Bellman equation of optimality as

$$V^*(i) = \max_{a \in A(i)} \left(k(i, a) + \gamma \sum_{j \in S} p(i, j, a) V^*(j)\right) \qquad (9)$$

for all $i \in S$.

D. Dynamic Programming

The classical approach for solving MDPs is using Dynamic Programming [15]. It converts the problem of finding an optimal policy to that of finding an optimal value function. The Bellman equation can be solved using value iteration or policy iteration. Value iteration solves the Bellman equation by improving the estimate of the optimal value function at each step. Policy iteration, on the other hand, breaks the problem at hand in two steps. The first step, policy evaluation(critic), determines the value of a given policy and the second step, policy improvement(actor), determines a better policy over the existing policy. These methods rely on the complete model of the system in the form of transition probabilities and the reward structure.

E. Reinforcement Learning (RL)

Reinforcement learning methods learn to control a stochastic environment(system) so as to maximize(minimize) a long-term reward(cost). They do so interacting with the environment. Typically the stochastic environment has states and the agent (or controller) has to learn to decide which action to take to achieve an objective. When an action is taken, the system evolves to a new state and a reward is obtained. Such stochastic sequential decision making problems can be posed in the framework of MDPs. Then the value function of MDP represents the objective and its optimization is the key to control the system. However the transition dynamics and reward structure are unknown. The idea in RL methods is that, to attain (8), we run a stochastic iterative algorithm using observations obtained from online samples. It is then shown using the theory of stochastic approximation, that the algorithm asymptotically converges to an optimal value function and policy tuple.

1) RL Algorithms: Methods used to solve the an RL problem are Monte-Carlo methods and Temporal difference(TD) methods. Monte Carlo methods are based on averaging sample returns obtained from simulated experience. Similar to Monte-Carlo, TD methods work with raw experience. They update estimates based in part on other learned estimates and ultimately learn the optimal value function and policy tuple. Examples of TD methods are Q-Learning [16], Sarsa and Actor-Critic algorithms [14].

2) Multi-agent RL: MDP is a single agent decision problem. An extension of MDP to multiagent systems is stochastic games, which essentially are n-agent MDPs. Stochastic games generalize Game theory and MDPs. It provides a framework for dynamic game theory with multiple agents, every agent being self-interested. Multiagent RL techniques integrate developments in the areas of single-agent RL, game theory and direct policy search techniques. This area has received lot of attention in the recent past.

F. Innovation Diffusion Modelling and RL

A reinforcement learning problem has a specific objective which the agent has to learn by interacting with the environment. In our view the problem of innovation diffusion cannot be looked at as a RL problem. The agent in this case is the firm selling the product. When predicting the future demand, the agent(firm) has no ultimate goal to achieve. Hence this problem does not fall in the paradigm of RL problems.

IV. Optimal Pricing Models

Firms grapple with the complex task of determining the right prices to charge a customer for a product or a service. This task requires that a company know not only its own operating costs and availability of supply but also what the future demand will be. A company therefore needs a wealth of information about its customers and also be able to adjust its prices at minimal cost. Advances in Internet technologies and e-commerce have dramatically increased the quantum of information the sellers can gather about customers, making it

easy to change the prices. Hence there is scope for dynamic adjustment of prices.

Optimal pricing is the dynamic adjustment of prices in order to maximize the long-term profit and also maintain or increase demand for the product over long-term. It renders itself as a sequential decision making problem and hence RL can be utilized.

A. Survey

Chen and Jain [11] examine a situation where a monopolist wants to choose a price for a product at every time t that maximizes the discounted profits. The demand $s(t)$ for the product is assumed to be a function of price $p(t)$, cumulative sales $x(t)$ and a state variable $y(t)$ whose movement is contingent on the occurrence of Poisson events. The behaviour of $y(t)$ is represented by a Poisson-driven differential equation as

$$dy(t) = k[x(t), y(t)]dq(t)$$

where $q(t)$ is a Poisson process with rate λ and it represents the random events which affect sales. They find that the movement of the expected optimal price depends on the relative strength of the contingent effects and the effects of the cumulative sales.

A stochastic model by Raman and Chatterjee [7] models the cumulative sales $x(t)$ as a SDE in the following manner-

$$dx(t) = f(x(t), p(t))dt + \sigma(x(t))dw$$

where $p(t)$ is the price function and $f(x(t), p(t))$ is referred to as a demand specification. They analyze the temporal trajectory of the optimal price using three different demand specifications. The first demand specification is linear in price, the second is a multiplicatively-separable function derived from Bass model and the third is a price-timing function. Some results for the effect of cumulative sales and change in unit production cost on the initial optimal price and the path of the optimal price are provided by the authors.

Gallego and Ryzin [12] deal with dynamic pricing of inventories with stochastic demand over finite horizon. This work models the demand as a Poisson process with intensity $\lambda(p)$ where $\lambda(p)$ is increasing in price p. By charging price p_t at time t, the firm controls the intensity of the demand. Its shown under suitable assumptions that: (a) more stock and/or longer remaining time to sell goods leads to higher expected revenues; (b) at a given point in time, the optimal price decreases as the inventory increases.

Carvalho and Puterman [26] consider the problem of a retailer who has to set the price of a good to optimize the total expected revenue over a period of time T . When the demand function is not known, the retailer has to rely on uncertain prior information to guide his pricing decisions. In this paper, a parametric model is considered in which the parameters are unknown. For example, after t days of sale, a seller knows the prices he has set on the preceding $t-1$ days and can observe the demands on the preceding $t-1$ days. The model is a simple log-linear regression model, where the logarithm of the demand is a linear function of the price. The seller can learn about the parameters of the demand function and use it to set prices so as to maximize the revenues over a given time horizon. It is shown that a one-step look-ahead rule performs fairly robustly for a single seller environment studied.

A detailed survey and classification of dynamic pricing models is provided by Y Narahari et al [17]. The machine learning models surveyed therein, are based on RL and are developed for the e-business environment. Raju et al [24] look at electronic retail markets with a single seller. The seller has an inventory of products which he replenishes according to a standard inventory policy. The seller is the learning agent in the system and uses reinforcement learning to learn from the environment. The problem is to determine dynamic prices that optimize the sellers performance metric. Based on simplifying assumptions about the arrival process of customers, valuations of the customers, inventory replenishment policy, and replenishment lead time distribution, the system becomes a MDP thus enabling the use of RL algorithms. Q-learning algorithm is used to solve the problem.

Multi-agent pricing models exist. These generally model competing firms. Ravikumar, Saluja, and Batra [25] study a service market environment with two competing sellers who service a stream of buyers. The buyers are considered to be of two different categories- informed and uninformed. They assume that both the sellers follow an RL-based adaptive behaviour and model the system as a general sum Markovian game. An actor-critic type of RL scheme is proposed for which convergence results are shown.

Hu [21] models three different types of pricing agents in a simulated market. The first agent uses RL to determine the prices, by learning an optimal action for one period based on the rewards it receives for that action. The second agent uses a traditional Q-learning method, by learning about Q-values which represent long-term optimal values for the agents own actions. The third agent uses a sophisticated Nash Q-learning algorithm, by learning about Q-values which represent long-term Nash equilibrium values for agents joint actions. The third agent performs better than the other two.

B. Directions for future research

The RL-based models considered in Section IV-A have to deal with large state spaces. Function approximation and feature selection can be used to mitigate this issue. Simulation-based methods like λ-Policy Iteration [18] can also be applied in the context of optimal pricing.

V. Conclusion

Developments in e-business have resulted in a significant change the way goods are marketed and sold. Firms have better information about buyers and customers too have better access to deals and discounts. This has led to a paradigm shift from fixed pricing to dynamic pricing. Machine-learning models based on dynamic pricing is an active area of research. Prominent models are based on RL and employ traditional RL solution methods like q-learning. However such models

have to deal with large state spaces. This provides interesting challenges and opportunities.

REFERENCES

[1] L. Fourt and J. Woodlock, "Early prediction of market success for new grocery products," *The Journal of Marketing*, pp. 31–38, 1960.

[2] E. Mansfield, "Early prediction of market success for new grocery products," *Econometrica*, vol. 29, no. 4, October 1961.

[3] F. Bass, "A new product growth model for consumer durables," *Management Science*, pp. 215–227, 1969.

[4] E. Rogers, *Diffusion of innovations*. Simon and Schuster, 1995.

[5] F. Bass, T. Krishnan, and D. Jain, "Why the bass model fits without decision variables," *Marketing science*, pp. 203–223, 1994.

[6] S. Kalish, "A new product adoption model with price, advertising, and uncertainty," *Management science*, pp. 1569–1585, 1985.

[7] K. Raman and R. Chatterjee, "Optimal monopolist pricing under demand uncertainty in dynamic markets," *Management Science*, pp. 144–162, 1995.

[8] R. Chatterjee and J. Eliashberg, "Stochastic issues in innovation diffusion models," *Innovation Diffusion models of New-Product Acceptance*, pp. 151–199, 1986.

[9] B. Robinson and C. Lakhani, "Dynamic price models for new-product planning," *Management Science*, pp. 1113–1122, 1975.

[10] S. Kalish, "Monopolist pricing with dynamic demand and production cost," *Marketing Science*, vol. 2, no. 2, pp. 135–159, 1983.

[11] Y. Chen and D. Jain, "Dynamic monopoly pricing under a poisson-type uncertain demand," *Journal of Business*, pp. 593–614, 1992.

[12] G. Gallego and G. Van Ryzin, "Optimal dynamic pricing of inventories with stochastic demand over finite horizons," *Management science*, vol. 40, no. 8, pp. 999–1020, 1994.

[13] S. Niu, "A piecewise-diffusion model of new-product demands," *Operations research*, pp. 678–695, 2006.

[14] R. Sutton and A. Barto, *Reinforcement learning: An introduction*. Cambridge Univ Press, 1998, vol. 1, no. 1.

[15] M. Puterman, *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, Inc., 1994.

[16] C. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3, pp. 279–292, 1992.

[17] Y. Narahari, C. Raju, K. Ravikumar, and S. Shah, "Dynamic pricing models for electronic business," *Sadhana*, vol. 30, no. 2, pp. 231–256, 2005.

[18] D. Bertsekas, "Lambda policy iteration: A review and a new implementation," *Lab. for Information and Decision Systems Report LIDS-P-2874, MIT*, 2011.

[19] G. Tesauro and J. Kephart, "Pricing in agent economies using multi-agent q-learning," *Autonomous Agents and Multi-Agent Systems*, vol. 5, no. 3, pp. 289–304, 2002.

[20] P. Kloeden and E. Platen, *Numerical solution of stochastic differential equations*. Springer, 2011, vol. 23.

[21] J. Hu and Y. Zhang, "Online reinforcement learning in multiagent systems," in *Proceedings of 8th Conference of American Association for Artificial Intelligence*, 2002.

[22] J. Hu, M. Wellman *et al.*, "Multiagent reinforcement learning: Theoretical framework and an algorithm," in *Proceedings of the Fifteenth International Conference on Machine Learning*, vol. 242. Citeseer, 1998, p. 250.

[23] D. Goswami *et al.*, "Transient bimodality and catastrophic jumps in innovation diffusion," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 38, no. 3, pp. 644–654, 2008.

[24] C. Raju, Y. Narahari, and K. Ravikumar, "Learning dynamic prices in electronic retail markets with customer segmentation," *Annals of Operations Research*, vol. 143, no. 1, pp. 59–75, 2006.

[25] K. Ravikumar, G. Batra, and R. Saluja, "Multi-agent learning for dynamic pricing games of service markets," 2002.

[26] A. Carvalho and M. Puterman, "Dynamic pricing and reinforcement learning," in *Neural Networks, 2003. Proceedings of the International Joint Conference on*, vol. 4. IEEE, 2003, pp. 2916–2921.