Online Learning for Inventory Problems

Massil HIHAT

Monday 30$^{\text{th}}$ May, 2022

In this document we start in Section 1 by introducing classical results from *inventory theory*. This sub-field of operations research provides models and solutions for the *inventory control problem*: the problem faced by a firm that must decide how much to order in each time period to meet a demand. This introduction is inspired by Snyder and Shen [2019].

In Section 2 we present a recent and powerful online learning framework called *Online Convex Optimization* (OCO). It tackles a very general sequential decision problem, provides algorithms to solve it and mathematical guarantees on their performances. This section follows Orabona [2019]. See also Shalev-Shwartz et al. [2011] or Hazan [2019].

Section 3 details the contributions of Huh and Rusmevichientong [2009] which is the first application of online learning to inventory problems. Then, in Section 4, applications of online learning to more complex inventory problems are briefly exposed. Finally, Section 5 concludes this document by discussing real-world applications and various research directions.

1 Classical Inventory Problems

1.1 The Newsvendor Problem

The newsvendor problem is a classical stochastic inventory problem: at the beginning of the day a newsboy needs to buy newspapers in order to fulfill an unknown random demand. Its modern formulation appeared in Arrow et al. [1951].

This problem is usually modeled as follows. We assume the decision-maker has to choose an *order quantity* $q \in \mathbb{R}_+$ in order to fulfill a *demand* $d \in \mathbb{R}_+$ which is the realization of a non-negative random variable D with cumulative distribution function (c.d.f.) F. The classical cost structure for this problem considers a *holding cost* $h > 0$ per unsold unit and a *penality cost* $p > 0$ per unmet demand unit, that is, given the order quantity q and the demand realization d the decision-maker incurs the cost:

$$\ell(q, d) = h\,[q - d]^+ \; + \; p\,[d - q]^+ \,. \tag{1}$$

The goal of the decision maker is to minimize the expected cost:

$$L(q) = \mathbb{E}_D\left[\ell(q, D)\right]. \tag{2}$$

Notice that both the cost function $\ell(\cdot, d)$ and the expected cost function $L(\cdot)$ are convex functions. In the newsvendor problem, the demand distribution is supposed to be *known*, thus the expected cost function $L(\cdot)$ defined by Equation (2) is also known. The problem is summarized in Problem 1.1 below.

Problem 1.1 (Newsvendor problem)**.**

1. *The decision-maker chooses* $q \in \mathbb{R}_+$,
2. *a random demand* $d \in \mathbb{R}_+$ *is realized,*
3. *the cost* $\ell(q, d)$ *is incurred.*

Goal: *minimize the expected cost* $L(q)$.
Knowledge: *the demand distribution.*

In fact, this optimization problem has a close-form solution known as the *critical fractile formula.*

Theorem 1.1. *Assume that either the c.d.f.* F *of* D *is continuous or* D *is integer-valued, then, a minimizer of* L *over* $\mathbb{R}_+$ *is given by:*

$$S^* = F^{-1}\left(\frac{p}{h+p}\right), \tag{3}$$

where F^{-1} *is the generalized inverse of* F*, that is:* $F^{-1}(\alpha) = \inf\{x \in \mathbb{R}_+ : F(x) \geq \alpha\}$ *for all* $\alpha \in [0, 1]$.

Proof. See the proofs of [Snyder and Shen, 2019, Theorem 4.1 and Theorem 4.3]. □

1.2 The Newsvendor Problem with Initial Inventory

Now let us assume the initial inventory is $x \in \mathbb{R}$. As it is done in most of the literature tackling similar inventory models, we introduce an equivalent parametrization of the problem. Instead of considering the *order quantity* $q \in \mathbb{R}_+$ as the decision variable we consider the *order-up-to level*[1], y, defined by:

$$y = x + q.$$

The order-up-to level y represent simply the on-hand inventory just before meeting the demand, i.e. the initial inventory x plus the order quantity q. In terms of order-up-to levels the problem of the decision-maker is described in Problem 1.2 below.

Problem 1.2 (Newsvendor problem with initial inventory)**.**

1. *The decision-maker starts with a known initial inventory level* $x \in \mathbb{R}$,
2. *the decision-maker chooses* $y \geq x$,
3. *a random demand* $d \in \mathbb{R}_+$ *is realized,*
4. *the cost* $\ell(y, d)$ *is incurred.*

Goal: *minimize the expected cost function* $L(y)$.
Knowledge: *the demand distribution and the initial inventory* x.

Notice that a constraint on the order-up-to level y appears: it needs to be greater or equal than the initial inventory level x. This is due to the fact that the order quantity q is non-negative.

A consequence of Theorem 1.1, is that this problem has also a close-form solution. This result is stated in Corollary 1.1.

Corollary 1.1. *Let* $x \in \mathbb{R}$. *Assume that either the c.d.f.* F *of* D *is continuous or* D *is integer-valued, then, a minimizer of* L *over* $[x, +\infty)$ *is given by:*

$$y^* = \max\{S^*, x\},$$

where S^* *is defined by Equation* (3).

In other words, the solution of the newsvendor problem with initial inventory is given by the order-up-to level $y^* = \max\{S^*, x\}$ or equivalently by the order quantity $q^* = [S^* - x]^+$. This kind of strategy are well-known in inventory theory and are called *base-stock policies*, a.k.a *order-up-to policies* or $S-$*policies.*

[1] Also called base-stock level.

1.3 Multi-period stochastic inventory problems

Let us now consider multi-period stochastic inventory problems. These can be seen as multi-period extensions of the newsvendor one. In order to state such problems, we will need to specify: the demand generating process and the transition from one period to another.

The demand generating process is assumed to be an *i.i.d* sequence of non-negative random variables $D_1, D_2, \dots$. This assumption leads to an analytically tractable model with a simple optimal policy. At the beginning of the first period we assume the initial inventory is 0. For a given period t, we need to define the *carryover* (or leftover) x_{t+1} which is the initial inventory of period $t+1$. To do so, we use a transition function $\mathcal{T}$ and set:

$$x_{t+1} = \mathcal{T}(y_t, d_t)$$

where y_t and d_t denote the order-up-to level and demand of period t respectively. Various transitions can be considered, let us mention three interesting examples.

Example 1.1 (Backordering)**.** *Excess demand can be satisfied later and these backorders are given by the negative part of the carryover. This corresponds to:* $\mathcal{T}(y,d) = y - d$.

Example 1.2 (Lost sales)**.** *Excess demand is lost, formally:* $\mathcal{T}(y,d) = [y-d]^+$.

Example 1.3 (No carryovers)**.** *The good has a lifetime of* 1 *period, thus, there are no carryovers and the transition function is the null function, i.e.* $\mathcal{T}(y,d) = 0$.

The multi-period stochastic inventory problem is summarized in Problem 1.3 below.

Problem 1.3 (Multi-period stochastic inventory problem)**.**

- *The initial inventory is zero, i.e.* $x_1 = 0$.
- *for* $t = 1, \dots, T$*:*
  1. *the decision-maker chooses* $y_t \geq x_t$,
  2. *a random demand* $d_t \in \mathbb{R}_+$ *is realized,*
  3. *the cost* $\ell(y_t, d_t)$ *is incurred,*
  4. *the carryover* $x_{t+1} = \mathcal{T}(y_t, d_t)$ *is observed.*

Goal: *minimize the expected cumulative cost* $\mathbb{E}\left[\sum_{t=1}^T \ell(y_t, D_t)\right] = \sum_{t=1}^T L(y_t)$.
Knowledge: *the demand distribution.*

It is well-known that this class of problems has a simple solution which is always ordering up to S^*. The statement below is more precise.

Theorem 1.2. *Let* $D_1, \dots, D_T$ *be i.i.d random variables drawn from a continuous or integer-valued distribution with c.d.f* F*. Assume that the transition function satisfies* $\mathcal{T}(y,d) \leq y$ *for all* $y, d \in \mathbb{R}_+$.

Consider the constant sequence $y_t^* = S^*$ *where* S^* *is defined by Equation* (3)*, then,*

- *the sequence* $(y_t^*)_{t=1,\dots,T}$ *is feasible in the sense that* $y_{t+1}^* \geq \mathcal{T}(y_t^*, d_t)$ *for* $t = 1, \dots, T-1$ *and* $d_t \geq 0$,
- *the sequence* $(y_t^*)_{t=1,\dots,T}$ *minimizes* $\mathbb{E}\left[\sum_{t=1}^T \ell(y_t, D_t)\right]$ *compared to any sequence* $y_1, \dots, y_T \in \mathbb{R}_+$.

Proof.

- Due to the assumption on the transition function we have for all $t = 1, \dots, T-1$ and $d_t \geq 0$,

$$y_{t+1}^* = S^* \geq \mathcal{T}(S^*, d_t) = \mathcal{T}(y_t^*, d_t).$$

- Also, we have for all $y_1, \dots, y_T \in \mathbb{R}_+$,

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(y_t, D_t)\right] = \sum_{t=1}^{T} L(y_t) \geq \sum_{t=1}^{T} L(S^*) = \mathbb{E}\left[\sum_{t=1}^{T} \ell(S^*, D_t)\right].$$

□

Notice that the condition on the transition function is very mild, it means that the carryover x_{t+1} should be less or equal than the order-up-to level y_t for any demand realization d_t. This assumption is satisfied for all the aforementioned examples of transitions.

The strategy described in Theorem 1.2 which is to maintain the order-up-to level to S^* is a base-stock policy with a time-invariant base-stock level S^*.

Towards online learning. All these classical inventory problems (see Problems 1.1, 1.2 and 1.3) are essentially optimization problems with close-form solutions. Their goal is to minimize a *known expected* loss, e.g. $L(\cdot)$. In more realistic inventory problems, the decision-maker does not know the demand distribution. Instead, he has access to partial information about the demand through realizations $d_1, d_2, \dots$ in a sequential way. In this case, he faces an *online learning* problem. The decision-maker has to learn somehow the distribution of the demand in order to minimize some loss function. In the next section, we introduce the Online Convex Optimization framework that will be the main tool to solve online inventory problems.

2 Online Convex Optimization

a) Problem statement

Introduced by Zinkevich [2003], Online Convex Optimization (OCO) tackles the sequential decision problem of Problem 2.1 below.

Problem 2.1 (Online Convex Optimization problem)**.**

- *for* $t = 1, \dots, T$*:*
  1. *the decision-maker choose* $u_t \in \mathcal{U} \subset \mathbb{R}^d$*,*
  2. *a convex loss* $\ell_t : \mathcal{U} \to \mathbb{R}$ *is revealed.*

Goal: *minimize the cumulative loss* $\sum_{t=1}^{T} \ell_t(u_t)$*.*
Knowledge at period t**:** *the previous losses* $\ell_1, \dots, \ell_{t-1}$*.*

In this section, we focus on the case where $\mathcal{U}$ is convex and bounded, and the losses are convex and Lipschitz. More precisely, let us consider the following assumptions:

Assumption 2.1. *The set* $\mathcal{U} \subset \mathbb{R}^d$ *is compact and convex with diameter* $D = \max_{u,v \in \mathcal{U}} \|u - v\|_2$*. The losses* $\ell_1, \dots, \ell_T$ *are convex and Lipschitz in the sense that there exists* $G \geq 0$ *such that:* $|\ell_t(u) - \ell_t(v)| \leq G \|u - v\|_2$ *for all* $u, v \in \mathcal{U}$ *and* $t \in \{1, \dots, T\}$*.*

To solve the OCO Problem 2.1, one has to provide an algorithm[2] that outputs at the beginning of period t a point u_t using the past observations $\ell_1, \dots, \ell_{t-1}$.

[2] We may also use the terms: *policy*, *strategy* or *method*.

b) Regret

The performance of an algorithm is usually measured by comparing the cumulative loss of the algorithm with respect to the performance of the best constant strategy in hindsight: $u^* \in \operatorname{argmin}_{u\in\mathcal{U}} \sum_{t=1}^T \ell_t(u_t)$. This performance measure is called the *regret* and it is defined as follows:

$$R_T = \sum_{t=1}^T \ell_t(u_t) - \min_{u\in\mathcal{U}} \sum_{t=1}^T \ell_t(u).$$

Notice that under Assumption 2.1, *any* algorithm has a regret bounded as follows:

$$R_T \leq DGT.$$

One may wonder, what is the minimal scaling of the regret under Assumption 2.1. It is in fact of order $O(\sqrt{T})$. The precise lower bound is given by the statement below.

Theorem 2.1 (Theorem 5.1 of Orabona [2019])**.** *Suppose Assumption 2.1 holds. Let $\mathcal{A}$ be any algorithm for OCO. There exist* linear *loss functions of the form $l_t(\cdot) = \langle g_t, \cdot\rangle$ where $\|g_t\|_2 \leq G$ so that the regret of $\mathcal{A}$ is lower bounded as follows:*

$$R_T \geq \frac{\sqrt{2}}{4} DG\sqrt{T}.$$

c) Online Gradient Descent

In Zinkevich [2003], the *Online Gradient Descent* (OGD) has been introduced to solve the OCO Problem 2.1 when the losses are differentiable. It is defined recursively as follows:

$$u_{t+1} = \operatorname{Proj}_{\mathcal{U}}\left(u_t - \eta_t \nabla \ell_t(u_t)\right),$$

with $u_1 \in \mathcal{U}$ arbitrary. The parameters $\eta_t \geq 0$ are called *learning rates*. $\operatorname{Proj}_{\mathcal{U}}$ denotes the usual projection operator with respect to the euclidean norm.

A classical analysis of the regret shows that, with the adequate choice of learning rates, OGD is optimal in the sense that its regret is of order $O(\sqrt{T})$. In Proposition 2.1 below, we gathered various optimal ways to tune these learning rates. Notice that they differ in the knowledge required for their computation.

Proposition 2.1. *Let Assumption 2.1 be satisfied and assume that $\ell_1, \dots, \ell_T$ are differentiable in open sets containing $\mathcal{U}$, then,*

(i) OGD with constant *learning rates $\eta_t = D/(G\sqrt{T})$ has a regret bounded as follows:*

$$R_T \leq DG\sqrt{T}$$

(ii) OGD with decreasing *learning rates $\eta_t = D/(\sqrt{2}G\sqrt{t})$ has a regret bounded as follows:*

$$R_T \leq \sqrt{2}DG\sqrt{T}$$

(iii) OGD with adaptive *learning rates $\eta_t = D/\left(\sqrt{2}\sqrt{\sum_{i=1}^t \|\nabla \ell_i(u_i)\|_2^2}\right)$ has a regret bounded as follows:*

$$R_T \leq \sqrt{2}D\sqrt{\sum_{t=1}^T \|\nabla \ell_t(u_t)\|_2^2} \leq \sqrt{2}DG\sqrt{T}.$$

Proof.

(i) The regret bound of the constant learning rate $\eta_t = D/(G\sqrt{T})$ follows directly from [Orabona, 2019, Theorem 2.13] as discussed there.

(ii) This regret bound is proposed as Problem 2.2 of Orabona [2019]. It can be proven using [Orabona, 2019, Theorem 2.13]. Indeed, this result guarantees for any non-increasing learning rates that:

$$R_T \leq \frac{D^2}{2\eta_T} + \sum_{t=1}^{T} \frac{\eta_t}{2} \|\nabla \ell_t(u_t)\|_2^2 \leq \frac{D^2}{2\eta_T} + \frac{G^2}{2} \sum_{t=1}^{T} \eta_t.$$

Therefore, taking $\eta_t = D/(\sqrt{2}G\sqrt{t})$ and noticing that $\sum_{t=1}^{T} 1/\sqrt{t} \leq 2\sqrt{T}$, we obtain the desired regret bound.

(iii) This result is a direct consequence of [Orabona, 2019, Theorem 4.14].

□

d) Online Subgradient Descent

When the loss functions are not differentiable, e.g. the absolute value function or the newsvendor cost $\ell(\cdot, d)$ defined by Equation (1), a more general method named *Online Subgradient Descent* (OSD) can be applied. It uses the concept of subgradient which generalizes the concept of gradient.

Definition 2.1 (Subgradient and subdifferential)**.** *A vector $g \in \mathbb{R}^d$ is a* subgradient *of a proper*[3] *function $f : \mathbb{R}^d \to (-\infty; +\infty]$ in $u \in \mathbb{R}^d$ if:*

$$f(v) \geq f(u) + \langle g, v - u \rangle, \forall v \in \mathbb{R}^d.$$

The set of subgradients of f in u is called the subdifferential *of f at u and it is denoted $\partial f(u)$. We say that f is* subdifferentiable *in u if there exists a subgradient of f in u, i.e. when $\partial f(u) \neq \emptyset$.*

Subgradients generalize the notion of gradient in the sense that when f is convex and differentiable at u, the only subgradient of f at u is the gradient $\nabla f(u)$, see e.g. [Rockafellar, 1970, Theorem 25.1].

Example 2.1. *The* absolute value function, $|\cdot|$, *is subdifferentiable over $\mathbb{R}$. Its subdifferential at x, $\partial|\cdot|(x)$, is given by: $\{1\}$ if $x > 0$, $[-1, 1]$ if $x = 0$ and $\{-1\}$ if $x < 0$.*

Example 2.2. *Given any demand realization $d \in \mathbb{R}$, the associated* newsvendor cost, *$\ell(\cdot, d)$, is subdifferentiable over $\mathbb{R}$. Its subdifferential at q, $\partial \ell(\cdot, d)(q)$, is given by: $\{h\}$ if $q > d$, $[-p, h]$ if $q = d$ and $\{-p\}$ if $q < d$.*

The Online Subgradient Descent (OSD) method is defined recursively as follows:

$$u_{t+1} = \text{Proj}_{\mathcal{U}}(u_t - \eta_t g_t),$$

with $u_1 \in \mathcal{U}$ arbitrary and g_t a subgradient of ℓ_t at u_t. Notice that the only difference with OGD is that we replaced the gradient $\nabla \ell_t(u_t)$ by a subgradient g_t.

Remarkably, all the guarantees stated for OGD in Proposition 2.1 hold for OSD replacing the term *differentiable* by *subdifferentiable* and the gradient $\nabla \ell_t(u_t)$ by a subgradient $g_t \in \partial \ell_t(u_t)$. See e.g. [Orabona, 2019, Paragraph 2.2.2 and Theorem 4.14].

[3] Meaning that f is nowhere $-\infty$ and finite somewhere.

3 Online Learning for Inventory Problems: the First Attempt

In this section we detail a simple censored demand inventory problem. We also detail the contributions of Huh and Rusmevichientong [2009] which provided the first algorithm with a provable convergence rate guarantee for this problem.

The censored demand inventory problem is very similar to the multi-period stochastic inventory problem described as Problem 1.3. It differs only in terms of observability structure. Here, we assume that the demand is the realization of an *i.i.d* sequence $D_1, \dots D_T$ with continuous c.d.f F, but this distribution is unknown, furthermore, the demand realizations $d_1, \dots, d_t$ are not directly observed. Only the sales which are censored demand information of the form $\min\{y_t, d_t\}$ are observed. The complete problem is summarized as Problem 3.1 below.

Problem 3.1 (Censored demand inventory problem of Huh and Rusmevichientong [2009])**.**

- *The initial inventory is zero, i.e.* $x_1 = 0$.
- *for* $t = 1, \dots, T$*:*
  1. *the decision-maker chooses* $y_t \geq x_t$,
  2. *a random demand* $d_t \in \mathbb{R}_+$ *is realized,*
  3. *the cost* $\ell(y_t, d_t)$ *is incurred.*
  4. *the carryover* $x_{t+1} = \mathcal{T}(y_t, d_t)$ *is observed.*

Goal: *minimize the expected cumulative cost* $\mathbb{E}\left[\sum_{t=1}^{T} \ell(y_t, D_t)\right] = \mathbb{E}\left[\sum_{t=1}^{T} L(y_t)\right]$.
Knowledge at period t**:** *the previous sales information* $\min\{y_1, d_1\}, \dots, \min\{y_{t-1}, d_{t-1}\}$.

a) Remarks

1. We notice some similarities with the OCO problem (see Problem 2.1) but also important differences. As mentioned in paragraph 1.1 of Huh and Rusmevichientong [2009] the major difficulty in applying OCO methods is the dependency of decisions from one period to another. Indeed, since $y_{t+1} \geq x_{t+1} = \mathcal{T}(y_t, d_t)$, one has to consider dynamic constraints in the design of the algorithm.

2. Another difference with the OCO problem is that, due to demand censoring, one does not observe the loss function $\ell(\cdot, d_t)$. Worst still, we may never observe the cost realization $\ell(y_t, d_t)$. More precisely, the holding part of the cost, $h\left[y_t - d_t\right]^+ = y_t - \min\{y_t, d_t\}$, is always observed but the stockout part of the cost, $p\left[d_t - y_t\right]^+$, may never be observed.

3. The good news is that due to the particular shape of the newsvendor cost function, $\ell(\cdot, d_t)$, one can always observe a particular subgradient $g_t \in \ell(\cdot, d_t)(y_t)$, which is defined by:

$$g_t = h\mathbb{1}_{\{y_t > d_t\}} - p\mathbb{1}_{\{y_t \leq d_t\}} = h\mathbb{1}_{\{y_t > \min\{y_t, d_t\}\}} - p\mathbb{1}_{\{y_t = \min\{y_t, d_t\}\}}.$$

4. Another important consequence of the shape of the newsvendor cost function, $\ell(\cdot, d_t)$, is that given any $\hat{y}_t \leq y_t$, one can compute a subgradient $\hat{g}_t \in \partial\ell(\cdot, d_t)(\hat{y}_t)$ using only the information y_t, $\hat{y}_t$ and $\min\{y_t, d_t\}$. Indeed, if $\min\{y_t, d_t\} = y_t$, then, $\hat{y}_t \leq y_t \leq d_t$ and $\hat{g}_t = -p$ works. Otherwise, $\min\{y_t, d_t\} < y_t$, then $\min\{y_t, d_t\} = d_t$ thus one can compute $\hat{g}_t = h\mathbb{1}_{\{\hat{y}_t > d_t\}} - p\mathbb{1}_{\{\hat{y}_t \leq d_t\}}$.

b) Algorithm and guarantees

Assume the decision-maker knows an upper bound $\bar{y} > 0$ on the optimal base-stock level S^* defined in Equation (3), i.e. $S^* \leq \bar{y}$. The method proposed in Huh and Rusmevichientong [2009] is based on two

sequences: an auxiliary sequence $\hat{y}_t$ that follows a subgradient descent schema and the implemented sequence y_t. These sequences are defined as follows. Set $y_1 = \hat{y}_1 \in [0, \bar{y}]$ and for $t \geq 1$, set:

$$\hat{y}_{t+1} = \text{Proj}_{[0,\bar{y}]}(\hat{y}_t - \eta_t \hat{g}_t), \qquad y_{t+1} = \max\{\hat{y}_{t+1}, x_{t+1}\},$$

where the learning rate $\eta_t = \gamma\bar{y}/(\max\{p,h\}\sqrt{t})$ for $\gamma > 0$ and the $\hat{g}_t$ is the subgradient of $\ell(\cdot, d_t)$ in $\hat{y}_t$ described in remark 3.

Let us present some guarantees proved in Huh and Rusmevichientong [2009]. Those are expressed in terms of the expected regret $\bar{R}_T$ defined as follows:

$$\bar{R}_T = \mathbb{E}\left[\sum_{t=1}^{T}(L(y_t) - L(S^*))\right].$$

Theorem 3.1 (Theorem 2 of Huh and Rusmevichientong [2009])**.** *Let $D_1, \dots, D_T$ be i.i.d non-negative random variables drawn from a continuous distribution. Assume $S^* \leq \bar{y}$, then, the sequence $(y_t)_t$ defined above has the following properties:*

(i) In the zero carryover *case, i.e. $\mathcal{T}(y,d) = 0$, it satisfies the following regret bound for all $T \geq 1$,*

$$\bar{R}_T \leq \left(\gamma + \frac{1}{\gamma}\right)\bar{y}\max\{p,h\}\sqrt{T}.$$

(ii) In the lost sales *case, i.e. $\mathcal{T}(y,d) = [y-d]^+$, and assuming $\mathbb{E}\left[D_1^6\right] < +\infty$ and $\gamma \leq (\rho\max\{p,h\})/(h\bar{y})$ with $\rho \in (0, \mathbb{E}[D_1])$, there exists a constant $C > 0$ such that for all $T \geq 1$ we have*

$$\bar{R}_T \leq C\sqrt{T}.$$

Proof.

(i) In the no carryover case, we have $x_t = 0$, thus, $\hat{y}_t = y_t$ for all $t \geq 1$. Therefore, this first claim is a direct consequence of OCO theory, that holds even with no probabilistic assumptions. Indeed, consider the OCO problem 2.1 with losses $\ell(\cdot, d_1), \dots, \ell(\cdot, d_T)$ over the set $\mathcal{U} = [0, \bar{y}]$. Then, by applying *(ii)* of Proposition 2.1 we obtain a regret bound of the form $R_T \leq \sqrt{2}\bar{y}\max\{p,h\}\sqrt{T}$ for $\gamma = 1/\sqrt{2}$. Taking the expectation leads to the bound on the expected regret $\bar{R}_T$. To obtain a more general result valid for all $\gamma > 0$ one can use instead Theorem 2.13 of Orabona [2019].

(ii) The proof of the second point is based on the following decomposition:

$$\bar{R}_T = \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}(\ell(\hat{y}_t, d_t) - \ell(S^*, d_t))\right]}_{\text{Expected regret of } \hat{y}_t} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}(\ell(y_t, d_t) - \ell(\hat{y}_t), d_t)\right]}_{\text{Error term}}.$$

The expected regret of the algorithm $\hat{y}_t$ is bounded using elementary results from OCO as in the zero carryover case. Bounding the second term is much more involved, it borrows techniques from queuing theory which are out of the scope of this document.

□

4 Other Inventory Problems Solved Using Online Learning

Many other inventory problems have been tackled using online learning techniques. In the following we briefly present various papers doing so.

- *Nonparametric data-driven algorithms for multiproduct inventory systems with censored demand*, Shi et al. [2016].

  This paper tackles a lost sales censored demand inventory problem with n products and a warehouse-capacity. They propose an algorithm which achieves the optimal rate of $O(\sqrt{T})$ in terms of expected regret.

  Each product demands are i.i.d across time periods, and the different products have independent demands. Additional regularity assumptions on the demand are introduced. The warehouse-capacity constraint defines the set of feasible base-stock levels: $\{\mathbf{y} \in \mathbb{R}^n_+ : \sum_{i=1}^n y_i \leq M\}$

  In this problem the optimal policy with complete distributional information is a time-invariant base-stock policy. See Ignall and Veinott Jr [1969]. This is similar to the single product problem of Huh and Rusmevichientong [2009] we introduced as Problem 3.1.

  As in Huh and Rusmevichientong [2009] the algorithm they propose is based on a time-dependent base-stock policy computed in an online fashion. Unconstrained base-stock levels $\hat{\mathbf{z}}$ are updated at some periods using a stochastic gradient step then projected into the feasible set defining target base-stock levels $\hat{\mathbf{y}}$. Due to the constraints and carry-over, $\hat{\mathbf{y}}$ may not be implementable, a transformation of it, $\mathbf{y}$, is computed and applied at every period. The gradient step are done during periods where $\mathbf{y} \geq \hat{\mathbf{y}}$, since it happens that this event makes the gradient direction for $\hat{\mathbf{z}}$ observable, although, demand is censored by the base-stock levels $\mathbf{y}$.

- *Perishable inventory systems: Convexity results for base-stock policies and learning algorithms under censored demand*, Zhang et al. [2018]. This paper provides the first online learning algorithm for a perishable inventory system with lost-sales and censored demand.

  In this perishable inventory system, newly ordered products are fresh units that have a fixed usable lifetime of m periods. The inventory is depleted according to an oldest first basis: oldest products are sold first. The demands across periods are, as usual, assumed to be realizations of *i.i.d.* continuous random variables. Unmet demand is lost. In addition to the usual holding cost h and penalty cost p an outdating cost θ per outdated unit is considered.

  It is known that for this perishable inventory system the optimal policy with complete distributional information of demand is rather complex. It has been shown that it depends on both the age distribution of the current inventory and the remaining period. See e.g. Nahmias [2011]. In particular, it is not a base-stock policy.

  The algorithm they propose follows a time-varying base-stock strategy, where at every "cycle update" the base-stock level is updated using a stochastic gradient step. A new cycle begins when lost sales occurs. The expected regret of this algorithm compared to the best time-invariant base-stock policy is $O(\sqrt{T})$. See [Zhang et al., 2018, Theorem 2].

- *On the hardness of inventory management with censored demand data*, Lugosi et al. [2017]. This paper considers the multi-period stochastic inventory problem with censored demand and no carryovers. They do not consider probabilistic assumptions whatsoever, that is, the demand generating process is arbitrary and can even be *adversarial*. It assumes however that the set of possible demands and possible order quantities are finite.

To solve this problem they relate it to the sequential decision problem of expert aggregation which resembles the OCO problem defined as Problem 2.1 but with a finite decision set $\mathcal{U}$. In this case randomized strategies are usually consider. They consider the so called Exponentially Weighted Forecaster in addition to carefully designed cost estimators because of demand censoring. They show that their policy achieves the optimal square root expected regret up to logarithmic factors.

Let us mention some more recent contributions of online learning to inventory problems. In Zhang et al. [2020] an algorithm are proposed to solve a lost sales censored demand inventory problem in the presence of order lead times. Also, Yuan et al. [2021] provides an algorithm based on a gradient descent approach and bandits to solve an inventory problem with fixed costs.

5 Conclusion: Real-World Applications and Future Work

Califrais is an innovative foodtech startup based in Paris. It operates the e-commerce platform Rungis Market which supplies restaurants and other food industry professionals from Rungis market: the world's biggest fresh produce market.

For research purposes, Califrais and prestigious academic instutions like the CNRS, Sorbonne University or Université Paris Cité cooperate through the joint lab: LabCom LOPF (Large-scale Optimization of Product Flows). Our work is part of this cooperation.

Amongst other machine learning problems such as demand prediction, Califrais faces inventory problems with many specific features like: large-scale multi-item system with inventory bounds, perishable goods, order lead times, non-stationary intermittent demand.

As we have seen in Section 4, there has been attempts to solve various inventory problems using online learning methods. Each attempt we presented tackles an inventory problem with a specific feature amongst: multiple products with inventory bounds, perishability, order lead times or additional fixed costs. For some of these models, the optimal strategy even with complete distributional information is complex and intractable, e.g. perishable systems or lost sales systems with lead times. In a future work, we would like to propose a method with theoretical guarantees such as regret bounds (lower and upper bounds) that includes simultaneously several problem features.

Also, almost all the literature assume i.i.d demands and provide bounds on the expected regret. However, we don't think this assumption is realistic. In a future work, we would like to consider non-stationary demands, through either novel stochastic assumptions or a deterministic framework as in Lugosi et al. [2017]. The latter considers the zero carryover case with a finite set of demands and order quantities. As highlighted by Huh and Rusmevichientong [2009] (which assumes i.i.d demands), considering lost sales transitions is rather challenging because of the dependency of decisions from one period to another. This becomes even more difficult in a non-stationary demand setup.

References

K. J. Arrow, T. Harris, and J. Marschak. Optimal inventory policy. *Econometrica: Journal of the Econometric Society*, pages 250–272, 1951. (Cited on page 1.)

E. Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019. (Cited on page 1.)

W. T. Huh and P. Rusmevichientong. A nonparametric asymptotic analysis of inventory planning with censored demand. *Mathematics of Operations Research*, 34(1):103–123, 2009. (Cited on pages 1, 7, 8, 9, and 10.)

E. Ignall and A. F. Veinott Jr. Optimality of myopic inventory policies for several substitute products. *Management Science*, 15(5):284–304, 1969. (Cited on page 9.)

G. Lugosi, M. G. Markakis, and G. Neu. On the hardness of inventory management with censored demand data. *arXiv preprint arXiv:1710.05739*, 2017. (Cited on pages 9 and 10.)

S. Nahmias. *Perishable inventory systems*, volume 160. Springer Science & Business Media, 2011. (Cited on page 9.)

F. Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019. (Cited on pages 1, 5, 6, and 8.)

R. Rockafellar. Convex analysis, princeton univ. press, 1970. (Cited on page 6.)

S. Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2011. (Cited on page 1.)

C. Shi, W. Chen, and I. Duenyas. Nonparametric data-driven algorithms for multiproduct inventory systems with censored demand. *Operations Research*, 64(2):362–370, 2016. (Cited on page 9.)

L. V. Snyder and Z.-J. M. Shen. *Fundamentals of supply chain theory*. John Wiley & Sons, 2019. (Cited on pages 1 and 2.)

H. Yuan, Q. Luo, and C. Shi. Marrying stochastic gradient descent with bandits: Learning algorithms for inventory systems with fixed costs. *Management Science*, 2021. (Cited on page 10.)

H. Zhang, X. Chao, and C. Shi. Perishable inventory systems: Convexity results for base-stock policies and learning algorithms under censored demand. *Operations Research*, 66(5):1276–1286, 2018. (Cited on page 9.)

H. Zhang, X. Chao, and C. Shi. Closing the gap: A learning algorithm for lost-sales inventory systems with lead times. *Management Science*, 66(5):1962–1980, 2020. (Cited on page 10.)

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003. (Cited on pages 4 and 5.)