Received: 20 September 2018 | Revised: 13 November 2019 | Accepted: 16 March 2020
DOI: 10.1002/nav.21902

RESEARCH ARTICLE

WILEY

Technical note: Finite-time regret analysis of Kiefer-Wolfowitz stochastic approximation algorithm and nonparametric multi-product dynamic pricing with unknown demand

L. Jeff Hong[1] | Chenghuai Li[2] | Jun Luo[3]

[1]School of Management and School of Data Science, Fudan University, Shanghai, China
[2]School of Mathematical Sciences and School of Data Science, Fudan University, Shanghai, China
[3]Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai, China

Correspondence
L. Jeff Hong, School of Management and School of Data Science, Fudan University, Shanghai 200433, China.
Email: hong_liu@fudan.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 71531010, 71722006, 71991473. The Hong Kong Research Grants Council, Grant/Award Number: GRF 11504017.

Abstract

We consider the problem of nonparametric multi-product dynamic pricing with unknown demand and show that the problem may be formulated as an online model-free stochastic program, which can be solved by the classical Kiefer-Wolfowitz stochastic approximation (KWSA) algorithm. We prove that the expected cumulative regret of the KWSA algorithm is bounded above by $\kappa_1\sqrt{T}+\kappa_2$ where κ_1, κ_2 are positive constants and T is the number of periods for any $T = 1, 2, \ldots$. Therefore, the regret of the KWSA algorithm grows in the order of $\sqrt{T}$, which achieves the lower bounds known for parametric dynamic pricing problems and shows that the nonparametric problems are not necessarily more difficult to solve than the parametric ones. Numerical experiments further demonstrate the effectiveness and efficiency of our proposed KW pricing policy by comparing with some pricing policies in the literature.

KEYWORDS

dynamic pricing and learning, Kiefer-Wolfowitz algorithm, nonparametric pricing policy, revenue management, stochastic approximation

1 | INTRODUCTION

Companies often need to make pricing decisions with little knowledge on the demand function. In such cases, they may experiment different selling prices in different time periods to learn the demand function (of the prices), and to find the optimal prices that maximize their revenues. The problem has been studied extensively in the operations research and management science literature recently (see Keskin and Zeevi (2014) and the references therein). Most of the papers consider a single product and assume the demand function of the product follows a parametric model with unknown parameters that need to be learned. For instance, den Boer and Zwart (2014) and Keskin and Zeevi (2014) both assume the demand function is linear in the price, and Broder and Rusmevichientong (2012) assume a parametric model of the customer's willingness-to-pay distribution. One critical insight learned from the literature is that a myopic pricing policy, which always sets the price as the current best based on all the information collected in all previous periods, typically leads to *incomplete learning* of the demand function, and is therefore not optimal. Here the incomplete learning means that the parameters of the demand functions cannot be learned consistently, that is, the parameter estimators do not converge to the true values as the number of periods goes to infinity (cf. Keskin and Zeevi (2018) for a detailed introduction on incomplete learning).

To avoid incomplete learning, different pricing policies have been proposed. Broder and Rusmevichientong (2012) propose the MLE-CYCLE policy which separates demand learning and revenue optimization. By balancing the two efforts in an optimal way, they show that the expected cumulative regret (ie, the difference between the expected cumulative revenues if we know the optimal price and if we use the proposed pricing policy for all periods from 1 to T) is of the order $\sqrt{T}$. den Boer and Zwart (2014) propose a controlled variance pricing policy that perturbs the myopic policy to prevent the price from concentrating too fast to ensure sufficient learning. By controlling the variance in an optimal way, they also show that the expected cumulative regret is of the order

$T^{1/2+\delta}$ for a small $\delta > 0$. Keskin and Zeevi (2014) also propose semimyopic pricing policies to ensure complete learning and show that the expected cumulative regret is of the order $\sqrt{T}$ and they also extend their results to the pricing of multiple products. In all three aforementioned pricing policies, the parameters of the demand function may be estimated consistently by using forced exploration, thus avoiding incomplete learning. Recently, Keskin and Zeevi (2018) propose a limited-memory learning scheme (ie, adaptively choosing the estimation windows) to improve the certainty-equivalence policy in both static and slowly time-varying environments without forced exploration. Furthermore, Broder and Rusmevichientong (2012) and Keskin and Zeevi (2014) prove respectively that the lower bounds of the expected cumulative regrets of their problems are $\sqrt{T}$, no matter what pricing policies are used. Therefore, the aforementioned pricing policies are either asymptotically optimal or asymptotically near optimal.

In practice, the parametric form of the demand function is typically unknown and using parametric models may cause model mis-specifications that cannot be removed. To solve the problem, Besbes and Zeevi (2015) consider a single-product pricing problem and take a nonparametric approach to approximate the demand function by a first-order Taylor's expansion (ie, a linear function) with a finite-difference derivative estimator. They show that, by controlling the size of the finite difference appropriately and by allowing the derivative at the current solution to be learned completely (ie, converging to the true value), the resulted optimal price converges to the true one and the expected cumulative regret is of the order $\sqrt{T}(\log T)^2$. This is a truly surprising result, because it shows that without a parametric assumption and without the complete learning of the demand function, the expected cumulative regret of the nonparametric approach is almost asymptotically optimal even for a parametric problem. Furthermore, Besbes and Zeevi (2015) *conjecture* that it is possible to achieve the optimal rate of $\sqrt{T}$ even for the nonparametric approach.

In this note we extend the nonparametric formulation of Besbes and Zeevi (2015) to a multi-product setting, and show that the problem is an example of model-free stochastic programs. We propose to use a variant of Kiefer-Wolfowitz stochastic approximation (KWSA) algorithm, the most famous model-free stochastic programming tool, to solve the problem and call it the Kiefer-Wolfowitz (KW) pricing policy. Instead of analyzing the regret of our specific problem, we analyze the regret of the KWSA algorithm for general stochastic programs and show that its expected cumulative regret is bounded by $\kappa_1\sqrt{T}+\kappa_2$ for some problem-dependent constants $\kappa_1 > 0$ and $\kappa_2 > 0$ for all $T = 1, 2, \ldots$. By applying this general result to the multi-product dynamic pricing problem, we show that the expected cumulative regret of the KW pricing policy is also bounded above by $\kappa_1\sqrt{T} + \kappa_2$, which is asymptotically optimal based on the lower bounds of Broder and Rusmevichientong (2012) and Keskin and Zeevi (2014) and which also proves the conjecture of Besbes and Zeevi (2015), not only for single-product cases but also for multiple-product cases, without assuming a parametric model of the demand function.

Our work is related to two different streams of literature. (There is an abundant literature on both dynamic pricing and stochastic approximation. Due to the space limitation of a technical note, we focus only on the works that are closely related to the problem studied in this paper.) The first stream is dynamic pricing with unknown demand learning in the revenue management area. Besides the aforementioned papers, that is, Broder and Rusmevichientong (2012), den Boer and Zwart (2014), Keskin and Zeevi (2014, 2018), and Besbes and Zeevi (2015), we would like to add a few closely related works. Lobo and Boyd (2003) is the first work that identifies through numerical studies that a myopic pricing policy is not optimal for a linear demand function, and price dithering, that is, adding noises to the myopic solutions, may improve the performance of the pricing policy. Harrison, Keskin, and Zeevi (2012) is one of the early works that demonstrate a myopic Bayesian policy may lead to incomplete learning under the simplest model uncertainty setting (ie, a binary demand model). Interesting readers may refer to Aviv and Vulcano (2012) and den Boer (2015) for comprehensive reviews on various dynamic pricing and learning problem formulations.

The second stream of literature is stochastic approximation (SA). SA algorithms, such as those of Robbins and Monro (1951) and Kiefer and Wolfowitz (1952), are typically used to solve offline stochastic optimization problems, and they have been used in many different areas (cf. the books of Benveniste, Priouret, and Metivier (1990) and Kushner and Yin (2003) for introductions). The rates of convergence of these algorithms have been studied extensively (see Chapter 10 of Kushner and Yin (2003)). In particular, Fabian (1967) proved that the asymptotic rate of convergence of the solutions is $T^{-1/4}$. This result implies that, if the algorithm is used to solve online stochastic optimization problems (such as ours), the asymptotic rate of the regret is $\sqrt{T}$. (The terms "offline" and "online" refer to two types of objectives when solving stochastic optimization problems. In offline problems, we only care about the quality of the final solution when the algorithm stops. In online problems, we care about the quality of every solution that the algorithm evaluates before it stops, in particular we want to minimize the cumulative optimality gaps (i.e., cumulative regrets) between every evaluated solution and the optimal.) The same asymptotic rate was also obtained by Cope (2009). However, the rate of convergence results is typically asymptotic. In this note, we borrow some ideas of Nemirovski, Juditsky, Lan, and Shapiro (2009), who propose a robust SA algorithm, to conduct a finite-time regret analysis of our variant of KWSA algorithm and show that the regret is upper bounded by $\kappa_1\sqrt{T} + \kappa_2$ for all $T = 1, 2, \ldots$.

KWSA algorithm, which is the fundamental algorithm behind our approach, is also an example of the zeroth-order (or derivative-free) stochastic convex optimization in the literature. Compared with first-order problems (ie, the gradient information is available) whose rate of convergence of the cumulative regret is lower bounded by $\mathcal{O}(\log T)$ for strongly convex cost functions (cf. Zinkevich, 2003), zeroth-order problems are in general more difficult and the rate of convergence of the cumulative regret is only lower bounded by $\mathcal{O}(\sqrt{T})$ (see, for instant, the algorithms in Agarwal, Dekel, and Xiao (2010) achieving the optimal bound $\mathcal{O}(\sqrt{T})$ for strongly convex functions and Agarwal, Foster, Hsu, Kakade, and Rakhlin (2013) achieving a near-optimal bound $\mathcal{O}(\sqrt{T}(\log T)^2)$ for general convex functions, respectively). Interested readers may refer to Table 1 in both Shamir (2013) and Besbes, Gur, and Zeevi (2015) for summaries on the rate of convergence of the average regret and cumulative regret under various settings of stochastic convex optimization problems.

Indeed, both den Boer and Zwart (2014) and Besbes and Zeevi (2015) also notice the connection of the dynamic pricing problem to SA algorithms. However, den Boer and Zwart (2014) point out that SA algorithms are used to solve offline problems instead of online problems and "the performance [of a SA algorithm] is measured by the quality of the estimate of the maximizer, and not by the cumulative costs [ie, regrets] that leads to the estimate." In this note, we show that SA algorithms can also be used to solve online optimization problems and achieve asymptotic optimality. Besbes and Zeevi (2015) have a subsection on KWSA algorithms (ie, Section 3.3). But as they point out, their focus is not on the nonparametric approach to solving the problem but "to understand whether policies that are typically used for learning and earning (and designed for well-specified scenarios) suffer from misspecification." In this note, our focus is to explore the asymptotical optimality of the KWSA algorithm in solving the multi-product dynamic pricing problem.

We want to emphasize that *the regret analysis of the KWSA algorithm is quite standard and it is not the main contribution of this technical note*. Instead, this technical note only tries to connect the two streams of literature, and shows that the dynamic pricing problem can indeed be solved by the KWSA algorithm with the optimal rate.

The rest of this technical note is organized as follows: in Section 2 we introduce our variant of the KWSA algorithm and analyze its finite-time regret. In Section 3 we formulate the nonparametric multi-product dynamic pricing problem and show that it may be solve by the KWSA algorithm and the resulted KW pricing policy is asymptotically optimal. Numerical studies demonstrate the effectiveness and efficiency of our KW pricing policy by comparing with some existing pricing policies in Section 4, followed by concluding remarks in Section 5. Some additional proofs are included in the Appendix.

ALGORITHM 1 Online KWSA algorithm

Initialization. Let $\mathbf{x}_0 \in \Omega$ be a starting solution. Let the iteration counter $n = 1$ and period counter $t = 0$.

Step 1. Function evaluations and information collection.

- Let $t = t + 1$. Set $\widetilde{\mathbf{x}}_t = \mathbf{x}_n$ and observe $F(\widetilde{\mathbf{x}}_t, \xi_{n,0})$;
- For $i = 1, 2, \ldots, d$,
  Let $t = t + 1$. Set $\widetilde{\mathbf{x}}_t = \mathbf{x}_n + c_n\mathbf{e}_i$ and observe $F(\widetilde{\mathbf{x}}_t, \xi_{n,i})$.
  End the for-loop.

Step 2. Updating.

Let
$$\mathbf{x}_{n+1} = \Pi_\Omega(\mathbf{x}_n - a_n\mathbf{G}(\mathbf{x}_n)),$$
where Π_Ω is a projection operator onto the set Ω, that is, $\Pi_\Omega(\mathbf{x}) = \operatorname{argmin}_{\mathbf{x}'\in\Omega}\|\mathbf{x} - \mathbf{x}'\|$, and
$$\mathbf{G}(\mathbf{x}_n) = \frac{1}{c_n}([F(\mathbf{x}_n + c_n\mathbf{e}_1, \xi_{n,1}) - F(\mathbf{x}_n, \xi_{n,0})], \ldots, [F(\mathbf{x}_n + c_n\mathbf{e}_d, \xi_{n,d}) - F(\mathbf{x}_n, \xi_{n,0})])^{\mathrm{T}}.$$

Let $n = n + 1$ and go back to Step 1.

2 | KWSA ALGORITHM AND FINITE-TIME REGRET

Consider the following online stochastic optimization problem:

$$\min_{\mathbf{x}\in\Omega} \{f(\mathbf{x}) := \mathrm{E}[F(\mathbf{x}, \xi)]\}, \tag{1}$$

where $\Omega \subset \Re^d$ is convex and compact. We propose to use a variant of KWSA algorithm (ie, Algorithm 1) to solve the problem. Let $\{a_n, n = 1, 2, \ldots\}$ and $\{c_n, n = 1, 2, \ldots\}$ be two positive sequences of real numbers, which will be used in the algorithm.

> *Remark* 1 Note that during each iteration round, Algorithm 1 conducts $d + 1$ periods of function evaluations. That is, in the nth iteration, the time period t changes from $k + 1$ to $k + d + 1$, where $k = n(d + 1)$, and the function evaluations have been taken down over $\widetilde{\mathbf{x}}_{k+1} = \mathbf{x}_n, \widetilde{\mathbf{x}}_{k+2} = \mathbf{x}_n + c_n\mathbf{e}_1, \ldots, \widetilde{\mathbf{x}}_{k+d+1} = \mathbf{x}_n + c_n\mathbf{e}_d$. Also note that $\{a_n, n = 1, 2, \ldots\}$ and $\{c_n, n = 1, 2, \ldots\}$ in Algorithm 1 as well as in Algorithm 2 can be properly selected to guarantee the convergence rate, which will be introduced in details in Theorems 1 and 2 and be discussed more in Remark 5.

Algorithm 1 is different from the classical KWSA algorithm of Kiefer and Wolfowitz (1952) in two aspects. First, the classical KWSA algorithm is for solving offline optimization problems where only the terminal solutions of the iterations, that is, $\mathbf{x}_1, \mathbf{x}_2, \ldots$, are of interest. While in our version of the algorithm, we are interested in all evaluated solutions $\widetilde{\mathbf{x}}_1, \widetilde{\mathbf{x}}_2, \ldots$ because they all produce regret. This is the difference between offline and online stochastic optimizations. Second, Algorithm 1 uses a forward finite-difference gradient estimator $\mathbf{G}(\mathbf{x}_n)$ instead of a central finite-difference estimator. We show in Remark 5 that the asymptotic growth rates of the expected cumulative regret of using both gradient estimators are the same. Therefore, we use the forward finite-difference estimator for simplicity.

Let $\mathbf{x}^* = \mathrm{argmin}_{\mathbf{x}\in\Omega} f(\mathbf{x})$ be the optimal solution of Problem (1). Then, we may define the expected cumulative regret of the first T periods as

$$R(T) = \sum_{t=1}^{T} \mathrm{E}[f(\widetilde{\mathbf{x}}_t) - f(\mathbf{x}^*)]$$

for any $T = 1, 2, \ldots$.

To analyze the regret $R(T)$ of Algorithm 1 in solving Problem (1), we make the following assumptions on the problem.

Assumption 1 There exists a finite constant $M > 0$ such that $\mathrm{E}\{[F(\mathbf{x}, \xi)]^2\} \le M$ for all $\mathbf{x} \in \Omega$.

Assumption 2 The expected value function $f(\mathbf{x})$ is twice continuously differentiable in Ω.

Assumption 3 The expected value function $f(\mathbf{x})$ is strongly convex in Ω, that is, there exists a finite constant $B_1 > 0$ such that $f(\mathbf{x}') \ge f(\mathbf{x}) + \nabla f(\mathbf{x})^{\mathrm{T}}(\mathbf{x}' - \mathbf{x}) + \frac{1}{2}B_1\|\mathbf{x}' - \mathbf{x}\|^2$.

Notice that Assumption 1 implies that $f(\mathbf{x})$ is well defined for all $\mathbf{x} \in \Omega$, Assumptions 2 implies that

$$\|\nabla^2 f(\mathbf{x})\| \le B_2 \tag{2}$$

for some finite $B_2 > 0$ for all $\mathbf{x} \in \Omega$, and Assumption 3 implies that

$$[\nabla f(\mathbf{x}') - \nabla f(\mathbf{x})]^{\mathrm{T}}(\mathbf{x}' - \mathbf{x}) \ge B_1\|\mathbf{x}' - \mathbf{x}\| \tag{3}$$

for all $\mathbf{x}, \mathbf{x}' \in \Omega$. Moreover, Assumption 3 also implies that there is a unique optimal solution to Problem (1). We denote it as $\mathbf{x}^*$ and make the following assumption to assume it is in the interior of Ω, that is, $\mathrm{int}(\Omega)$, which implies that $\nabla f(\mathbf{x}^*) = \mathbf{0}$.

Assumption 4 The optimal solution $\mathbf{x}^*$ satisfies $\mathbf{x}^* \in \mathrm{int}(\Omega)$.

Given the above assumptions, we can prove the following theorem that bounds $\mathrm{E}[\|\mathbf{x}_n - \mathbf{x}^*\|^2]$. The analysis used to prove the theorem is motivated by Kiefer and Wolfowitz (1952) and Nemirovski et al. (2009). In particular, Nemirovski et al. (2009) analyze the convergence of $\mathrm{E}[\|\mathbf{x}_n - \mathbf{x}^*\|^2]$ of Robbins-Monro SA algorithm, which assumes an unbiased estimator of $\nabla f(\mathbf{x}_n)$ is available, known as the first-order stochastic optimization problem in the literature. In our algorithm, it is not available and a forward finite-difference estimator is used instead.

Theorem 1 *Suppose that Algorithm 1 is used to solve Problem (1) and Assumptions 1–4 are satisfied. Let $a_n = \gamma n^{-1}$ and $c_n = \delta n^{-\frac{1}{4}}$ with $1/(4B_1) < \gamma < 1/(2B_1)$ and $\delta > 0$. Then, there exists a constant $\lambda > 0$ such that $\mathrm{E}[\|\mathbf{x}_n - \mathbf{x}^*\|^2] \le \lambda n^{-\frac{1}{2}}$ for all $n = 1, 2, \ldots$.*

Proof Let $b_n = \mathrm{E}[\|\mathbf{x}_n - \mathbf{x}^*\|^2]$. Notice that $\Pi_\Omega(\mathbf{x}^*) = \mathbf{x}^*$ and $\|\Pi_\Omega(\mathbf{x}) - \Pi_\Omega(\mathbf{x}')\| \le \|\mathbf{x} - \mathbf{x}'\|$. Then,

$$\begin{aligned}
b_{n+1} &= \mathrm{E}\{\|\Pi_\Omega(\mathbf{x}_n - a_n\mathbf{G}(\mathbf{x}_n)) - \mathbf{x}^*\|^2\} \\
&= \mathrm{E}\{\|\Pi_\Omega(\mathbf{x}_n - a_n\mathbf{G}(\mathbf{x}_n)) - \Pi_\Omega(\mathbf{x}^*)\|^2\} \\
&\le \mathrm{E}\{\|\mathbf{x}_n - a_n\mathbf{G}(\mathbf{x}_n) - \mathbf{x}^*\|^2\} \\
&= b_n + a_n^2\mathrm{E}(\|\mathbf{G}(\mathbf{x}_n)\|^2) - 2a_n\mathrm{E}[\mathbf{G}(\mathbf{x}_n)^{\mathrm{T}}(\mathbf{x}_n - \mathbf{x}^*)].
\end{aligned} \tag{4}$$

Let

$$\mathbf{g}(\mathbf{x}) = \frac{1}{c_n}([f(\mathbf{x}_n + c_n\mathbf{e}_1) - f(\mathbf{x}_n)], \ldots, ([f(\mathbf{x}_n + c_n\mathbf{e}_d) - f(\mathbf{x}_n)])^{\mathrm{T}}.$$

Notice that

$$\begin{aligned}
\mathrm{E}[\mathbf{G}(\mathbf{x}_n)^{\mathrm{T}}(\mathbf{x}_n - \mathbf{x}^*)] &= \mathrm{E}\{\mathrm{E}[\mathbf{G}(\mathbf{x}_n)^{\mathrm{T}}(\mathbf{x}_n - \mathbf{x}^*)\mathbf{x}_n]\} \\
&= \mathrm{E}[\mathbf{g}(\mathbf{x}_n)^{\mathrm{T}}(\mathbf{x}_n - \mathbf{x}^*)] \\
&= \mathrm{E}[\nabla f(\mathbf{x}_n)^{\mathrm{T}}(\mathbf{x}_n - \mathbf{x}^*)] + \mathrm{E}[(\mathbf{g}(\mathbf{x}_n) \\
&\quad - \nabla f(\mathbf{x}_n))^{\mathrm{T}}(\mathbf{x}_n - \mathbf{x}^*)].
\end{aligned} \tag{5}$$

By Assumption 4, $\nabla f(\mathbf{x}^*) = \mathbf{0}$. Then, by Equation (3),

$$\begin{aligned}
\mathrm{E}[\nabla f(\mathbf{x}_n)^{\mathrm{T}}(\mathbf{x}_n - \mathbf{x}^*)] &= \mathrm{E}\{[\nabla f(\mathbf{x}_n) - \nabla f(\mathbf{x}^*)]^{\mathrm{T}}(\mathbf{x}_n - \mathbf{x}^*)\} \\
&\ge B_1\mathrm{E}[\|\mathbf{x}_n - \mathbf{x}^*\|^2] = B_1 b_n.
\end{aligned} \tag{6}$$

Notice that, by Taylor's expansion and Equation (2),

$$\begin{aligned}
\mathbf{g}(\mathbf{x}_n) - \nabla f(\mathbf{x}_n) &= \frac{1}{2}c_n(\mathbf{e}_1^T\nabla^2(\boldsymbol{\eta}_1)\mathbf{e}_1, \ldots, \mathbf{e}_d^T\nabla^2(\boldsymbol{\eta}_d)\mathbf{e}_d)^{\mathrm{T}} \\
&\le \frac{1}{2}B_2c_n\mathbf{1},
\end{aligned}$$

where $\mathbf{1}$ is a d-dimensional vector with all elements being 1. Then,

$$\begin{aligned}
&\mathrm{E}[(\mathbf{g}(\mathbf{x}_n) - \nabla f(\mathbf{x}_n))^{\mathrm{T}}(\mathbf{x}_n - \mathbf{x}^*)] \ge -\frac{1}{2}B_2c_n\mathrm{E}[\mathbf{x}_n - \mathbf{x}^*x] \\
&\ge -\frac{1}{2}B_2c_n\sqrt{d}\sqrt{b_n},
\end{aligned} \tag{7}$$

where the last inequality follows from Jensen's inequality. Therefore, by Equations (5)-(7),

$$\mathrm{E}[\mathbf{G}(\mathbf{x}_n)^{\mathrm{T}}(\mathbf{x}_n - \mathbf{x}^*)] \ge B_1 b_n - \frac{1}{2}B_2c_n\sqrt{d}\sqrt{b_n}. \tag{8}$$

Furthermore, by Assumption 1,

$$\mathrm{E}(\|\mathbf{G}(\mathbf{x}_n)\|^2) \le \frac{4dM^2}{c_n^2}. \tag{9}$$

Then, by Equations (4), (8), and (9),

$$\begin{aligned}
b_{n+1} &\le (1 - 2a_nB_1)b_n + a_nc_nB_2\sqrt{d}\sqrt{b_n} \\
&\quad + \frac{4da_n^2}{c_n^2}M^2.
\end{aligned} \tag{10}$$

Notice that $a_n = \gamma n^{-1}$ and $c_n = \delta n^{-\frac{1}{4}}$ with $1/(4B_1) < \gamma < 1/(2B_1)$ and $\delta > 0$. Then, we have

$$b_{n+1} \leq \left(1 - \frac{2\gamma B_1}{n}\right) b_n + \gamma\delta B_2 \sqrt{d}\ n^{-\frac{5}{4}}\sqrt{b_n} + \frac{4d\gamma^2 M^2}{\delta^2} n^{-\frac{3}{2}}. \tag{11}$$

By induction (see Appendix), there exists $\lambda > 0$ such that

$$b_n \leq \lambda\ n^{-\frac{1}{2}}, \quad n = 1, 2, \ldots \tag{12}$$

This concludes the proof of the theorem. ∎

Remark 2 Fabian (1967) proves that $\mathrm{E}[\|\mathbf{x}_n - \mathbf{x}^*\|^2] = \mathcal{O}\left(n^{-\frac{1}{2}}\right)$. In Theorem 1 we provide a finite-time upper bound of $\mathrm{E}[\|\mathbf{x}_n - \mathbf{x}^*\|^2]$ which implies the result proved by Fabian (1967).

Remark 3 If we use Algorithm 1 to solve an offline stochastic program, we may consider using a central finite-difference gradient estimator. By Fabian (1967), if we set $a_n = \gamma n^{-1}$ and $c_n = \delta n^{-\frac{1}{6}}$, using a central finite-difference estimator may result in a better rate of convergence, that is, $\mathrm{E}[\|\mathbf{x}_n - \mathbf{x}^*\|^2] = \mathcal{O}\left(n^{-\frac{2}{3}}\right)$. However, in Remark 5, we show that using central finite-difference estimators does not lower the growth rate of the expected cumulative regret.

Based on Theorem 1, we can prove the following theorem that provides a finite-time bound on expected cumulative regret of Algorithm 1.

Theorem 2 *Suppose that Algorithm* 1 *is used to solve Problem* (1) *and Assumptions* 1-4 *are satisfied. Let $a_n = \gamma n^{-1}$ and $c_n = \delta n^{-\frac{1}{4}}$ with $1/(4B_1) < \gamma < 1/(2B_1)$ and $\delta > 0$. Then, there exists constant $\kappa_1 > 0$ and $\kappa_2 > 0$ such that $R(T) \leq \kappa_1\sqrt{T} + \kappa_2$ for all $T = 1, 2, \ldots$.*

Proof Notice that, in each iteration of Algorithm 1, $d+1$ solutions, $\mathbf{x}_n, \mathbf{x}_n + c_n\mathbf{e}_1, \ldots, \mathbf{x}_n + c_n\mathbf{e}_d$, need to be evaluated. Let $r(n)$ denote the expected regret of iteration n, $n = 1, 2, \ldots$. Then,

$$\begin{aligned} r(n) &= \mathrm{E}[F(\mathbf{x}_n, \xi_{n,0}) + F(\mathbf{x}_n + c_n\mathbf{e}_1, \xi_{n,1}) \\ &\quad + \cdots + F(\mathbf{x}_n + c_n\mathbf{e}_d, \xi_{n,d})] - (d+1)f(\mathbf{x}^*) \\ &= \mathrm{E}[f(\mathbf{x}_n) - f(\mathbf{x}^*)] + \mathrm{E}[f(\mathbf{x}_n + c_n\mathbf{e}_1) - f(\mathbf{x}^*)] \\ &\quad + \cdots + \mathrm{E}[f(\mathbf{x}_n + c_n\mathbf{e}_d) - f(\mathbf{x}^*)]. \end{aligned} \tag{13}$$

Notice that $\nabla f(\mathbf{x}^*) = \mathbf{0}$ (by Assumption 4) and $\|\nabla^2 f(\mathbf{x})\| \leq B_2$ for all $\mathbf{x} \in \Omega$ (by Equation (2)). Then, we have

$$\begin{aligned} f(\mathbf{x}_n) - f(\mathbf{x}^*) &\leq \frac{1}{2}B_2\|\mathbf{x}_n - \mathbf{x}^*\|^2, \\ f(\mathbf{x}_n + c_n\mathbf{e}_i) - f(\mathbf{x}^*) &\leq \frac{1}{2}B_2\|\mathbf{x}_n + c_n\mathbf{e}_i - \mathbf{x}^*\|^2 \\ &\leq B_2(\|\mathbf{x}_n - \mathbf{x}^*\|^2 + c_n^2), \quad i = 1, \ldots, d. \end{aligned}$$

Then, by Equation (13) and Theorem 1,

$$\begin{aligned} r(n) &\leq \left(\frac{1}{2} + d\right) B_2 b_n + dB_2 c_n^2 \\ &\leq \left(\frac{2d+1}{2}\lambda + d\delta^2\right) B_2\ n^{-\frac{1}{2}}. \end{aligned} \tag{14}$$

Then,

$$\begin{aligned} R(T) &\leq \sum_{n=1}^{\lceil T/(d+1)\rceil} r(n) = \left(\frac{2d+1}{2}\lambda + d\delta^2\right) B_2 \sum_{n=1}^{\lceil T/(d+1)\rceil} n^{-\frac{1}{2}} \\ &\leq \left(\frac{2d+1}{2}\lambda + d\delta^2\right) B_2 \int_0^{(T+d)/(d+1)} x^{-\frac{1}{2}} dx \\ &\leq \frac{[(2d+1)\lambda + 2d\delta^2]B_2}{\sqrt{d+1}}(\sqrt{T} + \sqrt{d}) \\ &\leq 2(\lambda + \delta^2)\sqrt{d+1} \cdot \sqrt{T} + 2(\lambda + \delta^2)(d+1). \end{aligned}$$

This concludes the proof of the theorem with $\kappa_1 = 2(\lambda + \delta^2)\sqrt{d+1}$ and $\kappa_2 = 2(\lambda + \delta^2)(d+1)$. ∎

Remark 4 Cope (2009) proves that $R(T) = \mathcal{O}(\sqrt{T})$. In Theorem 2 we provide a finite-time upper bound of $R(T)$ which implies the result of Cope (2009).

Remark 5 In Remark 3, we point out that using a central finite-difference gradient estimator may result in a better rate of convergence for $\mathrm{E}[\|\mathbf{x}_n - \mathbf{x}^*\|^2]$ if setting $c_n = \delta n^{-\frac{1}{6}}$. Notice that, by Equation (14), $r(n) \leq \nu_1\mathrm{E}[\|\mathbf{x}_n - \mathbf{x}^*\|^2] + \nu_2 c_n^2$ for some $\nu_1 > 0$ and $\nu_2 > 0$. Then, the regret $r(n) = \mathcal{O}(\max\{\mathrm{E}[\|\mathbf{x}_n - \mathbf{x}^*\|^2], c_n^2\})$. If using a central finite-difference estimator with $c_n = \delta n^{-\frac{1}{6}}$, $r(n)$ will be dominated by c_n^2 and $r(n) = \mathcal{O}\left(n^{-\frac{1}{3}}\right)$, causing $R(T) = \mathcal{O}\left(T^{\frac{2}{3}}\right)$, which is not as good as $\mathcal{O}(\sqrt{T})$. Therefore, even when using a central finite-difference gradient estimator, we also need to set $c_n = \delta n^{-\frac{1}{4}}$, which leads to the same $\mathcal{O}(\sqrt{T})$ of the regret.

3 | MULTI-PRODUCT DYNAMIC PRICING

Consider a firm that sells d different products over multiple periods. The firm may set the prices of the products, denoted by $\mathbf{p} = (p_1, \ldots, p_d)^{\mathrm{T}}$, at the beginning of each period and observe the random demands, denoted by $\mathbf{D}(\mathbf{p}) =$

$(D_1(\mathbf{p}), \ldots, D_d(\mathbf{p}))^{\mathrm{T}}$. The revenue of the period is then

$$\Theta(\mathbf{p}) = \mathbf{p}^{\mathrm{T}}\mathbf{D}(\mathbf{p}) = \sum_{i=1}^{d} p_i D_i(\mathbf{p}).$$

The objective of the firm is to choose the price $\mathbf{p}$ from a set of permissible prices, denoted by Ω, to maximize the expected revenue, that is,

$$\max_{\mathbf{p}\in\Omega}\{\theta(\mathbf{p}) := \mathrm{E}[\Theta(\mathbf{p})]\}. \tag{15}$$

A difficulty in solving Problem (15) is that the distribution of $\mathbf{D}(\mathbf{p})$ (and also the expected demand function $\mathrm{E}[\mathbf{D}(\mathbf{p})]$) may be unknown. Therefore, Problem (15) cannot be solved directly. Instead, one may design a pricing policy to determine the prices at the beginning of each period, based on the information up to the period, and observe the corresponding demands and the revenue. Let Ψ be a pricing policy which sets $\mathbf{p}_t = \Psi(\mathbf{p}_1, \ldots, \mathbf{p}_{t-1}, \mathbf{D}_1(\mathbf{p}_1), \ldots, \mathbf{D}_{t-1}(\mathbf{p}_{t-1}))$. Our goal is to find a pricing policy Ψ that maximizes the expected cumulative revenue of the first T periods, which is equivalent to minimize the expected cumulative regret

$$R(T,\Psi) = \sum_{t=1}^{T} \mathrm{E}[\theta(\mathbf{p}^*) - \theta(\mathbf{p}_t)], \tag{16}$$

where $\mathbf{p}^*$ is the true optimal price but unknown to us.

In this note we take a nonparametric approach that does not assume $\mathbf{D}(\mathbf{p})$ to follow any parametric model. Notice that our formulation is extremely general. It allows different cross elasticities among the demands, either complementary or substitutable, and it includes the parametric demand models of Broder and Rusmevichientong (2012), den Boer and Zwart (2014), and Keskin and Zeevi (2014) and the single-product nonparametric demand model of Besbes and Zeevi (2015) as special cases. Then, by den Boer and Zwart (2014) and Keskin and Zeevi (2014), we know that $R(T,\Psi)$ grows at least in the order of $\sqrt{T}$ for any feasible policy Ψ.

Based on the KWSA algorithm (ie, Algorithm 1), we propose the KW pricing policy (ie, Algorithm 2), denoted by Ψ^{KW}.

Notice that the KW pricing policy is very different from the existing policies in the literature. As pointed out by Keskin and Zeevi (2014), *complete learning* is critical to the pricing policies proposed by Broder and Rusmevichientong (2012), den Boer and Zwart (2014), and Keskin and Zeevi (2014). Here the complete learning means that the parameters of the demand models need to be learned consistently, that is, the parameter estimators converge to the true values as the number of periods goes to infinity. For the nonparametric approach proposed by Besbes and Zeevi (2015), complete learning of the demand function is not necessary. Nevertheless, the derivative estimator of the expected demand function still needs to be learned consistently. To achieve that, Besbes and Zeevi (2015) force the iterations become longer and longer, so that the finite-difference derivative estimator becomes more and more accurate. In the KW pricing policy, however, the length of the iteration is always fixed (ie, $d+1$) and the finite-difference gradient estimator is always noisy and inconsistent. To ensure convergence, the KW pricing policy uses a *slow optimizing strategy*, only changing the price vector by a size of a_n along the estimated (noisy) gradient direction and allowing the errors to cancel out in the limit, and the strategy leads to an asymptotically optimal pricing policy (as shown in Theorem 3). In contrast, pricing policies of Broder and Rusmevichientong (2012), den Boer and Zwart (2014), Keskin and Zeevi (2014), and Besbes and Zeevi (2015) all use an *aggressive optimizing strategy*, moving the price directly to the myopically optimal solution (with some adjustments to avoid incomplete learning). With the aggressive optimizing strategy, their policies need complete learning of either the demand function or the derivative. We use a multi-product linear demand model to illustrate the differences between the aggressive optimizing and the slow optimizing strategies in Section 4.4.

ALGORITHM 2 KW pricing policy

Initialization. Let $\mathbf{p}_0 \in \Omega$ by a starting price vector. Let the iteration counter $n = 1$ and period counter $t = 0$.

Step 1. Pricing and information collection.

- Let $t = t+1$. Set $\widetilde{\mathbf{p}}_t = \mathbf{p}_n$ and observe $\Theta_t = \Theta_t(\widetilde{\mathbf{p}}_t)$;
- For $i = 1, 2, \ldots, d$,

  Let $t = t+1$. Set $\widetilde{\mathbf{p}}_t = \mathbf{p}_n + c_n\mathbf{e}_i$ and observe $\Theta_t = \Theta_t(\widetilde{\mathbf{p}}_t)$.

  End the for-loop.

Step 2. Updating.

Let

$$\mathbf{p}_{n+1} = \Pi_\Omega(\mathbf{p}_n + a_n\mathbf{G}(\mathbf{p}_n)),$$

where Π_Ω is a projection operator onto the set Ω, that is, $\Pi_\Omega(\mathbf{p}) = \operatorname{argmin}_{\mathbf{p}'\in\Omega}\|\mathbf{p}-\mathbf{p}'\|$, and

$$\mathbf{G}(\mathbf{p}_n) = \frac{1}{c_n}[(\Theta_{k+2} - \Theta_{k+1}), \ldots, (\Theta_{k+d+1} - \Theta_{k+1})]^{\mathrm{T}}$$

with $k = n(d+1)$. (Note that the index k introduced here is only for notational simplification of the expressions of Θ_t, where $t = k+1, \ldots, k+d+1$. In order to make the expression of $\mathbf{G}(\mathbf{p}_n)$ in Algorithm 2 consistent with that of $\mathbf{G}(\mathbf{x}_n)$ Algorithm 1, in fact, we can rewrite $\mathbf{G}(\mathbf{x}_n) = \frac{1}{c_n}([F(\widetilde{\mathbf{x}}_{k+2}, \xi_{n,1}) - F(\widetilde{\mathbf{x}}_{k+1}, \xi_{n,0})], \ldots, [F(\widetilde{\mathbf{x}}_{k+d+1}, \xi_{n,d})$ $([-F(\widetilde{\mathbf{x}}_{k+1}, \xi_{n,0})])^{\mathrm{T}}$, with $k = n(d+1)$.) Let $n = n+1$ and go back to Step 1.

The following theorem establishes a finite-time upper bound of the expected cumulative regret for the KW pricing policy when solving the multi-product dynamic pricing problem. It is based on Theorems 1 and 2.

Theorem 3 *Suppose that the KW pricing policy is used to solve Problem* (15) *and that the following assumptions hold*:

1. $\Omega \subset \mathfrak{R}^d$ *is a convex and compact set and* $\mathbf{p}^* \in \mathrm{int}(\Omega)$;

2. $\mathrm{E}[\mathbf{D}(\mathbf{p})]$ *is twice continuously differentiable in* Ω *and* $\max_{\mathbf{p}\in\Omega}\mathrm{E}[\|\mathbf{D}(\mathbf{p})\|^2]<\infty$;
3. $\theta(\mathbf{p})$ is strongly concave.

Then, there exist constants $\lambda>0$, $\kappa_1>0$, *and* $\kappa_2>0$ *such that* $\mathrm{E}(\|\mathbf{p}_n-\mathbf{p}^*\|^2)\leq\lambda\, n^{-\frac{1}{2}}$ *for all* $n=1,2,\ \ldots$ *and* $R(T,\Psi^{\mathrm{KW}})\leq\kappa_1\sqrt{T}+\kappa_2$ *for all* $T=1,2,\ \ldots.$

By the lower bounds of the expected cumulative regret established by Broder and Rusmevichientong (2012) and Keskin and Zeevi (2014), Theorem 3 essentially shows that the KW pricing policy is asymptotically optimal. It also shows that the conjecture of Besbes and Zeevi (2015) is true not only for single-product dynamic pricing problems but also for multi-product dynamic pricing problems. It further confirms the surprising finding of Besbes and Zeevi (2015) that, in terms of the asymptotic growth rate of the expected cumulative regret, the nonparametric approach may be as good as the parametric one. Nevertheless, the nonparametric approach may avoid model misspecifications that always exist in parametric models.

4 | NUMERICAL EXPERIMENTS

In this section, we conduct numerical experiments to show the practical performance of the KW pricing policy under single- or multi-product linear or nonlinear demand models. We first demonstrate that the cumulative regret indeed grows in the order of $\sqrt{T}$ in Section 4.1. Ideally, we would like to compare our KW pricing policy with other nonparametric multi-product pricing policies. However, there are no such policies. Therefore, we compare to the multiple-product *parametric* policies of Keskin and Zeevi (2014) under their parametric demand models in Section 4.2. However, we do not assume that the parametric forms of the demand models are known to the KW policy. In Section 4.3, we compare the KW pricing policy to the *single-product* nonparametric policy of Besbes and Zeevi (2015). Finally, in Section 4.4, we illustrate the slow optimizing strategy by a comparison among the KW policy, Greedy ILS (iterated least squares) policy and the MCILS (multivariate constrained iterative least squares) policy of Keskin and Zeevi (2014).

4.1 | Illustration of the rate optimality of the total regret

We adopt the same multi-product linear demand model used in Keskin and Zeevi (2014), in which the random demand for product i in period t, denoted by $\mathbf{D}_t(\mathbf{p}_t)=(D_{1t}(\mathbf{p}_t),\ \ldots,D_{dt}(\mathbf{p}_t))^{\mathrm{T}}$, has the following linear form,

$$D_{it}(\mathbf{p}_t)=\alpha_i+\beta_i\cdot\mathbf{p}_t+\epsilon_{it},\quad \text{for}\quad i=1,\ \ldots,d,\quad \text{and}\quad t=1,2,\ \ldots,$$

where $\mathbf{p}_t=(p_{1t},\ \ldots,p_{dt})^{\mathrm{T}}$ is the price for all products at period t, and $\alpha_i\in\Re$, $\beta_i=(\beta_{i1},\ \ldots,\beta_{id})\in\Re^d$ are the unknown parameters, and ϵ_{it} are independent and identically distributed (i.i.d.) normal random variables with mean zero and variance σ^2. We assume that the compact set of all feasible prices is $\Omega=\{\mathbf{p}:\mathbf{p}\in[l_1,u_1]\times[l_2,u_2]\times\cdots\times[l_d,u_d]\}$, where $0\leq l_i<u_i<\infty$ for $i=1,2,\ \ldots,d$. In the matrix form, the demand model can be expressed as follows,

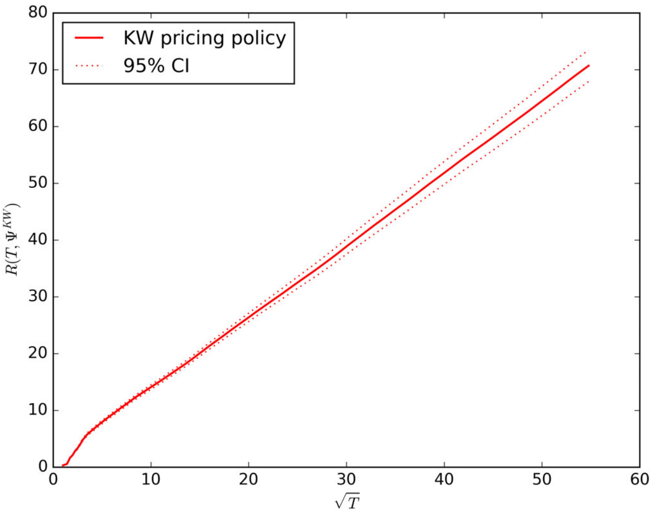

FIGURE 1 Average cumulative regret of the KW pricing policy [Colour figure can be viewed at wileyonlinelibrary.com]

$$\mathbf{D}_t(\mathbf{p}_t)=\alpha+\mathbf{B}\mathbf{p}_t+\epsilon_t,\quad \text{for}\quad t=1,2,\ \ldots,$$

where $\epsilon_t=(\epsilon_{1t},\ \ldots,\epsilon_{dt})^{\mathrm{T}}\in\Re^n$, $\alpha=(\alpha_1,\ \ldots,\alpha_d)^{\mathrm{T}}\in\Re^d$, and $\mathbf{B}$ is a $d\times d$ matrix

$$\mathbf{B}=\begin{bmatrix}\beta_{11}&\beta_{12}&\cdots&\beta_{1d}\\ \beta_{21}&\beta_{22}&\cdots&\beta_{2d}\\ \vdots&\vdots&\ddots&\vdots\\ \beta_{d1}&\beta_{d2}&\cdots&\beta_{dd}\end{bmatrix}.$$

Then, the optimal price $\mathbf{p}^*=-(\mathbf{B}+\mathbf{B}^{\mathrm{T}})^{-1}\alpha$.

In the following experiment, we all consider the 2-product case, that is, $d=2$. Other parameters are set as follows, $\alpha=(1.1,0.7)^{\mathrm{T}}$, $\mathbf{B}=\begin{bmatrix}-0.5&0.05\\0.05&-0.3\end{bmatrix}$, $\sigma^2=0.01$, and $[l_1,u_1]=[0.1,2.5]$, $[l_2,u_2]=[0.1,2.5]$. The initial price is $\mathbf{p}_0=(0.75,2.1)^{\mathrm{T}}$ and $\gamma=3$ and $\delta=1$ in the positive sequences of $\{a_n\}$ and $\{c_n\}$. Then the optimal price is $\mathbf{p}^*\approx(1.237,1.373)^{\mathrm{T}}$.

Figure 1 displays the average T-period regret $R(T,\Psi^{\mathrm{KW}})$ and its 95% confidence interval (CI) over 100 macro-replications with respect to $\sqrt{T}$, where $T=1,2,\ \ldots,3000$. From Figure 1, we find that, except for the cases of small T (roughly $T\leq 25$ in our setting), the expected cumulative regret of the KW pricing policy grows linearly in the order of $\sqrt{T}$, which verifies the performance guarantee obtained in Theorem 3.

4.2 | Comparison with multi-product parametric pricing policies

We consider the multi-product linear demand model and compare our KW pricing policy with two policies, that is, the ILS and MCILS policies, proposed by Keskin and

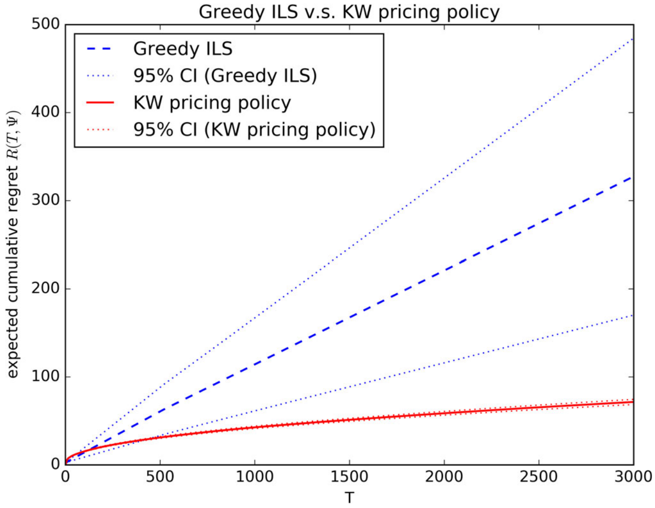

FIGURE 2 Comparisons between the ILS and KW pricing policies [Colour figure can be viewed at wileyonlinelibrary.com]

Zeevi (2014). The parameter settings are the same as used in Section 4.1.

Figure 2 displays the average T-period regret $R(T, \Psi^{\mathrm{KW}})$ and $R(T, \Psi^{\mathrm{ILS}})$, as well as their 95% CIs, respectively, over 100 macro-replications for $T = 1, 2, \ldots, 3000$. From Figure 2, we find that the cumulative regret for the KW policy increases slightly (in the order of $\sqrt{T}$) while that for the ILS policy increases linearly with respect to T. This finding demonstrates the phenomenon of incomplete learning, since the greedy ILS estimates may get stuck at some nonoptimal values without further exploration, resulting a constant fraction of revenue loss in each period, thus leading to a linearly increasing cumulative regret. However, in the KW pricing policy, during each iteration, we perturb the price by small amount for each single product once, observe the corresponding demand for all products, and then update the whole price using the stochastic gradient at the end of each iteration, which has inherently taken the tradeoff between exploration and exploitation into consideration. It is also worthwhile pointing out that, by a comparison of the 95% CI between the greedy ILS and KW pricing policies, KW policy seems more stable than ILS, which is not surprising since ILS often stops at different nonoptimal points, resulting a large variation of its solution quality.

To avoid the incomplete learning and achieve the asymptotic optimality, Keskin and Zeevi (2014) propose a MCILS(κ) policy, where κ is a threshold parameter, that adjusts the greedy ILS price whenever its deviation from the historical average price is not sufficiently large. By doing so, they show that MCILS(κ) policy achieves an $\mathcal{O}(\sqrt{T}\log T)$ of the cumulative regret. To compare the performances of the MCILS(κ) and KW pricing policy, we still consider a 2-product linear demand model under the same parameter settings as in Section 4.1. In order to implement the MCILS(κ) policy, we also need to choose the first three additional price vectors as $\mathbf{p}_1 = (2.0, 2.0)^{\mathrm{T}}$, $\mathbf{p}_2 = (2.0, 0.75)^{\mathrm{T}}$, $\mathbf{p}_3 = (0.75, 2.0)^{\mathrm{T}}$, and determine the value of the threshold κ, which is set as $\kappa = 0.8, 1.0, 1.2$, respectively.

Figure 3 presents the average T-period regret $R(T, \Psi^{\mathrm{KW}})$ and $R(T, \Psi^{\mathrm{MCILS}(\kappa)})$ with different κ, over 100 macro-replications for $T = 1, 2, \ldots, 300$. We notice that MCILS(κ) performs well without surprise because it is a specifically designed policy for the multi-product linear demand model. However, it is worthwhile pointing out that our nonparametric KW pricing policy is also competitive compared with the MCILS(κ) policy if the parameter κ is chosen improperly (eg, $\kappa = 1.2$ in our settings). Even through the $\mathcal{O}(\sqrt{T}\log T)$ of the growth rate of $R(T, \Psi^{\mathrm{MCILS}(\kappa)})$ will not change in terms of κ, the value of $R(T, \Psi^{\mathrm{MCILS}(\kappa)})$ varies significantly for different values of κ. In fact, we know that when κ is too large, MCILS(κ) reduces to the greedy ILS; meanwhile when κ is too small, the algorithm forces exploration in almost every step, causing a lot of computational efforts.

4.3 | Comparison with a single-product nonparametric pricing policy

In this subsection, we use the single-product general demand model of Besbes and Zeevi (2015) as follows. In each period, the seller chooses a price p_t from $\Omega = [l, u]$, and observes a demand response,

$$D_t(p_t) = \lambda(p_t) + \epsilon_t, \quad t = 1, 2, \ldots,$$

where $\lambda(\cdot)$ is a positive, strictly decreasing and twice continuously differentiable function, and ϵ_t, $t = 1, 2, \ldots$, are i.i.d. normal random variables with zero mean and variance σ^2.

Besbes and Zeevi (2015) propose a linear semimyopic policy, which is nonparametric and whose regret is upper bounded by the order of $\sqrt{T}(\log T)^2$. We compare the performances of the KW pricing policy and the linear semimyopic policy under three different demand models, that is, the linear, exponential, and logit models, as follows.

a. *Linear*: $\lambda(p) = (\alpha - \beta p)^+$, where $\alpha = 1$, $\beta = 0.5$, and $\sigma^2 = 0.05^2$. Let $\Omega = [l, u] = [0.5, 1.5]$, and then the optimal price is $p^* = 1$. Let $\hat{p}_1 = 0.7$, $I_i = 1$, $\delta_t = \rho t^{-1/4}$ where $\rho = 0.5$, and $\mathcal{T}_i = \{1, \ldots, t_{i+1}\}$ for the semimyopic policy; and let $p_0 = 0.7$, $\gamma = 3$, and $\delta = 1$ for the KW pricing policy.
b. *Exponential*: $\lambda(p) = \exp(\alpha - \beta p)$, where $\alpha = 1$, $\beta = 0.3$, and $\sigma^2 = 0.05^2$. Let $\Omega = [l, u] = [2.5, 3.5]$, and then the optimal price is $p^* = 3$. Let $\hat{p}_1 = 2.7$, $I_i = 1$, $\delta_t = \rho t^{-1/4}$ where $\rho = 0.2$, and $\mathcal{T}_i = \{1, \ldots, t_{i+1}\}$ for the semimyopic policy; and let $p_0 = 2.7$, $\gamma = 3$, and $\delta = 1$ for the KW pricing policy.
c. *Logit*: $\lambda(p) = \exp(\alpha - \beta p)/(1 + \exp(\alpha - \beta p))$, where $\alpha = 1$, $\beta = 0.3$, and $\sigma^2 = 0.05^2$. Let $\Omega = [l, u] = [3, 7]$, and then the ptimal price is

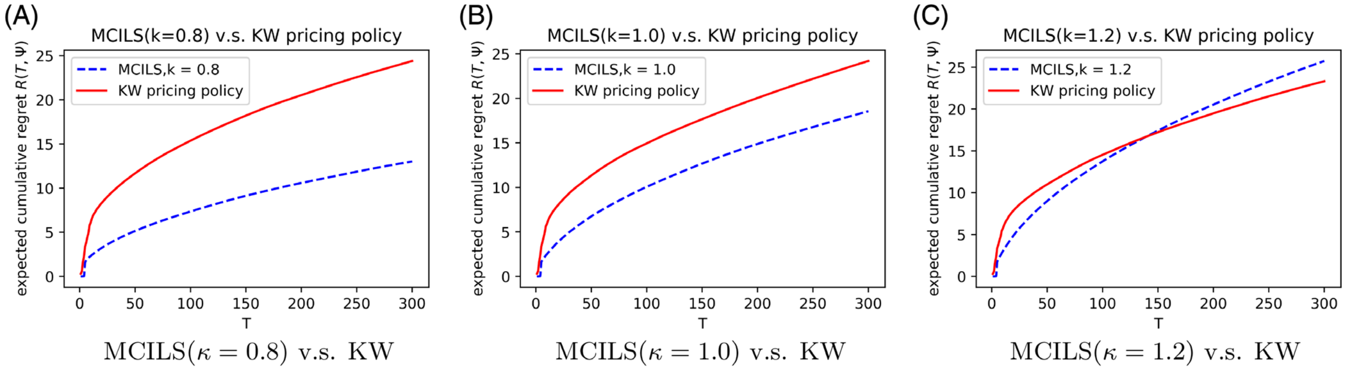

FIGURE 3 Comparisons between the MCILS(κ) and KW pricing policies. (A) MCILS($\kappa = 0.8$) versus KW. (B) MCILS($\kappa = 1.0$) versus KW. (C) MCILS($\kappa = 1.2$) versus KW [Colour figure can be viewed at wileyonlinelibrary.com]

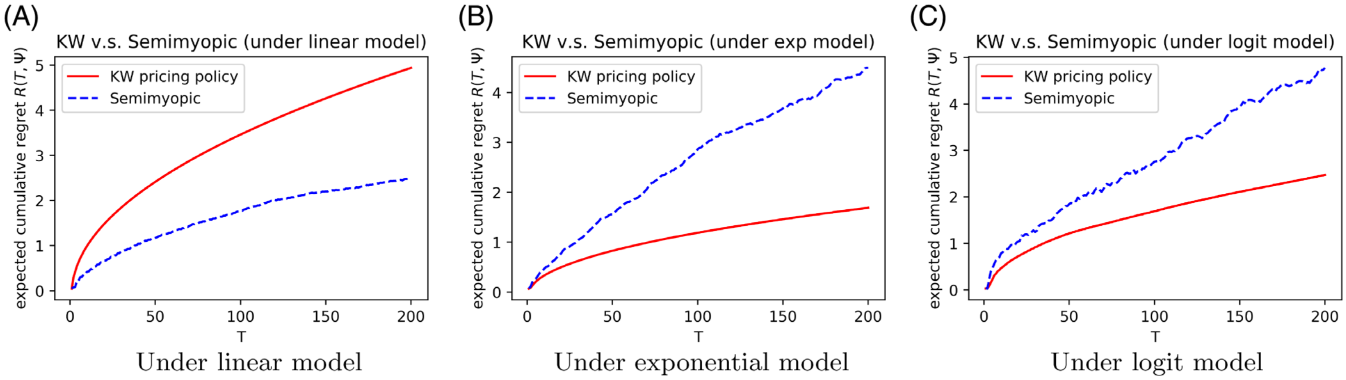

FIGURE 4 Comparisons between the semimyopic and KW pricing policies. (A) Under linear model. (B) Under exponential model [Colour figure can be viewed at wileyonlinelibrary.com]

$p^* \approx 5.2238$. Let $\hat{p}_1 = 4.5$, $I_i = 1$, $\delta_t = \rho t^{-1/4}$ where $\rho = 0.5$, and $\mathcal{T}_i = \{1, \ldots, t_{i+1}\}$ for the semimyopic policy; and let $p_0 = 4.5$, $\gamma = 10$, and $\delta = 1$ for the KW pricing policy.

Figure 4 presents the averaged cumulative regrets of the KW pricing policy and the semimyopic policy under linear, exponential and logit demand models, respectively, over 100 macro-replications for $T = 1, 2, \ldots, 200$. From Figure 4, we observe that the semimyopic policy outperforms the KW pricing policy when the underlying model is indeed linear, which is not surprising because the semimyopic policy presumes that the underlying demand model as linear and estimates the parameters based on a linear curve. However, when the underlying demand model is nonlinear (eg, exponential or logit), the performance of the KW pricing policy is generally better than that of the semimyopic policy, which indicates that the KW pricing policy may have advantages when applied to the dynamic pricing context with nonlinear demand models.

4.4 | Illustration of the aggressive and slow optimizing strategies

In Section 3, we mention that many existing pricing policies use an aggressive optimizing strategy, while the KW pricing policy uses a slow optimizing strategy. In this subsection, we continue to use the 2-product linear demand model in Section 4.2 and to compare the KW pricing policy with the ILS policy and MCILS ($\kappa = 1.2$) policy. The parameter settings remain the same.

In Figure 5, the horizontal axis p_1 and the vertical axis p_2 denote the prices of the two products, respectively, and the star denotes the theoretical optimal price $\mathbf{p}^*$. We present the scatter plots of the price evaluation points determined by ILS, MCILS, and KW pricing policies along different time periods during one macro-simulation run. In particular, we choose the same time period intervals for three policies, namely, $t \in [0, 2000]$, [8000, 10 000], and [28 000, 30 000].

From the top three subplots in Figure 5, we observe that the price points evaluated by the greedy ILS policy converge very quickly during the first time interval, which reflects the term of "greedy" in its name, but soon get stuck at a nonoptimal price point. This is the so-called incomplete learning phenomenon. In the middle three subplots, we observe that there exists a clear circular trajectory around the optimal price point. The radius of the circle diminishes as time period increases. Eventually, the circular trajectories gradually shrink to the optimal point. That is because MCILS policy enforces some computational effort for exploration to slow down the active learning speed towards the optimal directions to avoid incomplete learning. In the bottom three subplots, we observe that the price points evaluated by the KW policy appear three clusters of cloud shadows. The clusters will become more concentrated and closer to the optimal price

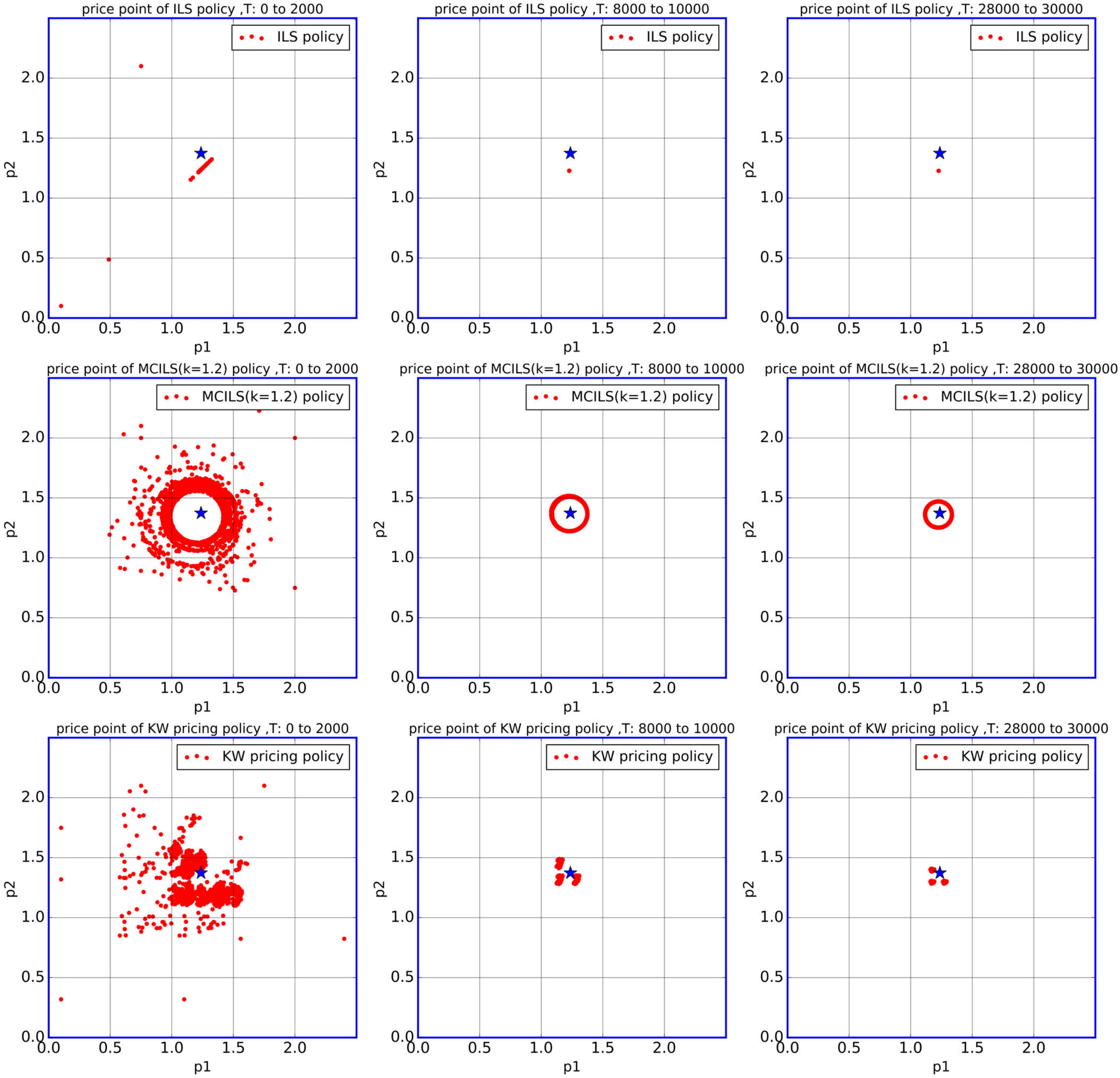

FIGURE 5 Illustration of the aggressive and slow optimizing strategies [Colour figure can be viewed at wileyonlinelibrary.com]

point over time, which typically reflects the inactive learning (or slow optimizing, in other words). It is worth mentioning that although slow optimizing strategy seems to make the evaluation points wonder around in early time periods without any clear trajectory, it will effectively reduce the scope of exploration and concentrate around the optimal solution over time.

5 | CONCLUSIONS

We study a nonparametric multi-product dynamic pricing problem with demand learning, and formulate it as an online stochastic optimization problem. We propose a nonparametric KW pricing policy, which is a variant of the classical KWSA algorithm, and show that the corresponding expected cumulative regret is upper bounded by $\kappa_1\sqrt{T} + \kappa_2$ for all $T = 1, 2, \ldots,$ where κ_1, κ_2 are positive constants and T is the number of time periods. Therefore, the KW pricing policy achieves the optimal $\mathcal{O}(\sqrt{T})$ order of regret. Finally, we conduct numerical studies to study the performance of the KW pricing policy and compare with other polices under various demand models.

There are some possible extensions of this work. First, the KWSA algorithm requires evaluating $d+1$ points at each iteration, which may slow the algorithm significantly when the dimension is high. When the dimension is high, that is, there are many products whose prices need to be determined by the decision maker, it would be better to use some other online optimization algorithms, such as the simultaneous perturbation stochastic approximation algorithm of Spall (1992), to remove the effect of the dimension. Second, there exists certain substitution or complementary effects among different products, resulting in different correlation structures among different products. Then, how to better utilize this type of correlation information to design dynamic pricing and learning policies is another interesting research direction. Third, the market environments may change over time, which may require a nonstationary optimal pricing strategy, then the nonstationary stochastic optimization algorithms, for example, the ones in Besbes et al. (2015) and Keskin and Zeevi (2016), may be adopted to solve this type of problems.

ACKNOWLEDGMENTS

The authors would like to thank the Department Editor, Prof Rene Caldentey, the Associate Editor, and two reviewers for their insightful and detailed comments that have significantly improved this paper. This research of the first author was supported in part by the Hong Kong Research Grants Council (GRF 11504017) and the Natural Science Foundation of China (Grant 71991473), and this research of the third author was supported in part by the Natural Science Foundation of China (Grants 71722006 and 71531010).

ORCID

L. Jeff Hong https://orcid.org/0000-0001-7011-4001
Jun Luo https://orcid.org/0000-0002-3639-916X

REFERENCES

Agarwal, A., Dekel, O., & Xiao, L. (2010). Optimal algorithms for online convex optimization with multi-point bandit feedback. *COLT*, 28–40.

Agarwal, A., Foster, D., Hsu, D., Kakade, S., & Rakhlin, A. (2013). Stochastic convex optimization with bandit feedback. *SIAM Journal on Optimization*, *23*(1), 213–240.

Aviv, Y., & Vulcano, G. (2012). *Dynamic list pricing*. In Ö. Özer & R. Phillips (Eds.), The Oxford handbook of pricing management (pp. 522–584). Oxford, England: Oxford University Press.

Benveniste, A., Priouret, P., & Metivier, M. (1990). Adaptive algorithms and stochastic approximations. New York: Springer.

Besbes, O., Gur, Y., & Zeevi, A. (2015). Non-stationary stochastic optimization. *Operations Research*, *63*(5), 1227–1244.

Besbes, O., & Zeevi, A. (2015). On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. *Management Science*, *61*, 723–739.

Broder, J., & Rusmevichientong, P. (2012). Dynamic pricing under a general parametric choice model. *Operations Research*, *60*, 965–980.

Cope, E. W. (2009). Regret and convergence bounds for a class of continuum-armed bandit problems. *IEEE Transactions on Automatic Control*, *54*, 1243–1253.

den Boer, A. V., & Zwart, B. (2014). Simultaneously learning and optimizing using controlled variance pricing. *Management Science*, *60*, 770–783.

den Boer, A. V. (2015). Dynamic pricing and learning: Historical origins, current research, and new directions. *Surveys in Operations Research and Management Science*, *20*(1), 1–18.

Fabian, V. (1967). Stochastic approximation of minima with improved asymptotic speed. *Annals of Mathematical Statistics*, *38*, 191–200.

Harrison, J. M., Keskin, N. B., & Zeevi, A. (2012). Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. *Management Science*, *58*(3), 570–586.

Keskin, N. B., & Zeevi, A. (2014). Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations Research*, *62*, 1142–1167.

Keskin, N. B., & Zeevi, A. (2016). Chasing demand: Learning and earning in a changing environment. *Mathematics of Operations Research*, *42*(2), 277–307.

Keskin, N. B., & Zeevi, A. (2018). On incomplete learning and certainty-equivalence control. *Operations Research*, *66*(4), 1136–1167.

Kiefer, J., & Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, *23*, 462–466.

Kushner, H., & Yin, G. (2003). Stochastic approximation and recursive algorithms and applications. New York: Springer-Verlag.

Lobo, M. S., & Boyd, S. (2003). *Pricing and learning with uncertain demand* (Working Paper). Duke University, Durham, NC.

Nemirovski, A., Juditsky, A., Lan, G., & Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, *19*, 1574–1609.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, *22*, 400–407.

Shamir, O. (2013). *On the complexity of bandit and derivative-free stochastic convex optimization*. In Conference on learning theory (pp. 3–24).

Spall, J. C. (1992). Multivariate stochastic approximation using simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, *37*, 332–341.

Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. *ICML*, 928–936.

How to cite this article: Hong LJ, Li C, Luo J. Technical note: Finite-time regret analysis of Kiefer-Wolfowitz stochastic approximation algorithm and nonparametric multi-product dynamic pricing with unknown demand. *Naval Research Logistics* 2020;1–12. https://doi.org/10.1002/nav.21902

APPENDIX: EXISTENCE OF λ OF EQUATION (12)

Let $\alpha = 2\gamma B_1$, $\beta = \gamma\delta B_2\sqrt{d}$, and $\omega = \frac{4d\gamma^2M^2}{\delta^2}$. By Equation (11), we have

$$b_{n+1} \leq \left(1 - \frac{\alpha}{n}\right) b_n + \beta n^{-\frac{5}{4}} \sqrt{b_n} + \omega n^{-\frac{3}{2}}. \tag{A1}$$

Let $\lambda = \max\{b_1, \lambda_0\}$, where

$$\lambda_0 = \left(\frac{\beta + \sqrt{\beta^2 + 2\omega(2\alpha - 1)}}{2\alpha - 1}\right)^2$$

and, because $2\alpha - 1 > 0$, λ_0 also satisfies

$$(2\alpha - 1)\kappa - 2\beta\sqrt{\kappa} - 2\omega \geq 0, \qquad \forall \kappa \geq \lambda_0. \tag{A2}$$

We prove by induction that $b_n \leq \lambda n^{-\frac{1}{2}}$. It is easy to see that it holds for $n = 1$. For any $n = 1, 2, \ldots$, suppose that $b_n \leq \lambda n^{-\frac{1}{2}}$. Then, by Equation (A1) and because $1 - \alpha/n > 0$ due to $\alpha < 1$,

$$\begin{aligned} b_{n+1} &\leq \left(1 - \frac{\alpha}{n}\right) \lambda n^{-\frac{1}{2}} + \beta\sqrt{\lambda} n^{-3/2} + \omega n^{-\frac{3}{2}} \\ &= \lambda n^{-\frac{1}{2}} - (\alpha\lambda - \beta\sqrt{\lambda} - \omega) n^{-\frac{3}{2}} \\ &= \lambda n^{-\frac{1}{2}} - \frac{\lambda}{2} n^{-\frac{3}{2}} - \frac{1}{2}[(2\alpha - 1)\lambda - 2\beta\sqrt{\lambda} - 2\omega] n^{-\frac{3}{2}} \\ &\leq \lambda\left(n^{-\frac{1}{2}} - \frac{1}{2} n^{-\frac{3}{2}}\right), \end{aligned} \tag{A3}$$

where the last inequality follows from Equation (A2) and the fact that $\lambda = \max\{b_0, \lambda_0\} \geq \lambda_0$. Let $g(x) = x^{-\frac{1}{2}}$. Then, $g'(x) = -\frac{1}{2} x^{-\frac{3}{2}}$. Notice that $g(x)$ is convex. Then,

$$g(x') - g(x) \geq g'(x)(x' - x).$$

Then,

$$(n + 1)^{-\frac{1}{2}} - n^{-\frac{1}{2}} = g(n + 1) - g(n) \geq g'(n) = -\frac{1}{2} n^{-\frac{3}{2}}.$$

Therefore,

$$n^{-\frac{1}{2}} - \frac{1}{2} n^{-\frac{3}{2}} \leq (n + 1)^{-\frac{1}{2}}.$$

Then, by Equation (A3), we have $b_{n+1} \leq \lambda(n + 1)^{-\frac{1}{2}}$. This concludes the induction proof and, therefore, $b_n \leq \lambda n^{-\frac{1}{2}}$ for all $n = 1, 2, \ldots$.