Learning to Order for Inventory Systems with Lost Sales and Uncertain Supplies

Boxiao Chen
College of Business Administration, University of Illinois at Chicago, Chicago, IL 60607, bbchen@uic.edu

Jiashuo Jiang
Stern School of Business, New York University, New York, NY 10012, jj2398@stern.nyu.edu

Jiawei Zhang
Stern School of Business, New York University, New York, NY 10012, jz31@stern.nyu.edu

Zhengyuan Zhou
Stern School of Business, New York University, New York, NY 10012, zzhou@stern.nyu.edu

We consider a stochastic lost-sales inventory control system with lead time L over a planning horizon T. Supply is uncertain, and is a function of the order quantity (due to random yield/capacity, etc). We aim to minimize the T-period cost, a problem that is known to be computationally intractable even under known distributions of demand and supply. In this paper, we assume that both the demand and supply distributions are unknown and develop a computationally efficient online learning algorithm. We show that our algorithm achieves a regret (i.e. the performance gap between the cost of our algorithm and that of an optimal policy over T periods) of $O(L+\sqrt{T})$ when $L \geq \log(T)$. We do so by 1) showing our algorithm's cost is higher by at most $O(L+\sqrt{T})$ for any $L \geq 0$ compared to an optimal constant-order policy under complete information (a well-known and widely-used algorithm) and 2) leveraging its known performance guarantee from the existing literature. To the best of our knowledge, a finite-sample $O(\sqrt{T})$ (and polynomial in L) regret bound when benchmarked against an optimal policy is not known before in the online inventory control literature.

A key challenge in this learning problem is that both demand and supply data can be censored; hence only truncated values are observable. We circumvent this challenge by showing that the data generated under an order quantity q^2 allows us to simulate the performance of not only q^2 but also q^1 for all $q^1 < q^2$, a key observation to obtain sufficient information even under data censoring. By establishing a high probability coupling argument, we are able to evaluate and compare the performance of different order policies at their steady state within a finite time horizon. Since the problem lacks convexity, commonly used learning algorithms such as SGD and bisection cannot be applied, and instead, we develop an active elimination method that adaptively rules out suboptimal solutions.

Key words: lost sales, lead time, supply uncertainty, online learning, censored data

1. Introduction

Matching supply with demand is one of the key concepts in supply chain management. However, achieving this is not easy because of uncertainties in the system. One source of uncertainties stems from the randomness on the demand side. Consider an inventory control system, if demand realizes to be higher than the inventory level, there will be left-over inventories, and if demand realizes to be lower than the inventory level, some customer needs cannot be fulfilled. Based on the behavior of unsatisfied customers, typically the inventory system is classified as either backlogging or lost-sales. The lost-sales system is usually more relevant than the backlogging system, especially in the retailing setting where customers can easily go to a competitor when facing a stockout (Bijvank and Vis 2011). However, the lost-sales system is in general much harder to analyze than the backlogging counterpart, due to more complicated system dynamics, loss of convexity, etc. One classic model is the lost-sales inventory system with positive lead times, that is, it takes multiple periods for an order to arrive once it is placed. It is well-known that the optimal solution for this problem is computationally intractable, because the dimension of the state space of the underlying Markov decision process (MDP) equals the length of the lead time, which is also known as the curse of dimensionality (Bu et al. 2020). Therefore, instead of striving to solve for the optimal solution, researches shift their focus to developing effective heuristic policies (Levi et al. 2008, Huh et al. 2009b, Xin and Goldberg 2016, Bu et al. 2020).

Another source of uncertainties resides in the supply side as a result of the probabilistic nature of machine/capacity availability, yield, quality, and processing times (Yano and Lee 1995, Angkiriwang et al. 2014). During and post the COVID-19 pandemic, supply uncertainties became more prominent worldwide, and supply chain runners need to adjust their operational strategies to better respond to supply disruptions, material shortages, and long lead times (Raj et al. 2022). The two most studied models on supply uncertainties are the proportional random yield model and the random capacity model. The former specifies that only a random proportion of the ordering quantity is fulfilled (Henig and Gerchak 1990), and the latter states that the ordering quantity is capped by a random capacity value (Ciarallo et al. 1994). Models with random yield are in general very hard to optimize, therefore, several heuristics were introduced in the literature and their performances were analyzed. See Bu et al. (2020) for more discussions on challenges of the random yield and random capacity problems.

Uncertainties from both the demand and the supply side contribute significantly to the difficulties of managing supply chains. In this paper, we consider both uncertainties. We study a periodic inventory control model over a planning horizon of T periods. At the beginning of every period t, the company determines an ordering quantity q_t, of which only a portion $s(q_t, Z_t)$ will be fulfilled. The lead time is L, i.e., $s(q_t, Z_t)$ will be delivered after L periods. Here, $s(\cdot,\cdot)$ is referred to as the supply function, and Z_t, $t = 1, \ldots, T$, are independent and identically distributed (i.i.d.) random variables. As illustrated in Section 3.2, $s(q_t, Z_t)$ considered in this paper include the random yield function $s(q_t, Z_t) = q_t Z_t$, the random capacity function $s(q_t, Z_t) = \min\{q_t, Z_t\}$, and so on. Note that when $s(q_t, Z_t) = q_t$, this is the classic lost-sales inventory system with lead times L and deterministic supply, which is a special case of our models. Demand for period t, D_t, $t = 1, \ldots, T$, are i.i.d. random variables. Demand D_t is satisfied as much as possible by available inventories during period t. If demand is smaller than supply, left-over inventories will be carried over to the beginning of period $t + 1$, for which a per-unit holding cost h is charged. If demand realizes to the higher than supply, unfulfilled demand will be lost and a per-unit shortage penalty cost of b will be incurred. Both the distributions of D_t and Z_t are unknown, and the company needs to infer their information only from historical data. The company would like to learn the demand and supply distributions and at the same time minimize the T period total cost.

Even under complete information, that is, both the distributions of D_t and Z_t are known to the company, solving for the optimal solution is incredibly hard due to the presence of lost sales, lead times, and supply uncertainty. As noted earlier in the discussion, even when $s(q_t, Z_t) = q_t$, this problem is already computationally intractable. The only heuristic policy reported in the literature is the constant-order policy, which prescribes a constant quantity to be ordered every period independent of the starting state. Bu et al. (2020) proves that the performance gap between the optimal constant-order policy and the true optimal policy (1) decays exponentially fast in the lead time L, (2) converges to 0 when the penalty cost b is large. In this paper, we consider a situation where neither the demand nor the supply distribution is known, and we develop online learning algorithms to learn the optimal constant-order policy using historical data.

Our main contributions are summarized as below.

1. Approaching the optimal policy with provable convergence rates. We prove that the cost incurred by our learning algorithm is higher than that of the optimal constant-order policy by at most $O(L + \sqrt{T})$ for any $L \geq 0$. On the other hand, it is shown in Bu et al. (2020) that the cost of the optimal constant-order policy converges to the optimal policy exponentially fast in the lead time L, implying that the cost of our learning algorithm is higher than the *optimal policy* by at most $O(L + \sqrt{T})$ when $L \geq O(\log T)$. This is the first learning algorithm that provably approaches the optimal policy. In fact, even for the special case of lost-sales inventory system with lead times and deterministic supply, which has been extensively studied in the learning literature, there is no existing learning algorithm that approaches the optimal policy under any parameter regime. As will be discussed in Section 2, algorithms developed for this special case in Huh et al. (2009a), Zhang et al. (2020), Agrawal and Jia (2022) and Lyu et al. (2021) are designed to approach various heuristic policies, and the best regret convergence rate in terms of its dependence in L and

T is $O(L\sqrt{T})$, also benchmarked against certain heuristic policy. Our regret rate of $O(L+\sqrt{T})$ dominates the rate of $O(L\sqrt{T})$, not to mention that our regret also holds with supply uncertainty, and when $L \geq O(\log T)$, our regret is benchmarked against the optimal policy.

2. Learning demand and supply distributions using censored data. We learn distributions for both the demand D_t and the supply Z_t, departing from the existing learning literature that assumes the supply is deterministic. Both demand and supply data can be censored. Demand data is truncated by inventory levels, and the company can only observe the sales data instead of the true demand data. When supply has a random capacity, i.e., $s(q_t, Z_t) = \max\{q_t, Z_t\}$, the company can only observe the realized supply $s(q_t, Z_t)$ instead of the capacity value Z_t, i.e., the capacity data is truncated by the ordering decision. Because data is censored, it is not efficient to estimate the distributions of D_t and Z_t directly. We circumvent the data censoring issue using the following observation: for any two constant-order quantities $q^1 < q^2$, utilizing the censored supply and demand data generated under q^2, we can construct approximate pseudo-costs to evaluate the performance of not only q^2 but also q^1 (Observation 1, Lemma 1). Based on this critical observation, we develop simulation-based algorithms that simulate the performance of all order quantities after implementing the largest reasonable order quantity. This approach enables us to reduce the time spent on exploration and quickly focus on near optimal solutions.

3. Maximizing in an inventory system without convexity. In the literature of inventory control with online learning, a common property researchers rely on is that the objective function or part of it is convex in the decision variables. Based on convexity, some popular approaches can be applied, such as stochastic gradient decent (SGD) used in Huh and Rusmevichientong (2009), Huh et al. (2009a), Zhang et al. (2018, 2020), Chen and Shi (2019), and bisection used in Agrawal and Jia (2022), Chen and Shi (2019), Chen et al. (2020). However, convexity does not hold in our problem due to the complex structure of the random supply functions, therefore, commonly adopted approaches cannot be applied in our setting. Instead, we develop an adaptive elimination procedure that keeps eliminating suboptimal ordering values using historical censored data and maintains a shrinking active set. By letting the procedure to proceed in *exponentially increasing* time intervals, we show that values that remain in the active set will perform very close to the true optimal constant-order policy with high probability as more data accumulates.

4. Applying high probability coupling arguments to evaluate policy performance at steady state. In order to solve for the optimal constant-order policy, we need to evaluate and compare the performance of different ordering policys at their steady state. However, this task can be highly nontrivial, because it may take a long time for an MDP to converge to its steady state and we only have a finite number of periods. Using a stochastic coupling argument, we prove that the MDP reaches its steady state after only $O(\log T)$ periods (Lemma 3) with high probability. Based

on this result, we are able to adequately explore the inventory space and approach the optimal constant-order policy at a fast speed.

2. Literature Review

Supply chain management with supply uncertainties has been widely studied in the literature. See Yano and Lee (1995) and Feng and Shanthikumar (2018) for a detailed review of the area. As discussed earlier, one well-studied model is the proportional random yield model (Henig and Gerchak 1990, Kazaz 2004, Federgruen and Yang 2008, 2009, Li et al. 2013). Models with random yield are in general very hard to solve for the optimal solution, therefore, papers such as Bollapragada and Morton (1999), Huh and Nagarajan (2010), Inderfurth and Kiesmüller (2015) develop heuristics and demonstrate their performance. A special case of the proportional random yield model is the all-or-nothing supply model (Anupindi and Akella 1993, Tomlin 2006, Babich et al. 2007, Yang et al. 2009, Gümüş et al. 2012), based on which the random proportion takes either 0 or 1. Another commonly used model is the random capacity model. The random capacity model is first introduced in Ciarallo et al. (1994), then extended to more general settings in papers such as Wang and Gerchak (1996), Chao et al. (2008), Feng (2010) and Bu et al. (2020). All of the above-mentioned papers assume the distribution of the random supply is known, and none of them consider learning.

There is a stream of literature that studies the lost-sales inventory system with positive lead times and no supply uncertainties, that is $s(q_t, Z_t) = q_t$. Even for this simpler problem, as discussed earlier in the paper, the optimal solution cannot be directly solved for. Well-known heuristics developed for this problem include the base-stock policy (Janakiraman and Roundy 2004, Huh et al. 2009b), the capped base-stock policy (Xin 2021), and the constant-order policy. Reiman (2004) and Zipkin (2008) first put forth the constant-order policy for the continuous review and periodic review lost-sales inventory system with positive lead times, respectively, and show that it performs favorably compared with other heuristics. Goldberg et al. (2016) applies the constant-order policy to the finite horizon lost-sales model and proves that it is asymptoticly optimal with large lead times. Xin and Goldberg (2016) studies the constant-order policy for the infinite horizon lost-sales model and proves that the performance gap between it and the true optimal policy decays exponentially fast in the lead time. The constant-order policy is then generalized to the joint pricing and inventory control problem with lead times by Chen et al. (2019b) and to several MDP settings in Bai et al. (2020).

For the lost-sales inventory system with lead times and random supply functions that is considered in this paper, the constant-order policy is the only reported heuristic in the literature, and Bu et al. (2020) proves that its performance converges to that of the true optimal policy exponentially fast in the lead time and is asymptotically optimal with large lost-sales penalty cost. All the above results regarding the constant-order policy assume the demand distribution is known, and there exist no existing algorithms that learn the constant-order policy when the demand distribution is unknown.

The area of inventory control with online demand learning has been flourishing in recent years. However, all existing studies assume supply is deterministic and demand is the only source of uncertainties. Algorithms are then proposed to learn only the demand distribution, which cannot be directly applied to the case when supply also has uncertainties. A special case of our problem is the well-known lost-sales inventory system with positive lead times L, but with deterministic supply. Because this special case is already too complex to have tractable optimal solutions, researchers propose online learning algorithms to learn various heuristics. Huh et al. (2009a) propose a gradient based learning algorithm that converges to the optimal base-stock heuristic policy. Their results are then improved by the SGD based learning algorithm in Zhang et al. (2020) whose cost is proved to be higher than the optimal base-stock heuristic policy by at most $O(\exp(L)\sqrt{T})$. Agrawal and Jia (2022) develops a bisection based algorithm and further improves the result to $O(L\sqrt{T})$. By developing a UCB based learning algorithm for discrete demand, Lyu et al. (2021) proves that the cost of their algorithm is higher than the optimal capped base-stock heuristic policy by at most $O(L\sqrt{T})$. Different from the existing results, when $L \geq O(\log T)$, the regret of our algorithm, $O(L+\sqrt{T})$, is obtained by comparing to the *true optimal policy*. Other inventory control models with online demand learning include Huh and Rusmevichientong (2009) considering the lost-sales inventory system with zero lead time, Zhang et al. (2018) considering perishable products, Chen and Shi (2019) studying the dual sourcing inventory system, Yuan et al. (2021) exploring inventory control problems with fixed setup cost. These works all develop SGD based learning algorithms and achieve a regret rate of $O(\sqrt{T})$. Problems with joint pricing and inventory decisions are explored in Chen et al. (2019a) and Chen et al. (2021) with regret $O(\sqrt{T})$.

3. Problem Formulation

Consider a periodic-review lost-sales inventory system of a single product over a finite horizon of T periods. The demand at each period t is denoted by D_t, which belongs to the interval $[0, \bar{D}]$ and is assumed to be drawn independently from an unknown distribution $F(\cdot)$. At each period t, the company places an order of q_t, which will arrive after L (a positive integer) periods. We consider the case where the company may not receive exactly what it orders. To model the supply uncertainty, we introduce a supply function $s(q,z): \mathbb{R}^2 \to \mathbb{R}$, and we assume that the company at period t receives a quantity given by $s(q_{t-L}, Z_t)$, where $Z_1, \ldots, Z_T$ are i.i.d. non-negative random variables with a common distribution function $G(\cdot)$, which is assumed to belong to the interval $[\underline{\alpha}, \bar{\alpha}]$ and is assumed to be unknown. We also denote by h the per-unit holding cost and denote by

b the per-unit lost-sale penalty cost. We have the following sequence of events happening at each period t:

1. At the beginning of period t, we observe the on-hand inventory level denoted by I_t and all the inventories in pipeline ordered from the supplier, denoted by $(x_{1,t}, x_{2,t}, \ldots, x_{L,t})$ where $x_{i,t}$ is the order quantity placed at period $t-L+i-1$ for $i=1,\ldots,L$. The system state is $(I_t, x_{1,t}, x_{2,t}, \ldots, x_{L,t})$.
2. The inventory placed L periods ago arrives and the random variable Z_t is realized. Then, the on-hand inventory is increased to $I_t + s(x_{1,t}, Z_t)$.
3. The company placed an order with amount q_t that will arrive at the beginning of period $t+L$.
4. The demand D_t is realized and is satisfied as much as possible by the on-hand inventory. We assume that unsatisfied demand is lost and unobservable.

The objective of the company is to minimize the cumulative holding and penalty costs. The system state is updated as follows:

$$I_{t+1} = (I_t + s(x_{1,t}, Z_t) - D_t)^+, x_{i,t+1} = x_{i+1,t} \ \forall 1 \le i \le L-1 \text{ and } x_{L,t+1} = q_t$$

A policy π for the company is specified by the order quantities $q_1^\pi, \ldots, q_T^\pi$. Since the lost demand is assumed to be unobserved, we assume that only the *censored demand* is known by the company, i.e., the company can only observe the sales quantity $\min\{I_t + s(x_{1,t}, Z_t), D_t\}$ instead of the realization of D_t, and when $I_{t+1} = 0$, the company does not know the volume of lost sales. Note that the *supply data can be censored* as well, since only $s(x_{1,t}, Z_t)$ can be observed rather than Z_t, and $s(x_{1,t}, Z_t)$ may only contain truncated information about Z_t (see more discussions on this issue in Section 3.2). A policy π is *feasible* if and only if π is non-anticipative, i.e, for each t, q_t^π can only depend on the system state $(I_\tau^\pi, x_{1,\tau}^\pi, \ldots, x_{L,\tau}^\pi)$ for $\tau \le t$ and the realized values of supply $(s_1, \ldots, s_t)$. Note that the distribution functions $F(\cdot)$ and $G(\cdot)$ are assumed to be unknown by the company and need to be learned on-the-fly. Then, the cost incurred at period t for the policy π is denoted by

$$C_t^\pi = h \cdot (I_t + s(x_{1,t}, Z_t) - D_t)^+ + b \cdot (D_t - I_t - s(x_{1,t}, Z_t))^+$$

and the expected cumulative cost for the policy π is denoted by

$$C^\pi(T, L) = \sum_{t=1}^{T} \mathbb{E}[C_t^\pi] = \sum_{t=1}^{T} \mathbb{E}\left[h \cdot (I_t + s(x_{1,t}, Z_t) - D_t)^+ + b \cdot (D_t - I_t - s(x_{1,t}, Z_t))^+\right] \tag{1}$$

where T is used to indicate the dependency on the number of periods in the entire horizon and L is used to indicate the dependency on the lead time. Following this notation, the long-term average cost of the policy π is denoted by

$$C_\infty^\pi = \limsup_{T \to \infty} \frac{1}{T} \cdot C^\pi(T, L) \tag{2}$$

Following the standard conditions (Bu et al. 2020), we will assume that the initial inventory is $I_1 = 0$ and the initial pipeline is also 0, i.e., $x_{i,1} = 0$ for all $1 \le i \le L$.

3.1. Constant Order Policies and Notion of Regret

The optimal policy for minimizing the long-term reward in (2) is known to be very complex and computationally intractable due to the curse of dimensionality caused by the lead time L. Thus, heuristics have been developed to solve the problem approximately. In this section, we introduce the heuristics studied in this paper, namely the *constant order policies*, where the company places the same order in every period, regardless of the system state.

When the demand distribution and the supply function are unknown to the company, the optimal order quantity q^* for minimizing (2) cannot be directly computed. Our goal is to develop a feasible learning algorithm π. Using the optimal constant order policy π_{q^*} as the benchmark, we measure the performance of the learning algorithm π using the following notion of regret:

$$\text{Regret}_T^\pi = C^\pi(T,L) - T \cdot C_\infty^{\pi_{q^*}} \tag{3}$$

An alternative way to define regret of online policy π is to measure the additive difference between $C^\pi(T,L)$ and $C^{\pi_{q^*}}(T,L)$. We remark that for each policy π, the alternative regret will be at the same order of the regret defined in (3) by noting that the gap between $C^{\pi_{q^*}}(T,L)$ and $T \cdot C_\infty^{\pi_{q^*}}$ can be bounded by $O(\sqrt{T})$ following standard concentration inequality for Markov chain with stationary distributions (see Lemma 8).

3.2. Random Supply Function

In this paper, we consider the random supply function $s(q,Z)$ that takes one of the following four formulations:

1. $s(q,Z) = q \cdot Z$.
2. $s(q,Z) = \min\{q,Z\}$.
3. $s(q,Z) = qZ/(q+\alpha Z^\rho)$ for $\rho \le 1$ and $\alpha > 0$.
4. $s(q,Z) = (qk)/(q+Z)$, for some $k > 0$.

Formulation 1 and 2 covers the well-known random yield model and the random capacity model. Formulation 3 is introduced in Dada et al. (2007) to model a non-linear relationship between the order quantity and the supply, which covers an increasing concave relation of the output to the input over a wide range of parameters. Formulation 4 has been used in Cachon (2003), Tang and Kouvelis (2014) to study a situation where the supplier serves multiple firms and allocates the total output quantity, denoted by k, proportional to the firms' order quantities, denoted by q. Note that the firm is not able to observe the order quantities required by other firms, which is captured by the random variable Z.

The above four formulations have been studied in Feng and Shanthikumar (2018), which proves that all these four formulations are *stochasticlly linear in mid-point* (Definition 1 in Feng and

Shanthikumar (2018)). It has been shown in Bu et al. (2020) that the long-run average cost of the optimal policy converges to the long-run average cost of the optimal constant order policy as $L \to \infty$, with the gap decreases exponentially in the lead time L. This result justifies the efficiency of the constant order policy when the lead time L is large, and is also the reason why we set the optimal constant order policy as the benchmark in the definition of regret in (3), which we further discuss in Remark 2.

Note that in formulations 1, 3, and 4, after observing $s(q,Z)$, the value of Z can be inferred. However, this does not hold for formulation 2, where the value of Z is truncated by the ordering quantity q. That is, if Z realizes to be higher than q, then the company can only observe q. This data censoring issue for supply uncertainty creates extra challenges for estimating the distribution of Z.

The following observation plays a critical role in addressing the supply data censoring issue, and it is a key step to develop our learning algorithm (further explained in Section 4.1).

Observation 1 *If the random supply function takes one of the Formulation 1, 2, 3 and 4, then for any q and Z, as long as we observe the value of $s(q,Z)$, we know the value of $s(q',Z)$ for any $q' \leq q$.*

Clearly, for formulation 1, 2 and 4, this observation holds true by noting that the value of Z can actually be derived backward from the value of q and $s(q,Z)$. For formulation 3, $q = s(q,Z)$ implies $q \leq Z$. Then, for any $q' \leq q$, we must have $s(q',Z) = q'$. Also, $q > s(q,Z)$ implies $q > Z = s(q,Z)$. Then, for any $q' \leq q$, we have $s(q',Z) = \min\{q', s(q,Z)\}$. Therefore, we justify Observation 1 also holds for formulation 3.

Remark 1. Note that when the supply is deterministic, i.e., $s(q,Z) = q$, it is a special case of our models. This is the classic lost-sales inventory system with positive lead times that is extensively studied in the literature (Bijvank and Vis 2011).

4. Algorithm and General Description of Our Approach

In this section, we propose our learning algorithm to achieve the regret of optimal order. We being with re-formulating the long-run average cost of a constant order policy π_q. Note that in expression (1), the true value of D_t is unobservable due to lost sales and censored demand. However, we now show that in order to learn the optimal order quantity q^*, it is enough to focus solely on the on-hand inventory I_t and the supply $s(x_{1,t}, Z_t)$, which can be directly observed.

First, under the constant order policy π_q, the on-hand inventory is updated as follows:

$$I_{t+1}^{\pi_q} = (I_t^{\pi_q} + s(q, Z_t) - D_t)^+. \tag{4}$$

From queueing theory Asmussen (2008), the sequence $\{I_t^{\pi_q}\}_{t=1}^{\infty}$ converges in probability to a random variable $I_\infty^{\pi_q}$, which we refer to as the limiting inventory level under the constant order policy π_q, as long as the following condition is satisfied for the order quantity q:

$$\mathbb{E}_{Z\sim G}[s(q,Z)] < \mathbb{E}_{D\sim F}[D]. \tag{5}$$

If condition in (5) is not met, the system will explode and the on-hand inventory level will approach infinity. Therefore, we also impose the same condition for our analyses as shown in Assumption 1 below.

Assumption 1. *The company knows an upper bound of the optimal order quantity q^*, denoted by $\bar{q}$, that satisfies $\mathbb{E}[s(\bar{q},Z)] < \mathbb{E}[D]$.*

Assumption 1 is very mild, because any value not satisfying the condition in (5) will cause the system to explode, therefore those values can be easily detected as suboptimal.

We have

$$C_\infty^{\pi_q} = h\cdot\mathbb{E}[I_\infty^{\pi_q} + s(q,Z) - D]^+ + b\cdot\mathbb{E}[D - s(q,Z) - I_\infty^{\pi_q}]^+$$

Moreover, it holds that

$$I_\infty^{\pi_q} =^d (I_\infty^{\pi_q} + s(q,Z) - D)^+$$

where $=^d$ denotes identical in distribution. By taking expectation over both sides of the following equation,

$$I_\infty^{\pi_q} + s(q,Z) - D = [I_\infty^{\pi_q} + s(q,Z) - D]^+ - [D - s(q,Z) - I_\infty^{\pi_q}]^+,$$

we have that

$$\begin{aligned}\mathbb{E}[I_\infty^{\pi_q}] + \mathbb{E}[s(q,Z)] - \mathbb{E}[D] &= \mathbb{E}[I_\infty^{\pi_q} + s(q,Z) - D]^+ - \mathbb{E}[D - s(q,Z) - I_\infty^{\pi_q}]^+ \\ &= \mathbb{E}[I_\infty^{\pi_q}] - \mathbb{E}[D - s(q,Z) - I_\infty^{\pi_q}]^+\end{aligned}$$

which implies that

$$C_\infty^{\pi_q} = h\cdot\mathbb{E}[I_\infty^{\pi_q}] + b\cdot\mathbb{E}[D] - b\cdot\mathbb{E}[s(q,Z)]. \tag{6}$$

Note that $C_\infty^{\pi_q}$ is unobservable, because the term $\mathbb{E}[D]$ in (6) is unobservable due to demand censoring. However, the term $\mathbb{E}[D]$ is independent of the order quantity q. In order to obtain the optimal order quantity q^*, it is equivalent to minimize the *pseudo-cost* defined as follows:

$$\hat{C}_\infty^{\pi_q} = h\cdot\mathbb{E}[I_\infty^{\pi_q}] - b\cdot\mathbb{E}[s(q,Z)] \tag{7}$$

over the set $Q = \{q : q \le \bar{q}\}$. Here, the pseudo-cost $\hat{C}_\infty^{\pi_q}$ is observable. However, $\hat{C}_\infty^{\pi_q}$ is not convex in the order quantity q, in which case commonly used learning approaches such as SGD and bisection cannot be applied. For more discussions regarding this issue, see Section 1 point 3.

We now describe our learning algorithm for solving (7). Speaking at a high level, we transfer our problem into a multi-arm bandit problem by specifying $K+1$ points uniformly over the interval $[0,\bar{q}]$ i.e., we specify a set of points $\mathcal{A}=\{a_1,\ldots,a_{K+1}\}$ such that $a_k=\frac{k-1}{K}\cdot\bar{q}$ for any $k=1,\ldots,K+1$. Then, our algorithm proceeds in epochs $n=1,2,\ldots$ by maintaining an active set $\mathcal{A}_n\subset\mathcal{A}$ for each epoch n. The key element of our algorithm is to guarantee that for each epoch n and each point $a\in\mathcal{A}_n$, the gap between $\hat{C}_\infty^{\pi_a}$ and $\hat{C}_\infty^{\pi_{q^*}}$ is upper bounded by γ_n, where $\{\gamma_n\}_{n\geq 1}$ is a decreasing sequence to be determined later. To be specific, we let each epoch n contain $\max\{\frac{1}{\gamma_{n+1}^2}\cdot\log T, 3L\}$ number of time periods and the implementation of our algorithm at epoch n can be classified into the following three steps:

1. We implement the constant order policy $\pi_{a^{n*}}$, where a^{n*} is the largest element in the active set $\mathcal{A}_n$.
2. We use the censored demand to *simulate* the pseudo-cost of the policies π_a for each $a\in\mathcal{A}_n$ and we construct a confidence interval of $\hat{C}_\infty^{\pi_a}$ for each $a\in\mathcal{A}_n$ (simulation step further discussed in Section 4.1).
3. We use the constructed confidence intervals to identify $\mathcal{A}_{n+1}\subset\mathcal{A}_n$ such that for each element $a\in\mathcal{A}_{n+1}$, the gap between $\hat{C}_\infty^{\pi_a}$ and $\hat{C}_\infty^{\pi_{a^*}}$ is upper bounded by $(h+b)\cdot\gamma_n$, where $\hat{C}_\infty^{\pi_{a^*}}=\min_{a\in\mathcal{A}}\hat{C}_\infty^{\pi_a}$.

Following the steps outlined above, as our learning algorithm proceeds and n increases, the active set $\mathcal{A}_n$ shrinkages and the optimal order quantity q^* is gradually approximated. Our algorithm is formally described in Algorithm 1. Note that the implementation of Algorithm 1 depends on a fixed constant κ_2. We provide further discussion on how to select κ_2 in Section 4.2. By specifying the value of K and the sequence $\{\gamma_n\}_{n\geq 1}$, we are able to prove the following theorem regarding the regret upper bound of our algorithm, which is the main theorem of our paper.

Theorem 1. *Denote by π Algorithm 1 with input $K=\sqrt{T}$ and $\gamma_n=2^{-n}$ for each $n\geq 1$. Suppose that the random supply function takes one of the four formulations specified in Section 3.2. Then, under Assumption 1, the regret of π has the following upper bound:*

$$\textit{Regret}(\pi)\leq\kappa\cdot\kappa_2\cdot(L+\sqrt{T})\cdot\log T$$

where κ is a constant that is independent of L and T, and κ_2 is the constant used in Algorithm 1.

Remark 2. We note that Theorem 1 implies a regret bound of Algorithm 1 even compared to the *optimal policy*, when the lead time L is sufficiently large. To see this, we apply Theorem 1 in Bu et al. (2020) to show that $C_\infty^{\pi_{q^*}}-C_\infty^{\pi^*}\leq\kappa_3\cdot\gamma^L$, where κ_3 and $\gamma\in(0,1)$ are constants and π^* stands for the optimal policy, i.e. $\pi^*=\text{argmin}_\pi C_\infty^\pi$. Therefore, we have $C^\pi(T,L)-T\cdot C_\infty^{\pi^*}\leq\kappa\cdot\kappa_2\cdot(L+\sqrt{T})\cdot\log T+\kappa_3\cdot T\cdot\gamma^L$, which implies that $C^\pi(T,L)-T\cdot C_\infty^{\pi^*}\leq O(L+\sqrt{T})$ when $L\geq O(\log T)$. This is the *first* time that a sublinear regret bound is derived for an online policy

Algorithm 1 Learning-based Constant Order Policy

```
Input: $K$ and $\{\gamma_n\}_{n\geq 1}$.
Initialize $\mathcal{A}_1 = \mathcal{A}$, where $\mathcal{A} = \{a_1, \dots, a_{K+1}\}$ such that $a_k = \frac{k-1}{K} \cdot \bar{D}$ for any $k = 1, \dots, K+1$.
Set $\tau_n = \sum_{n'=1}^{n-1} \kappa_2 \cdot \max\{\frac{1}{\gamma_{n'+1}^2} \cdot \log T, 3L\} + 1$ as the start of epoch $n$ for each $n \geq 1$, where $\kappa_2$ is
  a fixed constant.
for epoch $n = 1, 2, \dots,$ do
    Identify $a^{n*}$ as the largest element in the active set $\mathcal{A}_n$.
    for time period $t = \tau_n$ to $\tau_{n+1} - 1$ do
        Implement the constant order policy $\pi_{a^{n*}}$.
        Observe the value of the supply $s(x_{1,t}, Z_t)$ and the on-hand inventory level $I_t$.
    end for
    For each $a \in \mathcal{A}_n$, we construct $\tilde{C}_n^a$ as follows:
      • obtain the simulated supply $s(a, Z_t)$ under policy $\pi_a$ for each $t = \tau_n + L, \dots, \tau_{n+1} - 1$;
      • starting from $I_{\tau_n+L}^a = I_{\tau_n+L}$, for $t = \tau_n + L, \dots, \tau_{n+1} - 1$, do the following:
        —if $I_{t+1} > 0$, then $I_{t+1}^a = (I_t^a + s(a, Z_t) + I_{t+1} - I_t - s(a^{n*}, Z_t))^+$;
        —if $I_{t+1} \leq 0$, then $I_{t+1}^a = 0$.
      • compute
          $\tilde{C}_n^a = h \cdot \frac{1}{\tau_{n+1} - \tau_n - \kappa_2 \max\{\log T, 2L\}} \cdot \sum_{t=\tau_n+\kappa_2 \max\{\log T, 2L\}}^{\tau_{n+1}-1} I_t^a$
          $\qquad - b \cdot \frac{1}{\tau_{n+1} - \tau_n - \kappa_2 \max\{\log T, 2L\}} \cdot \sum_{t=\tau_n+\kappa_2 \max\{\log T, 2L\}}^{\tau_{n+1}-1} s(a, Z_t).$
    Denote by $\tilde{C}_n^* = \min_{a \in \mathcal{A}_n} \tilde{C}_n^a$ and identify the active set for epoch $n+1$.
          $\mathcal{A}_{n+1} = \{a \in \mathcal{A}_n : \tilde{C}_n^a \leq \tilde{C}_n^* + (h+b) \cdot \frac{\gamma_n}{2}\}$        (8)
end for
```

with respect to the optimal policy. As a result, our result justifies the efficiency of constant order policies for inventory control systems with large lead time, under an online learning environment.

In the literature, the most related results on regret convergence rates are derived from Huh et al. (2009a), Zhang et al. (2020), Agrawal and Jia (2022) and Lyu et al. (2021), which study a special case of our problem with deterministic supply. The state-of-the-art regret convergence rate is $O(L\sqrt{T})$, derived in Agrawal and Jia (2022) for continuous demand benchmarked against the optimal base-stock heuristic policy and Lyu et al. (2021) for discrete demand benchmarked against the optimal capped base-stock heuristic policy. Our regret rate of $O(L + \sqrt{T})$ compares favorably with the existing results in this special case in terms of the dependence on L and T, and it is derived benchmarked against the optimal policy (instead of a heuristic policy) when $L \geq O(\log T)$.

REMARK 3. Algorithm 1 can be carried out efficiently. Note that for each quantity $a \in \mathcal{A}$, the inventory level I_t^a is simulated at most once for each period $t = 1, \ldots, T$ and $|\mathcal{A}| = \sqrt{T}$. Therefore, the overall computation complexity of Algorithm 1 is upper bounded by $O(T^{\frac{3}{2}})$.

4.1. Discussion on the Simulation Step

In this section, we discuss why we could use the censored demand of the constant order policy $\pi_{a^{n*}}$ to simulate the pseudo-cost of the policies π_a for each $a \in \mathcal{A}_n$, as outlined in step 10 in Algorithm 1. Following Observation 1, since a^{n*} is the largest element in the active set $\mathcal{A}_n$, after observing the value of $s(a^{n*}, Z_t)$, we know the value of $s(a, Z_t)$ for all $a \in \mathcal{A}_n$, for any $t = \tau_n + L, \ldots, \tau_{n+1} - 1$. Thus, we can use $s(a, Z_t)$ for any $t = \tau_n + L, \ldots, \tau_{n+1} - 1$ to approximate the term $\mathbb{E}[s(a, Z)]$ in the expression (7) for $\hat{C}_{\infty}^{\pi_a}$, for all $a \in \mathcal{A}_n$. Following Hoeffding's inequality, the approximation error can be bounded with a high probability (formalized in Section 5.3).

For any $a \in \mathcal{A}_n$, we can approximate $\mathbb{E}[I_{\infty}^{\pi_a}]$. We define a stochastic process $\{I_t^a\}_{t=\tau_n+L}^{\tau_{n+1}-1}$ revolving in the following way:

$$I_{\tau_n+L}^a = I_{\tau_n+L} \text{ and } I_{t+1}^a = (I_t^a + s(a, Z_t) - D_t)^+ \text{ for all } t = \tau_n + L, \ldots, \tau_{n+1} - 2 \tag{9}$$

Clearly, when the value of D_t is censored, we can not directly obtain the value of I_{t+1}^a. However, we now show that if the random supply function takes one of the four formulations specified in Section 3.2, we can use the on-hand inventory level I_t to derive the value of I_t^a, for any $t = \tau_n + L + 1, \ldots, \tau_{n+1} - 1$. Note that $\{I_t\}_{t=\tau_n+L}^{\tau_{n+1}-1}$ evolves in the following way:

$$I_{t+1} = (I_t - s(a^{n*}, Z_t) - D_t)^+. \tag{10}$$

Suppose that the value of I_t^a is known, we derive the value of I_{t+1}^a under the following two cases.

1. If $I_{t+1} > 0$, then from (10), we can obtain the value of D_t and we can derive the value of I_{t+1}^a directly following (9).

2. If $I_{t+1} \leq 0$, then we have $I_t \leq D_t - s(a^{n*}, Z_t)$. Note that $s(a, Z_\tau) \leq s(a^{n*}, Z_\tau)$ for all $\tau \leq t$, we must have $I_t^a \leq I_t$. Thus, we have

$$I_t^a \leq I_t \leq D_t - s(a^{n*}, Z_t) \leq D_t - s(a, Z_t)$$

which implies that $I_{t+1}^a = 0$.

The above two steps are formalized in the following lemma.

LEMMA 1. *Suppose that the stochastic process $\{I_t^a\}_{t=\tau_n+L}^{\tau_{n+1}-1}$ is defined in (9) and denote by $\{I_t\}_{t=\tau_n+L}^{\tau_{n+1}-1}$ the on-hand inventory level evolving in (10). Then, the value of I_t^a can be computed iteratively for $t = \tau_n + L, \ldots, \tau_{n+1} - 2$ in the following way:*

- *if* $I_{t+1} > 0$, *then* $I^a_{t+1} = (I^a_t + s(a, Z_t) + I_{t+1} - I_t - s(a^{n*}, Z_t))^+$;
- *if* $I_{t+1} \le 0$, *then* $I^a_{t+1} = 0$.

After deriving the value of $\{I^a_t\}_{t=\tau_n+L}^{\tau_{n+1}-1}$, we use this sequence to approximate $\mathbb{E}[I^{\pi_a}_\infty]$. The key is to establish the coupling between the stochastic process $\{I^a_t\}_{t=\tau_n+L}^{\tau_{n+1}-1}$ and another stochastic process, which we further explain in Section 5.3.

4.2. Discussion on the constant κ_2

Note that the implementation of Algorithm 1 depends on a fixed constant κ_2. We now discuss how should we select the value of κ_2.

In order for the regret bound in Theorem 1 to hold, a condition on the constant κ_2 would be $\kappa_2 \ge \delta(F, G, \bar{q})$, where $\delta(F, G, \bar{q})$ is a constant that depends solely on F, G and $\bar{q}$, and is indepedent of L and T. Though the value of $\delta(F, G, \bar{q})$ is unknown at the beginning since we assume the distributions F and G is unknown, we can simply set $\kappa_2 = \log T$ and the condition $\kappa_2 \ge \delta(F, G, \bar{q})$ will automatically be satisfied when T is large enough. Such an operation will only induce an additional multiplicative $\log T$ term into the final regret bound in Theorem 1. Another way is to spend the first $O(\sqrt{T})$ periods as a pure learning phase to learn the distributions F and G, and estimate an upper bound of $\delta(F, G, \bar{q})$, which is a constant independent of T and L. Such an operation will only induce an additional additive $O(\sqrt{T})$ term into the final regret bound in Theorem 1, which arises from the learning phase.

5. Proof of Regret Bound

In this section, we prove the regret bound in Theorem 1. Our analysis can be classified into the following four steps:

1) we establish the Lipschitz continuity of the pseudo-cost $\hat{C}^{\pi_q}_\infty$ over q. As a result, instead of comparing with $\hat{C}^{\pi_{q^*}}_\infty$, we can compare with $\hat{C}^{\pi_{a^*}}_\infty$ where $a^* = \text{argmin}_{a\in\mathcal{A}} \hat{C}^{\pi_a}_\infty$. We show that the additional regret term caused by this replacement of benchmark is at most $O(\sqrt{T})$.

2) we provide a bound over the gap between the actual pseudo-cost incurred at each epoch n and the long-term average $\hat{C}^{\pi_{a^{n*}}}_\infty$. The proof of the bound relies on a novel coupling argument between two stochastic process, which is explained in Section 5.2.

3) we denote by $\mathcal{E}$ the event that for each epoch n (except the last epoch), the pesudo-cost of each $a \in \mathcal{A}_n$ falls into the confidence interval $[\tilde{C}^a_n - (h+b)\cdot\frac{\gamma_n}{2}, \tilde{C}^a_n + (h+b)\cdot\frac{\gamma_n}{2}]$, i.e.,

$$\mathcal{E} = \{|\tilde{C}^a_n - \hat{C}^{\pi_a}_\infty| \le (h+b)\cdot\frac{\gamma_n}{2},\ \forall a \in \mathcal{A}_n, \forall 1 \le n \le N-1\} \tag{11}$$

where N denotes the total number of epochs. We show that event $\mathcal{E}$ occurs with a high probability.

4) we show how a^* is approximated by the revolution of the active set $\mathcal{A}_n$ in (8), which leads to our final regret bound.

Following the above four steps, we decompose the regret of our policy π as follows:

$$\begin{aligned}\text{Regret}(\pi) &= \sum_{n}\sum_{t=\tau_n}^{\tau_{n+1}}(h\cdot\mathbb{E}[I_t^{\pi}] - b\cdot\mathbb{E}[s(a^{n*},Z_t)] - \hat{C}_{\infty}^{\pi_{q^*}}) \\ &= \underbrace{\sum_{t=1}^{T}(\hat{C}_{\infty}^{\pi_{a^*}} - \hat{C}_{\infty}^{\pi_{q^*}})}_{I} + \underbrace{\sum_{n}\sum_{t=\tau_n}^{\tau_{n+1}}(h\cdot\mathbb{E}[I_t^{\pi}] - b\cdot\mathbb{E}[s(a^{n*},Z_t)] - \hat{C}_{\infty}^{\pi_{a^{n*}}})}_{II} \\ &\quad + \underbrace{\sum_{n}\sum_{t=\tau_n}^{\tau_{n+1}}(\hat{C}_{\infty}^{\pi_{a^{n*}}} - \hat{C}_{\infty}^{\pi_{a^*}})}_{III}\end{aligned} \tag{12}$$

We use the Lipschitz continuity established in the first step to bound the term I in (12). We use the high probability bound established in the second step to bound the term II in (12). We finally use step three and step four to bound the term III in (12).

5.1. Proof of Lipschitz Continuity

In this section, we establish the Lipschitz continuity of $\mathbb{E}[I_{\infty}^{\pi_q}]$ over order quantity q. We denote by $\hat{s}(\mu, Z) = s(q, Z)$ for q satisfying $\mathbb{E}[s(q, Z)] = \mu$. Our approach relies on existing result (Lemma 9) showing that if we interpret the psuedo-cost as a function over μ, then this function is a convex function, which implies Lipschitz continuity since μ belongs to a bounded region. Moreover, for the random supply function taking one of the four formulations specified in Section 3.2, one can check that if we interpret μ as a function of q, then this function is Lipschitz continuous. Therefore, we prove the Lipschitz continuity of $\mathbb{E}[I_{\infty}^{\pi_q}]$. We summarize our result in the following lemma, where the proof is relegated to Section B.

Lemma 2. *There exists a constant $\beta > 0$ such that for any $q_1, q_2 \in [0, \bar{q}]$, we have*

$$|\hat{C}_{\infty}^{\pi_{q_1}} - \hat{C}_{\infty}^{\pi_{q_2}}| \le \beta\cdot|q_1 - q_2|$$

5.2. Gap Between Actual Pseudo Cost and Long-term Average Pseudo Cost

We provide the bound over the gap between the actual pseudo cost incurred during each epoch n and the pseudo-cost $\hat{C}_{\infty}^{\pi_{a^{n*}}}$. Our proof relies on establishing the coupling between the stochastic process $\{I_t\}_{t=\tau_n}^{\tau_{n+1}-1}$ and the stochastic process defined as follows:

$$\tilde{I}_{\tau_n}^{a^{n*}} =^d I_{\infty}^{\pi_{a^{n*}}} \text{ and } \tilde{I}_{t+1}^{a^{n*}} = (\tilde{I}_t^{a^{n*}} + s(a^{n*}, Z_t) - D_t)^+ \text{ for } t = \tau_n, \ldots, \tau_{n+1} - 2 \tag{13}$$

It is clear to see that the distribution of $\tilde{I}_t^{a^{n*}}$ is identical to the distribution of $I_{\infty}^{\pi_{a^{n*}}}$, for each $t = \tau_n, \ldots, \tau_{n+1} - 1$. The coupling argument is formalized in the following lemma.

Lemma 3. *Denote by N the total number of epochs and denote by $\mathcal{B}$ the event that $I_{\tau_n} \leq \kappa_1 \cdot \log T$ and $\tilde{I}_{\tau_n}^{a^{n*}} \leq \kappa_1 \cdot \log T$, where $\kappa_1 > 0$ is a fixed constant, and $\{I_{\tau_n+\kappa_2 \cdot \max\{\log T, 2L\}} = \tilde{I}^{a^{n*}}_{\tau_n+\kappa_2 \cdot \max\{\log T, 2L\}}\}$, for every epoch $n \in [N]$, i.e.,*

$$\mathcal{B} = \{I_{\tau_n} \leq \kappa_1 \cdot \log T, \tilde{I}_{\tau_n}^{a^{n*}} \leq \kappa_1 \cdot \log T \text{ and } \{I_{\tau_n+\kappa_2 \cdot \max\{\log T, 2L\}} = \tilde{I}^{a^{n*}}_{\tau_n+\kappa_2 \cdot \max\{\log T, 2L\}}\}, \ \forall n\}.$$

Then, we have that

$$P(\mathcal{B}) \geq 1 - \frac{3N}{T^2}.$$

The proof is relegated to Section B.

From Lemma 3, we know that conditioning on the event $\mathcal{B}$, it holds that $I_t = \tilde{I}_t^{a^{n*}}$ for any $t = \tau_n + \kappa_2 \cdot \max\{\log T, 2L\}, \dots, \tau_{n+1} - 1$ and any epoch $n \in [N]$. Moreover, note that the distribution of $\tilde{I}_t^{a^{n*}}$ is identical to the distribution of $I_\infty^{a^{n*}}$. It holds that

$$\hat{C}_\infty^{\pi_{a^{n*}}} = h \cdot \mathbb{E}[\tilde{I}_t^{a^{n*}}] - b \cdot \mathbb{E}[s(a^{n*}, Z_t)], \quad \forall t = \tau_n, \dots, \tau_{n+1} - 1, \ \forall n \in [N] \tag{14}$$

As a result, the expected value of I_t for $t = \tau_n + \kappa_2 \cdot \max\{\log T, 2L\}, \dots, \tau_{n+1} - 1$ will be the same as the expected value of $I_\infty^{a^{n*}}$, which implies that the expected actual cost should be the same as the long-term average cost for $t = \tau_n + \kappa_2 \cdot \max\{\log T, 2L\}, \dots, \tau_{n+1} - 1$. Thus, we can obtain an upper bound over the gap between the actual pseudo cost and the long-term average pseudo cost $\hat{C}_\infty^{\pi_{a^{n*}}}$, for each epoch $n \in [N]$. By summing up the bound for each epoch $n \in [N]$, we get an upper bound of the term II in (12) for the entire horizon, which is formalized in the following lemma.

Lemma 4. *It holds that*

$$\sum_n \sum_{t=\tau_n}^{\tau_{n+1}} (h \cdot \mathbb{E}[I_t^\pi] - b \cdot \mathbb{E}[s(a^{n*}, Z_t)] - \hat{C}_\infty^{\pi_{a^{n*}}}) \leq hN \cdot \kappa_1 \kappa_2 \log T \cdot \max\{\log T, 2L\} + 3hN\bar{D}$$

The proof is relegated to Section B.

5.3. Probability Bound on the Event $\mathcal{E}$

We now show that the pesudo-cost of each $a \in \mathcal{A}_n$ at each epoch n falls into the confidence interval $[\tilde{C}_n^a - \gamma_n, \tilde{C}_n^a + \gamma_n]$ with a high probability and we provide a bound over the probability that event $\mathcal{E}$ happens. The key is to establish the stochastic coupling between the stochastic process $\{I_t^a\}_{t=\tau_n}^{\tau_{n+1}-1}$ defined in (9) and the stochastic process $\{\tilde{I}_t^a\}_{t=\tau_n}^{\tau_{n+1}-1}$ defined as follows:

$$\tilde{I}_{\tau_n}^a =^d I_\infty^{\pi_a} \text{ and } \tilde{I}_{t+1}^a = (\tilde{I}_t^a + s(a, Z_t) - D_t)^+ \text{ for } t = \tau_n, \dots, \tau_{n+1} - 2 \tag{15}$$

We formalize the coupling argument in the following lemma, which generalizes the stochastic coupling established in Lemma 3 from the implemented order quantity a^{n*} to all quantity $a \in \mathcal{A}_n$.

Lemma 5. *Denote by N the total number of epochs and denote by $\mathcal{C}$ the event that $I^a_{\tau_n} \leq \kappa_1 \cdot \log T$ and $\tilde{I}^a_{\tau_n} \leq \kappa_1 \cdot \log T$, where $\kappa_1 > 0$ is a fixed constant, and $\{I^a_{\tau_n+\kappa_2\cdot\max\{\log T, 2L\}} = \tilde{I}^a_{\tau_n+\kappa_2\cdot\max\{\log T, 2L\}}\}$, for every epoch $n \in [N]$ and every $a \in \mathcal{A}_n$, i.e.,*

$$\mathcal{C} = \{I^a_{\tau_n} \leq \kappa_1 \cdot \log T,\ \tilde{I}^a_{\tau_n} \leq \kappa_1 \cdot \log T \ \textit{and}\ \{I^a_{\tau_n+\kappa_2\cdot\max\{\log T, 2L\}} = \tilde{I}^a_{\tau_n+\kappa_2\cdot\max\{\log T, 2L\}}\},\ \forall n \in [N], \forall a \in \mathcal{A}_n\}$$

Then, we have that

$$P(\mathcal{C}) \geq 1 - \frac{3(K+1)N}{T^2}.$$

where K is given as the input of Algorithm 1 to denote $|\mathcal{A}|$.

The proof is relegated to Section B.

For each epoch n and each action $a \in \mathcal{A}_n$, it is clear to see that the distribution of $\tilde{I}^a_t$ is identical to the distribution of $I^{\pi_a}_\infty$. Therefore, we can use the average value of $\tilde{I}^a_t$ for $t = \tau_n + \kappa_2 \cdot \max\{\log T, 2L\}$ to $\tau_{n+1} - 1$ to approximate the value of $\mathbb{E}[I^{\pi_a}_\infty]$, where the length of the confidence interval can be given by γ_n. Further note that conditioning on the event $\mathcal{C}$ happens, the value of $\{I^a_t\}_{t=\tau_n+\kappa_2\cdot\max\{\log T, 2L\}}^{\tau_{n+1}-1}$ equals the value of $\{\tilde{I}^a_t\}_{t=\tau_n+\kappa_2\cdot\max\{\log T, 2L\}}^{\tau_{n+1}-1}$, which implies that $\hat{C}^{\pi_a}_\infty \in [\tilde{C}^a_n - \gamma_n, \tilde{C}^a_n + \gamma_n]$ with a high probability.

Lemma 6. *We have the following bound over the probability that event $\mathcal{E}$ happens, where event $\mathcal{E}$ is defined in (11),*

$$P(\mathcal{E}) \geq 1 - \frac{7(K+1)N}{T^2}.$$

The proof is relegated to Section B.

5.4. Proof of Theorem 1

We are now ready to prove our main theorem. Following (12), we have

$$\text{Regret}(\pi) = \underbrace{\sum_{t=1}^{T}(\hat{C}^{\pi_{a^*}}_\infty - \hat{C}^{\pi_{q^*}}_\infty)}_{I} + \underbrace{\sum_n \sum_{t=\tau_n}^{\tau_{n+1}} (h \cdot \mathbb{E}[I^\pi_t] - b \cdot \mathbb{E}[s(a^{n*}, Z_t)] - \hat{C}^{\pi_{a^{n*}}}_\infty)}_{II}$$
$$+ \underbrace{\sum_n \sum_{t=\tau_n}^{\tau_{n+1}} (\hat{C}^{\pi_{a^{n*}}}_\infty - \hat{C}^{\pi_{a^*}}_\infty)}_{III}$$

We use the Lipschitz continuity established in Lemma 2 to bound the term I. We denote by $a' \in \mathcal{A}$ the nearest one to q^*. Clearly, from the construction of the set $\mathcal{A}$, we know that $|q^* - a'| \leq \frac{\bar{q}}{2K}$. Therefore, from Lemma 2, we know that

$$\text{I} = T \cdot (\hat{C}^{\pi_{a^*}}_\infty - \hat{C}^{\pi_{q^*}}_\infty) \leq T \cdot (\hat{C}^{\pi_{a'}}_\infty - \hat{C}^{\pi_{q^*}}_\infty) \leq T \cdot \frac{\beta \bar{q}}{2K} = \frac{\beta \bar{q}\sqrt{T}}{2}. \tag{16}$$

where we note $K = \sqrt{T}$.

We now bound the term II. From Lemma 4, we know that

$$\text{II} = \sum_{n}\sum_{t=\tau_n}^{\tau_{n+1}} (h\cdot \mathbb{E}[I_t^\pi] - b\cdot \mathbb{E}[s(a^{n*}, Z_t)] - \hat{C}_\infty^{\pi_{a^{n*}}}) \le hN\cdot \kappa_1\kappa_2 \log T \cdot \max\{\log T, 2L\} + 3hN\bar{D} \quad (17)$$

We now proceed to bound the term III with the help of the probability bound established in Section 5.3.

We now assume that the event

$$\mathcal{E} = \{|\tilde{C}_n^a - \hat{C}_\infty^{\pi_a}| \le (h+b)\cdot \frac{\gamma_n}{2},\ \forall a \in \mathcal{A}_n, \forall 1 \le n \le N-1\}$$

happens. For each epoch n and each $a \in \mathcal{A}_{n+1}$, from (8) and the conditions of event $\mathcal{E}$, we have

$$\hat{C}_\infty^{\pi_a} - \hat{C}_\infty^{\pi_{a^*}} \le \tilde{C}_n^a - \tilde{C}_n^{a^*} + (h+b)\cdot\gamma_n \le \frac{3}{2}\cdot(h+b)\cdot\gamma_n.$$

Note that $a^{(n+1)*} \in \mathcal{A}_{n+1}$, we have that

$$\hat{C}_\infty^{\pi_{a^{(n+1)*}}} - \hat{C}_\infty^{\pi_{a^*}} \le \frac{3}{2}\cdot(h+b)\cdot\gamma_n.$$

which implies the following inequality conditional on the event $\mathcal{E}$ happens,

$$\text{III} = \sum_n \sum_{t=\tau_n}^{\tau_{n+1}} (\hat{C}_\infty^{\pi_{a^{n*}}} - \hat{C}_\infty^{\pi_{a^*}}) \le \frac{3(h+b)}{2}\cdot \sum_{n=1}^{N}\sum_{t=\tau_n}^{\tau_{n+1}} \gamma_{n-1}$$

Moreover, denote by N the total number of epochs. We have

$$\kappa_2 \cdot \sum_{n=1}^{N-1} \frac{1}{\gamma_n^2}\cdot \log T \le \sum_{n=1}^{N-1} \kappa_2 \cdot \max\{\frac{1}{\gamma_n^2}\cdot \log T, 3L\} \le T$$

which implies that

$$\sum_{n=1}^{N-1}\frac{1}{\gamma_n^2} \le \frac{T}{\kappa_2\cdot \log T}$$

By specifying $\gamma_n = 2^{-n}$, we have that $N \le \log_4 \frac{3T+\log T}{\kappa_2 \cdot \log T}$. Therefore, conditional on the event $\mathcal{E}$ happens, we have that

$$\begin{aligned}
\sum_{n=1}^{N}\sum_{t=\tau_n}^{\tau_{n+1}-1} \hat{C}_\infty^{a^{n*}} - \hat{C}_\infty^{a^*} &\le \frac{3(h+b)}{2}\cdot \sum_{n=1}^{N}\sum_{t=\tau_n}^{\tau_{n+1}-1}\gamma_{n-1} = 3(h+b)\cdot \sum_{n=1}^{N}\gamma_n \cdot \max\{\frac{1}{\gamma_n^2}\cdot \log T, 3L\} \quad (18)\\
&= 3(h+b)\cdot \sum_{n=1}^{\lfloor \log_4 \frac{3L}{\log T}\rfloor} \gamma_n \cdot 3L + 3(h+b)\cdot \sum_{n=\lfloor \log_4 \frac{3L}{\log T}\rfloor+1}^{N} \frac{\log T}{\gamma_n}\\
&\le 3(h+b)\cdot \sum_{n=1}^{\lfloor \log_4 \frac{3L}{\log T}\rfloor} \gamma_n \cdot 3L + 3(h+b)\cdot \sum_{n=1}^{N}\frac{\log T}{\gamma_n}\\
&\le 9(h+b)L + 3(h+b)(2^{N+1}-1)\log T\\
&\le 9(h+b)L + 6(h+b)\cdot \sqrt{\frac{(3T+\log T)\log T}{\kappa_2}}
\end{aligned}$$

If the event $\mathcal{E}$ does not happen, clearly, we have that

$$\mathrm{III} = \sum_{n} \sum_{t=\tau_n}^{\tau_{n+1}} (\hat{C}_{\infty}^{\pi_{a^{n*}}} - \hat{C}_{\infty}^{\pi_{a^*}}) \leq T \cdot (h+b) \cdot \bar{D}$$

where we note that $\hat{C}_{\infty}^{\pi_{a^{n*}}} \leq (h+b) \cdot \bar{D}$ for each n. Therefore, we have the following upper bound over the term III,

$$\begin{aligned} \mathrm{III} &= \mathbb{E}\left[\sum_{n} \sum_{t=\tau_n}^{\tau_{n+1}} (\hat{C}_{\infty}^{\pi_{a^{n*}}} - \hat{C}_{\infty}^{\pi_{a^*}}) \mid \mathcal{E} \right] \cdot P(\mathcal{E}) + \mathbb{E}\left[\sum_{n} \sum_{t=\tau_n}^{\tau_{n+1}} (\hat{C}_{\infty}^{\pi_{a^{n*}}} - \hat{C}_{\infty}^{\pi_{a^*}}) \mid \mathcal{E}^c \right] \cdot (1 - P(\mathcal{E})) \qquad (19) \\ &\leq \mathbb{E}\left[\sum_{n} \sum_{t=\tau_n}^{\tau_{n+1}} (\hat{C}_{\infty}^{\pi_{a^{n*}}} - \hat{C}_{\infty}^{\pi_{a^*}}) \mid \mathcal{E} \right] + T \cdot (h+b) \cdot \bar{D} \cdot (1 - P(\mathcal{E})) \\ &\leq 9(h+b)L + 6(h+b) \cdot \sqrt{\frac{(3T + \log T)\log T}{\kappa_2}} + \frac{7(K+1)N\bar{D}(h+b)}{T} \end{aligned}$$

where the last inequality follows from (18) and the probability bound on the event $\mathcal{E}$ from Lemma 6. Combining (16), (17), and (19), we have that

$$\mathrm{Regret}(\pi) = \mathrm{I} + \mathrm{II} + \mathrm{III} \leq \kappa \cdot \kappa_2 \cdot (L + \sqrt{T}) \cdot \log T$$

where κ is a constant that is independent of L and T. Therefore, our proof of our main result Theorem 1 is completed.

6. Conclusion

In this paper, we study the lost-sales inventory system with lead times L. Both demand and supply have uncertainties, for which the distributions are unknown. This departs from the existing literature on online learning that assume supply is deterministic and only considers demand uncertainty. The company needs to learn the demand and supply distributions from historical censored data. Demand censoring is caused by the fact that demand data is truncated by the inventory level, and supply censoring stems from the fact that capacity data is truncated by the ordering quantity. Because of demand and supply data censoring, it is not feasible to measure the performance of a policy directly, which requires the knowledge of the full demand and supply distributions. To circumvent this obstacle, we adopt a pseudo cost measure and prove that for any two constant-order quantities $q^1 < q^2$, using the censored data generated under q^2, we can simulate the pseudo cost for not only q^2 but also q^1. The critical observation enables us to significantly reduce the time spent on exploration. In order to evaluate the performance of a policy under stead state, we develop a high probability coupling argument to show that the MDP under our policy approaches its steady state within $O(\log T)$ periods. Note that the objective function of our problem lacks convexity, a property that is utilized by almost all existing papers in the literature of inventory control with learning. We propose an active elimination based algorithm to achieve a regret of $O(L+\sqrt{T})$ when compared with the optimal constant-order policy, and when $L \geq O(\log T)$, our algorithm approaches the optimal policy at the same rate.

There are many interesting directions for future research. For example, in the current setting, pricing is not considered. It would be a set of nice results if pricing can be included in the decision process and a learning algorithm can be developed accordingly. Another direction is to consider multiple products, where there exist substitution effects between different products and learning algorithms need to learn the substitution behavior of customers.

References

Agrawal, S. and Jia, R. (2022). Learning in structured mdps with convex cost functions: Improved regret bounds for inventory management. *Operations Research*.

Angkiriwang, R., Pujawan, I. N., and Santosa, B. (2014). Managing uncertainty through supply chain flexibility: reactive vs. proactive approaches. *Production & Manufacturing Research*, 2(1):50–70.

Anupindi, R. and Akella, R. (1993). Diversification under supply uncertainty. *Management science*, 39(8):944–963.

Asmussen, S. (2008). *Applied probability and queues*, volume 51. Springer Science & Business Media.

Babich, V., Burnetas, A. N., and Ritchken, P. H. (2007). Competition and diversification effects in supply chains with supplier default risk. *Manufacturing & Service Operations Management*, 9(2):123–146.

Bai, X., Chen, X., Li, M., and Stolyar, A. (2020). Asymptotic optimality of semi-open-loop policies in markov decision processes with large lead times. *Available at SSRN 3685551*.

Bijvank, M. and Vis, I. F. (2011). Lost-sales inventory theory: A review. *European Journal of Operational Research*, 215(1):1–13.

Bollapragada, S. and Morton, T. E. (1999). Myopic heuristics for the random yield problem. *Operations Research*, 47(5):713–722.

Bu, J., Gong, X., and Yao, D. (2020). Constant-order policies for lost-sales inventory models with random supply functions: Asymptotics and heuristic. *Operations Research*, 68(4):1063–1073.

Cachon, G. P. (2003). Supply chain coordination with contracts. *Handbooks in operations research and management science*, 11:227–339.

Chao, X., Chen, H., and Zheng, S. (2008). Joint replenishment and pricing decisions in inventory systems with stochastically dependent supply capacity. *European Journal of Operational Research*, 191(1):142–155.

Chen, B., Chao, X., and Ahn, H.-S. (2019a). Coordinating pricing and inventory replenishment with non-parametric demand learning. *Operations Research*, 67(4):1035–1052.

Chen, B., Chao, X., and Shi, C. (2021). Nonparametric learning algorithms for joint pricing and inventory control with lost sales and censored demand. *Mathematics of Operations Research*, 46(2):726–756.

Chen, B. and Shi, C. (2019). Tailored base-surge policies in dual-sourcing inventory systems with demand learning. *Available at SSRN 3456834*.

Chen, B., Wang, Y., and Zhou, Y. (2020). Optimal policies for dynamic pricing and inventory control with nonparametric censored demands. *Available at SSRN 3750413*.

Chen, X., Stolyar, A., and Xin, L. (2019b). Asymptotic optimality of constant-order policies in joint pricing and inventory control models. *Available at SSRN 3375203*.

Ciarallo, F. W., Akella, R., and Morton, T. E. (1994). A periodic review, production planning model with uncertain capacity and uncertain demand—optimality of extended myopic policies. *Management science*, 40(3):320–332.

Dada, M., Petruzzi, N. C., and Schwarz, L. B. (2007). A newsvendor's procurement problem when suppliers are unreliable. *Manufacturing & Service Operations Management*, 9(1):9–32.

Federgruen, A. and Yang, N. (2008). Selecting a portfolio of suppliers under demand and supply risks. *Operations research*, 56(4):916–936.

Federgruen, A. and Yang, N. (2009). Optimal supply diversification under general supply risks. *Operations Research*, 57(6):1451–1468.

Feng, Q. (2010). Integrating dynamic pricing and replenishment decisions under supply capacity uncertainty. *Management Science*, 56(12):2154–2172.

Feng, Q. and Shanthikumar, J. G. (2018). Supply and demand functions in inventory models. *Operations Research*, 66(1):77–91.

Goldberg, D. A., Katz-Rogozhnikov, D. A., Lu, Y., Sharma, M., and Squillante, M. S. (2016). Asymptotic optimality of constant-order policies for lost sales inventory models with large lead times. *Mathematics of Operations Research*, 41(3):898–913.

Gümüş, M., Ray, S., and Gurnani, H. (2012). Supply-side story: Risks, guarantees, competition, and information asymmetry. *Management Science*, 58(9):1694–1714.

Healy, A. D. (2008). Randomness-efficient sampling within nc1. *Computational Complexity*, 17(1):3–37.

Henig, M. and Gerchak, Y. (1990). The structure of periodic review policies in the presence of random yield. *Operations Research*, 38(4):634–643.

Huh, W. T., Janakiraman, G., Muckstadt, J. A., and Rusmevichientong, P. (2009a). An adaptive algorithm for finding the optimal base-stock policy in lost sales inventory systems with censored demand. *Mathematics of Operations Research*, 34(2):397–416.

Huh, W. T., Janakiraman, G., Muckstadt, J. A., and Rusmevichientong, P. (2009b). Asymptotic optimality of order-up-to policies in lost sales inventory systems. *Management Science*, 55(3):404–420.

Huh, W. T. and Nagarajan, M. (2010). Linear inflation rules for the random yield problem: Analysis and computations. *Operations research*, 58(1):244–251.

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

Inderfurth, K. and Kiesmüller, G. P. (2015). Exact and heuristic linear-inflation policies for an inventory model with random yield and arbitrary lead times. *European Journal of Operational Research*, 245(1):109–120.

Janakiraman, G. and Roundy, R. O. (2004). Lost-sales problems with stochastic lead times: Convexity results for base-stock policies. *Operations Research*, 52(5):795–803.

Kazaz, B. (2004). Production planning under yield and demand uncertainty with yield-dependent cost and price. *Manufacturing & Service Operations Management*, 6(3):209–224.

Levi, R., Janakiraman, G., and Nagarajan, M. (2008). A 2-approximation algorithm for stochastic inventory control models with lost sales. *Mathematics of Operations Research*, 33(2):351–374.

Li, T., Sethi, S. P., and Zhang, J. (2013). Supply diversification with responsive pricing. *Production and Operations Management*, 22(2):447–458.

Lyu, C., Zhang, H., and Xin, L. (2021). Ucb-type learning algorithms for lost-sales inventory models with lead times. *Available at SSRN 3944354*.

Raj, A., Mukherjee, A. A., de Sousa Jabbour, A. B. L., and Srivastava, S. K. (2022). Supply chain management during and post-covid-19 pandemic: Mitigation strategies and practical lessons learned. *Journal of business research*, 142:1125–1139.

Reiman, M. I. (2004). A new and simple policy for the continuous review lost sales inventory model. *Unpublished manuscript*.

Tang, S. Y. and Kouvelis, P. (2014). Pay-back-revenue-sharing contract in coordinating supply chains with random yield. *Production and Operations Management*, 23(12):2089–2102.

Tomlin, B. (2006). On the value of mitigation and contingency strategies for managing supply chain disruption risks. *Management science*, 52(5):639–657.

Wang, Y. and Gerchak, Y. (1996). Periodic review production models with variable capacity, random yield, and uncertain demand. *Management science*, 42(1):130–137.

Xin, L. (2021). Understanding the performance of capped base-stock policies in lost-sales inventory models. *Operations Research*, 69(1):61–70.

Xin, L. and Goldberg, D. A. (2016). Optimality gap of constant-order policies decays exponentially in the lead time for lost sales models. *Operations Research*, 64(6):1556–1565.

Yang, Z., Aydın, G., Babich, V., and Beil, D. R. (2009). Supply disruptions, asymmetric information, and a backup production option. *Management science*, 55(2):192–209.

Yano, C. A. and Lee, H. L. (1995). Lot sizing with random yields: A review. *Operations research*, 43(2):311–334.

Yuan, H., Luo, Q., and Shi, C. (2021). Marrying stochastic gradient descent with bandits: Learning algorithms for inventory systems with fixed costs. *Management Science*, 67(10):6089–6115.

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

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

Zipkin, P. (2008). Old and new methods for lost-sales inventory systems. *Operations research*, 56(5):1256–1263.

Appendix A: Useful Previous Results

We first state the well-known Hoeffdeing's inequality, which establishes concentration bound for i.i.d. random variables.

Lemma 7 **(Hoeffding's Inequality)**. *Let $X_1, \ldots, X_m$ be independent random variables such that $a_i \le X_i \le b_i$ almost surely for each $i \in [m]$. Then, denote by $S_n = \sum_{i=1}^m X_i$. It holds that*

$$P(S_n - \mathbb{E}[S_n] \ge t) \le \exp(-\frac{2t^2}{\sum_{i=1}^m (b_i - a_i)^2})$$

We then state a general concentration bound for Markov chain with stationary distributions from Healy (2008).

Lemma 8 **(Theorem 1.1 in Healy (2008))**. *Let $\boldsymbol{X} = (X_i, i \ge 1)$ be a Markov chain with a stationary distribution ϕ. Suppose that the distribution of X_1 is identical to the distribution of ϕ. Then, there exists a constant $\lambda > 0$ such that for any $\epsilon > 0$, it holds that*

$$P\left(\left|\sum_{i=1}^m X_i - \mathbb{E}\left[\sum_{i=1}^m X_i\right] \ge \sqrt{m} \cdot \epsilon\right|\right) \le 2 \exp\left(-\frac{\epsilon^2(1-\lambda)}{4}\right) \tag{20}$$

We also state the following lemma regarding the convexity of the pseudo-cost.

Lemma 9 **(Proposition 1 in Bu et al. (2020))**. *We denote by $\hat{s}(\mu, Z) = s(q(\mu), z)$ where $q(\mu) = \min_q\{q : \mathbb{E}[s(q, Z)] \ge \mu\}$. We also denote the transformed cost function $TC(\mu) = \hat{C}_\infty^{\pi_{q(\mu)}}$. Suppose that the random supply function takes one of the four formulations specified in Section 3.2. Then, $TC(\mu)$ is a convex function over $[0, \bar{\mu}]$, where $\bar{\mu}$ satisfying $q(\bar{\mu}) = \bar{q}$.*

We finally state the following result, showing how the limiting inventory level can be bounded.

Lemma 10 **(Lundberg's Inequality)**. *Denote by I_∞ as the limiting distribution of the stochastic process $I_{t+1} = (I_t + Q - D)^+$, where Q and D are two positive random variables. Then, there exists a constant ρ such that for any $a > 0$, we have*

$$P(I_\infty \ge a) \le \exp(-\rho a).$$

Moreover, ρ is the adjustment coefficient of the random variable $Q - D$, which is defined as the solution to $\lambda(z) = 1$, where $\lambda(z) = \mathbb{E}[\exp(z \cdot (Q - D))]$.

Appendix B: Missing Proofs

Proof of Lemma 2. From Lemma 9, we know that the transformed cost function $\mathrm{TC}(\mu) = \hat{C}_\infty^{\pi_{q(\mu)}}$ is a convex function over $\mu \in [0, \bar{\mu}]$, which is a bounded region. Thus, we know that there exists a constant $\beta' > 0$ such that

$$|\mathrm{TC}(\mu_1) - \mathrm{TC}(\mu_2)| \le \beta' \cdot |\mu_1 - \mu_2|, \ \forall \mu_1, \mu_2 \in [0, \bar{\mu}]. \tag{21}$$

For any $q_1, q_2 \in [0, \bar{q}]$, we now denote by $\mu_1 = \mathbb{E}[s(q_1, Z)]$ and $\mu_2 = \mathbb{E}[s(q_2, Z)]$. Moreover, for the random supply function taking one of the four formulations specified in Section 3.2, it is direct to check that there exists a constant $\alpha' > 0$ such that

$$|\mu_1 - \mu_2| \le \alpha' \cdot |q_1 - q_2| \tag{22}$$

Plugging (22) into (21), we know that

$$|\hat{C}_\infty^{\pi_{q_1}} - \hat{C}_\infty^{\pi_{q_2}}| = |\mathrm{TC}(\mu_1) - \mathrm{TC}(\mu_2)| \le \beta' \cdot |\mu_1 - \mu_2| \le \alpha'\beta' \cdot |q_1 - q_2|$$

Therefore, we prove that $\hat{C}_\infty^{\pi_q}$ is Lipschitz continuous over q with a Lipschitz constant $\beta = \alpha'\beta'$. □

Proof of Lemma 3. For each epoch n, we denote by

$$\mathcal{B}_n = \{I_{\tau_{n'}} \le \kappa_1 \cdot \log T,\ \tilde{I}^{a^{n*}}_{\tau_{n'}} \le \kappa_1 \cdot \log T \text{ and } I_{\tau_{n'}+\kappa_2\cdot\max\{\log T, 2L\}} = \tilde{I}^{a^{n'*}}_{\tau_{n'}+\kappa_2\cdot\max\{\log T,2L\}}\},\ \forall n' \le n\}.$$

In order to prove the lemma, it is sufficient to prove that

$$P(\mathcal{B}_n) \ge 1 - \frac{3n}{T^2}. \tag{23}$$

We prove (23) by using induction on the epoch n.

Clearly, when $n = 1$, we have that $P(I_{\tau_1} = 0) = 1$. From Lemma 10, there exists a constant $\kappa_1 > 0$ such that

$$P(\tilde{I}^{a^{1*}}_{\tau_1} \le \kappa_1 \cdot \log T) \ge 1 - \frac{1}{T^2}$$

by noting that the distribution of $\tilde{I}^{a^{1*}}_{\tau_1}$ is identical to the distribution of $I^{a^{1*}}_{\infty}$.

Now conditioning on the event $\{\tilde{I}^{a^{1*}}_{\tau_1} \le \kappa_1 \cdot \log T\}$, we proceed to bound the probability that event $\{I_{\tau_1+\kappa_2\cdot\max\{\log T,2L\}} = \tilde{I}^{a^{1*}}_{\tau_1+\kappa_2\cdot\max\{\log T,2L\}}\}$ happens. Note that the evolution of the stochastic process I_t in (10) is identical to the evolution of the stochastic process $\tilde{I}^{a^{1*}}_t$ in (13) for $t = \tau_1, \dots, \tau_2 - 1$. Therefore, it is clear to see that the event $\{I_{\tau_1+\kappa_2\cdot\max\{\log T,2L\}} = \tilde{I}^{a^{1*}}_{\tau_1+\kappa_2\cdot\max\{\log T,2L\}}\}$ happens as long as

$$I_t = \tilde{I}^{a^{1*}}_t = 0, \quad \text{for some } t = \tau_1, \dots, \tau_1 + \kappa_2 \cdot \max\{\log T, 2L\}. \tag{24}$$

We note that $I_{\tau_1} \le \tilde{I}^{a^{1*}}_{\tau_1}$ implies that $I_t \le \tilde{I}^{a^{1*}}_t$ for all $t = \tau_1, \dots, \tau_1 + \kappa_2 \cdot \max\{\log T, 2L\}$. From the non-negativity of I_t and $\tilde{I}^{a^{1*}}_t$, we have that (24) holds as long as $\tilde{I}^{a^{1*}}_t = 0$ for some $t = \tau_1, \dots, \tau_1 + \kappa_2 \cdot \max\{\log T, 2L\}$. As a result, a sufficient condition for (24) to hold is that

$$\sum_{t=\tau_1}^{\tau_1+\kappa_2\cdot\max\{\log T,2L\}} D_t - s(a^{1*}, Z_t) \ge \kappa_1 \cdot \log T$$

Note that $D_t - s(a^{1*}, Z_t)$ are i.i.d. random varibles for $t = \tau_1, \dots, \tau_1 + \kappa_2 \cdot \max\{\log T, 2L\}$. We also denote by $\delta = \mathbb{E}[D] - \mathbb{E}[s(\bar{q}, Z)]$. Following Hoeffding's inequality (Lemma 7), we have that

$$P\left(\sum_{t=\tau_1}^{\tau_1+\kappa_2\cdot\max\{\log T,2L\}} D_t - s(a^{1*}, Z_t) \ge \kappa_1 \cdot \log T\right) \ge 1 - \exp(-\frac{2(\delta\kappa_2 \max\{\log T, 2L\} - \kappa_1 \log T)^2}{\kappa_2 \cdot \max\{\log T, 2L\} \cdot \bar{D}}) \ge 1 - \frac{1}{T^2}$$

where $\kappa_2 \ge \max\{\frac{2\kappa_1}{\delta}, \frac{4\bar{D}}{\delta^2}\} \ge \max\{\frac{2\kappa_1 \log T}{\delta\cdot\max\{\log T,2L\}}, \frac{4\bar{D}\log T}{\delta^2\cdot\max\{\log T,2L\}}\}$.

The above derivation implies that

$$\begin{aligned}
P(\mathcal{B}_1) &= P\left(\mathcal{B}_1 \mid \tilde{I}^{a^{1*}}_{\tau_1} \le \kappa_1 \cdot \log T\right) \cdot P(\tilde{I}^{a^{1*}}_{\tau_1} \le \kappa_1 \cdot \log T) \\
&= P\left(\sum_{t=\tau_1}^{\tau_1+\kappa_2\cdot\max\{\log T,2L\}} D_t - s(a^{1*}, Z_t) \ge \kappa_1 \cdot \log T\right) \cdot P(\tilde{I}^{a^{1*}}_{\tau_1} \le \kappa_1 \cdot \log T) \\
&\ge (1 - \frac{1}{T^2}) \cdot (1 - \frac{1}{T^2}) \ge 1 - \frac{2}{T^2} \ge 1 - \frac{3}{T^2}
\end{aligned}$$

Therefore, we prove (23) for $n = 1$.

Now assume that (23) holds for epoch $n - 1$. We consider epoch n. Clearly, from the definition of the stochastic process $\tilde{I}^{a^{n*}}_t$ in (13), $\tilde{I}^{a^{n*}}_t$ refreshes when $t = \tau_n$. As a result, the distribution of $\tilde{I}^{a^{n*}}_{\tau_n}$ is independent of the event $\mathcal{B}_{n-1}$ and is identical to the distribution of $I^{a^{n*}}_{\infty}$, which implies that

$$P(\tilde{I}^{a^{n*}}_{\tau_n} \le \kappa_1 \cdot \log T \mid \mathcal{B}_{n-1}) = P(\tilde{I}^{a^{n*}}_{\tau_n} \le \kappa_1 \cdot \log T) \ge 1 - \frac{1}{T^2} \tag{25}$$

where the second inequality follows from Lemma 10. Moreover, conditioning on $\mathcal{B}_{n-1}$, since I_{τ_n-1} couples with $\tilde{I}^{a^{(n-1)*}}_{\tau_n-1}$, we have that

$$P(I_{\tau_n-1} \leq \kappa_1 \cdot \log T \,|\, \mathcal{B}_{n-1}) = P(\tilde{I}^{a^{(n-1)*}}_{\tau_n-1} \leq \kappa_1 \cdot \log T \,|\, \mathcal{B}_{n-1}) = P(\tilde{I}^{a^{(n-1)*}}_{\tau_n-1} \leq \kappa_1 \cdot \log T)/P(\mathcal{B}_{n-1}) \tag{26}$$

Note that the distribution of $\tilde{I}^{a^{(n-1)*}}_{\tau_n-1}$ is identical to the distribution of $I^{a^{(n-1)*}}_{\infty}$, which implies that

$$P(\tilde{I}^{a^{(n-1)*}}_{\tau_n-1} \leq \kappa_1 \cdot \log T) = P(I^{a^{(n-1)*}}_{\infty} \leq \kappa_1 \cdot \log T) \geq 1 - \frac{1}{T^2}$$

Therefore, by noting that $P(\mathcal{B}_{n-1}) \leq 1$, from (26), we have that

$$P(I_{\tau_n-1} \leq \kappa_1 \cdot \log T \,|\, \mathcal{B}_{n-1}) \geq 1 - \frac{1}{T^2} \tag{27}$$

From (25), (27) and the union bound, we have that

$$P(I_{\tau_n-1} \leq \kappa_1 \cdot \log T \text{ and } \tilde{I}^{a^{n*}}_{\tau_n} \leq \kappa_1 \cdot \log T \,|\, \mathcal{B}_{n-1}) \geq 1 - \frac{2}{T^2} \tag{28}$$

As a result, conditioning on $\mathcal{B}_{n-1}$, we know that

$$I_{\tau_n+L} \leq L \cdot \bar{D} + \kappa_1 \cdot \log T \text{ and } \tilde{I}^{a^{n*}}_{\tau_n+L} \leq L \cdot \bar{D} + \kappa_1 \cdot \log T \tag{29}$$

happens with a probability at least $1 - \frac{2}{T^2}$. It is clear to see that the event $\{I_{\tau_n+\max\{\log T, 2L\}} = \tilde{I}^{a^{n*}}_{\tau_n+\max\{\log T, 2L\}}\}$ happens as long as

$$I_t = \tilde{I}^{a^{n*}}_t = 0 \tag{30}$$

for some $t = \tau_n + L, \ldots, \tau_n + \max\{\log T, 2L\}$.

Suppose that $I_{\tau_n} \leq \tilde{I}^{a^{n*}}_{\tau_n}$ (resp. $I_{\tau_n} \geq \tilde{I}^{a^{n*}}_{\tau_n}$), from the evolution of the stochastic process in (10) and (13), we have that $I_t \leq \tilde{I}^{a^{n*}}_t$ (resp. $I_t \geq \tilde{I}^{a^{n*}}_t$) for any $t = \tau_n + L, \ldots, \tau_n + \max\{\log T, 2L\}$. Given that I_t and $\tilde{I}^{a^{n*}}_t$ must be non-negative (from definition), we conclude that if $I_{\tau_n} \leq \tilde{I}^{a^{n*}}_{\tau_n}$ (resp. $I_{\tau_n} \geq \tilde{I}^{a^{n*}}_{\tau_n}$), then (30) happens as long as $\tilde{I}^{a^{n*}}_t = 0$ (resp. $I_t = 0$). Thus, a sufficient condition for (30) to happen is that

$$\sum_{t=\tau_n+L}^{\tau_n+\max\{\log T, 2L\}} D_t - s(a^{n*}, Z_t) \geq L \cdot \bar{D} + \kappa_1 \cdot \log T$$

Since $D_t - s(a^{n*}, Z_t)$ are i.i.d. random variable for $t = \tau_n + L, \ldots, \tau_n + \max\{\log T, 2L\}$, we denote by $\delta_n = \mathbb{E}_{D\sim F}[D] - \mathbb{E}_{Z\sim G}[s(a^{n*}, Z)] \geq \delta$. Following Hoeffding's inequality (Lemma 7), we have that

$$\begin{aligned} P\left(\sum_{t=\tau_n+L}^{\tau_n+\kappa_2\max\{\log T, 2L\}} D_t - s(a^{n*}, Z_t) \geq L \cdot \bar{D} + \kappa_1 \cdot \log T\right) &\geq 1 - \exp(-\frac{2(\kappa_2 \max\{\log T, 2L\} - L(\bar{D}+1) - \kappa_1 \log T)^2}{\kappa_2 \max\{\log T, 2L\} - L}) \\ &\geq 1 - \frac{1}{T^2} \end{aligned}$$

where $\kappa_2 \geq \max\{4, 2(\bar{D} + 1 + \kappa_1)\} \geq \max\{\frac{4\log T}{\max\{\log T, 2L\}}, 2(\bar{D} + 1 + \kappa_1)\}$. Therefore, we have that

$$P\left(I_{\tau_n+\kappa_2\max\{\log T, 2L\}} = \tilde{I}^{a^{n*}}_{\tau_n+\kappa_2\max\{\log T, 2L\}} \,|\, \mathcal{B}_{n-1} \text{ and (29) happens}\right) \geq 1 - \frac{1}{T^2}.$$

Combining (28) and the induction hypothesis that $P(\mathcal{B}_{n-1}) \geq 1 - \frac{3(n-1)}{T^2}$, we have that

$$\begin{aligned} P(\mathcal{B}_n) =& P(\mathcal{B}_{n-1}) \cdot P(I_{\tau_n-1} \leq \kappa_1 \cdot \log T \text{ and } \tilde{I}^{a^{n*}}_{\tau_n} \leq \kappa_1 \cdot \log T \,|\, \mathcal{B}_{n-1}) \\ &\cdot P\left(I_{\tau_n+\kappa_2\max\{\log T, 2L\}} = \tilde{I}^{a^{n*}}_{\tau_n+\kappa_2\max\{\log T, 2L\}} \,|\, \mathcal{B}_{n-1} \text{ and (29) happens}\right) \\ \geq& (1 - \frac{3(n-1)}{T^2}) \cdot (1 - \frac{2}{T^2}) \cdot (1 - \frac{1}{T^2}) \geq (1 - \frac{3(n-1)}{T^2}) \cdot (1 - \frac{3}{T^2}) \\ \geq& 1 - \frac{3n}{T^2} \end{aligned}$$

which completes our proof of the induction of (23) for each epoch n. Therefore, our proof of the lemma is completed. □

Proof of Lemma 4. Clearly, from (14), it is enough to compare the value of I_t and $\tilde{I}_t^{a^{n*}}$ for each epoch n and each period t in the epoch n. Note that we identify an event $\mathcal{B}$ in Lemma 3 that I_t and $\tilde{I}_t^{a^{n*}}$ couple with each other. We consider two situations where $\mathcal{B}$ happens or $\mathcal{B}$ not happens.

Case 1: We now assume that $\mathcal{B}$ happens. Then, we know that for each epoch $n \in [N]$ and each $t = \tau_n + \kappa_2 \cdot \max\{\log T, 2L\}, \dots, \tau_{n+1} - 1$, the value of I_t and $\tilde{I}_t^{a^{n*}}$ are identical. Therefore, only when $t = \tau_n, \dots, \tau_n + \kappa_2 \cdot \max\{\log T, 2L\}$, the value of I_t and $\tilde{I}_t^{a^{n*}}$ can be different. Moreover, note that the evolution of I_t in (10) is the same as the evolution of $\tilde{I}_t^{a^{n*}}$ in (13), except that the intial value I_{τ_n} is different from $\tilde{I}_{\tau_n}^{a^{n*}}$. We know that the gap between I_t and $\tilde{I}_t^{a^{n*}}$ can only become smaller. Therefore, we get that

$$|I_t - \tilde{I}_t^{a^{n*}}| \le |I_{\tau_n} - \tilde{I}_{\tau_n}^{a^{n*}}| \le \kappa_1 \cdot \log T \tag{31}$$

where the last inequality follows from the condition in the event $\mathcal{B}$. We have that

$$\begin{aligned}\left|\mathbb{E}\left[\sum_n \sum_{t=\tau_n}^{\tau_{n+1}-1} (I_t - \tilde{I}_t^{a^{n*}}) \mid \mathcal{B}\right]\right| &\le \sum_n \sum_{t=\tau_n}^{\tau_n+\kappa_2\cdot\max\{\log T, 2L\}} \mathbb{E}[|I_t - \tilde{I}_t^{a^{n*}}| \mid \mathcal{B}] \\ &\le \sum_n \kappa_1 \cdot \log T \cdot \kappa_2 \cdot \max\{\log T, 2L\} \\ &= N \cdot \kappa_1\kappa_2 \log T \cdot \max\{\log T, 2L\}\end{aligned} \tag{32}$$

Case 2: We now assume that $\mathcal{B}$ does not happen. Clearly, a direct upper bound on both I_t and $\tilde{I}_t^{a^{n*}}$ is that

$$I_t \le \bar{D} \cdot t \text{ and } \mathbb{E}[\tilde{I}_t^{a^{n*}} \mid \mathcal{B}^c] \le \bar{D} \cdot t$$

Therefore, we have that

$$\left|\mathbb{E}\left[\sum_n \sum_{t=\tau_n}^{\tau_{n+1}-1} (I_t - \tilde{I}_t^{a^{n*}}) \mid \mathcal{B}^c\right]\right| \le \bar{D} \cdot \sum_{t=1}^{T} t \le \bar{D} \cdot T^2 \tag{33}$$

However, from Lemma 3, we know that $P(\mathcal{B}^c) \le \frac{3N}{T^2}$. As a result, combining (32) and (33), we get that

$$\begin{aligned}\left|\mathbb{E}\left[\sum_n \sum_{t=\tau_n}^{\tau_{n+1}-1} (I_t - \tilde{I}_t^{a^{n*}})\right]\right| &\le \left|\mathbb{E}\left[\sum_n \sum_{t=\tau_n}^{\tau_{n+1}-1} (I_t - \tilde{I}_t^{a^{n*}}) \mid \mathcal{B}\right]\right| \cdot P(\mathcal{B}) + \left|\mathbb{E}\left[\sum_n \sum_{t=\tau_n}^{\tau_{n+1}-1} (I_t - \tilde{I}_t^{a^{n*}}) \mid \mathcal{B}^c\right]\right| \cdot P(\mathcal{B}^c) \\ &\le \left|\mathbb{E}\left[\sum_n \sum_{t=\tau_n}^{\tau_{n+1}-1} (I_t - \tilde{I}_t^{a^{n*}}) \mid \mathcal{B}\right]\right| + \left|\mathbb{E}\left[\sum_n \sum_{t=\tau_n}^{\tau_{n+1}-1} (I_t - \tilde{I}_t^{a^{n*}}) \mid \mathcal{B}^c\right]\right| \cdot \frac{3N}{T^2} \\ &\le N \cdot \kappa_1\kappa_2 \log T \cdot \max\{\log T, 2L\} + 3N\bar{D}\end{aligned}$$

which completes our proof. □

Proof of Lemma 5. The proof generalizes the proof of Lemma 3. For each epoch n, we denote by

$$\mathcal{C}_n = \{I^a_{\tau_{n'}} \le \kappa_1 \cdot \log T,\ \tilde{I}^a_{\tau_{n'}} \le \kappa_1 \cdot \log T \text{ and } I^a_{\tau_{n'}+\kappa_2\cdot\max\{\log T, 2L\}} = \tilde{I}^{a^{n'*}}_{\tau_{n'}+\kappa_2\cdot\max\{\log T, 2L\}}\},\ \forall a \in \mathcal{A}_{n'},\ \forall n' \le n\}.$$

In order to prove the lemma, it is sufficient to prove that

$$P(\mathcal{C}_n) \ge 1 - \frac{3(K+1)n}{T^2}. \tag{34}$$

We prove (34) by using induction on the epoch n.

Clearly, when $n = 1$, we have that $P(I^a_{\tau_1} = 0) = 1$ for all $a \in \mathcal{A}_1$. From Lemma 10, there exists a constant $\kappa_1 > 0$ such that for each $a \in \mathcal{A}_1$, it holds that

$$P(\tilde{I}^a_{\tau_1} \le \kappa_1 \cdot \log T) \ge 1 - \frac{1}{T^2}$$

by noting that the distribution of $\tilde{I}^a_{\tau_1}$ is identical to the distribution of I^a_∞ for each $a \in \mathcal{A}_1$.

Now conditioning on the event $\{\tilde{I}^a_{\tau_1} \le \kappa_1 \cdot \log T, \ \forall a \in \mathcal{A}_1\}$, we proceed to bound the probability that event $\{I^a_{\tau_1+\kappa_2\cdot\max\{\log T, 2L\}} = \tilde{I}^a_{\tau_1+\kappa_2\cdot\max\{\log T, 2L\}}, \ \forall a \in \mathcal{A}_1\}$ happens. Note that the evolution of the stochastic process I^a_t in (9) is identical to the evolution of the stochastic process $\tilde{I}^a_t$ in (15) for $t = \tau_1, \dots, \tau_2 - 1$. Therefore, for each $a \in \mathcal{A}_1$, it is clear to see that the event $\{I^a_{\tau_1+\kappa_2\cdot\max\{\log T, 2L\}} = \tilde{I}^a_{\tau_1+\kappa_2\cdot\max\{\log T, 2L\}}\}$ happens as long as

$$I^a_t = \tilde{I}^a_t = 0, \quad \text{for some } t = \tau_1, \dots, \tau_1 + \kappa_2 \cdot \max\{\log T, 2L\}. \tag{35}$$

We note that $I^a_{\tau_1} \le \tilde{I}^a_{\tau_1}$ implies that $I^a_t \le \tilde{I}^a_t$ for all $t = \tau_1, \dots, \tau_1 + \kappa_2 \cdot \max\{\log T, 2L\}$. From the non-negativity of I^a_t and $\tilde{I}^a_t$, we have that (35) holds as long as $\tilde{I}^a_t = 0$ for some $t = \tau_1, \dots, \tau_1 + \kappa_2 \cdot \max\{\log T, 2L\}$. As a result, a sufficient condition for (35) to hold for a $a \in \mathcal{A}_1$ is that

$$\sum_{t=\tau_1}^{\tau_1+\kappa_2\cdot\max\{\log T, 2L\}} D_t - s(a, Z_t) \ge \kappa_1 \cdot \log T$$

Note that $D_t - s(a, Z_t)$ are i.i.d. random varibles for $t = \tau_1, \dots, \tau_1 + \kappa_2 \cdot \max\{\log T, 2L\}$. We also denote by $\delta = \mathbb{E}[D] - \mathbb{E}[s(\bar{q}, Z)]$. Following Hoeffding's inequality (Lemma 7), we have that

$$P\left(\sum_{t=\tau_1}^{\tau_1+\kappa_2\cdot\max\{\log T, 2L\}} D_t - s(a, Z_t) \ge \kappa_1 \cdot \log T\right) \ge 1 - \exp(-\frac{2(\delta\kappa_2 \max\{\log T, 2L\} - \kappa_1 \log T)^2}{\kappa_2 \cdot \max\{\log T, 2L\} \cdot \bar{D}}) \ge 1 - \frac{1}{T^2}$$

where $\kappa_2 \ge \max\{\frac{2\kappa_1}{\delta}, \frac{4\bar{D}}{\delta^2}\} \ge \max\{\frac{2\kappa_1 \log T}{\delta\cdot\max\{\log T, 2L\}}, \frac{4\bar{D}\log T}{\delta^2\cdot\max\{\log T, 2L\}}\}$.

The above derivation implies that

$$\begin{aligned}
P(\mathcal{B}_1) &= P\left(\mathcal{B}_1 \mid \tilde{I}^a_{\tau_1} \le \kappa_1 \cdot \log T, \ \forall a \in \mathcal{A}_1\right) \cdot P(\tilde{I}^a_{\tau_1} \le \kappa_1 \cdot \log T, \ \forall a \in \mathcal{A}_1) \\
&= P\left(\sum_{t=\tau_1}^{\tau_1+\kappa_2\cdot\max\{\log T, 2L\}} D_t - s(a, Z_t) \ge \kappa_1 \cdot \log T, \ \forall a \in \mathcal{A}_1\right) \cdot P(\tilde{I}^{a^{1*}}_{\tau_1} \le \kappa_1 \cdot \log T, \ \forall a \in \mathcal{A}_1) \\
&\ge (1 - \frac{K+1}{T^2}) \cdot (1 - \frac{K+1}{T^2}) \ge 1 - \frac{2(K+1)}{T^2} \ge 1 - \frac{3(K+1)}{T^2}
\end{aligned}$$

where the first inequality follows from the union bound by noting that $|\mathcal{A}_1| \le K + 1$. Therefore, we prove (34) for $n = 1$.

Now assume that (34) holds for epoch $n - 1$. We consider epoch n. Clearly, from the definition of the stochastic process $\tilde{I}^a_t$ in (15), $\tilde{I}^a_t$ refreshes when $t = \tau_n$. As a result, the distribution of $\tilde{I}^a_{\tau_n}$ is independent of the event $\mathcal{C}_{n-1}$ and is identical to the distribution of I^a_∞, which implies that

$$P(\tilde{I}^a_{\tau_n} \le \kappa_1 \cdot \log T \mid \mathcal{C}_{n-1}) = P(\tilde{I}^a_{\tau_n} \le \kappa_1 \cdot \log T) \ge 1 - \frac{1}{T^2}, \quad \forall a \in \mathcal{A}_n \tag{36}$$

where the second inequality follows from Lemma 10. Moreover, conditioning on $\mathcal{B}_{n-1}$, since $I^a_{\tau_n-1}$ couples with $\tilde{I}^a_{\tau_n-1}$ for each $a \in \mathcal{A}_n \subset \mathcal{A}_{n-1}$, we have that

$$P(I^a_{\tau_n-1} \le \kappa_1 \cdot \log T \mid \mathcal{C}_{n-1}) = P(\tilde{I}^a_{\tau_n-1} \le \kappa_1 \cdot \log T \mid \mathcal{C}_{n-1}) = P(\tilde{I}^a_{\tau_n-1} \le \kappa_1 \cdot \log T)/P(\mathcal{C}_{n-1}), \ \forall a \in \mathcal{A}_n \tag{37}$$

Note that the distribution of $\tilde{I}^a_{\tau_n-1}$ is identical to the distribution of I^a_∞, which implies that

$$P(\tilde{I}^a_{\tau_n-1} \le \kappa_1 \cdot \log T) = P(I^a_\infty \le \kappa_1 \cdot \log T) \ge 1 - \frac{1}{T^2}, \ \forall a \in \mathcal{A}_n$$

Therefore, by noting that $P(\mathcal{C}_{n-1}) \le 1$, from (37), we have that

$$P(I^a_{\tau_n - 1} \le \kappa_1 \cdot \log T \,|\, \mathcal{C}_{n-1}) \ge 1 - \frac{1}{T^2}, \ \forall a \in \mathcal{A}_n. \tag{38}$$

From (36), (38) and the union bound, we have that

$$P(I_{\tau_n - 1} \le \kappa_1 \cdot \log T \text{ and } \tilde{I}^{a^{n*}}_{\tau_n} \le \kappa_1 \cdot \log T, \ \forall a \in \mathcal{A}_n \,|\, \mathcal{C}_{n-1}) \ge 1 - \frac{2(K+1)}{T^2} \tag{39}$$

where we note that $|\mathcal{A}_n| \le K+1$. As a result, conditioning on $\mathcal{C}_{n-1}$, we know that

$$I^a_{\tau_n + L} \le L \cdot \bar{D} + \kappa_1 \cdot \log T \text{ and } \tilde{I}^a_{\tau_n + L} \le L \cdot \bar{D} + \kappa_1 \cdot \log T, \ \forall a \in \mathcal{A}_n \tag{40}$$

happens with a probability at least $1 - \frac{2(K+1)}{T^2}$. It is clear to see that for each $a \in \mathcal{A}_n$, the event $\{I^a_{\tau_n + \max\{\log T, 2L\}} = \tilde{I}^a_{\tau_n + \max\{\log T, 2L\}}\}$ happens as long as

$$I^a_t = \tilde{I}^a_t = 0 \tag{41}$$

for some $t = \tau_n + L, \dots, \tau_n + \max\{\log T, 2L\}$.

For each $a \in \mathcal{A}_n$, suppose that $I^a_{\tau_n} \le \tilde{I}^a_{\tau_n}$ (resp. $I^a_{\tau_n} \ge \tilde{I}^a_{\tau_n}$), from the evolution of the stochastic process in (9) and (15), we have that $I^a_t \le \tilde{I}^a_t$ (resp. $I^a_t \ge \tilde{I}^a_t$) for any $t = \tau_n + L, \dots, \tau_n + \max\{\log T, 2L\}$. Given that I^a_t and $\tilde{I}^a_t$ must be non-negative (from definition), we conclude that if $I^a_{\tau_n} \le \tilde{I}^a_{\tau_n}$ (resp. $I^a_{\tau_n} \ge \tilde{I}^a_{\tau_n}$), then (30) happens as long as $\tilde{I}^a_t = 0$ (resp. $I^a_t = 0$). Thus, a sufficient condition for (41) to happen is that

$$\sum_{t=\tau_n+L}^{\tau_n + \max\{\log T, 2L\}} D_t - s(a, Z_t) \ge L \cdot \bar{D} + \kappa_1 \cdot \log T$$

Since $D_t - s(a, Z_t)$ are i.i.d. random variable for $t = \tau_n + L, \dots, \tau_n + \max\{\log T, 2L\}$, we denote by $\delta_{n,a} = \mathbb{E}_{D \sim F}[D] - \mathbb{E}_{Z \sim G}[s(a, Z)] \ge \delta$. Following Hoeffding's inequality (Lemma 7), for each $a \in \mathcal{A}_n$, we have that

$$\begin{aligned} P\left(\sum_{t=\tau_n+L}^{\tau_n + \kappa_2 \max\{\log T, 2L\}} D_t - s(a, Z_t) \ge L \cdot \bar{D} + \kappa_1 \cdot \log T\right) &\ge 1 - \exp(-\frac{2(\kappa_2 \max\{\log T, 2L\} - L(\bar{D}+1) - \kappa_1 \log T)^2}{\kappa_2 \max\{\log T, 2L\} - L}) \\ &\ge 1 - \frac{1}{T^2} \end{aligned}$$

where $\kappa_2 \ge \max\{4, 2(\bar{D} + 1 + \kappa_1)\} \ge \max\{\frac{4 \log T}{\max\{\log T, 2L\}}, 2(\bar{D} + 1 + \kappa_1)\}$. Therefore, from the union bound, we have that

$$P\left(I^a_{\tau_n + \kappa_2 \max\{\log T, 2L\}} = \tilde{I}^a_{\tau_n + \kappa_2 \max\{\log T, 2L\}}, \ \forall a \in \mathcal{A}_n \,|\, \mathcal{C}_{n-1} \text{ and (40) happens}\right) \ge 1 - \frac{K+1}{T^2}.$$

Combining (39) and the induction hypothesis that $P(\mathcal{C}_{n-1}) \ge 1 - \frac{3(K+1)(n-1)}{T^2}$, we have that

$$\begin{aligned} P(\mathcal{C}_n) =& P(\mathcal{C}_{n-1}) \cdot P(I^a_{\tau_n - 1} \le \kappa_1 \cdot \log T \text{ and } \tilde{I}^a_{\tau_n} \le \kappa_1 \cdot \log T, \ \forall a \in \mathcal{A}_n \,|\, \mathcal{C}_{n-1}) \\ & \cdot P\left(I^a_{\tau_n + \kappa_2 \max\{\log T, 2L\}} = \tilde{I}^a_{\tau_n + \kappa_2 \max\{\log T, 2L\}}, \ \forall a \in \mathcal{A}_n \,|\, \mathcal{C}_{n-1} \text{ and (40) happens}\right) \\ \ge& (1 - \frac{3(K+1)(n-1)}{T^2}) \cdot (1 - \frac{2(K+1)}{T^2}) \cdot (1 - \frac{K+1}{T^2}) \\ \ge& (1 - \frac{3(K+1)(n-1)}{T^2}) \cdot (1 - \frac{3(K+1)}{T^2}) \\ \ge& 1 - \frac{3(K+1)n}{T^2} \end{aligned}$$

which completes our proof of the induction of (34) for each epoch n. Therefore, our proof of the lemma is completed. □

Proof of Lemma 6. We first show that for each epoch $n \in [N]$ and each action $a \in \mathcal{A}_n$, we can use the average value of $\tilde{I}^a_t$ for $t = \tau_n + \kappa_2 \cdot \max\{\log T, 2L\}$ to $\tau_{n+1} - 1$ to approximate the value of $\mathbb{E}[I^{\pi_a}_\infty]$, where the length of the confidence interval can be given by γ_n.

Clearly, $\{\tilde{I}^a_t\}_{t=\tau_n+\kappa_2\cdot\max\{\log T,2L\}}^{\tau_{n+1}-1}$ forms a Markov chain. We denote by $\boldsymbol{I}$ a vector such that

$$\boldsymbol{I} = (\tilde{I}^a_t, \forall t = \tau_n + \kappa_2 \cdot \max\{\log T, 2L\}, \dots, \tau_{n+1} - 1)$$

We apply Lemma 8 to derive a concentration bound for $\boldsymbol{I}$. To be specific, for each epoch $n \leq N-1$, we regard $\tilde{I}^a_{\tau_n+\kappa_2\cdot\max\{\log T,2L\}+i}$ as X_i for $i = 1, \dots, \tau_{n+1} - 1 - \tau_n - \kappa_2 \cdot \max\{\log T, 2L\}$. Clearly, $\boldsymbol{I}$ is a Markov chain with stationary distributions and satisfies the conditions in Lemma 8.

We now denote $m = \tau_{n+1} - \tau_n - 1 - \kappa_2 \cdot \max\{\log T, 2L\}$. Then, from Lemma 8, there exists a constant λ such that for any $\epsilon > 0$, it holds

$$P\left(\left|\sum_{i=1}^m \tilde{I}^a_{\tau_n+\kappa_2\cdot\max\{\log T,2L\}+i} - \mathbb{E}\left[\sum_{i=1}^m \tilde{I}^a_{\tau_n+\kappa_2\cdot\max\{\log T,2L\}+i}\right]\right| \geq \sqrt{m}\cdot\epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2(1-\lambda)}{4}\right).$$

We now set $\epsilon = \frac{\gamma_n\cdot\sqrt{m}}{2}$. Then, we have

$$\exp\left(-\frac{\epsilon^2(1-\lambda)}{4}\right) \leq \exp(-\frac{(1-\lambda)m\gamma_n^2}{16}) \tag{42}$$

We proceed to give a lower bound on $m\gamma_n^2$, which will imply an upper bound for (42). Note that

$$m = \tau_{n+1} - \tau_n - 1 - \kappa_2 \cdot \max\{\log T, 2L\} = \kappa_2 \cdot (\max\{\frac{1}{\gamma_n^2}\cdot\log T, 3L\} - \max\{\log T, 2L\})$$

If $\frac{1}{\gamma_2}\cdot\log T \geq 3L$, then we have

$$\max\{\frac{1}{\gamma_n^2}\cdot\log T, 3L\} - \max\{\log T, 2L\} = \frac{1}{\gamma_n^2}\cdot\log T - \max\{\log T, 2L\} \geq \frac{1}{3\gamma_n^2}\cdot\log T$$

If $\frac{1}{\gamma_2}\cdot\log T < 3L$, then we have

$$\max\{\frac{1}{\gamma_n^2}\cdot\log T, 3L\} - \max\{\log T, 2L\} = L \geq \frac{1}{3\gamma_n^2}\cdot\log T$$

Therefore, it holds that

$$m = \kappa_2 \cdot (\max\{\frac{1}{\gamma_n^2}\cdot\log T, 3L\} - \max\{\log T, 2L\}) \geq \frac{\kappa_2}{3\gamma_n^2}\cdot\log T \tag{43}$$

Plugging (43) into (42), we have

$$\exp\left(-\frac{\epsilon^2(1-\lambda)}{4}\right) \leq \exp\left(-\frac{(1-\lambda)\kappa_2\log T}{12}\right) \leq \frac{1}{T^2}$$

where $\kappa_2 \geq 24/(1-\lambda)$. We have

$$\begin{aligned} &P\left(\left|\sum_{i=1}^m \tilde{I}^a_{\tau_n+\kappa_2\cdot\max\{\log T,2L\}+i} - \mathbb{E}\left[\sum_{i=1}^m \tilde{I}^a_{\tau_n+\kappa_2\cdot\max\{\log T,2L\}+i}\right]\right| \geq m\cdot\frac{\gamma_n}{2}\right) \\ =&P\left(\left|\sum_{i=1}^m \tilde{I}^a_{\tau_n+\kappa_2\cdot\max\{\log T,2L\}+i} - m\cdot\mathbb{E}\left[I^{\pi_a}_\infty\right]\right| \geq m\cdot\frac{\gamma_n}{2}\right) \\ \leq&\frac{2}{T^2} \end{aligned} \tag{44}$$

where the second inequality follows from the fact that the distribution of $\tilde{I}_t^a$ is identical to the distribution of $I_\infty^{\pi_a}$. Moreover, from Hoeffding's inequality (Lemma 7), it holds that

$$\begin{aligned} P\left(\left|\sum_{t=\tau_n+\kappa_2\cdot\max\{\log T,2L\}}^{\tau_{n+1}-1} s(a,Z_t)) - m\cdot\mathbb{E}[s(a,Z)]\right| \geq m\cdot\frac{\gamma_n}{2}\right) &\leq 2\exp(-\frac{m\gamma_n^2}{2\bar{D}^2}) \leq 2\exp(-\frac{\kappa_2\log T}{6\bar{D}^2}) \\ &\leq \frac{2}{T^2} \end{aligned} \tag{45}$$

where the second inequality follows from (43) and the third inequality follows from $\kappa_2 \geq 12\bar{D}^2$. Therefore, conditional on the event $\mathcal{C}$ happens, we have that

$$P\left(\left|\tilde{C}_n^a - \hat{C}_\infty^{\pi_a}\right| \leq (h+b)\cdot\frac{\gamma_n}{2} \mid \mathcal{C}\right) \geq 1 - \frac{4}{T^2}$$

which implies that (from union bound over all $a \in \mathcal{A}$ and all $n \leq N-1$)

$$P(\mathcal{E} \mid \mathcal{C}) = P\left(\{|\tilde{C}_n^a - \hat{C}_\infty^{\pi_a}| \leq (h+b)\cdot\frac{\gamma_n}{2},\ \forall a \in \mathcal{A}_n, \forall 1 \leq n \leq N-1\}\right) \geq 1 - \frac{4(K+1)N}{T^2}$$

From Lemma 5, we know that $P(\mathcal{C}) \geq 1 - \frac{3(K+1)N}{T^2}$. Therefore, we have that

$$P(\mathcal{E}) = P(\mathcal{E} \mid \mathcal{C}) \cdot P(\mathcal{C}) \geq 1 - \frac{7(K+1)N}{T^2}$$

which completes our proof. □