Learning the Stackelberg Equilibrium in a Newsvendor Game

Nicolò Cesa-Bianchi
Università degli Studi di Milano
and Politecnico di Milano
Milan, Italy
nicolo.cesa-bianchi@unimi.it

Tommaso Cesari
University of Ottawa
Ottawa, Canada
tcesari@uottawa.ca

Takayuki Osogami
IBM Research – Tokyo
Tokyo, Japan
osogami@jp.ibm.com

Marco Scarsini
Luiss University
Rome, Italy
marco.scarsini@luiss.it

Segev Wasserkrug
IBM Research – Israel
Haifa, Israel
segevw@il.ibm.com

ABSTRACT

We study a repeated newsvendor game between a supplier and a retailer who want to maximize their respective profits without full knowledge of the problem parameters. After characterizing the uniqueness of the Stackelberg equilibrium of the stage game with complete information, we show that even with partial knowledge of the joint distribution of demand and production cost, natural learning dynamics guarantee convergence of the supplier and retailer's joint strategy profile to the Stackelberg equilibrium of the stage game. We also prove finite-time bounds on the supplier's regret and asymptotic bounds on the retailer's regret, where the specific rates depend on the type of knowledge preliminarily available to the players. Finally, we empirically confirm our theoretical findings on synthetic data.

KEYWORDS

Regret minimization; Supply chain analysis; Newsvendor game; Online learning; Stackelberg equilibrium

ACM Reference Format:
Nicolò Cesa-Bianchi, Tommaso Cesari, Takayuki Osogami, Marco Scarsini, and Segev Wasserkrug. 2023. Learning the Stackelberg Equilibrium in a Newsvendor Game. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 9 pages.

1 INTRODUCTION

The newsvendor problem is a central topic in inventory theory and, more generally, in the analysis of supply chains. In its classical version [2], a retailer orders a certain quantity of a perishable good from a supplier. The decision of how much to order is made before the realization of the unknown demand for the good. If all costs are linear, the optimal decision is a quantile of the demand distribution that depends on the parameters of the model (i.e., the wholesale price and the retail market price). Even in this simple framework, the retailer can compute the optimal quantity only if the demand distribution and the model parameters are known.

In a practical newsvendor scenario, it is rarely the case that the retailer is the only decision maker. For instance, the wholesale price could be determined by a supplier incurring an exogenous production cost that is unknown to the retailer. These multi-agent versions of the newsvendor problem can be analyzed via a game-theoretic approach, where the optimal choices are expressed in terms of equilibria of a game. As it is unreasonable to assume that supplier and retailer have full knowledge of the distributions of production cost and demand, it is important to find strategies that perform well even when only partial knowledge of these quantities is available to each player.

1.1 Our Contribution

In this paper we consider a newsvendor model where players do not have full knowledge of the relevant parameters. The problem is modeled as a repeated game between a supplier and a retailer whose goal is to earn, in the long term, at least as much as their utility at a Stackelberg equilibrium (SE).

In Section 2, we start by considering the stage game with complete information, which we model as a Stackelberg game. Here, the supplier chooses the wholesale price and reveals it to the retailer, who in turn chooses the quantity to order by solving a newsvendor problem. Under weak conditions (Assumption 1) on the joint distribution of production cost C, retail price P, and demand D, we characterize (Theorem 2.1) the uniqueness of the SE of the game, which we show to be in pure strategies.

In Section 3.1, we consider a repeated game in which the supplier only knows the marginal distribution of C, and the retailer only knows the distribution of (P, D). Assuming that, at each time t, the retailer chooses the quantity q_t by best-responding to the supplier's choice of wholesale price w_t, in Theorem 3.1 we show that the supplier's time-averaged expected utility after T interactions converges to that of the SE at rate $T^{-1/2}$, while the players' strategy profile (w_t, q_t) converges to the SE of the stage game asymptotically at the same $T^{-1/2}$ rate. If the supplier is given an upper bound on the Lipschitz constant of their own utility, then their time-average expected utility converges to that of the SE at a faster rate $(\ln T)/T$. Notably, in this case the rate of convergence to the SE $(w^\star, q^\star)$ of the players' strategy profile can be arbitrarily slow (Theorem 3.3).

Our most interesting contributions are in Section 3.2, where we drop the assumption that the supplier and retailer know the distributions of C and (P, D), respectively. We analyze the game dynamics when no *a priori* distributional knowledge is given to any player and they are both required to learn the relevant information through regret-minimization techniques. If the supplier runs

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

Explore-Then-Commit (ETC, Algorithm 4) and the retailer runs Follow-The-Leader (FTL, Algorithm 3), we show that the supplier's time-averaged expected utility converges to that of the SE at rate $T^{-1/3}$ and, the retailer's time-averaged expected utility converges to that at the SE, and the player's strategy profile (w_t, q_t) converges to $(w^\star, q^\star)$ with probability 1. Finally, in Section 4, we run some experiments on synthetic data to gain additional insights. Our experiments reveal how our algorithms can sometimes outperform the SE, and the role that discretization plays in the system's dynamics.

Our analyses of convergence of the learning dynamics combine ideas from online learning and bandits (e.g., ETC and FTL) with tools from zeroth-order optimization (e.g., the Piyavskii–Shubert algorithm).

1.2 Related Work

The newsvendor problem, also known as the newsboy problem, goes back to Edgeworth [10]. The formalization used in this work is due to Arrow et al. [2]; we refer the reader to the handbook [8] for a survey of the many variants of Arrow's model.

A game-theoretic formulation of the newsvendor problem with competing retailers is proposed by Parlar [17]—see also [13, 14, 16]. Wang and Gerchak [22] use a Stackelberg game to model a situation where an assembler has to buy components from different suppliers. Lariviere and Porteus [12] study a model where a supplier and a retailer interact through a price-only contract, and compare its efficiency with the efficiency of an integrated system. Adida and DeMiguel [1] consider a competitive inventory model with several suppliers and several retailers, and prove equilibrium uniqueness under some symmetry conditions. We refer the reader to Cachon and Netessine [6] for a survey of the literature on game-theoretic models in supply chain analysis and to Silbermayr [21] for a more recent and specific survey on newsvendor games.

The problem of learning equilibria is investigated by Balcan et al. [4] for Stackelberg security games. They prove bounds on the leader's regret when the follower has a type that changes over time in a known and finite class. Their results hold both in the full information setting (where the leader can observe the type of the follower) and in the bandit setting (where the follower's type is not observed). Sessa et al. [19] study general repeated games between a leader with a finite number of actions and an follower with a finite number of types which may adversarially change over time. They prove bounds on the leader's regret in the full information setting when the utility of each follower (which is determined by its type) is only known to satisfy certain regularity assumptions. Bai et al. [3] show that in the bandit setting with finitely many actions for leader and follower, there exist expected utility functions such that any leader's algorithm suffers non-vanishing regret with probability at least 1/3. They also show leader algorithms that converge to SE up to a certain suboptimality gap. Deng et al. [9] prove some interesting non-constructive results. Let V be the utility of the leader in a SE. Under some mild assumptions, they show that for any $\varepsilon > 0$ the leader can always obtain a utility of at least $(V - \varepsilon)T - o(T)$ in T rounds, against any no-regret algorithm of the follower (note that the convergence rate is not explicit in their results). Mansour et al. [15] extend these results to Bayesian games.

Note that our results take advantage of the specific structure of the utility functions to obtain good rates for the leader's regret in a bandit setting. Note also that, unlike previous works, the follower's best response in our setting is not determined by a type, but rather learned from observed data. This allows us to prove that the follower's regret vanishes too. As a consequence, we are also able to prove convergence to SE of the players' strategy profile.

2 STAGE GAME AND UNIQUE SE

In this section, we present and analyze the (one-shot) stage version of our newsvendor game and characterize the uniqueness of its SE (formally defined below). We also provide some insights on the learning results proven in the following sections.

The stage game. An instance of the stage game is characterized by a known distribution $\mathcal{D}$ on $[0, \infty)^3$ that governs the (possibly correlated) *production cost* C of the supplier, the *retail price* P dictated by the market, and the *demand* D. We make the following assumption on $\mathcal{D}$.

Assumption 1. *The distribution $\mathcal{D}$ of $(C, P, D) \in [0, \infty)^3$ satisfies the following:*

(1) $\mathbb{E}[C], \mathbb{E}[P], \mathbb{E}[D]$, and $\mathbb{E}[PD]$ are all finite.
(2) $\mathbb{E}[C] < \mathbb{E}[P]$.
(3) The conditional distribution of D given (C, P) admits a density (w.r.t. the Lebesgue measure) such that $f(d \mid c, p) > 0$, for all $(c, p, d) \in [0, \infty)^3$.

Item 1 guarantees that the expected utilities of the supplier and the retailer (see below for a definition) are well-defined and finite for any action profile. Item 2 is an economic assumption stating that, on average, the supplier's cost is lower than the retail price, thus eliminating trivial scenarios. Item 3 is a mild technical condition that simplifies the presentation of the proof of Theorem 2.1.

We denote the conditional cumulative distribution function and survival function of the demand, given the supplier's cost and retail price, for all $c, p, d \geq 0$, by

$$F(d \mid c, p) \coloneqq \int_0^d f(x \mid c, p)\, \mathrm{d}x \quad \text{and} \quad \bar{F}(d \mid c, p) \coloneqq 1 - F(d \mid c, p).$$

In this section, we assume that the structure of the model (namely, $\mathcal{D}$ in Assumption 1) is common knowledge to both players. The game proceeds as follows. First, the supplier (S) selects a wholesale price $w \in [0, \infty)$ and reveals it to the retailer. Then, the retailer (R) selects a quantity $q \in [0, \infty)$. Their expected *utilities* are respectively defined, for any $(w, q) \in [0, \infty)^2$, by[1]

$$u_{\mathrm{S}}(w, q) \coloneqq qw - q\mathbb{E}\big[C\big] \quad \text{and} \quad u_{\mathrm{R}}(w, q) \coloneqq \mathbb{E}\big[\min\{q, D\}P\big] - qw.$$

where the expectations are with respect to $(C, P, D) \sim \mathcal{D}$.

Finally, the Stackelberg Equilibria (SE) of this game are defined as strategy pairs $(w^\star, q^\star) \in [0, \infty)^2$ such that

$$w^\star \in \operatorname*{argmax}_{w \in [0,\infty)} u_{\mathrm{S}}\big(w, \mathrm{BR}(w)\big) \quad \text{and} \quad q^\star = \mathrm{BR}(w^\star).$$

where BR is a best-response of the retailer, i.e., for all $w \in [0, \infty)$, $\mathrm{BR}(w) \in \operatorname{argmax}_{q \in [0,\infty)} u_{\mathrm{R}}(w, q)$.[2]

[1] All that matters regarding C is its expectation: we can strengthen all statements where we assume knowledge of the distribution of C by only assuming that of $\mathbb{E}[C]$.
[2] At this stage we allow arbitrary tie-breaking rules to determine the two argmax's, but we preemptively note that in all of our results the maximizers will be unique.

Uniqueness of the SE.. We can now state our characterization of the uniqueness of the SE in the stage game.

Theorem 2.1. *Under Assumption 1, let $g(w) := h^{-1}(w)$ be the inverse of*

$$h(x) := \mathbb{E}\big[P\bar{F}(x \mid C, P)\big]. \tag{1}$$

Then the following conditions are equivalent:

(1) The stage game in Section 2 admits a unique SE $\big(w^\star, g(w^\star)\big)$.
(2) $\{w^\star\} \equiv \operatorname{argmax}_{w\in A} g(w)\big(w - \mathbb{E}[C]\big)$, where

$$A := \left\{ w \in \big(\mathbb{E}[C], \mathbb{E}[P]\big) : -\frac{g'(w)}{g(w)} = \frac{1}{w - \mathbb{E}[C]} \right\}.$$

As an example, the SE is unique when D has a Weibull distribution with nondecreasing failure rate and (C, P) is deterministic (for more, see the extended version of this work [7]).

To prove the theorem, we begin with a simple but key lemma, whose proof can be found in the extended version [7].

Lemma 2.2. *Under Assumption 1, the function h in Equation (1) is differentiable on $x \in (0, \infty)$ and has a strictly negative derivative, hence it is invertible and its inverse $h^{-1}(w)$ is differentiable and strictly decreasing on $w \in \big(0, \mathbb{E}[P]\big)$.*

We are now ready to prove Theorem 2.1.

Proof of Theorem 2.1. We prove the theorem by backward induction. First, we show that under Assumption 1, the retailer has a unique best response. Then, we show that given that the retailer best-responds, the supplier has a unique optimal move $w^\star$ if and only if the condition of the theorem holds.

Retailer's move. Fix an arbitrary wholesale price $w \geq 0$. For any $q \geq 0$, the retailer's utility $u_{\mathrm{R}}(w, q)$ is

$$\mathbb{E}\left[P\left(\int_0^q x f(x \mid C, P)\,\mathrm{d}x + q\big(1 - F(q \mid C, P)\big)\right)\right] - qw.$$

To maximize it, we compute its derivative, which is justified by Assumption 1 combined with the Leibniz integral rule. For any $q > 0$, we obtain

$$\frac{\partial}{\partial q} u_{\mathrm{R}}(w, q) = \mathbb{E}\big[P\,\bar{F}(q \mid C, P)\big] - w, \tag{2}$$

which is non-negative if and only if $\mathbb{E}\big[P\bar{F}(q \mid C, P)\big] \geq w$. By the arbitrariness of w and Lemma 2.2, we can conclude that, for any choice of the wholesale price $w > 0$, there exists a unique maximizer $q^\star_w = \mathrm{BR}(w)$ of $q \mapsto u_{\mathrm{R}}(w, q)$ where

$$\mathrm{BR}(w) := \begin{cases} g(w) & \text{if } w < \mathbb{E}[P], \\ 0 & \text{if } w \geq \mathbb{E}[P]. \end{cases} \tag{3}$$

Supplier's move. Given the retailer's best response $q^\star_w = \mathrm{BR}(w)$, the supplier's utility is, for any $w > 0$,

$$u_{\mathrm{S}}(w, q^\star_w) = q^\star_w\big(w - \mathbb{E}[C]\big). \tag{4}$$

Since $q^\star_w = 0$ for all $w \geq \mathbb{E}[P]$ and $w - \mathbb{E}[C] \leq 0$ for all $w \leq \mathbb{E}[C]$, to maximize the supplier's expected utility, we can restrict our search to $w \in \big(\mathbb{E}[C], \mathbb{E}[P]\big)$, where $u_{\mathrm{S}}(w, q^\star_w) = g(w)\big(w - \mathbb{E}[C]\big)$ is strictly positive and differentiable. To find the maximum, then, we can study the sign of the derivative of the supplier's expected utility, obtaining, for all $w \in \big(\mathbb{E}[C], \mathbb{E}[P]\big)$,

$$\frac{\partial}{\partial w} u_{\mathrm{S}}(w, q^\star_w) = g'(w)\big(w - \mathbb{E}[C]\big) + g(w). \tag{5}$$

Thus, A (in the statement of the theorem) is the set of all stationary points of $u_{\mathrm{S}}(w, q^\star_w)$ and the condition that $\operatorname{argmax}_{w\in A} g(w) \cdot \big(w - \mathbb{E}[C]\big)$ is a singleton is exactly stating that there exists a unique maximizer of $g(w)\big(w - \mathbb{E}[C]\big) = u_{\mathrm{S}}(w, q^\star_w)$, which coincides with the existence of a unique SE. □

In general, the payoffs under SE may be unique under weaker conditions than the ones for the uniqueness of the SE, but such weaker conditions do not appear to be simply stated in our case.

We also note that efficiency of the Stackeleberg equilibrium can be measured using the price of anarchy. Some preliminary results on this topic can be found in the extended version [7].

3 LEARNING THE SE

A limitation of Theorem 2.1 is that, even when a stage game has a unique SE, in order to compute it, both players need to know the underlying distribution $\mathcal{D}$. In this section, we show how to circumvent this issue and achieve convergence to the unique SE without relying on the knowledge of $\mathcal{D}$. We do this by reconstructing the salient features of $\mathcal{D}$ through a learning technique in a repeated game. An instance of the repeated game is characterized by a distribution $\mathcal{D}$ on[3] $[0, 1]^3$, that governs the (possibly correlated) production cost, retail price, and demand.

We study the following online protocol. At each round $t = 1, 2, \ldots$:

(1) Nature draws (C_t, P_t, D_t) i.i.d. according to $\mathcal{D}$.
(2) The supplier (S) selects a wholesale price $W_t \in [0, 1]$ and reveals it to the retailer (R).
(3) The production cost C_t is revealed to S.
(4) R buys a quantity $Q_t \in [0, 1]$, paying $Q_t W_t$ to S.
(5) The retail price P_t and the demand D_t are revealed to R.
(6) The market buys a quantity $\min\{Q_t, D_t\}$, paying the corresponding $\min\{Q_t, D_t\}P_t$ to R.

The individual goals of the supplier and the retailer are to maximize, for any time horizon T, their individual long-term expected utilities in reference to a SE $(w^\star, q^\star)$; more precisely:

$$\mathbb{E}\big[\sigma(w^\star, q^\star, C_1)\big] - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\big[\sigma(W_t; Q_t, C_t)\big],$$

$$\mathbb{E}\big[\rho(q^\star; w^\star, P_1, D_1)\big] - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\big[\rho(Q_t; W_t, P_t, D_t) \mid W_t\big],$$

where we define $\sigma(w; q, c) := qw - qc$ for $(w, q, c) \in [0, 1]^3$ and $\rho(q; w, p, d) := \min\{q, d\}p - qw$ for $(q, w, p, d) \in [0, 1]^4$. The conditional expected utility of the retailer is to be maximized with high probability with respect to $(W_t)_{t\in\mathbb{N}}$. The asymmetry in the objectives of the supplier (S) and retailer (R) is due to the fact that W_t is revealed to R before R makes a decision at time t, while S has to act before observing Q_t.

[3] In contrast to Section 2, we assume here that $\mathcal{D}$ is bounded (without loss of generality, by 1). This assumption is for simplifying the presentation; all the following results can be extended to the unbounded case simply by assuming subgaussianity.

3.1 Convergence to SE With Partial Information (ETC vs BR)

In Section 2, we showed that the stage game admits a unique SE under the assumptions in Theorem 2.1, and one can obtain the SE if $\mathcal{D}$ is known by both the supplier (S) and the retailer (R). We now show that a repeated interaction between S and R who act rationally and selfishly can lead to a convergence to the SE, even assuming that S and R have only partial information on $\mathcal{D}$. In Section 3.2, we will show that the same result can be obtained even when S and R have essentially no information on $\mathcal{D}$ at a cost of a slower convergence rate. We make the following assumption.

Assumption 2. *The distribution $\mathcal{D}$ of $(C, P, D) \in [0,1]^3$ satisfies the following:*

(1) $\mathbb{E}[C] < \mathbb{E}[P]$.
(2) D and C are conditionally independent, given P.
(3) The conditional distribution of D given P admits a density (with respect to the Lebesgue measure) such that $f(d \mid p) > L$, for some $L > 0$ and all $(p, d) \in [0,1]^2$.
(4) Condition 2 of Theorem 2.1 holds (i.e., the SE is unique).

Condition 1 states that, on average, the supplier's cost is lower than the retail price, eliminating trivial scenarios. Condition 2 states that the demand D may depend on the supplier's cost C only via the retail price P. Condition 3 is a mild technical condition guaranteeing that the learning problem is at least Lipschitz (see below). Essentially, Assumption 2 implies that our base assumption (Assumption 1) is satisfied and guarantees that Theorem 2.1 can be applied, so that the learning problem is tractable.

We now assume partial knowledge of the supplier and the retailer, where S has more information on the production cost, while R has more information about the direct interaction with the market.

Assumption 3. *The supplier has access to the marginal distribution of C, but not to that of (P, D); the retailer has access to the marginal of (P, D), but not to that of C.*

The retailer's strategy is to best-respond to any wholesale price. Note that $q^\star_w = \mathrm{BR}(w)$ can be computed exactly via Equation (3) thanks to Assumptions 2 and 3.

Under these assumptions on the setting, and given that the retailer best-responds, the supplier can compute their expected cost $\mathbb{E}[C]$ and is left with solving a zeroth-order Lipschitz optimization problem (i.e., maximizing Equation (4)) without the knowledge of the Lipschitz constant of its objective. This can be done with a simple Explore-Then-Commit algorithm (Algorithm 1), which we assume to be the supplier's strategy.[4]

Theorem 3.1. *Under Assumptions 2 and 3, for any horizon T, if the supplier runs Algorithm 1 with input T and the retailer best-responds, then*

$$\mathbb{E}\big[\sigma(w^\star; q^\star, C)\big] - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\sigma(w_t; q_t, C_t)\big] \le \left(\frac{1-\mathbb{E}[C]}{\mathbb{E}[P]L} + 2\right) T^{-1/2}, \tag{6}$$

where $(w^\star, q^\star)$ is the unique SE of the stage game. Also, for all sufficiently large T:

[4] We denote by $\lfloor x \rfloor$ the floor of x, i.e., the largest integer $n \le x$.

input: Time horizon T
for $t = 1, \dots, \lfloor T^{1/2} \rfloor$ **do**
  Select the wholesale price $w_t := t/\big(\lfloor T^{1/2} \rfloor + 1\big)$
  Observe the quantity q_t
end
for $t = \lfloor T^{1/2} \rfloor + 1, \dots, T$ **do**
  Select a wholesale price $w_t = w_{s^\star}$, where
  $s^\star \in \operatorname*{argmax}_{s=1,\dots,\lfloor T^{1/2} \rfloor} q_s\big(w_s - \mathbb{E}[C]\big)$
  Observe the quantity q_t
end

Algorithm 1: Explore-Then-Commit

(1) $\mathbb{E}\big[\rho(q^\star; w^\star, P, D)\big] - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\rho(q_t; w_t, P_t, D_t)\big] \le \big(L^{-1} + 2\big)T^{-1/2}$.
(2) $\big\|(w^\star, q^\star) - (w_T, q_T)\big\|_1 \le \big((\mathbb{E}[P]L)^{-1} + 1\big)T^{-1/2}$.

Item 2 shows *last-iterate* convergence to the unique SE (in contrast to the weaker *time-average* convergence that is typically obtained in regret minimization). Equation (6) and Item 1 give performance guarantees for both the supplier and the retailer, showing that not only their utilities converge to that of the SE, but that both their cumulative utilities match (up to lower-order terms) those that would be gathered by consistently selecting the SE $(w^\star, q^\star)$ at all time steps with full knowledge of $\mathcal{D}$. Crucially, while for the supplier it is possible to obtain finite-time regret guarantees, for the retailer these only hold asymptotically. Finally, given that in this setting the supplier's actions w_t are deterministic, we omitted from Item 1 the redundant conditioning on w_t and both Item 1 and Item 2 hold deterministically.

Proof. By Theorem 2.1, under Assumption 2, the best-response function BR defined in (3) maps each wholesale price w into its unique best-response $q^\star_w = \mathrm{BR}(w)$, which the retailer (R) can compute by Assumption 3. Since R is best-responding, the utility of the supplier (S), for any $w > 0$, is $q^\star_w\big(w - \mathbb{E}[C]\big)$. Note that, under Assumption 2, the unique SE is $(w^\star, q^\star_{w^\star})$, where $w^\star$ is the unique maximizer of $q^\star_w\big(w - \mathbb{E}[C]\big)$, which S cannot compute directly because Assumption 3 is not sufficient for S to determine $q^\star_w$. However, S can calculate $\mathbb{E}[C]$ with the knowledge of the marginal distribution of C. Hence, S gets a noise-free evaluation $q^\star_{w_t}\big(w_t - \mathbb{E}[C]\big)$ at each round t, *after* selecting the wholesale price w_t for the round.

Now, note that S's objective $w \mapsto q^\star_w\big(w - \mathbb{E}[C]\big)$ is Lipschitz. Indeed, recalling (4), (5), and Assumption 2 (for more details, see the extended version [7]), we get that, for any $w \in (0,1)$,

$$\left|\frac{\partial}{\partial w} q^\star_w\big(w - \mathbb{E}[C]\big)\right| \le \frac{1}{\mathbb{E}[P]L} \cdot \big(1 - \mathbb{E}[C]\big) + 1.$$

Since for any time horizon T, S selects the best point in a grid of step-size at most $T^{-1/2}$, we have that

$$\lim_{T\to\infty} \big|w_T - w^\star\big| \le \lim_{T\to\infty} T^{-1/2} = 0. \tag{7}$$

Then, using again (5) and Assumption 2 (for more details, see the extended version [7]), we get that R's best-response function BR is $1/(\mathbb{E}[P]L)$-Lipschitz. Recalling (2), we also have that for any fixed wholesale price w, R's instantaneous utility at time t, $q \mapsto$

$\mathbb{E}\big[\rho(q; w, P_t, D_t)\big]$ is 1-Lipschitz. Hence

$$\lim_{T\to\infty} \left|q_T - q^\star_{w^\star}\right| \le \lim_{T\to\infty} T^{-1/2}/(\mathbb{E}[P]L) = 0. \tag{8}$$

Putting (8) and (7) together, gives Item 2. Equation (6) is an immediate consequence of the $\big((1-\mathbb{E}[C])/(\mathbb{E}[P]L)+1\big)$-Lipschitzness of S's objective $w \mapsto q^\star_w\big(w-\mathbb{E}[C]\big)$. Item 1 is an immediate consequence of the $(L^{-1}+1)$-Lipschitzness of R's utility $w \mapsto \mathbb{E}\big[\rho(q^\star_w; w, P_t, D_t)\big]$, which follows directly from the chain rule. □

The previous result yields sublinear regret guarantees for S even when S is oblivious to the expected retail price $\mathbb{E}[P]$ and the lower bound L on the conditional density of the demand. Since the Lipschitz constant of S's objective is a deterministic function of $\mathbb{E}[C]$, L, and $\mathbb{E}[P]$, the reader might wonder if improved regret guarantees could be achieved if these quantities were known to S. We show now that this is indeed the case.

We refine Assumption 3 as follows.

Assumption 4. *The supplier has access to the $\mathbb{E}[C]$, L, and $\mathbb{E}[P]$; the retailer has access to the marginal distribution of (P, D).*

Under Assumptions 2 and 4, and given that R best-responds to any wholesale price w, the supplier S can compute their expected cost and solve their zeroth-order Lipschitz optimization problem with the knowledge of (an upper bound of) the Lipschitz constant of its objective. This can be done with the Piyavskii–Shubert algorithm (Algorithm 2). We now assume that this is S's strategy.

input: Time horizon T, Lipschitz constant $M > 0$
initialization: Let $w_1 := 1$
for $t = 1, \dots, T$ **do**
  Select the wholesale price w_t
  Observe the quantity q_t
  Update the proxy function
    $\hat\rho_t(\cdot) := \min_{s\in[t]}\big\{q_s\big((\cdot)-\mathbb{E}[C]\big) + M\big|w_s - (\cdot)\big|\big\}$
  Let $w_{t+1} \in \operatorname{argmax}_{w\in[0,1]} \hat\rho_t(w)$
end

Algorithm 2: Piyavskii–Shubert

The Piyavskii–Shubert algorithm has been known for half a century [18, 20] but only recently proven to enjoys outstanding theoretical guarantees for its query complexity, regret, and robustness [5, 11]. In particular, the following theorem follows directly by specializing [5, Theorem 3.5] and [11, Theorem 1] to our setting.

Theorem 3.2 ([5, 11]). *Under Assumptions 2 and 4, for any horizon T, if the supplier runs Algorithm 2 with inputs T and $M := (1-\mathbb{E}[C])/(\mathbb{E}[P]L)+1$, and the retailer best-responds, then the function $w \mapsto q^\star_w\big(w-\mathbb{E}[C]\big)$ is M-Lipschitz and, for all $t \in [T]$,*

$$\max_{w\in[0,1]}\big\{q^\star_w\big(w-\mathbb{E}[C]\big)\big\} - q^\star_{w_t}\big(w_t-\mathbb{E}[C]\big) \le 9M\frac{\log_2(Mt)}{t}$$

$$\max_{w\in[0,1]}\big\{q^\star_w\big(w-\mathbb{E}[C]\big)\big\} - \frac{1}{T}\sum_{t=1}^{T} q^\star_{w_t}\big(w_t-\mathbb{E}[C]\big) \le 2M\frac{\ln(4T)}{T}$$

Theorem 3.2 allows us to prove the following result.

Theorem 3.3. *Under Assumptions 2 and 4, for any horizon T, if the supplier runs Algorithm 2 with inputs T and $M := (1-\mathbb{E}[C])/(\mathbb{E}[P]L)+1$, and the retailer best-responds, then:*

$$\mathbb{E}\big[\sigma(w^\star; q^\star, C)\big] - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\sigma(w_t; q_t, C_t)\big] \le 2M\frac{\ln(4T)}{T} \tag{9}$$

where $(w^\star, q^\star)$ is the unique SE of the stage game. Moreover,

(1) $\lim_{T\to\infty}\left(\mathbb{E}\big[\rho(q^\star; w^\star, P, D)\big] - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\rho(q_t; w_t, P_t, D_t)\big]\right) = 0$

(2) $\lim_{T\to\infty}\big\|(w^\star, q^\star) - (w_T, q_T)\big\|_1 = 0$

Proof. Proceeding as in the proof of Theorem 3.1 and applying Theorem 3.2, we get that the retailer's instantaneous utility at time t, $q \mapsto \mathbb{E}\big[\rho(q; w, P_t, D_t)\big]$ is 1-Lipschitz for any fixed wholesale price w, and the supplier's instantaneous utility at time t, $w \mapsto q^\star_w\big(w-\mathbb{E}[C_t]\big)$ is M-Lipschitz, where $q^\star_w$ is defined as in (3), for all $w \in [0,1]$. As above, under Assumption 2, the unique SE is precisely $(w^\star, q^\star_{w^\star})$, where $w^\star$ is the unique maximizer of $q^\star_w\big(w-\mathbb{E}[C]\big)$. Applying again Theorem 3.2, we obtain immediately the result. □

3.2 Convergence to SE With No Information (ETC vs FTL)

If the retailer R had access to the distribution $\mathcal{D}$, or at least to the marginal distribution of (P, D), R could best-respond to the supplier's move w_t at each time t, as described in Sections 2 and 3.1. Since, in this section, none of these is available to R, we assume that R acts according to the next-best available strategy, i.e., best-responding to an empirical distribution that can be maintained by gathering samples. In the literature, this strategy is known as Follow-the-Leader (FTL) and is detailed in Algorithm 3. For $t = 1$,

input: Time horizon $T \ge 12$
initialization: Let
  $\mathcal{Q} := \big\{1/(\lceil T^{1/3}\rceil+1), \dots, \lceil T^{1/3}\rceil/(\lceil T^{1/3}\rceil+1)\big\}$
Observe the wholesale price W_1
Draw a quantity Q_1 from $\mathcal{Q}$ uniformly at random
Observe the demand D_1 and the retail price P_1
for $t = 2, 3, \dots$ **do**
  Observe the wholesale price W_t
  Select a quantity
    $Q_t \in \operatorname{argmax}_{q\in\mathcal{Q}}\big(\frac{1}{t-1}\sum_{s=1}^{t-1}\min\{q, D_s\}P_s - qW_t\big)$
  Observe the demand D_t and the retail price P_t
end

Algorithm 3: Follow-the-Leader (FTL)

R picks a quantity at random and observes the demand D_1. During each time step $t \ge 2$, define, for all $w, q \in [0,1]$, the auxiliary function

$$\widehat{\rho}_t(w, q) := \frac{1}{t-1}\sum_{s=1}^{t-1}\min\{q, D_s\}P_s - qw.$$

Note that this is not built to maximize the empirical average of the utility gained in last $t-1$ interactions, i.e., it differs from $q \mapsto \frac{1}{t-1}\sum_{s=1}^{t-1}\big(\min\{q,\ D_s\}P_s - qW_s\big)$. Indeed, the retailer is not interested

in maximizing their expected utilities at time steps t but rather, their more challenging expected utility given W_t. Equivalently stated, the retailer is not maximizing an expected revenue computed with respect to the empirical distribution of the sequence of random vectors $(C_1, P_1, D_1, W_1), \ldots, (C_{t-1}, P_{t-1}, D_{t-1}, W_{t-1})$ at time t, but rather, that of $(C_1, P_1, D_1, W_t), \ldots, (C_{t-1}, P_{t-1}, D_{t-1}, W_t)$ given W_t. This way, $\widehat{\rho}_t(w, q)$ is an unbiased estimate of $\mathbb{E}\big[\rho(q; w, P_t, D_t)\big]$ for all $w, q \geq 0$ and $t \geq 2$, which in turn implies that $\mathbb{E}\big[\widehat{\rho}_t(W_t, q) \mid W_t\big] = \mathbb{E}\big[u_S(q; W_t, P_t, D_t) \mid W_t\big]$ for all $q \geq 0$ and $t \geq 2$. This corresponds precisely to the instantaneous objective of the retailer. Therefore, naturally, the choice of a discretized retailer at time $t \geq 2$ is

$$Q_t \in \operatorname*{argmax}_{q \in Q} \big(\widehat{\rho}_t(W_t, q)\big) = \operatorname*{argmax}_{q \in Q} \left(\frac{1}{t-1} \sum_{s=1}^{t-1} \min\{q,\, D_s\} P_s - q W_t \right).$$

Similarly to the previous section, we assume here that the supplier adopts an Explore-Then-Commit strategy (Algorithm 4), with the *caveat* that, in this section, the expected production cost $\mathbb{E}[C]$ is not available to S but has to be estimated.

input: Time horizon $T \geq 12$
for $t = 1, \ldots, \lceil T^{1/3} + 1 \rceil$ **do**
  for $s = 1, \ldots, \lceil T^{1/3} \rceil$ **do**
    Select the wholesale price $W_{(t-1)\lceil T^{1/3} \rceil + s} \coloneqq s / \big(\lceil T^{1/3} \rceil + 1\big)$
    Observe the quantity $Q_{(t-1)\lceil T^{1/3} \rceil + s}$ and production cost $C_{(t-1)\lceil T^{1/3} \rceil + s}$
  end
end
Compute $S^\star \in \operatorname{argmax}_{s \in \{\lceil T^{1/3} \rceil^2 + 1, \ldots, \lceil T^{1/3} + 1 \rceil \lceil T^{1/3} \rceil\}} Q_s \big(W_s - \frac{1}{\lceil T^{1/3} \rceil^2} \sum_{j=1}^{\lceil T^{1/3} \rceil^2} C_j\big)$
for $t = \lceil T^{1/3} + 1 \rceil \lceil T^{1/3} \rceil + 1, \ldots, T$ **do**
  Select the wholesale price $W_t \coloneqq W_{S^\star}$
  Observe the quantity Q_t
end

Algorithm 4: Explore-Then-Commit (without knowledge of $\mathbb{E}[C]$)

Theorem 3.4. *Under Assumption 2, for any horizon $T \geq 12$, if the supplier runs Explore-Then-Commit (Algorithm 4) with input T and the retailer runs Follow-the-Leader (Algorithm 3) with input T, then:*

$$\mathbb{E}\big[\sigma(w^\star; q^\star, C)\big] - \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\big[\sigma(W_t; Q_t, C_t)\big] \leq \left(16 + \frac{1 - \mathbb{E}[C]}{\mathbb{E}[P] L} + 7\sqrt{\ln T}\right) T^{-1/3}, \quad (10)$$

where $(w^\star, q^\star)$ is the unique SE of the stage game. Moreover:

(1) $\lim_{T \to \infty} \Big(\mathbb{E}\big[\rho(q^\star; w^\star, P, D)\big] - \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\big[\rho(Q_t; W_t, P_t, D_t \mid W_t)\big] \Big) = 0$ *with probability* 1.

(2) $\lim_{T \to \infty} \big\| (w^\star, q^\star) - (W_T, Q_T) \big\|_1 = 0$ *with probability* 1.

Proof. Fix any time horizon $T \geq 12$. Proceeding as in the proof of Theorem 3.1 and applying Theorem 3.2, we get that the retailer's instantaneous utility at time t, $q \mapsto \mathbb{E}\big[\rho(q; w, P_t, D_t)\big]$ is 1-Lipschitz for any fixed wholesale price w, and the supplier's instantaneous utility at time t, $w \mapsto q^\star_w\big(w - \mathbb{E}[C_t]\big)$ is M-Lipschitz, where $q^\star_w$ is defined as in (3), for all $w \in [0, 1]$, and $M \coloneqq \big(1 - \mathbb{E}[C]\big) / \big(\mathbb{E}[P] L\big) + 1$. As above, under Assumption 2, the unique SE is precisely $(w^\star, q^\star_{w^\star})$, where $w^\star$ is the unique maximizer of $w \mapsto q^\star_w\big(w - \mathbb{E}[C]\big)$. Now, fix an arbitrary $\delta \in \big(0, 1/\lceil T^{1/3} + 1 \rceil\big)$. Observing that for any $t \geq 2$, given W_t, the retailer's quantity Q_t is the argmax of an empirical average translated by a constant, applying Hoeffding's inequality $\lceil T^{1/3} + 1 \rceil$ times and the fact that the retailer's discretization has step-size $1/\lceil T^{1/3} + 1 \rceil$, we obtain that

$$\left| Q_s - q^\star_{s/(\lceil T^{1/3} \rceil + 1)} \right| \leq \sqrt{\frac{\ln 2/\delta}{2\lceil T^{1/3} \rceil^2}} + \frac{1}{\lceil T^{1/3} + 1 \rceil} \quad (11)$$

$$\text{for all } s = \lceil T^{1/3} \rceil^2 + 1, \ldots, \lceil T^{1/3} + 1 \rceil \lceil T^{1/3} \rceil$$

$$\left| \frac{1}{\lceil T^{1/3} \rceil^2} \sum_{j=1}^{\lceil T^{1/3} \rceil^2} C_j - \mathbb{E}[C] \right| \leq \sqrt{\frac{\ln 2/\delta}{2\lceil T^{1/3} \rceil^2}} \quad (12)$$

hold simultaneously with probability at least $1 - \lceil T^{1/3} + 1 \rceil \delta$. Thus,

$$\left| Q_s \left(W_s - \frac{1}{\lceil T^{1/3} \rceil^2} \sum_{j=1}^{\lceil T^{1/3} \rceil^2} C_j \right) - q^\star_{W_s}\big(W_s - \mathbb{E}[C]\big) \right| \leq 2\sqrt{\frac{2 \ln 2/\delta}{\lceil T^{1/3} \rceil^2}} + \frac{3}{\lceil T^{1/3} + 1 \rceil}$$

hold simultaneously for all $s = \lceil T^{1/3} \rceil^2 + 1, \ldots, \lceil T^{1/3} + 1 \rceil \lceil T^{1/3} \rceil$, with probability at least $1 - \lceil T^{1/3} + 1 \rceil \delta$. Consequently, letting

$$\tilde{w} \in \operatorname*{argmax}_{w \in \{1/\lceil T^{1/3}+1 \rceil, \ldots, \lceil T^{1/3} \rceil / \lceil T^{1/3}+1 \rceil\}} q^\star_w\big(w - \mathbb{E}[C]\big)$$

and given that the supplier's discretization has a step size $1/\lceil T^{1/3} + 1 \rceil$ and $w \mapsto q^\star_w\big(w - \mathbb{E}[C]\big)$ is M-Lipschitz, the triangular inequality yields, for any $t \geq \lceil T^{1/3} + 1 \rceil \lceil T^{1/3} \rceil + 1$,

$$\begin{aligned}
&\left| Q_t \left(W_t - \frac{1}{\lceil T^{1/3} \rceil^2} \sum_{j=1}^{\lceil T^{1/3} \rceil^2} C_j \right) - \mathbb{E}\big[\sigma(w^\star, q^\star, C)\big] \right| \leq \\
&\leq \left| Q_t \left(W_{S^\star} - \frac{1}{\lceil T^{1/3} \rceil^2} \sum_{j=1}^{\lceil T^{1/3} \rceil^2} C_j \right) - q^\star_{W_{S^\star}}\big(W_{S^\star} - \mathbb{E}[C]\big) \right| \\
&\quad + \left| q^\star_{W_{S^\star}}\big(W_{S^\star} - \mathbb{E}[C]\big) - q^\star_{\tilde{w}}\big(\tilde{w} - \mathbb{E}[C]\big) \right| \\
&\quad + \left| q^\star_{\tilde{w}}\big(\tilde{w} - \mathbb{E}[C]\big) - \mathbb{E}\big[\sigma(w^\star, q^\star, C)\big] \right| \\
&\leq 6\sqrt{\frac{2 \ln 2/\delta}{\lceil T^{1/3} \rceil^2}} + \frac{9 + M}{\lceil T^{1/3} + 1 \rceil}
\end{aligned}$$

with probability at least $1 - \lceil T^{1/3} + 1 \rceil \delta$. Charging regret 1 to the supplier for the first $\lceil T^{1/3} + 1 \rceil \lceil T^{1/3} \rceil$ rounds, then summing the

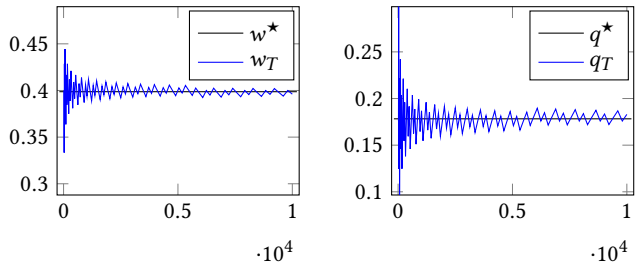

Figure 1: On the x axis, the time horizon T. On the y axis of the left (resp., right) pane, the wholesale price w_T (resp., the quantity q_T) at time T when the suppliers runs Algorithm 1 and the retailer best-responds. The black horizontal line on the left (resp., right) pane represents the wholesale price $w^\star$ (resp., the quantity $q^\star$) at the SE.

previous bound over all remaining rounds and upper bounding $T - \lceil T^{1/3} + 1 \rceil \lceil T^{1/3} \rceil$ with T yields

$$\begin{aligned}
&\mathbb{E}\big[\sigma(w^\star; q^\star, C)\big] - \frac{1}{T}\sum_{t=1}^{T} \sigma(W_t; Q_t, C_t) \\
&\qquad \le \frac{\lceil T^{1/3} + 1 \rceil \lceil T^{1/3} \rceil}{T} + 6\sqrt{\frac{2 \ln 2/\delta}{\lceil T^{1/3} \rceil^2}} + \frac{9 + M}{\lceil T^{1/3} + 1 \rceil} \\
&\qquad \le \big(15 + M + 6\sqrt{2 \ln 2/\delta}\big) T^{-1/3}
\end{aligned}$$

with probability at least $1 - \lceil T^{1/3} + 1 \rceil \delta$. Thus, (10) follows directly by choosing, e.g., $\delta = 2T^{-2/3}$ and upper bounding $12/\sqrt{3}$ with 7. The proof of Item 2 is a simple consequence of (11) and the fact that the retailer's best-response function BR is Lipschitz. Finally, Item 1 follows directly by (11), (12), and Lipschitzness of the retailer's utility $w \mapsto \mathbb{E}\big[\rho(q^\star_w; w, P_t, D_t)\big]$ as a function of the wholesale price (which is implied by the chain rule). □

Again, the previous result shows *last-iterate* convergence of the supply chain to the unique SE (in contrast to the weaker *time-average* convergence that is typically obtained in regret minimization). In contrast to our previous results, this theorem holds under much weaker assumptions on the prior knowledge of the retailer and the supplier. Indeed, we do not assume anything other than the knowledge that the support of $\mathcal{D}$ is included in $[0, 1]^3$.

4 EXPERIMENTS

In this section, we describe experiments on synthetic data that provide further insights on the theoretical results of Section 3. In our experiments, we consider a constant cost $C = 0.3$, a retail price P uniform on $[0, 1]$, and for all $p \in [0, 1]$, we assume that the probability density function given $P = p$ of the demand is $f(x \mid p) = (1 - 2p)\, x + \frac{1}{2} + p$.

4.1 ETC vs BR

In the first experiment, we study the supplier-retailer dynamics in the setting of Section 3.1, when the supplier runs Explore-Then-Commit (Algorithm 1) and the retailer best-responds (the resulting dynamics is deterministic).

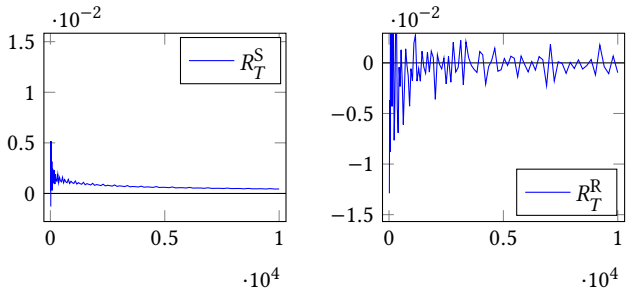

Figure 2: On the x axis (black horizontal line), the time horizon T. On the y axis of the left (resp., right) pane, the difference R_T^{S} (resp., R_T^{R}) after T time steps between the cumulative utility at the SE and that obtained by the supplier (resp., retailer) when the supplier runs Algorithm 1 and the retailer best-responds.

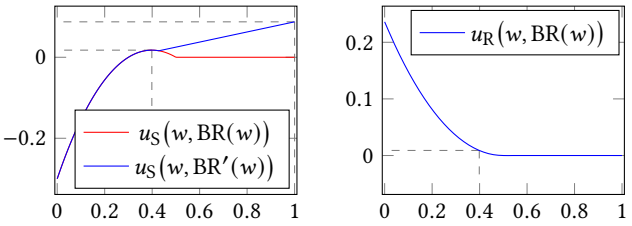

Figure 3: On the x axis, the wholesale price w. Left pane: in red, the utility $u_{\mathrm{S}}(w, \mathrm{BR}(w))$ of the supplier when the retailer best-responds, illustrating the objective of the supplier both in the noiseless setting of Section 4.1 and and the noisy one of Section 4.2, where the discretization of the retailer includes 0; in blue, the utility $u_{\mathrm{S}}(w, \mathrm{BR}'(w))$ of the supplier when the best response $\mathrm{BR}'(w) = \max\{\mathrm{BR}(w), q_{\min}\}$ of the retailer is constrained to be larger than $q_{\min} > 0$, illustrating the objective of the supplier in the noisy setting of Section 4.2, where the discretization of the retailer does not include 0. Right pane: the utility $u_{\mathrm{R}}(w, \mathrm{BR}(w))$ of the retailer when best-responding to w.

In Figure 1 the wholesale price w_T and the corresponding quantity q_T can be seen converging to the SE $(w^\star, q^\star)$. The corresponding long-term differences in utilities with respect to the utilities at the SE appear in Figure 2. While the performance of the supplier is as expected, the fact that the retailer frequently outperforms the SE is somewhat surprising. This behavior can be explained by noting that the utility $u_{\mathrm{R}}(w, \mathrm{BR}(w))$ of the retailer when the supplier selects w and the retailer best-responds with $\mathrm{BR}(w)$ is concave and decreasing (Figure 3, right pane). Hence, whenever the supplier undershoots with respect to $w^\star$, the potential gain of the retailer is higher that their potential loss due to the supplier overshooting. By observing Figure 1, we see that that in many rounds the supplier commits to a wholesale price lower than $w^\star$. Although for the supplier undershooting and overshooting is equivalent (see the symmetric form of the supplier's objective in a neighborhood of its maximum in Figure 3, left pane, red plot), each $w < w^\star$ gives the retailer an opportunity to outperform the utility of the SE. Being

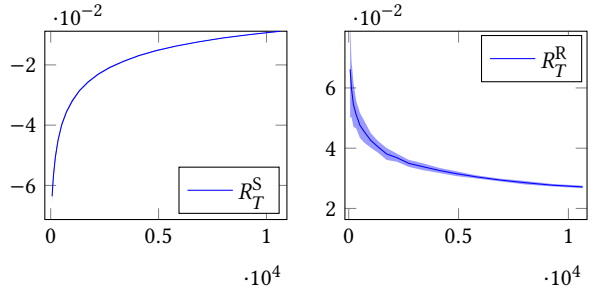

Figure 5: On the x axis, the time horizon T. On the y axis of the left (resp., right) pane, the difference R_T^{S} (resp., R_T^{R}) after T time steps between the cumulative utility at the SE and that obtained by the supplier (resp., retailer) when the supplier runs Algorithm 4 and the retailer runs Algorithm 3.

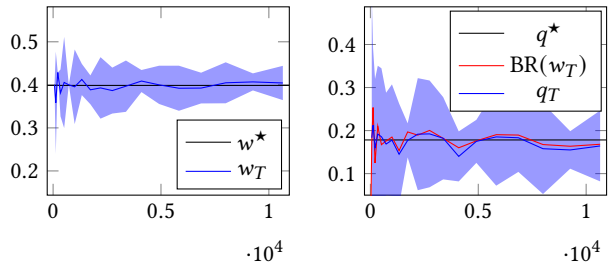

Figure 6: On the x axis, the time horizon T. On the y axis of the left (resp., right) pane, the wholesale price w_T (resp., the quantity q_T) at time T when the suppliers runs Algorithm 4 and the retailer runs Algorithm 3 on a discretization $0, 1/\lceil T^{1/3} - 1\rceil, \ldots, \lceil T^{1/3}\rceil/\lceil T^{1/3} - 1\rceil, 1$. The black horizontal line represents the wholesale price $w^\star$ (resp., the quantity $q^\star$) at the SE. In red, the plot of the sequence of exact best responses $\mathrm{BR}(w_T)$.

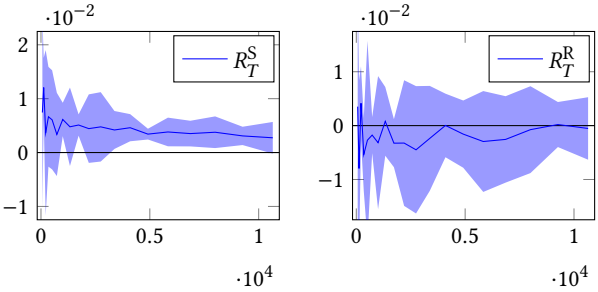

Figure 7: On the x axis (black horizontal line), the time horizon T. On the y axis of the left (resp., right) figure, the difference R_T^{S} (resp., R_T^{R}) after T time steps between the cumulative utility at the SE and that obtained by the supplier (resp., retailer) when the supplier runs Algorithm 4 and the retailer runs Algorithm 3 on a discretization $0, 1/\lceil T^{1/3} - 1\rceil, \ldots, \lceil T^{1/3}\rceil/\lceil T^{1/3} - 1\rceil, 1$.

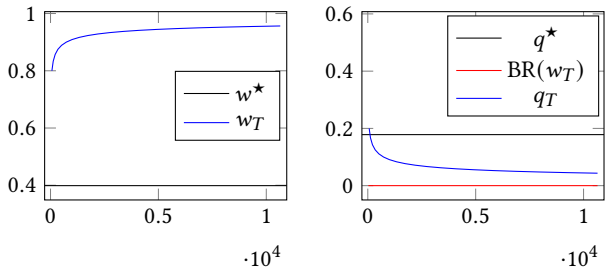

Figure 4: On the x axis, the time horizon T. On the y axis of the left (resp., right) figure, the wholesale price w_T (resp., the quantity q_T) at time T when the suppliers runs Algorithm 4 and the retailer runs Algorithm 3. The black horizontal line represents the wholesale price $w^\star$ (resp., the quantity $q^\star$) at the SE. In red, the plot of the sequence of exact best responses $\mathrm{BR}(w_T)$. It is always zero because the supplier correctly learns to always play $w > 0.5$ and $\mathrm{BR}(w) = 0$ for all $w > 0.5$. Note that the retailer is indeed learning that they should play the smallest possibly quantity.

able to compute the best response exactly, the retailer can take advantage of this fact.

4.2 ETC vs FTL

We now consider the setting of Section 3.2, where no player has any prior distributional information and all knowledge is learned by them running ETC (Algorithm 4) and FTL (Algorithm 3), respectively. As in this case the dynamics is stochastic, the plots are averages over 20 runs of each simulation.

The role of the discretization. Although not immediately apparent, the retailer's choice on how to implement their discretization has significant consequences on the dynamics of the system. Consider first Algorithm 3, where quantities are discretized as $i/\lceil T^{1/3} + 1\rceil$, for $i = 1, \ldots, \lceil T^{1/3}\rceil$, T being the time horizon. This effectively sets a hard *positive* lower bound to the best-response of the retailer. As Figure 3 (left pane, blue curve) shows, a consequence of this is that $w^\star$ is no longer the maximizer of the seller's utility, unless T is sufficiently large. A direct computation using the parameters chosen for the experiments shows that $T > 6 \cdot 10^4$ is needed to recover $w^\star$ as the solution to the supplier's optimization problem. For smaller horizons, the supplier should learn that their optimal move is to post the highest possible w. Figure 4 shows that this is exactly what happens. Therefore, unlike the previous setting, here the supplier is always outperforming the SE (Figure 5).

This artifact in the dynamics due to the retailer's discretization can be completely eliminated with a simple trick. Assume that the retailer discretizes the unit interval using $i/\lceil T^{1/3} - 1\rceil$, for $i = 0, \ldots, \lceil T^{1/3} - 1\rceil$. Now the quantity $q = 0$ is available to the retailer, therefore the behavior of Figure 3 (left pane, blue) is prevented, and the supplier has no longer any incentive of choosing prices far away from $w^\star$. As Figure 6 shows, in this case the dynamics converges directly to $(w^\star, q^\star)$ and the performance of the two agents change accordingly (Figure 7).

ACKNOWLEDGMENTS

Marco Scarsini is a member of GNAMPA-INdAM. This research project received partial support from the COST action CA16228 GAMENET and the Italian MIUR PRIN 2017 Project ALGADIMAR "Algorithms, Games, and Digital Markets." NCB and TC were partially supported by an IBM Global University Program Academic Award. NCB was also partially supported by the FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program.

REFERENCES

[1] Elodie Adida and Victor DeMiguel. 2011. Supply chain competition with multiple manufacturers and retailers. *Oper. Res.* 59, 1 (2011), 156–172.
[2] Kenneth J. Arrow, Theodore Harris, and Jacob Marschak. 1951. Optimal inventory policy. *Econometrica* 19 (1951), 250–272.
[3] Yu Bai, Chi Jin, Huan Wang, and Caiming Xiong. 2021. Sample-efficient learning of Stackelberg equilibria in general-sum games. *Advances in Neural Information Processing Systems* 34 (2021), 25799–25811.
[4] Maria-Florina Balcan, Avrim Blum, Nika Haghtalab, and Ariel D Procaccia. 2015. Commitment without regrets: Online learning in Stackelberg security games. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*. Association for Computing Machinery, New York, NY, USA, 61–78.
[5] Clément Bouttier, Tommaso Cesari, and Sébastien Gerchinovitz. 2020. *Regret analysis of the Piyavskii-Shubert algorithm for global Lipschitz optimization.* Technical Report. arXiv2002.02390.
[6] Gérard P. Cachon and Serguei Netessine. 2006. *Game theory in supply chain analysis.* INFORMS, Online, Chapter 8, 200–233.
[7] Nicolò Cesa-Bianchi, Tommaso Cesari, Takayuki Osogami, Marco Scarsini, and Segev Wasserkrug. 2022. *Online Learning in Supply-Chain Games.* Technical Report. arXiv:2207.04054.
[8] Tsan-Ming Choi (Ed.). 2012. *Handbook of Newsvendor Problems. Models, Extensions and Applications.* Springer, New York, NY.
[9] Yuan Deng, Jon Schneider, and Balasubramanian Sivan. 2019. Strategizing against No-regret Learners. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc., Online.
[10] F. Y. Edgeworth. 1888. The mathematical theory of banking. *Journal of the Royal Statistical Society* 51, 1 (1888), 113–127.
[11] Kaan Gokcesu and Hakan Gokcesu. 2021. *Regret analysis of global optimization in univariate functions with Lipschitz derivatives.* Technical Report. arXiv:2108.10859.
[12] Martin A. Lariviere and Evan L. Porteus. 2001. Selling to the Newsvendor: An Analysis of Price-Only Contracts. *Manufacturing & Service Operations Management* 3, 4 (2001), 293–305.
[13] Steven A. Lippman and Kevin F. McCardle. 1997. The competitive newsboy. *Operations Research* 45, 1 (1997), 54–65.
[14] Siddharth Mahajan and Garrett van Ryzin. 2001. Inventory competition under dynamic consumer choice. *Oper. Res.* 49, 5 (2001), 646–657.
[15] Yishay Mansour, Mehryar Mohri, Jon Schneider, and Balasubramanian Sivan. 2022. Strategizing against Learners in Bayesian Games. In *Proceedings of Thirty Fifth Conference on Learning Theory*, Po-Ling Loh and Maxim Raginsky (Eds.). PMLR, Online, 5221–5252.
[16] Serguei Netessine, Nils Rudi, and Yunzeng Wang. 2006. Inventory competition and incentives to back-order. *IIE Transactions* 38, 11 (2006), 883–902.
[17] Mahmut Parlar. 1988. Game theoretic analysis of the substitutable product inventory problem with random demands. *Naval Res. Logist.* 35, 3 (1988), 397–409.
[18] S.A. Piyavskii. 1972. An algorithm for finding the absolute extremum of a function. *U. S. S. R. Comput. Math. and Math. Phys.* 12, 4 (1972), 57–67.
[19] Pier Giuseppe Sessa, Ilija Bogunovic, Maryam Kamgarpour, and Andreas Krause. 2020. Learning to play sequential games versus unknown opponents. *Advances in Neural Information Processing Systems* 33 (2020), 8971–8981.
[20] Bruno O Shubert. 1972. A sequential method seeking the global maximum of a function. *SIAM J. Numer. Anal.* 9, 3 (1972), 379–388.
[21] Lena Silbermayr. 2020. A review of non-cooperative newsvendor games with horizontal inventory interactions. *Omega* 92 (2020), 102148.
[22] Yunzeng Wang and Yigal Gerchak. 2003. Capacity games in assembly systems with uncertain demand. *Manufacturing & Service Operations Management* 5, 3 (2003), 252–267.