On the Hardness of Learning from Censored and Nonstationary Demand

Gábor Lugosi[*] Mihalis G. Markakis[†] Gergely Neu[‡]

Abstract

We consider a repeated newsvendor problem where the inventory manager has no prior information about the demand, and can access only censored/sales data. In analogy to multi-armed bandit problems, the manager needs to simultaneously "explore" and "exploit" with her inventory decisions, in order to minimize the cumulative cost. Our goal is to understand the hardness of the problem disentangled from any probabilistic assumptions on the demand sequence–importantly, independence or time stationarity–and, correspondingly, to develop policies that perform well with respect to the regret criterion. We design a cost estimator that is tailored to the special structure of the censoring problem and we show that, if coupled with the classic Exponentially Weighted Forecaster, it achieves optimal scaling of the expected regret (up to logarithmic factors) with respect to both the number of time periods and available actions. This result also leads to two important insights: the benefit from "information stalking" as well as the cost of censoring are both negligible, at least in terms of the regret. We demonstrate the flexibility of our technique by combining it with the Fixed Share Forecaster to provide strong guarantees in terms of tracking regret, a powerful notion of regret that uses a large class of time-varying action sequences as benchmark. Numerical experiments suggest that the resulting policy outperforms existing policies (that are tailored to, or facilitated by time stationarity) on nonstationary demand models, with time-varying noise, trend, and seasonality components. Finally, we consider the "combinatorial" version of the repeated newsvendor problem, that is, single-warehouse multi-retailer inventory management of a perishable product. We extend the proposed approach so that, again, it achieves near-optimal performance in terms of the regret.

Keywords: repeated newsvendor problem; demand learning; censored observations; regret minimization; Exponentially Weighted Forecaster.

[*] ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain; Department of Economics and Business, Universitat Pompeu Fabra & Barcelona Graduate School of Economics; gabor.lugosi@gmail.com

[†] IESE Business School, University of Navarra; MMarkakis@iese.edu

[‡] Department of Information and Communication Technologies, Universitat Pompeu Fabra; gergely.neu@gmail.com

1 Introduction

We consider the multi-period inventory management problem of a perishable product, like newspapers, fresh food, or certain pharmaceutical products, where no prior information is available about the demand for the product over the different periods. This may be the case when a new product is introduced to the market, or when the market conditions for an existing product change drastically, e.g., due to a major competitor entering/exiting the market, due to a major economic downturn etc. An additional complication comes from the fact that, often times, the inventory manager cannot observe the actual demand for the product due to *censoring*: the firm is unable to measure or estimate accurately lost sales, so the manager has access only to sales data. However, the sales depend on the manager's prior inventory decisions, making inferences about the underlying demand much harder. In such scenarios, the inventory manager is faced with a dynamic learning problem, having to simultaneously "explore" with her inventory decisions in order to learn the underlying demand, as well as "exploit," that is, focus mostly on decisions that are likely to incur low cost. Due to its practical importance and intellectual challenge, the problem of inventory management with demand learning through censored data has attracted significant attention from the academic community, leading to valuable insights as we detail below.

Our main motivation, and point of departure from existing literature, stems from the fact that the demand for a product may very well be nonstationary: trends and seasonalities are very common in a demand time series; competition in the market that a firm operates may change over time, in terms of both the assortments and the prices offered; consumers may time their decisions strategically. We aim to develop a framework that incorporates in a tractable way the potentially nonstationary nature of demand, and to explore the fundamental limits of performance, that is, to understand the hardness of the dynamic learning problem at hand in a mathematically precise way. Accordingly, we wish to design inventory management policies that perform near-optimally in this setting, and to shed light on the performance loss compared to the case where the demand is time-stationary, as well as the case where the demand is not censored. Hence, we view the problem of inventory management under censored demand as a repeated game between the inventory manager and the market, without making any probabilistic assumptions on the mechanism via which the market generates the demand; more formally, we are interested in providing performance guarantees that hold for every individual sequence of demands. We evaluate the performance of different policies with respect to the regret criterion, that is, the difference between the cumulative cost of a policy and the cumulative cost of the best fixed action in hindsight, and we provide performance guarantees that hold uniformly over all feasible demand sequences. This setting is often termed in the literature as one of "adversarial" demand, due to the minimax regret-optimality guarantees that it lends itself to. Nevertheless, this term may be somewhat misleading, since

Demand	**Stationary, Continuous**	**Stationary, Continuous, density $\geq \alpha$**	**Stationary, Discrete, lost-sales indicator**	**Stationary, Discrete, ϵ-separation**	**Nonstationary, Discrete** (our work)
Upper bound	$O\left(\sqrt{T}\right)$	$O\left(\log T/\alpha\right)$	$O\left(\sqrt{T}\right)$	$O\left(\log T/\epsilon^2\right)$	$O\left(\sqrt{T\log T}\right)$
Lower Bound	$\Omega\left(\sqrt{T}\right)$	$\Omega\left(\log T\right)$	$\Omega\left(\sqrt{T}\right)$	$\Omega\left(\log T\right)$	$\Omega\left(\sqrt{T}\right)$

Table 1: A summary of existing results, contrasted to our findings, regarding the minimax optimal scaling of the regret in the repeated newsvendor problem with demand learning via censored data, with respect to the number of time periods, T; for different sets of probabilistic assumptions on the demand. The reported results constitute a combination of the findings in Huh and Rusmevichientong [2009], which is based on online convex optimization, and in Besbes and Muharremoglu [2013], which is based on alternating exploration and exploitation intervals.

there is no one strategizing against the inventory manager and her decisions in our model.

1.1 Main Contributions

Our primary goal is to understand the hardness of the problem disentangled from any probabilistic assumptions on the demand sequence, which can play a big role in the minimax optimal scaling of the regret, as we illustrate in Table 1. Moreover, we devise inventory management policies that learn from censored data without making any parametric assumptions, and which have guaranteed performance under discrete and nonstationary demand; in fact, near-optimal performance in terms of the regret criterion. Both features are quite important in practice and, jointly, they require a different methodological approach than the ones existing in the literature on stochastic inventory theory: approaches based on alternating exploration and exploitation intervals (AEE) depend crucially on time stationarity, whereas approaches based on online convex optimization (OCO) rely heavily on the continuity of state and action spaces, as well as on the existence of a direction of cost "descent"; for more details, we refer the reader to the Literature Review section. Moreover, in existing works, action space (orders) and outcome space (demand) coincide. In practice, however, there may be only few, predetermined ordering levels, for example, due to fixed ordering costs. Accordingly, we disentangle the two, and provide more refined results that highlight the scaling of the expected regret not only with respect to the number of time periods, but also with respect to the number of ordering decisions available. We note that our approach has guaranteed performance not only with respect to the standard notion of regret, which is based on the best fixed action in hindsight, but also with respect to the tracking regret, which is a much stronger benchmark.

More specifically, we make a direct connection between the *hardness* of the problem at hand and certain properties of its cost and feedback structure. On that end, we employ the theory of partial monitoring, which on the one hand provides a tractable framework for analyzing dynamic learning problems in nonstationary environments; while, on the other hand, admits a clear and complete classification of finite games based on the notion of *local observability*. We establish that the repeated newsvendor problem with censored feedback is a locally observable game, i.e., the difference in cost between any two inventory levels, for any given demand realization, can be determined based just on feedback (sales) from these two decisions. This implies that the correct scaling of the minimax optimal regret is $\Theta\left(\sqrt{T}\right)$, where T is the number of time periods. Nevertheless, the generic algorithms proposed in Bartók et al. [2014] for locally observable games, CBP and Neighborhood Watch, are quite complicated, hard to interpret, and do not provide optimality guarantees regarding the scaling of the regret with respect to the number of available actions, N. The foundations of our approach lie in extending the framework of partial monitoring to achieve the fundamental limits of performance.

Finally, in terms of *prescriptions*, we leverage the special structure of the problem in order to show that simple policies can achieve near-optimal performance; specifically, we show that the Exponentially Weighted Forecaster (EWF) achieves regret that scales as $O\left(\sqrt{T\log T}\right)$, which is optimal up to the logarithmic factor. Even in the case where the demand over different time periods is i.i.d., no better scaling than $\Omega\left(\sqrt{T}\right)$ can be achieved, unless stronger probabilistic assumptions are made; see Table 1. In terms of the number of actions N, we show that the regret of the same policy scales very mildly as $O\left(\log N\right)$. This dependence can be also seen to be near-optimal; see Table 2. Notably, this improves significantly over the existing results in Bartók [2013] and Bartók et al. [2014], whose upper bounds depend polynomially on N. We note that the EWF is a generic algorithm, and any good performance under limited feedback is contingent upon the design of a cost estimator that is tailored to the special structure of the problem. For instance, the case of bandit feedback is a relatively easy one, since the cost estimator relies on updating only the inferred cost of the action taken. The case of censored feedback is more complicated, interpolating, to some extent, the cases of full information and bandit feedback as, in principle, the feedback from the action taken can provide information about the cost of other actions. A cost estimator has not been known for censored feedback, which precisely constitutes our main algorithmic innovation.

We extend the proposed approach so that it has guaranteed performance with respect to the tracking regret, that is, using as benchmark the best sequence of inventory decisions that switches a limited number of times, through the Fixed-Share Forecaster (FSF) algorithm. The tracking regret is a much stronger benchmark, particularly suitable for nonstationary demand models as our numerical experiments suggest. The tradeoff is a somewhat looser bound: while the scaling of the expected tracking regret with respect to T and N remains optimal up to logarithmic terms, the upper bound now includes a multiplicative term that relates to the number of times that the reference sequence is allowed to switch

Feedback Structure	**Censored**	**Full Information**
	(our work)	(Cesa-Bianchi and Lugosi [2006])
Regret Scaling with T	$O\left(\sqrt{T \log T}\right)$	$O\left(\sqrt{T}\right)$
Lower Bound	$\Omega\left(\sqrt{T}\right)$	
Regret Scaling with N	$O\left(\log N\right)$	$O\left(\sqrt{\log N}\right)$
Lower Bound	$\Omega\left(\sqrt{\log N}\right)$	

Table 2: A summary of our setting and findings: we consider a repeated newsvendor problem where the demand is learned from censored observations. We adopt a nonstochastic viewpoint, i.e., the demand is an individual sequence on a finite space. We show that the regret scales as $O\left(\sqrt{T \log T}\right)$ with the number of time periods, and as $O\left(\log N\right)$ with the number of available inventory decisions. The results for the full information case, as well as the lower bounds are obtained from Cesa-Bianchi and Lugosi [2006].

actions.

Our results establish two important *insights* about the dynamic learning problem at hand. First, the cost of censoring is negligible, at least with respect to the regret criterion. In particular, the scaling of the minimax optimal regret would only improve by a root-log factor, with respect to both T and N, if full information on the demand was available; see Table 2. Second, the amount of exploration of the proposed near-optimal policy, in expectation, does not scale either with T or with N. Hence, there is limited benefit to “information stalking,” in contrast to other dynamic learning problems. This explains why myopic policies, with Bayesian updating of parameters, tend to do well in stochastic versions of the repeated newsvendor problem. Intuitively speaking, the two insights solidify and extend the findings in Besbes et al. [2022], where analogous results have been established in a time-stationary setting and for the class of “newsvendor” distributions. Overall though, our findings illustrate that the (non) hardness of the problem is not the outcome of any convenient set of probabilistic assumptions regarding the demand, but rather of the special cost and feedback structure that it possesses.

Finally, in practical terms, we compare the performance of the proposed EWF and FSF policies to the AEE and OCO policies in the existing literature, on numerical experiments with nonstationary demand models, including time-varying noise, trend, and seasonality components. Overall, the performance of the AEE and EWF policies is comparable, but inferior to that of the other two policies, presumably, for different reasons: in the case of the AEE policy, due to its inability to accommodate nonstationarities, by design; in the case of the EWF policy, due to the very conservative updating of beliefs. On the other hand, the FSF policy, with suitably tuned parameters, achieves the best performance in all experiments, while the performance of the OCO policy is consistently good but somewhat inferior.

1.2 Outline of the Paper

The remainder of the paper is organized as follows. Section 2 reviews the related literature on the repeated newsvendor problem with demand learning via censored data, both from the stochastic and the non-stochastic/adversarial viewpoints. Section 3 provides a detailed description of our benchmark model, gives some necessary background on online learning, and presents the proposed inventory management policy accompanied by a regret analysis. Section 4 introduces the notion of tracking regret, and shows that a modification of the proposed policy has guaranteed performance with respect to the latter criterion. This is followed by extensive numerical experiments in Section 5. Section 6 presents the "combinatorial" version of the repeated newsvendor problem, that is, single-warehouse multi-retailer inventory management of a perishable product, again incorporating demand learning through censored observations from a nonparametric viewpoint. We conclude the paper with a brief commentary in Section 7 on problems that are structurally similar to the one treated in this paper, as well as on potential extensions of the proposed approach. Most proofs are relegated to an appendix, at the end of the paper.

2 Literature Review

Stochastic inventory theory with demand learning is a field with a long history and rich literature, going back to early classics such as Scarf [1959] and Karlin [1960]. A detailed account of this literature is beyond the scope of our paper. What is crucial to our work though is the issue of censoring: during stock-outs, however, it is often the case that excess demand is lost, making it very hard to measure or estimate the realized demand. In other words, on many occasions it may be more realistic to assume that the inventory manager has access only to the sales, that is, *censored demand* data. The main insight here is that a dynamic analysis is required even in the case of a perishable product, and that the optimal inventory decision is higher than that of a Bayesian myopic policy, a phenomenon that is referred to in the literature as *information stalking*. The intuition behind it is that this additional inventory gives, occasionally, some extra uncensored demand samples, which contribute towards learning the true parameter values and are, thus, useful in the future. Consequently, some level of "experimentation/exploration" is necessary when dealing with censored demand. This result was first proved in Harpaz et al. [1982] in the context of a perfectly competitive firm making output decisions in the presence of demand uncertainty, and later cast in an inventory management setting and strengthened in Ding et al. [2002] and Lu et al. [2008]. A similar insight has been derived in a dynamic pricing context in Braden and Oren [1994]. Lariviere and Porteus [1999] derives a closed-form expression for the Bayesian optimal inventory level if the demand belongs to the class of "newsvendor distributions" developed in Braden and Freimer [1991], and confirms that it is optimal to enhance learning through stocking higher. Recently, Besbes et al. [2022], building on the framework of Lariviere and Porteus [1999], provides both analytical and numerical evidence to the fact that, while there is cost in being myopic (instead of far-sighted, in the Dynamic Programming sense), this cost is actually quite small. Hence, Bayesian myopic policies are near-optimal in a certain sense, besides being easily implementable. Moreover, the cost of censoring, despite being not too large either, is about an order of magnitude greater than the cost of being myopic, so the inventory manager should direct her efforts in measuring or estimating lost sales.

A fundamental limitation, and the standard criticism against the parametric approach, is that if the parametric family adopted is not broad/flexible enough to capture the underlying demand process, estimating the best parameter values is of little help in really learning the demand, and consequently managing the inventory in a cost-effective way. Hence, in parallel to the aforementioned parametric approach to stochastic inventory theory with uncertainty regarding the demand distribution, a literature following a *nonparametric approach* has also been developing. The papers that come closest to our work are Huh and Rusmevichientong [2009] and Besbes and Muharremoglu [2013]. Both consider the inventory management of a perishable product, in other words, a repeated newsvendor problem over T time periods, and follow the nonparametric approach: the manager has no prior information on the demand - assumed to be independent and identically distributed over different time periods, drawn from an unknown distribution - and has access only to censored demand data. The objective is to minimize the expected regret, that is, the difference between the expected incurred cost and the optimal expected cost, had the demand distribution been known a priori.

Huh and Rusmevichientong [2009] proposes an adaptive inventory management algorithm based on the methodology of online convex optimization (OCO); see Zinkevich [2003]. This algorithm has expected regret that scales as $O\left(\sqrt{T}\right)$, which is the minimax optimal scaling. This can be improved to $O\left(\log T\right)$ if the demand has a continuous density, uniformly bounded away from zero. We note that the stationarity of the underlying demand is not crucial for OCO–in fact, Zinkevich [2003] does not make this assumption–but the fact that demand and inventory are continuous quantities is important: this is a gradient descent-type algorithm, and the continuity of state and action spaces implies that a direction of cost improvement is available almost surely, irrespective of censoring. Consequently, in the case of discrete demand, their methodology requires the existence of a lost-sales indicator to recover the $O\left(\sqrt{T}\right)$ scaling of the expected regret. As we illustrate in our numerical experiments, in the absence of such an indicator, this approach cannot guarantee sublinear regret.

On the other hand, the objective in Besbes and Muharremoglu [2013] is to understand the impact of the available information/feedback structure (fully observable/censored/partially censored demand) on the optimal scaling of the expected regret. In the case of discrete demand and censored observations, which is most relevant to our work and often the case in practice, the authors develop an algorithm based on alternating exploration and exploitation intervals whose expected regret scales as $O\left(\log T\right)$, which is the minimax optimal scaling. The discrepancy between the results in Huh and Rusmevichientong [2009] and Besbes and Muharremoglu [2013] in the case of discrete demand stems from the fact that the latter study makes stronger assumptions regarding the families of distributions allowed. In particular, it is assumed that the expected cost function is not "too flat" around the optimal ordering quantity, that is, the separation between the optimal and the best suboptimal ordering quantity is bounded from below by some ϵ, and the upper bound on the expected regret of their algorithm also scales as $O\left(1/\epsilon^2\right)$. Overall, the proposed algorithm is well equipped to deal with the discreteness of the demand, and the issues it may create when combined with censoring, but the stationarity of the underlying demand process seems to be crucial here: as we illustrate in our numerical experiments, if the demand is nonstationary, then the exploration intervals could lead to very poor inferences, and hence performance during the subsequent exploitation intervals.

2.1 Partial Monitoring

From the vast literature on Online Learning originating form the Machine Learning community, of particular interest to us is the literature on *partial monitoring*, where the information available to the learner is limited in some way. A notable member of this class is the (nonstochastic) multi-armed bandit problem studied in Auer et al. [2002], where the learner knows with certainty the loss of the actions that she takes, but she has no information on the losses of other actions she could have chosen instead. Hence, she has to "explore" in order to learn the losses associated with different actions, and to "exploit" by converging sooner rather than later to the ones she believes have the smallest loss. The main result is that a simple randomized policy, termed the *Exponentially Weighted Forecaster* (EWF), achieves expected regret that scales as $O\left(\sqrt{T}\right)$, which is also the optimal scaling for the particular problem.

The question of the exact dependence of the minimax regret on the problem structure was settled in Bartók et al. [2014], where a complete classification of finite partial monitoring games is provided: "trivial" games where the expected regret does not scale with T; "easy" games where the scaling is $\Theta\left(\sqrt{T}\right)$; "hard" games where the scaling is $\Theta\left(T^{2/3}\right)$; and "hopeless" games with scaling $\Theta\left(T\right)$. Importantly, there can be no other scaling apart from the aforementioned four, despite the fact that the structure of the game, in terms of both the loss and the feedback functions, can be chosen arbitrarily. A geometric condition, termed local observability, is shown to distinguish "easy" from "hard" games, and generic algorithms are developed for each case, albeit significantly more complicated than the EWF.

The framework of *bandits with graph-structured feedback*, first proposed by Mannor and Shamir [2011] and further developed by Alon et al. [2013, 2014], is also related to our work, as it interpolates the cases of full information/experts and bandit feedback: whenever the learner takes an action, she observes not only the loss of the particular action, but also the losses corresponding to an arbitrary subset of the remaining actions. The feedback system characterizing the game is captured by a directed graph, and the fundamental limits of performance as well as the algorithms for this setting relate to properties of the graph. In particular, no policy can achieve expected regret that scales better than $O\big(\sqrt{\alpha(G)T}\big)$, where $\alpha(G)$ is the independence number of the feedback graph.

It is worthwhile to mention the sole paper on the repeated newsvendor problem with demand learning that adopts the nonstochastic framework, Levina et al. [2010]. In their setting, however, there is no censoring, while demand and orders are continuous quantities; both crucial features as explained earlier.

3 Inventory Management with Censored Demand Data

Fix $T, D \in \mathbb{N}$, and define the sets $\mathcal{T} = \{1, 2, \dots, T\}$ and $\mathcal{D} = \{0, 1, \dots, D\}$. Also, fix $N \in \mathbb{N}$, with $N \le |\mathcal{D}| = D + 1$, and let $\mathcal{I}$ be an arbitrary subset of $\mathcal{D}$ with cardinality N. We denote by $\mathbb{I}_E$ the indicator variable of event E, and by $(x)^+$ the maximum of scalar x and 0.

Consider a firm that sells a single perishable product to the market. The following "game" between the firm's inventory manager and the market is repeated over T time periods: at the beginning of period $t \in \mathcal{T}$, the inventory manager chooses an inventory level $I_t \in \mathcal{I}$ to have in stock. (So, implicitly, we assume that there is zero lead time between placing and receiving an order.) Simultaneously, the market chooses the demand that the firm experiences during that period, $d_t \in \mathcal{D}$, which is covered up to the extent that the available inventory allows. At the end of the period any remaining inventory perishes, and the firm incurs a cost

$$c(I_t, d_t) = h(I_t - d_t)^+ + b(d_t - I_t)^+, \tag{1}$$

where $h, b > 0$ represent the overage and underage cost rates, respectively, which are known to the manager and fixed for all time periods.

An important characteristic of our model is that the inventory manager has no information about the demand prior to the beginning of the game. Moreover, she has to "learn" the demand via *censored data*: at the end of period t, the inventory manager can only observe the sales during that period, $\min\{I_t, d_t\}$. In particular, if the inventory I_t turns out to be less than or equal to the demand, then the inventory manager does not know with certainty the exact demand that the firm experienced, d_t, nor the exact cost that it incurred, $c(I_t, d_t)$.

Fix a sequence of demand realizations $\{d_t\}$. We define the *regret* of the inventory manager, for any sequence of inventory decisions $\{I_t\}$, to be the difference between the cumulative cost that is actually incurred and the cost that would have been incurred under the best fixed inventory decision, in hindsight:

$$\mathcal{R}(T) = \sum_{t\in\mathcal{T}} c(I_t, d_t) - \min_{i\in\mathcal{I}} \sum_{t\in\mathcal{T}} c(i, d_t),$$

where $c(i, d_t)$ is defined similarly to Eq. (1). We denote by i^* the minimizer in the equation above, omitting its dependence on the sequence $\{d_t\}$ for convenience.

An important point of our work is that when no probabilistic assumptions are made about the demand, that is, if one adopts the so-called *nonstochastic* viewpoint, then in many cases randomization is the only way to achieve low regret. Of course, under a randomized inventory management policy the regret is a random variable, so our goal is to design policies that have low expected regret:

$$\mathbb{E}[\mathcal{R}(T)] = \mathbb{E}\left[\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}} p_i(t)\big(c(i, d_t) - c(i^*, d_t)\big)\right], \tag{2}$$

where $p_i(t)$ denotes the probability of selecting inventory level $i \in \mathcal{I}$ at the beginning of time period $t \in \mathcal{T}$, conditional on all previous decisions made, that is, $p_i(t) = \mathbb{P}(I_t = i \mid I_1, I_2, \ldots, I_{t-1})$.

3.1 Feedback Structure and Local Observability

The principal challenge associated with our setting is having to deal with censored feedback: instead of the actual costs associated with its decisions, the manager only gets to observe the censored demand. Coupled with our (non-)assumptions about the adversarial demand sequence, this challenge makes it practically impossible to estimate the true demand sequence and the associated costs. Our key observation is that, despite the impossibility of estimating the costs, it is possible to construct a *cost surrogate* that enables us to directly estimate the *differences* between the costs associated with each order level. Partially observable online learning problems where such constructions are possible are called *locally observable* by Bartók et al. [2011], who show that this condition enables regret guarantees of order $\sqrt{T}$. We explain the construction of our cost surrogate below.

Consider an arbitrary pair of inventory decisions $i, j \in \mathcal{I}$, and a demand realization $d \in \mathcal{D}$. We wish to compute the difference between the cost of the two actions. Without loss of generality, assume that $i > j$. (If $i = j$ then, obviously, the difference in cost is zero.) We have that:

(1) $c(i, d) - c(j, d) = h(i - j)$, if $d \le j$;

(2) $c(i,d) - c(j,d) = h(i-d) - b(d-j) = hi + bj - (h+b)d$, if $j < d < i$;

(3) $c(i,d) - c(j,d) = b(j-i)$, if $d \geq i$.

Intuitively, the methodology of partial monitoring in the case of locally observable games expresses the regret between any pair of actions in terms of feedback obtained solely by these actions. Here, in particular, the newsvendor cost function can be written in the form:

$$c(i,d) = hi + bd - (h+b)\min\{i,d\}.$$

Note that this is not known to the manager, in general, as the actual demand may be censored. The essence of local observability is to introduce the surrogate cost structure

$$c'(i,d) = hi - (h+b)\min\{i,d\},$$

which can be computed using only observable quantities (inventory and sales). Importantly, this surrogate cost satisfies:

$$\begin{aligned}\mathcal{R}(T) &\equiv \sum_{t\in\mathcal{T}} c(I_t,d_t) - \min_{i\in\mathcal{I}} \sum_{t\in\mathcal{T}} c(i,d_t)\\ &= \sum_{t\in\mathcal{T}} \big(hI_t - (h+b)\min\{I_t,d_t\}\big) - \min_{i\in\mathcal{I}} \sum_{t\in\mathcal{T}} \big(hi - (h+b)\min\{i,d_t\}\big)\\ &= \sum_{t\in\mathcal{T}} c'(I_t,d_t) - \min_{i\in\mathcal{I}} \sum_{t\in\mathcal{T}} c'(i,d_t).\end{aligned}$$

This transforms the problem into a bandit-like one, something that can be leveraged to achieve expected regret that scales as $O\big(\sqrt{T}\big)$.

3.2 The Exponentially Weighted Forecaster

The Exponentially Weighted Forecasting (EWF) is a well-studied online learning methodology that simultaneously “explores” and “exploits,” in a randomized way. The main idea behind it is to keep track of not only the cost of actions that are actually taken, but also of the estimated cost of all other actions that could have been taken instead. Of course, the specifics of cost estimation are context-specific, as they are closely tied to the type of feedback that the learner receives. Based on the cumulative estimated cost of the different actions, the learner forms beliefs about the chances each of them has being the best one, in hindsight, and prioritizes future actions accordingly.

More concretely, let $\widetilde{c}(i, d_t)$ be the estimated cost that inventory decision $i \in \mathcal{I}$ would have incurred at period $t \in \mathcal{T}$ under demand d_t. Note that, implicitly, $\widetilde{c}(i, d_t)$ may also be a function of the actual inventory decision I_t that was made at period t. In fact, that is the case in the cost estimator that we propose below. Similarly, we define $\widetilde{C}_i(t)$ as the cumulative estimated cost of (fixed) inventory decision $i \in \mathcal{I}$ at period $t \in \mathcal{T}$, with $\widetilde{C}_i(0) = 0$. The cumulative estimated cost can be computed through the recursion:

$$\widetilde{C}_i(t) = \widetilde{C}_i(t-1) + \widetilde{c}(i, d_t), \qquad i \in \mathcal{I}.$$

For convenience, let us also define $W_i(t) = e^{-\eta \widetilde{C}_i(t)}$ and $W(t) = \sum_{i \in \mathcal{I}} W_i(t)$, where η is a positive constant whose exact value depends on the primitives of the problem in a way that is specified later on. Using this notation, we have that

$$W_i(t) = W_i(t-1) e^{-\eta \widetilde{c}(i, d_t)}, \qquad i \in \mathcal{I}. \tag{3}$$

The EWF policy chooses inventory $I_t = i$ with probability

$$p_i(t) = (1-\gamma)\frac{W_i(t-1)}{W(t-1)} + \frac{\gamma}{N}, \qquad i \in \mathcal{I}, \tag{4}$$

where γ is another parameter, in the $(0,1)$ interval, whose precise value will be determined later.

Note that the EWF policy simultaneously "explores" the available action space by making every inventory decision with probability at least γ/N, and "exploits" by assigning higher probability to decisions that have low cumulative estimated cost. The precise way that the inventory manager prioritizes between exploration and exploitation depends on the exact values of the η and γ parameters.

While the EWF is a generic and well-studied policy, what takes advantage of the special structure of the problem at hand is the design of the proper cost estimator $\widetilde{c}(i, d_t)$. To get some insight into what type of estimator may be suitable, let us assume that at period $t \in \mathcal{T}$ the inventory manager decides to hold inventory I_t. At the end of period t, the firm gets (potentially censored) feedback about the demand, that is, the sales $\min\{I_t, d_t\}$. Importantly, this feedback also gives information about the sales that the firm would have had during the particular period, had the inventory manager chosen any $i \le I_t$:

(1) if the feedback was censored, that is, $d_t \ge I_t$, then $\min\{i, d_t\} = i$, for all $i \le I_t$;

(2) if the feedback was not censored, that is, $d_t < I_t$, then the demand d_t is known with certainty and the sales $\min\{i, d_t\}$ can be computed, for all $i \in \mathcal{I}$.

We use this insight to define the estimated cost of action i under demand d_t as follows:

$$\widetilde{c}(i,d_t) = \frac{\mathbb{I}_{\{I_t \geq i\}}}{\mathbb{P}_t(I_t \geq i)} \left(q_i e_{d_t} + \beta\right), \qquad i \in \mathcal{I}, \tag{5}$$

where $\beta = D \cdot \max\{h,b\}$; $\mathbb{P}_t(I_t \geq i) = \sum_{j \in \mathcal{I}: j \geq i} p_j(t)$ according to Eq. (4); e_d is the $(D+1)$-dimensional column vector, with $e_d(j) = 1$ if $j = d$, and 0 otherwise; and q_i is a $(D+1)$-dimensional row vector, whose entries depend only on the primitives of the problem, and which satisfies:

$$q_i e_{d_t} = \begin{cases} h(i - d_t + 1) - b(d_t - 1), & d_t \leq i, \\ -bd_t, & \text{otherwise.} \end{cases}$$

In the Electronic Companion, we show how q_i can be constructed as the product of a cost vector and a signal matrix, so that the problem is formally embedded in the framework of Partial Monitoring.

Theorem 1. *Consider the repeated newsvendor problem described above. The expected regret of the EWF policy with the cost estimator in Eq.* (5) *and parameters*

$$\gamma = \frac{1}{2\beta T}, \qquad \eta = \sqrt{\frac{\log N}{10\beta^2 T \log\left(\frac{3N}{\gamma} + 3\right)}},$$

is bounded from above as follows:

$$\mathbb{E}[\mathcal{R}(T)] \leq 7\beta\sqrt{T \log N \log\left(6\beta T N + 3\right)} + 1.$$

Proof. See Electronic Companion. ■

Theorem 1 implies that the expected regret of the EWF policy scales as $O\left(\sqrt{T \log T}\right)$. On the other hand, the expected regret of any policy in the particular setting scales as $\Omega\left(\sqrt{T}\right)$, even if the demand over different time periods is i.i.d.; see Section 2.5 in Huh and Rusmevichientong [2009] and Section 2.3 in Besbes and Muharremoglu [2013].

Moreover, the expected regret of the EWF policy scales as $O(\log N)$. While the correct scaling with respect to N is not known, the EWF policy cannot be further than a root-log factor off the optimal scaling, as we have a $\Omega\left(\sqrt{\log N}\right)$ lower bound from the full-information case, for the same cost structure; see the discussion in Section 3.3. We note that the best known scaling of the expected regret for locally observable partial monitoring problems is $O\left(\sqrt{N}\right)$, achieved by the LocalExp3 algorithm in Bartók [2013].

Note that since $\beta = D \cdot \max\{h, b\}$, the expected regret of the EWF policy scales as $O\left(D\sqrt{\log D}\right)$. It can be easily verified that the expected regret of any policy scales as $\Omega(D)$ so, again, the performance achieved by the EWF policy is near-optimal.

Observe that the suggested choice of the parameters η and γ involves the total number of time periods T. Thus, in order to implement the suggested policy, the inventory manager needs to know T—or at least an estimate of it—in advance. On the other hand, there are standard ways to relax this assumption in the literature of online learning. In order to avoid tedious but straightforward technicalities, we assume that the inventory manager knows T in advance and refer the reader to Section 2.3 of Cesa-Bianchi and Lugosi [2006].

3.3 Benefit from "Information Stalking" and Cost of Censoring

"Information stalking," an informal term that means to capture the additional exploration that an optimal policy performs compared to reasonable myopic ones in a dynamic learning setting, can be measured in our case by the right-most term in Eq. (4), which determines the frequency of purely exploratory decisions made by the EWF policy. (The other term on the right-hand side of Eq. (4) captures the beliefs of the forecaster about each fixed action being the best one, in hindsight. So, the randomization that it induces is not equivalent to exploration. It is, rather, due to the absence of an underlying probabilistic structure.) Note that with the optimal selection of the γ parameter, this term scales like $1/T$. So, roughly speaking, across all periods the proposed policy is expected to explore only a constant number of times, differing little in that sense from a myopic policy. Moreover, the amount of exploration can be reduced arbitrarily: by choosing $\gamma = 1/T^k$, for any $k > 1$, the number of expected exploratory decisions is decreasing in k, at the cost of a constant term in the upper bound on the expected regret.

Let us also remark on the cost of censoring, that is, the additional cost incurred by having censored observations instead of pure demand samples, as captured by the regret criterion. If there is no censoring in the demand, then one can still use the EWF policy, simply replacing the estimator in Eq. 5 with $c(i, d_t)$, the actual cost that would have been incurred if the manager had held inventory $i \in \mathcal{I}$ at period $t \in \mathcal{T}$. By following similar arguments to the proof of Theorem 1, it can be verified that the expected regret of the EWF policy in that case scales as $O\left(\sqrt{T}\right)$, $O\left(\sqrt{\log N}\right)$, and $O(D)$, respectively, in terms of the three key primitives of our problem. We note that this setting falls into the class of online learning problems with full information, and matching lower bounds are known: for a cost function that is defined as the absolute value of the difference between action and outcome, so essentially the newsvendor cost function in Eq. (1), no policy can achieve scaling of the expected regret that is better than the scaling achieved by the EWF policy, for all three key primitives of the problem; see Theorem 3.7 in Cesa-Bianchi and Lugosi

[2006]. Comparing these to Theorem 1, we have that censoring does not cost more than a root-log factor off the optimal scaling, for all three primitives. In that sense, the cost of censoring is negligible with respect to the regret criterion.

4 Tracking Regret and the Fixed-Share Forecaster

One can argue that the standard notion of regret, which evaluates performance against the best *fixed* action in hindsight, is not suitable for nonstationary environments, such as the ones motivating our work. Hence, in this section, we consider a stronger notion of regret that compares the total cost incurred by a given policy to that of a changing sequence of inventory decisions. Of course, establishing non-trivial performance guarantees is only possible under some restrictions on either the sequence of demands, or the reference sequence of decisions. Besbes et al. [2015] follows the former approach, albeit in an abstract setting of nonstationary stochastic optimization. Inspired by Herbster and Warmuth [1998] we follow the latter approach, and we consider reference sequences that switch between decisions at most S times. Clearly, we cannot expect strong guarantees for values of S comparable to T, so we focus on the case where $S \ll T$, which is also the more relevant in practice: achieving low regret against such comparators intuitively translates to good performance in nonstationary environments, where the demand distribution may change abruptly several times, but otherwise remains roughly stationary for long periods of time; think, for instance, of the demand for seasonal goods, or the demand for a product before and after the entrance of a major supplier in the market.

More formally, let $i_{[T]} = (i_1, i_2, \ldots, i_T) \in \mathcal{I}^T$ be a sequence of inventory decisions, and let

$$\mathcal{C}\left(i_{[T]}\right) = \sum_{t \in \mathcal{T}} \mathbb{I}_{\{i_t \neq i_{t+1}\}}$$

be the complexity of that sequence, that is, the number of times that $i_{[T]}$ switches between two actions. We denote the class of sequences of complexity at most S by

$$\mathcal{I}_S^T = \left(i_{[T]} = (i_1, i_2, \ldots, i_T) : \mathcal{C}(i_{[T]}) \leq S\right).$$

Then, we can define the *tracking regret* against class $\mathcal{I}_S^T$ as

$$\mathcal{R}_S(T) = \sum_{t \in \mathcal{T}} c(I_t, d_t) - \min_{i_{[T]} \in \mathcal{I}_S^T} \sum_{t \in \mathcal{T}} c(i_t, d_t).$$

We propose a simple variant of the EWF policy that aims to minimize the tracking regret. This variant is based on the *Fixed-Share Forecaster* (FSF) introduced in Herbster and Warmuth [1998]; see also Auer et al. [2002], Bousquet and Warmuth [2002], and Cesa-Bianchi et al. [2012]. The key difference compared to the standard EWF is that instead of using Equation (3) for updating the weights, the FSF policy uses the update rule:

$$W_i(t) = W_i(t-1)e^{-\eta\tilde{c}(i,d_t)} + \frac{\alpha}{N}\sum_{j\in\mathcal{I}} W_j(t-1), \tag{6}$$

for all $i \in \mathcal{I}$, where $\alpha > 0$ is a suitably chosen constant. Otherwise, the probabilities $p_i(t)$ and the cost estimates $\tilde{c}(i,d_t)$ are computed as described in the previous section. The following result summarizes the performance guarantee that we can prove for the FSF policy.

Theorem 2. *Consider the repeated newsvendor problem. The expected tracking regret of the FSF policy with the cost estimator in Eq.* (5) *and parameters*

$$\alpha = \frac{1}{T}, \qquad \gamma = \frac{1}{2\beta T}, \qquad \eta = \sqrt{\frac{\log(N/\alpha)}{10\beta^2 T\log\big(3N/\gamma+3\big)}},$$

is bounded from above, for any S, as follows:

$$\mathbb{E}[\mathcal{R}_S(T)] \le 4(S+1)\beta\sqrt{T\log(NT)\log\big(6\beta TN+3\big)}+2.$$

Furthermore, if S is known in advance, then by choosing parameters

$$\alpha = \frac{1}{T}, \qquad \gamma = \frac{1}{2\beta T}, \qquad \eta = \sqrt{\frac{S\log(N/\alpha)}{10\beta^2 T\log\big(3N/\gamma+3\big)}},$$

we have the improved bound:

$$\mathbb{E}[\mathcal{R}_S(T)] \le 7\beta\sqrt{ST\log(NT)\log\big(6\beta TN+3\big)}+2.$$

Proof. See Electronic Companion. ■

We remark that the scaling of the expected tracking regret with respect to both T and N remains optimal up to logarithmic terms. Contrasting the results in Theorems 1 and 2 though, we note that the upper bound in the latter includes a multiplicative term that relates to the number of times that the reference sequence is allowed to switch actions; effectively, the price that one has to pay for having guaranteed performance against a richer class of comparator sequences.

5 Numerical Experiments

In this section, we conduct a numerical study of the EWF and FSF policies, comparing them to the most relevant existing methods in the literature by employing the total cost of each policy as performance criterion. The goal of our experiments is to illustrate the feasibility of the proposed approach in challenging scenarios of nonstationary demand, e.g., time-varying trends, seasonalities, noise, and occasionally big jumps, as well as relatively sparse ordering points, e.g., due to fixed ordering costs. We particularly focus on making comparisons to an inventory management policy based on Alternating Exploration and Exploitation phases (AEE), introduced in Besbes and Muharremoglu [2013], which is designed for the same learning problem, albeit tailored to time-stationary demand settings; and to the online gradient descent-based Adaptive Inventory Management policy (AIM), introduced in Huh and Rusmevichientong [2009], originally designed for continuous state and action spaces.

Sanity Check: A Simple Stationary Demand Model

As a sanity check, we start by reproducing the first numerical experiment in Section 5 of Besbes and Muharremoglu [2013], where the sequence of demands is i.i.d., with each d_t generated independently from a binomial distribution representing 30 independent trials with a success probability of 1/2. We choose $T = 100,000$ and $\mathcal{I} = \mathcal{D} = \{1, 2, \dots, 30\}$. While the performance evaluation criterion in our numerical experiments is consistently the total cost of each policy, we make an exception here by plotting the regret, so that the results can be directly compared to the findings in Besbes and Muharremoglu [2013]. Moreover, to gain some insight into the cost of censoring, we study both the uncensored and the censored versions of each of the policies described above. Note that in the full-information case, the policy proposed in Besbes and Muharremoglu [2013] has no reason to explore. Rather, it exploits in a greedy fashion by ordering the empirical critical quantile. We term the resulting policy "greedy-full." We exclude the FSF policy from this set of experiments, since a nonstationary comparator sequence does not make sense in stationary environments.

In Figure 1, each reported curve represents the average of 100 simulation runs, with shaded areas representing the standard deviations. In this setting, the AEE policy outperforms the others by a wide margin. In particular, in both the censored and the uncensored cases, the regret of the EWF policy is about an order of magnitude greater than that of the AEE policy. This observation is not surprising: EWF is decidedly more conservative because it is primarily designed for nonstationary environments. The inferior performance of general no-regret algorithms, like the EWF, in time-stationary environments has been widely acknowledged in the online learning literature, with some remedies offered by the works of

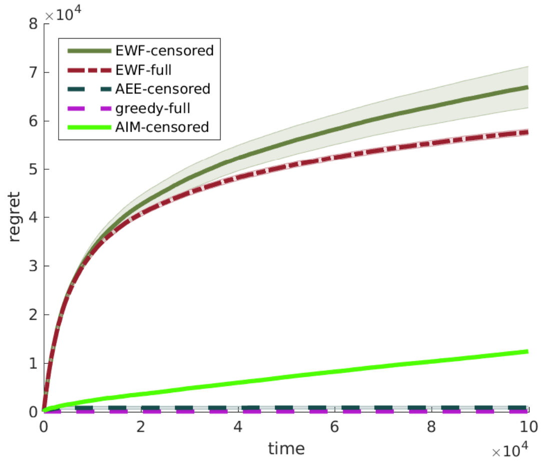

Figure 1: The regret of the different policies on a stationary demand series.

Sani et al. [2014], Van Erven et al. [2014]. Nevertheless, the empirical regret of the EWF policy grows in a square-root fashion, in line with our theoretical guarantees.

As a side note, we observe that in this simple, time-stationary scenario, the cost of censoring is negligible for the AEE policy while it scales mildly for the proposed EWF policy, in agreement with our theoretical analysis.

A Simple Nonstationary Demand Model

Our second set of numerical experiments considers a simple nonstationary demand sequence. As before, the demands are generated independently from a binomial distribution with 30 trials, albeit with a time-varying success probability q_t. Similarly to the previous experiment, we set $q_t = 1/2$ for most values of t, however, this probability drops to 0.1 for $t \in [T/5, T/2]$. Again, we choose $T = 100{,}000$ and $\mathcal{I} = \mathcal{D} = \{1, 2, \dots, 30\}$. The performance of each policy is shown in Figure 2.

The first lesson from this experiment is that the AEE policy fails to cope with the shifting demand distribution, incurring a linearly increasing regret. The intuition behind this is simple: since the policy, effectively, collects data in deterministically (and scarcely) scheduled exploration periods, it is easily thrown off its tracks by a shifting demand distribution. In the particular experiment, the AEE policy bases all its decisions during the interval $[T/2, T]$ on data that has nothing to do with the actual demand distribution. The policy has no mechanism to recover from such mistakes, and introducing such a mechanism while maintaining strong performance guarantees, is far from trivial.

In contrast, the EWF policy is robust to this nonstationary behavior. Notably, its cumulative cost in the censored case, using our carefully designed cost estimator, is remarkably close to that in the full-

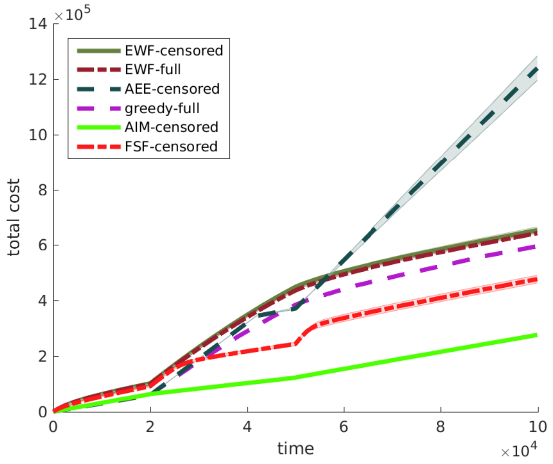

Figure 2: The total cost of the different policies on a simple nonstationary demand sequence.

feedback case. This confirms our main insight, that censoring has minimal impact on the performance of well-designed learning policies in this setting. We also highlight the good performance of the FSF policy: ran with $S = 3$, that is, correctly anticipating 3 shifts in the comparator sequence, this policy is seen to react much quicker to the distributional shifts compared to the standard EWF variant, thus achieving superior performance.

Finally, the performance of the AIM policy is pretty good in both experiments; in fact, superior to that of the EWF/FSF policies. Given its algorithmic simplicity, it presents itself as a good candidate for practical use. The issue with AIM is the absence of guaranteed performance without a lost-sales index and the problems that may arise then, which we detail in a following section.

More Challenging Nonstationary Demand Models

The bulk of our numerical experiments concerns progressively more challenging nonstationary demand models. In the sets of experiments that we term "mild nonstationarity," we consider time-varying trend, seasonality, and noise terms, in different combinations; while in the "significant nonstationarity" experiments, we add big jumps on top of the aforementioned terms. We repeat those sets of experiments for the parameter values that theory prescribes (i.e., for which the policies have guaranteed performance), as well as for heuristically optimized parameter values in an attempt to boost overall performance.[1] We assume that ordering points are sparse, as is often the case in practice. Specifically, we choose $T = 1000$, $\mathcal{D} = \{0, 1, \dots, 200\}$, and $\mathcal{I} = \{0, 100, 200\}$.

[1] Regarding the important S parameter of the FSF policy, which can be selected arbitrarily, we use the same values in all sets of experiments (1, 2, 2, and 3, respectively, in the four experiments of each set, to reflect roughly their increasing complexity), and we do *not* optimize further.

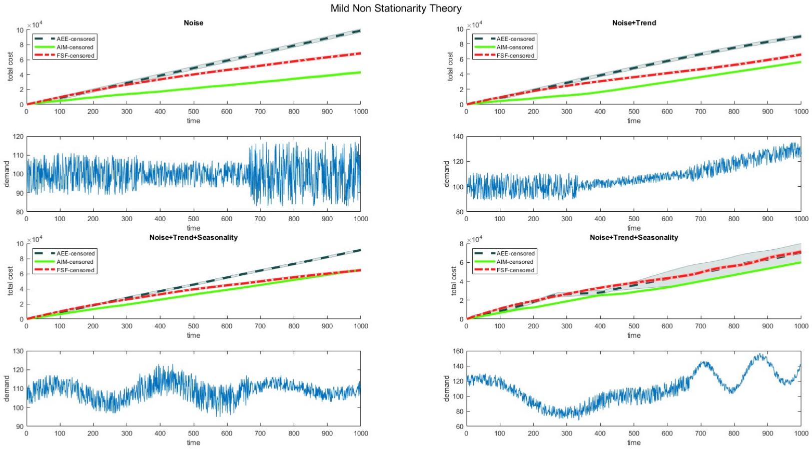

Figure 3: The total cost of the different policies for a variety of demand patterns that exhibit "mild" nonstationarity, including noise and/or trend and/or seasonality terms. The parameters of all policies are set to the values for which they have guaranteed performance.

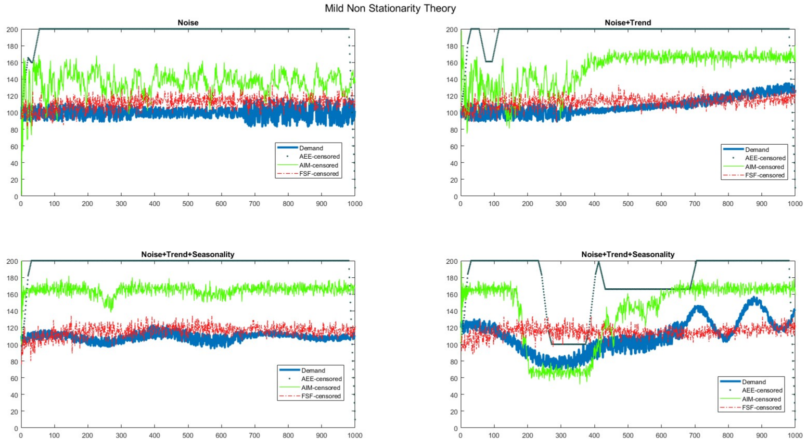

Figure 4: The actions of the different policies (average over 100 runs) for a variety of demand patterns that exhibit "mild" nonstationarity, including noise and/or trend and/or seasonality terms. The parameters of all policies are set to the values for which they have guaranteed performance.

The performance of the three policies (AEE, AIM, and FSF) is shown in Figures 3, 5, 7, and 9, while in Figures 4, 6, 8, and 10, we present the corresponding actions in the various experiments, in an attempt to provide some insight into their behaviour. We note that, in the latter figures, we present the average of the actions over 100 simulation runs, smoothed out to facilitate visualization. For that reason, what appears in the figures as actions is not confined to the $\mathcal{I} = \{0, 100, 200\}$ action space.

We observe that, under "mild stationarity," AIM has a small advantage over FSF, while both outper-

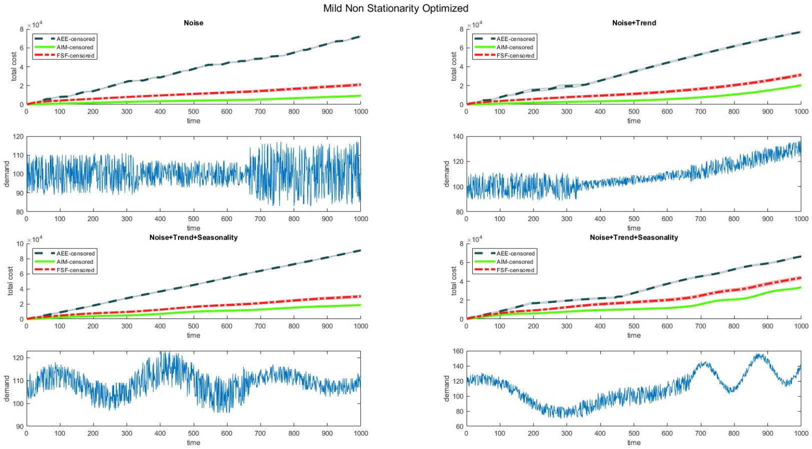

Figure 5: The total cost of the different policies for a variety of demand patterns that exhibit “mild” nonstationarity, including noise and/or trend and/or seasonality terms. The parameters of all policies are heuristically optimized, i.e., for each policy, a grid search has been performed to identify the values that give the best overall performance in the four experiments.

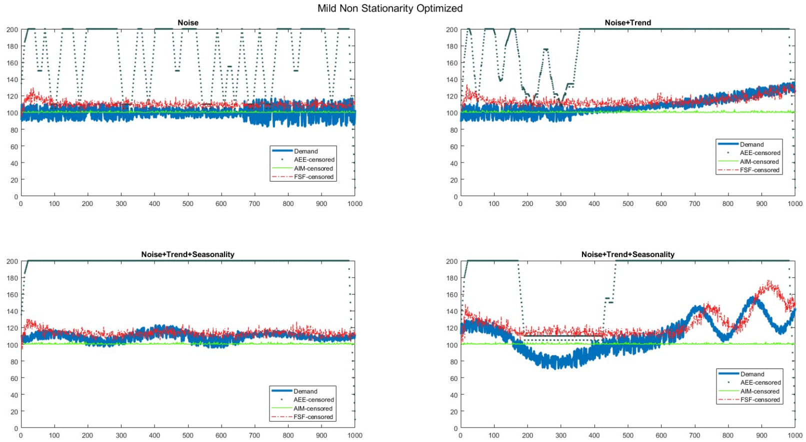

Figure 6: The actions of the different policies (average over 100 runs) for a variety of demand patterns that exhibit “mild” nonstationarity, including noise and/or trend and/or seasonality terms. The parameters of all policies are heuristically optimized, i.e., for each policy, a grid search has been performed to identify the values that give the best overall performance in the four experiments.

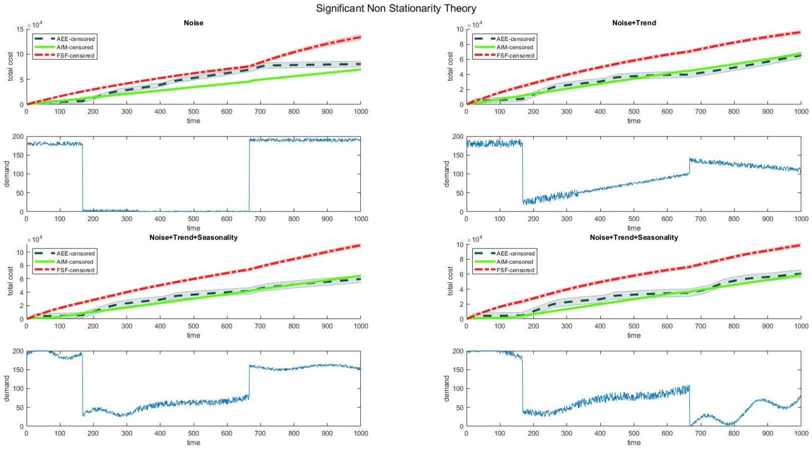

Figure 7: The total cost of the different policies for a variety of demand patterns that exhibit "significant" nonstationarity, including big jumps on top of noise and/or trend and/or seasonality terms. The parameters of all policies are set to the values for which they have guaranteed performance (for AEE and AIM, this performance concerns time-stationary settings).

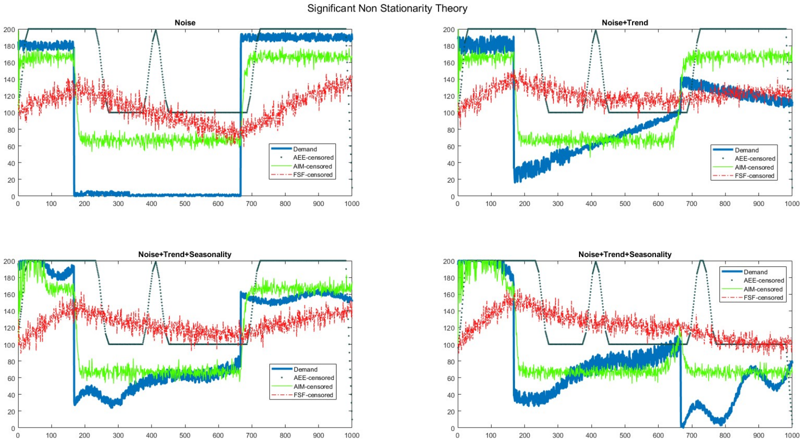

Figure 8: The actions of the different policies (average over 100 runs) for a variety of demand patterns that exhibit "significant" nonstationarity, including big jumps on top of noise and/or trend and/or seasonality terms. The parameters of all policies are set to the values for which they have guaranteed performance (for AEE and AIM, this performance concerns time-stationary settings).

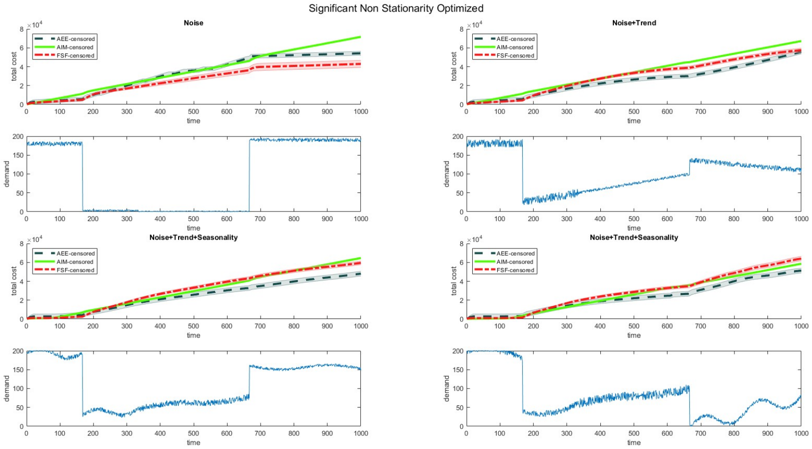

Figure 9: The total cost of the different policies for a variety of demand patterns that exhibit “significant” nonstationarity, including big jumps on top of noise and/or trend and/or seasonality terms. The parameters of all policies are heuristically optimized, i.e., for each policy, a grid search has been performed to identify the values that give the best overall performance in the four experiments.

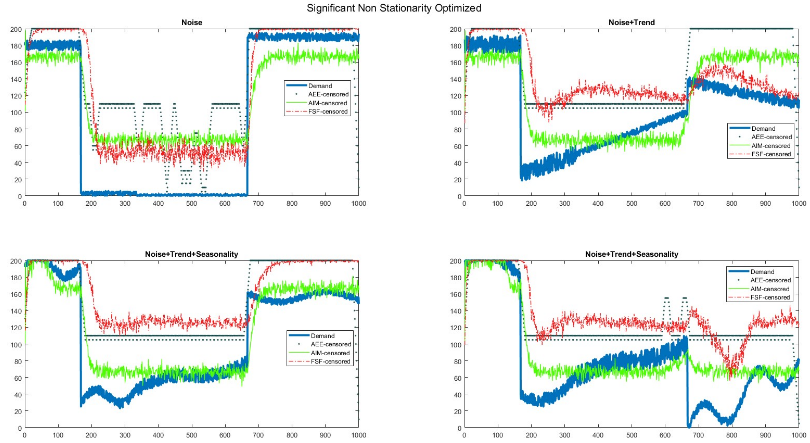

Figure 10: The actions of the different policies (average over 100 runs) for a variety of demand patterns that exhibit “significant” nonstationarity, including big jumps on top of noise and/or trend and/or seasonality terms. The parameters of all policies are heuristically optimized, i.e., for each policy, a grid search has been performed to identify the values that give the best overall performance in the four experiments.

form significantly the AEE policy, both under the theoretically prescribed and the heuristically optimized parameter values. None of the two findings is too surprising: FSF is conservative by design, in order to maintain reasonable performance even in very complex demand patterns, which comes at a performance cost; while AEE is not designed to accommodate nonstationarities, in general.

On the other hand, under "significant stationarity," AIM and AEE outperform the FSF policy for the theoretically prescribed parameter values, while there is no clear winner or loser among the three for optimized parameter values, with their performance being close overall. These findings merit some discussion. Regarding the relatively poor performance of FSF in the former set of experiments, Figure 8 shows that the parameter values for which FSF has guaranteed performance force it to adapt quite slowly to the big jumps in the demand, relative to the size of those jumps and the length of the horizon. AIM and AEE adapt much faster to these jumps, as their theory-prescribed parameter values are designed for time-stationary settings. Moreover, the mere size of those jumps seems reduce the importance of the other nonstationary terms (and the ability, or lack thereof, of the different policies to accommodate them). Finally, a more aggressive tuning of the parameters of the FSF policy helps it to catch up to the other policies, performance-wise. We note that the performance of AEE and AIM does not change much when their parameter values change from theory to optimized. Again, the reason is that these policies do not have guaranteed performance in nonstationary settings, so their theory-prescribed parameters are borrowed from their analyses in time-stationary settings which, typically, is relatively aggressive already.

The Misbehavior of the AIM Policy

The AIM policy maintains consistently good performance in all numerical experiments so far. Combined with the fact that it is algorithmically simpler than AEE and FSF, this makes it the logical candidate for practical implementation. While this may very well be the case in many real-life scenarios, one should be mindful of the potential consequences of the absence of guaranteed performance in discrete nonstationary environments, on which we elaborate below.

In the case of discrete (i.i.d.) demand—see the AIM-discrete variant, in Section 3.4 of Huh and Rusmevichientong [2009]—the policy has a guaranteed performance only in the "partially censored" case, that is, if the learner has access to a lost-sales indicator. It is based on the online gradient descent method of Zinkevich [2003]: the key idea is to compute recursively the sequence of auxiliary points $x_{t+1} = \Pi\left(x_t - \alpha_t g_t\right)$, where g_t is an estimate of the left derivative of the loss function $c(\cdot, d_t)$, evaluated at x_t, and Π is a projection operator on the interval $[0, D]$. Recall that $c(x, d_t)$, as defined in Equation (1), is convex, so it has a well-defined left derivative: $-h$ for $x < d_t$, b for $x > d_t$, and $[-h, b]$ for $x = d_t$. With the help of this sequence of auxiliary points, the actual (discrete) inventory levels are chosen randomly

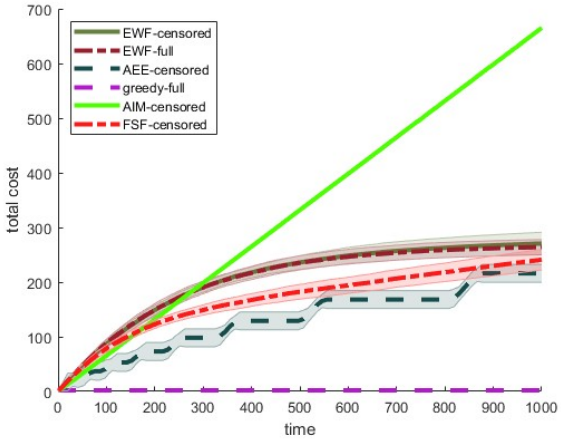

Figure 11: The misbehavior of the AIM policy.

according to $I_t = \lfloor x_t \rfloor + B_t$, where B_t is a Bernoulli random variable with expectation $x_t - \lfloor x_t \rfloor$, so that $\mathbb{E}_t[I_t] = x_t$.

The most pertinent question concerns the estimation of the gradients of $c(\cdot, d_t)$ in a reliable way. Huh and Rusmevichientong [2009] suggest to use

$$g_t = \begin{cases} -b + (h+b)\,\mathbb{1}_{\{d_t \le I_t\}}, & \text{if } I_t = \lfloor x_t \rfloor \\ -b + (h+b)\,\mathbb{1}_{\{d_t \le I_t - 1\}}, & \text{if } I_t = \lceil x_t \rceil, \end{cases}$$

which is an unbiased estimator of the left derivative. Crucially, however, the construction of this operator requires access to the lost sales indicator $\mathbb{1}_{\{d_t \le I_t\}}$, which on many occasions may not be available. In the absence of such an indicator, the most natural choice for estimating the derivative would be

$$\widehat{g}_t = -b + (h+b)\,\mathbb{1}_{\{d_t \le I_t - 1\}},$$

instead of g_t in the policy described above; this is the scheme that we implement in the numerical experiments above. We note that, to our knowledge, there is no alternative gradient estimator in the literature that uses just censored/sales data. While $\widehat{g}_t$ can be computed with just censored data, it is easy to see that it is a *biased* estimator of the left derivative. In particular, since $\mathbb{1}_{\{d_t \le I_t - 1\}} > \mathbb{1}_{\{d_t \le I_t\}}$, $\widehat{g}_t$ consistently underestimates the true derivative. This shortcoming severely impacts the behavior of the policy: even for constant demand realizations, it always overshoots the optimal inventory level. This effect is illustrated in a very simple scenario in Figure 11, with three possible order levels, $\{0, 1, 2\}$, and a constant demand of 1: the AIM policy has a linearly increasing regret, which is the same as the total cost in this case.

Summarizing, in the absence of performance guarantees, which is the case for AIM when a lost-sales indicator is not available, it is hard to know in advance how well will the policy perform. It is reasonable, of course, to conjecture that as the demand becomes more fine-grained, the value of the lost-sales indicator decreases and, thus, the AIM policy performs better; see also Section 5.2 in Besbes and Muharremoglu [2013]. However, apart from few numerical experiments, little else is known in terms of quantifying the rate at which the discrete setting "converges" to a continuous one, and the corresponding effect on the performance of online gradient descent-based policies. This simply reiterates the point about performance guarantees.

6 The Single-Warehouse Multi-Retailer Problem

In this section we extend our benchmark model to include the inventory management problem of a perishable product, for a vertically integrated supply chain with a single warehouse and K retailers. Again, the demand in each of the K retailers needs to be learned from censored/sales data, and a nonstochastic view of the problem is adopted. As this setting has many commonalities with the model of Section 2, for brevity, we only present the points at which they differ.

Let us denote by $\mathcal{K}$ the set of retailers $\{1,2,\ldots,K\}$. At the beginning of period $t \in \mathcal{T}$, the inventory manager allocates $I_t^{(k)} \in \mathcal{I}$ units of inventory from the warehouse to retailer $k \in \mathcal{K}$. If $I_t^{(k)}$ is not zero, then the retailer incurs a fixed ordering cost of $f^{(k)}$. We assume zero lead times, so the inventory is delivered to the retailer instantaneously. The retailer experiences demand $d_t^{(k)} \in \mathcal{D}$ during the particular period. At the end of the period, and depending on the initial inventory and the demand, the retailer incurs overage or underage cost at rates $h^{(k)}$ and $b^{(k)}$, respectively, and any remaining inventory perishes. Thus, at the end of time period $t \in \mathcal{T}$, retailer $k \in \mathcal{K}$ incurs a total cost of

$$c_k\left(I_t^{(k)}, d_t^{(k)}\right) = f^{(k)}\mathbb{1}_{\left\{I_t^{(k)}>0\right\}} + h^{(k)}\left(I_t^{(k)} - d_t^{(k)}\right)^+ + b^{(k)}\left(d_t^{(k)} - I_t^{(k)}\right)^+.$$

We assume that the supply chain operates in a "push" manner, from upstream to downstream. More specifically, the upstream supplier replenishes the inventory of the warehouse with $r > 0$ units at the beginning of every time period. The inventory manager allocates different parts of this inventory to the different retailers, but may also have an incentive to keep a part of it at the warehouse, if she believes that the total demand at the retailer level is low. (The overage cost rate usually increases as one moves downstream.) Hence, the cost that the warehouse incurs over time period t is equal to

$$c_0\left(I_t^{(1)},\dots,I_t^{(K)},r\right) = f^{(0)} + h^{(0)}\left(r - \sum_{k\in\mathcal{K}} I_t^{(k)}\right)^+ .$$

The allocation that the inventory manager can make must belong to the set

$$\mathcal{A}_r = \left\{\left(i^{(1)},\dots,i^{(K)}\right)\in\mathcal{I}^K : \sum_{k\in\mathcal{K}} i^{(k)} \le r\right\}.$$

The goal of the inventory manager is to minimize the total cost incurred by the warehouse and the retailers throughout the T periods. More concretely, the demand sequences $\left\{d_t^{(k)}\right\}$, $k\in\mathcal{K}$, and the inventory replenishment level r are exogenously determined by the market and the upstream supplier, respectively, and the manager wishes to minimize her expected regret, $\mathbb{E}[\mathcal{R}_c(T)]$, over the best fixed ($K$-dimensional) inventory decision, in hindsight.

We refer to this setting as the “combinatorial” version of the repeated newsvendor problem, due to the fact that the manager’s inventory decisions have a combinatorial nature, taking values in $\mathcal{A}_r$. To the best of our knowledge, this setting has not been studied before from the angle of demand learning via censored data, and from a nonstochastic viewpoint. It may be worthwhile to mention, however, some high-level similarity to Shi et al. [2016], where the results in Huh and Rusmevichientong [2009] are extended to the capacitated multi-product case. The model analyzed in Shi et al. [2016] is also a repeated newsvendor model with a combinatorial action space, but the approach is quite different: in order to obtain results for the (arguably harder) case of nonperishable products, strong probabilistic assumptions are made about the demand for the different products.

For convenience, let us define the quantities

$$\beta = D\cdot\max\left\{\max_{k\in\{0\cup\mathcal{K}\}}\left\{h^{(k)}\right\},\max_{k\in\mathcal{K}}\left\{b^{(k)}\right\}\right\},$$

and

$$f = \max_{k\in\{0\cup\mathcal{K}\}}\left\{f^{(k)}\right\}.$$

In what follows, we mimic the approach of Section 2, to construct an inventory management policy that performs well with respect to the regret criterion. At a high level, the proposed policy follows the EWF scheme, but unlike the versions discussed in previous sections, it draws actions in a non-uniform way during the exploration rounds.

Specifically, in each round the policy explores with probability γ. The exploration procedure generates a random allocation by selecting a retailer index κ uniformly at random from $\mathcal{K}$, and by choosing the order level $I^{(\kappa)}$ uniformly at random from $\mathcal{I}$. The rest of the allocations can be completed arbitrarily; for simplicity, we choose an order level of 0 for the remaining retailer. The probability of choosing the allocation $\left(i^{(1)},\dots,i^{(K)}\right)$ is denoted by $\mu_{\left(i^{(1)},\dots,i^{(K)}\right)}$.

Now let us describe the main component of the policy, which is computing the weights of the assignments. Let $\widetilde{c}\left(i^{(1)},\dots,i^{(K)},r,d_t^{(1)},\dots,d_t^{(K)}\right)$ be the estimated cost that inventory decision $\left(i^{(1)},\dots,i^{(K)}\right)$ would have incurred at period $t\in\mathcal{T}$. We compute, recursively, the quantities

$$W_{\left(i^{(1)},\dots,i^{(K)}\right)}(t) = W_{\left(i^{(1)},\dots,i^{(K)}\right)}(t-1)\,e^{-\eta\widetilde{c}\left(i^{(1)},\dots,i^{(K)},r,d_t^{(1)},\dots,d_t^{(K)}\right)},$$

where $\left(i^{(1)},\dots,i^{(K)}\right)\in\mathcal{A}_r$, with $W_{\left(i^{(1)},\dots,i^{(K)}\right)}(0)=1$. We also use the shorthand notation

$$W(t) = \sum_{\left(i^{(1)},\dots,i^{(K)}\right)\in\mathcal{A}_r} W_{\left(i^{(1)},\dots,i^{(K)}\right)}(t).$$

In this setting, the EWF policy chooses inventory $\left(I_t^{(1)},\dots,I_t^{(K)}\right)=\left(i^{(1)},\dots,i^{(K)}\right)\in\mathcal{A}_r$ with probability

$$p_{\left(i^{(1)},\dots,i^{(K)}\right)}(t) = (1-\gamma)\frac{W_{\left(i^{(1)},\dots,i^{(K)}\right)}(t-1)}{W(t-1)} + \gamma\mu_{\left(i^{(1)},\dots,i^{(K)}\right)}.$$

Similarly to Section 2, the proper values for parameters η and γ are specified during the analysis stage.

The special structure of the problem is exploited by designing a suitable cost estimator. For every $\left(i^{(1)},\dots,i^{(K)}\right)\in\mathcal{A}_r$, let

$$\widetilde{c}_k\left(i^{(k)},d_t^{(k)}\right) = \frac{\mathbb{1}_{\left\{I_t^{(k)}\geq i^{(k)}\right\}}}{\mathbb{P}_t\left(I_t^{(k)}\geq i^{(k)}\right)}\left(q_i e_{d_t^{(k)}} + f^{(k)}\mathbb{1}_{\{i^{(k)}>0\}} + \beta\right),$$

where q_i is the same as in Eq. (5), and

$$\mathbb{P}_t\left(I_t^{(k)}\geq i^{(k)}\right) = \sum_{\left(j^{(1)},\dots,j^{(K)}\right)\in\mathcal{A}_r : j^{(k)}\geq i^{(k)}} p_{\left(j^{(1)},\dots,j^{(K)}\right)}(t).$$

Through this, we define the estimated cost of inventory decision $\left(i^{(1)},\dots,i^{(K)}\right)\in\mathcal{A}_r$ as

$$\widetilde{c}\left(i^{(1)},\dots,i^{(K)},r,d_t^{(1)},\dots,d_t^{(K)}\right) = c_0\left(i^{(1)},\dots,i^{(K)},r\right) + \sum_{k\in\mathcal{K}}\widetilde{c}_k\left(i^{(k)},d_t^{(k)}\right). \tag{7}$$

Note that the proposed estimator is designed in order to be unbiased in terms of inferring the difference in expected cost between two decisions:

$$\begin{aligned}&\mathbb{E}_t\left[\widetilde{c}\left(i^{(1)},\ldots,i^{(K)},r,d_t^{(1)},\ldots,d_t^{(K)}\right)-\widetilde{c}\left(j^{(1)},\ldots,j^{(K)},r,d_t^{(1)},\ldots,d_t^{(K)}\right)\right]\\&=c\left(i^{(1)},\ldots,i^{(K)},r,d_t^{(1)},\ldots,d_t^{(K)}\right)-c\left(j^{(1)},\ldots,j^{(K)},r,d_t^{(1)},\ldots,d_t^{(K)}\right),\end{aligned}\tag{8}$$

for every $\left(i^{(1)},\ldots,i^{(K)}\right),\left(j^{(1)},\ldots,j^{(K)}\right)\in\mathcal{A}_r$; a consequence of the local observability property of the feedback structure in the single-retailer setting. Furthermore, it is easy to show that the mean of the estimator satisfies

$$0\le\mathbb{E}_t\left[\widetilde{c}\left(i^{(1)},\ldots,i^{(K)},r,d_t^{(1)},\ldots,d_t^{(K)}\right)\right]\le f+K(2\beta+f).\tag{9}$$

The following theorem establishes a performance guarantee of the proposed allocation policy.

Theorem 3. *Consider the "combinatorial" version of the repeated newsvendor problem described above. The expected regret of the EWF policy with the cost estimator in Eq.* (7) *and parameters*

$$\gamma=\frac{1}{T},\qquad\eta=\sqrt{\frac{\log N}{\beta^2KT\log(TNK+3)}},$$

is bounded from above as

$$\mathbb{E}[\mathcal{R}(T)]=O\left(K^{3/2}\beta\sqrt{T\log N\log(TNK)}\right).$$

Proof. See Electronic Companion. ■

As in the single-warehouse setting, the scaling of the expected regret of the EWF policy is near-optimal with respect to both T and N, similarly to the single-retailer setting in Section 2. Again, we highlight the fact that the logarithmic scaling with respect to N is made possible by our carefully designed estimator that assigns non-zero entries to all inventory levels below the realized inventory decision. On the other hand, the $O\left(K^{3/2}\right)$ scaling with respect to the number of retailers is unlikely to be optimal. To see this, we note that the combinatorial version of the repeated newsvendor problem falls into the broader class of online combinatorial optimization, for which the limitations of the EWF policy are now clearly understood. In particular, Audibert et al. [2014] shows that the EWF policy has, in general, suboptimal performance guarantees in terms of the size of the decision set, essentially translating to a suboptimal scaling with respect to K in our case. We conjecture that the optimal scaling with respect to K is linear, and it can be achieved by a suitable adaptation of the Component Hedge algorithm in Koolen et al. [2010] (also called Online Stochastic Mirror Descent in Audibert et al. [2014]). We omit a detailed treatment of that direction in order to maintain the clarity of our presentation. We also remark that a similar combination of feedback graphs and combinatorial decision sets is studied in Kocák et al. [2014]. An adaptation of their algorithm, termed FPL-IX, combined with our cost estimates, can be shown to satisfy a regret bound identical to our Theorem 3. Details are again omitted for brevity.

Note that from a mathematical standpoint, the role of the cost that the warehouse incurs at every time period is to couple the inventory management problems of the different retailers. However, the exact functional form of $c_0\left(I_t^{(1)}, \ldots, I_t^{(K)}\right)$ is never used. Thus, the approach presented above extends in a straightforward way to the case where the supply chain operates in a “pull” manner, from downstream to upstream. In that case, the inventory manager has no incentive to keep any inventory at the warehouse, since the product is perishable. The only cost that the warehouse incurs is a fixed ordering cost of $f^{(0)}$ for procuring the required inventory from upstream:

$$c_0\left(I_t^{(1)}, \ldots, I_t^{(K)}\right) = f^{(0)} \mathbb{1}_{\left\{\sum_{k\in\mathcal{K}} I_t^{(k)}>0\right\}}.$$

With this minor modification, the rest of the analysis follows verbatim.

7 Concluding Remarks

We conclude the paper by drawing connections between our setting and two related problems that have recently attracted the attention of academic research.

The first problem is bidding in repeated second-price auctions with valuation learning: a bidder participates in second-price auctions for different products, with her valuations of these products being unknown a priori. The bidder learns her true valuation of a given product only if she wins the respective auction, that is, if she submits the highest bid, and her reward in that case is equal to the difference between that valuation and the second highest bid; otherwise, the bidder learns nothing about her valuation and earns no reward. The goal of the bidder is to maximize her expected reward, which translates naturally into an exploration-exploitation tradeoff in her bidding strategy. Weed et al. [2016] studies this problem from a nonstochastic viewpoint, and proposes a bidding policy whose expected regret scales optimally with respect to the number of auctions. There is an intriguing similarity to our problem, in that the learner receives feedback only when her action is greater than the opponent’s, in which case the problem reduces to a full-information one. The cost structure, however, is quite different: in a newsvendor setting the amount of cost incurred depends on the inventory decision of the manager, whereas in a second-price auction the size of the reward is independent of the bid, conditional on winning/not winning the auction. The authors take advantage of this property by introducing a variation of the EWF policy that is based on interval splitting. Their approach does not seem applicable to our setting, and our cost estimator is, not surprisingly, quite different.

The second problem is sequential stock allocation to dark pools: a trader receives amounts of stocks to liquidate at different periods, and allocates these amounts among different dark pools, whose demand for the particular stocks is unknown a priori. The trader learns the demand at a certain dark pool only if the amount of stocks allocated there exceeds it. This problem is formulated as a dynamic learning problem with limited feedback in Ganchev et al. [2010], and analyzed from a nonstochastic viewpoint with regret guarantees in Agarwal et al. [2010]. It is not hard to see that this is almost identical to the "combinatorial" newsvendor problem in Section 6, with the main difference being that the amount of inventory that the warehouse receives is constant, whereas the amount of stocks that the trader tries to liquidate may vary. (Relatedly, the notion of regret that the authors adopt is somewhat different than ours.) Hence, one can argue that we analyze a special case of the problem in Agarwal et al. [2010]. On the other hand, the expected regret of the policy proposed in the latter paper, for the case of integral allocations, scales as $O\left(T^{2/3}\right)$, a result which we improve considerably upon by taking advantage of the local observability property.

Acknowledgements

The authors would like to thank Claudio Sánchez for his assistance with the numerical experiments of the paper. G. Lugosi was supported by the Spanish Ministry of Economy, Industry and Competitiveness through the grant Grant MTM2015-67304-P (AEI/FEDER, UE). M.G. Markakis was supported by the Spanish Ministry of Economy and Competitiveness through the ECO2016-75905-R (AEI/FEDER, UE) grant and a Juan de la Cierva fellowship, as well as the Spanish Ministry of Science and Innovation through a Ramón y Cajal fellowship. G. Neu was supported by the UPFellows Fellowship (Marie Curie COFUND program n° 600387).

References

Alekh Agarwal, Peter Bartlett, and Max Dama. Optimal allocation strategies for the dark pool problem. *Proceedings of the* 13^{th} *International Conference on Artificial Intelligence and Statistics*, pages 9–16, 2010.

Noga Alon, Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. From Bandits to Experts: A Tale of Domination and Independence. In *Neural Information Processing Systems*, pages 1610–1618, 2013.

Noga Alon, Nicolò Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback, September 2014. URL `https://arxiv.org/abs/1409.8428`.

Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2014.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Gábor Bartók. A near-optimal algorithm for finite partial-monitoring games against adversarial opponents. *Proceedings of the* 26^{th} *Annual Conference on Learning Theory*, 30:1–15, 2013.

Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Minimax regret of finite partial-monitoring games in stochastic environments. In S. Kakade and U. von Luxburg, editors, *Proceedings of the 24th Annual Conference on Learning Theory*, 2011.

Gábor Bartók, Dean P. Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring – classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014.

Omar Besbes and Alp Muharremoglu. On implications of demand censoring in the newsvendor problem. *Management Science*, 59(6):1407–1424, 2013.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations Research*, 63(5):1227–1244, 2015.

Omar Besbes, Juan Chaneton, and Ciamac C. Moallemi. The exploration-exploitation trade-off in the newsvendor problem. *Stochastic Systems*, Articles in Advance, 2022.

Olivier Bousquet and Manfred K Warmuth. Tracking a small set of experts by mixing past posteriors. *Journal of Machine Learning Research*, 3(Nov):363–396, 2002.

David J. Braden and Marshall Freimer. Informational dynamics of censored observations. *Management Science*, 37(11):1390–1404, 1991.

David J. Braden and Shmuel S. Oren. Nonlinear pricing to produce information. *Marketing Science*, 13 (3):310–326, 1994.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

Nicolò Cesa-Bianchi, Pierre Gaillard, Gábor Lugosi, and Gilles Stoltz. Mirror descent meets fixed share (and feels no regret). In P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 989–997. 2012.

Xiaomei Ding, Martin L. Puterman, and Arnab Bisi. The censored newsvendor and the optimal acquisition of information. *Operations Research*, 50(3):517–527, 2002.

Kuzman Ganchev, Yuriy Nevmyvaka, Michael Kearns, and Jennifer Wortman Vaughan. Censored exploration and the dark pool problem. *Communications of the ACM*, 53(5):99–107, 2010.

Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors. *Advances in Neural Information Processing Systems 27*, 2014.

Giora Harpaz, Wayne Y. Lee, and Robert L. Winkler. Learning, experimentation, and the optimal output decisions of a competitive firm. *Management Science*, 28(6):589–603, 1982.

M. Herbster and M. Warmuth. Tracking the best expert. *Machine Learning*, 32:151–178, 1998.

Woonghee Tim Huh and Paat Rusmevichientong. A nonparametric asymptotic analysis of inventory planning with censored demand. *Mathematics of Operations Research*, 34(1):103–123, 2009.

Samuel Karlin. Dynamic inventory policy with varying stochastic demands. *Management Science*, 6(3): 231–258, 1960.

Tomás Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. In Ghahramani et al. [2014], pages 613–621.

Wouter Koolen, Manfred K. Warmuth, and Jyrki Kivinen. Hedging structured concepts. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, pages 93–105, 2010.

Martin A. Lariviere and Evan L. Porteus. Stalking information: bayesian inventory management with unobserved lost sales. *Management Science*, 45(3):346–363, 1999.

Tatsiana Levina, Yuri Levin, Jeff McGill, Mikhail Nediak, and Vladimir Vovk. Weak aggregating algorithm for the distribution-free perishable inventory problem. *Operations Research Letters*, 38:516–521, 2010.

Xiangwen Lu, Jing-Sheng Song, and Kaijie Zhu. Technical note: Analysis of perishable-inventory systems with censored demand data. *Operations Research*, 56(4):1034–1038, 2008.

S Mannor and O Shamir. From Bandits to Experts: On the Value of Side-Observations. In *Neural Information Processing Systems*, 2011.

Amir Sani, Gergely Neu, and Alessandro Lazaric. Exploiting easy data in online optimization. In Ghahramani et al. [2014], pages 810–818.

Herbert Scarf. Bayes solutions of the statistical inventory problem. *The Annals of Mathematical Statistics*, 30(2):490–508, 1959.

Cong Shi, Weidong Chen, and Izak Duenyas. Technical note—nonparametric data-driven algorithms for multiproduct inventory systems with censored demand. *Operations Research*, 64(2):362–370, 2016.

Tim Van Erven, Manfred Warmuth, and Wojciech Kotłowski. Follow the leader with dropout perturbations. In Feldman V. Szepesvári Cs. Balcan, M. F., editor, *Proceedings of the 27th Annual Conference on Learning Theory*, pages 949–974, 2014.

Jonathan Weed, Vianney Perchet, and Philippe Rigollet. Online learning in repeated auctions. *Proceedings of the* 29^{th} *Annual Conference on Learning Theory*, 49:1–31, 2016.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. *Proceedings of the* 20^{th} *International Conference on Machine Learning*, pages 928–936, 2003.

Electronic Companion
“On the Hardness of Learning from Censored and Nonstationary Demand”

Proofs of Main Results

Let $k \in \mathbb{N}$, with $k \leq D$, and denote by $\mathbb{L}_k$ the $k \times k$ identity matrix, and by $\mathbb{M}_k$ the $k \times (D-k)$ matrix, where $\mathbb{M}_k(i,j) = 1$ if $i = k$, and 0 otherwise. Finally, let e_d be the $(D+1)$-dimensional column vector, with $e_d(j) = 1$ if $j = d$, and 0 otherwise.

The *signal matrix* of inventory decision $i \in \mathcal{I}$, denoted by $\mathbb{S}_i$, is a $(i+1) \times (D+1)$ matrix, where element $\mathbb{S}_i(k,j) = 1$ if the sales of the firm are equal to $k \in \{0,1,\ldots,i\}$, assuming that the demand is equal to $j \in \mathcal{D}$ and the inventory is equal to i, and 0 otherwise. It can be verified that $\mathbb{S}_i$ is equal to the concatenated matrix $\mathbb{S}_i = \left[\mathbb{L}_{i+1} \mid \mathbb{M}_{i+1}\right]$.

Lemma 1: Local Observability

Lemma 1. *Let $i, j \in \mathcal{I}$ be arbitrary inventory decisions. There exist vectors $v_i \in \mathcal{R}^{i+1}$ and $v_j \in \mathcal{R}^{j+1}$ such that*

$$\left(v_i^T \mathbb{S}_i - v_j^T \mathbb{S}_j\right) e_d = c(i,d) - c(j,d), \qquad d \in \mathcal{D}.$$

Proof. Consider the $(i+1)$-dimensional column vector v_i, where

$$v_i(k) = hi - (h+b)(k-1), \qquad k \in \{1,2,\ldots,i+1\}.$$

Define similarly the vector v_j. The result follows through straightforward calculations. ■

Hence, the game between the inventory manager and the market is locally observable, in the sense of Definition 6 in Bartók et al. [2014]. This classifies the repeated newsvendor problem with demand learning via censored data as an “easy” partial monitoring problem (see Section 2.1), which implies that the correct scaling of the expected regret is $\Theta\left(\sqrt{T}\right)$.

Lemma 2: Bias and Variance of the Estimator

Lemma 2. *The cost estimator in Eq.* (5) *satisfies:*

$$\mathbb{E}_t\left[\widetilde{c}(i,d_t)\right] = v_i^T \mathbb{S}_i e_{d_t} + \beta,$$

so that $\mathbb{E}_t\left[\widetilde{c}(i,d_t)\right] \in (0, 2\beta)$, *and*

$$\mathbb{E}_t\left[\widetilde{c}(i,d_t)^2\right] \leq \frac{4\beta^2}{\mathbb{P}_t(I_t \geq i)},$$

where $\mathbb{P}_t(\cdot)$ *and* $\mathbb{E}_t[\cdot]$ *respectively denote the probability and the expectation conditioned on the history of interaction up until the beginning of round* t.

Proof. The first part of the lemma follows directly by noting that

$$\mathbb{E}_t\left[\widetilde{c}(i,d_t)\right] = \frac{\mathbb{E}_t\left[\mathbb{1}_{\{I_t \geq i\}}\right]}{\mathbb{P}_t(I_t \geq i)}\left(v_i^T \mathbb{S}_i e_{d_t} + \beta\right) = v_i^T \mathbb{S}_i e_{d_t} + \beta.$$

The fact that $\mathbb{E}_t\left[\widetilde{c}(i,d_t)\right] \in (0, 2\beta)$ is a direct consequence of β being an upper bound on the absolute value of $v_i^T S_i e_{d_t}$, for every $i \in \mathcal{I}$ and d_t. Hence, regarding the second part of the lemma, we have that

$$\mathbb{E}_t\left[\widetilde{c}(i,d_t)^2\right] = \frac{\mathbb{E}_t\left[\mathbb{1}_{\{I_t \geq i\}}\right]}{\mathbb{P}_t(I_t \geq i)^2}\left(v_i^T \mathbb{S}_i e_{d_t} + \beta\right)^2 \leq \frac{4\beta^2}{\mathbb{P}_t(I_t \geq i)}.$$

■

Note that the proposed estimator is biased, as $\mathbb{E}_t\left[\widetilde{c}(i,d_t)\right] \neq c(i,d_t)$. In particular, the estimator is pessimistic in the sense that it always overestimates the actual cost incurred. A direct corollary of Lemmas 1 and 2 is the following result.

Lemma 3: Unbiased Estimator

Lemma 3. *The cost estimator in Eq.* (5) *is unbiased when inferring the difference in cost between two actions:*

$$\mathbb{E}_t\left[\widetilde{c}(i,d_t) - \widetilde{c}(j,d_t)\right] = c(i,d_t) - c(j,d_t), \qquad i,j \in \mathcal{I}.$$

The significance of Lemma 3 lies in the fact that the regret, by definition, is a metric that is based on cost differences. This facilitates the analysis in our main result, which characterizes the performance of the proposed inventory management policy with respect to the regret criterion.

Lemma 4

Lemma 4. *Let* $\mathbf{p} = (p_1, p_2, \dots, p_N)$ *be any probability distribution satisfying* $p_i \geq \omega > 0$ *for all* $i \in \{1, 2, \dots, N\}$. *Then, the following inequality is satisfied:*

$$\sum_{i=1}^{N} \frac{p_i}{\sum_{j=i}^{N} p_j} \leq 5 \log(1 + 1/\omega) + 3.$$

Proof. We fix $\omega > 0$ and partition the actions into the sets $M > 1$ sets, with the m-th set holding probability mass of at least 2^m for all m. Precisely, we let $k_0 = N + 1$ and define k_m for $m > 0$ recursively as

$$k_{m+1} = \max \left\{ i : \sum_{j=i}^{k_m - 1} p_j \geq 2^m \omega \right\},$$

with $\max\{\emptyset\} = 1$, and we let $M = \min\{m : k_m = 1\}$. Then, we set $\mathcal{I}_m = [k_m, k_{m-1}]$ for all $m = 1, \dots, M$. Notice that, for any $m < M$, this choice guarantees that

$$\sum_{j=k_m}^{N} p_j = \sum_{n=0}^{m-1} \sum_{j=k_{n+1}}^{k_n - 1} p_j \geq \sum_{n=0}^{m-1} 2^n \omega = \left(2^m - 1\right) \omega,$$

which can be seen to imply an upper bound on M since

$$1 = \sum_{j=k_M}^{N} p_j \geq \sum_{j=k_{M-1}}^{N} p_j \geq \left(2^{M-1} - 1\right) \omega \geq 2^{M-1} \omega.$$

Indeed, this implies that $M \leq \log_2(1 + 1/\omega) + 1$. Furthermore, we can write the following:

$$\begin{aligned}
\sum_{i=1}^{N} \frac{p_i}{\sum_{j=i}^{N} p_j} &= \sum_{m=0}^{M-1} \sum_{i=k_{m+1}}^{k_m - 1} \frac{p_i}{\sum_{j=i}^{N} p_j} \\
&= \sum_{m=0}^{M-1} \left(\frac{p_{k_{m+1}}}{\sum_{j=k_{m+1}}^{N} p_j} + \sum_{i=k_{m+1}+1}^{k_m - 1} \frac{p_i}{\sum_{j=i}^{N} p_j} \right) \\
&\leq \sum_{m=0}^{M-1} \left(1 + \sum_{i=k_{m+1}+1}^{k_m - 1} \frac{p_i}{\sum_{j=k_m}^{N} p_j} \right) \\
&\leq \sum_{m=0}^{M-1} \left(1 + \sum_{i=k_{m+1}+1}^{k_m - 1} \frac{p_i}{(2^m - 1)\,\omega} \right) \\
&\leq \sum_{m=0}^{M-1} \left(1 + \frac{2^m \omega}{(2^m - 1)\,\omega} \right) = \sum_{m=0}^{M-1} \frac{2^m + 2^m - 1}{2^m - 1} = \sum_{m=0}^{M-1} \frac{2^{m+1} - 1}{2^m - 1} \leq 3M,
\end{aligned}$$

where we used the fact that $\sum_{i=k_{m+1}+1}^{k_m - 1} p_i \leq 2^m \omega$ holds by definition of the index k_m. The proof is concluded by putting this together with the upper bound on M and bounding $3/\log(2) \leq 5$. ■

Proof of Theorem 1

Our analysis largely follows the steps of the proof of Theorem 3.1 by Auer et al. [2002], combined with our Lemmas 2 and 3. We analyze a slightly more general version of the EWF algorithm that uses an arbitrary exploration distribution μ, with μ_i being the probability of taking action i in exploration rounds. More precisely, we consider a version of EWF that computes its probability distributions over the actions as

$$p_i(t) = (1-\gamma)\frac{W_i(t-1)}{W(t-1)} + \gamma\mu_i, \qquad i \in \mathfrak{I}.$$

The statement will follow from setting $\mu_i = 1/N$ for all actions i.

The key idea of the analysis is studying the term $\log\bigl(W(T)/W(0)\bigr)$ which, as we show shortly, relates closely to the regret. We start by constructing a lower bound:

$$\begin{aligned}
\log\left(\frac{W(T)}{W(0)}\right) &= \log\bigl(W(T)\bigr) - \log\bigl(W(0)\bigr) \\
&= \log\left(\sum_{i\in\mathfrak{I}} W_i(T)\right) - \log\left(\sum_{i\in\mathfrak{I}} W_i(0)\right) \\
&= \log\left(\sum_{i\in\mathfrak{I}} e^{-\eta\widetilde{C}_i(T)}\right) - \log N \\
&\geq \log\left(e^{-\eta\widetilde{C}_{i^*}(T)}\right) - \log N \\
&= -\eta\sum_{t\in\mathfrak{T}} \widetilde{c}(i^*, d_t) - \log N, \qquad (10)
\end{aligned}$$

where i^* is the best fixed action in hindsight, for the particular demand sequence.

Then, we derive an upper bound on $\log\bigl(W(T)/W(0)\bigr)$:

$$\begin{aligned}
\log\left(\frac{W(T)}{W(0)}\right) &= \log\left(\prod_{t\in\mathfrak{T}} \frac{W(t)}{W(t-1)}\right) \\
&= \sum_{t\in\mathfrak{T}} \log\left(\frac{W(t)}{W(t-1)}\right) \\
&= \sum_{t\in\mathfrak{T}} \log\left(\sum_{i\in\mathfrak{I}} \frac{W_i(t)}{W(t-1)}\right) \\
&= \sum_{t\in\mathfrak{T}} \log\left(\sum_{i\in\mathfrak{I}} \frac{W_i(t-1)}{W(t-1)} e^{-\eta\widetilde{c}(i,d_t)}\right) \quad \text{(by Eq. (3))}.
\end{aligned}$$

Note that $e^{-x} \leq 1 - x + x^2/2$, for all $x \geq 0$, and our estimators for the cost of the different decisions are nonnegative. Thus,

$$\begin{aligned}\log\left(\frac{W(T)}{W(0)}\right) &\leq \sum_{t\in\mathfrak{T}} \log\left(\sum_{i\in\mathfrak{I}} \frac{W_i(t-1)}{W(t-1)}\left(1-\eta\widetilde{c}(i,d_t)+\frac{\eta^2}{2}\widetilde{c}(i,d_t)^2\right)\right)\\ &= \sum_{t\in\mathfrak{T}} \log\left(1-\eta\sum_{i\in\mathfrak{I}}\frac{W_i(t-1)}{W(t-1)}\widetilde{c}(i,d_t)+\frac{\eta^2}{2}\sum_{i\in\mathfrak{I}}\frac{W_i(t-1)}{W(t-1)}\widetilde{c}(i,d_t)^2\right).\end{aligned}$$

Moreover, $\log(1+x)\leq x$, for all $x>-1$. Since this is the case with the right-hand side of the expression above (being an upper bound to a sum of exponential terms), we have, using Eq. (4), that

$$\begin{aligned}\log\left(\frac{W(T)}{W(0)}\right) &\leq \sum_{t\in\mathfrak{T}}\left(-\eta\sum_{i\in\mathfrak{I}}\frac{W_i(t-1)}{W(t-1)}\widetilde{c}(i,d_t)+\frac{\eta^2}{2}\sum_{i\in\mathfrak{I}}\frac{W_i(t-1)}{W(t-1)}\widetilde{c}(i,d_t)^2\right)\\ &= \sum_{t\in\mathfrak{T}}\left(-\eta\sum_{i\in\mathfrak{I}}\frac{p_i(t)-\gamma\mu_i}{1-\gamma}\widetilde{c}(i,d_t)+\frac{\eta^2}{2}\sum_{i\in\mathfrak{I}}\frac{p_i(t)-\gamma\mu_i}{1-\gamma}\widetilde{c}(i,d_t)^2\right). \qquad (11)\end{aligned}$$

Eqs. (10) and (11) imply that

$$\frac{\eta}{1-\gamma}\sum_{t\in\mathfrak{T}}\sum_{i\in\mathfrak{I}}p_i(t)\widetilde{c}(i,d_t)-\eta\sum_{t\in\mathfrak{T}}\widetilde{c}(i^*,d_t)\leq\log N+\frac{\eta\gamma}{1-\gamma}\sum_{t\in\mathfrak{T}}\sum_{i\in\mathfrak{I}}\mu_i\widetilde{c}(i,d_t)+\frac{\eta^2}{2(1-\gamma)}\sum_{t\in\mathfrak{T}}\sum_{i\in\mathfrak{I}}p_i(t)\widetilde{c}(i,d_t)^2.$$

Since $\widetilde{c}(i^*,d_t)\geq 0$, for all $t\in\mathfrak{T}$, we have that

$$\sum_{t\in\mathfrak{T}}\sum_{i\in\mathfrak{I}}p_i(t)\widetilde{c}(i,d_t)-\sum_{t\in\mathfrak{T}}\widetilde{c}(i^*,d_t)\leq\frac{\log N}{\eta}+\gamma\sum_{t\in\mathfrak{T}}\sum_{i\in\mathfrak{I}}\mu_i\widetilde{c}(i,d_t)+\frac{\eta}{2}\sum_{t\in\mathfrak{T}}\sum_{i\in\mathfrak{I}}p_i(t)\widetilde{c}(i,d_t)^2.$$

This further implies that

$$\mathbb{E}\left[\sum_{t\in\mathfrak{T}}\sum_{i\in\mathfrak{I}}p_i(t)\big(\widetilde{c}(i,d_t)-\widetilde{c}(i^*,d_t)\big)\right]\leq\frac{\log N}{\eta}+\gamma\mathbb{E}\left[\sum_{t\in\mathfrak{T}}\sum_{i\in\mathfrak{I}}\mu_i\widetilde{c}(i,d_t)\right]+\frac{\eta}{2}\mathbb{E}\left[\sum_{t\in\mathfrak{T}}\sum_{i\in\mathfrak{I}}p_i(t)\widetilde{c}(i,d_t)^2\right].$$

The tower rule of expectations along with Eq. (2) and Lemma 3 imply that

$$\begin{aligned}\mathbb{E}\left[\sum_{t\in\mathfrak{T}}\sum_{i\in\mathfrak{I}}p_i(t)\big(\widetilde{c}(i,d_t)-\widetilde{c}(i^*,d_t)\big)\right] &= \mathbb{E}\left[\sum_{t\in\mathfrak{T}}\sum_{i\in\mathfrak{I}}p_i(t)\mathbb{E}_t\big[\widetilde{c}(i,d_t)-\widetilde{c}(i^*,d_t)\big]\right]\\ &= \mathbb{E}\left[\sum_{t\in\mathfrak{T}}\sum_{i\in\mathfrak{I}}p_i(t)\big(c(i,d_t)-c(i^*,d_t)\big)\right]=\mathbb{E}[\mathfrak{R}(T)].\end{aligned}$$

Combined with Lemma 2, and the fact that $\mu_i=1/N$, we get:

$$\mathbb{E}[\mathfrak{R}(T)]\leq\frac{\log N}{\eta}+\frac{\gamma}{N}\sum_{t\in\mathfrak{T}}\sum_{i\in\mathfrak{I}}2\beta+\frac{\eta}{2}\sum_{t\in\mathfrak{T}}\sum_{i\in\mathfrak{I}}p_i(t)\frac{4\beta^2}{\mathbb{P}_t(I_t\geq i)}$$

$$= \frac{\log N}{\eta} + 2\beta\gamma T + 2\beta^2 \eta \sum_{t\in\mathfrak{T}} \sum_{i\in\mathfrak{I}} \frac{p_i(t)}{\mathbb{P}_t(I_t \geq i)}.$$

The final step in the proof is to bound the term $\sum_{i\in\mathfrak{I}} p_i(t) \big/ \mathbb{P}_t(I_t \geq i)$. This can be directly bounded by applying Lemma 4 with $\omega = \frac{\gamma}{N}$, which gives

$$\sum_{i\in\mathfrak{I}} \frac{p_i(t)}{\mathbb{P}_t(I_t \geq i)} = \sum_{i\in\mathfrak{I}} \frac{p_i(t)}{\sum_{j=i}^{N} p_j(t)} \leq 5\left(\log\left(\frac{N}{\gamma}+1\right)+1\right) \leq 5\log\left(\frac{3N}{\gamma}+3\right).$$

Consequently,

$$\mathbb{E}[\mathcal{R}(T)] \leq \frac{\log N}{\eta} + 2\beta\gamma T + 10\beta^2\eta T \log\left(\frac{3N}{\gamma}+3\right).$$

By choosing

$$\eta = \sqrt{\frac{\log N}{10\beta^2 T \log\left(\frac{3N}{\gamma}+3\right)}},$$

and bounding $2\sqrt{10} \leq 7$, we have that

$$\mathbb{E}[\mathcal{R}(T)] \leq 7\beta\sqrt{T \log N \log\left(\frac{3N}{\gamma}+3\right)} + 2\beta\gamma T.$$

Finally, by setting $\gamma = 1/\left(2\beta T\right)$, we get:

$$\mathbb{E}[\mathcal{R}(T)] \leq 7\beta\sqrt{T \log N \log\left(6\beta T N + 3\right)} + 1.$$

This concludes the proof.

Proof of Theorem 2

We follow the steps of the proof of Theorem 8.1 in Auer et al. [2002]. To this end, fix a comparator sequence $i_{[T]} \in \mathfrak{I}_S^T$, and partition the interval $[1,T]$ into a number of subintervals $I_1 = [1, T_1]$, $I_2 = [T_1+1, T_2]$, $\ldots$, $I_C = [T_{C-1}+1, T]$, such that i_t remains constant within each interval. Since $i_{[T]} \in \mathfrak{I}_S^T$, we have that $C \leq S$. In the remainder of the proof, we bound the regret within each interval, and then combine the obtained bounds to prove a guarantee about the (global) tracking regret.

Fix an arbitrary interval I_s, $s \in \{1, \ldots, C\}$, and let j_s be the action taken during that interval (i.e., $i_t = j_s$, for all $t \in I_s$). Also, let $\tau_s = T_s - T_{s-1}$ be the length of I_s. As in the proof of Theorem 1, we study the term

$\log\big(W(T_s)/W(T_{s-1})\big)$. First, note that

$$\begin{aligned} W_{j_s}(T_s) &\geq W_{j_s}(T_{s-1}+1)\exp\left(-\eta\sum_{t=T_{s-1}+2}^{T_s}\tilde{c}(j_s,d_t)\right) \\ &\geq \frac{\alpha}{N}W(T_{s-1})\exp\left(-\eta\sum_{t=T_{s-1}+2}^{T_s}\tilde{c}(j_s,d_t)\right) \\ &\geq \frac{\alpha}{N}W(T_{s-1})\exp\left(-\eta\sum_{t=T_{s-1}+1}^{T_s}\tilde{c}(j_s,d_t)\right), \end{aligned}$$

where the first and second inequalities follow from the (recursive) definition of the sequence of weights $W_j(t)$, and the last one is a consequence of the non-negativity of the loss estimates $\tilde{c}(i,d_t)$. Hence,

$$\log\left(\frac{W(T_s)}{W(T_{s-1})}\right) \geq \log\left(\frac{W_{j_s}(T_s)}{W(T_{s-1})}\right) \geq \log\left(\frac{\alpha}{N}\right) - \eta\sum_{t=T_{s-1}+1}^{T_s}\tilde{c}(j_s,d_t).$$

On the other hand, we have that

$$\frac{W(t+1)}{W(t)} = \sum_{i\in\mathcal{I}}\frac{W_i(t)e^{-\eta\tilde{c}(i,d_t)}+\frac{\alpha}{N}W(t)}{W(t)} = \sum_{i\in\mathcal{I}}\frac{W_i(t)e^{-\eta\tilde{c}(i,d_t)}}{W(t)}+\alpha,$$

so by a similar line of reasoning as in the proof of Theorem 1, it can be verified that

$$\log\frac{W(T_s)}{W(T_{s-1})} \leq \sum_{t=T_{s-1}+1}^{T_s}\left(-\eta\sum_{i\in\mathcal{I}}\frac{p_i(t)-\gamma/N}{1-\gamma}\tilde{c}(i,d_t)+\frac{\eta^2}{2}\sum_{i\in\mathcal{I}}\frac{p_i(t)-\gamma/N}{1-\gamma}\tilde{c}(i,d_t)^2+\alpha\right).$$

Combining the two bounds together, taking expectations, and appealing to Lemma 2, we obtain:

$$\mathbb{E}\left[\sum_{t=T_{s-1}+1}^{T_s}\sum_{i\in\mathcal{I}}p_i(t)\big(\tilde{c}(i,d_t)-\tilde{c}(j_s,d_t)\big)\right] \leq \frac{\log\left(\frac{N}{\alpha}\right)}{\eta}+2\beta\gamma\tau_s+2\beta^2\eta\mathbb{E}\left[\sum_{t=T_{s-1}+1}^{T_s}\sum_{i\in\mathcal{I}}\frac{p_i(t)}{\mathbb{P}_t(I_t\geq i)}\right]+\alpha\tau_s. \quad (12)$$

As in the proof of Theorem 1, the third term on the right-hand side can be bounded from above using Lemma 4 as follows:

$$\sum_{i\in\mathcal{I}}\frac{p_i(t)}{\mathbb{P}_t(I_t\geq i)} \leq 5\log\left(\frac{3N}{\gamma}+3\right).$$

Using this bound, we can add over all intervals $s\in\{1,\ldots,C\}$ both sides of Eq. (12), and use Lemma 3 to obtain:

$$\mathbb{E}\left[\sum_{t=1}^{T}\big(c(I_t,d_t)-c(i_t,d_t)\big)\right] \leq \frac{S\log\left(\frac{N}{\alpha}\right)}{\eta}+2\beta\gamma T+10\beta^2\eta T\log\left(\frac{3N}{\gamma}+3\right)+\alpha T.$$

The statement of the theorem follows from taking the supremum over all $i_{[T]}\in\mathcal{I}_S^T$, and substituting for the chosen values of γ, η, and α. We note that supremum and expectation can be interchanged in our case, since the comparator sequence, that the supremum is taken with respect to, is deterministic. This would not have been the case, e.g., if the firm competed against an adaptive adversary.

Lemma 5: Variance of Estimator for the Combinatorial Case

Lemma 5. *The second moment of the estimator defined in Equation* (7) *satisfies*

$$\sum_{(i^{(1)},\ldots,i^{(K)})\in\mathcal{A}_r} p_{(i^{(1)},\ldots,i^{(K)})}(t)\mathbb{E}_t\left[\widetilde{c}\left(i^{(1)},\ldots,i^{(K)},r,d_t^{(1)},\ldots,d_t^{(K)}\right)^2\right] \\ \leq 20K^2\beta^2\log\left(\frac{3KN}{\gamma}+3\right)+2(f+K\beta)^2.$$

Proof. For simplicity, let us introduce the notation

$$\ell_k\left(i^{(k)},d_t^{(k)}\right)=v_i^T\mathbb{S}_i e_{d_t^{(k)}}+f^{(k)}\mathbb{1}_{\{i^{(k)}>0\}}+\beta,$$

so that each retailer's cost estimate can be written as

$$\widetilde{c}_k\left(i^{(k)},d_t^{(k)}\right)=\frac{\mathbb{1}_{\left\{I_t^{(k)}\geq i^{(k)}\right\}}\ell_k\left(i^{(k)},d_t^{(k)}\right)}{\mathbb{P}_t\left(I_t^{(k)}\geq i^{(k)}\right)},$$

for all $k\in\mathcal{K}$. Also, let $\tilde{I}_t$ be an independent copy of I_t. With this notation, we have that

$$\begin{aligned}&\sum_{(i^{(1)},\ldots,i^{(K)})\in\mathcal{A}_r} p_{(i^{(1)},\ldots,i^{(K)})}(t)\mathbb{E}_t\left[\widetilde{c}\left(i^{(1)},\ldots,i^{(K)},r,d_t^{(1)},\ldots,d_t^{(K)}\right)^2\right]\\ =&\mathbb{E}_t\left[\left(c_0\left(\tilde{I}_t^{(1)},\ldots,\tilde{I}_t^{(K)},r\right)+\sum_{k\in\mathcal{K}}\widetilde{c}_k\left(\tilde{I}_t^{(k)},d_t^{(k)}\right)\right)^2\right]\\ \leq&2\mathbb{E}_t\left[c_0\left(\tilde{I}_t^{(1)},\ldots,\tilde{I}_t^{(K)},r\right)^2+\left(\sum_{k\in\mathcal{K}}\widetilde{c}_k\left(\tilde{I}_t^{(k)},d_t^{(k)}\right)\right)^2\right],\end{aligned}$$

where the last step follows from the inequality $(a+b)^2\leq 2\left(a^2+b^2\right)$, which holds for all $a,b\in\mathbb{R}$. The first term can be trivially bounded by $(f+K\beta)^2$. Regarding the second term, we have that

$$\mathbb{E}_t\left[\left(\sum_{k\in\mathcal{K}}\widetilde{c}_k\left(\tilde{I}_t^{(k)},d_t^{(k)}\right)\right)^2\right]$$

$$
\begin{aligned}
=&\mathbb{E}_t\left[\left(\sum_{j\in\mathcal{K}}\frac{\mathbb{1}_{\left\{I_t^{(j)}\geq\tilde{I}_t^{(j)}\right\}}}{\mathbb{P}_t\left(I_t^{(j)}\geq\tilde{I}_t^{(j)}\right)}\ell_j\left(\tilde{I}_t^{(j)},d_t^{(j)}\right)\right)\cdot\left(\sum_{k\in\mathcal{K}}\frac{\mathbb{1}_{\left\{I_t^{(k)}\geq\tilde{I}_t^{(k)}\right\}}}{\mathbb{P}_t\left(I_t^{(k)}\geq\tilde{I}_t^{(k)}\right)}\ell_k\left(\tilde{I}_t^{(k)},d_t^{(k)}\right)\right)\right]\\
=&\mathbb{E}_t\left[\sum_{j\in\mathcal{K}}\sum_{k\in\mathcal{K}}\frac{\mathbb{1}_{\left\{I_t^{(j)}\geq\tilde{I}_t^{(j)}\right\}}\mathbb{1}_{\left\{I_t^{(k)}\geq\tilde{I}_t^{(k)}\right\}}}{\mathbb{P}_t\left(I_t^{(j)}\geq\tilde{I}_t^{(j)}\right)\mathbb{P}_t\left(I_t^{(k)}\geq\tilde{I}_t^{(k)}\right)}\ell_j\left(\tilde{I}_t^{(j)},d_t^{(j)}\right)\ell_k\left(\tilde{I}_t^{(k)},d_t^{(k)}\right)\right]\\
\leq&\frac{1}{2}\mathbb{E}_t\left[\sum_{j\in\mathcal{K}}\sum_{k\in\mathcal{K}}\left(\frac{1}{\mathbb{P}_t\left(I_t^{(j)}\geq\tilde{I}_t^{(j)}\right)^2}+\frac{1}{\mathbb{P}_t\left(I_t^{(k)}\geq\tilde{I}_t^{(k)}\right)^2}\right)\mathbb{1}_{\left\{I_t^{(j)}\geq\tilde{I}_t^{(j)}\right\}}\mathbb{1}_{\left\{I_t^{(k)}\geq\tilde{I}_t^{(k)}\right\}}\ell_j\left(\tilde{I}_t^{(j)},d_t^{(j)}\right)\ell_k\left(\tilde{I}_t^{(k)},d_t^{(k)}\right)\right],
\end{aligned}
$$

again using $(a+b)^2\leq 2\left(a^2+b^2\right)$. We further have that

$$
\begin{aligned}
&\mathbb{E}_t\left[\left(\sum_{k\in\mathcal{K}}\widetilde{c}_k\left(\tilde{I}_t^{(k)},d_t^{(k)}\right)\right)^2\right]\\
=&\mathbb{E}_t\left[\sum_{j\in\mathcal{K}}\sum_{k\in\mathcal{K}}\frac{1}{\mathbb{P}_t\left(I_t^{(j)}\geq\tilde{I}_t^{(j)}\right)^2}\mathbb{1}_{\left\{I_t^{(j)}\geq\tilde{I}_t^{(j)}\right\}}\mathbb{1}_{\left\{I_t^{(k)}\geq\tilde{I}_t^{(k)}\right\}}\ell_j\left(\tilde{I}_t^{(j)},d_t^{(j)}\right)\ell_k\left(\tilde{I}_t^{(k)},d_t^{(k)}\right)\right],
\end{aligned}
$$

due to the symmetry between j and k, which implies that

$$
\begin{aligned}
&\mathbb{E}_t\left[\left(\sum_{k\in\mathcal{K}}\widetilde{c}_k\left(\tilde{I}_t^{(k)},d_t^{(k)}\right)\right)^2\right]\\
=&\mathbb{E}_t\left[\sum_{j\in\mathcal{K}}\frac{1}{\mathbb{P}_t\left(I_t^{(j)}\geq\tilde{I}_t^{(j)}\right)^2}\mathbb{1}_{\left\{I_t^{(j)}\geq\tilde{I}_t^{(j)}\right\}}\ell_j\left(\tilde{I}_t^{(j)},d_t^{(j)}\right)\sum_{k\in\mathcal{K}}\mathbb{1}_{\left\{I_t^{(k)}\geq\tilde{I}_t^{(k)}\right\}}\ell_k\left(\tilde{I}_t^{(k)},d_t^{(k)}\right)\right]\\
\leq&2K\beta\mathbb{E}_t\left[\sum_{j\in\mathcal{K}}\frac{1}{\mathbb{P}_t\left(I_t^{(j)}\geq\tilde{I}_t^{(j)}\right)^2}\mathbb{1}_{\left\{I_t^{(j)}\geq\tilde{I}_t^{(j)}\right\}}\ell_j\left(\tilde{I}_t^{(j)},d_t^{(j)}\right)\right]\\
=&2K\beta\mathbb{E}_t\left[\sum_{j\in\mathcal{K}}\frac{\ell_j\left(\tilde{I}_t^{(j)},d_t^{(j)}\right)}{\mathbb{P}_t\left(I_t^{(j)}\geq\tilde{I}_t^{(j)}\right)}\right]\\
\leq&4K\beta^2\sum_{j=1}^{K}\mathbb{E}_t\left[\sum_{i=1}^{N}\frac{\mathbb{P}_t\left(I_t^{(j)}=i\right)}{\mathbb{P}_t\left(I_t^{(j)}\geq i\right)}\right],
\end{aligned}
$$

where the inequalities follow from bounding from above $\ell_j(\cdot,\cdot)$ by 2β.

It remains to bound the sums within the expectation, for all j. To this end, we observe that our exploration distribution μ guarantees that $\mathbb{P}\left[I_t^{(j)}=i\right]\geq\frac{\gamma}{NK}$ holds for all i,j. Given this lower bound, we can apply Lemma 4 with $\omega = \frac{\gamma}{NK}$ as done in the proof of Theorem 1:

$$\sum_{i=1}^{N} \frac{\mathbb{P}_t\left(I_t^{(j)} = i\right)}{\mathbb{P}_t\left(I_t^{(j)} \geq i\right)} \leq 5 \log\left(\frac{3KN}{\gamma} + 3\right).$$

Putting everything together, we obtain the desired result. ■

Proof of Theorem 3

By following closely the proof of Theorem 1 with our definition of μ and applying Eq. (8), we have that

$$\begin{aligned}
\mathbb{E}[\mathfrak{R}_c(T)] \leq & \frac{\log|\mathcal{A}_r|}{\eta} + \gamma \sum_{t\in\mathfrak{T}} \sum_{(i^{(1)},\ldots,i^{(K)})\in\mathcal{A}_r} \mu_{(i^{(1)},\ldots,i^{(K)})} \mathbb{E}_t\left[\widetilde{c}\left(i^{(1)},\ldots,i^{(K)},r,d_t^{(1)},\ldots,d_t^{(K)}\right)\right] \\
& + \frac{\eta}{2} \sum_{t\in\mathfrak{T}} \sum_{(i^{(1)},\ldots,i^{(K)})\in\mathcal{A}_r} p_{(i^{(1)},\ldots,i^{(K)})}(t) \mathbb{E}_t\left[\widetilde{c}\left(i^{(1)},\ldots,i^{(K)},r,d_t^{(1)},\ldots,d_t^{(K)}\right)^2\right] \\
\leq & \frac{K\log N}{\eta} + \gamma\big(f + K(2\beta + f)\big)T + \eta T\left(10K^2\beta^2 \log\left(\frac{3KN}{\gamma} + 3\right) + (f + K\beta)^2\right),
\end{aligned}$$

where the last step uses Eq. (9) and Lemma 5. Substituting the prescribed values for γ and η yields the statement of the theorem.