Dynamic Inventory and Price Controls Involving Unknown Demand on Discrete Nonperishable Items

Michael N. Katehakis, Jian Yang, and Tingting Zhou
Department of Management Science and information Systems
Business School, Rutgers University, Newark, NJ 07102
June 2017

Abstract

We study adaptive policies that handle dynamic inventory and price controls while the random demand for discrete nonperishable items is unknown. Pure inventory control is achieved by targeting newsvendor ordering quantities that correspond to empirical demand distributions learned over time. On the basis of it we conduct the more complex joint inventory-price control, whereupon demand-affecting prices are chosen. We identify policies that strive to balance between exploration and exploitation, and measure their performances by regrets, i.e., the prices to pay for not knowing demand distributions a priori. Also derived are bounds in the orders of $T^{1/2}$, $T^{2/3}$, $T^{3/4}$, and $T^{5/6}$ for the regrets, depending on how thoroughly unknown the demand distributions are and whether nonperishability has indeed been accounted for. Our simulation sheds lights on the true growth rate of regrets, and also hints at directions for future research.

Keywords: Inventory Control; Newsvendor; Empirical Distribution; Adaptive Policy; Regret; Joint Inventory-price Control; Large Deviation; Information Theory
1 Introduction

For a given firm, inventory control is about dynamically adjusting ordering quantities to minimize the total long-run expected cost. When unit prices influence demand it faces, the firm can further exert joint inventory-price control to attain the maximum total long-run expected profit. In traditional models, demand levels faced by a firm are often assumed to be random, however, with known probabilistic distributions. Even the knowledge on demand distributions, as it often turns out, can be too optimistic a proposition. When the firm has just introduced a new product or when its external environment has just transitioned to a previously unfamiliar phase (such as a severe economic downturn), it will not be sure of the demand patterns to come forth. One way out is adopting the Bayesian approach. In it, the firm possesses prior distributions on potential demand patterns. Then, posterior understanding on demand is updated by its realized levels. Inventory management taking this approach can be found in, for instance, Scarf [33] and Lariviere and Porteus [29].

Most other times, even prior distributions on the demand pattern can be elusive. What meager information one possesses might just be a collection of potential demand distributions. Now, the concerned firm has still to make decisions based on its past observations. But its goal is no longer about catering to specific demand distributions or even sequences of posterior demand distributions. Rather, its history-dependent (henceforward called adaptive) control policy should better yield results that are reasonably good under all potential demand distributions from the given collection. A given policy’s regret under a given demand pattern and over a fixed time horizon measures the price paid for ambiguity; namely, it registers the difference between the policy’s performance and that of the best policy tailor-made for the demand pattern had it been known. A policy will be considered good when its worst regret over all demand patterns in a collection grows over time as slowly as possible.

We suppose items being ordered and demanded are discrete. Thus, the underlying firm can be, for instance, a car manufacturer or a wholesaler of bulky items such as bags of grain. Often, the granularity of items does not cause much difference in the analysis of inventory and price controls, and continuous- and discrete-item settings are good approximations of each other. The latter choice has the advantage that, with the finest tuning of ordering decisions all but impossible, it reflects better the inability to be absolutely certain on minuscule details in real inventory management.

We start with pure inventory control, where demand can have an unknown distribution $f$ over the discrete support $\{0, 1, \ldots, \bar{d}\}$, where $\bar{d}$ is a given maximum demand. We adopt a
very simple and natural policy that has also been considered by Besbes and Muharremoglu [6]. Recall that the optimal ordering quantity for a newsvendor problem involving effective holding cost rate \( \bar{h} \) and effective backlogging cost rate \( \bar{b} \) is the \( \beta \)-quantile of the distribution \( f \), where \( \beta = \bar{b}/(\bar{h} + \bar{b}) \). In every period \( t \), the heuristic policy advocates ordering up to the \( \beta \)-quantile of the empirical demand distribution \( \hat{f}_{t-1} \) that is learned from past demand levels in periods \( 1, 2, ..., t-1 \). We consider two cases, in both of which nonperishability poses as the main difficulty. For our first case, though the flat growth rate of regret over time is impressive, the result can be faulted by its requirement on the unknown distribution’s behavior around the \( \beta \)-quantile. We thus go on to a second, more involved, case where every potential demand distribution is anticipated. We show that the worst regret over all distributions will not grow faster than the rate \( T^{1/2} \cdot (\ln T)^{3/2} \). The part of the bound attributable to items’ nonperishability is later repeatedly used in joint inventory-price control.

A good portion of the paper is then devoted to the more complex joint inventory-price control. Here, the firm can choose from prices \( \bar{p}^1, \bar{p}^2, ..., \bar{p}^k \). But the demand distribution \( f^k \) under each choice \( k = 1, 2, ..., k \) is unknown. Now the empirical distribution \( \hat{f}^k_{t-1} \) of demand under price choice \( k \) by the beginning of period \( t \) depends on how many times \( k \) has been chosen in periods \( 1, 2, ..., t - 1 \), and is in turn the product of both demand realizations and the policy being adopted. We propose learning while doing policies \( \text{LwD}(\mu) \) that are parameterized by some \( \mu \in [1/2, 1) \). For a given \( \mu \), the policy ensures that the number of times that each price choice \( k \) is visited by time \( t \) grows at least in the order of \( t^\mu \). Meanwhile, the more competitive a price is, the more likely it will be chosen. Let \( V^k_f \) be the best average single-period profit the firm can achieve under price choice \( k \) and demand distribution \( f \). As the actual \( f^k \) under choice \( k \) is unknowable to the firm, the policy advocates that, as much as possible, the \( k \) achieving the maximum \( V^k_{\hat{f}^k_{t-1}} \) be chosen in period \( t \).

As in pure inventory control, the analysis dealing with perishable items draws upon established results in information theory and large deviation, such as Hoeffding’s inequality. Basically, we take advantage of the fact that empirical distributions will get infinitesimally close to their generating distributions as increasingly more realizations are observed. The issue of nonperishability necessitates more innovations on our part. We take up the nonperishability-induced bound in pure control and for this to work well, introduce virtual learning periods that accumulate at rates roughly proportional to \( t^\mu \). This trick allows us to establish that the dominant price, if there ever is one, will be used in long sequences of periods; consequently, the sub-linear bound from pure control will be patched up over
various sequences to deliver a reasonable bound for joint control.

The most challenging case is when demand is so utterly unknown that even the existence of a price leading others by a tiny margin can not be taken for granted. To handle this, we build on the virtual-learning idea to obtain stickier modifications of the LwD(µ) policies which we call LwD′(µ, ν, ψ), for ν > 0 and ψ ∈ (0, µ). The latter policies are aware of built-in virtual learning periods that accumulate at rates proportional to t1−ψ and in most periods, favor incumbent price choices at degrees expressible in terms of the parameters µ, ν, and ψ, as well as the times t. The extra lengths that single prices linger on permit the pure-control bound to take its effect. In addition to performance guarantees for particular policies, we also attempt to identify lower bounds on regrets that no policy can ever beat.

Now, let us detail our main theoretical results in the following.

**Pure inventory control**

- when the probability f(¯d) is bounded below by some ϵ > 0 and an indicator of the separation between the distribution f and the coefficient β is bounded below by some δ > 0:
  - * regrets are bounded by an (ϵ, δ, β)-dependent constant (Theorem 1);

- when demand distribution f can roam absolutely freely:
  - * regrets are of the form O(T1/2 · (ln T)3/2) (Theorem 2), with the non perishability-induced part coming from Proposition 2 and remaining useful for joint control;

**Joint inventory-price control**

- when items are perishable:
  - when Vk is uniquely maximized and the second best choice is at least δ > 0 away:
    - * the LwD(µ) policy with the best known performance guarantee happens at µ = 1/2, whence regrets are of the form O(T1/2 · (ln T)1/2) (Proposition 5);
  - when demand-distribution vector f ≡ (f1, f2, ..., fK) can roam absolutely freely:
    - * the LwD(µ) policy with the best known performance guarantee happens at µ = 2/3, whence regrets are of the form O(T2/3) (Proposition 6);

- when items are non perishable and Vk is uniquely maximized with a δ margin:
  - * the LwD(µ) policy with the best known performance guarantee happens at µ = 1/2, whence regrets are of the form O(T3/4 · (ln T)5/2) (Theorem 3);

- when items are non perishable and demand patterns are truly arbitrary:
  - * it takes the LwD′(µ, ν, ψ) policies to achieve known performance guarantees; the best known one is attained at µ = 2/3 and ψ = 1/3, whence regrets are of the form O(T5/6 · (ln T)5/2) (Theorem 4);
the tightest lower bound so far achievable is of the form $\Omega(T^{1/2})$ (Theorem 5).

Moreover, we have conducted a simulation study. Its pure inventory control part demonstrates the competitiveness of the newsvendor-based policy. Through the study, we also confirm that the underlying distribution’s separation from $\beta$ plays a prominent role in determining the regret generated by this policy. The part concerning joint inventory-price control suggests that more work is probably needed on both the upper-bounding Theorem 4 and lower-bounding Theorem 5. It is likely that $T^{3/5}$- or $T^{2/3}$-sized bounds should prevail for joint inventory-price control just as $T^{1/2}$-sized ones do for pure inventory control. It also indicates that nonperishability does not contribute as much to regret bounds as suggested by our theoretical bounds, on which we have spent major efforts. Therefore, new ideas, especially those that do not rely on the pure-control bound, are still wanted.

The remainder of the paper is organized as follows. We put our contribution in the perspective of existing literature in Section 2. Then, Section 3 introduces pure inventory control along with the newsvendor-based policy; whereas, Section 4 provides various upper bounds. For joint inventory-price control, we use Section 5 to introduce the problem and the LwD($\mu$) policies. We then spend the next three sections on detailed analyses. Section 6 focuses on the case with perishable items; moreover, Section 7 moves on to nonperishable items, however, with a slight restriction on demand patterns. In Section 8, we achieve upper bounds for the modified policies LwD($\mu, \nu, \psi$) under the most general conditions involving nonperishable items and unrestricted demand patterns. A lower bound is also derived there. We present our simulation study in Section 9 and conclude the paper in Section 10.

2 Literature Survey

Pioneering works on regret analysis started with adaptive allocation, where the main concern is on dynamically selecting the most promising pool of samples to draw so as to maximize their total sum; see, e.g., Robbins [31], Lai and Robbins [28], Katehakis and Robbins [26], and Auer, Cesa-Bianchi, and Fischer [2]. Also, Auer et al. [3] treated variants where each choice’s output is not necessarily an independent sample from a predetermined though unknown distribution; meanwhile, Burnetas, Kanavetas, and Katehakis [11] introduced constraints on the total costs of sampling from various pools. The dynamic pricing portion of our work is akin to picking a winning pool of samples. However, subsequent ordering decisions and inventory carry-overs pose additional challenges.
Regret analysis was also conducted on adaptive Markov decision processes (MDPs) that often involve unknown reward patterns and unknown random state transitions; see, e.g., Burnetas and Katehakis [12], Auer and Ortner [4], Tewari and Bartlett [34], and Jacksh, Ortner, and Auer [25]. Regret bounds derived in this body of literature are often dependent on particular MDPs or to lesser degrees, characteristics of MDPs such as their so-called diameters. So even inventory and price controls are MDP by nature, we need new insights and techniques to achieve regret bounds that countenance all or nearly all possible demand distributions that underlie our particular Markov processes.

Adaptive policies for inventory control have been considered. Huh and Rusmevichientong [24] analyzed a gradient-based policy suitable to the continuous-item case. Their policy could also be thought of as an extension of stochastic approximation (SA), which was started by Robbins and Monro [32] and Kiefer and Wolfowitz [27]. More recently, Besbes and Muharremoglu [6] studied the implications of demand censoring in pure inventory control involving unknown demand. They focused on the discrete-item case of the repeated newsvendor problem and proposed policies with provably good performance guarantees. In a revenue management setup, Besbes and Zeevi [7] studied the dynamic selection of prices while learning demand on the fly. For the newsvendor problem and its multi-period version involving nonperishable items, Levi, Roundy, and Shmoys [30] relied on randomly generated demand samples to reach solutions with relatively good qualities at high probabilities.

We study both pure inventory and joint inventory-price controls involving the real-time learning of unknown demand patterns. In pure inventory control, we deal with the nonperishability of items as Huh and Rusmevichientong [24] did. When items are discrete, however, their SA-based approach requiring precise, often non-integer resolutions of order-up-to levels, is no longer applicable. Rather, the newvendor-based policy relying on empirical demand distributions, as was considered by Besbes and Muharremoglu [6] in their study of perishable items, comes as a natural choice. Our contribution is rather technical, in bounding the policy’s nonperishability-induced regrets. Proposition 2, especially, provides a bound when no restriction is placed on the demand distribution $f \text{ a priori}$. It is also an essential component that enables the analysis of nonperishability-induced regrets in joint inventory-price control.

Pure price control involving unknown demand patterns, as was treated in Besbes and Zeevi [7], is a problem transient in nature. In it, prices are adjusted over time to reap the highest profit from selling a given initial stock in a fixed time horizon. This area is seeing rapid progresses in recent years. For instance, Wang, Deng, and Ye [35] proposed
a policy that conducts learning and doing intermittently and achieves very tight regret bounds. Besbes and Zeevi [8] demonstrated that a firm can pretend its demand function to be linear and still manage to avoid severe regrets. Ferreira, Simchi-Levi and Wang [22] applied Thompson Sampling to a network revenue management, where different products consume a given set of resources. Also, Aviv and Pazgal [5], Araman and Caldentey [1], Farias and van Roy [20], Broder and Rusmevichientong [10], and den Boer and Zwart [9] took parametric approaches to such problems. Meanwhile, Cheung, Simchi-Levi, and Wang [16] limited the number of price changes in a setting without inventory constraints.

In contrast, joint inventory-price control is recurrent in nature. It deals with the repeated use of pricing and ordering for the attainment of the highest profit in the long run. Our counterpart with known demand patterns and strict discounts over time is Federgruen and Heching [21]. Assuming unknown demand patterns, Burnetas and Smith [13] studied such a problem where demand distributions are continuous and the only information available is whether sales have exceeded ordering quantities. They developed an SA-based method that reached consistency, i.e., asymptotic convergence in time-average profit to the truly optimal; however, rates of convergence were left untouched.

Like us, Chen, Chao, and Ahn [14] also dealt with joint inventory-price control involving unknown demand patterns while providing analysis on regret’s growth rate. Whereas we allow a finite number of price choices and assume an arbitrary price-demand relationship, they let prices come from a compact interval, adopted an either additive or multiplicative demand pattern with average demand decreasing over price, and also made other assumptions like the twice differentiability of the demand-price relationship’s deterministic part, concavity of the average revenue function, and strict positivity of average demand levels. They were able to achieve $T^{1/2}$-sized regret bounds. Between the slightly earlier work and ours, we believe there is some trade-off between model flexibility and result quality. Our avoidance of any assumption on the price-demand relationship reduces the risk of model mis-specification to the minimum. On the flip side, this probably contributes to our less ideal, $T^{3/4}$- and $T^{5/6}$-sized bounds, as illustrated in Theorems 3 and 4, respectively. Working with a setup similar to that of Chen, Chao, and Ahn [14], recently Chen, Chao, and Shi [15] established a $T^{4/5}$-sized regret bound for the case involving lost sales and censored demand observation.

Our results, especially those allowing demand patterns total freedom in their ranges, such as Theorems 2 and 4, will offer guidance to production, inventory, and sales managers at those critical junctures when for instance, new products have just been rolled out or the economy
has just entered a new phase. Besides effective uses of the empirical distribution’s properties known in the theories of information and large deviations, we contribute methodologically in the LwD(µ) policies that balance between exploration and exploitation, their stickier variants the LwD′(µ, ν, ψ) policies that favor incumbent prices to measured degrees, the virtual-learning trick that regulate frequencies of learning in both directions, and various other regret-analysis techniques. In both pure and joint controls, nonperishability of discrete items poses as one of the main challenges. Our treatments of Propositions 10 and 14 to this effect might lend ideas to other applications.

Due to presently unsurmountable difficulties, there is still a sizable gap between the $T^{5/6}$-sized upper bound in Theorem 4 and the $T^{1/2}$-sized lower bound in Theorem 5. Our simulation study adds to the credulity of a $T^{2/3}$-sized bound. In addition, it questions whether nonperishable items contribute as much to the final bound as has been depicted in Propositions 10 and 14. We probably need a new proof idea other than the current one focusing on long-lasting single dominant prices.

3 Pure Inventory Control

We consider a multi-period inventory control problem in which unsatisfied demand is either backlogged or lost. Also, items are nonperishable so that those unsold in one period are carried over to the next period. Demand $D_t$ in each period $t = 1, 2, ..., T$ is a random draw from a distribution with discrete support $\{0, 1, ..., \bar{d}\}$, where $\bar{d}$ is some positive integer. A distribution is basically a vector $f \equiv (f(0), f(1), ..., f(\bar{d}))$ in the $\bar{d}$-dimensional simplex

$$\Delta \equiv \left\{ f \in [0, 1]^{\bar{d}+1} : \sum_{d=0}^{\bar{d}} f(d) = 1 \right\} \subset \mathbb{R}^{\bar{d}+1}. \quad (1)$$

We use $F_f$ to denote the cumulative distribution function (cdf) associated with any given $f \in \Delta$. It satisfies $F_f(x) = \sum_{d=0}^{[x]} f(d')$ for $x \in \mathbb{R}$.

Suppose the planning horizon constitutes periods 1, 2, ..., $T$, and the firm starts with nothing at the beginning of period 1. We can handle all four combinations where unsatisfied demands are either backlogged or lost and where leftover items are either nonperishable or perishable. Suppose in any period $t = 1, 2, ..., T$, the order-up-to level is $y_t$ and the realized demand level is $d_t$. Then, based on the discussion in Section 1 of Zhou, Katehakis, and Yang [36], some $\sum_{t=1}^{T} q(y_t, d_t)$ will serve as the relevant cost term over the $T$-period horizon for
any one of the four setups. All we need is that

\[ q(y, d) \equiv \bar{h} \cdot (y - d)^+ + \bar{b} \cdot (d - y)^+, \tag{2} \]

with \( \bar{h} \) and \( \bar{b} \) as strictly positive constants. In (2), \( \bar{h} \) is interpretable as the unit holding cost rate when leftover items are nonperishable. When items are perishable, \( \bar{h} \) can be understood as the difference \( \bar{c} - \bar{s} \), where \( \bar{c} \) is the unit production cost and \( \bar{s} \) the unit salvage value. Meanwhile, \( \bar{b} \) is interpretable as the unit backlogging cost rate when unsatisfied demands are backlogged. When those demands are lost, it can be treated as the difference \( \bar{l} - \bar{c} \), where \( \bar{l} \) is the cost for not satisfying a unit demand. Another signature of items' nonperishability is the requirement that

\[ y_t \geq y_{t-1} - d_{t-1}, \quad \forall t = 2, 3, ..., T. \tag{3} \]

Define \( Q_f(y) \) for every \( f \in \Delta \) and \( y = 0, 1, ..., \bar{d} \), so that

\[
Q_f(y) \equiv \mathbb{E}_f[q(y, D)] = \sum_{d=0}^{\bar{d}} f(d) \cdot [\bar{h} \cdot (y - d)^+ + \bar{b} \cdot (d - y)^+]
= \bar{h} \cdot \sum_{d=0}^{y-1} F_f(d) + \bar{b} \cdot \sum_{d=y}^{\bar{d}-1} (1 - F_f(d)), \tag{4}
\]

as the single-period average cost under order-up-to level \( y \). Let \( Q_f^* = \min_{y=0,1,...,\bar{d}} Q_f(y) \) be the minimum cost in one period under \( f \). Suppose \( y_f^* \) is an order-up-to level that achieves the one-period minimum. Then, when facing a \( T \)-period horizon, an optimal policy will be to repeatedly order up to this level. Thus, the minimum cost over \( T \) periods is \( Q_f^* \cdot T \).

A salient feature of our current problem, however, is that \( f \) is not known beforehand. So instead of any \( f \)-dependent policy, we seek a good \( f \)-independent policy which takes advantage of demand levels observed in the past. A deterministic policy \( y \equiv (y_1, y_2, ...) \) is such that, for \( t = 1, 2, ..., \) each \( y_t = 0, 1, ..., \bar{d} \) is a function of the historical demand vector \( d_{[1,t-1]} \equiv (d_1, ..., d_{t-1}) \in \{0, 1, ..., \bar{d}\}^{t-1} \). Under it, the \( T \)-period total average cost is

\[
Q_f^T(y) \equiv \sum_{t=1}^{T} \mathbb{E}_f[\bar{h} \cdot (y_t(D_{[1,t-1]}) - D_t)^+ + \bar{b} \cdot (D_t - y_t(D_{[1,t-1]}))^+], \tag{5}
\]

which, due to the independence between \( D_{[1,t-1]} \) and \( D_t \), is equal to \( \sum_{t=1}^{T} \mathbb{E}_f[Q_f(y_t(D_{[1,t-1]}))] \). Now define \( T \)-period regret \( R_f^T(y) \) of using policy \( y \) against the unknown distribution \( f \):

\[
R_f^T(y) \equiv Q_f^T(y) - Q_f^* \cdot T = \sum_{t=1}^{T} \mathbb{E}_f[Q_f(y_t(D_{[1,t-1]}))] - Q_f^* \cdot T. \tag{6}
\]

Here, the ultimate goal should be that of identifying adaptive policies \( y \) that prevent \( R_f^T(y) \) from growing too fast in \( T \) under all or at least “most” \( f \)'s within \( \Delta \).
We concentrate on one policy inspired by an optimal \( y_f^* \) when \( f \) is known. From (4), we see that necessary and also sufficient conditions for optimality of any \( y \) are

\[
Q_f(y + 1) - Q_f(y) = (\bar{h} + \bar{b}) \cdot F_f(y) - \bar{b} \geq 0,
\]

(7)

and

\[
Q_f(y) - Q_f(y - 1) = (\bar{h} + \bar{b}) \cdot F_f(y - 1) - \bar{b} \leq 0.
\]

(8)

Let \( \beta = \bar{b}/(\bar{h} + \bar{b}) \) be the famous newsvendor parameter that lies in \((0,1)\). For \( f \in \Delta \), let \( y_f^* \) be the associated newsvendor order-up-to level, so that

\[
y_f^* \equiv F_f^{-1}(\beta) \equiv \min\{d = 0, 1, ..., \bar{d} : F_f(d) \geq \beta\}.
\]

(9)

For simplicity, we have omitted \( y_f^* \)'s dependency on the cost parameter \( \beta \). By definition, \( F_f(y_f^*) \geq \beta \) and hence \( Q_f(y_f^* + 1) - Q_f(y_f^*) \geq 0 \) by (7); also, \( F_f(y_f^* - 1) < \beta \) and hence \( Q_f(y_f^*) - Q_f(y_f^* - 1) < 0 \) by (8). Therefore, \( Q_f(y_f^*) = Q_f^* \), meaning that \( y_f^* \) is an optimal order-up-to level for the one-period problem when \( f \) is known.

Instead of a finite upper bound \( \bar{d} \) on demand, the discrete-item case of Besbes and Moharremoglu [6] allowed demand to come from positive integers. On the flip side, they limited demand distributions \( f \) to those with \( y_f^* \)'s below some constant \( \bar{y} \) for the given \( \beta \). Practically, both restrictions are well founded. For a grocery store serving a community of ten thousand people, one can either impose \( \bar{d} = 10,000 \) as the maximum daily demand for gallons of milk or \( \bar{y} = 10,000 \) as the maximum order-up-to level. Neither will be even remotely challenged in real life. Technically, the two constants would play quite interchangeable roles.

We could have replaced \( \bar{d} \) with \( \bar{y} \) in many of our constants had we adopted the earlier approach. The choice made here permits us to present the \( \beta \)-independent case of \( f \in \Delta \), with \( \bar{d} \) already embedded in \( \Delta \) at (1), as one where demand distribution roams freely.

Now with \( f \) unknown, we might adopt level \( y_{f_{t-1}}^* \) with \( f_{t-1} \) being a good estimate of \( f \). The prime candidate for \( f_{t-1} \) is the empirical distribution \( \hat{f}_{t-1} \). For \( t = 2, 3, ..., \), define \( \hat{f}_{t-1} \in \Delta \) by \( (\hat{f}_{t-1}(0), \hat{f}_{t-1}(1), ..., \hat{f}_{t-1}(\bar{d})) \), so that for every \( d = 0, 1, ..., \bar{d} \),

\[
\hat{f}_{t-1}(d) = \frac{\sum_{s=1}^{t-1} \mathbf{1}(d_s = d)}{t - 1},
\]

(10)

Each \( \hat{f}_{t-1} \) has its corresponding cdf \( \hat{F}_{t-1} \equiv F_{\hat{f}_{t-1}} \). Our heuristic policy applies the newsvendor formula to the empirical demand distribution. It lets the firm order nothing in period 1; that is, \( y_1 = \hat{y}_1 = 0 \). For any \( t = 2, 3, ..., \) it advises the firm to order up to

\[
y_t = \hat{y}_t \lor (y_{t-1} - d_{t-1}),
\]

(11)
in period \( t \) so that (3) is satisfied, in which
\[
\hat{y}_t = y_{f_{t-1}}^* = \hat{F}^{-1}_{f_{t-1}}(\beta) = \min \left\{ d = 0, 1, \ldots, \bar{d} : \sum_{s=1}^{t-1} 1(d_s \geq d) \geq \beta \cdot (t-1) \right\}. \tag{12}
\]

For the lost sales case, we need to guarantee that \( y_t \geq 0 \) and hence enhance (11) to
\[
y_t = \hat{y}_t \vee (y_{t-1} - d_{t-1})^+. \tag{13}
\]
However, the current heuristic through (12) has ensured that \( \hat{y}_t \geq 0 \). So the same (11) can still be used.

4 Bounds for Pure Control

We show that the \( f \)-blind and yet adaptive policy \( y \) described by (11) and (12) will incur regret \( R_T^f(y) \) as defined by (6) that is slow-growing in the planning length \( T \). By (6) and (11),
\[
R_T^f(y) = R_{fT1}^T(y) + R_{fT2}^T(y), \tag{13}
\]
where
\[
R_{fT1}^T(y) = \sum_{t=1}^{T} \mathbb{E}_f[Q_f(\hat{y}_t)] - Q_f^* \cdot T, \tag{14}
\]
and, since \( y_1 = \hat{y}_1 = 0 \) by design and hence \( y_2 = \hat{y}_2 \),
\[
R_{fT2}^T(y) = \sum_{t=3}^{T} \mathbb{E}_f[Q_f(y_t) - Q_f(\hat{y}_t)]. \tag{15}
\]

In view of (4) and (15), over-payment in holding might be more than offset by under-payment in backlogging. So \( R_{fT2}^T(y) \) for an arbitrary policy \( y \) might even be strictly negative. Still, it can be said that \( R_{fT1}^T(y) \) represents the price paid for the regrettable fact that the policy \( y \) was not designed with the particular distribution \( f \) in mind; meanwhile, \( R_{fT2}^T(y) \) captures the additional “cost” due to the nonperishability of items.

We first establish a bound for \( R_T^f(y) \) when \( f \) is quite separated from \( \beta \). For a given \( f \in \Delta \), define \( \epsilon_f \equiv f(\bar{d}) \). Also, let
\[
\alpha_f \equiv \max \{ F_f(d) : F_f(d) < \beta \}, \quad \gamma_f \equiv \min \{ F_f(d) : F_f(d) > \beta \}. \tag{16}
\]
Then, define \( f \)'s separation from \( \beta \) by
\[
\delta_f \equiv (\beta - \alpha_f) \wedge (\gamma_f - \beta). \tag{17}
\]
Now given \( \epsilon, \delta > 0 \), we can define \( \Delta_{\epsilon, \delta} \subset \Delta \) so that
\[
\Delta_{\epsilon, \delta} \equiv \{ f \in \Delta : \epsilon_f \geq \epsilon, \delta_f \geq \delta \}. \tag{18}
\]
Every \( f \in \Delta_{\epsilon,\delta} \) has at least an \( \epsilon \) chance of having the top demand \( \overline{d} \) realized; more importantly, every \( f \in \Delta_{\epsilon,\delta} \) is at least “\( \delta \)-separated” from the parameter \( \beta \). We can achieve an upper bound for \( R_T^f(y) \) as long as \( f \) stays within \( \Delta_{\epsilon,\delta} \). Details are left to Section 2 of Zhou, Katehakis, and Yang [36].

**Theorem 1** Let \( y \) be the newsvendor-based adaptive policy. Then, for any \( \epsilon,\delta > 0 \), there is a positive constant \( A_{Them 1}^{\epsilon,\delta} \) so that

\[
\sup_{f \in \Delta_{\epsilon,\delta}} R_T^f(y) \leq A_{Them 1}^{\epsilon,\delta}.
\]

Theorem 1 shows that the regret \( R_T^f(y) \) is bounded from above by a constant independent of time \( T \), so long as there are known lower bounds on both \( \epsilon_f \equiv f(\overline{d}) \) and \( \delta_f \), the separation between \( f \) and \( \beta \). It poses as an extension of Besbes and Muharremoglu’s [6] Theorem 2 to the case involving items’ nonperishability. Between the two requirements in the result, the first one appears more reasonable as \( \overline{d} \) can always be the highest level that demand can ever reach. The second requirement, on the other hand, straddles between both demand distributions and cost parameters. It seems far-fetched to exclude a priori distributions \( f \) satisfying \( \delta_f \in (0,\delta) \) from consideration.

In view of the above, we feel compelled to derive a bound on \( R_T^f(y) \) that requires no prior knowledge on \( f \), let alone that about its relative positioning with respect to the cost parameter \( \beta \). Our derivation for the case where \( f \) can issue from anywhere in \( \Delta \) will rely on the convergence of the empirical distribution \( \hat{f}_{t-1} \) to the true distribution \( f \) as \( t \) tends to \( +\infty \). Let \( \delta_V(f,g) \) be the total variation between distributions \( f \) and \( g \); i.e.,

\[
\delta_V(f,g) \equiv \max_{K \subseteq \{0,1,\ldots,d\}} \left| \sum_{d \in K} f(d) - \sum_{d \in K} g(d) \right|,
\]

which also equals \( ||f - g||_1/2 \equiv \sum_{d=0}^d |f(d) - g(d)| / 2 \). It can be shown that

\[
P_f \left[ \delta_V(f,\hat{f}_{t-1}) \geq \epsilon \right] \leq 2 \cdot (\overline{d} + 1) \cdot \exp \left( -2\epsilon^2 \cdot (t - 1) \right).
\]

Indeed, let \( X_1,\ldots,X_n \) be independent random variables where each \( X_i \) is bounded by the interval \([a_i,b_i]\). Then for any \( \epsilon \geq 0 \), Theorem 2 of Hoeffding [23] stated that

\[
P \left[ \left| \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[\frac{1}{n} \sum_{i=1}^n X_i] \right| \geq \epsilon \right] \leq 2 \cdot \exp \left( -\frac{2n^2 \epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2} \right) \cdot (\overline{d} + 1) \cdot \exp \left( -2\epsilon^2 \cdot (t - 1) \right).
\]

According to \( \hat{f}_{t-1} \)’s definition through (10), \( \hat{f}_{t-1}(d) \) at any \( d \) is the \( (t - 1) \)-period mean of independent random variables \( 1(d_s = d) \), with the latter bounded by \([0,1]\) and with mean
f(d) under the probability \( \mathbb{P}_f \). Therefore, (21) will lead to

\[
\mathbb{P}_f[|\hat{f}_{t-1}(d) - f(d)| \geq \varepsilon] \leq 2 \cdot \exp(-2\varepsilon^2 \cdot (t-1)).
\]

By (19), this translates into

\[
\mathbb{P}_f[\delta_V(f, \hat{f}_{t-1}) \geq \varepsilon] \leq \mathbb{P}_f[\max_{d=0}^{\hat{d}} |f(d) - \hat{f}_{t-1}(d)| \geq \varepsilon] \leq \sum_{d=0}^{\hat{d}} \mathbb{P}_f[|f(d) - \hat{f}_{t-1}(d)| \geq \varepsilon],
\]

which, by (22), results in (20).

Without any prior knowledge on the \( f \in \Delta \), we can manage to obtain a \( T^{1/2} \cdot (\ln T)^{1/2} \)-sized bound on the \( R^{T1}_f(y) \) defined in (14). Due to (12), the key is to show that \( Q_f(y^*_f) - Q_f(y^*_f) \) will converge to 0 quickly. This will be achievable if we can show that \( Q_f(y^*_g) - Q_f(y^*_f) \) will be small when \( g \) and \( f \) are close by. This is when (20) and other properties related to the inventory management problem, such as the optimality of \( y^*_f \) to \( Q_f( \cdot ) \) and the linearity of \( Q_f(y) - Q_g(y) \) in the distance between \( f \) and \( g \), will be useful. The final form of the bound comes from the estimation of certain summations through integrations.

**Proposition 1** There are positive constants \( A^{Prop1} \) and \( B^{Prop1} \), so that

\[
R^{T1}_f(y) \leq A^{Prop1} + B^{Prop1} \cdot T^{1/2} \cdot (\ln T)^{1/2}.
\]

All remaining proofs of this section have been relegated to Appendix A. Next, we obtain a bound in the same order of magnitude for \( R^{T2}_f(y) \) as defined by (15). The process is much more involved than that for \( R^{T1}_f(y) \). From (11), we see that

\[
y_t = \hat{y}_t \lor (\hat{y}_{t-1} - d_{t-1}) \lor (\hat{y}_{t-2} - d_{t-2} - d_{t-1}) \lor \cdots \lor (\hat{y}_1 - d_1 - d_2 - \cdots - d_{t-1}).
\]

There is a latest \( s \) so that

\[
y_t = \hat{y}_s - d_s - d_{s+1} - \cdots - d_{t-1},
\]

which occurs exactly when either \( s = t \) or \( s \leq t - 1 \) and

\[
\hat{y}_s - d_s - d_{s+1} - \cdots - d_{t-1} - 1 \geq \hat{y}_t \lor (\hat{y}_{t-1} - d_{t-1}) \lor \cdots \lor (\hat{y}_{s+1} - d_{s+1} - \cdots - d_{t-1}),
\]

and regardless,

\[
\hat{y}_s \geq (\hat{y}_{s-1} - d_{s-1}) \lor (\hat{y}_{s-2} - d_{s-2} - d_{s-1}) \lor \cdots \lor (\hat{y}_1 - d_1 - d_2 - \cdots - d_{s-1}).
\]
Inspired by the above, we define random variables $I \geq 1$ and $S_1, S_2, \ldots, S_I, S_{I+1}$ in an iterative fashion as follows. First, let $S_1 = 1$. Now for some $i = 1, 2, \ldots$, suppose $S_i$ has been settled. Then, let $S_{i+1}$ be the first $t$ after $S_i$ so that

$$\hat{y}_t \geq \hat{y}_{S_i} - D_{S_i} - D_{S_i+1} - \cdots - D_{t-1},$$

(28)

if such a $t \leq T$ can be identified. If not, mark the latest $i$ as $I$ and let $S_{I+1} = T + 1$. For any $t$, let $L(t)$ be the largest $S_i \leq t$. This $L(t)$ can serve as the earlier $s$ satisfying (26) and (27) that corresponds to $t$. Note that $L(t)$ along with $D_{L(t)}, D_{L(t)+1}, \ldots, D_{t-1}$ are independent of $D_t$. So by (15), as well as (25) to (28),

$$R_f^{T2}(y) = \sum_{t=3}^{T} E_f[Q_f(\hat{y}_{L(t)} - D_{L(t)} - \cdots - D_{t-1}) - Q_f(\hat{y}_t)].$$

(29)

Now we are in a position to derive the bound.

**Proposition 2** There exist positive constants $A^{Prop2}$ and $B^{Prop2}$ so that

$$R_f^{T2}(y) \leq A^{Prop2} + B^{Prop2} \cdot T^{1/2} \cdot (\ln T)^{3/2}.$$

This is one of our most demanding results. It is also a building block for the analysis of the joint-control case involving nonperishable items. For its proof, we exploit the observations made from (24) to (29) to the fullest extent, with the basic understanding that the actual order-up-to level $y_t$ will be $\hat{y}_s - D_s - \cdots - D_{t-1}$ for some $s \leq t$. We are tasked to show that the term $Q_f(\hat{y}_s - D_s - \cdots - D_{t-1}) - Q_f(\hat{y}_t)$ can be bounded. For $\gamma = 1 - f(0)$, we divide the proof into two cases, the one with $\gamma \geq (1 - \beta)/2$ and the other one with $\gamma < (1 - \beta)/2$.

In the former large-$\gamma$ case, demand will accumulate over time with a guaranteed speed and $\hat{y}_t \geq \hat{y}_s - D_s - \cdots - D_{t-1}$ will occur ever more surely as $t - s$ increases. This is an effect similar to that achieved by Chen, Chao, and Ahn [14]'s assumption on the strict positivity of average demand levels. Then, for the minority case where $t - s$ is small, by exploiting natures of the empirical distribution and the newsvendor formula, we can come up with bounds related to $|Q_f(\hat{y}_s - D_s - \cdots - D_{t-1}) - Q_f(\hat{y}_t)| \cdot 1(\hat{y}_t \leq \hat{y}_s - D_s - \cdots - D_{t-1} - 1)$. Especially important is the observation that $\hat{y}_t \leq \hat{y}_s - D_s - \cdots - D_{t-1} - 1$ only if

$$\beta \leq \hat{F}_{t-1}(\hat{y}_t) \leq \hat{F}_{t-1}(\hat{y}_s - D_s - \cdots - D_{t-1} - 1) < \beta + \frac{t - s}{s}.$$  

(30)

We will end up with a trade-off already encountered in the proof of Proposition 1. This is the source of the $T^{1/2} \cdot (\ln T)^{3/2}$-sized growth rate. However, the constants will grow as $\gamma$ shrinks, because it takes ever longer for demand to accumulate.
Therefore, we seek a different approach for the latter small-$\gamma$ case, when $\gamma < (1 - \beta)/2 < 1 - \beta$. This is the time when $y_f^* = \hat{F}_f^{-1}(\beta) = 0$ because $F_f(0) = f(0) = 1 - \gamma > \beta$. We utilize the fact that $\hat{y}_s - D_s - \cdots - D_{t-1} \geq 1$ is the bare minimum for $\hat{y}_s - D_s - \cdots - D_{t-1} \geq \hat{y}_t + 1$. But the latter will be true only if both $\hat{y}_s \geq 1$ and for some $d = 1, 2, \ldots, \bar{d}$, both $\hat{y}_s \geq d$ and $D_s + \cdots + D_{t-1} \leq d - 1$. For all $\gamma$'s in the interval $[0, (1 - \beta)/2)$, we achieve a uniform bound in the order of $\ln T$, which is dominated by the one obtained in the first case.

Besides Hoeffding’s inequality, the proof also exploits Markov’s inequality in bounding $\mathbb{P}_f[\hat{y}_s \geq d]$, which, through (9) to (12), is the chance for the portion of earlier demand levels at or exceeding $d$ to be greater than $1 - \beta$. Our proof has not been helped by the fact that the $\hat{y}_t$’s as defined through (12) can be time-varying. Had it not been so, a simpler proof like that for Proposition 5 of Chen, Chao, and Shi [15] might have been achievable.

Combining Propositions 1 and 2, we get a bound for $R_T^f(y)$ that is not tangled up with how $f$ positions with $\beta$.

**Theorem 2** Let $y$ be the newsvendor-based adaptive policy. Then, there are positive constants $A_{Them}^2$ and $B_{Them}^2$ so that

$$\sup_{f \in \Delta} R_T^f(y) \leq A_{Them}^2 + B_{Them}^2 \cdot T^{1/2} \cdot (\ln T)^{3/2}.$$ 

The constants involved can depend on the problem’s parameters $\bar{h}$, $\bar{b}$, and $\bar{d}$. However, they are uniform across all $f$’s in $\Delta$. For the repeated newsvendor problem, Besbes and Muharremoglu [6] has already shown a $T^{1/2}$-sized lower bound (Lemma 4, with $\varepsilon$ replaced by $1/T^{1/2}$ in its (C-8)). The example used for the bound involves distributions $f$ with small separations from $\beta$ but in opposite directions. According to (11), the current case merely adds the restriction $y_t(d_{[1,t-1]}) \geq y_{t-1}(d_{[1,t-2]}) - d_{t-1}$ to the adaptive policy considered. So the lower bound can be no better.

In view of this, the above is almost the best one can hope for. Huh and Rusmevichientong’s [24] SA-based policy was shown to have a $T^{1/2}$-sized bound when items are continuous. The policy’s adaptation to the discrete-item setting, as to occur in (84) and (85), appears to be more complicated. Its full analysis awaits further research.

## 5 Joint Inventory-price Control

We now combine pricing with inventory control. For some finite integer $\bar{k} = 2, 3, \ldots$, suppose prices can be chosen from among the different $\bar{p}_1, \bar{p}_2, \ldots, \bar{p}_{\bar{k}}$, with all of them above the unit
production cost \(c\). Under any price \(p\) for \(k = 1, 2, \ldots, k\), demand will be randomly drawn from some \(f^k \equiv (f^k(0), f^k(1), \ldots, f^k(\bar{d}))\) in the simplex \(\Delta\) defined at (1). Let the firm immediately earn revenue \(p \cdot d\) when it charges price \(p\) in any period and demand realization happens to be \(d\). In the backlogging case, the total cost of dealing with the demanded units is certainly captured by \(\bar{c} \cdot \sum_{t=1}^{T} d_t + \sum_{t=1}^{T} q(y_t, d_t)\), with \(q(y, d)\) defined at (2).

This can also be true for a lost sales case slightly different from the more standard setting as studied in Chen, Chao, and Shi [15]. Here, we assume that any demand unit not satisfied within a period will cost the firm an extra \(\bar{b} + \bar{c} - p\) on top of the price not earned. Basically, we are assuming a fixed total lost sales cost of \(\bar{l}\); hence, it will be reasonable to assume a \(\bar{b}\) above the highest gross profit margin \(p - \bar{c}\). In the standard setting, however, the extra cost on top of the lost revenue is assumed to be independent of the price charged.

For either the backlogging or nontraditional lost sales case, the firm’s \(T\)-period profit is

\[
\sum_{t=1}^{T} p_t \cdot d_t - \bar{c} \cdot \sum_{t=1}^{T} d_t - \bar{h} \cdot \sum_{t=1}^{T} (y_t - d_t)^+ - \bar{b} \cdot \sum_{t=1}^{T} (d_t - y_t)^+.
\]

That is, it is equal to \(\sum_{t=1}^{T} v(p_t, y_t, d_t)\), where

\[
v(p, y, d) \equiv (p - \bar{c}) \cdot d - q(y, d) = (p - \bar{c}) \cdot d - \bar{h} \cdot (y - d)^+ - \bar{b} \cdot (d - y)^+,
\]

with \(q(y, d)\) as defined at (2). Here, \(v(p, y, d)\) can be understood as the profit the firm can make in one single period when it charges price \(p\) and orders \(y\) items, and the realized demand is \(d\). As before, whether or not items are nonperishable depends on (i) whether the constant \(\bar{h}\) is treated merely as the holding cost rate or the difference between unit production cost \(\bar{c}\) and unit salvage value \(\bar{s}\) and (ii) whether or not (3) is enforced.

Now, let \(V_f(p, y)\) be the average of the profit \(v(p, y, d)\) defined at (32) that the firm can make when it faces demand distribution \(f\):

\[
V_f(p, y) \equiv \mathbb{E}_f[v(p, y, D)] = (p - \bar{c}) \cdot \mathbb{E}_f[D] - Q_f(y),
\]

where \(Q_f(y)\) is defined at (4). For any price choice \(k\) and demand distribution \(f\), let

\[
V^k_f \equiv \max_{y=0,1,\ldots,\bar{d}} V_f(p^k, y) = V_f(p^k, y^*_f),
\]

be the most that the firm can earn while charging price \(p^k\) and facing demand distribution \(f\). In (34), the best order level \(y^*_f\) is given in (9). We have so far not been concerned with the price-demand relationship, namely, the vector \(\mathbf{f} \in \Delta^k\) in our current setting.
When presented with demand-distribution vector \( \mathbf{f} \equiv (f^1, f^2, ..., f^k) \), let
\[
V^*_{\mathbf{f}} \equiv \max_{k=1}^{\bar{k}} V^k_{f^k},
\]
be the maximum average profit that the firm can earn in one period when it uses the best price; meanwhile, each price \( \bar{p}^k \) corresponds to demand distribution \( f^k \). Suppose \( k^*_{\mathbf{f}} \) solves (35); namely, \( V^*_{\mathbf{f}} = V^{k^*_{\mathbf{f}}} f^{k^*_{\mathbf{f}}} \). Then, when facing a \( T \)-period horizon with a known demand-distribution vector \( \mathbf{f} \), an optimal policy will be to repeatedly charge the price \( \bar{p}^k \) and order up to \( y^{k^*_{\mathbf{f}}} \). Thus, the maximum profit over \( T \) periods is \( V^*_{\mathbf{f}} \cdot T = V^{k^*_{\mathbf{f}}} f^{k^*_{\mathbf{f}}} \cdot T = V^{k^*_{\mathbf{f}}} \left( \bar{p}^{k^*_{\mathbf{f}}}, y^{k^*_{\mathbf{f}}} \right) \cdot T \).

When \( \mathbf{f} \in \Delta^k \) is unknown, we again seek a good adaptive policy. Such a policy \((k, y) \equiv (k_1, y_1, k_2, y_2, ...)\) satisfies that, for \( t = 1, 2, ... \), each price choice \( k_t = 1, 2, ..., k \) is a function of the historical demand vector \( \mathbf{d}_{[1,t-1]} \), and so is each order level \( y_t = 0, 1, ..., \bar{d} \). Under it, the \( T \)-period total average profit is
\[
V^T_{\mathbf{f}}(k, y) \equiv \sum_{t=1}^{T} \mathbb{E}_f \left[ V^{k_t}_{f^{k_t}}(\bar{p}^{k_t}, y_t) \right].
\]
Note the average profit \( V^{k_t}_{f^{k_t}}(\bar{p}^{k_t}, y_t) \) as defined in (33) can be used for each period \( t \) because, given price choice \( k_t \), the demand in that period is independent of earlier demands which determine the pricing and ordering decisions \( k_t \) and \( y_t \). Now define \( T \)-period regret \( R^T_{\mathbf{f}}(k, y) \) of using the adaptive policy \((k, y)\) under demand-distribution vector \( \mathbf{f} \):
\[
R^T_{\mathbf{f}}(k, y) \equiv V^*_{\mathbf{f}} \cdot T - V^T_{\mathbf{f}}(k, y).
\]
We aim to identify adaptive policies \((k, y)\) that prevent \( R^T_{\mathbf{f}}(k, y) \) from growing too fast in \( T \) for “most” or even all \( \mathbf{f} \)'s within \( \Delta^k \). The quest is in a sense harder than the multi-armed bandit problem because we do not get to generate samples whose average makes up each \( V^k_{f^k} \). By (34), the latter is itself the product of an optimization problem.

A good policy should test each price \( \bar{p}^k \) often enough to learn the corresponding distribution \( f^k \) well; yet, it should not dwell on the price for too long if \( V^k_{f^k} \) is strictly below \( V^*_{\mathbf{f}} \). As the \( f^k \)'s are unknown, substitutes that are acquirable from past experience can be used in their stead. Since inventory-related errors already amount to the order of \( t^{1/2} \), we can use a term roughly proportional to \( t^\mu \) at some \( \mu \in [1/2, 1) \) as the guaranteed number that any price \( \bar{p}^k \) will have been visited by time \( t \). At the same time, we should limit the visits to \( \bar{p}^k \) when \( V^k_{f^k} \), as approximated from its surrogate, does not seem promising.

The above inspires our LwD(\( \mu \) \) (learning while doing at parameter \( \mu \)) policy at \( \mu \in
[1/2, 1). It keeps track of the number of times price \( p^k \) is charged in periods 1 through \( t \):

\[
\mathcal{N}_t^k \equiv \sum_{s=1}^{t} 1(p_s = p^k). \tag{38}
\]

The policy also designates the mode of each period \( t = 1, 2, \ldots \) as either learning, with \( m_t = 0 \) or doing, with \( m_t = 1 \). It is certainly true that

\[
\mathcal{N}_t^k = \mathcal{N}_{t,0}^k + \mathcal{N}_{t,1}^k, \tag{39}
\]

with

\[
\mathcal{N}_{t,m}^k \equiv \sum_{s=1}^{t} 1(m_s = m \text{ and } p_s = p^k), \quad \forall m = 0, 1. \tag{40}
\]

In addition, the policy keeps track of profits achievable by various prices under their respective empirical distributions. That is, for \( k = 1, 2, \ldots, \bar{k} \), the policy is aware of the value \( V_{t-1}^k \) as given by (34) as long as \( \mathcal{N}_{t-1}^k \geq 1 \), while \( \hat{f}_{k-1}^t \in \Delta \) stands for the empirical distribution of demand under price \( \hat{p}^k \) observed over the past \( t-1 \) periods: for \( d = 0, 1, \ldots, \bar{d} \),

\[
\hat{f}_{k-1}^t(d) = \frac{\sum_{s=1}^{t-1} 1(p_s = \hat{p}^k \text{ and } d_s = d)}{\mathcal{N}_{t-1}^k}. \tag{41}
\]

Note that the empirical distribution \( \hat{f}_{k-1}^t \) is knowledge gained over \( \mathcal{N}_{t-1}^k \) rather than \( t-1 \) periods. Thus, it is important that none of the \( \mathcal{N}_{t-1}^k \)'s, which together satisfy \( \sum_{k'=1}^{\bar{k}} \mathcal{N}_{t-1}^{k'} = t-1 \), lags too much behind \( t-1 \).

Initially, the policy lets \( \mathcal{N}_{0,0}^k = \mathcal{N}_{0,1}^k = \mathcal{N}_0^k = 0 \) for \( k = 1, 2, \ldots, \bar{k} \). Then, in every period \( t = 1, 2, \ldots \), suppose \( \kappa_{t-1}(1) \) is a \( k \) that minimizes the number \( \mathcal{N}_{t-1}^k \). If \( \mathcal{N}_{t-1}^{\kappa_{t-1}(1)} < (t/\bar{k})^\mu \), the policy will recommend the following:

0.1. set the mode of period \( t \) as learning, with \( m_t = 0 \);
0.2. also, let price choice \( k_t = \kappa_{t-1}(1) \);
0.3. next, update the book in the fashion of

\[
\mathcal{N}_{t,0}^{k_t} = \mathcal{N}_{t-1,0}^{k_t} + 1, \quad \mathcal{N}_{t,1}^{k_t} = \mathcal{N}_{t-1,1}^{k_t}, \quad \mathcal{N}_t^{k_t} = \mathcal{N}_{t-1}^{k_t} + 1. \tag{42}
\]

Otherwise, with \( \mathcal{N}_{t-1}^{\kappa_{t-1}(1)} \geq (t/\bar{k})^\mu \) which necessitates that \( \mathcal{N}_{t-1}^k \geq 1 \) at every \( k = 1, 2, \ldots, \bar{k} \), the policy will recommend the following:

1.1. set the mode of period \( t \) as doing, with \( m_t = 1 \);
1.2. also, let price choice \( k_t \) be a maximizer of \( V_{t-1}^k \) from \( k = 1, 2, \ldots, \bar{k} \);
1.3. next, update the book in the fashion of

\[
\mathcal{N}_{t,0}^{k_t} = \mathcal{N}_{t-1,0}^{k_t}, \quad \mathcal{N}_{t,1}^{k_t} = \mathcal{N}_{t-1,1}^{k_t} + 1, \quad \mathcal{N}_t^{k_t} = \mathcal{N}_{t-1}^{k_t} + 1. \tag{43}
\]
Finally, for those $k$’s unequal to $k_t$, we keep

$$
N_{t,0}^k = N_{t-1,0}^k, \quad N_{t,1}^k = N_{t-1,1}^k, \quad N_t^k = N_{t-1}^k. \quad (44)
$$

After pricing choice $k_t$ has been settled, the policy lets the firm charge the price $\bar{p}^k$. When $t = 1$, the policy has ordering facilitated through $y_1 = \hat{y}_1 = 0$. For $t = 2, 3, \ldots$, it advises on following (11) and in the place of (12),

$$
\hat{y}_t = \hat{y}_{t-1}^* = \min \left\{ d = 0, 1, \ldots, \bar{d} : \sum_{s=1}^{t-1} 1(\tilde{p}_s = \bar{p}^k \text{ and } d_s \geq d) \geq \beta \cdot N_{t-1}^k \right\}, \quad (45)
$$

when $N_{t-1}^k \geq 1$ and $\hat{y}_t = 0$ when $N_{t-1}^k = 0$.

In the policy, it is easy to see that the updatings (42) to (44) will ensure the satisfaction of (38) to (40) by the $N_{t,0}^k$’s, $N_{t,1}^k$’s, $N_t^k$’s, and $m_t$ in every period $t = 1, 2, \ldots$. In that exploration is done at controlled paces and exploitation is intended for the maximization of profit per period, the current LwD($\mu$) bears some resemblance to the pricing policy proposed in Section 4 of Burnetas and Smith [13]. However, our pacing using the $(t/\bar{k})^\mu$-function is different; it leads to Propositions 3 and 4 which are essential for our regret analysis. In addition, while we compare the potential profits under observed empirical distributions $V_{t-1}^k$ in the exploitation step 1.2, the earlier work compared the average profits truly experienced:

$$
\hat{V}_t^k = \frac{\sum_{s=1}^{t-1} 1(p_s = \bar{p}^k) \cdot v(\bar{p}^k, y_s, d_s)}{N_t^k}, \quad (46)
$$

where $v(p, y, d)$ is defined at (32). Our choice is realizable in the current discrete-item setting and being less affected by earlier errors, could encourage faster convergence.

Because $\mu < 1$, there exists the smallest $j = 2, 3, \ldots$ such that $j \geq (j + 1/\bar{k})^\mu$. Also, it will happen that $N_{n_k}^k = n$ for $n = 1, \ldots, j$ and $k = 1, \ldots, \bar{k}$. Basically, there are initially a $j$ number of $\bar{k}$-long cycles in each of which all prices are tried once in the learning mode.

### 6 Joint-control Bounds when Items are Perishable

For implementation and analysis purposes, it is actually beneficial to keep track of not only a minimizer of $N_t^k$ for $t = 0, 1, \ldots$, but also an entire sequence in the ascending order. Thus, let $\kappa_t \equiv (\kappa_t(1), \kappa_t(2), \ldots, \kappa_t(\bar{k}))$ be a permutation of the numbers $1, 2, \ldots, \bar{k}$ such that

$$
N_t^{\kappa_t(1)} \leq N_t^{\kappa_t(2)} \leq \cdots \leq N_t^{\kappa_t(\bar{k})}. \quad (47)
$$

The policy can maintain such a $\kappa_t$ for $t = 0, 1, \ldots$. Details are left to Section 3 of Zhou, Katehakis, and Yang [36]. We now make two important observations about LwD($\mu$).
Proposition 3 For any period \( t = 1, 2, \ldots \), it is true that

\[
\mathcal{N}_{t,0}^k < \left( \frac{t}{k} \right)^\mu + 1, \quad \forall k = 1, 2, \ldots, \bar{k}.
\]

Proposition 4 For any period \( t = 1, 2, \ldots \), it is true that

\[
\mathcal{N}_{t-1}^k \geq \left( \frac{t}{k} \right)^\mu - 1, \quad \forall k = 1, 2, \ldots, \bar{k}.
\]

All proofs of this section have been relegated to Appendix B. While Proposition 3 gives an upper bound on the time spent on pure learning, Proposition 4 gives a guarantee on the amount of time each price choice \( k \) will be experienced. For the regret of any policy \((k, y)\),

\[
R_t^T(k, y) = V^*_{T} \cdot T - V_t^T(k, y) = \sum_{t=1}^{T} (V_{f^k_{t}}^k - \mathbb{E}_t[V_{f^k_{t}}(\hat{y}_{t}, y_t)])
\]

\[
= \sum_{t=1}^{T} \mathbb{E}_t[V_{f^k_{t}}^k - V_{f^k_{t}}^k] + \sum_{t=1}^{T} \mathbb{E}_t[V_{f^k_{t}}^k - V_{f^k_{t}}(\hat{y}_{t}, y_t)],
\]

where the first equality is from (37), the second equality is from (35), (36), and the \( V_{f^k_{t}}^k \) - maximizing nature of \( k_\star^k \), while the third equality is just an identity.

Suppose furthermore that, the policy \((k, y)\) represents LwD(\( \mu \)). Then, due to (11), (33), and (45), we can rewrite (48) into something similar to (13):

\[
R_t^T(k, y) = R_t^{T_1}(k, y) + R_t^{T_2}(k, y),
\]

where

\[
R_t^{T_1}(k, y) = \sum_{t=1}^{T} \mathbb{E}_t \left[ V_{f^k_{t}}^k - V_{f^k_{t}}^k \right] + \sum_{t=1}^{T} \mathbb{E}_t \left[ V_{f^k_{t}}^k - V_{f^k_{t}}(\hat{y}_{t}, y_{\hat{y}_{t-1}}) \right],
\]

in which \( y_{\hat{y}_{t-1}} \) should be understood as 0 when \( \mathcal{N}_{t-1}^{k_{t-1}} = 0 \), and

\[
R_t^{T_2}(k, y) = \sum_{k=1}^{k} \sum_{t=1}^{T} \mathbb{E}_t \left[ 1(k_t = k) \cdot (Q_{f^k}(y_t) - Q_{f^k}(\hat{y}_t)) \right],
\]

with \( y_t \) and \( \hat{y}_t \) provided by (11) and (45). In (49), the first term \( R_t^{T_1}(k, y) \) stands for the regret that an LwD(\( \mu \)) policy will accrue if items are allowed to perish at the end of each period; the second term \( R_t^{T_2}(k, y) \) captures additional regret due to inventory carry-over.

We first bound the first term \( R_t^{T_1}(k, y) \) given in (50). Define

\[
\delta V_t^k \equiv V^*_{T} - V_{f^k_{t}}^k = V_{f^k_{t}}^k - V_{f^k_{t}}^k \geq 0.
\]

It measures the difference in average single-period profit between using price choice \( k \) and making the best choice \( k_\star^k \). Now due to the nature of the LwD(\( \mu \)) policies, (50) will become

\[
R_t^{T_1}(k, y) = T_1 + T_2 + T_3,
\]

(53)
with
\begin{align}
T_1 &= \sum_{k \neq k^*} \delta V^k_f \cdot \mathbb{E}_f[N^k_{T,0}], \\
T_2 &= \sum_{k \neq k^*} \delta V^k_f \cdot \sum_{t=1}^T \mathbb{P}_f \left[ m_t = 1 \text{ and } V^k_{f_{j_{t-1}}} \text{ achieves the maximum} \right], \\
T_3 &= \sum_{k=1}^k \sum_{t=1}^T \mathbb{E}_f \left[ 1(k_t = k) \cdot \left( V^k_{f_j} - V^k_{f_{j_{t-1}}} \left( p^k, y^*_f \right) \right) \right].
\end{align}

where \( \delta V^k_f \) has been defined at (52) and \( m_t = 0 \) or \( 1 \) stands for learning or doing, and

\begin{align}
\delta V^*_f \equiv \min_{k \neq k^*} \delta V^k_f,
\end{align}

be the minimum gap in single-period profits between optimal and non-optimal price choices.

For any \( \delta > 0 \), we use \( \Delta^k_{\delta} \) to denote the subset of \( f \)'s in \( \Delta^k \) that have one price leading other choices by at least a \( \delta \)-margin:
\begin{align}
\Delta^k_{\delta} \equiv \left\{ f \in \Delta^k : k^*_f \text{ is unique and } \delta V^*_f \geq \delta \right\} \subset \Delta^k.
\end{align}

The performance of LwD(\( \mu \)) can be well bounded when \( f \) is known to come from \( \Delta^k_{\delta} \).

**Proposition 5** Let \((k, y)\) be the adaptive policy generated from following LwD(\( \mu \)) for some \( \mu \in [1/2, 1] \). Then, for any \( \delta > 0 \), there are constants \( A^{\text{Prop5}}_\delta \), \( B^{\text{Prop5}} \), and \( C^{\text{Prop5}} \), so that
\begin{align}
R^{T_1}_f(y, k) \leq A^{\text{Prop5}}_\delta + B^{\text{Prop5}} \cdot T^\mu + C^{\text{Prop5}} \cdot T^{1/2} \cdot (\ln T)^{1/2},
\end{align}
for any \( f \in \Delta^k_{\delta} \); however, with \( \lim_{\delta \rightarrow 0^+} A^{\text{Prop5}}_\delta = +\infty \). Also, \( \mu = 1/2 \) is the choice with the tightest guarantee among the LwD(\( \mu \)) policies. It will achieve an \( O(T^{1/2} \cdot (\ln T)^{1/2}) \)-bound.
Bounding will be more demanding when $f$ is truly free to roam in $\Delta^k$. After a detailed analysis which relies on the proof of Proposition 5, we can arrive at the following bound.

**Proposition 6** Let $(k, y)$ be the adaptive policy generated from following LwD($\mu$) for some $\mu \in [1/2, 1)$. Then, there are positive constants $A_{Prop6}$ and $B_{Prop6}$ so that

$$\sup_{f \in \Delta} R_T^1(k, y) \leq A_{Prop6} + B_{Prop6} \cdot T^{\mu/(1-\mu/2)}.$$  

The choice $\mu = 2/3$ will achieve an $O(T^{2/3})$-bound.

Comparing Propositions 5 and 6, one might say that those incidences $f$ residing in $\Delta^k \setminus \Delta^k_\delta$ for ever smaller $\delta$’s are “trouble makers” that render the optimal price illusory to catch. Earlier, between Theorems 1 and 2, those $f$’s in $\Delta \setminus \Delta_{\epsilon, \delta}$ for ever smaller $\epsilon$’s and $\delta$’s are the culprits that make ordering difficult. However, this time, the most suitable policy will itself have to adapt: it is LwD(1/2) when a bottom to the $\delta V^k_f$’s is known to be there; whereas, it is LwD(2/3) when no such a bottom is known.

7 Nonperishability with Restricted Demand Patterns

We first bound $R_T^2(k, y)$ given in (51) when $f$ is restricted to $\Delta^k_\delta$ for some $\delta > 0$. Our plan is to show that the dominant price say $\bar{p}^1$ will be used uninterruptedly for long sequences of periods; meanwhile, the sub-linear growth of nonperishability-induced regrets in sequences’ lengths, as evidenced by Proposition 2, will help to bound the $T$-period regret overall.

To this end, let $r_f^t(x_1)$ be almost the same nonperishability-induced pure-control regret under demand distribution $f$ as the $R_T^2(y)$ defined at (15). But this time, we let it be from period 1 to a variable period $t$, and let the starting inventory level be some arbitrary $x_1 = 0, 1, ..., \bar{d}$. Proposition 2 can be understood as a bound for $r_f^t(0)$. But a close scrutiny of it would reveal that a bound of the same form works for $r_f^t(x_1)$ regardless of the valuation of the initial $x_1$. Actually, the only changes needed in the proof would be to replace “$t = 3$” with “$t = 1$” and “$T - 2$” with “$T$”. Hence, for the positive constants $A_{Prop2}$ and $B_{Prop2},$

$$r_f^t(x_1) \leq A_{Prop2} + B_{Prop2} \cdot t^{1/2} \cdot (\ln t)^{3/2},$$  \hspace{1cm} (59)

for any $t = 1, 2, ..., f \in \Delta$, and $x_1 = 0, 1, ..., \bar{d}$.

We also find it convenient to condition on when learning has happened. So for any $t = 1, 2, ...,$, let $\mathcal{M}(t) \equiv \{0, 1\}^t$ be the set of all potential learning/doing-mode sequences
m \equiv (m_1, \ldots, m_t) \text{ over the first } t \text{ periods. Given } m \equiv (m_1, \ldots, m_t) \in \mathcal{M}(t), \text{ we can use } 
abla_{t,0}(m) = \sum_{s=1}^{t}(1-m_s) \text{ to denote the total number of learning periods under mode sequence } m \text{ of the first } t \text{ periods. Due to Proposition 3,}
\begin{equation}
\nabla_{t,0}(m) \leq \bar{k} \cdot \left( \left( \frac{t}{k} \right)^\mu + 1 \right) = \bar{k}^{1-\mu} \cdot t^\mu + \bar{k}, \quad \forall m \in \mathcal{M}(t).
\end{equation}

Let \( s_1(m), s_2(m), \ldots, s_{\nabla_{t,0}(m)}(m) \) be the periods at which learning takes place. Certainly,
\begin{equation}
s_i(m) = \min \left\{ s = 1, \ldots, t : \sum_{\tau=1}^{s}(1-m_\tau) \geq i \right\}, \quad \forall i = 1, \ldots, \nabla_{t,0}(m),
\end{equation}
and hence
\begin{equation}
\nabla_{t,0}(m) = \max\{i = 1, 2, \ldots : s_i(m) \leq t\}.
\end{equation}

For convenience, use \( M(t) \) for the random mode sequence that has actually been realized. Note that \( \sum_{m \in \mathcal{M}(t)} \mathbb{P}_f[M(t) = m] = 1 \). Without loss of generality, designate choice 1 as \( k^*_1 \).
Conditioned on \( M(t) = m \), we can show two properties. First, the chance for \( V^{k}_{f_t} \) to be greater than any other \( V^{k}_{f_t} \) plus a margin say \( \delta V^*_t / 2 \) will be ever closer to one as \( t \) increases. Second, given that \( t \) is like that and given that there is no interruption from learning, the status of \( V^{k}_{f_t} > V^{k}_{f_t} \) can be kept for a long time, for \( \tau = t, t+1, \ldots \).

**Proposition 7** There exist positive constants \( A^{Prop7} \) and \( E^{Prop7} \), such that
\begin{equation}
\mathbb{P}_f \left[ V_{f_t} \geq \max_{k=2}^{k} V^{k}_{f_t} + \delta V^*_t / 2 | M(t) = m \right] \geq 1 - A^{Prop7} \cdot \exp \left( -E^{Prop7} \cdot (\delta V^*_t)^2 \cdot t^\mu \right),
\end{equation}
regardless of the mode sequence \( m \in \mathcal{M}(t) \) being realized for \( M(t) \).

Proofs of this section can be found in Appendix C. Towards the second point, let \( E_{t'} \) be the event that \( V^{k}_{f_t} > \max_{k=2}^{k} V^{k}_{f_t} + \delta V^*_t / 2 \) and there is no interruption from learning in periods \( t, t+1, \ldots, t+t' - 1 \).

**Proposition 8** There is a strictly positive constant \( A^{Prop8} \), say \( 1/(4d^2 \cdot (p^k - c + 3 \cdot (h \vee d))) \), so that for \( t' \) as large as \( A^{Prop8} \cdot \delta V^*_t \cdot (t/\bar{k})^\mu - 1 \), one will expect \( k_t = k_{t+1} = \cdots = k_{t+t'-1} = 1 \) whenever \( E_{t'} \) is true.

Proposition 7 states that the future profit indicator of price choice \( k^*_1 = 1 \) will lead those of the other prices by a comfortable margin with a probability that converges to 1 quickly as \( t \) grows; also, Proposition 8 predicts that once leading by a comfortable margin, price choice
1 will be revisited for a long time so long as it is not interrupted by learning. So far we have upper bounds on learning frequencies in the forms of Proposition 3 and (60). To utilize Propositions 7 and 8, it will help to have lower bounds as well. To this end, we introduce virtual learning periods that will allow the frequencies of learning, be it actual or virtual, to be regulated on both sides. Details now follow.

Let \( G_{\mu,\delta} \) be a constant strictly above \( (4 \cdot \bar{k}^{\mu})/(A_{Prop}^{8} \cdot \delta) \), where \( A_{Prop}^{8} \) is the constant in Proposition 8 and \( \delta \) is the margin that appeared in (58); for instance,

\[
G_{\mu,\delta} = \frac{4 \cdot \bar{k}^{\mu}}{A_{Prop}^{8} \cdot \delta} + 1. \tag{63}
\]

Also, let \( I_{\mu,\delta} \geq \bar{k} \) be large enough so that both \( [(1 + I_{\mu,\delta})^{1/\mu}/G_{\mu,\delta}^{1/\mu}] \geq 1 \) and \( [(2 + I_{\mu,\delta})^{1/\mu}/G_{\mu,\delta}^{1/\mu}] - [(1 + I_{\mu,\delta})^{1/\mu}/G_{\mu,\delta}^{1/\mu}] \geq 2 \). Note that both \( G_{\mu,\delta} \) and \( I_{\mu,\delta} \) will grow to \(+\infty\) when \( \delta \) approaches \( 0^{+} \). Now define \( s'_{i} = [(i + I_{\mu,\delta})^{1/\mu}/G_{\mu,\delta}^{1/\mu}] \) for \( i = 1, 2, \ldots \) as virtual learning periods. We have omitted the dependence of the \( s'_{i} \)'s on \((\mu, \delta)\) for simplicity. The size of \( I_{\mu,\delta} \) will guarantee that \( 1 \leq s'_{1} < s'_{2} < \cdots \). Similarly to (62), let

\[
N'_{i,0} = \max\{i = 1, 2, \ldots : s'_{i} \leq t\}. \tag{64}
\]

Let \( L'(t) = \{s'_{1}, s'_{2}, \ldots, s'_{N'_{i,0}}\} \) be the set of virtual learning periods up to \( t \).

For \( m \in \mathcal{M}(t) \), we could let \( L(m, t) = \{s_{1}(m), s_{2}(m), \ldots, s_{N'_{i,0}(m)}(m)\} \) be the set of actual learning periods up to \( t \). A virtual learning period could be either an actual learning or actual doing period. Now consider the combined set \( L''(m, t) = L(m, t) \cup L'(t) \). We can write \( L''(m, t) \) as \( \{s''_{1}(m), s''_{2}(m), \ldots, s''_{N'_{i,0}(m)}(m)\} \) with \( 1 \leq s''_{1}(m) < s''_{2}(m) < \cdots < s''_{N'_{i,0}(m)}(m) \leq t \) and each \( s''_{i}(m) \) being either some \( s_{j}(m) \) or some \( s'_{i} \) or both. The frequencies at which the combined learning periods \( s''_{i}(m) \) arise are constrained both ways.

**Proposition 9** We have the following useful inequalities:

\[
N''_{i,0}(m) \leq H_{\mu,\delta} \cdot t^{\mu}, \tag{65}
\]

for some constant \( H_{\mu,\delta} \) which is above \( G_{\mu,\delta} \) and hence in satisfaction of \( \lim_{\delta \to 0^{+}} H_{\mu,\delta} = +\infty \);

\[
s''_{i}(m) \geq \left( \frac{1}{H_{\mu,\delta}^{1/\mu}} \right) \cdot i^{1/\mu}; \tag{66}
\]

\[
s''_{i}(m) \leq \left( \frac{1}{G_{\mu,\delta}^{1/\mu}} \right) \cdot (i + I_{\mu,\delta})^{1/\mu} + 1; \tag{67}
\]

\[
s''_{i+1}(m) - s''_{i}(m) \leq \left[ \left( \frac{4}{G_{\mu,\delta}^{1/\mu} \cdot (s''_{i}(m))^{1-\mu} + 1} \right) \vee \left( \frac{(1 + I_{\mu,\delta})^{1/\mu}}{G_{\mu,\delta}^{1/\mu}} \right) \right]. \tag{68}
\]
Note that (65) and (66) bound combined learning frequencies from above, while (67) and (68) bound them from below.

We now utilize (59) and Propositions 7 to 9 to bound $R^T_f(k, y)$ when $(k, y)$ comes from the LwD($\mu$) policy and $f \in \Delta^k$. For convenience, let $s''_{N_T,0}(m) = T + 1$. Now for any $i = 1, 2, ..., N_T(m)$, let $N_i(m)$ be the random number of consecutive same-price doing sequences from period $s''_i(m) + 1$ to period $s''_{i+1}(m) - 1$. We do allow $N_i(m) = 0$ when $s''_{i+1}(m) = s''_i(m) + 1$; otherwise, $N_i(m)$ is integer-valued between 1 and $s''_{i+1}(m) - s''_i(m) - 1$. For instance, when $s''_i(m) = 10$, $s''_{i+1}(m) = 16$, price choice 1 is used in periods 11 and 12, price choice 2 is used in periods 13 and 14, and price choice 1 is used again in period 15, then $N_i(m)$ would be 3 because there have been three same-price consecutive sequences in periods 11 to 15. Define $U_{i,1}(m), ..., U_{i,N_i(m)+1}(m)$ so that

$$s''_i(m) + 1 = U_{i,1}(m) < U_{i,2}(m) < \cdots < U_{i,N_i(m)}(m) < U_{i,N_i(m)+1}(m) = s''_{i+1}(m),$$

(69)

and for each $j = 1, ..., N_i(m)$, periods $U_{i,j}(m), U_{i,j}(m) + 1, ..., U_{i,j+1}(m) - 1$ form the $(i, j)$-segment, i.e., a consecutive same-price doing sequence. For the segment, denote the price choice by $K_{i,j}(m) \equiv k_{U_{i,j}(m)} = k_{U_{i,j}(m)+1} = \cdots = k_{U_{i,j+1}(m)} - 1$ and the starting inventory level by $X_{i,j}(m)$. For the previous example, $U_{i,1}(m) = 11$, $U_{i,2}(m) = 13$, $U_{i,3}(m) = 15$, and $U_{i,4}(m) = 16$; also, $K_{i,1}(m) = 1$, $K_{i,2}(m) = 2$, and $K_{i,3}(m) = 1$.

Now (51) can be rewritten as

$$R^T_f(k, y) = T_1 + T_2,$$

(70)

where

$$T_1 = \sum_{m \in M(T)} \mathbb{P}_f[M(T) = m] \cdot \mathbb{E}_f \left[ \sum_{i=1}^{N_T,0(m)} r_{K_{i,j}(m)}(X_{i,j}(m)) \right],$$

(71)

$$T_2 = \sum_{m \in M(T)} \mathbb{P}_f[M(T) = m] \cdot \theta_2(m),$$

(72)

and

$$\theta_2(m) = \mathbb{E}_f \left[ \sum_{i=1}^{N_T,0(m)} \frac{N_i(m)}{m} \sum_{j=1}^{U_{i,j+1}(m) - U_{i,j}(m)} r_{K_{i,j}(m)}(X_{i,j}(m)) \right].$$

(73)

In (70), the nonperishability-induced regret is partitioned into two parts, the part $T_1$ that is accrued over combined-learning periods and the part $T_2$ that is accrued over doing periods that are not even virtual learning ones. In both (71) and (72), we sum over all potential $m \in M(t)$ that could be realized for the random $M(t)$. The way the term $\theta_2(m)$ is expressed in (73) captures price-switching epochs.
A $T^\mu$-sized bound can be easily identified for $T_1$ due to (65). From (73), we also have
\[
\theta_2(m) = \eta_2(m) + \zeta_2(m),
\]
where
\[
\eta_2(m) = \sum_{i=1}^{N_{\mu,0}(m)} P_r \left[ V_{f_{\mu,0}(m)+1} \leq \max_{k=2} V_{f_{\mu,0}(m)+1}^k + \delta V^*/2 | M(T) = m \right] \times \mathbb{E}_r \left[ \sum_{j=1}^{N_i(m)} r^{U_{i,j+1}(m)-U_{i,j}(m)}(X_{i,j}(m)) | V_{f_{\mu,0}(m)+1} \leq \max_{k=2} V_{f_{\mu,0}(m)+1}^k + \delta V^*/2 \right],
\]
and
\[
\zeta_2(m) = \sum_{i=1}^{N_{\mu,0}(m)} P_r \left[ V_{f_{\mu,0}(m)+1}^k > \max_{k=2} V_{f_{\mu,0}(m)+1}^k + \delta V^*/2 | M(T) = m \right] \times \mathbb{E}_r \left[ \sum_{j=1}^{N_i(m)} r^{U_{i,j+1}(m)-U_{i,j}(m)}(X_{i,j}(m)) | V_{f_{\mu,0}(m)+1} > \max_{k=2} V_{f_{\mu,0}(m)+1}^k + \delta V^*/2 \right].
\]

But by Proposition 7, the probabilities within (74) are small. This point will eventually lead to a constant bound for $\eta_2(m)$. By Proposition 8 and some bounds in Proposition 9, we can establish that $N_i(m) = 1$ under the $m$’s specified in (75). We can then utilize (59) and other bounds in Proposition 9 to bound $\zeta_2(m)$ by a $T^{(1+\mu)/2}$-sized term.

**Proposition 10** It turns out that $T_1$ of (71) has a $T^\mu$-sized bound and $T_2$ of (72) has a $T^{(1+\mu)/2}$-sized bound. Overall, there are positive constants $A_{\text{Prop}10}^{\mu,\delta}$, $B_{\text{Prop}10}^{\mu,\delta}$, and $C_{\text{Prop}10}^{\mu,\delta}$, so that for the LwD($\mu$) policy $(k, y)$ and any $f \in \Delta_{\delta}$,
\[
R_f^{T_2}(k, y) \leq A_{\text{Prop}10}^{\mu,\delta} + B_{\text{Prop}10}^{\mu,\delta} \cdot T^\mu + C_{\text{Prop}10}^{\mu,\delta} \cdot T^{(1+\mu)/2} \cdot (\ln T)^{5/2},
\]
however, $\lim_{\delta \to 0^+} B_{\text{Prop}10}^{\mu,\delta} = +\infty$ while $\lim_{\delta \to 0^+} C_{\text{Prop}10}^{\mu,\delta} = 0$.

Even though the third term on the right-hand side of Proposition 10’s signature inequality dominates the second term, the limiting behaviors of the coefficients have prompted us to keep the dominated term. They also give us some hope that a bound somewhere between the orders of $T^\mu$ and $T^{(1+\mu)/2}$ might be achievable for the limiting case where $\delta = 0$. For the time being, by putting Propositions 5 and 10 together, we can obtain a bound for the LwD($\mu$) policy that accounts for the nonperishability of items.

**Theorem 3** Let $(k, y)$ be the policy generated from following LwD($\mu$). Then, for any $\delta > 0$, there are constants $A_{\text{Them}3}^{\mu,\delta}$, $B_{\text{Them}3}^{\mu,\delta}$, $C_{\text{Th}}^{\mu,\delta}$, and $C_{\text{Prop}10}^{\mu,\delta}$, so that
\[
\sup_{f \in \Delta_{\delta}^k} R_f^{T}(y, k) \leq A_{\text{Them}3}^{\mu,\delta} + B_{\text{Them}3}^{\mu,\delta} \cdot T^\mu + C_{\text{Prop}5} \cdot T^{1/2} \cdot (\ln T)^{1/2} + C_{\text{Prop}10}^{\mu,\delta} \cdot T^{(1+\mu)/2} \cdot (\ln T)^{5/2}.
\]
Also, the choice $\mu = 1/2$ will achieve an $O(T^{3/4} \cdot (\ln T)^{5/2})$-bound.
Our next target is to bound the case involving both nonperishable items and truly arbitrary demand patterns. For the LwD(µ) policies, there is now no guarantee that any single price choice will be kept for sufficiently long. While still utilizing (59), our remedy is to consider stickier variants that allow longer reigns of incumbent price choices.

8 Nonperishability with Arbitrary Demand Patterns

A redefinition of virtual learning periods is needed before we can come to the modified policies. Fix some \( \nu > 0 \) and \( \psi \in (0, \mu) \). Let

\[
G_{\mu, \nu} = \frac{4 \cdot \bar{k}^{\mu}}{A^{\text{Prop8}} \cdot \mu} + 1, \tag{76}
\]

where \( A^{\text{Prop8}} \equiv 1/(4d^2 \cdot (\bar{\rho} \bar{\kappa} - \bar{c} + 3 \cdot (\bar{h} \lor \bar{d}))) \) is a constant that fits Proposition 8. Also, let \( I_{\mu, \nu, \psi} \geq \bar{k} \) be the smallest integer so that both \( \lceil (1 + I_{\mu, \nu, \psi})^{1/(1-\psi)/G_{\mu, \nu}} \rceil \geq 1 \) and \( \lceil (2 + I_{\mu, \nu, \psi})^{1/(1-\psi)/G_{\mu, \nu}} \rceil - \lceil (1 + I_{\mu, \nu, \psi})^{1/(1-\psi)/G_{\mu, \nu}} \rceil \geq 2 \). Now redefine \( s'_i = \lceil (i + I_{\mu, \nu, \psi})^{1/(1-\psi)/G_{\mu, \nu}} \rceil \) for \( i = 1, 2, \ldots \) as virtual learning periods. We have omitted the dependence of the \( s'_i \)'s on \( (\mu, \nu, \psi) \) for simplicity. Again, define \( N'_{t,0} \) through (64).

Now consider policy LwD'(\( \mu, \nu, \psi \)) that favors incumbent price choices in most of the doing periods. In particular, this new stickier policy shares the same learning periods as LwD(\( \mu \)). Also, it behaves the same as the original one except in doing periods \( t \) satisfying

\[
\text{neither } m_{t-1} = 0 \text{ nor } t - 1 = s'_i \text{ for some } i. \tag{77}
\]

For such a period, the only difference lies in replacing the original step 1.2 with the following:

1.2'. let price choice \( k^* \) be a maximizer of \( V^k_{f_{t-1}} \) from \( k = 1, 2, \ldots, \bar{k} \). If

\[
V^k_{f_{t-1}} \geq V^{k_{t-1}}_{f_{t-1}} + \frac{\nu}{t^{\mu-\psi}},
\]

let \( k_t = k^* \); otherwise, let \( k_t = k_{t-1} \).

Certainly, we will have \( k_t = k_{t-1} \) when the incumbent choice \( k_{t-1} \) happens to be \( k^* \). Otherwise, now there is a \( \nu/t^{\mu-\psi} \)-sized threshold for the prospective profit to cross before the price choice switches from the incumbent \( k_{t-1} \) to the new \( k^* \). Note that LwD(\( \mu \)) could somehow be understood as LwD'(\( \mu, 0, \psi \)) for any \( \psi \in (0, \mu) \). Propositions 3 and 4 hinge on the schedule of learning epochs; they are not affected by the change at some other periods. So the two propositions still apply to the new policy.
To go further, let \( L'(t) = \{s'_1, s'_2, \ldots, s'_{N_{t,0}}\} \) be the set of virtual learning periods up to \( t \). For \( m \in \mathcal{M}(t) \), consider the combined set \( L''(m, t) = L(m, t) \cup L'(t) \), where \( L(m, t) \) is still the set of actual learning periods as specified by LwD(\( \mu \)) under \( m \). We can write \( L''(m, t) \) as \( \{s''_1(m), s''_2(m), \ldots, s''_{N_{t,0}}(m)\} \) with \( 1 \leq s''_1(m) < s''_2(m) < \cdots < s''_{N_{t,0}}(m) \leq t \) and each \( s''_i(m) \) being either some \( s_j(m) \) or some \( s'_i \) or both. Now the condition (77) for step 1.2’ to be executed in a doing period \( t \) is that it not be some \( s''(m) + 1 \) under mode sequence \( m \).

For \( R^{T1}_f(k, y) \) defined at (50), the decomposition at (53) can be kept intact without any change on \( T_1 \) at (54) or \( T_3 \) at (56). The only difference is that (55) should be updated to

\[
T_2 \leq \sum_{k \neq k_{f'}} \sum_{t=1}^{T} \left( \delta \nu_{k_f} + \frac{\nu}{f^{\mu-\psi}} \right) \cdot \mathbb{P}_f \left[ m_t = 1 \text{ and } V^{k_{f'}}_{t-1} \text{ achieves the maximum} \right].
\]

The extra term \( \nu/f^{\mu-\psi} \) comes from the fact that now the incumbent can be worse off by as much as this amount and yet still keep its place. Its contribution to \( R^{T1}_f(k, y) \) is \( T^{1-\mu+\psi} \). We thus have the following adaptation of Proposition 6.

**Proposition 11** Let \((k, y)\) be the policy generated from LwD(\( \mu, \nu, \psi \)). Then, there are positive constants \( A_{\text{Prop6}}, B_{\text{Prop6}}, \) and \( C^{\text{Prop11}}_{\mu, \nu, \psi} \) so that

\[
\sup_{f \in \Delta^k} R^{T1}_f(k, y) \leq A_{\text{Prop6}} + B_{\text{Prop6}} \cdot T^{\nu/(1-\mu/2)} + C^{\text{Prop11}}_{\mu, \nu, \psi} \cdot T^{1-\mu+\psi}.
\]

Any choice with \( \mu = 2/3 \) and \( \psi \leq 1/3 \) will achieve an \( O(T^{2/3}) \)-bound.

We have put proofs of this section, except that for Theorem 5, into Appendix D. Propositions 6 and 11 both have \( T^{2/3} \)-sized bounds at their respective bests. So the LwD(\( \mu, \nu, \psi \)) policies do not give much away when items are perishable.

We now attempt to bound \( R^{T2}(k, y) \) defined at (51), the task to which the new policies are tailored. It turns out that a new version of Proposition 7 is not necessary. On the other hand, we need a modification of Proposition 8. This time around, let \( E_{t'}^{k} \) be the event that \( \hat{V}^{k_{f'}}_{t'-1} \geq \max_{k \neq k_t} \hat{V}^{k}_{f_t-1} \) and there is no interruption from learning in periods \( t, t+1, \ldots, t+t' -1 \).

**Proposition 12** Suppose \( t \) is large enough so that

\[
2 \cdot t^{\mu-\psi} > \left( t + A^{\text{Prop8}} \cdot \delta \cdot \nu \cdot \left( \frac{t^{\psi}}{k^\mu} - 1 \right) - 1 \right)^{\mu-\psi},
\]

where \( A^{\text{Prop8}} \) is the same one used in Proposition 8. Then, for \( t' \) as large as \( A^{\text{Prop8}} \cdot \delta \cdot \nu \cdot (t^{\psi}/k^\mu - 1) \), one can expect \( k_t = k_{t+1} = \cdots = k_{t+t' -1} \) whenever \( E_{t'}^{k_t} \) is true.
As in Proposition 9, the frequencies at which the new combined learning periods $s''_i(m)$ arise are also constrained both ways.

**Proposition 13** We have the following useful inequalities:

$$N''_{t,0}(m) \leq H_{\mu,\nu} \cdot t^{\mu \nu (1-\psi)}, \quad (79)$$

for some constant $H_{\mu,\nu}$ which is above $G_{\mu,\nu}$;

$$s''_i(m) \leq \left( \frac{1}{G_{\mu,\nu}^{1/(1-\psi)}} \right) \cdot (i + I_{\mu,\nu,\psi})^{1/(1-\psi)} + 1; \quad (80)$$

$$s''_{i+1}(m) - s''_i(m) \leq \left[ \left( \frac{4}{G_{\mu,\nu}} \right) \cdot (s''_i(m))^\psi + 1 \right] \vee \left[ \frac{(1 + I_{\mu,\nu,\psi})^{1/(1-\psi)}}{G_{\mu,\nu}^{1/(1-\psi)}} \right]. \quad (81)$$

In Proposition 13, (79), (80), and (81) correspond, respectively, to (65), (67), and (68) in Proposition 9. However, the counterpart to the earlier (66) is not needed now. We can still decompose $R^{T^2}(k,y)$ in the fashion of (70) to (73). By leveraging the fact that $V^k_{j\neq k} \geq \max_{k \neq k} V^k_{j_{k-1}}$ for a doing period $t$ which happens to be some $s''_i(m) + 1$, as well as Propositions 12 and 13 in similar manners in which we used Propositions 7 to 9 in the proof of Proposition 10, we can come to the following counterpart to the latter result.

**Proposition 14** There are positive constants $A^\text{Prop14}_{\mu,\nu,\psi}$, $B^\text{Prop14}_{\mu,\nu,\psi}$, and $C^\text{Prop14}_{\mu,\nu,\psi}$, so that the LwD$(\mu, \nu, \psi)$ policy $(k, y)$ would satisfy, for any $f \in \Delta_k$,

$$R^T_f(k,y) \leq A^\text{Prop14}_{\mu,\nu,\psi} + B^\text{Prop14}_{\mu,\nu,\psi} \cdot T^{\mu \nu (1-\psi)} + C^\text{Prop14}_{\mu,\nu,\psi} \cdot T^{(2-\psi)\nu (1-\psi)/(2-2\psi)} \cdot (\ln T)^{5/2}. \quad (82)$$

Note that the choice $\mu = 2/3$ and $\psi = 1/3$ will achieve an $O(T^{5/6} \cdot (\ln T)^{5/2})$-bound.

When combining Propositions 11 and 14, we obtain a complete bound.

**Theorem 4** Let $(k, y)$ be the policy generated from following LwD$(\mu, \nu, \psi)$. Then, there are constants $A^\text{Them4}_{\mu,\nu,\psi}$, $B^\text{Them4}_{\mu,\nu,\psi}$, and $C^\text{Them4}_{\mu,\nu,\psi}$, so that $\sup_{f \in \Delta_k} R^T_f(y, k)$ is below

$$A^\text{Them4}_{\mu,\nu,\psi} + B^\text{Them4}_{\mu,\nu,\psi} \cdot T^{\mu \nu (1-\mu/2) \nu (1-\nu)(1-\mu+\nu)} + C^\text{Them4}_{\mu,\nu,\psi} \cdot T^{(2-\psi)\nu (1-\psi)/(2-2\psi)} \cdot (\ln T)^{5/2}. \quad (83)$$

Also, the choice $\mu = 2/3$ and $\psi = 1/3$ will achieve an $O(T^{5/6} \cdot (\ln T)^{5/2})$-bound.

Now come to the lower bound. Without loss of generality, we suppose that $\bar{p}^1 < \bar{p}^2 < \cdots < \bar{p}^k$. Our best effort so far has achieved a result of $\Omega(T^{1/2})$ when, either because $\bar{b} + \bar{c} - \bar{p}^k \leq 0$ or because $\bar{b} + \bar{c} - \bar{p}^k > 0$ but $\bar{d}$ is large enough,

$$(\bar{p}^k - \bar{c}) \cdot (\bar{d} - 1) > (\bar{b} + \bar{c} - \bar{p}^k) \cdot (1 - \beta). \quad (82)$$
Theorem 5 Under (82), there is a constant $A^{Them5}$ so that for any adaptive policy $(k,y)$,

$$\sup_{r \in \Delta^k} R^T_t(k,y) \geq A^{Them5} \cdot T^{1/2}.$$ 

Also, this is true even if we relax the requirement (3).

The bound is no tighter than what Besbes and Muharremoglu [6] can achieve for the pure inventory control case. However, in the hope that our current derivation can be improved upon to reach a tighter bound, we have pursued in our own direction. Details can be found in Section 4 of Zhou, Katehakis, and Yang [36]. There remains a sizable gap between Theorem 4’s $T^{5/6}$-sized upper bound for LwD′(2/3, ν, 1/3) policies and this $T^{1/2}$-sized lower bound for arbitrary policies.

In the simulation study to be presented in Section 9, regrets involving nonperishable items are found to grow at close to the rate of $T^{2/3}$, one that is predicted for the perishable-item case by Proposition 6 while using the LwD(2/3) policy and Proposition 11 while using the LwD′(2/3, ν, 1/3) policies. Therefore, besides working with the temptingly improvable lower bound Theorem 5, we might also look beyond techniques employing (59) for clues to further improve Proposition 14 and then Theorem 4 concerning nonperishable items.

Theorems 3 and 4, as well as to less degrees, Propositions 5, 6, and 11 for the perishable-item special case, are complementary to the upper bounds achieved by Chen, Chao, and Ahn [14]. Dealing with continuous demands that enjoy specific relations with prices, the earlier work obtained $T^{1/2}$-sized bounds. We, on the other hand, have treated the discrete-item case where the finest tuning of ordering decisions are impossible. In addition, we have allowed demand-distribution vector $f$ to come from virtually anywhere in $\Delta^k$ in case of the $T^{3/4}$-sized bounds and literally everywhere in case of the $T^{5/6}$-sized bounds. This helps us to largely shrug off the perils of model mis-specification.

9 Simulation Study

We use a simulation study to gain more insights on the growth trends of regrets in both pure inventory and joint price-inventory controls. For the latter, the study also hints at the best possible rate of the growth rate, an issued yet unresolved by our theoretical analysis.

In the study, we fix $\bar{h} + \bar{b} = 10$ and $\bar{d} = 20$, and let $\bar{b}$’s variation drive changes in $\beta = \bar{b}/(\bar{h} + \bar{b})$. Smaller $\beta$ values are suitable for the backlogging case and larger $\beta$ values the
lost sales case. We first deal with pure inventory control. At each parameter combination, we randomly generate \( M = 1,000 \) distributions \( f \) in \( \Delta \) uniformly. Under each \( f \) out of the \( M \) possibilities, we test policies \( a \) for \( T = 10,000 \) periods on \( L = 100 \) sample paths of demand levels. Let \( y^a_t \) be the order-up-to level in period \( t \) under policy \( a \). Due to (4) to (6), the following is approximately the policy's regret by time \( t \):

\[
r^a_{f,t} = \text{AVG} \left\{ \sum_{s=1}^{t} [\bar{h} \cdot (y^a_{s} - d_s)^+ + \bar{b} \cdot (d_s - y^a_{s})^+] \right\} - Q^*_f \cdot t, \tag{83}
\]

where \( Q^*_f \) is defined around (4) and AVG stands for average over the \( L = 100 \) demand paths.

For any \( \alpha \in (0,1) \), we let \( R^a_{\alpha,t} \) be the conditional value at risk at the \( \alpha \)-quantile of the \( M \) regrets \( r^a_{f,t} \). For instance, \( R^a_{0.100,t} \) would stand for the average of the top 50 highest \( r^a_{f,100,t} \) values, where each is the regret of policy \( a \) by time \( t = 100 \), out of the \( M = 1,000 \) randomly generated \( f \)'s. Also, \( R^a_{0.0,t} \) would be the average of all the \( r^a_{f,100,t} \)'s sampled from all over \( \Delta \). Suppose policy \( a \) is the \( y \) studied in Sections 3 and 4. Then, when \( M \) and \( L \) both approach \( +\infty \) and \( \alpha \) approaches 100%, \( R^a_{\alpha,t} \) will approach \( \sup_{f \in \Delta} R_f(y) \), the focal point of Theorem 2. Since \( L \) is finite, each \( r^a_{f,t} \) is merely an approximation of the regret \( R_f(y) \). Moreover, the finiteness of \( M \) means that the \( f \) in \( \Delta \) generating the worst regret will most likely be missed. Nevertheless, the \( R^a_{\alpha,t} \) values will yield insights into regrets of policies \( a \).

We mainly test two policies \( a \) for pure inventory control, with \( a = 0 \) and 1. Policy 0 is our newservendor-based one defined through (11) and (12). Policy 1 is adapted from the SA-based one proposed by Huh and Rusmevichientong [24] to the current discrete-item setting. It still follows (11), although its generation of the \( \hat{y} t \)'s is different from (12). For the latter, let \( \varepsilon_t = \bar{d}/((\bar{h} \lor \bar{b}) \cdot t^{1/2}) \). There is also an auxiliary process \( \hat{z}_t \). In the beginning, \( y_1 = \hat{y}_1 = \hat{z}_1 = 0 \). Then for \( t = 2, 3, \ldots \),

\[
\hat{y}_t = \begin{cases} 
[\hat{z}_t], & \text{with probability } \lfloor \hat{z}_t \rfloor - \hat{z}_t, \\
[\hat{z}_t], & \text{with probability } 1 - (\lfloor \hat{z}_t \rfloor - \hat{z}_t),
\end{cases} \tag{84}
\]

where \( \hat{z}_t \) equals

\[
\begin{cases} 
(\hat{z}_{t-1} - \bar{h} \varepsilon_{t-1}) \lor 0 \land \bar{d}, & \text{when } \hat{y}_{t-1} = \lfloor \hat{z}_{t-1} \rfloor \text{ and } d_{t-1} \leq y_{t-1}, \text{ or} \\
(\hat{z}_{t-1} + \bar{b} \varepsilon_{t-1}) \lor 0 \land \bar{d}, & \text{when } \hat{y}_{t-1} = \lfloor \hat{z}_{t-1} \rfloor \text{ and } d_{t-1} \geq y_{t-1} + 1, \text{ or}
\end{cases} \tag{85}
\]

We have also considered a policy inspired by both stochastic approximation and Derman’s [19] up-and-down idea. However, our simulation study indicates that policy 2 is not com-
petitive against either of the previous two policies, except when $\beta$ is close to 0.5, at which time it is better than policy 1 in some occasions. So we omit presenting its performances.

At various $\alpha$ and $\beta$ values, and at different $t$ points, we compare $R^{a,t}_\alpha$ for the two policies. At various $\beta$ values, we can evaluate $R^{a,t}_\alpha$ for policies $a = 0$ and 1 at times $t = 1^2, 2^2, \ldots, 100^2$ and $\alpha$ values at 0%, 95%, and 99.9%. Except for scaling differences, we find the basic findings do not depend much on $\beta$. In Figures 1 to 3, we just present results at the three $\beta$ values of 0.1, 0.5, and 0.9. Instead of $t$, we have used $t^{1/2}$ as our horizontal axis.

![Figure 1: $R^{a,t}_\alpha$ Values when $\beta = 0.1$](image)

From these figures, we can observe that policy 0 generates much smaller regrets, in both average and worst senses, than policy 1. This should be anticipated, as policy 0 utilizes more information regarding past observations than policy 1. We also see that regrets for both policies grow at approximately the rate of $t^{1/2}$.

For joint inventory-price control, our main purpose is to determine where the growth rate of the optimal regret stands against the backdrops of the $T^{5/6}$-sized upper bound of Theorem 4 and the $T^{1/2}$-sized lower bound of Theorem 5. In addition, we want to seek out competitive policies. Recall that each LwD($\mu$) could be understood as an LwD($\mu, 0, \psi$). Besides
the general LwD'(µ, ν, ψ) policies, we also test random policies which we call rLwD(υ, ω), where υ and ω are positive constants. In every period t, a particular rLwD(υ, ω) chooses the price choice k that maximizes

\[ V^k_{t-1} + \nu \cdot \frac{|Z|}{(N^k_{t-1} + 1)^\omega}, \]

where Z is sampled from the standard Normal distribution. This policy also strives to balance between exploration and exploitation, albeit in a fashion different than that employed by any LwD'(µ, ν, ψ). Let its ordering be newsvendor-based as well.

To compare policies, we randomly generate \( M = 2,000 \) demand-distribution vectors \( f \equiv (f^1, f^2, ..., f^k) \) in \( \Delta^k \) uniformly, and at each selected vector f, randomly generate \( L = 200 \) demand-vector sample paths. We consider \( \bar{k} = 2 \) price choices, with \( \bar{d} = 20, \bar{h} = 5, \bar{b} = 5, \bar{c} = 50, \bar{p}^1 = 80, \) and \( \bar{p}^2 = 100. \) We only study \( \bar{k} = 2 \) while using a bigger \( M \) than the pure inventory control case due to the need to search thoroughly for bad cases in the space \( \Delta^k \), which grows exponentially in size with \( \bar{k} \). Now demand-distribution vectors are merely pairs. Given demand-distribution pair \( f \equiv (f^1, f^2) \), let \( k^a_t \) be the price choice and \( y^a_t \) the order-up-to level in period t under a given policy a for a particular demand-pair sample path.
Due to (33), (36), and (37), we use the following as an approximation to the policy’s regret on demand-distribution pair $f$ by time $t$:

$$
R_{\alpha,t}^f = V^*_f \cdot t - \text{AVG} \left\{ \sum_{s=1}^{t} \left[ \left( \bar{p}^k s - \bar{c} \right) \cdot d_s - \bar{h} \cdot (y^a_s - d_s)^+ - \bar{b} \cdot (d_s - y^a_s)^+ \right] \right\}, \tag{87}
$$

where $V^*_f$ is defined at (35) and AVG stands for average over the $L$ demand-pair paths.

At various $t$ points up to $T = 15,000$, we compare the $R_{\alpha,t}^f$ values among the different policies. With $M = 2,000$, each $R_{99\%}^{0,t}$ captures the average of the 20 worst regrets $r_{t}^{\alpha,t}$ as computed in (87). An $\alpha$ even closer to 100% would certainly produce results that reflect the true worst regret more faithfully. However, we observe that average of a sufficient number of demand-distribution pairs is needed for our regret trend to be smooth. In addition, constraints on computational resources have prevented us from pursuing even larger $M$ values. Thus, we settle with $M = 2,000$ and $\alpha = 99\%$. The similar trends and proximities of the regret curves at $\alpha = 95\%$ and $\alpha = 99.9\%$ in Figures 1 to 3 can serve as testimonies to the soundness of using the $R_{\alpha,t}^{99\%}$’s to evaluate the worst regrets.

Using the $R_{\alpha=99\%}^{0,t}$ values for $t = 1, \ldots, T = 15,000$, we examine various policies $a$ where
each $a$ represents either some $LwD'(\mu, \nu, \psi)$ or some $rLwD(\nu, \omega)$. For the former, we have tried $\mu = 1/2, 2/3, 3/4, \nu = 0, 10, 100, \text{and } \psi = \mu/4, \mu/2, 3\mu/4$; whereas, for the latter, we have tried $\nu = 100, 500, 1000, 2000$ and $\omega = 1/2, 2/3, 1, \text{and } 3/2$. Although having offered help to our theoretical development in Section 8, we find the general cases with $\nu > 0$ do not provide substantial improvements in practice. Hence, we could just as well revert back to the simpler $LwD(\mu)$ policies. Among the $rLwD(\nu, \omega)$ policies, on the other hand, we find the one with $(2000, 1)$ to be particularly competitive. From now on, let us narrow down to policies $a = 0$ to $3$, with policy 0 being $LwD(2/3)$, policy 1 $LwD(1/2)$, policy 2 $LwD(3/4)$, policy 3 $rLwD(2000, 1)$. For each of the policies 0 to 2, we have also tested the variants inspired by Burnetas and Smith [13], in which the $V_{k,t}^k$ term used in step 1.2 is replaced by $\hat{V}_{t-1}^k$, defined at (46). However, the changes do not much improve performances.

Now in Figures 4 to 6, we present results on $R_{99\%}^{a,t}$ for policies $a = 0$ to $3$ at various time points with, respectively, the horizontal axis being scaled at $t^{5/6}$, $t^{1/2}$, and $t^{2/3}$.

![Graph](image)

**Figure 4:** $R_{99\%}^{a,t}$ Values when Horizontal Scale is $t^{5/6}$

In Figure 4 where $t^{5/6}$ serves as the horizontal axis, the growth rates of regrets of all policies are downward-sloping, and in Figure 5 where $t^{1/2}$ serves as the same, those rates are
all upward-sloping. Meanwhile, in Figure 6 where \( t^{2/3} \) serves as the horizontal axis, all regrets grow almost linearly. These suggest that the growth rate of the regret of any “reasonable policy” is in the vicinity of \( T^{2/3} \). So far, after much effort, we have not found a policy whose growth rate of regret can be significantly slower than the \( T^{2/3} \)-pace.

For the most promising policy 3, we choose \( M = 10,000 \) and continue to simulate the regrets \( R^{3,t}_{99.8\%} \) (20 out of 10,000 is 0.2%) up to \( t = 20,000 \). We then conduct a linear regression analysis in the form of

\[
\ln(R^{3,t}_{99.8\%}) = a + b \cdot \ln t,
\]

for \( t = 2,001 \) to 20,000. With the R-square at 99.4% and the \( p \)-value practically zero, the least-squares estimate for the slope \( b \) turns out to be 0.62. Hence, neither Theorem 4 nor Theorem 5 seems to have had the final say yet. Since being worse than 99.8% of distributions is not the worst yet, the regret growth of policy 3 could well be \( T^{2/3} \)-sized. On the other hand, there still might be more competitive policies. We tend to conjecture that the signature regret growth rate for joint inventory-price control is somewhere between \( T^{3/5} \)- and \( T^{2/3} \)-sized; this sets it apart from pure inventory control, whose signature rate is \( T^{1/2} \)-sized.
Our final task is to advance understandings, separately, on the perishable-item portion \( R_f^{T_1}(k, y) \) and the nonperishability-induced portion \( R_f^{T_2}(k, y) \). Define \( r^{a,t,1} \) in the same fashion of (87) as \( r^{a,t} \) was, but with the newsvendor order-up-to level \( \hat{y}_s^a \) replacing the actual level \( y_s^a \) which has also to satisfy (3). Then, let \( r^{a,t,2} = r^{a,t} - r^{a,t,1} \). These will be regrets for perishable items and those stemming from the nonperishability issue, respectively. Following the comment after (15), it is not necessary that \( r^{a,t,2} \geq 0 \). In the same manner that \( R_a^{a,t} \) was defined from \( r^{a,t} \), we can define \( R_a^{a,t,1} \) and \( R_a^{a,t,2} \). For policies \( a = 0 \) to 3, we now present the former measures in Figure 7 and the latter measures in Figure 8, both at \( \alpha = 99\% \) and using the horizontal scale \( t^{2/3} \).

From Figure 7, we see that Propositions 6 and 11 are fairly competent at handling perishable items. From Figure 8, however, we can tell that Proposition 14 has overestimated the effect of nonperishability. As for constants, those in Figure 7 representing perishable items actually dominate those in Figure 8 reflecting nonperishability. This adds doubts on whether the best approach to handle nonperishability is the current one based on (59).
Concluding Remarks

We have worked on both pure inventory and joint inventory-price controls involving discrete nonperishable items and unknown demand. For the former, we contribute on the case where demand is completely unknown. For the latter joint control case, we have proposed $LwD(\mu)$ policies and their variants that build on the newsvendor-based policy and balance between learning/exploration and doing/exploitation. The nonperishability issue here is dealt with using a bound developed for pure control. Our emphasis on the case with completely thorough ambiguity in demand can help users to avoid model mis-specification. This, however, may have come at the expense of potentially less desirable bounds.

Certainly, more await to be done. The newsvendor-based policy requires higher observability of historical demand levels than other policies, say the SA-based one. This compromises its suitability for situations involving demand censoring. When pricing is involved, the gap between upper and lower bounds need to be narrowed. Furthermore, the complete removal of any relation between price and demand might have overshot. For instance, it is utterly reasonable that the demand distribution, though otherwise unknown, be decreasing.
in price in the stochastic sense. It will thus be interesting to know how such knowledge will help to tighten the regret bounds.

References


Appendices

A Proofs of Section 4

Proof of Proposition 1: For $f, g \in \Delta$, note that $Q_f(y^*_g) - Q_f(y^*_f)$ can be written as

$$[Q_f(y^*_g) - Q_g(y^*_g)] + [Q_g(y^*_g) - Q_g(y^*_f)] + [Q_g(y^*_f) - Q_f(y^*_f)].$$

(89)

While the first and third terms can be made small when $f$ and $g$ are close, the second term is always negative due to $y^*_g$’s optimality when the underlying demand distribution is $g$. Let us investigate how small the first and third terms can be. For any $y = 0, 1, ..., \bar{d},$

$$Q_f(y) - Q_g(y) = \bar{h} \cdot \sum_{d=0}^{\bar{d}-1} [f(d) - g(d)](y - d) + \bar{b} \cdot \sum_{d=y}^{\bar{d}} [f(d) - g(d)](d - y)$$

$$\leq (\bar{h} \vee \bar{b}) \cdot \bar{d} \cdot ||f - g||_1 = 2 \cdot (\bar{h} \vee \bar{b}) \cdot \bar{d} \cdot \delta_V(f, g).$$

(90)

In view of the discussion around (89),

$$Q_f(y^*_g) - Q_f(y^*_f) \leq 4 \cdot (\bar{h} \vee \bar{b}) \cdot \bar{d} \cdot \delta_V(f, g).$$

(91)

Note (4) also leads to

$$Q_f(y^*_g) - Q_f(y^*_f) \leq (\bar{h} \vee \bar{b}) \cdot \bar{d}.$$  

(92)

Consider $R^{T1}_f(y)$ defined at (14). By (12), we have

$$R^{T1}_f(y) = \sum_{t=1}^{T} \{ E_f[Q_f(y^*_{f_{t-1}})] - Q_f(y^*_f) \}.$$  

(93)

Let $\varepsilon_t$ be a sequence of positive constants. We then see that $R^{T1}_f(y)$ is below

$$\sum_{t=1}^{T} \{ \mathbb{P}_f[\delta_V(f, \hat{f}_{t-1}) < \varepsilon_t] \cdot E_f[Q_f(y^*_{f_{t-1}})] - Q_f(y^*_f)] \delta_V(f, \hat{f}_{t-1}) < \varepsilon_t \}$$

$$+ \mathbb{P}_f[\delta_V(f, \hat{f}_{t-1}) \geq \varepsilon_t] \cdot E_f[Q_f(y^*_{f_{t-1}})] - Q_f(y^*_f)] \delta_V(f, \hat{f}_{t-1}) \geq \varepsilon_t]) \}$$

$$\leq 2 \cdot (\bar{h} \vee \bar{b}) \cdot \bar{d} \cdot \sum_{t=1}^{T} [2 \varepsilon_t + \exp(-2 \varepsilon_t^2 \cdot (t - 1))].$$

(94)

where the inequality comes from (20), (91), and (92). Suppose $\varepsilon_1 = 0$ and $\varepsilon_t = (\ln t / (t-1))^{1/2}$ for $t = 2, 3, ..., T$. Then, after plugging this into (94), we get

$$R^{T1}_f(y) \leq 2 \cdot (\bar{h} \vee \bar{b}) \cdot \bar{d} \cdot [2T_1 + T_2],$$

(95)

where

$$T_1 = \sum_{t=1}^{T-1} \frac{(\ln (t+1))^{1/2}}{t^{1/2}}, \quad \text{and} \quad T_2 = \sum_{t=1}^{T} \frac{1}{t^2}.$$  

(96)
Clearly, \[ T_1 \leq (\ln T)^{1/2} \cdot \int_0^T \frac{1}{t^{1/2}} \cdot dt = \frac{T^{1/2} \cdot (\ln T)^{1/2}}{2}, \] (97)
while \( T_2 \) is known to be bounded by a constant.

**Proof of Proposition 2:** Let \( \gamma = 1 - f(0) \). We divide the proof into two cases, with respectively, \( \gamma \in [(1 - \beta)/2, 1] \) and \( \gamma \in [0, (1 - \beta)/2] \).

Consider the first case with \( \gamma \in [(1 - \beta)/2, 1] \). For any positive integer \( \tau_T \), we show how \( \mathbb{P}_f[S_{i+1} - S_i - 1 \geq \tau_T + 1] \) can be bounded. By the definition of the \( S_i \)'s around (28),
\[ \hat{y}_{S_{i+1}} - \hat{y}_S - D_S - \cdots - D_{S_{i+1} - 2} - 1. \] (98)
Since both \( \hat{y}_S \) and \( \hat{y}_{S_{i+1}} \) are between 0 and \( \hat{d} \), the above necessitates that
\[ D_S + D_{S+1} + \cdots + D_{S_{i+1} - 2} \leq \hat{d} - 1. \] (99)
This is only possible when there are at least \( S_{i+1} - S_i - \hat{d} \) zeros among the \( S_{i+1} - S_i - 1 \) demand levels \( D_S, D_{S+1}, \ldots, D_{S_{i+1} - 2} \). When \( S_{i+1} - S_i - 1 = \tau + 1 \geq \hat{d} - 1 \), the latter event's chance under \( f \) with \( f(0) = 1 - \gamma \) is, by the binomial formula,
\[ \sum_{k=\tau - \hat{d} + 1}^{\tau + 1} \frac{(\tau + 1)!}{k! \cdot (\tau + 1 - k)!} \cdot (1 - \gamma)^k \cdot \gamma^{\tau + 1 - k} \times (1 - \gamma)^{\hat{d} - 1} \cdot (1 + \gamma + \cdots + \gamma^{\hat{d}}), \] (100)
which is less than \( (\tau + 1)^{\hat{d}} \cdot (1 - \gamma)^{\tau - \hat{d} + 1} \). There exists \( \theta_\gamma = \hat{d} - 1, \hat{d}, \ldots \) so that when \( \tau \geq \theta_\gamma \), the aforementioned term will decrease with \( \tau \). For \( \tau_T \geq \theta_\gamma \), we can thus deduce that
\[ \mathbb{P}_f[S_{i+1} - S_i - 1 \geq \tau_T + 1] < (\tau_T + 1)^{\hat{d}} \cdot (1 - \gamma)^{\tau_T - \hat{d} + 1}. \] (101)
Each of the terms in (29) is between 0 and \( (\hat{h} + \hat{b}) \cdot \hat{d} \). So by (26) and (101),
\[ R_f^{T_2}(y) \leq \sum_{t=3}^{T} \sum_{s=2}^{(t-\tau_T)} \mathbb{E}_f[\left| Q_f(\hat{y}_s - D_s - \cdots - D_{t-1}) - Q_f(\hat{y}_t) \right| \times \times 1(\hat{y}_t \leq \hat{y}_s - D_s - \cdots - D_{t-1} - 1)] \times (T - 2) \cdot (\hat{h}\hat{d} + \hat{b}\hat{d}) \cdot (\tau_T + 1)^{\hat{d}} \cdot (1 - \gamma)^{\tau_T - \hat{d} + 1}. \] (102)
The above right-hand side can be written as
\[ \sum_{\tau=1}^{\tau_T} R_f^{T_2,\tau}(y) + (T - 2) \cdot (\hat{h}\hat{d} + \hat{b}\hat{d}) \cdot (\tau_T + 1)^{\hat{d}} \cdot (1 - \gamma)^{\tau_T - \hat{d} + 1}, \] (103)
where for \( \tau = 1, 2, \ldots, \tau_T \),
\[ R_f^{T_2,\tau}(y) = \sum_{t=\tau+2}^{T} \mathbb{E}_f[\left| Q_f(\hat{y}_{t-\tau} - D_{t-\tau} - \cdots - D_{t-1}) - Q_f(\hat{y}_t) \right| \times \times 1(\hat{y}_t \leq \hat{y}_{t-\tau} - D_{t-\tau} - \cdots - D_{t-1} - 1)]. \] (104)
By (9) and (12), we have 
\[ \hat{F}_{t-\tau}(\hat{y}_{t-\tau}) - \beta \leq \hat{F}_{t-1}(\hat{y}_t) \leq \hat{F}_{t-1}(\hat{y}_{t-\tau} - D_{t-\tau} - \cdots - D_{t-1} - 1). \] 
(105)

Also, due to the nature of the empirical distribution as illustrated in (10),
\[ \hat{F}_{t-1}(\hat{y}_{t-\tau} - D_{t-\tau} - \cdots - D_{t-1} - 1) \leq \hat{F}_{t-1}(\hat{y}_{t-\tau} - 1) \leq \hat{F}_{t-1}(\hat{y}_{t-\tau} - 1) + \frac{\tau}{t - \tau}. \] 
(106)

Therefore, \( \hat{y}_t \leq \hat{y}_{t-\tau} - D_{t-\tau} - \cdots - D_{t-1} - 1 \) only if
\[ \beta \leq \hat{F}_{t-1}(\hat{y}_t) \leq \hat{F}_{t-1}(\hat{y}_{t-\tau} - D_{t-\tau} - \cdots - D_{t-1} - 1) < \beta + \frac{\tau}{t - \tau} \leq \beta + \frac{\tau}{t - \tau}, \] 
(107)

an inequality alluded to earlier in (30). On the other hand, (4) has that, for \( y \leq z, \)
\[ Q_f(z) - Q_f(y) = \bar{h} \cdot \sum_{d=0}^{z} \sum_{d=1}^{z} (1 - F_f(d)) - \bar{h} \cdot \sum_{d=0}^{z} (1 - F_f(d)) = (\bar{h} + \bar{b}) \cdot \sum_{d=0}^{z} (1 - F_f(d)) \] 
(108)

Now by (104), (107), and (108), \( R_f^{T, \tau}(y) \) is less than \( (\bar{h} + \bar{b}) \cdot \sum_{t=\tau+2}^{T} E_f[Z_t] \), where
\[ Z_t = \sum_{d=\hat{y}_t}^{\hat{y}_{t-\tau} - D_{t-\tau} - \cdots - D_{t-1} - 1} | F_f(d) - \beta | \times 1(\beta \leq \hat{F}_{t-1}(\hat{y}_t) \leq \hat{F}_{t-1}(\hat{y}_{t-\tau} - D_{t-\tau} - \cdots - D_{t-1} - 1) < \beta + \tau/(t - \tau)). \] 
(109)

In general,
\[ Z_t \leq \bar{d}. \] 
(110)

Noting in turn that \( | F_f(d) - \hat{F}_{t-1}(d) | \leq \delta_V(f, \hat{f}_{t-1}), \) we also have
\[ Z_t \leq \sum_{d=\hat{y}_t}^{\hat{y}_{t-\tau} - D_{t-\tau} - \cdots - D_{t-1} - 1} | F_f(d) - \hat{F}_{t-1}(d) | + | \hat{F}_{t-1}(d) - \beta | \times 1(\beta \leq \hat{F}_{t-1}(\hat{y}_t) \leq \hat{F}_{t-1}(\hat{y}_{t-\tau} - D_{t-\tau} - \cdots - D_{t-1} - 1) < \beta + \tau/(t - \tau)) \] 
(111)

So for a sequence \( \varepsilon_t, \)
\[ R_f^{T, \tau}(y) \leq (\bar{h} + \bar{b}) \cdot \sum_{t=\tau+2}^{T} E_f[Z_t] \cdot \delta_V(f, \hat{f}_{t-1}) \leq \varepsilon_t \cdot P_f[\delta_V(f, \hat{f}_{t-1}) \leq \varepsilon_t] \] 
(112)

which by (20), (110), and (111), is less than
\[ (\bar{h}d + \bar{b}d) \cdot \sum_{t=1}^{T} \left[ \varepsilon_t + 2 \cdot (d + 1) \cdot \exp(-2\varepsilon_t^2 \cdot (t - 1)) + \frac{\tau}{t} \right]. \] 
(113)
The situation we face is very similar to (94) except for the $\tau_T/t$-term. So as in Proposition 1, there are constants $C''$, $D''$, and $E''$ so that

$$R_f^{T_2}(y) \leq C'' + D'' \cdot T^{1/2} \cdot (\ln T)^{1/2} + E'' \cdot T^{\gamma} \cdot \ln T.$$  \hspace{1cm} (114)

Note the $E''$-term stems from the $\tau_T/t$-term in (113). In view of (102) and (103), $R_f^{T_2}(y)$ is bounded by

$$C'' \cdot T \cdot (\ln T)^{1/2} + E'' \cdot \frac{T^{1/2} \cdot (\ln T)^{3/2}}{\ln(1/(1 - \gamma))} + T^{\gamma} \cdot \ln T$$

when $\tau_T$ is above the $\theta_T$ defined right after (100). Otherwise, we have almost the same inequality, albeit with the last term replaced by $(T - 2) \cdot (\ln T \cdot (1/(1 - \gamma)))$. Then, as long as $T$ is large enough, say greater than some $T_0$, we can ensure that $\tau_T$ is above $\theta_T$. Very importantly, just because $\gamma \in (0, 1]$, we can make sure that the last term, regardless whether $\tau_T$ is below or above $\theta_T$, is always bounded from above by a positive constant $F''$. Thus, $R_f^{T_2}(y)$ is less than

$$R_f^{T_2}(y) \leq \frac{C'' \cdot \ln T}{\ln(1/(1 - \gamma))} + \frac{D'' \cdot T^{1/2} \cdot (\ln T)^{3/2}}{\ln(1/(1 - \gamma))} + E'' \cdot \left(\frac{1}{\ln(1/(1 - \gamma))}\right)^2 \cdot (\ln T)^3 + F'' \cdot (\ln T)^2.$$  \hspace{1cm} (116)

However, as long as $T$ is large enough, the $T^{1/2} \cdot (\ln T)^{3/2}$-sized term will dominate all other terms. A constant term can certainly cover the case when $T$ is not that large. Therefore, positive constants $C''_T$ and $D''_T$ exist for the intended inequality

$$R_f^{T_2}(y) \leq C''_T + D''_T \cdot T^{1/2} \cdot (\ln T)^{3/2}.$$  \hspace{1cm} (117)

Since $(C''_{(1-\beta)/2}, D''_{(1-\beta)/2})$ can be used as $(C''_T, D''_T)$ for cases with $\gamma \geq (1 - \beta)/2 > 0$, we can have the intended bound, namely,

$$R_f^{T_2}(y) \leq A^{Prop} + B^{Prop} \cdot T^{1/2} \cdot (\ln T)^{3/2},$$  \hspace{1cm} (118)

as long as $\gamma$ stays above $(1 - \beta)/2$.

We now turn to the second case with $\gamma \in [0, (1 - \beta)/2)$. From (29), $R_f^{T_2}(y)$ is equal to

$$\sum_{t=3}^{T} \sum_{s=2}^{t-1} E_f[Q_f(\bar{y}_s - D_s - \cdots - D_{t-1}) - Q_f(\bar{y}_t) | L(t) = s] \cdot P_f[L(t) = s]$$

$$\leq (\hat{h}d + \hat{b}d) \cdot \sum_{t=3}^{T} \sum_{s=2}^{t-1} E_f[1(\bar{y}_s - D_s - \cdots - D_{t-1} \geq 1) | L(t) = s] \cdot P_f[L(t) = s]$$

$$= (\hat{h}d + \hat{b}d) \cdot \sum_{t=3}^{T} \sum_{s=2}^{t-1} P_f[\bar{y}_s - D_s - \cdots - D_{t-1} \geq 1 \text{ and } L(t) = s]$$

$$\leq (\hat{h}d + \hat{b}d) \cdot \sum_{t=3}^{T} \sum_{s=2}^{t-1} P_f[\bar{y}_s - D_s - \cdots - D_{t-1} \geq 1],$$  \hspace{1cm} (119)
in which the first inequality is attributable to the facts that \( \hat{y}_t \geq 0 \) and \( Q_f(z) - Q_f(y) \leq (\hat{h}d + \hat{b}d) \cdot 1(z \geq y + 1) \) for \( z \geq y \), which is easy to see from (4). But

\[
\mathbb{P}_f[\hat{y}_s - D_s - \cdots - D_{t-1} \geq 1] \leq \mathbb{P}_f[\hat{y}_s \geq 1] \wedge \sum_{d=1}^{\tilde{d}} \mathbb{P}_f[\hat{y}_s \geq d] \cdot \mathbb{P}_f[D_s + \cdots + D_{t-1} \leq d-1]. \tag{120}
\]

Meanwhile, by (12) and the current range of \( \gamma \),

\[
\mathbb{P}_f[\hat{y}_s \geq 1] = \mathbb{P}_f[\hat{F}_{s-1}(0) < \beta] \leq \mathbb{P}_f[\delta_V(f, \hat{f}_{s-1}) > 1 - \beta - \gamma], \tag{121}
\]

which, due to (20), is below \( 2 \cdot (\tilde{d} + 1) \cdot \exp(-2 \cdot (1 - \beta - \gamma)^2 \cdot (s - 1)) \). Thus,

\[
\mathbb{P}_f[\hat{y}_s \geq 1] \leq 2 \cdot (\tilde{d} + 1) \cdot \exp(-2 \cdot (1 - \beta - \gamma)^2 \cdot (s - 1)). \tag{122}
\]

Now for \( d = 1, 2, \ldots, \tilde{d} \), let \( \gamma_d = 1 - F_d(d - 1) \). Our setup is such that \( 0 \leq \gamma_d \leq \gamma_{d-1} \leq \cdots \leq \gamma_1 = \gamma < (1 - \beta)/2 \). Again due to (12),

\[
\mathbb{P}_f[\hat{y}_s \geq d] = \mathbb{P}_f[\hat{F}_{s-1}(d - 1) < \beta] = \mathbb{P}_f\left[\sum_{\tau=1}^{s-1} 1(D_{\tau} \geq d) > (1 - \beta) \cdot (s - 1)\right]. \tag{123}
\]

Note that \( 1(D_1 \geq d), 1(D_2 \geq d), \ldots, 1(D_{s-1} \geq d) \) are independent Bernouli random variables with mean \( \gamma_d \), and hence \( \sum_{\tau=1}^{s-1} 1(D_{\tau} \geq d) \) is a Binomial random variable with mean \( \gamma_d \cdot (s - 1) \). So by Markov’s inequality, the rightmost term in (123) is below

\[
\frac{\mathbb{E}_f[\sum_{\tau=1}^{s-1} 1(D_{\tau} \geq d)]}{(1 - \beta) \cdot (s - 1)} = \frac{\gamma_d}{1 - \beta}. \tag{124}
\]

Therefore,

\[
\mathbb{P}_f[\hat{y}_s \geq d] \leq \frac{\gamma_d}{1 - \beta}. \tag{125}
\]

Also, it is easy to see that

\[
\mathbb{P}_f[D_s + \cdots + D_{t-1} \leq d - 1] \leq (1 - \gamma_d)^{t-s}. \tag{126}
\]

Combining (120), (122), (125), and (126), we obtain

\[
\sum_{s=2}^{t-1} \mathbb{P}_f[\hat{y}_s - D_s - \cdots - D_{t-1} \geq 1] \leq \sum_{s=2}^{t-1} \{[2 \cdot (\tilde{d} + 1) \cdot \exp(-2 \cdot (1 - \beta - \gamma)^2 \cdot (s - 1))] \wedge \sum_{d=1}^{\tilde{d}} (\gamma_d/(1 - \beta)) \cdot (1 - \gamma_d)^{t-s}\}. \tag{127}
\]

Consider \( a(\gamma, \tau) \equiv 2 \cdot (\tilde{d} + 1) \cdot \exp(-2 \cdot (1 - \beta - \gamma)^2 \cdot (\tau - 1)) \). There exists a \( t_0 \geq 1 \) so that for any \( t \geq t_0 \),

\[
a\left(\frac{1 - \beta}{2}, t\right) = 2 \cdot (\tilde{d} + 1) \cdot \exp\left(-\frac{(1 - \beta)^2}{2} \cdot (t - 1)\right) < \exp\left(-\frac{(1 - \beta)^2 \cdot t}{4}\right). \tag{128}
\]
Note also that \( a(\gamma, s) < a((1 - \beta)/2, s) \) for \( \gamma \in (0, (1 - \beta)/2) \). Next, consider \( b(\gamma', \tau) \equiv \gamma' \cdot (1 - \gamma')^\tau \). Note that
\[
\frac{\partial b(\gamma', \tau)}{\partial \gamma'} = (1 - \gamma')^{\tau - 1} \cdot [1 - (\tau + 1) \cdot \gamma'],
\]
and
\[
\frac{\partial^2 b(\gamma', \tau)}{\partial (\gamma')^2} = -\tau \cdot (1 - \gamma')^{\tau - 2} \cdot [2 - (\tau + 1) \cdot \gamma'].
\]
So the \( b \)-maximizing \( \gamma' \) is \( \gamma^*_\tau = 1/(\tau + 1) \). Plugging back, we have
\[
b(\gamma^*_\tau, \tau) = \frac{1}{\tau + 1} \cdot \frac{1}{\tau + 1} = \frac{1}{\tau + 1} \cdot \frac{1}{(1 + 1/\tau)^\tau}.
\]
Note that \( \lim_{\tau \to +\infty} (1 + 1/\tau)^\tau = e \), the natural logarithmic base which is above 2. So when \( \tau \) is large enough, say greater than some \( t_1 \), the above will be below \( 1/(2\tau + 2) \).

For \( T \geq 2 \cdot (t_0 + t_1)^2 \), the upper bound in (127) is further bounded by a constant plus
\[
\sum_{t=2}^{T} \sum_{s=2}^{\lfloor t/2 \rfloor} \sum_{d=1}^{\bar{d}} \frac{1}{\sqrt{t}} (\gamma_d/(1 - \beta)) \cdot (1 - \gamma_d)^{t-s} + \sum_{s=\lfloor t/2 \rfloor+1}^{t-1} 2 \cdot (\bar{d} + 1) \cdot \exp(-2 \cdot (1 - \beta - \gamma)^2 \cdot (s - 1)),
\]
which, according to the above from (128) to (131), is below
\[
\sum_{t=2}^{T} \sum_{s=2}^{\lfloor t/2 \rfloor} \frac{\bar{d}}{1 - \beta} \cdot \frac{1}{2t - 2s + 2} + \sum_{s=\lfloor t/2 \rfloor+1}^{t-1} \exp\left(-\frac{(1 - \beta)^2 \cdot s}{4}\right).
\]
But this is smaller than
\[
\sum_{t=2}^{T} \frac{\bar{d}}{2 \cdot (1 - \beta) \cdot t^{1/2}} + \frac{4}{(1 - \beta)^2} \cdot \exp\left(-\frac{(1 - \beta)^2 \cdot t^{1/2}}{8}\right),
\]
which has a constant-plus-\( T^{1/2} \) bound. So, there exist positive constants \( E'' \) and \( F'' \) so that
\[
R_f^{T^2}(y) \leq E'' + F'' \cdot T^{1/2},
\]
for any \( \gamma \in (0, (1 - \beta)/2) \). Now between (118) and (135), only the former has to be used when \( T \) is made large enough. We therefore have the intended bound.

\[ \blacksquare \]

**B Proofs of Section 6**

**Proof of Proposition 3:** Fix some \( k = 1, 2, \ldots, \bar{k} \) and \( t = 1, 2, \ldots \). If \( N_{s,0}^k = N_{s-1,0}^k + 1 \) never occurred for \( s = 1, 2, \ldots, t \), we can conclude that \( N_{t,0}^k = 0 \) from the policy’s initialization.
Otherwise, let \( s = 1, 2, \ldots, t \) be the latest time for the update (42) to occur. Note this must have coincided with \( m_s = 0 \) and \( k_s = \kappa_{s-1}(1) = k \). To have triggered this in the policy, it must follow that \( N_{s-1}^k < (s/k)^\mu \). Thus,

\[
N_{t,0}^k = N_{s,0}^k = N_{s-1,0}^k + 1 \leq N_{s-1}^k + 1 < \left(\frac{s}{k}\right)^\mu + 1 \leq \left(\frac{t}{k}\right)^\mu + 1. \tag{136}
\]

For either case, we see that the desired inequality is valid. \[\blacksquare\]

**Proof of Proposition 4:** We first use induction to prove that, for \( t = 0, 1, \ldots \),

\[
N_t^{\kappa_t(k)} \geq \left(\frac{t + k}{k}\right)^\mu - 1, \quad \forall k = 1, 2, \ldots, \bar{k}. \tag{137}
\]

For \( t = 0 \), we have

\[
N_0^{\kappa_0(k)} = N_0^k = 0 = \left(\frac{\bar{k}}{k}\right)^\mu - 1 \geq \left(\frac{k}{k}\right)^\mu - 1, \quad \forall k = 1, 2, \ldots, \bar{k}, \tag{138}
\]

which is exactly (137) at \( t = 0 \). Now suppose (137) is true for \( t - 1 \). That is,

\[
N_{t-1}^{\kappa_{t-1}(k)} \geq \left(\frac{t - 1 + \bar{k}}{k}\right)^\mu - 1, \quad \forall k = 1, 2, \ldots, \bar{k}. \tag{139}
\]

To show that (137) is true, we discuss whether \( N_{t-1}^{\kappa_{t-1}(1)} < (t/\bar{k})^\mu \) and hence \( m_t = 0 \) and \( l_t = 1 \), or \( N_{t-1}^{\kappa_{t-1}(1)} \geq (t/\bar{k})^\mu \) and hence \( m_t = 1 \). For the former case, we have

\[
N_t^{\kappa_t(1)} = N_{t-1}^{\kappa_{t-1}(1)} + 1 \geq \left(\frac{t}{k}\right)^\mu \geq \left(\frac{t + \bar{k}}{k}\right)^\mu - 1, \tag{140}
\]

where the equality comes from (42), the first inequality comes from (139) when applied \( k = 1 \), and the second inequality is due to the fact that \( x^\mu + 1 \geq (x + 1)^\mu \). Now,

\[
N_t^{\kappa_t(k)} = N_{t-1}^{\kappa_{t-1}(k+1)} = N_{t-1}^{\kappa_{t-1}(k+1)} \geq \left(\frac{t + \bar{k}}{k}\right)^\mu - 1, \quad \forall k = 1, 2, \ldots, j_t - 1, \tag{141}
\]

where the first equality is due to (267), the second equality is due to (44), and the inequality is due to (139); for \( k = j_t \),

\[
N_t^{\kappa_t(j_t)} = N_{t-1}^{\kappa_{t-1}(1)} \geq \left(\frac{t + \bar{k}}{k}\right)^\mu - 1 \geq \left(\frac{t + j_t}{k}\right)^\mu - 1, \tag{142}
\]

where the equality is due to the assignment that \( \kappa_t(j_t) = \kappa_{t-1}(l_t) = \kappa_{t-1}(1) \) as carried out between (267) and (268), the first inequality comes from (140), and the second inequality comes from the fact that \( j_t \leq \bar{k} \); also, for \( k = j_t + 1, j_t + 2, \ldots, \bar{k} \),

\[
N_t^{\kappa_t(k)} = N_t^{\kappa_{t-1}(k)} \geq N_t^{\kappa_{t-1}(1)} \geq \left(\frac{t + \bar{k}}{k}\right)^\mu - 1 \geq \left(\frac{t + \bar{k}}{k}\right)^\mu - 1, \tag{143}
\]

49
where the equality comes from (268), the first inequality is due to the definition of \( j_t \) with respect to \( l_t = 1 \), the second inequality comes from (140), and the third inequality is due to the fact that \( k \leq \bar{k} \). For the latter case, for any \( k = 1, 2, ..., \bar{k} \),

\[
N^k_t \geq N^k_{t-1} \geq N^\kappa_{t-1}^{(1)} \geq \left( \frac{t}{k} \right)^\mu \geq \left( \frac{t + \bar{k}}{k} \right)^\mu - 1 \geq \left( \frac{t + \bar{k}}{\bar{k}} \right)^\mu - 1, \tag{144}
\]

where the first inequality comes from (42) to (44), the second inequality comes from (47) at \( t - 1 \), the third inequality stems from the definition of this case with \( m_t = 1 \), the fourth and last inequalities come from reasons already stated. When combining (141) to (144), we can derive that \( \kappa_t \) satisfies (137).

Combining (47) and (137) at \( t - 1 \), we have

\[
N^k_{t-1} \geq N^\kappa_{t-1}^{(1)} \geq \left( \frac{t}{\bar{k}} \right)^\mu - 1, \quad \forall k = 1, 2, ..., \bar{k}, \tag{145}
\]

which is the desired inequality.

**Proof of Proposition 5:** To bound \( T_1 \) at (54), note from Proposition 3 that

\[
N^k_{T,0} < \frac{1}{k^\mu} \cdot T^\mu + 1. \tag{146}
\]

Since \( \delta V^k_f \) is bounded by \( \max_{k=1}^{\bar{k}} p^k \cdot \bar{d} \), (54) will lead to

\[
T_1 \leq A''_k + B' \cdot T^\mu, \tag{147}
\]

for some positive \( \mu \)-independent constants \( A''_k \) and \( B^{Prop5} \).

To bound \( T_2 \) at (55), we make the simplifying assumption that \( k^*_f = 1 \) to ease presentation. Note \( m_t = 1 \) is not always true in period \( t \); also, for \( V^k_{f_t} \) to achieve the maximum among \( k = 1, 2, ..., \bar{k} \), it must happen that \( V^k_{f_t} \geq V^1_{f_t} \). So in view of (55),

\[
T_2 \leq \sum_{k=2}^{\bar{k}} \delta V^k_f \cdot \left\{ \left[ \sum_{i=1}^{T} \mathbb{P}_f \left[ V^k_{f_{t-1}} \geq V^1_{f_{t-1}} \right] \right] \right\} \wedge T. \tag{148}
\]

Now, note that

\[
\mathbb{P}_f \left[ V^k_{f_{t-1}} \geq V^1_{f_{t-1}} \right] \leq \mathbb{P}_f \left[ V^k_{f_{t-1}} \geq \frac{V^1_{f_{t-1}} + V^{k}_{f_{t-1}}}{2} \right] + \mathbb{P}_f \left[ \frac{V^1_{f_{t-1}} + V^{k}_{f_{t-1}}}{2} \geq V^1_{f_{t-1}} \right]. \tag{149}
\]

But by the definition of \( \delta V^k_f \) in (52),

\[
\mathbb{P}_f \left[ V^k_{f_{t-1}} \geq \frac{V^1_{f_{t-1}} + V^{k}_{f_{t-1}}}{2} \right] = \mathbb{P}_f \left[ V^k_{f_{t-1}} - V^{k}_{f_{t-1}} \geq \frac{\delta V^k_f}{2} \right], \tag{150}
\]
while
\[
\mathbb{P}_r \left[ \frac{V_{f_1}^k + V_{f_k}^k}{2} \geq V_{f_1}^l ight] = \mathbb{P}_r \left[ V_{f_1}^l - V_{f_1}^l \geq \frac{\delta V_k^k}{2} \right].
\] (151)

On the other hand, by (33) and (34), for \( l = 1, k, \)
\[
|V_{f_1}^l - V_{f_1}^l| \leq (\bar{p}^l - c) \cdot \sum_{d=0}^d d \cdot |f_{l-1}(d) - f^l(d)| + |Q_{f_{l-1}}(y_{f_{l-1}}^* - Q_{f_{l-1}}(y_{f*}^l) - Q_{f_{l-1}}(y_{f*}^l)|,
\] (152)
which, by (19), (90), and (91), is bounded by a constant, say \( 2 \bar{C} \), times \( \delta_V(f^l, \hat{f}_{l-1}^l) \). Putting (149) to (152) together, we obtain
\[
\mathbb{P}_r \left[ \delta_V(f^l, \hat{f}_{l-1}^l) \geq \varepsilon \right] \leq 2 \cdot (\bar{d} + 1) \cdot \exp(-2\varepsilon^2 \cdot N_{l-1}^l).
\] (153)

Combining this with Proposition 4, we can then obtain,
\[
\mathbb{P}_r \left[ \delta_V(f^l, \hat{f}_{l-1}^l) \geq \varepsilon \right] \leq 2 \cdot (\bar{d} + 1) \cdot \exp(-2(C'')^2 \cdot ((t/k)^\mu - 1) \cdot (\delta V_k^k)^2).
\] (154)
Plug this back into (153), and we get
\[
\mathbb{P}_r \left[ \delta_V(f^l, \hat{f}_{l-1}^l) \geq \varepsilon \right] \leq D'' \cdot \exp(-E''_k \cdot (\delta V_k^k)^2 \cdot t^\mu),
\] (155)
for some positive \( \mu \)-independent constants \( D'' \) and \( E''_k \). Therefore,
\[
\sum_{t=1}^T \mathbb{P}_r \left[ V_{f_1}^k \geq V_{f_1}^l \right] \leq D'' \cdot \sum_{t=1}^T \exp(-E''_k \cdot (\delta V_k^k)^2 \cdot t^\mu)
\leq D'' \cdot \int_0^{+\infty} \exp(-E''_k \cdot (\delta V_k^k)^2 \cdot t^\mu) \cdot dt
\leq \int_0^{+\infty} y^{1/\mu-1} \cdot \exp(-y) \cdot dy/(\mu \cdot (E''_k)^{1/\mu} \cdot (\delta V_k^k)^{2/\mu}) \leq F''/(\delta V_k^k)^{2/\mu},
\] (157)
for some positive \( \mu \)-independent constant \( F'' \). In (157), the first inequality follows from (156), the second inequality comes from \( \exp(-E''_k \cdot (\delta V_k^k)^2 \cdot t^\mu) \)'s positivity and decreasing trend in \( t \), the first equality is achieved by a change of variables, and the last equality is due to the boundedness of \( \int_0^{+\infty} y^{1/\mu-1} \cdot \exp(-y) \cdot dy \) for \( \mu \in [1/2, 1] \). Now since \( \delta V_k^k \geq \delta > 0 \), it along with (148) will result in
\[
T_2 \leq \sum_{k=2}^\bar{k} \delta V_k^k \cdot \sum_{t=1}^T \mathbb{P}_r \left[ V_{f_1}^k \geq V_{f_1}^l \right] \leq \frac{C''}{\delta^{2/\mu}} \leq C''_k \left( \frac{1}{\delta^2} \vee 1 \right),
\] (158)
for some positive $\mu$-independent constant $G''_k$, where the last inequality is due to $\delta > 0$ and $\mu \in [1/2, 1)$. This can translate into

$$T_2 \leq A_{\text{Prop5}}^\delta,$$

for some $\delta$-related constant $A_{\text{Prop5}}^\delta$ that satisfy $\lim_{\delta \to 0^+} A_{\text{Prop5}}^\delta = +\infty$.

To bound $T_3$ at (56), note from (33) and (34) that

$$V^k_{f_k} - V^k_f \left( \bar{p}^k, y^*_{f_{k-1}} \right) = Q_{f_k} \left( y^*_{f_{k-1}} \right) - Q_{f_k} (y^*_k).$$

So (56) can be written as

$$T_3 = \sum_{k=1}^k \mathbb{E}_f \left[ R^N_{k,1} (y) \right],$$

where each $R^N_{k,1} (y)$ as defined in (14), with $N_k$ here replacing $T$ there and $f^k$ here replacing $f$ there. Due to Proposition 1, we have

$$T_3 \leq H''_k + I''_k \cdot T^1/2 \cdot (\ln T)^{1/2},$$

(162) for some positive constants $H''_k$ and $I''_k$.

Combining (53), (147), (159), and (162), we can obtain

$$R^T_{1} (k, y) \leq A_{\text{Prop5}}^\delta + B_{\text{Prop5}} \cdot T^\mu + C_{\text{Prop5}} \cdot T^{1/2} \cdot \ln T,$$

for some positive constants $A_{\text{Prop5}}^\delta$, $B_{\text{Prop5}}$, and $C_{\text{Prop5}}$. When $\mu = 1/2$, the term involving $B_{\text{Prop5}}$ is also not necessary, and the regression is at its lowest growth rate.

**Proof of Proposition 6:** We still have (147) for bounding $T_1$. To bound $T_2$, we can still resort to (157). In view of (148), we will have

$$T_2 \leq \sum_{k=2}^k \left\{ \left( \frac{F''}{(\delta V^k_f)^{2/\mu - 1}} \right) \wedge (\delta V^k_f \cdot T) \right\} \leq (\bar{k} - 1) \cdot (F'' \cdot T^{1-\mu/2} \leq D'' \cdot T^{1-1/2},$$

(164) for some positive $\mu$-independent constant $D''$. The first inequality in (164) comes from the fact that the maximum of $g(\delta V^k_f) \equiv [F'' / ((\delta V^k_f)^{2/\mu - 1})] \wedge (\delta V^k_f \cdot T)$ is achieved at $\delta V^k_f = (F'' / T)^{\mu/2}$, at which time $g(\delta V^k_f) = (F'' \cdot T^{1-\mu/2})$. On the other hand, we can still use (162) to bound $T_3$.

Combining (53), (147), (162), and (164), while noting that $\mu \vee (1 - \mu/2) > 1/2$ for $\mu \in [1/2, 1)$, we can obtain

$$R^T_{1} (k, y) \leq A_{\text{Prop6}} + B_{\text{Prop6}} \cdot T^{\mu \vee (1-\mu/2)},$$

(165)
for some positive \( \mu \)-independent constants \( A_{\text{Prop6}} \) and \( B_{\text{Prop6}} \). The choice for \( \mu \in [1/2, 1) \) that ensures the slowest guaranteed growth rate for \( R_{T_\delta}^{T_1}(k, y) \) is certainly 2/3.

## C Proofs of Section 7

**Proof of Proposition 7:** Note that \( \Pr\left[ \max_{k=2}^{\tilde{k}} V_{f_{t-1}}^k + \delta V_t^*/2 \geq V_{f_{t-1}}^1 | M(t) = m \right] \) is below

\[
\Pr\left[ V_{f_1}^1 - \frac{\delta V_t^*}{4} \geq V_{f_{t-1}}^1 | M(t) = m \right] + \Pr\left[ \max_{k=2}^{\tilde{k}} V_{f_{t-1}}^k \geq V_{f_1}^1 - \frac{3\delta V_t^*}{4} | M(t) = m \right],
\]

which is further below

\[
\Pr\left[ V_{f_1}^1 - V_{f_{t-1}}^1 \geq \frac{\delta V_t^*}{4} | M(t) = m \right] + \sum_{k=2}^{\tilde{k}} \Pr\left[ V_{f_{t-1}}^k \geq V_{f_1}^1 - \frac{3\delta V_t^*}{4} | M(t) = m \right].
\]

But by the definition of \( \delta V_t^k \) in (52) and that of \( \delta V_t^* \) in (57), for \( k = 2, ..., \tilde{k} \),

\[
\Pr[V_{f_{t-1}}^k \geq V_{f_1}^1 - 3\delta V_t^*/4 | M(t) = m] \leq \Pr[V_{f_{t-1}}^k \geq V_{f_1}^1 - 3\delta V_t^k/4 | M(t) = m] = \Pr[V_{f_{t-1}}^k - V_{f_1}^k \geq \delta V_t^k/4 | M(t) = m].
\]

So by (152) in the proof of Proposition 5, the earlier probability is below

\[
\Pr[\delta V(f_1, f_{t-1}) \geq C'' \cdot \delta V_t^* | M(t) = m] + \sum_{k=2}^{\tilde{k}} \Pr[\delta V(f^k, f_{t-1}) \geq C'' \cdot \delta V_t^k | M(t) = m],
\]

for some positive constant \( C'' \). Following (154) and (155) in the proof of Proposition 5, especially noting the validity of Proposition 4 under any realization \( m \in M(t) \) for \( M(t) \), we can then show the probability to be below

\[
D'' \cdot \exp\left( -E_{\text{Prop7}} \cdot (\delta V_t^*)^2 \cdot t^\mu \right) + \sum_{k=2}^{\tilde{k}} D'' \cdot \exp\left( -E_{\text{Prop7}} \cdot (\delta V_t^k)^2 \cdot t^\mu \right),
\]

for some constants \( D'' \) and \( E_{\text{Prop7}} \). But this is below \( \tilde{k}D'' \cdot \exp\left( -E_{\text{Prop7}} \cdot (\delta V_t^*)^2 \cdot t^\mu \right). \]

**Proof of Proposition 8:** From arguments right after (152) in the proof of Proposition 5, we know there is a positive constant \( C'' \), say \( 2\tilde{d} \cdot (\tilde{p}\tilde{k} - \tilde{c} + 3 \cdot (\tilde{h} \lor \tilde{d}) \), so that

\[
V_{f_{t-1}}^1 \geq V_{f_{t-1}}^1 - C'' \cdot \delta V(f_{t-1}, f_{t-1}).
\]
So when $E_{tt'}$ is true, we will have $k_t = k_{t+1} = \cdots = k_{t+t'-1} = 1$ as long as
\[
\delta V(\hat{f}_{t-1}, \hat{f}_{t+t'-1}) = \frac{\delta V^*_t}{2C''}, \quad \forall \tau = 1, \ldots, t'.
\] (172)

By (19), this can be achieved if
\[
|\hat{f}_{t-1}(d) - \hat{f}_{t+t'-1}(d)| \leq A^{\text{Prop8}} \cdot \delta V^*_t, \quad \forall d = 0, 1, \ldots, \bar{d}, \tau = 1, \ldots, t',
\] (173)
for some other constant $A^{\text{Prop8}}$, say \(1/(4\bar{d}^2 \cdot (\bar{p}^k - \bar{c} + 3 \cdot (\bar{h} \vee \bar{d})))\). But in view of (41),
\[
\mathcal{N}_{t-1}^1 \cdot \hat{f}_{t-1}(d) \leq \mathcal{N}_{t+t'-1}^1 \cdot \hat{f}_{t+t'-1}(d) = (\mathcal{N}_{t-1}^1 + \tau) \cdot \hat{f}_{t+t'-1}(d) \leq \mathcal{N}_{t-1}^1 \cdot \hat{f}_{t-1}(d) + \tau,
\] (174)
and hence
\[
-\frac{\tau}{\mathcal{N}_{t-1}^1} \leq (1 - \hat{f}_{t+t'-1}(d)) \leq \frac{\tau}{\mathcal{N}_{t-1}^1} \cdot \hat{f}_{t-1}(d) - \hat{f}_{t+t'-1}(d) \leq \frac{\tau}{\mathcal{N}_{t-1}^1} \cdot \hat{f}_{t+t'-1}(d) \leq \frac{\tau}{\mathcal{N}_{t-1}^1}. \tag{175}
\]
Thus, (173) would be true if $\tau \leq A^{\text{Prop8}} \cdot \delta V^*_t \cdot \mathcal{N}_{t-1}^1$. Due to Proposition 4, this in turn can be guaranteed when $\tau \leq A^{\text{Prop8}} \cdot \delta V^*_t \cdot ((\bar{t}/\bar{k})^\mu - 1)$.

**Proof of Proposition 9:** Since $s'_i \geq (i + I_{\mu, \delta})^{1/\mu}/G_{\mu, \delta}^{1/\mu}$, (64) will lead to
\[
\frac{(\mathcal{N}_{t,0}^\prime + I_{\mu, \delta})^{1/\mu}}{G_{\mu, \delta}^{1/\mu}} \leq s'_{N_{t,0}^\prime} \leq t,
\] (176)
and hence
\[
\mathcal{N}_{t,0}^\prime \leq G_{\mu, \delta} \cdot t^\mu - I_{\mu, \delta}. \tag{177}
\]
Due to (60) and (177), as well as the facts that $I_{\mu, \delta} \geq \bar{k}$ and
\[
\mathcal{N}_{t,0}^\prime(m) \leq \mathcal{N}_{t,0}(m) + \mathcal{N}_{t,0}^\prime,
\] (178)
we can ensure (65); for instance, we could let $H_{\mu, \delta} = G_{\mu, \delta} + \bar{k}^{1-\mu}$.

For $t = s''_i(m)$ at any given $i$, (65) would lead to
\[
i = \mathcal{N}_{t,0}^\prime(m) \leq H_{\mu, \delta} \cdot t^\mu = H_{\mu, \delta} \cdot (s''_i(m))^\mu,
\] (179)
which is just (66). We have (67) because $s''_i(m) \leq s'_i$ and $s'_i \leq (i + I_{\mu, \delta})^{1/\mu}/G_{\mu, \delta}^{1/\mu} + 1$.

For any $x \in (0, 1)$, we have from Taylor expansion that
\[
(1 + x)^{1/\mu} = 1 + \left(\frac{1}{\mu}\right) \cdot x + \frac{1}{2} \cdot \left(\frac{1}{\mu}\right) \cdot \left(\frac{1}{\mu} - 1\right) \cdot x^2 + \sum_{k=1}^{+\infty} (-T_{1k} + T_{2k}),
\] (180)
where for $k = 1, 2, \ldots$,

$$T_{1k} = \left( \frac{1}{\mu} \right) \cdot \left( \frac{1}{\mu - 1} \right) \cdot \frac{1}{(2k + 1)!} \cdot \prod_{j=2}^{2k} \left( j - \frac{1}{\mu} \right) \cdot x^{2k + 1},$$

and

$$T_{2k} = \left( \frac{1}{\mu} \right) \cdot \left( \frac{1}{\mu - 1} \right) \cdot \frac{1}{(2k + 2)!} \cdot \prod_{j=2}^{2k + 1} \left( j - \frac{1}{\mu} \right) \cdot x^{2k + 2}.$$  

If $\mu = 1/2$, we have $T_{1k} = T_{2k} = 0$ throughout. Suppose $\mu \in (1/2, 1)$. Then,

$$\frac{T_{2k}}{T_{1k}} = \frac{2k + 1 - 1/\mu}{2k + 2} \cdot x < 1.$$  

Either way, we can conclude from (180) that

$$(1 + x)^{1/\mu} - 1 \leq \left( \frac{1}{\mu} \right) \cdot x + \frac{1}{2} \cdot \left( \frac{1}{\mu} \right) \cdot \left( \frac{1}{\mu - 1} \right) \cdot x^2 < \left( \frac{1}{\mu^2} \right) \cdot x \leq 4x.$$  

For $y > 1$, we then have

$$(y + 1)^{1/\mu} - y^{1/\mu} = y^{1/\mu} \cdot \left[ \left( 1 + \frac{1}{y} \right)^{1/\mu} - 1 \right] < 4 \cdot y^{1/\mu - 1}. $$

By the definition that $s'_{i} = [(i + I_{\mu, \delta})^{1/\mu}/G_{\mu, \delta}^{1/\mu}]$, it follows that

$$s'_{i+1} - s'_{i} \leq \left[ \frac{(i + I_{\mu, \delta} \cdot G_{\mu, \delta}^{1/\mu}) - (i + I_{\mu, \delta})^{1/\mu}}{G_{\mu, \delta}^{1/\mu}} \right] + 1 < \frac{4 \cdot (i + I_{\mu, \delta})^{1-\mu}}{G_{\mu, \delta}^{1/\mu}} + 1.$$  

Since $(i + I_{\mu, \delta})^{1/\mu}/G_{\mu, \delta}^{1/\mu} \leq s'_{i}$ and hence $i \leq G_{\mu, \delta} \cdot (s'_{i})^{\mu} - I_{\mu, \delta}$, the above would entail

$$s'_{i+1} - s'_{i} \leq \left( \frac{4}{G_{\mu, \delta}} \right) \cdot (s'_{i})^{1-\mu} + 1.$$  

For any $i$ with $s''_{i+1}(m) \geq s'_{i} + 1$, we can identify $j$ such that $s'_{j} \leq s''_{i}(m)$ and $s'_{j+1} \geq s''_{i+1}(m)$. So due to (187),

$$s''_{i+1}(m) - s''_{i}(m) \leq s'_{j+1} - s'_{j} \leq \left( \frac{4}{G_{\mu, \delta}} \right) \cdot (s'_{j})^{1-\mu} + 1,$$

and hence (68) as by choice, $s'_{j} \leq s''_{i}(m)$. Otherwise, we still have $s''_{i+1}(m) - s''_{i}(m) \leq s'_{i}$.  

**Proof of Proposition 10:** Due to (65) and the fact that $r_{j}^{1}(x)$ is bounded by a constant, (71) would lead to

$$T_{1} \leq A^{\nu} \cdot E_{t}[N_{T,0}^{\nu}] \leq B_{\mu, \delta}^{\nu} \cdot T^{\nu},$$

55
for some constants $A''$ and $B''_{\mu,\delta}$. Since $B''_{\mu,\delta} \geq A'' \cdot H_{\mu,\delta}$, we know $\lim_{\delta \to 0^+} B''_{\mu,\delta} = +\infty$.

In view of Proposition 7 as well as the facts that regrets are positive and each $r_f^\prime(x)$ is bounded by a constant times $t$, (74) will lead to

$$
\eta_2(m) \leq C'' \cdot \sum_{i=1}^{N''_{T,0}(m)} \exp\left(-E_{\text{Prop}}^{\mu,\delta} \cdot \delta^2 \cdot (s''_i(m) + 1)^\mu \cdot (s''_{i+1}(m) - s''_i(m) - 1)\right),
$$

(190)

where $C''$ is some new constant and $E_{\text{Prop}}^{\mu,\delta}$ is the one associated with Proposition 7. By (65) and (68), as well as the fact that $\mu \geq 1 - \mu$, we further have

$$
\eta_2(m) < \left(\frac{4C''}{G_{\mu,\delta}}\right) \cdot \sum_{i=1}^{[H_{\mu,\delta} \cdot T''/\mu]} \exp\left(-E_{\text{Prop}}^{\mu,\delta} \cdot \delta^2 \cdot (s''_i(m) + 1)^\mu \cdot (s''_i(m) + 1)^\mu\right).
$$

(191)

Note that $u(x) = x \cdot \exp(-E_{\text{Prop}}^{\mu,\delta} \cdot \delta^2 \cdot x)$ is decreasing in $x$ when the latter is large enough. So in view of (66),

$$
\eta_2(m) < \left(\frac{4C''}{G_{\mu,\delta}}\right) \cdot \left[\sum_{i=1}^{[H_{\mu,\delta} \cdot T''/\mu]} \exp\left(-E_{\text{Prop}}^{\mu,\delta} \cdot \delta^2 \cdot (i/H_{\mu,\delta}) \cdot (i/H_{\mu,\delta}) + D''_{\delta}\right)\right],
$$

(192)

where $D''_{\delta}$ is a constant that grows with $1/\delta$ which accounts for the occasion when $s''_i(m)$ is not large enough. Therefore,

$$
\eta_2(m) \leq \left(\frac{4C''}{G_{\mu,\delta}}\right) \cdot \left[\int_0^{[H_{\mu,\delta} \cdot T''/\mu]} \exp\left(-E_{\text{Prop}}^{\mu,\delta} \cdot \delta^2 \cdot x\right) \cdot x \cdot dx + F''_{\delta}\right],
$$

(193)

where $F''_{\delta}$ is another constant that has to grow with $1/\delta$. Since the integral is bounded by $\int_0^{+\infty} y \cdot \exp(-y) \cdot dy/[(E_{\text{Prop}}^{\mu,\delta})^2 \cdot \delta^4]$, there exists a constant $J''_{\mu,\delta}$ that grows with $1/\delta$, so that

$$
\eta_2(m) \leq J''_{\mu,\delta}.
$$

(194)

Meanwhile, (75) will lead to

$$
\zeta_2(m) \leq \sum_{i=1}^{N''_{T,0}(m)} \mathbb{E}_{\mathbf{r}} \left[\sum_{j=1}^{N_{x}(m)} \int_{F_{k,j}(m)}^{U_{i,j+1}(m) - U_{i,j}(m)} (X_{i,j}(m)) |V_{x_i}^1|_{f_{x_i}(m)} V_{x_i}^1 + \frac{k}{2} F_{x_i}^k + \frac{\delta V^*_{x_i}}{2}\right].
$$

(195)

By the way in which $G_{\mu,\delta}$ is specified in (63) and the facts that $\mu \geq 1 - \mu$ and that $\mathbf{f} \in \Delta_{\delta}$ and hence $\delta V^*_{x_i} \geq \delta$ as according to (58), we would have

$$
(4/G_{\mu,\delta}) \cdot (s''_i(m))^{1-\mu} \leq (A_{\text{Prop}}^{*8} \cdot \delta/\bar{k}^\mu) \cdot [(s''_i(m) + 1)^\mu - \bar{k}^\mu]
$$

$$
\leq (A_{\text{Prop}}^{*8} \cdot \delta V^*_f/\bar{k}^\mu) \cdot [(s''_i(m) + 1)^\mu - \bar{k}^\mu],
$$

(196)

56
as long as \( s_i''(m) \) is large enough. When this happens, (68) will then lead to
\[
s_{i+1}''(m) - s_i''(m) - 1 \leq A^{\text{Prop}8} \cdot \delta V_t^* \cdot \left( \frac{(s_i''(m) + 1)^\mu}{k^\mu} - 1 \right).
\]

Due to Proposition 8, we now have \( N_i(m) = 1 \) and \( K_{i,1}(m) = 1 \). Thus, (195) would lead to
\[
\zeta_2(m) \leq \sum_{i=1}^{N_{i,0}''(m)} \mathbb{E}_r \left[ r_i^{s_i''(m)-s_{i+1}''(m)-1} (X_{s_i''(m)+1}) | V_{f_i''(m)}^1 > \max_k V_{f_i''(m)}^k + \frac{\delta V_t^*}{2} \right] + K''_{\mu,\delta},
\]
where \( K''_{\mu,\delta} \) is a constant. But by (59), we further have, for \( w(t) = t^{1/2} \cdot (\ln t)^{3/2} \),
\[
\zeta_2(m) \leq \sum_{i=1}^{N_{i,0}''(m)} \left[ A^{\text{Prop}2} + B^{\text{Prop}2} \cdot w \left( s_{i+1}'(m) - s_i''(m) - 1 \right) \right] + K''_{\mu,\delta}.
\]

With the help of (65), we can come to
\[
\zeta_2(m) \leq K''_{\mu,\delta} + A^{\text{Prop}2} H_{\mu,\delta} \cdot T^\mu + B^{\text{Prop}2} \cdot \gamma_2(m),
\]
where
\[
\gamma_2(m) = \sum_{i=1}^{N_{i,0}''(m)} w \left( s_{i+1}'(m) - s_i''(m) - 1 \right).
\]

The key know lies in bounding \( \gamma_2(m) \).

Due to (65), (68), and the monotonicity of \( w(\cdot) \),
\[
\gamma_2(m) \leq \sum_{i=1}^{[H_{\mu,\delta} T^\mu]} w \left( \frac{4}{G_{\mu,\delta}} \cdot (s_i''(m))^{1-\mu} \right),
\]
which by (67) leads further to
\[
\gamma_2(m) \leq \sum_{i=1}^{[H_{\mu,\delta} T^\mu]} w \left( \frac{4}{G_{\mu,\delta}} \cdot \left( \frac{1}{G_{\mu,\delta}^{1/\mu}} \cdot (i + I_{\mu,\delta})^{1/\mu} + 1 \right)^{1-\mu} \right).
\]

When \( i \) is large, the term inside \( w(\cdot) \) will be below \( M_{\mu,\delta}^{\mu} \cdot i^{1/\mu - 1} \) for some \( M_{\mu,\delta}^{\mu} \geq 1 \). So
\[
\gamma_2(m) \leq N''_{\mu,\delta} + \sum_{i=1}^{[H_{\mu,\delta} T^\mu]} w(M_{\mu,\delta}^{\mu} \cdot i^{1/\mu - 1}) \leq N''_{\mu,\delta} + \int_1^{H_{\mu,\delta} T^{\mu+2}} w(M_{\mu,\delta}^{\mu} \cdot x^{1/\mu - 1}) \cdot dx,
\]
where \( N''_{\mu,\delta} \) is a constant which safeguards against the occasion when \( i \) is not yet large enough.

If we let \( y = \ln(M_{\mu,\delta}^{\mu} \cdot x^{1/\mu - 1}) \geq 0 \), the integral in (204) would become
\[
\frac{\mu}{(1-\mu) \cdot (M_{\mu,\delta}^{\mu})^{\mu/(1-\mu)}} \cdot \int_0^{(1 + \mu) \cdot y} \exp \left( \frac{(1 + \mu) \cdot y}{2 - 2\mu} \right) \cdot y^{3/2} \cdot dy.
\]

57
After letting $z$ equal $(1 + \mu) \cdot y/(2 - 2\mu)$, the above is equal to

$$O''_{\mu, \delta} \cdot \int_0^{(1+\mu)\ln(M''_{\mu, \delta} \cdot (H_{\mu, \delta} \cdot T^\mu + 2)^{1/\mu - 1})/(2 - 2\mu)} \exp(z) \cdot z^{3/2} \cdot dz,$$

for some constant $O''_{\mu, \delta}$. Through integration by parts, the above is further bounded by a term proportional to

$$u \left( \frac{(1 + \mu) \cdot \ln \left( M''_{\mu, \delta} \cdot (H_{\mu, \delta} \cdot T^\mu + 2)^{1/\mu - 1} \right)}{2 - 2\mu} \right),$$

for $u(z) = \exp(z) \cdot z^{5/2}$. Thus, for large enough constants $P''_{\mu, \delta}$ and $Q''_{\mu, \delta}$, we have

$$\gamma_2(m) \leq P''_{\mu, \delta} + Q''_{\mu, \delta} \cdot T^{(1+\mu)/2} \cdot (\ln T)^{5/2}. \quad (208)$$

In view of (200) and (208),

$$\zeta_2(m) \leq K''_{\mu, \delta} + B^{Prop10}_{\mu, \delta} P''_{\mu, \delta} + A^{Prop2}_{\mu, \delta} H_{\mu, \delta} \cdot T^\mu + B^{Prop2}_{\mu, \delta} Q''_{\mu, \delta} \cdot T^{(1+\mu)/2} \cdot (\ln T)^{5/2}. \quad (209)$$

Combining (70) to (75), (189), (194), (209), and the facts that $\sum_{m \in M(T)} \mathbb{P}_r[M(T) = m] = 1$, $\mu \geq 1/2$, and $\theta_2(m) = \eta_2(m) + \zeta_2(m)$, we can obtain the desired result for $R^T_{\mu}(k, y)$, with $A^{Prop10}_{\mu, \delta} = J''_{\mu, \delta} + K''_{\mu, \delta} + B^{Prop2}_{\mu, \delta} P''_{\mu, \delta}$, $B^{Prop10}_{\mu, \delta} = B''_{\mu, \delta} + A^{Prop2}_{\mu, \delta} H_{\mu, \delta}$, and $C^{Prop10}_{\mu, \delta} = B^{Prop2}_{\mu, \delta} Q''_{\mu, \delta}$.

Since both $B''_{\mu, \delta}$ and $H_{\mu, \delta}$ grow to $+\infty$ when $\delta$ approaches $0^+$, we know $\lim_{\delta \to 0^+} B^{Prop10}_{\mu, \delta} = +\infty$. From (207), we know it not necessary that $Q''_{\mu, \delta} > 2 \cdot (M''_{\mu, \delta})^{(1+\mu)/(2-2\mu)} \cdot H_{\mu, \delta}^{(1+\mu)/(2\mu)}$. Due to (203) and (204), it is also not required that $M''_{\mu, \delta} > 8/G_{\mu, \delta}^{1/\mu}$. Since $H_{\mu, \delta}$ can be made less than $2G_{\mu, \delta}$, these translate into it being possible that $C^{Prop10}_{\mu, \delta}$ be made less than a $\delta$-independent constant times $1/G_{\mu, \delta}^{(1+\mu)/(2-2\mu)}$. As $\lim_{\delta \to 0^+} G_{\mu, \delta} = +\infty$, it is possible that $\lim_{\delta \to 0^+} C^{Prop10}_{\mu, \delta} = 0$.

\section*{D Proofs of Section 8}

**Proof of Proposition 11:** Compared to the bound in Proposition 6, now there is just the new term from (78):

$$\theta_2 = \sum_{k \neq k^*} \sum_{t=1}^T \frac{\nu}{t^{\mu-\psi}} \cdot \mathbb{P}_r \left[ m_t = 1 \text{ and } V^k_{f_{t-1}} \text{ achieves the maximum} \right]. \quad (210)$$

Since $1/t^{\mu-\psi}$ is decreasing in $t$, it follows that

$$\theta_2 \leq \nu \cdot \int_0^T \frac{1}{t^{\mu-\psi}} \cdot dt = \nu \cdot \frac{T^{1-\mu+\psi}}{1-\mu+\psi}. \quad (211)$$
We can have the desired bound by combining (211) with (147), (162), and (164).

Proof of Proposition 12: Similarly to (171) in the proof of Proposition 8, we know there is the same constant $C''$ as used in the earlier proposition, so that

$$V_{j_{t+1}}^{k_t} \geq V_{j_{t+1}}^{k_t} - C'' \cdot \delta_V(j_{t+1}, \hat{j}_t^{k_t}).$$

(212)

According to step 1.2' of the new policy LwD$'(\mu, \nu, \psi)$, when $E'_{t+1}$ is true, we will have $k_t = k_{t+1} = \cdots = k_{t+t'-1}$ as long as

$$\delta_V(j_{t+1}, \hat{j}_t^{k_t}) \leq \frac{\nu}{2(C'' \cdot t^{\mu-\psi})} < \frac{\nu}{C'' \cdot (t + t' - 1)^{\mu-\psi}}, \quad \forall \tau = 1, \ldots, t'.

(213)

The second inequality comes from the large size of $t$ as specified in the proposition’s statement. The rest of the proof can follow that of Proposition 8 with $\nu/(t^{\mu-\psi})$ replacing $\delta V^*_t$.

Proof of Proposition 13: Since $s''_i \geq (i + I_{\mu,\nu,\psi})^{1/(1-\psi)} / G_{\mu,\nu}^{1/(1-\psi)}$, (64) will lead to

$$\frac{(N'_{t,0} + I_{\mu,\nu,\psi})^{1/(1-\psi)}}{G_{\mu,\nu}^{1/(1-\psi)}} \leq s''_{N'_{t,0}} \leq t,$$

(214)

and hence

$$N'_{t,0} \leq G_{\mu,\nu} \cdot t^{1-\psi} - I_{\mu,\nu,\psi}.$$

(215)

Due to (60), (178), and (215), as well as the fact that $I_{\mu,\nu,\psi} \geq \bar{k}$, we can ensure (79); for instance, we could let $H_{\mu,\nu} = G_{\mu,\nu} + \bar{k}^{1-\mu}$. By treating $1 - \psi$ here as $\mu$ in Proposition 9, we can obtain (80) and (81) as we did (67) and (68) in the proof of the earlier proposition.

Proof of Proposition 14: Due to (79) and the fact that $r^1_f(x)$ is bounded by a constant, (71) would lead to

$$T_1 \leq A'' \cdot \mathbb{E}T'_{T,0} \leq B''_{\mu,\nu} \cdot T^{\mu(1-\psi)},$$

(216)

for some constants $A''$ and $B''_{\mu,\nu}$.

By the way in which $G_{\mu,\nu}$ is specified in (76), we would have

$$\left(\frac{4}{G_{\mu,\nu}}\right) \cdot (s''_i(m))^{\psi} \leq \frac{A^{\text{Prop8}} \cdot \nu}{k^\mu} \cdot [(s''_i(m) + 1)^{\psi} - \bar{k}^{\mu}],$$

(217)

as long as $s''_i(m)$ is large enough. When this happens, (81) will then lead to

$$s''_{i+1}(m) - s''_i(m) - 1 \leq A^{\text{Prop8}} \cdot \nu \cdot \left(\frac{(s''_i(m) + 1)^{\psi}}{k^\mu} - 1\right).$$

(218)
Due to Proposition 12 and the fact that $k_{s_i^r(m)+1}$ achieves the maximum $V_{f_{i-1}^k}$ among $k = 1, \ldots, \bar{k}$, we now have $N_i(m) = 1$ and $K_{i,1}(m) = 1$. Thus, (73) would lead to

$$\theta_2(m) \leq \sum_{i=1}^{N'_{F,0}(m)} \text{Er} \left[ \sum_{i=1}^{N'_F(m)} \left( s''_{i+1}^{s_i^r(m)} - s''_{i}^{s_i^r(m)} - 1 \right) (X s''_{i+1}^{s_i^r(m)}(m+1)) + K'''_{\mu,\nu,\psi} \right], \quad (219)$$

where $K'''_{\mu,\nu,\psi}$ is a constant. But by (59), we further have, for $w(t) = t^{1/2} \cdot (\ln t)^{3/2}$,

$$\theta_2(m) \leq \sum_{i=1}^{N'_{F,0}(m)} \left[ A^{\text{Prop}^2} + B^{\text{Prop}^2} : w \left( s''_{i+1}^{s_i^r(m)} - s''_{i}^{s_i^r(m)} - 1 \right) \right] + K'''_{\mu,\nu,\psi}. \quad (220)$$

With the help of (79), we can come to

$$\theta_2(m) \leq K'''_{\mu,\nu,\psi} + A^{\text{Prop}^2} H_{\mu,\nu} \cdot T^{\nu(1-\psi)} + B^{\text{Prop}^2} \cdot \gamma_2(m), \quad (221)$$

where $\gamma_2(m)$ follows the same definition as (201).

Due to (79), (81), and the monotonicity of $w(\cdot)$,

$$\gamma_2(m) \leq \sum_{i=1}^{N'_{F,0}(m)} w \left( \frac{4}{G_{\mu,\nu}} \cdot (s''_{i}^{s_i^r(m)})^{\psi} \right), \quad (222)$$

which by (80) leads further to

$$\gamma_2(m) \leq \sum_{i=1}^{N'_{F,0}(m)} w \left( \frac{4}{G_{\mu,\nu}} \cdot \left( \frac{1}{G_{\mu,\nu}^{1/(1-\psi)}} \cdot (i + I_{\mu,\nu,\psi})^{1/(1-\psi)} + 1 \right)^{\psi} \right). \quad (223)$$

When $i$ is large, the term inside $w(\cdot)$ will be below $M'''_{\mu,\nu,\psi} \cdot i^{\psi/(1-\psi)}$ for some constant $M'''_{\mu,\nu,\psi} \geq 1$. Therefore, for some constant $N'''_{\mu,\nu,\psi}$,

$$\gamma_2(m) \leq N'''_{\mu,\nu,\psi} + \sum_{i=1}^{N'_{F,0}(m)} \left[ H_{\mu,\nu} \cdot T^{\nu(1-\psi)} \right] w(M'''_{\mu,\nu,\psi} \cdot i^{\psi/(1-\psi)}) \leq N'''_{\mu,\nu,\psi} + J_{1}\frac{1}{\psi} \cdot \frac{1}{M'''_{\mu,\nu,\psi}} \cdot \frac{1}{(H_{\mu,\nu} \cdot T^{\nu(1-\psi)} + 2)^{\psi/(1-\psi)}} \cdot dx. \quad (224)$$

If we let $y = \ln(M'''_{\mu,\nu,\psi} \cdot x^{\psi/(1-\psi)}) \geq 0$, the integral in (224) would become

$$\frac{1 - \psi}{\psi} \cdot \left( M'''_{\mu,\nu,\psi} \right)^{1/\psi-1} \cdot \int_{0}^{\ln(M'''_{\mu,\nu,\psi} \cdot (H_{\mu,\nu} \cdot T^{\nu(1-\psi)} + 2)^{\psi/(1-\psi)})} \exp \left( \frac{(2 - \psi) \cdot y}{2 \psi} \right) \cdot y^{3/2} \cdot dy. \quad (225)$$

After letting $z$ equal $(2 - \psi) \cdot y/(2 \psi)$, the above is equal to

$$O'''_{\mu,\nu,\psi} \cdot \int_{0}^{(2 - \psi) \cdot \ln(M'''_{\mu,\nu,\psi} \cdot (H_{\mu,\nu} \cdot T^{\nu(1-\psi)} + 2)^{\psi/(1-\psi)})/(2 \psi)} \exp(z) \cdot z^{3/2} \cdot dz, \quad (226)$$

60
for some constant $O''_{\mu,\nu,\psi}$. Through integration by parts, the above is further bounded by a term proportional to

$$u \left( \frac{(2 - \psi) \cdot \ln (M''_{\mu,\nu,\psi} \cdot (H_{\mu,\nu} \cdot T^{\mu\nu(1-\psi)} + 2)^{\psi/(1-\psi)})}{2^{\psi}} \right),$$

(227)

for $u(z) = \exp(z) \cdot z^{5/2}$. Thus, for large enough constants $P''_{\mu,\nu,\psi}$ and $Q''_{\mu,\nu,\psi}$, we have

$$\gamma_2(m) \leq P''_{\mu,\nu,\psi} + Q''_{\mu,\nu,\psi} \cdot T^{(2-\psi)(\mu\nu(1-\psi))/(2-2\psi)} \cdot \left( \ln T \right)^{5/2}.$$  

(228)

In view of (221) and (228),

$$\theta_2(m) \leq K''_{\mu,\nu,\psi} + B^{Prop}_{\mu,\nu,\psi} P''_{\mu,\nu,\psi} + A^{Prop}_{\mu,\nu,\psi} H_{\mu,\nu} \cdot T^{\mu\nu(1-\psi)} + B^{Prop}_{\mu,\nu,\psi} Q''_{\mu,\nu,\psi} \cdot T^{(2-\psi)(\mu\nu(1-\psi))/(2-2\psi)} \cdot \left( \ln T \right)^{5/2}.$$  

(229)

Combining (70) to (73), (216), (229), and the fact that $\sum_{m \in M(T)} P_{f}[M(T) = m] = 1$, we can obtain the desired result for $R_{T}^{2}(k, y)$, with $A^{Prop}_{\mu,\nu,\psi} = K''_{\mu,\nu,\psi} + B^{Prop}_{\mu,\nu,\psi}$, $B^{Prop}_{\mu,\nu,\psi} = B''_{\mu,\nu} + A^{Prop}_{\mu,\nu,\psi}$, and $C^{Prop}_{\mu,\nu,\psi} = B^{Prop}_{\mu,\nu,\psi}$. 

\[\blacksquare\]