IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 28, 2021, accepted June 3, 2021, date of publication June 8, 2021, date of current version June 28, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3087577

Multi-Step Look-Ahead Optimization Methods for Dynamic Pricing With Demand Learning

DINA ELREEDY, AMIR F. ATIYA, (Senior Member, IEEE), AND SAMIR I. SHAHEEN, (Life Member, IEEE)
Computer Engineering Department, Cairo University, Giza 12613, Egypt
Corresponding author: Samir I. Shaheen (sshaheen@ieee.org)

ABSTRACT Dynamic pricing is a beneficial strategy for firms seeking to achieve high revenues. It has been widely applied to various domains such as the airline industry, the hotel industry, and e-services. Dynamic pricing is basically the problem of setting time-varying prices for a certain product or service for the purpose of optimizing revenue. However, a major challenge encountered when applying dynamic pricing is the lack of knowledge of the demand-price curve. The demand-price curve defines the customer's response towards changing the price. In this work, we address the dynamic pricing problem in case of unknown demand-price relation. This work introduces a less myopic pricing approach based on looking ahead for one or several future steps in our quest to optimize revenue. Specifically, the proposed formulation maximizes the summation of the immediate revenue and the expected future revenues of one or multiple look-ahead steps. A key benefit of the proposed approach is that it automatically strikes a balance between the conflicting goals of revenue maximization and demand learning, by producing less myopic and profitable prices. We provide a formulation for the presented look-ahead pricing approach, and we implement two variants of it: one-step and two-step look-ahead methods. Experiments are conducted on synthetic and real datasets to compare the proposed pricing methods to other pricing strategies in literature. The experimental results indicate that the proposed look-ahead methods outperform their counterparts in terms of the achieved revenue gain.

INDEX TERMS Demand learning, dynamic pricing, optimization, revenue management, sequential decision making.

I. INTRODUCTION

In the last few decades dynamic pricing has been considered an active area of research having substantial contributions in the fields of operations research, management sciences, economics, and computer science [1]. Dynamic pricing is defined as setting a time-varying price for a certain product or service [2], [3]. Recently, dynamic pricing is broadly applicable in various domains such as: hotel revenue management [4], airline industry [5], [6], mobile data services [7], electricity [8], [9] and [10], and e-services [11]. However, there is a fundamental challenge when applying dynamic pricing, which is how to set prices optimally in order to maximize revenue returns. Determining the optimal price is an arduous problem since in most cases, the customer's behavior in response to a price change is not known beforehand. In other words, firms usually do not know the demand-price relation, and this is particularly the case for new products/services [12].

Dynamic pricing in case of unknown demand behavior has attained a growing interest in the literature [13]–[16]. In such a problem, the ultimate objective of firms is to maximize the total gained revenues over a finite sales time horizon T. Therefore, given an initial estimate of demand behavior, the simplest strategy is to choose the price maximizing the revenue given the estimated demand model parameters. This simple pricing policy is known as greedy or myopic pricing. However, as its name indicates, the "myopic" pricing yields "myopic" decisions, and accordingly sub-optimal revenue returns.

The myopic pricing policy obtains sub-optimal revenues since it does not devote any attention to learning the demand model parameters. In particular, myopic pricing is a pure exploitation strategy where it fully emphasizes revenue maximization. Accordingly, myopic pricing obtains sub-optimal revenues since it relies on initial inaccurate

The associate editor coordinating the review of this manuscript and approving it for publication was Utku Kose.

estimates of the demand behavior. The trade-off between learning demand curve parameters and maximizing revenue is known as the learning versus earning trade-off [17], [18]. It is akin to the exploration-exploitation trade-off encountered in machine learning and evolutionary optimization algorithms [19]–[22].

In this work, we address the dynamic pricing problem with demand learning [16] in a less myopic scheme. Most of the pricing policies in case of unknown demand are essentially variants of the myopic pricing such as: myopic pricing with dithering [23] and controlled variance pricing [24]. However, in this work, we propose a different less myopic approach. Specifically, instead of maximizing the immediate revenue, our proposed formulation seeks to maximize the summation of the immediate revenue in addition to one or more look-ahead steps of the expected future revenues. The multi-step look-ahead formulation not only yields less myopic pricing decisions, but also boosts the overall revenue gains. Furthermore, the proposed formulation automatically balances the trade-off between learning demand parameters and earning revenues. The proposed approach achieves this balance since it produces more diverse prices than the myopic approach. In contrast, the myopic approach obtains prices that are centered on the price maximizing the immediate revenue, therefore such prices do not enhance demand learning. In addition, our proposed formulation is simple and closed-form. Furthermore, the presented formulation is efficient and computationally tractable.

In the proposed formulation, we employ a simple parametric model, assuming a linear demand curve. We apply a parametric model for several reasons: first, at early time steps, limited information is available, which hinders the performance of non-parametric models. Moreover, generally, parametric models are less computationally expensive than non-parametric ones. Furthermore, linear demand models are the most dominant models adopted in the operations research and economics literature [15], [23], [25]–[27]. Finally, as argued by Keskin and Zeevi in [15], the linear demand function could approximate any demand function, especially that firms usually do not consider a very broad range of prices. They rather experiment with prices within a certain range, where such predefined prices are set according to business considerations and marketing conditions. Even if the true demand curve is nonlinear, in a narrow range a linear model would be a good approximation. Thus, operating in a narrow range of prices supports the validity of using a linear demand model.

In this work, we apply the recursive linear regression model developed in [28] for estimating the demand curve parameters after acquiring each new price and its corresponding demand. We adopt the recursive linear regression model because it fits the recursive nature of the problem in case of considering multiple future time steps, where each time step updates and improves over the previous time step. Moreover, the recursive linear regression model is very computationally efficient due to its incremental updates of model parameters. Section IV presents the formulation of the recursive linear regression model proposed in [28].

We conduct a set of experiments to our proposed pricing formulation with one and two look-ahead steps. In addition, we compare the proposed pricing methods to myopic pricing and to some other pricing strategies in the literature. The experiments demonstrate that the proposed look-ahead formulation outperforms not only the myopic pricing policy, but also other competing methods in terms of the achieved revenue.

An analogous problem that experiences a similar debate between single-step and multi-step is the problem of adaptive control. There have been approaches for one-step-ahead adaptive control, and also other approaches for multi-step adaptive control, the so-called model predictive control. The multi-step look-ahead optimization has also proved its efficacy in different applications such as active search and surveying [29], [30]. However, to the best of our knowledge, the multi-step look-ahead optimization has not been applied to the dynamic pricing problem. Generally, most of the dynamic pricing methods essentially maximize the immediate revenues such as: myopic pricing with dithering [23] and controlled variance pricing [24]. Accordingly, such pricing methods yield myopic prices since they do not consider future revenues. On the other hand, the proposed multi-step look-ahead pricing provides less myopic pricing by incorporating the immediate revenue and look-ahead future revenues. Thus, the proposed pricing approach is expected to achieve higher revenue gains than other pricing methods [23], [24].

The main contributions of this work are summarized as follows:

- In this work, we develop a novel multi-step look-ahead formulation for the dynamic pricing problem in case of unknown demand curve.
- The presented formulation is simple, easy to implement, and computationally tractable.
- We apply our proposed pricing formulation to real and synthetic datasets. The proposed pricing approach achieves good performance in terms of the gained revenue compared to the other pricing methods in the literature including: myopic pricing, myopic pricing with dithering [23], and controlled variance pricing (CVP) [24].

The paper is organized as follows: Section II presents a literature review. Section III defines the problem formulation. Section IV briefly describes the recursive formulation of the linear regression model that is applied in our experiments. Then, the proposed look-ahead model formulation is introduced in Section V. After that, Section VI demonstrates the experimental results. Then, the results are further analyzed in Section VII. Section VIII concludes the paper. Finally, potential future research directions are presented in Section IX.

II. RELATED WORK

A. DYNAMIC PRICING WITH DEMAND LEARNING

In this section, we review the main contributions related to the considered problem: dynamic pricing in case of unknown demand. Several comprehensive reviews are provided in [16], [31]–[33].

Carvalho and Puterman propose a simple single-step look-ahead method for revenue maximization in case of log-normal demand curve [12]. Their proposed pricing method maximizes the one-step look-ahead revenue using Taylor series expansion to approximate the next step revenue. Their proposed method outperforms myopic pricing. Further, Carvalho and Puterman [34] extend their work of [12] and apply it to online pricing over the internet. The authors' model uses a binomial demand distribution since they model individual customer's response to a price change as a binary random variable.

A major difference between Carvalho and Puterman's work and the proposed work is that our one-step look-ahead formulation is exact. However, Carvalho and Puterman use Taylor series approximation in their objective function formulation. Another significant difference is that Carvalho and Puterman present and experiment with a single-step look-ahead revenue maximization policy. On the other hand, our formulation is general and can be applied to multiple future steps ahead as illustrated in Section V. For example, in the conducted experiments, we apply one and two steps look-ahead policies as presented in Section VI. Finally, unlike the work of [12], [34], we adopt the linear demand model which is the most ubiquitous model in the operations research literature [16].

Besbes and Zeevi investigate how model misspecification could affect revenue loss [27]. They consider a multi-period single product pricing problem, and prove that some pricing strategies based on a two-parameter linear demand models could converge to near-optimal pricing decisions even in case of model misspecification.

Harrison *et al.* provide a mathematical analysis to study the impacts of myopic pricing on demand learning [35]. In their study, the authors consider a binary demand variable for each price, and they develop a Bayesian formulation with a binary prior distribution. Their results emphasize the negative consequences of myopic pricing on learning the demand model, which is named as ''incomplete learning''.

To alleviate incomplete learning, many variants of myopic pricing are developed. For example, the authors of [23] propose a basic simple pricing policy for linear demand learning of a single product based on the simple myopic pricing policy. The authors adjust the myopic pricing and introduce some exploration to it by adding a random perturbation to the myopic price.

Another extension of the simple myopic pricing is introduced in [24]. The proposed pricing policy, named Controlled Variance Pricing (CVP), chooses the optimal price given the current estimate of the model (like myopic greedy pricing). However, the CVP policy imposes a constraint that the chosen price should not be very close to the average of the previously selected prices. This constraint typically ensures the diversity of the chosen prices. Consequently, the CVP policy incorporates some exploration to enhance the accuracy of demand learning.

Furthermore, Keskin and Zeevi address both single and multiple-product pricing along a finite sales horizon [15]. They propose some variants of the greedy iterative least squares strategy which utilize sequential model learning, and myopic price optimization given the learned model.

Some dynamic pricing approaches use customized prices for different customers by adjusting the price according to each customer's buying behavior. This pricing scheme is commonly known in literature as personalized pricing [36], [37]. As an example, the authors of [38] assume a different potential buying probability per customer. They develop two different pricing policies. The first policy seeks to maximize the expected revenue improvement. The second pricing method selects the price maximizing the summation of the expected immediate revenue and the expected revenue of the next time step. However, both of these methods fail to outperform the myopic greedy pricing policy.

B. MULTI-STEP LOOK-AHEAD UTILITY OPTIMIZATION

Multi-step look-ahead optimization has been studied in different (non-pricing) optimization contexts since it provides less myopic solutions. For example, Garnett *et al.* propose a customized multi-step look-ahead formulation for two classification problems: active search and active surveying [29]. In addition, their work is extended in [30] and applied to the active search problem. In order to emphasize exploration, the authors of [30] approximate the final cumulative expected utility by assuming independence of the future points till the end of the time horizon. However, their proposed approach is computationally intensive since they evaluate the objective function representing look-ahead formulation over all the feasible points in the dataset. Then, the point maximizing the objective function is queried. On the other hand, our proposed approach employs an optimization algorithm in order to pick the point maximizing the objective function. In fact, using an optimization algorithm is much faster than iterating over the feasible data points.

Multi-step look-ahead formulations have been applied to information allocation and ranking applications. A simple Bayesian single-step look-ahead method named knowledge gradient (KG) is proposed in [39]. The knowledge gradient method is applied to sampling allocation and ranking. The knowledge gradient method myopically maximizes the expected improvement of a certain utility function at each time step. This method maintains a Bayesian predictive distribution for each alternative's utility, and the posterior distributions are updated according to new observations. However, the KG method is proved to be optimal in extreme cases: for a single step horizon, and when the time horizon tends to infinity [40]. Furthermore, the KG method could be computationally intensive in case of large number of alternatives.

C. REINFORCEMENT LEARNING

Reinforcement learning is analogous to multi-step look-ahead optimization. Specifically, the proposed look-ahead formulation could be regarded as an approximation of the Bellman equation of the Markov decision process (MDP) using a certain number of look-ahead steps. However, there are some principal differences: first, the utility function in the considered dynamic pricing problem incurs inherent uncertainty. Moreover, the dynamic pricing problem has a continuous space of actions and states. Furthermore, most of the reinforcement learning approaches in literature tackle problems where the time horizon T is infinite, or sufficiently large [41], [42]. However, the dynamic pricing problem is generally studied in a finite sales horizon setting [12].

Reinforcement learning is extensively applied in the dynamic pricing framework [6], [26], [43]–[47]. For example, the authors of [26] employ Q-learning for dynamic pricing and demand learning. In their work, the authors use Q-learning for learning the value function aiming to maximize revenue. However, one of the main pitfalls of the reinforcement learning approach is that it is computationally expensive. Consequently, within the constraint of having limited price experimentation, reinforcement learning could be incompatible with the considered dynamic pricing problem.

D. STUDIES HANDLING THE EXPLOITATION-EXPLORATION TRADE-OFF

The exploration-exploitation trade-off is studied in various fields including: dynamic pricing ([14], [24], [27], [35]), evolutionary optimization [21], and sequential optimization [48]. In addition, the exploration-exploitation trade-off is investigated in the multi-armed bandit problem literature [49]–[52].

Multi-armed bandit (MAB) is a class of sequential decision making problems originally introduced in [53], [54]. The multi-armed bandit problems seek to maximize rewards, but under uncertainty and incomplete feedback about rewards. Consequently, multi-armed bandit problems incur a trade-off between performing an action that retrieves information regarding reward (exploration), and making a decision that maximizes the immediate reward given the information gained so far (exploitation) [55]. Many problems can be formulated using the multi-armed bandit setting such as: our target problem: dynamic pricing with unknown demand [33], online advertising [56], and clinical trials [57].

Trovo *et al.* utilize the multi-armed bandit formulation for developing pricing policies that maximize revenue in case of an unknown demand model [58]. The authors propose two pricing policies that are typically refined versions of the Upper Confidence Bound (UCB) algorithm proposed in [59] to adapt to the pricing problem. Nevertheless, the proposed methods do not achieve better regret bounds than the UCB algorithm.

Another piece of work that exploits multi-armed bandit algorithms for dynamic pricing of e-commerce applications is presented by Ganti *et al.* in [1]. The authors apply two different multi-armed bandit algorithms: Upper Confidence Bound (UCB) [59] and Thomson Sampling (TS) [53]. However, the revenue improvement of the proposed methods over myopic pricing essentially depends on the number of products they actually price.

Recently, active learning has been an effective paradigm whenever the cost of data collection is substantial [60]. In [61], the authors develop an active learning framework that aims to balance the exploration-exploitation trade-off in optimization problems. They apply their framework to the dynamic pricing with demand learning problem. Their proposed methods surpass myopic pricing and the upper confidence bound (UCB) algorithm [59].

In order to highlight the features of the different pricing policies for the case of uncertain demand parameters in a concise way, Table 1 compares these policies with respect to their approaches, contributions, and limitations. In order to stay focused, we limit the table to the core approaches.

III. PROBLEM FORMULATION

In this work, we formulate the dynamic pricing problem in case of unknown demand curve as an iterative optimization problem. At each time step, we choose the price maximizing a certain utility function which is essentially the summation of the immediate revenue and the expected future revenue(s) for one or multiple time steps ahead. Then, the corresponding demand for the chosen price is observed. In order to estimate the demand model parameters, we adopt the recursive formulation of the weighted least squares algorithm proposed in [28].

In this work, we employ a linear price demand model (or price elasticity model), like typically used in the economics and operations research literature. The price is the key controlling variable for demand. We assume a monopolist seller who has a sufficient inventory to satisfy all potential demand, which is known in literature as infinite inventory. The presented work addresses pricing a single product over a finite sales horizon T.

The adopted linear price demand model is defined as follows:

$$y = a + bp + \epsilon \tag{1}$$

where p is the price, y is the corresponding demand, and a and b denote the demand model parameters. In addition, the price sensitivity parameter b is negative ($b < 0$), and ϵ is a normally distributed random error term such that $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

Let $x = [1 \quad p]^T$, then the linear regression problem can be expressed as follows:

$$y = \beta^T x + \epsilon \tag{2}$$

where $\beta = [a \quad b]^T$.

Assume that at any time step n, the cumulative utility function given n data points $\mathcal{D}_n$ is denoted as $u(\mathcal{D}_n)$, such that $\mathcal{D}_n = \{(x_i, y_i) \ \ \forall i \ \ 1 \leq i \leq n\}$. The utility function $u(\mathcal{D}_n)$ can be defined as the total cumulative revenue or the cumulative discounted revenue gained from time step 1 to n.

TABLE 1. Comparison between the main dynamic pricing strategies in the literature.

Paper	Approach	Contributions	Limitations
Carvahlo and Puterman [34]	Optimization	Proposing a one step look-ahead revenue optimization method using Taylor series expansion.	• Not exact. • Limited to one look-ahead step only.
Lobo and Boyd [23]	Variant of myopic pricing	Adding random exploration to the myopic pricing.	• The amount of dithering needs to be tuned. • Not effective in case of unknown demand due to its greedy nature.
den Boer and Zwart [24]	Variant of myopic pricing	Enhancing diversity of selected prices.	• Choosing non profitable prices for exploration. • Does not consider future revenues.
Morales-Enciso and Branke[38]	Personalized pricing	Developing two pricing methods based on efficient global optimization and dynamic programming.	• Does not outperform the myopic pricing.
Cheng [26]	Reinforcement learning	Using Q-learning for learning the value function aiming to maximize revenue.	• Conflicting with limited price experimentation constraint. • Computationally expensive.
Trovo et al. [58]	Multi-armed Bandits	Proposing two refined versions of the Upper Confidence Bound (UCB) algorithm.	• Does not improve the UCB regret bounds.
Ganti et al. [1]	Multi-armed Bandits	Developing two methods based on the Upper Confidence Bound (UCB) algorithm and Thomson Sampling (TS).	• Revenue improvement depends on the number of products they actually price.

We formulate the dynamic pricing problem as a sequential optimization problem where at each time step n, we choose the price p_n that maximizes the expected utility. The expected immediate utility with no look-ahead steps, $\mathbb{E}[u(\mathcal{D}_n)|x_n, \mathcal{D}_{n-1}]$ is computed as follows:

$$\mathbb{E}[u(\mathcal{D}_n)|x_n, \mathcal{D}_{n-1}] = \int_{y_n} u(\mathcal{D}_n) Pr[y_n|x_n, \mathcal{D}_{n-1}]\, \mathrm{d}y_n \quad (3)$$

where $x_n = [1 \;\; p_n]^T$ and y_n is the corresponding demand for p_n. The look-ahead formulation is presented in Section V.

IV. PRELIMINARIES: RECURSIVE FORMULATION OF WEIGHTED LINEAR REGRESSION

In this section, we briefly describe the weighted linear regression model developed in [28] and used in the proposed formulation. We apply such a recursive regression model for several reasons. First, it suits the recursive nature of our proposed formulation in case of multiple look-ahead steps. Moreover, the dynamic pricing with demand learning problem exhibits a sequential behavior. At each time step, a new price is tested, and the demand model is updated accordingly. Consequently, using the recursive linear regression model with incremental updates conforms to the sequential operation of the dynamic pricing. Moreover, the recursive regression model is computationally efficient due to its incremental updates. Furthermore, the weighted linear regression converges to a global minimum since it is a least-squares formulation.

For the notation used in the recursive linear regression, let x_n be the d-dimensional vector chosen at time n. In the dynamic pricing problem x_n is defined as: $x_n = [1 \;\; p_n]^T$ where p_n is the price chosen at time step n. Let y_n be the corresponding predicted output, which is essentially the demand in the dynamic pricing problem. In addition, let $\hat{\beta}$ be the d-dimensional vector of the estimated coefficients of the regression model. Since we adopt a linear demand model where $d = 2$ as defined in Eq. (1), the estimated model's parameters vector can be defined as $\hat{\beta} = [\hat{a} \;\; \hat{b}]^T$.

The discounted error function for T time steps is defined as follows:

$$E(T) = \sum_{n=1}^{T} \gamma^{T-n} \left({x_n}^T \hat{\beta} - y_n\right)^2 \quad (4)$$

where γ is a discount factor such that $0 < \gamma \leq 1$, and generally, γ is set close to 1.

A. ESTIMATING THE REGRESSION MODEL PARAMETERS

In this section, we present the formulas for evaluating the regression model's coefficient vector and the covariance of the regression model's parameters. The estimated model's parameter vector $\hat{\beta}$ is given by the least squares formula as follows:

$$\hat{\beta} = (X^T W X)^{-1} X^T W y \quad (5)$$

where the rows of matrix X represent the input vectors ${x_n}^T$, and y is defined as the vector of target output variables y_n. The matrix W denotes the discount matrix, which is an $n \times n$ diagonal matrix with the diagonal entries $W_{nn} = \gamma^{T-n}$.

The covariance matrix of β is computed as follows:

$$\Sigma_\beta = \sigma^2 (X^T W X)^{-1} \quad (6)$$

Evaluating Eq. (5) and Eq. (6) in a continuous manner is computationally extensive, therefore we use the recursive formulation instead.

According to the work of [28], the recursive formula for updating the model's parameters vector β_n in terms of previous estimates is:

$$\beta_n = \beta_{n-1} + \frac{\Sigma_{\beta_{n-1}} x_n \left(y_n - {x_n}^T \beta_{n-1}\right)}{\sigma^2 \gamma + {x_n}^T \Sigma_{\beta_{n-1}} x_n} \tag{7}$$

Similarly, the covariance matrix of the model's parameter vector, Σ_{β_n} is recursively updated as follows:

$$\Sigma_{\beta_n} = \frac{1}{\gamma}\Sigma_{\beta_{n-1}} - \frac{\Sigma_{\beta_{n-1}} x_n {x_n}^T \Sigma_{\beta_{n-1}}}{\sigma^2\gamma^2 + \gamma {x_n}^T \Sigma_{\beta_{n-1}} x_n} \tag{8}$$

B. ESTIMATING THE VARIANCE OF THE RANDOM ERROR TERM (σ^2)

In the last section, we have presented the recursive formulas of the regression model's parameter vector β, and the covariance matrix Σ_β using the work presented in [28]. In this section, we estimate the variance σ^2 of the random error term ϵ defined in the adopted linear demand model (see Eq. (1)). We use the maximum likelihood estimator [62] for evaluating the variance parameter σ^2.

The likelihood function is defined as:

$$\mathcal{L}(\sigma^2, \beta) = \prod_{n=1}^{T} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{\sum_{n=1}^{T} -\gamma^{T-n}(y_n - \beta^T x_n)^2}{2\sigma^2}} \tag{9}$$

where T denotes the number of data points used in the estimate, and γ is the discount factor of the weighted linear regression. Accordingly, the log-likelihood function is calculated as follows:

$$l(\sigma^2, \beta) = -T \log\sigma - T \log\sqrt{2\pi} - \frac{\sum_{n=1}^{T} \gamma^{T-n}(y_n - \beta^T x_n)^2}{2\sigma^2} \tag{10}$$

Maximizing the log-likelihood in Eq. (10) results in the following estimate $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{\sum_{n=1}^{T} \gamma^{T-n}(y_n - \beta^T x_n)^2}{T} \tag{11}$$

It is evident from Eq. (11) that this estimate represents the variance of data.

In order to unify all the update equations to be recursive, we derive a recursive version for estimating the variance of the error term as follows:

$$\sigma_n^2 = \frac{\gamma(n-1)}{n}\sigma_{n-1}^2 + \frac{{e_n}^2}{n} \tag{12}$$

where $e_n = y_n - {\beta^T}_{n-1} x_n$.

C. EVALUATING THE PREDICTIVE DISTRIBUTION

In this section, we present the formulas for the predictive distribution of the adopted regression model. The regression coefficients' vector β follows a multivariate Gaussian distribution [63]. The expectation of this Gaussian distribution is evaluated in Eq. (7), and the covariance matrix is estimated in Eq. (8).

From Eq. (2), it can be inferred that the predicted value at time step n (y_n) follows a Gaussian distribution as follows:

$$y_n \sim \mathcal{N}(\mathbb{E}[y_n|x_n, \beta_{n-1}], {\sigma^2}_{y_n|x_n,\beta_{n-1}}) \tag{13}$$

According to Eq. (2), the expected value of y_n is evaluated as follows:

$$\mathbb{E}[y_n|x_n, \beta_{n-1}] = {x_n}^T \beta_{n-1} \tag{14}$$

The variance of y_n is calculated as:

$${\sigma^2}_{y_n|x_n,\beta_{n-1}} = {x_n}^T \Sigma_{\beta_{n-1}} x_n + {\sigma^2}_{n-1} \tag{15}$$

V. MODEL FORMULATION

In this section we present the formulation of the proposed multi-step look-ahead approach. First, we define the formulation for any general utility function u. Then, we apply the proposed formulation to the dynamic pricing application by setting the target utility function to the revenue function. We provide the detailed formulation of the one-step look-ahead case, and we use recursive updates for multiple look-ahead steps.

A. GENERAL FORMULATION

Typically, the greedy approach, which maximizes the immediate utility, obtains myopic and sub-optimal solutions, especially when the utility function incurs some uncertainty. In this work, we propose less myopic solutions for utility optimization by looking ahead for one or more future steps of the considered utility function u. The proposed approach seeks to achieve near-optimal solutions.

We first consider a one-step look-ahead formulation where the objective choosing the point x_{n-1} that maximizes the expected utility at step n, for a general utility function u. Then, we apply the proposed formulation to the dynamic pricing problem where the utility function is the attained revenue.

For a general utility function u, the expected one-step look-ahead utility function at time step n given the $n-2$ previously acquired data points and the underlying point (the optimization variable) x_{n-1}, $\mathbb{E}[u(\mathcal{D}_n)|\mathcal{D}_{n-2}, x_{n-1}]$, is evaluated as:

$$\begin{aligned}&\mathbb{E}[u(\mathcal{D}_n)|\mathcal{D}_{n-2}, x_{n-1}] \\ &= \int_{y_n}\int_{x_n}\int_{y_{n-1}} u(\mathcal{D}_n) \\ &\quad \times Pr[y_n|x_n, \mathcal{D}_{n-2}, x_{n-1}, y_{n-1}]Pr[x_n|\mathcal{D}_{n-2}, x_{n-1}, y_{n-1}]] \\ &\quad \times Pr[y_{n-1}|\mathcal{D}_{n-2}, x_{n-1}]\, \mathrm{d}y_{n-1}\, \mathrm{d}x_n\, \mathrm{d}y_n\end{aligned} \tag{16}$$

However, the probability $Pr[x_n|\mathcal{D}_{n-2}, x_{n-1}, y_{n-1}]$ can be safely ignored since the selection rule of x_n given $\mathcal{D}_{n-1}$ is deterministic as will be further illustrated in Section V-B.

Thus, Eq. (16) can be simplified into:

$$\begin{aligned}&\mathbb{E}[u(\mathcal{D}_n)|\mathcal{D}_{n-2}, x_{n-1}] \\ &= \int_{y_n}\int_{y_{n-1}} u(\mathcal{D}_n) Pr[y_n|x_n, \mathcal{D}_{n-1}] \\ &\quad \times Pr[y_{n-1}|\mathcal{D}_{n-2}, x_{n-1}]\, \mathrm{d}y_{n-1}\, \mathrm{d}y_n\end{aligned} \tag{17}$$

where the probabilities $Pr[y_n|x_n, \mathcal{D}_{n-1}]$ and $Pr[y_{n-1}|\mathcal{D}_{n-2}, x_{n-1}]$ are evaluated according to the regression model as defined in Eq. (13) (see Section IV-C).

B. DYNAMIC PRICING FORMULATION

In this section, we apply the general formulation presented in Section V-A to the dynamic pricing problem. We use the cumulative revenue function $R(\mathcal{D}_n)$ as the utility function to be maximized.

The cumulative revenue function at time step n is recursively defined as:

$$R(\mathcal{D}_n) = R(\mathcal{D}_{n-1}) + p_n y_n \tag{18}$$

where $R(\mathcal{D}_{n-1})$ is the cumulative revenue accomplished so far from time step 1 till time step $n-1$, p_n is the experimented price at time step n, and y_n is the corresponding demand.

The expected value of the cumulative revenue function at time step n given $n-1$ price-demand pairs is defined as:

$$\mathbb{E}[R(\mathcal{D}_n)|x_n, \mathcal{D}_{n-1}] = R(\mathcal{D}_{n-1}) + p_n \mathbb{E}[y_n|x_n, \mathcal{D}_{n-1}] \tag{19}$$

where $x_n = [1 \;\; p_n]^T$.

Using the look-ahead utility definition at Eq. (16) in Section V-A, and the recursive definition of the revenue function presented in Eq. (18), the expected one-step look-ahead revenue is evaluated as follows:

$$\begin{aligned}
&\mathbb{E}[R(\mathcal{D}_n)|x_{n-1}, \mathcal{D}_{n-2}] \\
&\quad = \mathbb{E}[R(\mathcal{D}_{n-1})|x_{n-1}, \mathcal{D}_{n-2}] \\
&\qquad + \int_{y_n}\int_{p_n}\int_{y_{n-1}} p_n y_n Pr[p_n|\mathcal{D}_{n-2}, x_{n-1}, y_{n-1}] \\
&\qquad \times Pr[y_n|x_n, x_{n-1}, y_{n-1}, \mathcal{D}_{n-2}] \\
&\qquad \times Pr[y_{n-1}|\mathcal{D}_{n-2}, x_{n-1}]\, \mathrm{d}y_{n-1}\, \mathrm{d}p_n \mathrm{d}y_n
\end{aligned} \tag{20}$$

Substituting for the first term $\mathbb{E}[R(\mathcal{D}_{n-1})|x_{n-1}, \mathcal{D}_{n-2}]$ using the recursive definition of Eq. (19), the expected one-step look-ahead revenue can be expressed as:

$$\begin{aligned}
&\mathbb{E}[R(\mathcal{D}_n)|\mathcal{D}_{n-2}, x_{n-1}] \\
&\quad = R(\mathcal{D}_{n-2}) + p_{n-1}\mathbb{E}[y_{n-1}|x_{n-1}, \mathcal{D}_{n-2}] \\
&\qquad + \int_{p_n}\int_{y_{n-1}} p_n Pr[p_n|\mathcal{D}_{n-2}, x_{n-1}, y_{n-1}] \\
&\qquad \times Pr[y_{n-1}|x_{n-1}, \mathcal{D}_{n-2}] \\
&\qquad \times \mathbb{E}[y_n|x_n, x_{n-1}, y_{n-1}, \mathcal{D}_{n-2}]\, \mathrm{d}y_{n-1}\, \mathrm{d}p_n
\end{aligned} \tag{21}$$

However, as previously mentioned in Section V-A, the probability $Pr[p_n|\mathcal{D}_{n-2}, x_{n-1}, y_{n-1}]$ is simply evaluated since the selection rule of p_n given $\mathcal{D}_{n-1}$ is deterministic. Convincingly, x_n is set so as to maximize the immediate revenue at time step n according to Eq. (19).

Accordingly, Eq.(21) could be further simplified into:

$$\begin{aligned}
&\mathbb{E}[R(\mathcal{D}_n)|\mathcal{D}_{n-2}, x_{n-1}] \\
&\quad = R(\mathcal{D}_{n-2}) + p_{n-1}\mathbb{E}[y_{n-1}|x_{n-1}, \mathcal{D}_{n-2}] \\
&\qquad + \int_{y_{n-1}} p_n Pr[y_{n-1}|\mathcal{D}_{n-2}, x_{n-1}] \\
&\qquad \times \mathbb{E}[y_n|x_n, x_{n-1}, y_{n-1}, \mathcal{D}_{n-2}]\, \mathrm{d}y_{n-1}
\end{aligned} \tag{22}$$

Using the dataset $\mathcal{D}$ definition presented in Section III, the expected one-step look-ahead revenue is expressed as follows:

$$\begin{aligned}
&\mathbb{E}[R(\mathcal{D}_n)|\mathcal{D}_{n-2}, x_{n-1}] \\
&\quad = R(\mathcal{D}_{n-2}) + p_{n-1}\mathbb{E}[y_{n-1}|x_{n-1}, \mathcal{D}_{n-2}] \\
&\qquad + \int_{y_{n-1}} p_n Pr[y_{n-1}|\mathcal{D}_{n-2}, x_{n-1}] \\
&\qquad \times \mathbb{E}[y_n|x_n, \mathcal{D}_{n-1}]\, \mathrm{d}y_{n-1}
\end{aligned} \tag{23}$$

Inspired by reinforcement learning literature [42], we apply a discount factor γ_r to future revenues in order to boost instantaneous revenue gain, as follows:

$$\begin{aligned}
&\mathbb{E}[R(\mathcal{D}_n)|\mathcal{D}_{n-2}, x_{n-1}] \\
&\quad = R(\mathcal{D}_{n-2}) + p_{n-1}\mathbb{E}[y_{n-1}|x_{n-1}, \mathcal{D}_{n-2}] \\
&\qquad + \gamma_r \int_{y_{n-1}} p_n Pr[y_{n-1}|\mathcal{D}_{n-2}, x_{n-1}] \\
&\qquad \times \mathbb{E}[y_n|x_n, \mathcal{D}_{n-1}]\, \mathrm{d}y_{n-1}
\end{aligned} \tag{24}$$

The probability $Pr[y_{n-1}|\mathcal{D}_{n-2}, x_{n-1}]$ presented in the integral of Eq. (24) is evaluated according to the linear regression model as defined in Eq. (13). Similarly, the expectation $\mathbb{E}[y_n|x_n, \mathcal{D}_{n-1}]$ is computed using Eq. (14) (see Section IV-C).

Concerning the price p_n presented in Eq. (24), it is set so as to maximize the expected immediate revenue at time step n, which is denoted as $\mathbb{E}[r(p_n)|\mathcal{D}_{n-1}]$, as follows:

$$\begin{aligned}
p_n &= argmax_{p^*}\; \mathbb{E}[r(p^*)|\mathcal{D}_{n-1}] \\
&= argmax_{p^*}\; p_n \mathbb{E}[y_n|x^*, \mathcal{D}_{n-1}] \\
&= argmax_{p^*}\; p^*(x^{*T}\beta_{n-1}) \\
&= argmax_{p^*}\; b_{n-1}p^{*2} + a_{n-1}p^*
\end{aligned} \tag{25}$$

Accordingly, the price p_n can be evaluated in terms of β_{n-1} components as:

$$p_n = \frac{-a_{n-1}}{b_{n-1}} \tag{26}$$

Then, substituting for the price p_n from Eq. (26) into Eq. (25), the maximum expected immediate revenue is computed as:

$$\mathbb{E}[r(p_n)|\mathcal{D}_{n-1}] = \frac{-a^2{}_{n-1}}{4b_{n-1}} \tag{27}$$

Substituting from Eq. (27) into the third term of Eq. (24) results in:

$$\begin{aligned}
&\mathbb{E}[R(\mathcal{D}_n)|\mathcal{D}_{n-2}, x_{n-1}] \\
&\quad = R(\mathcal{D}_{n-2}) + p_{n-1}\mathbb{E}[y_{n-1}|x_{n-1}, \mathcal{D}_{n-2}] \\
&\qquad + \gamma_r \int_{y_{n-1}} \frac{-a^2{}_{n-1}}{4b_{n-1}} Pr[y_{n-1}|\mathcal{D}_{n-2}, x_{n-1}]\mathrm{d}y_{n-1}
\end{aligned} \tag{28}$$

After several substitutions presented in the appendix, Eq. (28) can be formulated as:

$$\begin{aligned}&\mathbb{E}[R(\mathcal{D}_n)|\mathcal{D}_{n-2}, x_{n-1}]\\&= R(\mathcal{D}_{n-2}) + a_{n-2}p_{n-1} + b_{n-2}{p^2}_{n-1}\\&\quad+\gamma_r\int_{y_{n-1}}\mathcal{N}(y_{n-1};\mu_{y_{n-1}(p_{n-1})},\sigma^2_{y_{n-1}}(p_{n-1}))\\&\quad\times\frac{-\big(A(p_{n-1})y_{n-1}+B(p_{n-1})\big)^2}{4\big(C(p_{n-1})y_{n-1}+D(p_{n-1})\big)}\mathrm{d}y_{n-1}\end{aligned}\tag{29}$$

where the expectation $\mu_{y_{n-1}}$ and the variance $\sigma^2_{y_{n-1}}$ are evaluated using Eq. (14) and Eq. (15), respectively as defined in Section IV-C.

The four terms: $A(p_{n-1})$, $B(p_{n-1})$, $C(p_{n-1})$, and $D(p_{n-1})$ are four polynomials of p_{n-1} which are evaluated as follows:

$$A(p_{n-1}) = \sigma^2_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1}\tag{30}$$

$$\begin{aligned}B(p_{n-1}) =\ &\big(b_{n-2}\sigma_{ab_{n-2}} + a_{n-2}\sigma^2_{b_{n-2}}\big)p^2_{n-1}\\&+\big(a_{n-2}\sigma_{ab_{n-2}} + b_{n-2}\sigma^2_{a_{n-2}}\big)p_{n-1}\\&+\gamma a_{n-2}\sigma^2_{n-2}\end{aligned}\tag{31}$$

$$\begin{aligned}C(p_{n-1}) =\ &\sigma^4_{b_{n-2}}p^3_{n-1} + 3\sigma_{ab_{n-2}}\sigma^2_{b_{n-2}}p^2_{n-1}\\&+\big(2{\sigma^2}_{ab_{n-2}} + \sigma^2_{b_{n-2}}(\gamma\sigma^2_{n-2} + {\sigma^2}_{a_{n-2}})\big)p_{n-1}\\&+\sigma_{ab_{n-2}}(\gamma\sigma^2_{n-2} + {\sigma^2}_{a_{n-2}})\end{aligned}\tag{32}$$

$$\begin{aligned}D(p_{n-1}) =\ &2b_{n-2}{\sigma^4}_{b_{n-2}}p^4_{n-1}\\&+\big(7b_{n-2}\sigma^2_{b_{n-2}}\sigma_{ab_{n-2}} - a_{n-2}\sigma^4_{b_{n-2}}\big)p^3_{n-1}\\&+\big(4b_{n-2}\sigma^2_{ab_{n-2}} + 3b_{n-2}\sigma^2_{b_{n-2}}(\gamma\sigma^2_{n-2} + {\sigma^2}_{a_{n-2}})\\&+2b_{n-2}{\sigma^2}_{ab_{n-2}} - 3a_{n-2}\sigma_{ab_{n-2}}\sigma^2_{b_{n-2}}\big)p^2_{n-1}\\&+\big((5b_{n-2}\sigma_{ab_{n-2}} - a_{n-2}\sigma^2_{b_{n-2}})(\gamma\sigma^2_{n-2} + {\sigma^2}_{a_{n-2}})\\&\quad-2a_{n-2}\sigma^2_{ab_{n-2}}\big)\\&\times p_{n-1} + \big(b_{n-2}(\gamma\sigma^2_{n-2} + {\sigma^2}_{a_{n-2}}) - a_{n-2}\sigma_{ab_{n-2}}\big)\\&\times(\gamma\sigma^2_{n-2} + {\sigma^2}_{a_{n-2}})\end{aligned}\tag{33}$$

Finally, it is worth noting that the key objective of the one-step look-ahead formulation is to choose the pricing point p_{n-1} maximizing the expected revenue at time step n. Using Eq. (29), the underlying optimization objective can be formulated as:

$$\begin{aligned}{p^*}_{n-1} =\ &argmax_{p_{n-1}}\ R(\mathcal{D}_{n-2}) + a_{n-2}p_{n-1} + b_{n-2}{p^2}_{n-1}\\&+\gamma_r\int_{y_{n-1}}\mathcal{N}(y_{n-1};\mu_{y_{n-1}(p_{n-1})},\sigma^2_{y_{n-1}}(p_{n-1}))\\&\times\frac{-\big(A(p_{n-1})y_{n-1}+B(p_{n-1})\big)^2}{4\big(C(p_{n-1})y_{n-1}+D(p_{n-1})\big)}\mathrm{d}y_{n-1}\end{aligned}\tag{34}$$

The term $R(\mathcal{D}_{n-2})$ is constant, thus it is independent of the optimization variable p_{n-1}, and Eq. (34) can be simplified into:

$$\begin{aligned}{p^*}_{n-1} =\ &argmax_{p_{n-1}}\ a_{n-2}p_{n-1} + b_{n-2}{p^2}_{n-1}\\&+\gamma_r\int_{y_{n-1}}\mathcal{N}(y_{n-1};\mu_{y_{n-1}(p_{n-1})},\sigma^2_{y_{n-1}}(p_{n-1}))\\&\times\frac{-\big(A(p_{n-1})y_{n-1}+B(p_{n-1})\big)^2}{4\big(C(p_{n-1})y_{n-1}+D(p_{n-1})\big)}\mathrm{d}y_{n-1}\end{aligned}\tag{35}$$

However, firms generally prefer that prices belong to a controllable range. Thus, we assume price bounds: p_l and p_u such that any potential price belongs to the range $[p_l, p_u]$.

Accordingly, the optimization problem defined in Eq. (35) turns to be a constrained optimization problem as follows:

$$\begin{aligned}{p^*}_{n-1} =\ &argmax_{p_{n-1},\ p_l\le p_{n-1}<p_u}\ \ a_{n-2}p_{n-1} + b_{n-2}{p^2}_{n-1}\\&+\gamma_r\int_{y_{n-1}}\mathcal{N}(y_{n-1};\mu_{y_{n-1}(p_{n-1})},\sigma^2_{y_{n-1}}(p_{n-1}))\\&\times\frac{-\big(A(p_{n-1})y_{n-1}+B(p_{n-1})\big)^2}{4\big(C(p_{n-1})y_{n-1}+D(p_{n-1})\big)}\mathrm{d}y_{n-1}\end{aligned}\tag{36}$$

To solve the optimization problem in Eq. (36), a constrained optimization algorithm or a simple grid search over the potential pricing values could be used. In our implementation, we employ the interior point optimization algorithm [64].

For multiple look-ahead steps, we apply the look-ahead formulation recursively. For example, for a two-step look-ahead pricing, assume that at time step $n-1$, the objective is to maximize the expected revenue at step $n+1$. This can be recursively expressed as:

$$\begin{aligned}&\mathbb{E}[R(\mathcal{D}_{n+1})|x_{n-1}, \mathcal{D}_{n-2}]\\&\quad=\mathbb{E}[R(\mathcal{D}_n)|x_{n-1}, \mathcal{D}_{n-2}]\\&\qquad+{\gamma_r}^2\int_{y_n}p_{n+1}Pr[y_n|x_n, \mathcal{D}_{n-1}]\\&\qquad\times\mathbb{E}[y_{n+1}|x_{n+1}, x_n, y_n, \mathcal{D}_{n-1}]\,\mathrm{d}y_n\end{aligned}\tag{37}$$

The two-step look-ahead expected revenue is evaluated as follows: first, the first term of Eq. (37) is recursively evaluated using Eq. (29). Then, the second term of Eq. (37) is evaluated (see the appendix for more details). The second term is multiplied by ${\gamma_r}^2$ since it represents the two-step future revenue. The same recursive equation (Eq. (37)) can be applied in case of further number of look-ahead steps.

The final formulation of the price p_{n-1} maximizing the two-step look-ahead revenue is defined as:

$$\begin{aligned}&{p^*}_{n-1}\\&=argmax_{p_{n-1},\ p_l\le p_{n-1}<p_u}\ \ a_{n-2}p_{n-1} + b_{n-2}{p^2}_{n-1}\\&\quad+\gamma_r\int_{y_{n-1}}\mathcal{N}(y_{n-1};\mu_{y_{n-1}(p_{n-1})},\sigma^2_{y_{n-1}}(p_{n-1}))\\&\quad\times\frac{-\big(A(p_{n-1})y_{n-1}+B(p_{n-1})\big)^2}{4\big(C(p_{n-1})y_{n-1}+D(p_{n-1})\big)}\mathrm{d}y_{n-1}\\&\quad+{\gamma_r}^2\int_{y_n}\mathcal{N}\big(y_n;(\mu_{y_n}\circ W)(p_{n-1}),({\sigma^2}_{y_n}\circ W)(p_{n-1})\big)\end{aligned}$$

$$\times \frac{-\Big((J \circ W)(p_{n-1})y_n + (F \circ W)(p_{n-1})\Big)^2}{4\Big((G \circ W)(p_{n-1})y_n + (H \circ W)(p_{n-1})\Big)} \mathrm{d}y_n$$

where the four polynomials J, F, G, and H are defined in Eq. (70), Eq. (72), Eq. (74), and Eq. (76), respectively. In addition, the W function is defined in Eq. (78). The detailed formulation of the two-step look-ahead is provided in the appendix.

VI. EXPERIMENTS

We apply the proposed look-ahead formulation defined in Section V-B with two values for the number of steps: one step and two steps. We have found that increasing the number of steps beyond two steps does not lead to further improvement.

We have conducted experiments using synthetic and real datasets in order to assess the performance of the proposed approach.

A. BENCHMARKS

In order to perform a comprehensive analysis, we compare our proposed approach to the myopic pricing and two other pricing algorithms proposed in literature: myopic pricing with dithering [23], and the controlled variance pricing (CVP) policy [24].

The myopic pricing applies pure exploitation by choosing the price that maximizes the immediate revenue. However, myopic pricing yields sub-optimal revenues since it does not consider learning the demand function [35].

The myopic pricing with dithering is a simple variant of myopic pricing [23]. It is a myopic-type pricing that incorporates a random perturbation to the myopic price in order to enhance demand learning by introducing some exploration.

The controlled variance pricing strategy chooses the price value that maximizes the immediate revenue provided that it does not belong to a specific taboo interval around the average of the prices selected so far [24]. Thus, the CVP method injects some price diversity, which enhances exploration.

In addition, we apply another benchmark strategy that performs exploration and exploitation separately in two phases. In this strategy, the first phase is a pure exploration with the goal being to obtain an accurate estimate of model parameters. Exploration in this phase is performed by simple random sampling. In the next phase, the estimated demand model is used and pure exploitation is implemented using the myopic pricing policy. We name this benchmark as Random-Myopic policy.

B. PERFORMANCE METRICS

We assess the performance of the different pricing policies in terms of two major aspects: revenue maximization and demand model estimation accuracy. However, revenue maximization is essentially the primary objective. The revenue maximization objective is measured in terms of the gained revenue. We evaluate a normalized version of the cumulative discounted revenue R_T evaluated at the last time step T. We normalize R_T with respect to the optimal revenue if the true demand model parameters are known.

The revenue gain is defined as follows:

$$Rev\ Gain = \frac{R_T}{R_{opt}} = \frac{\sum_{n=1}^{T} {\gamma_r}^{n-1} r_n}{\sum_{n=1}^{T} {\gamma_r}^{n-1} r_{opt}} \tag{38}$$

where r_n is the revenue gained at step n, and r_{opt} is the optimal revenue given the true model parameters a and b, which is expressed as:

$$r_{opt} = p_{opt}(a + bp_{opt}) = b{p^2}_{opt} + ap_{opt} \tag{39}$$

where p_{opt} is the optimal price, which equals to $\frac{-a}{2b}$ in case of the adopted linear demand model (see Eq. (1)). The parameters a and b are the ground truth values of the demand model parameters.

Simplifying the summation term $\sum_{n=1}^{T} {\gamma_r}^{n-1}$ in the denominator of Eq. (38) using the summation of geometric series formula, we get:

$$Rev\ Gain = \frac{R_T}{R_{opt}} = \frac{\sum_{n=1}^{T} {\gamma_r}^{n-1} r_n}{(1 - {\gamma_r}^{T})/(1 - \gamma_r) r_{opt}} \tag{40}$$

In addition to evaluating the gained revenue, we test whether the final price converges to the true optimal price by measuring the deviation of the price p_T, at last time step T, from the true optimal price p_{opt}.

$$\delta_p = \frac{|p_T - p_{opt}|}{p_{opt}} \tag{41}$$

The demand model estimation accuracy is evaluated in terms of the model estimation error of the final estimated demand model's parameter vector $\hat{\beta}_T$, at time step T, as shown in Eq.(42):

$$\delta_\beta = \frac{||\beta - \hat{\beta}_T||_2}{||\beta||_2} \tag{42}$$

C. EXPERIMENTAL SETUP

The simulation proceeds as follows: after generating a pool of price-demand data, we start with a very limited number of points, three points (less than three points cannot give any sensible initial parameter estimate). Then, we train a regression model to obtain an initial estimate for the model parameters β_0, and the corresponding covariance matrix Σ_{β_0}. After that, we apply the proposed look-ahead optimization approach in order to obtain the look-ahead price at iteration n, denoted as p_n. Then, the corresponding demand y_n is observed. It follows the linear demand model (Eq. (1)), with the error term ϵ giving random fluctuations around the true demand line. We utilize this acquired point (p_n, y_n) to update the demand model estimates β and Σ_β using the recursive weighted linear regression update equations (Eq. (7) and Eq. (8)). The simulation loop continues till reaching a certain predefined number of iterations T.

In the conducted experiments, the number of iterations T is set to 100, and the discount factor of the weighted linear regression, γ is set to 0.99. Similarly, the discount value γ_r

used in the look-ahead formulation (Eq. (24)) is set to 0.99. To ensure the reliability of the results, we run each experiment 20 times and we present the average results over the runs.

In our implementation, for the considered two-phase random-myopic strategy, we use the same number of iterations for the exploration phase as for the exploitation phase, i.e. 50 for each. Regarding the myopic pricing with dithering method [23], we set the amount of dithering to 0.1.

To have a fair comparison among all the adopted pricing strategies, we employ a unified method for estimating the demand model parameters. Specifically, we apply the recursive weighted linear regression model described in Section IV.

D. EXPERIMENTS USING SYNTHETIC DATASETS

First, we apply our two proposed look-ahead pricing methods: one-step and two-step look-ahead, and the other benchmark pricing methods to synthetic datasets. The advantage of using synthetic data is that the true model's parameter vector β is known. Accordingly, the revenue gain can be accurately estimated with the knowledge of the true optimal revenue (see Eq. (40)). In addition, the demand model estimation error can be precisely evaluated (see Eq. (42)). We create synthetic datasets by generating several price points and then assuming linear demand model, we calculate the corresponding demands using Eq. (1). We generate twenty synthetic datasets using different values for parameters a, b, and σ.

We investigate different values for the demand price elasticity parameter b including the three demand elasticity cases of inelastic, neutral, and elastic demands [65]. Furthermore, we adopt two different values for the standard deviation σ of the error term representing low (5%) and high (40%) error settings. We use different values for the standard deviation σ of the error term to analyze the impact of the error term on the different pricing policies, and to quantify their robustness to errors. Additionally, in the dynamic pricing application, using different error settings could be considered as aggregating all other influencing factors that may be hard to incorporate in the model such as: competition, seasonality, and perishability of the products.

Tables 2-4 represent the gain in revenue, the estimation error of model parameters β, and the percentage error of the estimated price with respect to the optimal price, respectively. These tables show the results averaged over the twenty synthetic datasets in case of low and high error settings.

Figure 1 demonstrates the performance of different pricing policies over time horizon T in terms of the cumulative discounted revenue compared to the optimal revenue, using a synthetic dataset with parameters $a = 408.17$, $b = -1.32$, and $\sigma = 163$. Figure 2 shows the chosen prices over T iterations by different pricing strategies compared to the optimal price for the same synthetic dataset.

E. EXPERIMENTS USING REAL DATASETS

We apply our experiments to five real demand-price datasets described in Table 5. For the transport dataset, we gather it online through surveying. This dataset is essentially transportation ticket pricing data. We have asked users about the minimum and the maximum fares they are willing to pay for an economy class bus ticket between two certain cities. We have received 41 responses from different users. In order

TABLE 2. Revenue gain of different pricing methods, averaged over twenty different synthetic datasets for two different settings of the standard deviation of the error term. The methods are sorted in descending order according to their average revenue gain over the two settings of the standard deviation of the error term. The bold entries represent the maximum revenue gain per column over all methods. The myopic pricing chooses the price maximizing the immediate revenue. The myopic dithering adds a random perturbation to the myopic price. The controlled variance pricing (CVP) method selects the myopic price unless it is close to the previously chosen prices. The random-myopic baseline first applies exploration using random sampling, then it performs exploitation using myopic pricing.

Method	$\sigma = 5\%$	$\sigma = 40\%$	Average
Look-ahead (1s)	99.12 %	**84.05 %**	**91.58 %**
Look-ahead (2s)	**99.13 %**	83.72 %	91.42 %
CVP	96.89 %	81.34 %	89.12 %
Myopic-dithering	98.66 %	71.75 %	85.20 %
Myopic	98.59 %	70.61 %	84.60 %
Random-Myopic	79.44 %	77.47 %	78.45 %

TABLE 3. Percentage error in estimating model's parameter vector β of different pricing methods, averaged over twenty different synthetic datasets for two different settings of the standard deviation of the error term. The methods are sorted in ascending order according to their average percentage model error over the two settings of the standard deviation of the error term. The bold entries represent the minimum model error per column over all methods. The myopic pricing chooses the price maximizing the immediate revenue. The myopic dithering adds a random perturbation to the myopic price. The controlled variance pricing (CVP) method selects the myopic price unless it is close to the previously chosen prices. The random-myopic baseline first applies exploration using random sampling, then it performs exploitation using myopic pricing.

Method	$\sigma = 5\%$	$\sigma = 40\%$	Average
Random-Myopic	**1.39 %**	**9.81 %**	**5.60 %**
CVP	3.10 %	21.31 %	12.21 %
Look-ahead (1s)	3.61 %	25.08 %	14.34 %
Look-ahead (2s)	3.43 %	26.41 %	14.92 %
Myopic	3.71 %	28.88 %	16.29 %
Myopic-dithering	3.66 %	29.52 %	16.59 %

TABLE 4. Percentage error of the final estimated price p_T for all pricing methods, averaged over twenty different synthetic datasets for two different settings of the standard deviation of the error term. The methods are sorted in ascending order according to their average price deviation over the two settings of the standard deviation of the error term. The bold entries represent the minimum price deviation per column over all methods. The myopic pricing chooses the price maximizing the immediate revenue. The myopic dithering adds a random perturbation to the myopic price. The controlled variance pricing (CVP) method selects the myopic price unless it is close to the previously chosen prices. The random-myopic baseline first applies exploration using random sampling, then it performs exploitation using myopic pricing.

Method	$\sigma = 5\%$	$\sigma = 40\%$	Average
Random-Myopic	**0.59 %**	**4.96 %**	**2.77 %**
CVP	3.43 %	10.83 %	7.13 %
Look-ahead (1s)	1.48 %	16.10 %	8.79 %
Look-ahead (2s)	1.33 %	17.94 %	9.63 %
Myopic-dithering	1.40 %	34.62 %	18.01 %
Myopic	1.69 %	35.49 %	18.59 %

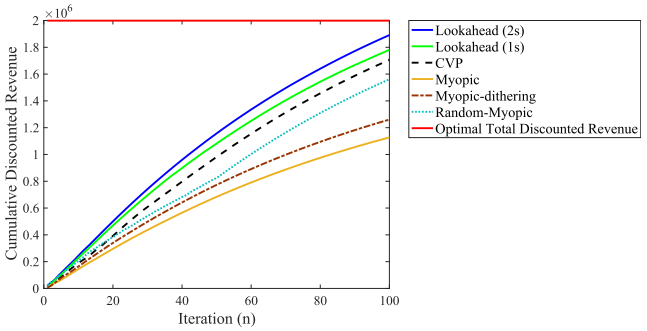

FIGURE 1. Cumulative discounted revenue of different pricing policies over 100 iterations for the synthetic dataset $a = 408.17$, $b = -1.32$, and $\sigma = 163$.

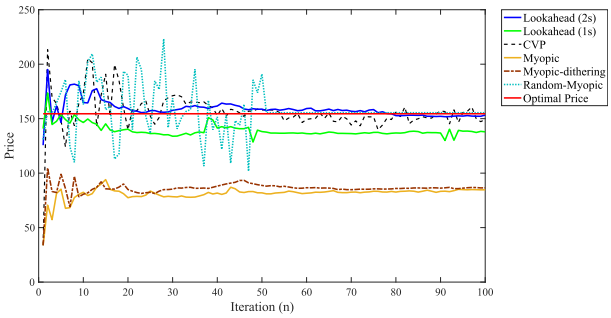

FIGURE 2. Prices of different pricing policies over 100 iterations for the synthetic dataset $a = 408.17$, $b = -1.32$, and $\sigma = 163$.

TABLE 5. A description for the real-world datasets.

Dataset	Size	$\hat{a}$	$\hat{b}$
Transport	41	41.3778	−0.1378
Beef	91	30.0515	−0.0465
Sugar	18	1.3576	−0.3184
Spirits	69	4.4651	−1.2723
Coke	20	50.5700	−0.3406

to have the data in the form of price and demand pairs, we perform the following. For each price, we calculate the corresponding demand as the number of users who can afford this price according to their stated minimum and maximum prices.

For the other four datasets: the beef dataset is obtained from the USDA Red Meats Yearbook [66]. Similarly, the spirits dataset is obtained from [67]. The sugar dataset is adopted from [68], and the coke dataset is acquired from [69].

In spite of the importance of performing our experiments on real datasets, there is one hurdle in applying dynamic pricing methods on real datasets. In dynamic pricing, at each time step n, a certain price p_n is chosen according to a certain criterion. A real dataset has a finite set of price-demand points. Thus, the chosen price p_n can be outside the available prices provided in the real dataset. Consequently, we use real datasets mainly for estimating linear demand model's parameters vector β only. Then, we generate data using the estimated parameters, with the same methodology described in Section VI-D. The regression model coefficients a and b are estimated using ordinary least squares linear regression.

TABLE 6. Revenue gain of all pricing methods, averaged over the five real datasets, using different error settings σ. The strategies are sorted in descending order according to their average revenue gain for the two settings of the standard deviation of the error term. The bold entries represent the maximum revenue gain per column over all methods. The myopic pricing chooses the price maximizing the immediate revenue. The myopic dithering adds a random perturbation to the myopic price. The controlled variance pricing (CVP) method selects the myopic price unless it is close to the previously chosen prices. The random-myopic baseline first applies exploration using random sampling, then it performs exploitation using myopic pricing.

Method	$\sigma = 5\%$	$\sigma = 40\%$	Average
Lookahead (2s)	**98.66 %**	**84.37 %**	**91.52 %**
Lookahead (1s)	98.62 %	83.50 %	91.06 %
Myopic-dithering	98.26 %	76.14 %	87.20 %
CVP	95.14 %	78.55 %	86.85 %
Myopic	98.32 %	75.31 %	86.81 %
Random-Myopic	81.12 %	78.02 %	79.57 %

TABLE 7. Percentage model error for all pricing methods, over the five real datasets, using different error settings σ. The strategies are sorted in ascending order according to their model estimation error for the different error settings. The bold entries represent the minimum model error per column over all methods. The myopic pricing chooses the price maximizing the immediate revenue. The myopic dithering adds a random perturbation to the myopic price. The controlled variance pricing (CVP) method selects the myopic price unless it is close to the previously chosen prices. The random-myopic baseline first applies exploration using random sampling, then it performs exploitation using myopic pricing.

Method	$\sigma = 5\%$	$\sigma = 40\%$	Average
Random-Myopic	**3.11 %**	**13.02 %**	**8.06 %**
Lookahead (1s)	6.24 %	23.72 %	14.98 %
Lookahead (2s)	5.87 %	24.34 %	15.11 %
CVP	5.06 %	25.68 %	15.37 %
Myopic-dithering	6.13 %	27.77 %	16.95 %
Myopic	7.06 %	28.84 %	17.95 %

TABLE 8. Price deviation from the optimal price for all methods, averaged over the five real datasets, using different error settings σ. The strategies are sorted in ascending order according to their average price deviation over the different error settings. The bold entries represent the minimum price deviation per column over all methods. The myopic pricing chooses the price maximizing the immediate revenue. The myopic dithering adds a random perturbation to the myopic price. The controlled variance pricing (CVP) method selects the myopic price unless it is close to the previously chosen prices. The random-myopic baseline first applies exploration using random sampling, then it performs exploitation using myopic pricing.

Method	$\sigma = 5\%$	$\sigma = 40\%$	Average
Random-Myopic	**0.40 %**	**9.61 %**	**5.01 %**
Look-ahead (1s)	2.32 %	14.92 %	8.62 %
Look-ahead (2s)	2.88 %	14.46 %	8.67 %
CVP	3.76 %	15.10 %	9.43 %
Myopic	2.80 %	22.76 %	12.78 %
Myopic-dithering	3.08 %	23.64 %	13.36 %

Similar to the synthetic datasets, two different values for the standard deviation σ of the error term are experimented to represent low (5%) and high (40%) error settings.

Tables 6-8 present the revenue gain, the demand model estimation error, and the deviation error of the final estimated price from the optimal price, respectively. These tables display the results averaged over the five real datasets described in Table 5, for low and high error settings.

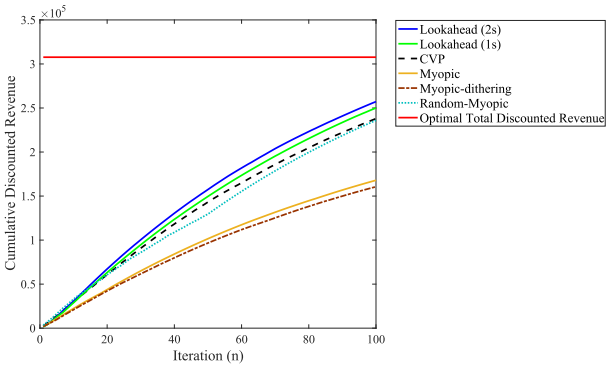

FIGURE 3. Cumulative discounted revenue of different pricing policies over 100 iterations for the beef dataset with the high error setting $\sigma = 12$.

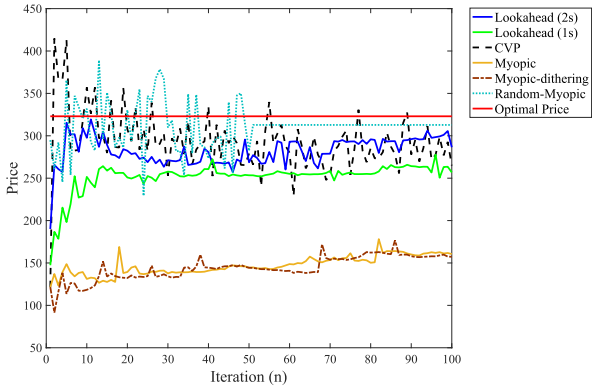

FIGURE 4. Prices of different pricing policies over 100 iterations for the beef dataset with the high error setting $\sigma = 12$.

Figure 3 shows the cumulative discounted revenue achieved by different pricing policies over time horizon T versus the optimal revenue using the beef dataset with high error setting $\sigma = 12$. Figure 4 represents the selected prices by different pricing methods over T steps versus the optimal price for the beef dataset with the high error setting $\sigma = 12$.

VII. DISCUSSION

In this section, we analyze the empirical results presented in Section VI. The main findings of the experimental results are summarized as follows:

- It can be observed from Table 2 and Table 6 that our proposed look-ahead methods outperform the myopic pricing and other benchmarks in terms of the achieved revenue gain in both cases of low and high error settings, for synthetic and real datasets.
- In particular, both of the one-step and two-step look-ahead methods surpass the performance of myopic pricing and all other benchmarks. For synthetic datasets, the one-step look-ahead achieves the highest revenue gain as indicated in Table 2. Concerning the real datasets, the two-step look-ahead method performs the best according to Table 6.
- The proposed look-ahead methods accomplish superior performance in terms of the gained revenue due to their less myopic behavior by considering not only the immediate revenue as myopic pricing and its variants do, but also future revenues. For example, the proposed look-ahead methods achieve above 2% and 4.5% revenue improvement over the CVP method on synthetic and real datasets, respectively (see Table 2 and Table 6). In addition, our proposed approach exceeds the myopic pricing by around 7% and 4.5% in case of synthetic and real datasets as indicated in Table 2 and Table 6, respectively.
- The myopic pricing with dithering policy incorporates some exploration to the myopic pricing, thus it enhances the myopic pricing performance with respect to the accomplished revenue. However, the look-ahead methods lead to around 6% and 4% revenue gain over myopic pricing with dithering, on synthetic and real datasets as presented in Table 2 and Table 6, respectively.
- Regarding the random-myopic pricing policy, it obtains the worst revenue gain in case of synthetic and real datasets. The reason for that is the long exploration phase which essentially compromises revenue by choosing non-rewarding prices.
- Table 2 and Table 6 demonstrate that our proposed look-ahead methods achieve significant revenue improvement over myopic pricing and other pricing policies, especially, in case of high error setting. Specifically, both of the look-ahead methods attain over 13% and $8-9\%$ revenue gain over the myopic pricing in case of high error setting, for synthetic and real datasets, respectively. Accordingly, one can conclude that the look-ahead pricing methods are robust towards high error settings.
- As presented in Table 2 and Table 6, the proposed look-ahead methods outperform the controlled variance pricing (CVP) method proposed in [24] in terms of the achieved revenue gain. Although both of the look-ahead and the CVP strategies consider exploitation in terms of immediate revenue maximization, they entirely differ in performing exploration. The look-ahead approach performs exploration automatically by incorporating future revenues as indicated in Eq. (24) (see Section V). However, the CVP method executes exploration by choosing diverse prices without considering their profitability. Consequently, the proposed look-ahead approach outperforms the CVP method in terms of revenue gain.
- Figure 1 and Figure 3 represent two examples of a synthetic and a real dataset, respectively, in the high error setting $\sigma = 40\%$. These two figures show the behavior of cumulative discounted revenue of different pricing methods over the time horizon T. It can be observed from these figures that the look-ahead methods are the top-performing methods for synthetic and real datasets. These results agree with the average results presented in Table 2 and Table 6.
- Concerning the demand model estimation error, Table 3 and Table 7 demonstrate that the random-myopic two-phase benchmark obtains the most accurate model estimates in case of both synthetic and real datasets. The

random-myopic pricing policy yields accurate parameter estimates since it devotes a whole long phase for exploration. However, this pricing policy compromises the gained revenue as indicated in Table 2 and Table 6.

- The proposed look-ahead pricing policies are comparable to the CVP method with respect to demand model estimation as presented in Table 3 and Table 7, for synthetic and real datasets, respectively. In particular, the look-ahead methods outperform the CVP method in case of real datasets, and the CVP method leads to less estimation error for synthetic datasets. Both of the look-ahead and the CVP methods outperform the myopic pricing. Accordingly, the proposed look-ahead approach boosts the revenue gain which is basically the main objective of pricing, without sacrificing demand model estimation.
- In addition to revenue gain and model estimation error, we evaluate the final price deviation from the optimal price. The random-myopic benchmark yields the least price deviation as shown in Table 4 and Table 8. These results are intuitive since random-myopic is the most accurate method in demand model estimation due to its long exploration phase. Therefore, its final price is close to the optimal price since it obtains accurate estimates of the demand model parameters. However, this is achieved at the expense of the gained revenue. As presented in Table 2 and Table 6, the random-myopic benchmark is the worst performing method in terms of the achieved revenue.
- The chosen prices by different pricing strategies over the time horizon T are shown in Figure 2 and Figure 4, for two examples of a synthetic and a real dataset, respectively. As previously illustrated, the random-myopic benchmark achieves the best convergence to the optimal price, at the expense of the gained revenue. Figure 2 and Figure 4 indicate that the two-step look-ahead method outperforms the remaining pricing strategies.
- One can conclude from Table 3 and Table 7 that the CVP method and the look-ahead methods achieve comparable performance in terms of model estimation error. Similarly, the look-ahead and the CVP methods have the same relative ranks with respect to price deviation from the optimal price. They both have comparable results for synthetic and real datasets as presented in Table 4 and Table 8, respectively.

VIII. CONCLUSION

Dynamic pricing with unknown demand function is a challenging problem. Firms seek to set prices that maximize their revenues. However, the demand-price relation is originally not known and it should be learned from the data as the firm continually carries out sales transactions. It is important to have estimates of the demand-price relation as accurate as possible, otherwise the revenues can be significantly impacted. In this work, we propose a look-ahead pricing approach for revenue maximization in case of unknown demand. The proposed approach considers not only the immediate revenue but also the revenues of future steps. Furthermore, the presented look-ahead approach automatically incorporates exploration by considering less myopic profitable prices to enhance future revenues. We implement two variants of the proposed approach: one-step and two-step look-ahead methods. We compare the proposed look-ahead methods to different benchmarks and popular methods in literature with respect to several performance metrics such as: revenue gain, model estimation accuracy, and price convergence to the optimal price. We conduct experiments applying our methods, in addition to the benchmark pricing policies, using different twenty synthetic datasets with different parameters and error settings, and five different real datasets. The experiments demonstrate a considerable improvement of the proposed look-ahead pricing methods in terms of the gained revenue over the other methods. For demand model estimation, the look-ahead approach achieves comparable performance to other pricing strategies. In addition, the proposed look-ahead pricing formulation is simple, easy to analyze and implement, and computationally efficient. The proposed approach mainly handles the case of the linear demand price curve, but it would also be beneficial to extend this work to the case of non-linear demand functions such as negative exponential and the power function.

IX. FUTURE WORK

We believe that the multi-step ahead approach is a promising methodology for the dynamic pricing problem, and that it can be developed further. Thus, there are several potential research directions for extending the proposed methods. For example, one can incorporate other factors that could affect the demand including: seasonality, competition, etc. Furthermore, one can explore different problem settings such as: having a finite inventory or pricing multiple products. Finally, we could investigate the density estimates of the quantities in question [70], in order to take into account non-Gaussian situations. By progressively adding more factors, the model can closely mimic real-world applications. In order to tackle these factors, perhaps the approach should involve some aspect of Monte Carlo simulation, so as to preserve the flexibility of the modeling.

APPENDIX
DETAILED FORMULATIONS
A. ONE-STEP LOOK-AHEAD FORMULATION

As presented in Section V-B, the expected one-step look-ahead revenue is evaluated as:

$$\begin{aligned}\mathbb{E}[R(\mathcal{D}_n)|\mathcal{D}_{n-2}, x_{n-1}] &= R(\mathcal{D}_{n-2}) + p_{n-1}\mathbb{E}[y_{n-1}|x_{n-1}, \mathcal{D}_{n-2}] \\ &\quad + \gamma_r \int_{y_{n-1}} \frac{-a^2{}_{n-1}}{4b_{n-1}} Pr[y_{n-1}|\mathcal{D}_{n-2}, x_{n-1}]\mathrm{d}y_{n-1} \end{aligned} \quad (43)$$

In order to express the demand model parameters a_{n-1} and b_{n-1} in Eq. (43) in terms of β_{n-2}, p_{n-1}, and y_{n-1}, i.e. the

available information so far using $\mathcal{D}_{n-2}$, substitute for β_{n-1} from Eq. (7) as described in Section IV.

Accordingly, the demand model parameters a_{n-1} and b_{n-1} are evaluated using Eq. (44) and Eq. (45), respectively:

$$a_{n-1} = a_{n-2} + (\sigma^2_{a_{n-2}} + \sigma_{ab_{n-2}} p_{n-1}) \times \left(\frac{y_{n-1} - x^T{}_{n-1}\beta_{n-2}}{\gamma\sigma^2_{n-2} + x^T_{n-1}\Sigma_{\beta_{n-2}} x_{n-1}} \right) \tag{44}$$

$$b_{n-1} = b_{n-2} + (\sigma_{ab_{n-2}} + \sigma^2_{b_{n-2}} p_{n-1}) \times \left(\frac{y_{n-1} - x^T{}_{n-1}\beta_{n-2}}{\gamma\sigma^2_{n-2} + x^T_{n-1}\Sigma_{\beta_{n-2}} x_{n-1}} \right) \tag{45}$$

Let the covariance matrix $\Sigma_{\beta_{n-2}}$ be:

$$\Sigma_{\beta_{n-2}} = \begin{pmatrix} \sigma^2{}_{a_{n-2}} & \sigma_{ab_{n-2}} \\ \sigma_{ab_{n-2}} & \sigma^2{}_{b_{n-2}} \end{pmatrix} \tag{46}$$

Using Eq. (46) and the definition of x_{n-1} as $x_{n-1} = [1 \;\; p_{n-1}]^T$, the demand model parameters a_{n-1} and b_{n-1} are expressed in terms of the previous estimates as the two following equations:

$$a_{n-1} = a_{n-2} + (\sigma^2_{a_{n-2}} + \sigma_{ab_{n-2}} p_{n-1}) \times \left(\frac{y_{n-1} - a_{n-2} + b_{n-2} p_{n-1}}{\gamma\sigma^2_{n-2} + \sigma^2{}_{a_{n-2}} + 2\sigma_{ab_{n-2}} p_{n-1} + \sigma^2_{b_{n-2}} p^2_{n-1}} \right) \tag{47}$$

$$b_{n-1} = b_{n-2} + (\sigma_{ab_{n-2}} + \sigma^2_{b_{n-2}} p_{n-1}) \times \left(\frac{y_{n-1} - a_{n-2} + b_{n-2} p_{n-1}}{\gamma\sigma^2_{n-2} + \sigma^2{}_{a_{n-2}} + 2\sigma_{ab_{n-2}} p_{n-1} + \sigma^2_{b_{n-2}} p^2_{n-1}} \right) \tag{48}$$

Then, evaluating the expectation $\mathbb{E}[y_{n-1}|x_{n-1}, \mathcal{D}_{n-2}]$ from Eq. (14), and substituting for a_{n-1} and b_{n-1} using Eq. (47) and Eq. (48), respectively, Eq. (43) turns to Eq. (49) as follows:

$$\begin{aligned} &\mathbb{E}[R(\mathcal{D}_n)|\mathcal{D}_{n-2}, x_{n-1}] \\ &= R(\mathcal{D}_{n-2}) + a_{n-2} p_{n-1} + b_{n-2} p^2{}_{n-1} \\ &\quad + \gamma_r \int_{y_{n-1}} Pr[y_{n-1}|\mathcal{D}_{n-2}, x_{n-1}] \\ &\quad \times \left(\frac{-\left(a_{n-2} + (\sigma^2_{a_{n-2}} + \sigma_{ab_{n-2}} p_{n-1})\right)}{4\left(b_{n-2} + (\sigma_{ab_{n-2}} + \sigma^2_{b_{n-2}} p_{n-1})\right)} \right) \\ &\quad \times \left(\frac{\left(\frac{y_{n-1} - a_{n-2} + b_{n-2} p_{n-1}}{\gamma\sigma^2_{n-2} + \sigma^2{}_{a_{n-2}} + 2\sigma_{ab_{n-2}} p_{n-1} + \sigma^2_{b_{n-2}} p^2_{n-1}} \right)^2}{\left(\frac{y_{n-1} - a_{n-2} + b_{n-2} p_{n-1}}{\gamma\sigma^2_{n-2} + \sigma^2{}_{a_{n-2}} + 2\sigma_{ab_{n-2}} p_{n-1} + \sigma^2_{b_{n-2}} p^2_{n-1}} \right)} \right) \mathrm{d}y_{n-1} \end{aligned} \tag{49}$$

Simplifying Eq. (49) results in Eq. (50), as shown at the bottom of the next page, where $z(p_{n-1})$ is defined as:

$$z(p_{n-1}) = \gamma\sigma^2_{n-2} + \sigma^2{}_{a_{n-2}} + 2\sigma_{ab_{n-2}} p_{n-1} + \sigma^2_{b_{n-2}} p^2_{n-1} \tag{51}$$

Rearranging Eq. (50) according to the integration variable y_{n-1} results in the following equation (Eq. (52)), as shown at the bottom of the next page.

Substituting for the probability distribution of y_{n-1}, $Pr[y_{n-1}|\mathcal{D}_{n-2}, x_{n-1}]$ from Eq. (13) results in Eq. (53) as shown at the bottom of the next page, where the expectation $\mu_{y_{n-1}}$ and the variance $\sigma^2_{y_{n-1}}$ are evaluated using Eq. (14) and Eq. (15) as presented in Section IV.

Thus, the expectation of y_{n-1} is evaluated as:

$$\mu_{y_{n-1}}(p_{n-1}) = x_{n-1}{}^T \beta_{n-2} = a_{n-2} + b_{n-2} p_{n-1} \tag{54}$$

The variance of y_{n-1} is defined as:

$$\sigma^2_{y_{n-1}}(p_{n-1}) = x_{n-1}{}^T \Sigma_{\beta_{n-2}} x_{n-1} + \sigma^2{}_{n-2} \tag{55}$$

which can be expressed as:

$$\sigma^2_{y_{n-1}}(p_{n-1}) = \sigma^2_{n-2} + \sigma^2{}_{a_{n-2}} + 2\sigma_{ab_{n-2}} p_{n-1} + \sigma^2_{b_{n-2}} p^2_{n-1} \tag{56}$$

Then, the integration in the last term of Eq. (53) can be formulated as:

$$\begin{aligned} &\mathbb{E}[R(\mathcal{D}_n)|\mathcal{D}_{n-2}, x_{n-1}] \\ &\quad = R(\mathcal{D}_{n-2}) + a_{n-2} p_{n-1} + b_{n-2} p^2{}_{n-1} \\ &\qquad + \gamma_r \int_{y_{n-1}} Pr[y_{n-1}|\mathcal{D}_{n-2}, x_{n-1}] \\ &\qquad \times \frac{-\big(A(p_{n-1}) y_{n-1} + B(p_{n-1})\big)^2}{4\big(C(p_{n-1}) y_{n-1} + D(p_{n-1})\big)} \mathrm{d}y_{n-1} \end{aligned} \tag{57}$$

where:

$$\begin{aligned} A(p_{n-1}) &= \sigma^2_{a_{n-2}} + \sigma_{ab_{n-2}} p_{n-1} \\ B(p_{n-1}) &= (\sigma^2_{a_{n-2}} + \sigma_{ab_{n-2}} p_{n-1})(b_{n-2} p_{n-1} - a_{n-2}) \\ &\quad + a_{n-2}\big(\gamma\sigma^2_{n-2} + \sigma^2{}_{a_{n-2}} + 2\sigma_{ab_{n-2}} p_{n-1} \\ &\qquad + \sigma^2_{b_{n-2}} p^2_{n-1}\big) \end{aligned}$$

This can be expressed as a second order polynomial of p_{n-1} as follows:

$$\begin{aligned} B(p_{n-1}) &= \big(b_{n-2}\sigma_{ab_{n-2}} + a_{n-2}\sigma^2_{b_{n-2}}\big) p^2_{n-1} \\ &\quad + \big(a_{n-2}\sigma_{ab_{n-2}} + b_{n-2}\sigma^2_{a_{n-2}}\big) p_{n-1} + \gamma a_{n-2}\sigma^2_{n-2} \\ C(p_{n-1}) &= (\sigma_{ab_{n-2}} + \sigma^2_{b_{n-2}} p_{n-1}) \\ &\quad \times \big(\gamma\sigma^2_{n-2} + \sigma^2{}_{a_{n-2}} + 2\sigma_{ab_{n-2}} p_{n-1} + \sigma^2_{b_{n-2}} p^2_{n-1}\big) \end{aligned}$$

Then, $C(p_{n-1})$ could be further simplified as a third order polynomial of p_{n-1}:

$$\begin{aligned} &C(p_{n-1}) \\ &= \sigma^4_{b_{n-2}} p^3_{n-1} + 3\sigma_{ab_{n-2}}\sigma^2_{b_{n-2}} p^2_{n-1} \\ &\quad + \big(2\sigma^2{}_{ab_{n-2}} + \sigma^2_{b_{n-2}}(\gamma\sigma^2_{n-2} + \sigma^2{}_{a_{n-2}})\big) p_{n-1} \\ &\quad + \sigma_{ab_{n-2}}(\gamma\sigma^2_{n-2} + \sigma^2{}_{a_{n-2}}) \\ &D(p_{n-1}) \\ &= \big(\gamma\sigma^2_{n-2} + \sigma^2{}_{a_{n-2}} + 2\sigma_{ab_{n-2}} p_{n-1} \\ &\quad + \sigma^2_{b_{n-2}} p^2_{n-1}\big)\Big((\sigma_{ab_{n-2}} + \sigma^2_{b_{n-2}} p_{n-1})(b_{n-2} p_{n-1} - a_{n-2}) \\ &\quad + b_{n-2}\big(\gamma\sigma^2_{n-2} + \sigma^2{}_{a_{n-2}} + 2\sigma_{ab_{n-2}} p_{n-1} + \sigma^2_{b_{n-2}} p^2_{n-1}\big)\Big) \end{aligned}$$

This can be expressed as a fourth order polynomial of p_{n-1} as follows:

$$\begin{aligned}
&D(p_{n-1}) \\
&= 2b_{n-2}{\sigma^4}_{b_{n-2}}p_{n-1}^4 + \big(7b_{n-2}\sigma^2_{b_{n-2}}\sigma_{ab_{n-2}} \\
&\quad - a_{n-2}\sigma^4_{b_{n-2}}\big)p_{n-1}^3 + \big(4b_{n-2}\sigma^2_{ab_{n-2}} + 3b_{n-2}\sigma^2_{b_{n-2}}(\gamma\sigma^2_{n-2} \\
&\quad + {\sigma^2}_{a_{n-2}}) + 2b_{n-2}{\sigma^2}_{ab_{n-2}} - 3a_{n-2}\sigma_{ab_{n-2}}\sigma^2_{b_{n-2}}\big)p_{n-1}^2 \\
&\quad + \big((5b_{n-2}\sigma_{ab_{n-2}} - a_{n-2}\sigma^2_{b_{n-2}})(\gamma\sigma^2_{n-2} + {\sigma^2}_{a_{n-2}}) \\
&\quad - 2a_{n-2}\sigma^2_{ab_{n-2}}\big)p_{n-1} + \big(b_{n-2}(\gamma\sigma^2_{n-2} + {\sigma^2}_{a_{n-2}}) \\
&\quad - a_{n-2}\sigma_{ab_{n-2}}\big) \times (\gamma\sigma^2_{n-2} + {\sigma^2}_{a_{n-2}})
\end{aligned}$$

Since the key objective of the one-step look-ahead formulation is to choose the pricing point p_{n-1} maximizing the expected revenue at time step n. Using Eq. (57), the underlying optimization objective can be formulated as:

$$\begin{aligned}
{p^*}_{n-1} &= argmax_{p_{n-1}}\, R(\mathcal{D}_{n-2}) + a_{n-2}p_{n-1} + b_{n-2}{p^2}_{n-1} \\
&\quad + \gamma_r \int_{y_{n-1}} \mathcal{N}(y_{n-1};\, \mu_{y_{n-1}(p_{n-1})}, \sigma^2_{y_{n-1}}(p_{n-1})) \\
&\quad \times \frac{-\big(A(p_{n-1})y_{n-1} + B(p_{n-1})\big)^2}{4\big(C(p_{n-1})y_{n-1} + D(p_{n-1})\big)}\mathrm{d}y_{n-1}
\end{aligned}$$

However, the term $R(\mathcal{D}_{n-2})$ is constant, then:

$$\begin{aligned}
{p^*}_{n-1} &= argmax_{p_{n-1}}\, a_{n-2}p_{n-1} + b_{n-2}{p^2}_{n-1} \\
&\quad + \gamma_r \int_{y_{n-1}} \mathcal{N}(y_{n-1};\, \mu_{y_{n-1}(p_{n-1})}, \sigma^2_{y_{n-1}}(p_{n-1})) \\
&\quad \times \frac{-\big(A(p_{n-1})y_{n-1} + B(p_{n-1})\big)^2}{4\big(C(p_{n-1})y_{n-1} + D(p_{n-1})\big)}\mathrm{d}y_{n-1}
\end{aligned}$$

Assume price bounds: p_l and p_u such that any potential price should belong to the range $[p_l, p_u]$.

Then, the problem can be formulated as a constrained optimization problem as follows:

$$\begin{aligned}
{p^*}_{n-1} &= argmax_{p_{n-1},\, p_l \le p_{n-1} < p_u}\quad a_{n-2}p_{n-1} + b_{n-2}{p^2}_{n-1} \\
&\quad + \gamma_r \int_{y_{n-1}} \mathcal{N}(y_{n-1};\, \mu_{y_{n-1}(p_{n-1})}, \sigma^2_{y_{n-1}}(p_{n-1})) \\
&\quad \times \frac{-\big(A(p_{n-1})y_{n-1} + B(p_{n-1})\big)^2}{4\big(C(p_{n-1})y_{n-1} + D(p_{n-1})\big)}\mathrm{d}y_{n-1}
\end{aligned}$$

B. TWO-STEPS LOOK-AHEAD FORMULATION

As presented in Section V-B, the two-step look-ahead revenue is formulated as:

$$\begin{aligned}
&\mathbb{E}[R(\mathcal{D}_{n+1})|x_{n-1}, \mathcal{D}_{n-2}] \\
&\quad = \mathbb{E}[R(\mathcal{D}_n)|x_{n-1}, \mathcal{D}_{n-2}] \\
&\qquad + {\gamma_r}^2 \int_{y_n} p_{n+1} Pr[y_n|x_n, \mathcal{D}_{n-1}] \\
&\qquad \times \mathbb{E}[y_{n+1}|x_{n+1}, x_n, y_n, \mathcal{D}_{n-1}]\,\mathrm{d}y_n
\end{aligned} \tag{58}$$

Substituting for the first term of the above equation from Eq. (29), and using the dataset D_n definition, then:

$$\begin{aligned}
&\mathbb{E}[R(\mathcal{D}_{n+1})|x_{n-1}, \mathcal{D}_{n-2}] \\
&= R(\mathcal{D}_{n-2}) + a_{n-2}p_{n-1} \\
&\quad + b_{n-2}{p^2}_{n-1} + \gamma_r \int_{y_{n-1}} \mathcal{N}(y_{n-1};\, \mu_{y_{n-1}(p_{n-1})}, \sigma^2_{y_{n-1}}(p_{n-1})) \\
&\quad \times \frac{-\big(A(p_{n-1})y_{n-1} + B(p_{n-1})\big)^2}{4\big(C(p_{n-1})y_{n-1} + D(p_{n-1})\big)}\mathrm{d}y_{n-1} \\
&\quad + {\gamma_r}^2 \int_{y_n} p_{n+1} Pr[y_n|x_n, \mathcal{D}_{n-1}]\mathbb{E}[y_{n+1}|x_{n+1}, \mathcal{D}_n]\,\mathrm{d}y_n
\end{aligned} \tag{59}$$

$$\begin{aligned}
&\mathbb{E}[R(\mathcal{D}_n)|\mathcal{D}_{n-2}, x_{n-1}] \\
&\quad = R(\mathcal{D}_{n-2}) + a_{n-2}p_{n-1} + b_{n-2}{p^2}_{n-1} \\
&\qquad + \gamma_r \int_{y_{n-1}} Pr[y_{n-1}|\mathcal{D}_{n-2}, x_{n-1}] \frac{-\Big(a_{n-2}z(p_{n-1}) + (\sigma^2_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})\big(y_{n-1} - a_{n-2} + b_{n-2}p_{n-1}\big)\Big)^2}{4z(p_{n-1})\Big(b_{n-2}z(p_{n-1}) + (\sigma_{ab_{n-2}} + \sigma^2_{b_{n-2}}p_{n-1})\big(y_{n-1} - a_{n-2} + b_{n-2}p_{n-1}\big)\Big)}\mathrm{d}y_{n-1}
\end{aligned} \tag{50}$$

$$\begin{aligned}
&\mathbb{E}[R(\mathcal{D}_n)|\mathcal{D}_{n-2}, x_{n-1}] \\
&\quad = R(\mathcal{D}_{n-2}) + a_{n-2}p_{n-1} + b_{n-2}{p^2}_{n-1} + \gamma_r \int_{y_{n-1}} Pr[y_{n-1}|\mathcal{D}_{n-2}, x_{n-1}] \\
&\qquad \times \frac{-\Big((\sigma^2_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})y_{n-1} + (\sigma^2_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})(b_{n-2}p_{n-1} - a_{n-2}) + a_{n-2}z(p_{n-1})\Big)^2}{4z(p_{n-1})\Big((\sigma_{ab_{n-2}} + \sigma^2_{b_{n-2}}p_{n-1})y_{n-1} + (\sigma_{ab_{n-2}} + \sigma^2_{b_{n-2}}p_{n-1})(b_{n-2}p_{n-1} - a_{n-2}) + b_{n-2}z(p_{n-1})\Big)}\mathrm{d}y_{n-1}
\end{aligned} \tag{52}$$

$$\begin{aligned}
&\mathbb{E}[R(\mathcal{D}_n)|\mathcal{D}_{n-2}, x_{n-1}] \\
&\quad = R(\mathcal{D}_{n-2}) + a_{n-2}p_{n-1} + b_{n-2}{p^2}_{n-1} + \gamma_r \int_{y_{n-1}} \mathcal{N}(y_{n-1};\, \mu_{y_{n-1}(p_{n-1})}, \sigma^2_{y_{n-1}}(p_{n-1})) \\
&\qquad \times \frac{-\Big((\sigma^2_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})y_{n-1} + (\sigma^2_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})(b_{n-2}p_{n-1} - a_{n-2}) + a_{n-2}z(p_{n-1})\Big)^2}{4z(p_{n-1})\Big((\sigma_{ab_{n-2}} + \sigma^2_{b_{n-2}}p_{n-1})y_{n-1} + (\sigma_{ab_{n-2}} + \sigma^2_{b_{n-2}}p_{n-1})(b_{n-2}p_{n-1} - a_{n-2}) + b_{n-2}z(p_{n-1})\Big)}\mathrm{d}y_{n-1}
\end{aligned} \tag{53}$$

The probability $Pr[y_n|x_n, \mathcal{D}_{n-1}]$ is evaluated using Eq. (13), and the expectation $\mathbb{E}[y_{n+1}|x_{n+1}, \mathcal{D}_n]$ is evaluated using Eq. (14). Thus, Eq. (59) can be formulated as:

$$\begin{aligned}
&\mathbb{E}[R(\mathcal{D}_{n+1})|x_{n-1}, \mathcal{D}_{n-2}] \\
&= R(\mathcal{D}_{n-2}) + a_{n-2}p_{n-1} \\
&\quad + b_{n-2}{p^2}_{n-1} + \gamma_r \int_{y_{n-1}} \mathcal{N}(y_{n-1}; \mu_{y_{n-1}(p_{n-1})}, \sigma^2_{y_{n-1}}(p_{n-1})) \\
&\quad \times \frac{-\big(A(p_{n-1})y_{n-1} + B(p_{n-1})\big)^2}{4\big(C(p_{n-1})y_{n-1} + D(p_{n-1})\big)} \mathrm{d}y_{n-1} \\
&\quad + {\gamma_r}^2 \int_{y_n} (a_n p_{n+1} + b_n {p^2}_{n+1}) \\
&\quad \times \mathcal{N}(y_n; \mu_{y_n}(p_n), {\sigma^2}_{y_n}(p_n))\, \mathrm{d}y_n
\end{aligned} \tag{60}$$

Similar to Eq. (54) and Eq. (56), the expectation μ_{y_n} and the variance ${\sigma^2}_{y_n}$ of y_n, are evaluated as follows:

$$\mu_{y_n}(p_n) = a_{n-1} + b_{n-1}p_n \tag{61}$$

$$\sigma^2_{y_n}(p_n) = \sigma^2_{n-1} + {\sigma^2}_{a_{n-1}} + 2\sigma_{ab_{n-1}}p_n + \sigma^2_{b_{n-1}}p_n^2 \tag{62}$$

It can be observed from Eq. (60), Eq. (61) and Eq. (62) that the future revenue at step $n+1$ is a function of β_{n-1}, $\Sigma_{\beta_{n-1}}$, and σ_{n-1}. Since the true value of y_{n-1} is not known, we assume that the model's parameter vector β_{n-1} is approximately equal to β_{n-2}. Similarly, we assume that the standard deviation σ_{n-1} of the error term is approximately equal to σ_{n-2}. However, we can evaluate the covariance matrix of model parameters $\Sigma_{\beta_{n-1}}$ since it is not a function of y_{n-1} as indicated in Eq. (8). Consequently, the elements of the covariance matrix $\Sigma_{\beta_{n-1}}$ are evaluated in terms of $\Sigma_{\beta_{n-2}}$ elements as follows:

$${\sigma^2}_{a_{n-1}} = \frac{1}{\gamma}{\sigma^2}_{a_{n-2}} - \frac{1}{\gamma z(p_{n-1})}({\sigma^2}_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})^2 \tag{63}$$

$${\sigma^2}_{b_{n-1}} = \frac{1}{\gamma}{\sigma^2}_{b_{n-2}} - \frac{1}{\gamma z(p_{n-1})}(\sigma_{ab_{n-2}} + {\sigma^2}_{b_{n-2}}p_{n-1})^2 \tag{64}$$

$$\begin{aligned}
\sigma_{ab_{n-1}} &= \frac{1}{\gamma}\sigma_{ab_{n-2}} - \frac{1}{\gamma z(p_{n-1})}({\sigma^2}_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1}) \\
&\quad \times (\sigma_{ab_{n-2}} + {\sigma^2}_{b_{n-2}}p_{n-1})
\end{aligned} \tag{65}$$

Then, substituting from Eq. (63), Eq. (64), and Eq. (65) into Eq. (62) results in:

$$\begin{aligned}
&\sigma^2_{y_n}(p_n) \\
&= \sigma^2_{n-2} + \frac{1}{\gamma}{\sigma^2}_{a_{n-2}} - \frac{1}{\gamma z(p_{n-1})}({\sigma^2}_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})^2 \\
&\quad + 2\Big(\frac{1}{\gamma}\sigma_{ab_{n-2}} - \frac{1}{\gamma z(p_{n-1})}({\sigma^2}_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1}) \\
&\qquad \times (\sigma_{ab_{n-2}} + {\sigma^2}_{b_{n-2}}p_{n-1})\Big)p_n \\
&\quad + \Big(\frac{1}{\gamma}{\sigma^2}_{b_{n-2}} - \frac{1}{\gamma z(p_{n-1})}(\sigma_{ab_{n-2}} + {\sigma^2}_{b_{n-2}}p_{n-1})^2\Big)p_n^2
\end{aligned} \tag{66}$$

Similar to the one-step look-ahead formulation, the price p_{n+1} is set so as to maximize the immediate revenue at step $n+1$. Accordingly, the price p_{n+1} is evaluated in terms of β_n as follows:

$$p_{n+1} = \frac{-a_n}{2b_n} \tag{67}$$

Substituting from Eq. (67) into Eq. (60):

$$\begin{aligned}
&\mathbb{E}[R(\mathcal{D}_{n+1})|x_{n-1}, \mathcal{D}_{n-2}] \\
&= R(\mathcal{D}_{n-2}) + a_{n-2}p_{n-1} + b_{n-2}{p^2}_{n-1} \\
&\quad + \gamma_r \int_{y_{n-1}} \mathcal{N}(y_{n-1}; \mu_{y_{n-1}(p_{n-1})}, \sigma^2_{y_{n-1}}(p_{n-1})) \\
&\quad \times \frac{-\big(A(p_{n-1})y_{n-1} + B(p_{n-1})\big)^2}{4\big(C(p_{n-1})y_{n-1} + D(p_{n-1})\big)} \mathrm{d}y_{n-1} \\
&\quad + {\gamma_r}^2 \int_{y_n} \frac{-{a^2}_n}{4b_n} \mathcal{N}(y_n; \mu_{y_n}(p_n), {\sigma^2}_{y_n}(p_n))\, \mathrm{d}y_n
\end{aligned} \tag{68}$$

The model parameters a_n and b_n are evaluated recursively using a_{n-1} and b_{n-1} according to Eq. (44) and Eq. (45), respectively.

Then, similar to the one-step look-ahead case, substituting for the immediate revenue at time step $n+1$, $\frac{-{a^2}_n}{4b_n}$ in the third term of Eq. (68), the two-step look-ahead revenue can be expressed in terms of p_n as follows:

$$\begin{aligned}
&\mathbb{E}[R(\mathcal{D}_{n+1})|x_{n-1}, \mathcal{D}_{n-2}] \\
&= R(\mathcal{D}_{n-2}) + a_{n-2}p_{n-1} + b_{n-2}{p^2}_{n-1} \\
&\quad + \gamma_r \int_{y_{n-1}} \mathcal{N}(y_{n-1}; \mu_{y_{n-1}(p_{n-1})}, \sigma^2_{y_{n-1}}(p_{n-1})) \\
&\quad \times \frac{-\big(A(p_{n-1})y_{n-1} + B(p_{n-1})\big)^2}{4\big(C(p_{n-1})y_{n-1} + D(p_{n-1})\big)} \mathrm{d}y_{n-1} \\
&\quad + {\gamma_r}^2 \int_{y_n} \mathcal{N}(y_n; \mu_{y_n}, {\sigma^2}_{y_n}) \frac{-\big(J(p_n)y_n + F(p_n)\big)^2}{4\big(G(p_n)y_n + H(p_n)\big)}\, \mathrm{d}y_n
\end{aligned} \tag{69}$$

where the four polynomials: $J(p_n)$, $F(p_n)$, $G(p_n)$, and $H(p_n)$ are similar to the four polynomials used to evaluate the one-step look-ahead revenue. However,the four polynomials: $J(p_n)$, $F(p_n)$, $G(p_n)$, and $H(p_n)$ are evaluated in terms of p_n, β_{n-2}, $\Sigma_{\beta_{n-1}}$, and σ_{n-2}. The equations below represent the four polynomials first as functions of $\Sigma_{\beta_{n-1}}$, then as functions of $\Sigma_{\beta_{n-2}}$ after substituting from Eq. (63), Eq. (64), and Eq. (65).

$$\begin{aligned}
&J(p_n) \\
&= \sigma^2_{a_{n-1}} + \sigma_{ab_{n-1}}p_n
\end{aligned} \tag{70}$$

$$\begin{aligned}
&J(p_n) \\
&= \frac{1}{\gamma}{\sigma^2}_{a_{n-2}} - \frac{1}{\gamma z(p_{n-1})}({\sigma^2}_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})^2 \\
&\quad + \Big(\frac{1}{\gamma}\sigma_{ab_{n-2}} - \frac{1}{\gamma z(p_{n-1})}({\sigma^2}_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1}) \\
&\quad \times (\sigma_{ab_{n-2}} + {\sigma^2}_{b_{n-2}}p_{n-1})\Big)p_n
\end{aligned} \tag{71}$$

$$
\begin{aligned}
&F(p_n)\\
&= \left(b_{n-1}\sigma_{ab_{n-1}} + a_{n-1}\sigma^2_{b_{n-1}}\right)p_n^2\\
&\quad + \left(a_{n-2}\sigma_{ab_{n-1}} + b_{n-1}\sigma^2_{a_{n-1}}\right)p_n + \gamma a_{n-1}\sigma^2_{n-1} \qquad (72)\\
&F(p_n)\\
&= \Big(b_{n-2}\Big(\frac{1}{\gamma}\sigma_{ab_{n-2}} - \frac{1}{\gamma z(p_{n-1})}(\sigma^2{}_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})\\
&\qquad \times (\sigma_{ab_{n-2}} + \sigma^2{}_{b_{n-2}}p_{n-1})\Big)p_n^2\\
&\quad + a_{n-2}\Big(\frac{1}{\gamma}\sigma_{ab_{n-2}} - \frac{1}{\gamma z(p_{n-1})}(\sigma^2{}_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})\\
&\qquad \times (\sigma_{ab_{n-2}} + \sigma^2{}_{b_{n-2}}p_{n-1})\Big)p^2{}_n\\
&\quad + a_{n-2}\Big(\frac{1}{\gamma}\sigma_{ab_{n-2}} - \frac{1}{\gamma z(p_{n-1})}(\sigma^2{}_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})\\
&\qquad \times (\sigma_{ab_{n-2}} + \sigma^2{}_{b_{n-2}}p_{n-1})\Big)p_n\\
&\quad + b_{n-2}\Big(\frac{1}{\gamma}\sigma^2{}_{a_{n-2}} - \frac{1}{\gamma z(p_{n-1})}(\sigma^2{}_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})^2\Big)p_n\\
&\quad + \gamma a_{n-2}\sigma^2_{n-2} \qquad (73)\\
&G(p_n)\\
&= \sigma^4_{b_{n-1}}p_n^3 + 3\sigma_{ab_{n-1}}\sigma^2_{b_{n-1}}p_n^2\\
&\quad + \left(2\sigma^2{}_{ab_{n-1}} + \sigma^2_{b_{n-1}}(\gamma\sigma^2_{n-1} + \sigma^2{}_{a_{n-1}})\right)p_n\\
&\qquad + \sigma_{ab_{n-1}}(\gamma\sigma^2_{n-1} + \sigma^2{}_{a_{n-1}}) \qquad (74)\\
&G(p_n)\\
&= \Big(\frac{1}{\gamma}\sigma_{ab_{n-2}} - \frac{1}{\gamma z(p_{n-1})}(\sigma^2{}_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})\\
&\qquad \times (\sigma_{ab_{n-2}} + \sigma^2{}_{b_{n-2}}p_{n-1})\Big)^2 p_n^3\\
&\quad + 3\Big[\Big(\frac{1}{\gamma}\sigma_{ab_{n-2}} - \frac{1}{\gamma z(p_{n-1})}(\sigma^2{}_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})\\
&\qquad \times (\sigma_{ab_{n-2}} + \sigma^2{}_{b_{n-2}}p_{n-1})\Big)\\
&\qquad \times \Big(\frac{1}{\gamma}\sigma^2{}_{b_{n-2}} - \frac{1}{\gamma z(p_{n-1})}(\sigma_{ab_{n-2}} + \sigma^2{}_{b_{n-2}}p_{n-1})^2\Big)\Big]\\
&\quad \times p_n^2 + 2\Big(\frac{1}{\gamma}\sigma_{ab_{n-2}} - \frac{1}{\gamma z(p_{n-1})}(\sigma^2{}_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})\\
&\quad \times (\sigma_{ab_{n-2}} + \sigma^2{}_{b_{n-2}}p_{n-1})\Big)^2 p_n\\
&\quad + \Big(\frac{1}{\gamma}\sigma^2{}_{b_{n-2}} - \frac{1}{\gamma z(p_{n-1})}(\sigma_{ab_{n-2}} + \sigma^2{}_{b_{n-2}}p_{n-1})^2\Big)\\
&\quad \times \Big(\gamma\sigma^2_{n-2} + \frac{1}{\gamma}\sigma^2{}_{a_{n-2}} - \frac{1}{\gamma z(p_{n-1})}\\
&\qquad \times (\sigma^2{}_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})^2\Big)p_n\\
&\quad + \Big(\frac{1}{\gamma}\sigma_{ab_{n-2}} - \frac{1}{\gamma z(p_{n-1})}(\sigma^2{}_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})\\
&\quad \times (\sigma_{ab_{n-2}} + \sigma^2{}_{b_{n-2}}p_{n-1})\Big)\\
&\quad \times \Big(\gamma\sigma^2_{n-2} + \Big(\frac{1}{\gamma}\sigma^2{}_{a_{n-2}} - \frac{1}{\gamma z(p_{n-1})}\\
&\qquad \times (\sigma^2{}_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})^2\Big)\Big) \qquad (75)
\end{aligned}
$$

$$
\begin{aligned}
&H(p_n)\\
&= 2b_{n-1}\sigma^4{}_{b_{n-1}}p_n^4 + \left(7b_{n-1}\sigma^2_{b_{n-1}}\sigma_{ab_{n-1}} - a_{n-1}\sigma^4_{b_{n-1}}\right)p_n^3\\
&\quad + \left(4b_{n-1}\sigma^2_{ab_{n-1}} + 3b_{n-1}\sigma^2_{b_{n-1}}(\gamma\sigma^2_{n-1} + \sigma^2{}_{a_{n-1}})\right.\\
&\quad \left. + 2b_{n-1}\sigma^2{}_{ab_{n-1}} - 3a_{n-1}\sigma_{ab_{n-1}}\sigma^2_{b_{n-1}}\right)p_n^2\\
&\quad + \left((5b_{n-1}\sigma_{ab_{n-1}} - a_{n-1}\sigma^2_{b_{n-1}})(\gamma\sigma^2_{n-1} + \sigma^2{}_{a_{n-1}})\right.\\
&\qquad \left. - 2a_{n-1}\sigma^2_{ab_{n-1}}\right)p_n\\
&\quad + \left(b_{n-1}(\gamma\sigma^2_{n-1} + \sigma^2{}_{a_{n-1}}) - a_{n-1}\sigma_{ab_{n-1}}\right)\\
&\quad \times (\gamma\sigma^2_{n-1} + \sigma^2{}_{a_{n-1}}) \qquad (76)\\
&H(p_n)\\
&= 2b_{n-2}\Big(\frac{1}{\gamma}\sigma^2{}_{b_{n-2}} - \frac{1}{\gamma z(p_{n-1})}\\
&\quad \times (\sigma_{ab_{n-2}} + \sigma^2{}_{b_{n-2}}p_{n-1})^2\Big)^2 p_n^4\\
&\quad + 7\Big[b_{n-2}\Big(\frac{1}{\gamma}\sigma^2{}_{b_{n-2}} - \frac{1}{\gamma z(p_{n-1})}(\sigma_{ab_{n-2}} + \sigma^2{}_{b_{n-2}}p_{n-1})^2\Big)\\
&\qquad \times \Big(\frac{1}{\gamma}\sigma_{ab_{n-2}} - \frac{1}{\gamma z(p_{n-1})}(\sigma^2{}_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})\\
&\qquad \times (\sigma_{ab_{n-2}} + \sigma^2{}_{b_{n-2}}p_{n-1})\Big)\Big]p_n^3\\
&\quad - a_{n-2}\Big(\frac{1}{\gamma}\sigma^2{}_{b_{n-2}} - \frac{1}{\gamma z(p_{n-1})}(\sigma_{ab_{n-2}} + \sigma^2{}_{b_{n-2}}p_{n-1})^2\Big)^2 p_n^3\\
&\quad + \Big(4b_{n-2}\Big(\frac{1}{\gamma}\sigma_{ab_{n-2}} - \frac{1}{\gamma z(p_{n-1})}(\sigma^2{}_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})\\
&\qquad \times (\sigma_{ab_{n-2}} + \sigma^2{}_{b_{n-2}}p_{n-1})\Big)^2 p_n^2\\
&\quad + 3b_{n-2}\Big(\frac{1}{\gamma}\sigma^2{}_{b_{n-2}} - \frac{1}{\gamma z(p_{n-1})}\\
&\qquad \times (\sigma_{ab_{n-2}} + \sigma^2{}_{b_{n-2}}p_{n-1})^2\Big)\\
&\quad \times \Big(\gamma\sigma^2_{n-2} + \Big(\frac{1}{\gamma}\sigma^2{}_{a_{n-2}} - \frac{1}{\gamma z(p_{n-1})}\\
&\qquad \times (\sigma^2{}_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})^2\Big)\Big)\\
&\quad \times p_n^2 + 2b_{n-2}\Big(\frac{1}{\gamma}\sigma_{ab_{n-2}} - \frac{1}{\gamma z(p_{n-1})}(\sigma^2{}_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})\\
&\qquad \times (\sigma_{ab_{n-2}} + \sigma^2{}_{b_{n-2}}p_{n-1})\Big)^2 p_n^2\\
&\quad - 3a_{n-2}\Big(\frac{1}{\gamma}\sigma_{ab_{n-2}} - \frac{1}{\gamma z(p_{n-1})}(\sigma^2{}_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})\\
&\qquad \times (\sigma_{ab_{n-2}} + \sigma^2{}_{b_{n-2}}p_{n-1})\Big)\\
&\quad \times \Big(\frac{1}{\gamma}\sigma^2{}_{b_{n-2}} - \frac{1}{\gamma z(p_{n-1})}(\sigma_{ab_{n-2}} + \sigma^2{}_{b_{n-2}}p_{n-1})^2\Big)p_n^2\\
&\quad + \Big[\Big(5b_{n-2}\Big(\frac{1}{\gamma}\sigma_{ab_{n-2}} - \frac{1}{\gamma z(p_{n-1})}(\sigma^2{}_{a_{n-2}} + \sigma_{ab_{n-2}}p_{n-1})\\
&\qquad \times (\sigma_{ab_{n-2}} + \sigma^2{}_{b_{n-2}}p_{n-1})\Big) - a_{n-2}\Big(\frac{1}{\gamma}\sigma^2{}_{b_{n-2}}\\
&\qquad - \frac{1}{\gamma z(p_{n-1})}(\sigma_{ab_{n-2}} + \sigma^2{}_{b_{n-2}}p_{n-1})^2\Big)\Big)\Big(\gamma\sigma^2_{n-2}
\end{aligned}
$$

$$+\frac{1}{\gamma}\sigma^2{}_{a_{n-2}}-\frac{1}{\gamma z(p_{n-1})}(\sigma^2{}_{a_{n-2}}+\sigma_{ab_{n-2}}p_{n-1})^2\Big)\Big]p_n$$
$$-2a_{n-2}\Big(\frac{1}{\gamma}\sigma_{ab_{n-2}}-\frac{1}{\gamma z(p_{n-1})}(\sigma^2{}_{a_{n-2}}+\sigma_{ab_{n-2}}p_{n-1})$$
$$\times(\sigma_{ab_{n-2}}+\sigma^2{}_{b_{n-2}}p_{n-1})\Big)^2p_n$$
$$+b_{n-2}\Big(\gamma\sigma^2_{n-2}+\Big(\frac{1}{\gamma}\sigma^2{}_{a_{n-2}}-\frac{1}{\gamma z(p_{n-1})}$$
$$\times(\sigma^2{}_{a_{n-2}}+\sigma_{ab_{n-2}}p_{n-1})\Big)^2\Big)^2-a_{n-2}\Big(\frac{1}{\gamma}\sigma_{ab_{n-2}}$$
$$-\frac{1}{\gamma z(p_{n-1})}(\sigma^2{}_{a_{n-2}}+\sigma_{ab_{n-2}}p_{n-1})(\sigma_{ab_{n-2}}+\sigma^2{}_{b_{n-2}}p_{n-1}))$$
$$\times\Big(\gamma\sigma^2_{n-2}+\Big(\frac{1}{\gamma}\sigma^2{}_{a_{n-2}}-\frac{1}{\gamma z(p_{n-1})}$$
$$\times(\sigma^2{}_{a_{n-2}}+\sigma_{ab_{n-2}}p_{n-1})^2\Big)\Big) \tag{77}$$

Substituting from Eq. (47), Eq. (48), and Eq. (51) into Eq. (26), the price p_n can be expressed in terms of p_{n-1} and β_{n-2} as follows:

$$p_n=\frac{M(p_{n-1})}{N(p_{n-1})} \tag{78}$$

where $M(p_{n-1})$ and $N(p_{n-1})$ are defined as follows:

$$M(p_{n-1})=(\sigma^2_{a_{n-2}}+\sigma_{ab_{n-2}}p_{n-1})y_{n-1}+(\sigma^2_{a_{n-2}}+\sigma_{ab_{n-2}}p_{n-1})$$
$$\times(b_{n-2}p_{n-1}-a_{n-2})+a_{n-2}z(p_{n-1}) \tag{79}$$
$$N(p_{n-1})=2z(p_{n-1})\Big((\sigma_{ab_{n-2}}+\sigma^2_{b_{n-2}}p_{n-1})y_{n-1}$$
$$+(\sigma_{ab_{n-2}}+\sigma^2_{b_{n-2}}p_{n-1})(b_{n-2}p_{n-1}-a_{n-2})$$
$$+b_{n-2}z(p_{n-1})\Big) \tag{80}$$

In Eq. (78), since the true value of y_{n-1} is not known, we use its expected value as defined in Eq. (14).

Accordingly, substituting from Eq. (78) into Eq. (69) results in:

$$\mathbb{E}[R(\mathcal{D}_{n+1})|x_{n-1},\mathcal{D}_{n-2}]$$
$$=R(\mathcal{D}_{n-2})+a_{n-2}p_{n-1}+b_{n-2}p^2{}_{n-1}$$
$$+\gamma_r\int_{y_{n-1}}\mathcal{N}(y_{n-1};\mu_{y_{n-1}(p_{n-1})},\sigma^2_{y_{n-1}}(p_{n-1}))$$
$$\times\frac{-\big(A(p_{n-1})y_{n-1}+B(p_{n-1})\big)^2}{4\big(C(p_{n-1})y_{n-1}+D(p_{n-1})\big)}\mathrm{d}y_{n-1}$$
$$+{\gamma_r}^2\int_{y_n}\mathcal{N}(y_n;\mu_{y_n}(W(p_{n-1})),\sigma^2{}_{y_n}(W(p_{n-1}))$$
$$\times\frac{-\Big(\big(J(W(p_{n-1}))y_n+F(W(p_{n-1}))\big)^2}{4\Big(G(W(p_{n-1}))y_n+H(W(p_{n-1}))\Big)}\mathrm{d}y_n \tag{81}$$

Since the first term of Eq. (81) is constant, it is safely ignored in optimization. Accordingly, the optimization problem of choosing the pricing point x_{n-1} maximizing the two-step look-ahead revenue is formulated as:

$$p^*{}_{n-1}$$
$$=argmax_{p_{n-1},\ p_l\le p_{n-1}<p_u}\quad a_{n-2}p_{n-1}+b_{n-2}p^2{}_{n-1}$$
$$+\gamma_r\int_{y_{n-1}}\mathcal{N}(y_{n-1};\mu_{y_{n-1}(p_{n-1})},\sigma^2_{y_{n-1}}(p_{n-1}))$$
$$\times\frac{-\big(A(p_{n-1})y_{n-1}+B(p_{n-1})\big)^2}{4\big(C(p_{n-1})y_{n-1}+D(p_{n-1})\big)}\mathrm{d}y_{n-1}$$
$$+{\gamma_r}^2\int_{y_n}\mathcal{N}\big(y_n;(\mu_{y_n}\circ W)(p_{n-1}),(\sigma^2{}_{y_n}\circ W)(p_{n-1})\big)$$
$$\times\frac{-\Big(\big(J\circ W)(p_{n-1})y_n+(F\circ W)(p_{n-1})\big)^2}{4\Big((G\circ W)(p_{n-1})y_n+(H\circ W)(p_{n-1})\Big)}\mathrm{d}y_n$$

REFERENCES

[1] R. Ganti, M. Sustik, Q. Tran, and B. Seaman, "Thompson sampling for dynamic pricing," 2018, *arXiv:1802.03050*. [Online]. Available: http://arxiv.org/abs/1802.03050

[2] M. Akan, B. Ata, and R. C. Savaşkan-Ebert, "Dynamic pricing of remanufacturable products under demand substitution: A product life cycle model," *Ann. Oper. Res.*, vol. 211, no. 1, pp. 1–25, Dec. 2013.

[3] M. N. Ibrahim and A. F. Atiya, "Analytical solutions to the dynamic pricing problem for time-normalized revenue," *Eur. J. Oper. Res.*, vol. 254, no. 2, pp. 632–643, Oct. 2016.

[4] A. E.-M. Bayoumi, M. Saleh, A. F. Atiya, and H. A. Aziz, "Dynamic pricing for hotel revenue management using price multipliers," *J. Revenue Pricing Manage.*, vol. 12, no. 3, pp. 271–285, May 2013.

[5] R. P. McAfee and V. Te Velde, "Dynamic pricing in the airline industry," in *Forthcoming in Handbook on Economics and Information Systems*, T. J. Hendershott, Ed. Amsterdam, The Netherlands: Elsevier, 2006.

[6] N. Bondoux, A. Q. Nguyen, T. Fiig, and R. Acuna-Agost, "Reinforcement learning applied to airline revenue management," *J. Revenue Pricing Manage.*, vol. 19, pp. 332–348, Jan. 2020.

[7] D. Elreedy, A. F. Atiya, H. Fayed, and M. Saleh, "A framework for an agent-based dynamic pricing for broadband wireless price rate plans," *J. Simul.*, vol. 13, no. 2, pp. 96–110, 2019.

[8] C. Triki and A. Violi, "Dynamic pricing of electricity in retail markets," *4OR*, vol. 7, no. 1, pp. 21–36, Mar. 2009.

[9] S. Chen, G. Sun, Z. Wei, and D. Wang, "Dynamic pricing in electricity and natural gas distribution networks: An EPEC model," *Energy*, vol. 207, Sep. 2020, Art. no. 118138.

[10] Z. Almahmoud, J. Crandall, K. Elbassioni, T. T. Nguyen, and M. Roozbehani, "Dynamic pricing in smart grids under thresholding policies," *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 3415–3429, May 2019.

[11] C. H. Xia and P. Dube, "Dynamic pricing in e-services under demand uncertainty," *Prod. Oper. Manage.*, vol. 16, no. 6, pp. 701–712, Jan. 2009.

[12] A. X. Carvalho and M. L. Puterman, "Dynamic pricing and reinforcement learning," in *Proc. Int. Joint Conf. Neural Netw.*, vol. 4, 2003, pp. 2916–2921.

[13] Y. Aviv and A. Pazgal, "Pricing of short life-cycle products through active learning," Olin School Bus., Under Revision Manage. Sci., Washington Univ., St. Louis, MO, USA, Work. Paper, Oct. 2002, pp. 1–32.

[14] V. F. Araman and R. Caldentey, "Dynamic pricing for nonperishable products with demand learning," *Oper. Res.*, vol. 57, no. 5, pp. 1169–1188, Oct. 2009.

[15] N. B. Keskin and A. Zeevi, "Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies," *Oper. Res.*, vol. 62, no. 5, pp. 1142–1167, Oct. 2014.

[16] A. V. den Boer, "Dynamic pricing and learning: Historical origins, current research, and new directions," *Surveys Oper. Res. Manage. Sci.*, vol. 20, no. 1, pp. 1–18, Jun. 2015.

[17] M. Rothschild, "A two-armed bandit theory of market pricing," *J. Econ. Theory*, vol. 9, no. 2, pp. 185–202, Oct. 1974.

[18] W. C. Cheung, D. Simchi-Levi, and H. Wang, "Dynamic pricing and demand learning with limited price experimentation," *SSRN Electron. J.*, vol. 65, no. 6, pp. 1722–1731, 2017.

[19] T. Osugi, D. Kun, and S. Scott, "Balancing exploration and exploitation: A new algorithm for active machine learning," in *Proc. 5th IEEE Int. Conf. Data Mining (ICDM)*, 2005, pp. 1–8.
[20] M. Tokic, "Adaptive ε-greedy exploration in reinforcement learning based on value differences," in *Proc. Annu. Conf. Artif. Intell.* Berlin, Germany: Springer, 2010, pp. 203–210.
[21] M. Črepinšek, S.-H. Liu, and M. Mernik, "Exploration and exploitation in evolutionary algorithms: A survey," *ACM Comput. Surveys*, vol. 45, no. 3, p. 35, 2013.
[22] J. Kelly, E. Hemberg, and U.-M. O'Reilly, "Improving genetic programming with novel exploration-exploitation control," in *Proc. Eur. Conf. Genetic Program.* Cham, Switzerland: Springer, 2019, pp. 64–80.
[23] M. S. Lobo and S. Boyd, "Pricing and learning with uncertain demand," Duke Univ., Durham, NC, USA, Work. Paper, 2003, pp. 1–35.
[24] A. V. den Boer and B. Zwart, "Simultaneously learning and optimizing using controlled variance pricing," *Manage. Sci.*, vol. 60, no. 3, pp. 770–783, Mar. 2014.
[25] D. Bertsimas and G. Perakis, "Dynamic pricing: A learning approach," in *Mathematical and Computational Models for Congestion Charging*. Boston, MA, USA: Springer, 2006, pp. 45–79.
[26] Y. Cheng, "Dynamic pricing decision for perishable goods: A Q-learning approach," in *Proc. 4th Int. Conf. Wireless Commun., Netw. Mobile Comput. (WiCOM)*, Oct. 2008, pp. 1–5.
[27] O. Besbes and A. Zeevi, "On the (surprising) sufficiency of linear models for dynamic pricing with demand learning," *Manage. Sci.*, vol. 61, no. 4, pp. 723–739, Apr. 2015.
[28] A. F. Atiya, M. A. Aly, and A. G. Parlos, "Sparse basis selection: New results and application to adaptive prediction of video source traffic," *IEEE Trans. Neural Netw.*, vol. 16, no. 5, pp. 1136–1146, Sep. 2005.
[29] R. Garnett, Y. Krishnamurthy, X. Xiong, J. Schneider, and R. Mann, "Bayesian optimal active search and surveying," 2012, *arXiv:1206.6406*. [Online]. Available: http://arxiv.org/abs/1206.6406
[30] S. Jiang, G. Malkomes, G. Converse, A. Shofner, B. Moseley, and R. Garnett, "Efficient nonmyopic active search," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1714–1723.
[31] V. F. Araman and R. Caldentey, "Revenue management with incomplete demand information," in *Wiley Encyclopedia of Operations Research and Management Science*. Hoboken, NJ, USA: Wiley, 2010.
[32] Y. Aviv and G. Vulcano, "Dynamic list pricing," in *The Oxford Handbook of Pricing Management*. London, U.K.: Oxford Univ. Press, 2012, pp. 522–584.
[33] A. den Boer, "Dynamic pricing and learning," Ph.D. dissertation, Dept. Math., Fac. Sci., Free Univ. Amsterdam, Amsterdam, The Netherlands, 2012.
[34] A. X. Carvalho and M. L. Puterman, "Learning and pricing in an Internet environment with binomial demands," *J. Revenue Pricing Manage.*, vol. 3, no. 4, pp. 320–336, Jan. 2005.
[35] J. M. Harrison, N. B. Keskin, and A. Zeevi, "Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution," *Manage. Sci.*, vol. 58, no. 3, pp. 570–586, Mar. 2012.
[36] G. Aydin and S. Ziya, "Personalized dynamic pricing of limited inventories," *Oper. Res.*, vol. 57, no. 6, pp. 1523–1531, Dec. 2009.
[37] J. Diao, K. Zhu, and Y. Gao, "Agent-based simulation of durables dynamic pricing," *Syst. Eng. Procedia*, vol. 2, pp. 205–212, Jan. 2011.
[38] S. Morales-Enciso and J. Branke, "Revenue maximization through dynamic pricing under unknown market behaviour," in *Proc. 3rd Student Conf. Oper. Res.* Wadern, Germany: Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012, pp. 11–20.
[39] S. S. Gupta and K. J. Miescke, "Bayesian look ahead one-stage sampling allocations for selection of the best population," *J. Stat. Planning Inference*, vol. 54, no. 2, pp. 229–244, Sep. 1996.
[40] P. I. Frazier, W. B. Powell, and S. Dayanik, "A knowledge-gradient policy for sequential information collection," *SIAM J. Control Optim.*, vol. 47, no. 5, pp. 2410–2439, Jan. 2008.
[41] T. G. Dietterich and X. Wang, "Batch value function approximation via support vectors," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 1491–1498.
[42] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
[43] E. Kutschinski, T. Uthmann, and D. Polani, "Learning competitive pricing strategies by multi-agent reinforcement learning," *J. Econ. Dyn. Control*, vol. 27, nos. 11–12, pp. 2207–2218, Sep. 2003.
[44] W. Jintian and Z. Lei, "Application of reinforcement learning in dynamic pricing algorithms," in *Proc. IEEE Int. Conf. Autom. Logistics*, Aug. 2009, pp. 419–423.
[45] R. Rana and F. S. Oliveira, "Dynamic pricing policies for interdependent perishable products or services using reinforcement learning," *Expert Syst. Appl.*, vol. 42, no. 1, pp. 426–436, Jan. 2015.
[46] R. Maestre, J. Duque, A. Rubio, and J. Arévalo, "Reinforcement learning for fair dynamic pricing," in *Proc. SAI Intell. Syst. Conf.* Cham, Switzerland: Springer, 2018, pp. 120–135.
[47] R. Lu, S. H. Hong, and X. Zhang, "A dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach," *Appl. Energy*, vol. 220, pp. 220–230, Jun. 2018.
[48] R. Martinez-Cantin, N. de Freitas, E. Brochu, J. Castellanos, and A. Doucet, "A Bayesian exploration-exploitation approach for optimal online sensing and planning with a visually guided mobile robot," *Auto. Robots*, vol. 27, no. 2, pp. 93–103, Aug. 2009.
[49] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, nos. 2–3, pp. 235–256, 2002.
[50] J. Vermorel and M. Mohri, "Multi-armed bandit algorithms and empirical evaluation," in *Proc. Eur. Conf. Mach. Learn.* Berlin, Germany: Springer, 2005, pp. 437–448.
[51] H. Valizadegan, R. Jin, and S. Wang, "Learning to trade off between exploration and exploitation in multiclass bandit prediction," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2011, pp. 204–212.
[52] O. Besbes, Y. Gur, and A. Zeevi, "Optimal exploration-exploitation in a multi-armed-bandit problem with non-stationary rewards," 2014, *arXiv:1405.3316*. [Online]. Available: http://arxiv.org/abs/1405.3316
[53] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, nos. 3–4, pp. 285–294, Dec. 1933.
[54] H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Amer. Math. Soc.*, vol. 58, no. 5, pp. 527–535.
[55] J.-Y. Audibert, R. Munos, and C. Szepesvári, "Exploration–exploitation tradeoff using variance estimates in multi-armed bandits," *Theor. Comput. Sci.*, vol. 410, no. 19, pp. 1876–1902, Apr. 2009.
[56] S. Pandey, D. Agarwal, D. Chakrabarti, and V. Josifovski, "Bandits for taxonomies: A model-based approach," in *Proc. SIAM Int. Conf. Data Mining*. Philadelphia, PA, USA: SIAM, Apr. 2007, pp. 216–227.
[57] S. S. Villar, J. Bowden, and J. Wason, "Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges," *Stat. Sci. A, Rev. J. Inst. Math. Statist.*, vol. 30, no. 2, p. 199, May 2015.
[58] F. Trovo, S. Paladino, M. Restelli, and N. Gatti, "Multi–armed bandit for pricing," in *Proc. Eur. Workshop Reinforcement Learn. (EWRL)*, 2015, pp. 1–9.
[59] P. Auer, "Using confidence bounds for exploitation-exploration tradeoffs," *J. Mach. Learn. Res.*, vol. 3, pp. 397–422, Nov. 2003.
[60] B. Settles, "Active learning literature survey," Dept. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, Tech. Rep. 1648, 2009.
[61] D. Elreedy, A. F. Atiya, and S. I. Shaheen, "A novel active learning regression framework for balancing the exploration-exploitation tradeoff," *Entropy*, vol. 21, no. 7, p. 651, Jul. 2019.
[62] R. J. Rossi, *Mathematical Statistics: An Introduction to Likelihood Based Inference*. Hoboken, NJ, USA: Wiley, 2018.
[63] E. Uriel, "3 Multiple linear regression: Estimation and properties," *Parameters*, vol. 1, no. 2, p. 3, 2013.
[64] R. H. Byrd, M. E. Hribar, and J. Nocedal, "An interior point algorithm for large-scale nonlinear programming," *SIAM J. Optim.*, vol. 9, no. 4, pp. 877–900, Jan. 1999.
[65] J. Gwartney, R. Stroup, R. Sobel, and A. Macpherson, *Economics: Private & Public Choice*. Orlando, FL, USA: Har-Court, Brace, Jovanovich, 1983.
[66] C. U. M. Library. (2001). *Musdaers Electronic Data Archive, Red Meats Yearbook*. [Online]. Available: http://usda.mannlib.cornell.edu/
[67] J. Durbin and G. S. Watson, "Testing for serial correlation in least squares regression: I," *Biometrika*, vol. 37, nos. 3–4, pp. 409–428, 1950.
[68] H. Schultz, "A comparison of elasticities of demand obtained by different methods," *Econometrica, J. Econ. Soc.*, vol. 37, pp. 274–308, Dec. 1933.
[69] Y. Sun. (2011). *Coke Demand Estimation Dataset*. [Online]. Available: http://leeds-faculty.colorado.edu/ysun/doc/Demand_estimation_worksheet.doc

[70] M. Magdon-Ismail and A. Atiya, "Density estimation and random variate generation using multilayer networks," *IEEE Trans. Neural Netw.*, vol. 13, no. 3, pp. 497–520, May 2002.

DINA ELREEDY received the B.Sc. and M.Sc. degrees from the Computer Engineering Department, Cairo University, in 2012 and 2015, respectively, the M.Sc. degree from the Computer Science and Engineering Department, Washington University in Saint Louis, in 2016, and the Ph.D. degree from the Computer Engineering Department, Cairo University, in 2020.

She is currently an Assistant Professor with the Computer Engineering Department, Cairo University. Her research interests include machine learning, pattern recognition, natural language processing, and data mining.

AMIR F. ATIYA (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Cairo University, and the M.S. and Ph.D. degrees in electrical engineering from Caltech, Pasadena, CA, USA. He is currently a Professor with the Department of Computer Engineering, Cairo University. His research interests include machine learning, theory of forecasting, computational finance, and dynamic pricing. He obtained several awards, such as the Kuwait Prize, in 2005, and the Egyptian State Appreciation Award, in 2018. He is a Handling Editor of *International Journal of Forecasting*.

SAMIR I. SHAHEEN (Life Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Cairo University, and the Ph.D. degree in electrical engineering from McGill University, Montreal, Canada. He is currently a Professor with the Department of Computer Engineering, Cairo University. His research interests include knowledge engineering, intelligent systems, and computer vision. He received the Egyptian State Excellence Prize in Advanced Technological Sciences in Engineering, in 2014, and the Cairo University Science Distinction Award, in 2020. He was the PI for several projects in e-learning, telemedicine, family medicine, and e-government. He is the founder of the e-learning and telemedicine initiatives in Egypt. He has several projects with the European Union in GIS and e-learning.