Learning To Maximize Welfare with a Reusable Resource

MATTHEW FAW*, The University of Texas at Austin, USA
ORESTIS PAPADIGENOPOULOS*, The University of Texas at Austin, USA
CONSTANTINE CARAMANIS, The University of Texas at Austin, USA
SANJAY SHAKKOTTAI, The University of Texas at Austin, USA

Considerable work has focused on optimal stopping problems where random IID offers arrive sequentially for a single available resource which is controlled by the decision-maker. After viewing the realization of the offer, the decision-maker irrevocably rejects it, or accepts it, collecting the reward and ending the game. We consider an important extension of this model to a dynamic setting where the resource is "renewable" (a rental, a work assignment, or a temporary position) and can be allocated again after a delay period d. In the case where the reward distribution is known a priori, we design an (asymptotically optimal) 1/2-competitive Prophet Inequality, namely, a policy that collects in expectation at least half of the expected reward collected by a prophet who a priori knows all the realizations. This policy has a particularly simple characterization as a thresholding rule which depends on the reward distribution and the blocking period d, and arises naturally from an LP-relaxation of the prophet's optimal solution. Moreover, it gives the key for extending to the case of unknown distributions; here, we construct a dynamic threshold rule using the reward samples collected when the resource is not blocked. We provide a regret guarantee for our algorithm against the best policy in hindsight, and prove a complementing minimax lower bound on the best achievable regret, establishing that our policy achieves, up to poly-logarithmic factors, the best possible regret in this setting.

CCS Concepts: • **Theory of computation → Online learning algorithms**; **Approximation algorithms analysis**.

Additional Key Words and Phrases: Prophet Inequalities; Online Learning; Lower Bounds; Regret

ACM Reference Format:
Matthew Faw, Orestis Papadigenopoulos, Constantine Caramanis, and Sanjay Shakkottai. 2022. Learning To Maximize Welfare with a Reusable Resource. *Proc. ACM Meas. Anal. Comput. Syst.* 6, 2, Article 27 (June 2022), 29 pages. https://doi.org/10.1145/3530893

1 INTRODUCTION

In a wide class of sequential decision-making environments, the decision-maker observes a sequence of random variables and decides on the most profitable time to take an action. Such scenarios, which arise in a number of different domains including economics, statistics and operation research, are the main focus of *optimal stopping theory* [11, 44]. One of the most studied problems in this area is the *Prophet Inequality*. Here, a *"gambler"* observes a sequence of random *"rewards"* $X_1, \ldots, X_n$ arriving in an arbitrary (or even adversarial) order. After observing the realization of each reward,

*Both authors contributed equally to this research.

Authors' addresses: Matthew Faw, matthewfaw@utexas.edu, The University of Texas at Austin, Austin, Texas, USA; Orestis Papadigenopoulos, papadig@cs.utexas.edu, The University of Texas at Austin, Austin, Texas, USA; Constantine Caramanis, constantine@utexas.edu, The University of Texas at Austin, Austin, Texas, USA; Sanjay Shakkottai, sanjay.shakkottai@utexas.edu, The University of Texas at Austin, Austin, Texas, USA.

2476-1249/2022/6-ART27 $15.00
https://doi.org/10.1145/3530893

the gambler has to choose whether to collect it and stop, or irrevocably reject it and continue to the next reward. Assuming distributional knowledge on the rewards, the gambler's objective is to maximize the expected reward collected.

The first results in this area (Krengel, Sucheston, and Garling [31, 32]) established the existence of a stopping-rule guaranteeing an expected reward of at least $1/2 \cdot \mathbb{E}[\max_i X_i]$ – that is, half of the expected reward collected by a *"prophet"*, who knows all the reward realizations from the beginning and simply stops at the maximum. Soon after, Samuel-Cahn [43] showed that a remarkably simple threshold-based policy which accepts the first reward in the order that is greater or equal to median$(\max_i X_i)$, if such a reward exists, also achieves the above guarantee. Almost three decades later, Kleinberg and Weinberg [30] show that the same guarantee holds by replacing the median of $\max_i X_i$ with $1/2 \cdot \mathbb{E}[\max_i X_i]$. The above type of problems (and associated guarantees) have drawn the attention of researchers from various fields [14, 37].

In this work, we initiate the study of the following dynamic prophet inequality setting: A gambler observes a sequence of random IID offers for a single available resource and, at each time step, decides whether to collect the observed reward, or to skip the round. However, once the gambler collects a reward, the resource becomes *"unavailable"* (or *"blocked"*) and, thus, she cannot collect or observe any other reward for a fixed and known number of subsequent time steps, known as the *"delay."* The gambler seeks to maximize her expected reward collected (*"welfare"*) within an unknown time horizon. Our high-level goal is to design an optimal prophet inequality for the above setting, even in the case where a priori distributional knowledge on the rewards is not assumed. In that case, a parallel objective is to minimize the regret against the best possible prophet inequality.

The above natural model captures many applications in task-allocation, ride-sharing platforms, online auctions, and matching platforms. As an example, consider a platform like Mechanical Turk or Upwork that is used for matching workers to tasks. From the perspective of a specific worker, when a new task arrives, the worker can choose to either work on this task and accept the associated payment, or pass, depending on whether the payment makes the task worthwhile for the worker. If the worker decides to accept the task, the worker must complete the task (and hence is "blocked") before accepting a new task. More generally, whenever tasks or assignments, rentals, etc. are for a fixed duration (as in [22, 29]), our model applies directly.

1.1 Main challenges and our contributions

In this work, we distinguish between the *"Bayesian"* variant of our recurrent prophet inequality problem, where the reward distribution is known to the gambler a priori, and the *"learning"* variant, where the gambler starts without any information, other than the delay. We remark that in both variants, the gambler is not aware of the time horizon of the instance. We now outline the key challenges encountered and our main technical contributions.

Optimal prophet inequality for the Bayesian problem. In the first part of our work, we provide an optimal prophet inequality for the Bayesian variant of our problem. Here, the main technical hurdle is that, due to the complex dynamical patterns induced by the delay, the expected reward collected by a prophet is hard to exactly characterize. Instead, we upper-bound the asymptotic prophet's expected reward by the solution of an *infinite-dimensional linear program* (LP), defined over the space of probability distributions. By analyzing the KKT conditions of this formulation, we show that the optimal solution is characterized through greedy waterfilling. This observation allows us to construct a threshold-based policy, which computes a threshold as a function of the reward distribution and the delay, and, then, accepts any element of reward larger than this threshold, as long as the resource is not blocked. Specifically, given knowledge of the reward distribution and the delay d, our policy collects (asymptotically and in expectation) at least a $\rho(d)$-fraction

of the prophet's expected reward, where $\rho(d) = \frac{d+1}{2d+1} \approx 1/2$. As a crucial insight for proving this guarantee, we model the resource availability as a time-homogeneous discrete time *Markov Chain* (DTMC) with states $\{0, 1, \dots d\}$, each representing the number of rounds that need to pass until the resource becomes available again. Using knowledge of the transition probabilities, which are time-invariant as a function of the fixed threshold of our policy, we first compute the stationary probability of the resource being available, and then we leverage a coupling argument in order to lower bound the availability of any round. Finally, we prove that the competitive guarantee of our prophet inequality is the best achievable by drawing a connection to a related bandits problem.

Regret upper bound for the learning problem. We then focus on the learning setting, where the gambler is initially unaware of the reward distribution, and observes the realized rewards *only* when the resource is available (i.e., not blocked). The objective here is to minimize the regret measured relative to the best possible gambler's policy. The naïve extension of the policy from the Bayesian setting – estimating the threshold of the Bayesian setting using an explore-then-commit type strategy (i.e., use a sub-linear number of samples purely to learn the threshold) – will not work, since estimating the threshold to a sufficiently high accuracy requires a *linear* number of samples (which would incur *linear* regret under an explore-then-commit strategy). Instead, we exploit the fact that our system is *guaranteed* to receive a *linear* number of samples (roughly $t/d+1$ samples after t time steps) as part of the state evolution, and continually update our estimated threshold over time using these samples. This procedure, however, correlates the state of the system with the estimate of the threshold, and thus poses an interesting technical challenge. Furthermore, modeling the availability state via a time-homogeneous Markov Chain, as in the Bayesian case, is no longer possible, since the transition probabilities of each round depend on the current trajectory through the estimated threshold.

In order to overcome these difficulties and provide an upper bound on the regret of our learning algorithm, we proceed as follows: As a first step, through an application of the "*compensating coupling*" technique due to Vera and Banerjee [46], we are able to link the regret accumulated with the error of the estimated threshold at each round. In order to control these errors, we establish that our algorithm satisfies two properties: *quality of the threshold estimator* and *sufficiency of samples*. In this direction, we provide "anytime" concentration guarantees for our estimator (namely, for the case where the number of collected samples is random), and show that at any round t, the algorithm has collected a linear in t (and independent of d) number of samples. To achieve the latter, we construct an "*eager*" (fictitious) version of our Bayesian prophet inequality, equipped with a carefully chosen threshold. This threshold must be large enough to guarantee that (with high probability) $\Omega(t)$ samples are observed by time t, while simultaneously small enough such that, after $\Theta(\log n)$ rounds, it will always be overestimated by the threshold of the learning algorithm (with high probability). Under the last condition, and through a deterministic *charging argument* on the coupled evolution of the two policies, we show that the number of samples collected by the learning policy *stochastically dominates* that of the eager prophet inequality. The above ideas culminate in a final regret guarantee of $O\left(\sqrt{n \cdot \log n}\right)$, where n is the time horizon.

Regret lower bound via time-aggregated suboptimalities. We provide a regret lower bound of $\Omega(\sqrt{n}/d^{3/2})$, thus proving that the regret of our policy is optimal up to poly-logarithmic factors and inverse scaling in d. Our construction has two main insights. (i) A fundamental difficulty in our setting compared to the bandit setting is that our policies observe the offered reward *before* deciding whether to accept or reject. This discrepancy invalidates several crucial steps in the standard lower bound argument (see, e.g., [34, Theorem 15.2]) However, we show that, *for the environments considered in our lower-bound construction, any* policy can simulate its decision of whether to accept

or reject the reward *before observing it*. (ii) Even though the *instantaneous* regret of an algorithm can potentially be *negative* (e.g., if the algorithm collects when the optimal policy is blocked), the regret over properly-chosen time windows is always non-negative. Using this insight, we introduce a technique we call *time-aggregated suboptimality gaps*, which allows us to obtain (up to small additive terms) a regret decomposition of a similar form as in the stochastic bandit setting [34, Lemma 4.5]. Taken together, these two insights allow us to reduce our lower bound construction to that of a stochastic (two-armed) bandit instance.

1.2 Related work

Prophet inequalities and secretary problems. In addition to the original prophet inequality problem [31, 32], numerous combinatorial extensions (where more than one reward can be collected, subject to feasibility constraints) have been studied. Examples include the choose-k variant [2, 26], matroid and packing constraints [9, 21, 30, 40], and (bipartite) matching environments [3, 18, 20, 25]. The problem has been also studied under non-linear (submodular or subadditive) objectives [41]. We remark that the type of feasibility constraints we consider in this work (i.e., where collecting a reward invalidates the subsequent d rewards) does not fall into any of the above categories.

Of particular importance is the IID case of the problem, where the rewards are drawn independently from the same distribution. In their seminal work, Hill and Kertz [27] prove a $(1-1/e) \approx 0.632$-competitive prophet inequality for this setting, and conjecture that no policy can collect (in expectation) more than a 0.731-fraction of the prophet's expected reward. More than three decades later, this conjecture was refuted by Abolhassani et al. [1] by proving the existence of a 0.738-competitive policy. The state-of-the-art in this regime is a 0.745-competitive prophet inequality due to Correa et al. [13], and this guarantee is known to be the best possible.

Recent work has also focused on designing prophet inequalities in the regime where the gambler has access only to a limited number of samples from each distribution. A number of results have been obtained in this setting, both for the IID case [12, 15], as well as for general distributions [4, 8, 42]. However, access to limited number of samples usually entails competitive guarantees that are far from the best achievable in the Bayesian setting.

The need for modeling reusable resources has motivated analogous models in the literature on the secretary problem. In this setting, Fiat et al. [22] introduce the "temp secretary problem" (see also [29]), where, every time a reward is collected, the resource becomes unavailable for a fixed duration of time. However, since the elements in their setting arrive stochastically in continuous time, this model is incomparable with ours.

Undiscounted reinforcement learning. The problem of regret minimization in undiscounted Reinforcement Learning (RL) has a rich and growing literature [28, 33, 38, 39, 45]. One distinguishing feature of prophet inequality problems from RL is that rewards are offered *before* the policy decides whether or not to accept, unlike in RL, where rewards are observed *after* an action has been selected. Nevertheless, our problem can still be cast as a Markov Decision Process (MDP) with a (possibly *infinite*) state space $S = [d] \cup \text{supp}(\mathcal{D})$, where the states $s \in [d]$ encode the time left until the resource is available, and the state $v \in \text{supp}(\mathcal{D})$ encodes that the resource is available and a reward of value v is currently offered to the gambler. In the case that $\mathcal{D}$ has finite support, the regret upper bounds generally depend on the size of the state space, and thus are far from optimal for our case. While there is literature on the RL problem in the case when the state space is continuous, the best-known regret upper bounds for this setting [33, 38, 39] have significantly worse dependence on the time horizon n (in particular, $n^{3/4}$ from [38], which can be improved to $n^{2/3}$ [33] under additional smoothness conditions on the distribution). Finally, the regret lower bounds in [28, 38] (in the discrete and continuous MDP setting, respectively) simply show that there *exists* a "hard" instance

where the claimed regret must be suffered. On the other hand, the lower bounds we present in our paper are tailored to our specialized setting.

Online matching, assortment optimization, and revenue management. The notion of reusable resources has also been studied in the context of online matching, assortment optimization, and revenue management. In particular, in [23, 24] the authors study the problem of online assortment optimization with reusable resources. Their model captures scenarios of arbitrary buyer arrival order and multiple reusable resources, each associated to a specific capacity (i.e., maximum number of parallel allocations) and stochastic delay. However, the reward associated with an allocation in these settings is resource-dependent (and not buyer-dependent, as in our case) – a fact that makes our model different in nature and our results incomparable. In [16], Dickerson et al. study the problem of online bipartite matching, where the left-side (offline) vertices are reusable, while the right-side (online) vertices arrive stochastically under some known distribution. Our setting can be thought of as a variation of this model, where we have a single offline vertex of deterministic delay (our resource), yet an infinite number of online vertices (and, thus, the LP-based algorithm of [16] does not apply). In [35], Levi and Radovanović consider a similar problem in the context of revenue management, where the number of buyer types is again assumed to be finite, and the arrival rate of each type follows an independent Poisson process. Finally, in a different spirit, Chen et al. [10] develop an online learning approach for maximizing the net profit in a service queue, which is defined as the service fee (i.e., the price times the demand) minus the capacity cost (i.e., the available number of resources times the cost of a resource) and penalty of congestion (which is a function of the number of resources and the demand).

2 PRELIMINARIES

Model. Let $X_1, X_2, \ldots, X_n$ be a sequence of n rewards drawn IID from a non-negative reward distribution $\mathcal{D}$. At each round t, the gambler first observes the reward realization $X_t \sim \mathcal{D}$, and then must decide whether to collect the reward or skip on it forever. Crucially, at each time t when a reward is collected, the gambler *has* to skip (*without observing*) the $d \geq 1$ subsequent rewards, $\{X_{t+1}, \ldots, X_{t+d}\}$, while the resource is unavailable. The goal of the gambler is to maximize the expected collected reward relative to that of a prophet, who has infinite computational power and knows a priori the reward realizations of all rounds. We assume that the gambler knows the delay d, but does not know the time horizon n.

For the rest of this text, we denote by F (resp., f) the cumulative distribution function (resp., probability density function) of $\mathcal{D}$. For any non-negative integer k, we denote $[k] = \{1, 2, \ldots, k\}$. For any policy $\mathcal{P}$, we denote by $\text{free}_{\mathcal{P}}(t)$ the event that, during a run of $\mathcal{P}$, the resource is available at time t, i.e., no reward has been collected by $\mathcal{P}$ in the past d time-steps. When the policy is clear from context, we abuse this notation slightly by referring to the event as $\text{free}(t)$. We denote by $\log(\cdot)$ the natural logarithm. Finally, we use $x \lesssim y$ (resp., $x \gtrsim y$) to denote that y is greater than (resp., less than) x up to constant factors.

Competitive guarantee and regret definition. Let ALG (resp., OPT) be the reward collected by an online policy (resp., the prophet). For brevity, we use ALG (resp., OPT) for referring to both the policy and the associated collected reward. The competitive guarantee of a policy ALG (which knows the underlying reward distribution) is defined as the minimum ratio between the expected reward collected by a gambler using this policy and that collected by the prophet over all possible instances. Formally,

$$\rho_{\text{ALG}} = \inf \frac{\mathbb{E}[\text{ALG}]}{\mathbb{E}[\text{OPT}]},$$

where the infimum is taken over all problem instances, including the distribution $\mathcal{D}$, the delay d, and the time horizon n.

Due to information-theoretic reasons, there exist cases where no policy can achieve a competitive guarantee greater than ρ. For these cases, using the prophet's expected reward as a baseline would inevitably lead to linear regret. Instead, for any ρ-competitive policy, we use the relaxed notion of *ρ-approximate regret* (or, simply, ρ-regret), defined relative to a learning policy $\mathcal{S}$ as

$$\text{Regret}_\rho(n) = \rho \cdot \mathbb{E}\left[\text{OPT}\right] - \sum_{t\in[n]} \mathbb{E}\left[X_t \cdot \mathbb{1}\left\{\mathcal{S} \text{ collects } X_t\right\}\right]. \tag{1}$$

Technical assumptions. In Sections 3 and 4, we assume w.l.o.g. that all the probability distributions involved are continuous (i.e., they do not contain point masses). This is a purely technical assumption and can be easily relaxed by convolving the given distribution and the obtained samples with Gaussian noise of infinitesimally small variance.

For the case of known reward distribution (Section 3), we do not make any assumptions on the distribution $\mathcal{D}$ (other than continuity). In the case of unknown distribution (Section 4), we make the standard assumption in online learning settings that $\mathcal{D}$ has bounded support in $[0, 1]$ (we note that our results can be readily extended to the case of subgaussian distributions).

3 CONSTRUCTING AN OPTIMAL POLICY IN THE BAYESIAN SETTING

The main result of this section is a $\rho(d)$-competitive (asymptotically) prophet inequality for the case where the reward distribution is known a priori, where $\rho(d) = \frac{d+1}{2d+1} \approx 1/2$, and d the delay of the instance. In Section 5, we show that this guarantee is the best one can hope for against a prophet who a priori knows all the reward realizations, and optimally solves the underlying packing problem using infinite computational power.

We design a simple policy (see Algorithm 1) that computes a threshold τ as a function of the reward distribution $\mathcal{D}$, and then accepts any reward that satisfies $X_t \geq \tau$ if and only if the resource is available at time t. Specifically, given the c.d.f. F of $\mathcal{D}$, the threshold is computed at the initialization phase as $\tau = F^{-1}(1 - 1/d+1)$.

Algorithm 1: Asymptotically optimal policy for the case of known reward distribution

Input: Reward distribution $\mathcal{D}$ with c.d.f. F and delay d;
Set threshold $\tau \leftarrow F^{-1}\left(1 - \frac{1}{d+1}\right)$;
for $t = 1, 2, \ldots$ **do**
    Observe reward X_t;
    if $X_t \geq \tau$ **and** *resource is available* **then**
        Collect X_t and make resource unavailable for rounds $t+1, \ldots, t+d$;
    else
        Skip on X_t;
    end
end

The intuition behind the choice of the threshold in Algorithm 1 is simple: assuming an infinite time horizon, the value $\tau = F^{-1}(1 - 1/d+1)$ corresponds exactly to the limiting case where the algorithm is indifferent between collecting or not a reward equal to τ, since any option would not affect its long-run average expected reward. As we can see, the threshold $F^{-1}(1 - 1/d+1)$ increases naturally with d, quantifying in that way the fact that a larger delay requires a larger reward to make the commitment worthwhile.

In addition, one might expect that an algorithm equipped with an adaptive threshold (i.e., one that changes as a function of time) would perform strictly better (since, e.g., the algorithm could be less conservative towards the end of the horizon). While this is true in the finite horizon regime, since we do not assume knowledge of the time horizon, it is unclear how to leverage such an improvement. Moreover, our approach of simply choosing a fixed threshold comes at the cost of only small additive losses in the competitive guarantee.

In the rest of this section, we prove the following result on the competitive guarantee of Algorithm 1:

Theorem 3.1. *Let* $\mathbb{E}[\mathsf{OPT}]$ *be the prophet's expected reward. For* $\rho(d) = \frac{d+1}{2d+1}$*, the expected reward of Algorithm 1 satisfies*

$$\mathbb{E}[\mathsf{ALG}] \;\geq\; \underbrace{\rho(d)\cdot\mathbb{E}[\mathsf{OPT}]}_{\text{competitive guarantee}} - \underbrace{\rho(d)\cdot(d+1)\cdot\mathbb{E}[X]}_{\text{loss due to LP upper bound}} - \underbrace{e\cdot d\cdot\mathbb{E}[X]}_{\text{loss due to mixing}}.$$

As we show in the rest of this section, the $\rho(d)$-factor in the first term of the above bound is the (asymptotic) competitive guarantee of our policy for instances of delay d. The second term is the additive loss due to the LP relaxation (see (MP) below) we use as an upper bound on the expected optimal reward, given that this relaxation becomes very loose when the time horizon is small comparing to the delay. The use of this LP to motivate our algorithm is also the reason why our asymptotic competitive guarantee does not match the best possible for the IID prophet inequality, in the case where $d \geq n$. The third loss is due to the mixing of the Markov Chain we use to lower bound the availability of the resource. Notice that the last two terms in the above bound are vanishing as the time horizon goes to infinity.

3.1 Competitive analysis of Algorithm 1

Characterizing the expected maximum reward. The first observation while attempting to design a competitive policy is that the prophet's expected reward lacks a simple characterization. In order to overcome this issue, we instead choose to upper-bound this reward by constructing the following infinite-dimensional LP formulation:

$$\begin{aligned} \textbf{maximize:} \quad & n\cdot\int_{x=0}^{\infty} x\cdot q(x)\,\mathrm{d}x \\ \textbf{s.t.:} \quad & \int_{x=0}^{\infty} q(x)\,\mathrm{d}x \;\leq\; \frac{1}{d+1} \\ & 0 \;\leq\; q(x) \;\leq\; f(x), \qquad \forall x \geq 0. \end{aligned} \tag{MP}$$

In the above formulation, each variable $q(x)$ can be thought of as the *expected fraction of time* the prophet collects a reward equal to x. Intuitively, the first set of constraints suggests that the expected fraction of time where the prophet collects *any* reward cannot be more than $1/(d+1)$ (asymptotically) for any instance of delay equal to d. Further, the second set of constraints indicates that the expected fraction of time a reward x is collected cannot be more than $f(x)$, namely, the p.d.f. of $\mathcal{D}$ at x.

In the next lemma, we show that the formulation (MP) asymptotically yields an upper bound on the prophet's expected reward.

Lemma 3.2. *Let* MP^* *be an optimal solution of* (MP) *and* OPT *be the reward collected by the prophet. Then, it is the case that*

$$\mathrm{MP}^* \geq \left(1 - \frac{d+1}{n+d+1}\right)\cdot\mathbb{E}[\mathsf{OPT}].$$

By simply observing (MP), it is not hard to see that an optimal solution can be greedily computed using a greedy waterfilling approach: Starting from $x = \infty$ and moving towards smaller x, we set $q(x) = f(x)$ up until $x = \tau$, where $\int_{x=\tau}^{\infty} q(x)\,\mathrm{d}x = \frac{1}{d+1}$. Notice that for continuous distributions (without point masses), such a τ uniquely exists and is equal to $\tau = F^{-1}(1 - \frac{1}{d+1})$. By analyzing the KKT conditions that follow from (MP), it is not hard to verify that the optimal solution q^* has a particularly simple form: $q^*(x) = f(x) \cdot \mathbb{1}\{x \geq \tau\}$, and, thus, the optimal value of (MP) equals $n \cdot \mathbb{E}[X \cdot \mathbb{1}\{X \geq \tau\}]$.

Lemma 3.3. *For any continuous distribution $\mathcal{D}$ and $\tau = F^{-1}(1 - \frac{1}{d+1})$, the optimal solution of* (MP) *equals $n \cdot \mathbb{E}[X \cdot \mathbb{1}\{X \geq \tau\}]$.*

Lower bounding the availability of the resource. In order to bound the expected reward collected by Algorithm 1, we first lower bound the probability that the resource is available at any time t. The key idea here is that, since the threshold for accepting a reward is fixed, we can associate the availability state of the system to the evolution of a $d+1$-state time-homogeneous MC, where each state encodes the amount of time until the system is available. Through this link, we first compute the probability that the resource is available when the MC is in stationarity, and then argue on the availability of each round via coupling arguments.

Lemma 3.4. *Let $\{\mathrm{free}_{\mathcal{A}}(t)\}$ be the event that, at the beginning of time t, the resource is available to Algorithm 1. We have that*

$$\Pr[\mathrm{free}_{\mathcal{A}}(t)] \geq \rho(d) - \left(1 - e^{-1}\right)^{\left\lfloor \frac{t-1}{d} \right\rfloor}.$$

Proof. In order to bound the probability that the resource is free, we study the evolution of a Markov Chain (MC) on state space $\Omega = \{0, 1, \ldots, d\}$, with transition probabilities $p_{\omega,\omega-1} = 1$ for each $\omega \in \{1, \ldots, d\}$, $p_{0,d} = \Pr[X \geq \tau] = \frac{1}{d+1}$, and $p_{0,0} = 1 - \frac{1}{d+1}$. Each state $\omega \in \Omega$ represents the number of time steps until the the resource is available to the algorithm, where state 0 implies that the resource is already free. The state transition probabilities reflect the fact that, once blocked, the resource is unavailable *deterministically* for the next d time steps, and, by construction of the algorithm, a reward is collected with probability $\frac{1}{d+1}$ only when the resource is available.

Since the above MC is aperiodic and irreducible, the (unique) stationary distribution can be computed through the following system of balance equations:

$$\sum_{\omega=0}^{d} \pi(\omega) = 1 \qquad \text{and} \qquad \pi(1) = \pi(2) = \ldots = \pi(d) = \frac{1}{d+1} \cdot \pi(0),$$

which gives $\pi^*(0) = \frac{d+1}{2d+1} = \rho(d)$ and $\pi^*(\omega) = \frac{1}{2d+1}$ for each $\omega > 0$.

Now, let $P^{(t)}(0, \cdot)$ denote the t-step transition distribution of the MC initialized at state 0, since, by construction, the resource is free at the first round. Observe that by the above construction, and by definition of the total variation distance[1], we have

$$\Pr[\mathrm{free}_{\mathcal{A}}(t)] = P^{(t-1)}(0,0) = \pi^*(0) + \left(P^{(t-1)}(0,0) - \pi^*(0)\right) \geq \rho(d) - \left\|P^{(t-1)}(0,\cdot) - \pi^*\right\|_{\mathrm{TV}}.$$

Thus, in order to prove the claim of the lemma, it suffices to upper-bound the above total variation distance. The first step is to use a standard upper bound on the total variation distance. By Lemma 4.11 in [36], for any time t, we have

$$\left\|P^{(t)}(0,\cdot) - \pi^*\right\|_{\mathrm{TV}} \leq \max_{s \in \Omega} \left\|P^{(t)}(0,\cdot) - P^{(t)}(s,\cdot)\right\|_{\mathrm{TV}}. \tag{2}$$

[1]Recall that for any two probability measures p and q over a sample space $\Omega = \{0, \ldots, d\}$ we have $\|p - q\|_{\mathrm{TV}} = \sup_{A \subseteq \Omega} |p(A) - q(A)|$.

Let $\omega^* \in \Omega$ denote the state where the above maximum is achieved. Recalling that

$$\left\| P^{(t)}(0,\cdot) - P^{(t)}(\omega^*,\cdot) \right\|_{\text{TV}} = \inf \left\{ \Pr\left[Z_t \neq Y_t\right] : (Z_t, Y_t) \text{ a coupling of } P^{(t)}(0,\cdot) \text{ and } P^{(t)}(\omega^*,\cdot) \right\},$$

we proceed by constructing a coupling in order to bound the total variation distance. Indeed, we construct $\{Z_\ell\}_{\ell=0}^t$ and $\{Y_\ell\}_{\ell=0}^t$ such that $Z_0 = 0$, $Y_0 = \omega^*$. If $Z_\ell = \omega$ for some $\omega > 0$, then $Z_{\ell+1} = \omega - 1$, and similarly for Y_ℓ. At every time ℓ, we flip a coin $A_\ell \sim \text{Bernoulli}\left(1/(d+1)\right)$. If $Z_\ell = 0$, then $Z_{\ell+1} = d$ when $A_\ell = 1$, and $Z_{\ell+1} = 0$ otherwise. Crucially, we use the outcome of the same coin-flip to determine the transition for $Y_{\ell+1}$ when $Y_\ell = 0$. Under this construction, we have that once $Z_\tau = Y_\tau$, then $Z_{\tau+\ell} = Y_{\tau+\ell}$ for *every* $\ell \geq 0$. We call this the *stickiness* property of our coupling. Additionally, since every transition in the MC is deterministic except for at state 0, if h is the first time such that $Z_h = Y_h$, then it must be the case that $Z_h = 0$.

We complete our claim using an amplification argument. Specifically, we first show a constant lower bound on the probability of coupling on every length-d interval, and, then, use the Markov property to "amplify" this bound.

Indeed, let us begin by showing (by induction on k) that, for any $k \geq 0$:

$$\Pr\left[Z_{kd} \neq Y_{kd}\right] \leq \left(1 - e^{-1}\right)^k.$$

Now, the base case of $k = 0$ holds trivially, since the RHS of the above becomes 1. Now, suppose that the claim holds at any fixed $k \geq 0$. Now, by the law of total probability, combined with the fact that our coupling is *sticky*, we may decompose

$$\Pr\left[Z_{(k+1)d} \neq Y_{(k+1)d}\right] = \sum_{z \neq y \in \{0,\ldots,d\}} \Pr\left[Z_{(k+1)d} \neq Y_{(k+1)d} \mid Z_{kd} = z, Y_{kd} = y\right] \Pr\left[Z_{kd} = z, Y_{kd} = y\right].$$

Now, for any fixed $0 \leq z < y \leq d$, by the Markov property and the structure of our Markov chain, we have

$$\begin{aligned}
\Pr\left[Z_{(k+1)d} = Y_{(k+1)d} \mid Z_{kd} = z, Y_{kd} = y\right] &= \Pr\left[Z_{(k+1)d} = Y_{(k+1)d} \mid Z_{kd+z} = 0, Y_{kd+z} = y - z\right] \\
&= \Pr\left[A_{kd+j} = 0 \;\forall\; j \in [z, y) \mid Z_{kd+z} = 0, Y_{kd+z} = y - z\right] \\
&= \left(1 - \frac{1}{d+1}\right)^{y-z} \geq \left(1 - \frac{1}{d+1}\right)^{d} \geq e^{-1},
\end{aligned}$$

where on the last line, we use the inequality $\log(1-x) \geq \frac{-x}{1-x}$ when $x < 1$. A symmetric argument covers the case where $y < z$. Combining the above two observations with our induction hypothesis and the stickiness property of our coupling, we conclude that

$$\Pr\left[Z_{(k+1)d} \neq Y_{(k+1)d}\right] = \Pr\left[Z_{(k+1)d} \neq Y_{(k+1)d} \mid Z_{kd} \neq Y_{kd}\right] \Pr\left[Z_{kd} \neq Y_{kd}\right] \leq \left(1 - e^{-1}\right)^{k+1},$$

as required. Therefore, by another application of the *stickiness* property of our coupling, we may conclude that

$$\left\| P^{(t-1)}(0,\cdot) - P^{(t-1)}(\omega^*,\cdot) \right\|_{\text{TV}} \leq \Pr\left[Z_{t-1} \neq Y_{t-1}\right] = \Pr\left[Z_{t-1} \neq Y_{t-1}, Z_{\lfloor \frac{t-1}{d} \rfloor d} \neq Y_{\lfloor \frac{t-1}{d} \rfloor d}\right] \leq \left(1 - e^{-1}\right)^{\lfloor \frac{t-1}{d} \rfloor},$$

as claimed. □

Wrapping up: proving Theorem 3.1. With the above results in place, we are now ready to prove Theorem 3.1.

Proof of Theorem 3.1. At any time t, for the expected reward collected by Algorithm 1 we have

$$\begin{aligned}\mathbb{E}\left[X_t \cdot \mathbb{1}\left\{X_t \text{ is collected}\right\}\right] &= \mathbb{E}\left[X_t \cdot \mathbb{1}\left\{X_t \geq \tau \text{ and free}_{\mathcal{A}}(t)\right\}\right] \\ &= \mathbb{E}\left[X_t \cdot \mathbb{1}\left\{X_t \geq \tau\right\}\right] \cdot \Pr\left[\text{free}_{\mathcal{A}}(t)\right] \\ &\geq \mathbb{E}\left[X_t \cdot \mathbb{1}\left\{X_t \geq \tau\right\}\right] \rho(d) - \mathbb{E}\left[X\right] \cdot \left(1 - e^{-1}\right)^{\left\lfloor \frac{t-1}{d} \right\rfloor},\end{aligned}$$

where the first equality follows by definition of our policy, and the second since the availability of the resource at time t is independent of the reward realization X_t. The inequality follows by Lemma 3.4 and the fact that $\mathbb{E}\left[X_t \cdot \mathbb{1}\left\{X_t \geq \tau\right\}\right] \leq \mathbb{E}\left[X\right]$.

Therefore, using the above, the expected reward collected can be lower bounded as

$$\begin{aligned}\mathbb{E}\left[\mathsf{ALG}\right] &= \sum_{t\in[n]} \mathbb{E}\left[X_t \cdot \mathbb{1}\left\{X_t \text{ is collected}\right\}\right] \\ &\geq \rho(d) \cdot n \cdot \mathbb{E}\left[X \cdot \mathbb{1}\left\{X \geq \tau\right\}\right] - \mathbb{E}\left[X\right] \sum_{t\in[n]} \left(1 - e^{-1}\right)^{\left\lfloor \frac{t-1}{d} \right\rfloor} \\ &\geq \rho(d) \cdot \mathrm{MP}^* - d \cdot \mathbb{E}\left[X\right] \cdot \sum_{k=0}^{\infty} \left(1 - e^{-1}\right)^{k} \\ &\geq \rho(d) \cdot \mathbb{E}\left[\mathsf{OPT}\right] - \rho(d)(d+1) \cdot \mathbb{E}\left[X\right] - e \cdot d \cdot \mathbb{E}\left[X\right],\end{aligned}$$

where we used the facts that, by Lemmas 3.2 and 3.3, and since $\mathbb{E}\left[\mathsf{OPT}\right] \leq n \cdot \mathbb{E}\left[X\right]$,

$$n \cdot \mathbb{E}\left[X \cdot \mathbb{1}\left\{X \geq \tau\right\}\right] = \mathrm{MP}^* \geq \left(1 - \frac{d+1}{n+d+1}\right) \mathbb{E}\left[\mathsf{OPT}\right] \geq \mathbb{E}\left[\mathsf{OPT}\right] - (d+1) \cdot \mathbb{E}\left[X\right]. \qquad \square$$

4 DESIGNING A REGRET-MINIMIZING POLICY FOR THE LEARNING SETTING

In this section, we study the problem of learning the (optimal) $\rho(d)$-competitive policy described in the previous section, when the reward distribution is initially unknown. In particular, we design a learning policy that estimates the threshold using the empirical distribution constructed from the observed samples. Our goal is to prove a sublinear regret upper bound against the optimal Bayesian prophet inequality of the previous section.

Let us denote the empirical c.d.f. based on samples $Y_1, \ldots, Y_s$ as $\widehat{F}_s(x) = \frac{1}{s}\sum_{i=1}^{s} \mathbb{1}\left\{Y_i \leq x\right\}$. We denote $\widehat{Q}_s(p) = \inf\{x : \widehat{F}_s(x) \geq p\}$ as the empirical quantile function based on s samples drawn from the distribution. For convenience, we assume that for zero samples, we have $\widehat{Q}_0(p) = 1$. Recall that, in this setting, we assume that the distributions are bounded in $[0, 1]$ and, thus, this choice forces the algorithm to reject the first reward almost surely.

With this notation in place, we are ready to present Algorithm 2, the learning variant of Algorithm 1 presented in the previous section. Let N_t be the number of samples available at the beginning of round t (and before observing the realization X_t). Here, at each time t, the gambler constructs the empirical distribution $\widehat{F}_{N_t}$ using N_t observed samples, and computes the empirical threshold $\widehat{\tau}_{N_t}$, where $\widehat{\tau}_s = \widehat{Q}_s\left(1 - 1/d+1\right)$. Then, if it is feasible, it accepts the reward X_t if and only if $X_t \geq \widehat{\tau}_{N_t}$.

We are interested in upper-bounding the regret of Algorithm 2 with respect to an asymptotically-optimal online policy, which in our case corresponds to the $\rho(d)$-approximate regret. The remainder of this section is devoted to proving the following regret guarantee for Algorithm 2:

Algorithm 2: Learning policy for the case of unknown reward distribution

Input: Delay d;
for $t = 1, 2, \ldots$ **do**
  if *resource is available* **then**
    Construct the empirical distribution $\widehat{F}_{N_t}$ using N_t samples (# of samples observed);
    Set threshold $\widehat{\tau}_{N_t} \leftarrow \widehat{Q}_{N_t}\left(1 - \frac{1}{d+1}\right)$;
    Observe reward X_t;
    if $X_t \geq \widehat{\tau}_{N_t}$ **then**
      Collect X_t, and make resource unavailable for rounds $t+1, \ldots, t+d$;
    else
      Skip on X_t;
    end
  else
    Skip the round (without observing X_t);
  end
end

Theorem 4.1. *For any distribution $\mathcal{D}$ bounded in $[0,1]$, and delay d, the $\rho(d)$-approximate regret of Algorithm 2 for n rounds can be upper-bounded as*

$$\mathrm{Regret}_{\rho(d)}(n) \lesssim \sqrt{n \cdot \log n} + d^3 \log(n).$$

4.1 Regret analysis of Algorithm 2

We now present the regret analysis of Algorithm 2. By definition of regret and the fact that Algorithm 1 is a $\rho(d)$-competitive policy (asymptotically), it suffices to measure the difference in total expected reward between Algorithm 1 and Algorithm 2. This provides an upper bound on the regret, as defined in Eq. (1), modulo an additive $O(d)$ loss following by Theorem 3.1. In the rest of this section and for simplicity, we refer to "regret" as the difference in expected reward between the two algorithms.

Recall that Algorithms 1 and 2 operate in the same manner, except for the fact that the latter uses an empirically constructed threshold as a surrogate for $\tau = F^{-1}(1 - 1/(d+1))$ at each round. As a first step in our analysis, we provide a simple upper bound on the regret using the compensated coupling technique due to [46]. This allows us to associate the regret of each round to the estimation error of our empirically constructed threshold. In order to control the estimation error of each round, we provide "anytime" concentration results for the estimator used by Algorithm 2, and we prove strong (high probability) lower bounds on the number of samples collected the beginning of each round. The desired regret upper bound follows by combining the above results.

Compensated Coupling. Let us use $\mathcal{A}$ and $\mathcal{L}$ for any reference to Algorithm 1 and Algorithm 2, respectively. By leveraging the technique of compensated coupling [46], instead of arguing on the regret between $\mathcal{A}$ and $\mathcal{L}$ directly, we first introduce a family of auxiliary (fictitious) policies $\mathcal{P}_t$ in order to facilitate our analysis. More specifically, for each time $t \in [n]$, we denote by $\mathcal{P}_t$ a policy that follows the decisions of $\mathcal{L}$ for the first t time steps (including t) and, after that, follows the decisions of $\mathcal{A}$. Using the above definition, it becomes clear that $\mathcal{P}_0 \equiv \mathcal{A}$ and $\mathcal{P}_n \equiv \mathcal{L}$.

Lemma 4.2. *The difference in expected reward between Algorithms 1 and 2 can be upper-bounded as*

$$\mathbb{E}\left[\mathrm{ALG}_{\mathcal{A}}\right] - \mathbb{E}\left[\mathrm{ALG}_{\mathcal{L}}\right] \leq \sum_{t\in[n]} \Pr\left[X_t \in [\widehat{\tau}_{N_t}, \tau) \cup [\tau, \widehat{\tau}_{N_t})\right].$$

Proof. By using the definition of $\mathcal{P}_t$, we can express the expected difference between $\mathrm{ALG}_{\mathcal{A}}$ and $\mathrm{ALG}_{\mathcal{L}}$ as a telescoping sum:

$$\mathbb{E}\left[\mathrm{ALG}_{\mathcal{A}} - \mathrm{ALG}_{\mathcal{L}}\right] = \mathbb{E}\left[\sum_{t\in[n]} \left(\mathrm{ALG}_{\mathcal{P}_{t-1}} - \mathrm{ALG}_{\mathcal{P}_t}\right)\right] = \sum_{t\in[n]} \mathbb{E}\left[\mathrm{ALG}_{\mathcal{P}_{t-1}} - \mathrm{ALG}_{\mathcal{P}_t}\right]. \tag{3}$$

For any fixed t, the reward collected by policies $\mathcal{P}_{t-1}$ and $\mathcal{P}_t$ running on the same sample path can differ only if the two policies take a different decision at time t (since, otherwise, they follow exactly the same trajectory, by construction). Recalling that $\widehat{\tau}_{N_t}$ is the estimated threshold by policy $\mathcal{L}$ at time t, $\mathcal{P}_{t-1}$ and $\mathcal{P}_t$ deviate at time t only if $X_t \in [\widehat{\tau}_{N_t}, \tau) \cup [\tau, \widehat{\tau}_{N_t})$.

For any t, by the above discussion, we have that

$$\begin{aligned}\mathbb{E}\left[\mathrm{ALG}_{\mathcal{P}_{t-1}} - \mathrm{ALG}_{\mathcal{P}_t}\right] &= \mathbb{E}\left[\left(\mathrm{ALG}_{\mathcal{P}_{t-1}} - \mathrm{ALG}_{\mathcal{P}_t}\right) \mathbb{1}\left\{X_t \in [\widehat{\tau}_{N_t}, \tau) \cup [\tau, \widehat{\tau}_{N_t})\right\}\right] \\ &= \mathbb{E}\left[\left(\mathrm{ALG}_{\mathcal{P}_{t-1}} - \mathrm{ALG}_{\mathcal{P}_t}\right) \left(\mathbb{1}\left\{X_t \in [\widehat{\tau}_{N_t}, \tau)\right\} + \mathbb{1}\left\{X_t \in [\tau, \widehat{\tau}_{N_t})\right\}\right)\right]. \end{aligned} \tag{4}$$

Let us denote by $\mathrm{ALG}_{\mathcal{P}}(t')$ the reward collected by some policy $\mathcal{P}$ at time t'. In the above expression, consider the case where $X_t \in [\widehat{\tau}_{N_t}, \tau)$. In this case, policy $\mathcal{P}_t$ collects the reward X_t (and becomes blocked until time $t + d + 1$), while $\mathcal{P}_{t-1}$ rejects it. Thus, we have that

$$\begin{aligned}&\mathbb{E}\left[\left(\mathrm{ALG}_{\mathcal{P}_{t-1}} - \mathrm{ALG}_{\mathcal{P}_t}\right) \mathbb{1}\left\{X_t \in [\widehat{\tau}_{N_t}, \tau)\right\}\right] \\ &\quad= \mathbb{E}\left[\left(\sum_{t'=t+1}^{n} \mathrm{ALG}_{\mathcal{P}_{t-1}}(t') - \sum_{t'=t+d+1}^{n} \mathrm{ALG}_{\mathcal{P}_t}(t') - X_t\right) \mathbb{1}\left\{X_t \in [\widehat{\tau}_{N_t}, \tau)\right\}\right] \\ &\quad\leq \mathbb{E}\left[\left(\sum_{t'=t+1}^{n-d} \mathrm{ALG}_{\mathcal{P}_{t-1}}(t') + \sum_{t'=n-d+1}^{n} \mathrm{ALG}_{\mathcal{P}_{t-1}}(t') - \sum_{t'=t+d+1}^{n} \mathrm{ALG}_{\mathcal{P}_t}(t')\right) \mathbb{1}\left\{X_t \in [\widehat{\tau}_{N_t}, \tau)\right\}\right].\end{aligned}$$

Consider now any time $t \leq n - d - 1$. Given that $X_t \in [\widehat{\tau}_{N_t}, \tau)$, we know that the resource is available at time $t+1$ for $\mathcal{P}_{t-1}$, and at time $t+d+1$ for $\mathcal{P}_t$. Thus, since the two policies coincide with $\mathcal{A}$ for $t' \geq t+1$, by a simple translation, it is easy to see that $\mathbb{E}\left[\sum_{t'=t+1}^{n-d} \mathrm{ALG}_{\mathcal{P}_{t-1}}(t')\mathbb{1}\left\{X_t \in [\widehat{t'}_{N_t}, \tau)\right\}\right] = \mathbb{E}\left[\sum_{t'=t+d+1}^{n} \mathrm{ALG}_{\mathcal{P}_t}(t')\mathbb{1}\left\{X_t \in [\widehat{\tau}_{N_t}, \tau)\right\}\right]$ for any t (note that when $t > n-d-1$, both expressions are 0, since the summation range is empty). Further, since the rewards are bounded in $[0, 1]$ and at most one reward can be collected on any interval of d rounds, we have that $\sum_{t'=n-d+1}^{n} \mathrm{ALG}_{\mathcal{P}_{t-1}}(t') \leq 1$, deterministically. Combining these observations, we conclude that, for any t,

$$\mathbb{E}\left[\left(\mathrm{ALG}_{\mathcal{P}_{t-1}} - \mathrm{ALG}_{\mathcal{P}_t}\right) \mathbb{1}\left\{X_t \in [\widehat{\tau}_{N_t}, \tau)\right\}\right] \leq \Pr\left[X_t \in [\widehat{\tau}_{N_t}, \tau)\right]. \tag{5}$$

By a similar line of reasoning as above, it is easy to see that

$$\mathbb{E}\left[\left(\mathrm{ALG}_{\mathcal{P}_{t-1}} - \mathrm{ALG}_{\mathcal{P}_t}\right) \mathbb{1}\left\{X_t \in [\tau, \widehat{\tau}_{N_t})\right\}\right] \leq \Pr\left[X_t \in [\tau, \widehat{\tau}_{N_t})\right]. \tag{6}$$

The proof follows by combining Eqs. (3) to (6). □

The above lemma implies that, to upper-bound the regret of Algorithm 2, it suffices to control the error probability $\Pr\left[X_t \in [\widehat{\tau}_{N_t}, \tau) \cup [\tau, \widehat{\tau}_{N_t})\right]$ for any $t \in [n]$.

Quantile estimation via the empirical distribution. For the rest of this section, for any number of samples s, we define:

$$\epsilon_s = \sqrt{\frac{\log\left(2/\delta_s\right) + \log\left(s \cdot (s+1)\right)}{2s}} \qquad \text{and} \qquad \delta_s = \frac{1}{s^2}.$$

Under this notation, we can show the following result:

Lemma 4.3. *Let N_t denote the number of samples collected by Algorithm 2 up to, but not including, time t, from distribution $\mathcal{D}$ with c.d.f. F. Let $X_t \sim \mathcal{D}$ be a sample drawn at time t independently from the past observation history. Then we have that*

$$\Pr\left[X_t \in [\widehat{\tau}_{N_t}, \tau) \cup [\tau, \widehat{\tau}_{N_t}) \text{ and } N_t \geq m\right] \leq \epsilon_m + \delta_m + m^{-1},$$

where $\widehat{\tau}_{N_t}$ is the empirical threshold computed using N_t samples.

We give the proof for Lemma 4.3 below; first, we establish some intermediate inequalities which are useful in proving this result.

Our proof relies on the well-known Dvoretzky–Kiefer–Wolfowitz (DKW) inequality [19]:

Theorem 4.4 (Dvoretzky–Kiefer–Wolfowitz inequality). *Given s samples from a distribution with c.d.f. F, for any $\epsilon > 0$, we have*

$$\Pr\left[\sup_{x \in \mathbb{R}} |\widehat{F}_s(x) - F(x)| > \epsilon\right] \leq 2 \cdot \exp\left(-2s \cdot \epsilon^2\right).$$

Given that the number of observed samples at each round is a random quantity, we can leverage standard techniques to convert Theorem 4.4 into the following "anytime" bound on the concentration of our estimator around its mean:

Lemma 4.5. *Given a random number N_t of samples from a distribution with c.d.f. F, for any $m \in [t-1]$, we have*

$$\Pr\left[\sup_{x \in \mathbb{R}} \left|\widehat{F}_{N_t}(x) - F(x)\right| > \epsilon_{N_t} \text{ and } N_t \geq m\right] \leq \delta_m.$$

Proof. The proof follows essentially from a union bound over the possible values of N_t, combined with Theorem 4.4. Indeed,

$$\begin{aligned}
\Pr\left[\sup_{x \in \mathbb{R}} \left|\widehat{F}_{N_t}(x) - F(x)\right| > \epsilon_{N_t} \text{ and } N_t \geq m\right] &= \sum_{s=m}^{t-1} \Pr\left[N_t = s \text{ and } \sup_{x \in \mathbb{R}} \left|\widehat{F}_s(x) - F(x)\right| > \epsilon_s\right] \\
&\leq \sum_{s=m}^{t-1} \Pr\left[\sup_{x \in \mathbb{R}} \left|\widehat{F}_s(x) - F(x)\right| > \epsilon_s\right] \\
&\leq \sum_{s=m}^{t-1} 2 \cdot \exp\left(-2s \cdot \left(\frac{\log\left(2/\delta_s\right) + \log\left(s \cdot (s+1)\right)}{2s}\right)\right) \\
&\leq \sum_{s=m}^{t-1} \delta_s \cdot \frac{1}{s(s+1)} \\
&\leq \delta_m,
\end{aligned}$$

where the second inequality follows by Theorem 4.4, and the last since δ_s is decreasing in s and $\sum_{s=1}^{\infty} \frac{1}{s(s+1)} = 1$. □

Using the above "anytime" version of the DKW inequality we are able to prove Lemma 4.3, through which we can bound the probability of policy $\mathcal{L}$ making a different decision from $\mathcal{A}$ when $\mathcal{L}$ has collected at least m samples:

Proof of Lemma 4.3. Consider a random sample $X_t \sim \mathcal{D}$ drawn independently from the past observation history. Let us denote

$$\mathcal{N}_t = \left\{ \sup_{x \in \mathbb{R}} \left| \widehat{F}_{N_t}(x) - F(x) \right| \leq \epsilon_{N_t} \right\}.$$

Intuitively, $\mathcal{N}_t$ is a "nice sampling" event when the empirical c.d.f. has been sufficiently well-estimated. Under this notation, we obtain the following decomposition:

$$\begin{aligned} &\Pr\left[X_t \in [\widehat{\tau}_{N_t}, \tau) \cup [\tau, \widehat{\tau}_{N_t}) \text{ and } N_t \geq m\right] \\ &\leq \Pr\left[X_t \in [\widehat{\tau}_{N_t}, \tau) \cup [\tau, \widehat{\tau}_{N_t}) \text{ and } N_t \geq m \mid \mathcal{N}_t\right] + \Pr\left[\neg \mathcal{N}_t \text{ and } N_t \geq m\right], \end{aligned}$$

where the inequality follows by upper-bounding $\Pr[\mathcal{N}_t]$ and $\Pr\left[X_t \in [\widehat{\tau}_{N_t}, \tau) \cup [\tau, \widehat{\tau}_{N_t}) \mid \neg \mathcal{N}_t \text{ and } N_t \geq m\right]$ by 1.

In order to bound the first term above, notice that the event $\{X_t \in [\widehat{\tau}_{N_t}, \tau) \cup [\tau, \widehat{\tau}_{N_t})\}$ is equivalent (by our assumption of continuity of $F(\cdot)$) to the event $\{F(X_t) \in [F(\widehat{\tau}_{N_t}), F(\tau)) \cup [F(\tau), F(\widehat{\tau}_{N_t}))\}$. By definition, $F(\tau) = 1 - \rho(d)$. Further, assuming $\mathcal{N}_t$, it follows that $\left|\widehat{F}_{N_t}(\widehat{\tau}_{N_t}) - F_{N_t}(\widehat{\tau}_{N_t})\right| \leq \epsilon_{N_t}$. By definition of $\widehat{\tau}_{N_t}$, we know that $\widehat{F}_{N_t}(\widehat{\tau}_{N_t}) - (1 - \rho(d)) \in [0, 1/N_t)$ almost surely. Therefore, since $\widehat{\tau}_{N_t}$ is computed *before* observing the sample X_t, we have that

$$\begin{aligned} &\Pr\left[X_t \in [\widehat{\tau}_{N_t}, \tau) \cup [\tau, \widehat{\tau}_{N_t}) \text{ and } N_t \geq m \mid \mathcal{N}_t\right] \\ &= \Pr\left[F(X_t) \in [F(\widehat{\tau}_{N_t}), F(\tau)) \text{ and } N_t \geq m \mid \mathcal{N}_t, \widehat{\tau}_{N_t} \leq \tau\right] \Pr\left[\widehat{\tau}_{N_t} \leq \tau \mid \mathcal{N}_t\right] \\ &\quad + \Pr\left[F(X_t) \in [F(\tau), F(\widehat{\tau}_{N_t})) \text{ and } N_t \geq m \mid \mathcal{N}_t, \widehat{\tau}_{N_t} > \tau\right] \Pr\left[\widehat{\tau}_{N_t} > \tau \mid \mathcal{N}_t\right] \\ &\leq \Pr\left[F(X_t) - (1 - \rho(d)) \in [-\epsilon_{N_t}, 0) \text{ and } N_t \geq m \mid \mathcal{N}_t, \widehat{\tau}_{N_t} \leq \tau\right] \Pr\left[\widehat{\tau}_{N_t} \leq \tau \mid \mathcal{N}_t\right] \\ &\quad + \Pr\left[F(X_t) - (1 - \rho(d)) \in [0, \epsilon_{N_t} + 1/N_t) \text{ and } N_t \geq m \mid \mathcal{N}_t, \widehat{\tau}_{N_t} > \tau\right] \Pr\left[\widehat{\tau}_{N_t} > \tau \mid \mathcal{N}_t\right] \\ &\leq \epsilon_m + m^{-1}, \end{aligned}$$

where in the last inequality we use the fact that $F(X_t)$ is uniformly distributed in $[0, 1]$, when X_t is drawn independently of $\mathcal{N}(t)$ and $\widehat{\tau}_{N_t}$.

The proof follows by combining the above inequalities with the fact that $\Pr[\neg \mathcal{N}_t \text{ and } N_t \geq m] \leq \delta_m$, which follows by Lemma 4.5. □

Sufficiency of samples. With Lemma 4.3 in place, to upper-bound the regret through (4.2), it suffices to find a sufficiently strong lower bound on N_t.

By noticing that any policy observes at least one sample every $d + 1$ rounds, the following deterministic lower bound on the number of samples obtained is immediate:

Fact 4.6. *For any instance of delay d, for the number of samples collected by the policy at the beginning of round t, we have $N_t \geq \lfloor \frac{t-1}{d+1} \rfloor$.*

As it turns out, simply using Fact 4.6 to lower bound N_t would yield an *unnecessary* dependence on the delay parameter d in our regret upper bound. To see that this dependence is unnecessary, let us first consider the following simpler question: how many samples does a policy $\mathcal{A}'$ with a *fixed* threshold $\tau' = F^{-1}\left(1 - 1.5/(d+1)\right)$ "typically" collect? Using the same arguments as in Lemma 3.4, it is straightforward to verify that the stationary resource availability distribution induced by $\mathcal{A}'$ satisfies $\pi'(0) = (d+1)/(2.5d+1)$, and, thus, $\mathcal{A}'$ collects a linear and independent of d number

of samples *in expectation*. Moreover, by appealing to a particular form of the Azuma-Hoeffding inequality [17, Corollary 5.20], we can show that this occurs not only in expectation, but also with *high probability*:

Lemma 4.7. *Let N'_t denote the number of samples available at the beginning of round t to the policy which uses as threshold $\tau' = F^{-1}(1 - {}^{1.5}/_{d+1})$, initialized in an arbitrary availability state $S'_0 \in \{0, \dots, d\}$. Then, with probability at least $1 - \delta$,*

$$N'_t \geq \left(\rho'(d) - e^3 \cdot d \sqrt{\frac{\log({}^1/\delta)}{2(t-1)}} \right) \cdot (t-1) - e^3 \cdot d,$$

where $\rho'(d) = {}^{(d+1)}/_{2.5d+1}$.

Notice that algorithm $\mathcal{A}'$ described above can be thought of as an "eager" version our Bayesian policy, which collects rewards *more frequently* than Algorithm 1. The key-idea is that we can view the threshold set by this eager policy (after a small number of rounds) as a high-probability lower bound on the estimated threshold used by Algorithm 2. As long as this high-probability event occurs, we can show that, by reasoning about a coupled version of $\mathcal{A}'$ and Algorithm 2, the latter collects more samples than $\mathcal{A}'$ (up to a small additive loss). Using these insights, we are able to establish the following key result:

Lemma 4.8. *Let N_t denote the number of samples available to Algorithm 2 at the beginning of round t. Then, with probability at least $1 - \delta$, for any time $t > t_0 := 2(d+1)^3 \log({}^{4n}/\delta)$,*

$$N_t \geq \underline{N}(t) := \left(\rho'(d) - e^3 \cdot d \sqrt{\frac{\log({}^2/\delta)}{2(t-t_0-1)}} \right) \cdot (t - t_0 - 1) - e^3 \cdot d,$$

where $\rho'(d) = {}^{(d+1)}/_{2.5d+1}$.

Proof. Recall that we use $\mathcal{L}$ to refer to Algorithm 2. For the sake of this proof, we define $\mathcal{A}'$ to be a policy that follows the decisions of $\mathcal{L}$ up to (and including) time t_0 and then uses as a fixed threshold the value $\tau' = F^{-1}\left(1 - {}^{1.5}/_{(d+1)}\right)$ for the remaining time steps. Thus, the two algorithms have the same availability state (i.e., number of rounds until the resource is available) at the beginning of time $t_0 + 1$.

The high-level idea behind our proof is the following: first, using the weak lower bound of Fact 4.6, we show that between rounds 1 and t_0, algorithm $\mathcal{L}$ has collected enough information so that its empirical threshold always overestimates that of $\mathcal{A}'$ in rounds $t_0 + 1$ to $t - 1$. Then, using a deterministic *charging argument* we show that, under the above assumption, the number of samples collected by $\mathcal{L}$ between rounds $t_0 + 1$ and $t - 1$ is lower bounded by that of $\mathcal{A}'$. Finally, using Lemma 4.7, we provide a high-probability lower bound on this number of samples collected by $\mathcal{A}'$ (and, hence, by $\mathcal{L}$) between rounds $t_0 + 1$ and $t - 1$.

Let us denote by $N_{[t_1,t_2]}$ (resp., $N'_{[t_1,t_2]}$) the number of samples observed by $\mathcal{L}$ (resp. $\mathcal{A}'$) between rounds t_1 and t_2 (including t_1 and t_2). Using this notation and recalling that N_t (resp., N'_t) is the number of samples collected by $\mathcal{L}$ (resp. $\mathcal{A}'$) at the beginning of round t, we have that

$$\Pr\left[N_t < \underline{N}(t)\right] = \Pr\left[N_{t_0} + N_{[t_0+1,t-1]} < \underline{N}(t)\right] \leq \Pr\left[N_{[t_0+1,t-1]} < \underline{N}(t)\right],$$

where the inequality follows by the fact that N_{t_0} and $N_{[t_0+1,t-1]}$ are non-negative integers. Hence, to prove the lemma it suffices to show that $\Pr\left[N_{[t_0+1,t-1]} < \underline{N}(t)\right] \leq \delta$.

Let us denote $\mathcal{G}_\ell = \mathbb{1}\left\{\widehat{\tau}_{N_\ell} \geq \tau'\right\}$ as the indicator of a "good" event at time ℓ – namely, that the threshold $\widehat{\tau}_{N_\ell}$ used by $\mathcal{L}$ overestimates $\tau' = F^{-1}(1 - {}^{1.5}/_{(d+1)})$. Further, for any $t_2 \geq t_1$, let us denote

by $\mathcal{G}_{[t_1,t_2]} = \bigcap_{\ell=t_1}^{t_2} \mathcal{G}_\ell$ the event that $\mathcal{G}_\ell$ is true for all rounds $\ell \in [t_1, t_2]$. Then, by using the above definitions, we have that

$$\begin{aligned}\Pr\left[N_{[t_0+1,t-1]} < \underline{N}(t)\right] &= \Pr\left[N_{[t_0+1,t-1]} < \underline{N}(t), \neg\mathcal{G}_{[t_0+1,t-1]}\right] + \Pr\left[N_{[t_0+1,t-1]} < \underline{N}(t), \mathcal{G}_{[t_0+1,t-1]}\right] \\ &\leq \underbrace{\Pr\left[\neg\mathcal{G}_{[t_0+1,t-1]}\right]}_{(A)} + \underbrace{\Pr\left[N_{[t_0+1,t-1]} < \underline{N}(t), \mathcal{G}_{[t_0+1,t-1]}\right]}_{(B)}.\end{aligned}$$

In the rest of this proof, we upper-bound each of the terms (A) and (B) above by $\frac{\delta}{2}$.

Upper-bounding term (A). We first note that, for $\tau = F^{-1}\left(1 - \frac{1}{d+1}\right)$, we have

$$\begin{aligned}\Pr\left[\neg\mathcal{G}_{[t_0+1,t-1]}\right] &\leq \sum_{\ell=t_0+1}^{t-1} \Pr\left[\widehat{\tau}_{N_\ell} < \tau'\right] \\ &\leq \sum_{\ell=t_0+1}^{t-1} \Pr\left[F(\widehat{\tau}_{N_\ell}) < F(\tau')\right] \\ &= \sum_{\ell=t_0+1}^{t-1} \Pr\left[F(\widehat{\tau}_{N_\ell}) - \widehat{F}_{N_\ell}(\widehat{\tau}_{N_\ell}) < F(\tau') - F(\tau) + F(\tau) - \widehat{F}_{N_\ell}(\widehat{\tau}_{N_\ell})\right].\end{aligned}$$

Now, by definition of $\widehat{\tau}_{N_\ell}$, we know that $\widehat{F}_{N_\ell}(\widehat{\tau}_{N_\ell}) \geq 1 - 1/(d+1)$ for every ℓ. Further, by definition of τ and τ', $F(\tau') - F(\tau) = -1/2(d+1)$. Thus, combining these facts with the above, we conclude that

$$\begin{aligned}\Pr\left[\neg\mathcal{G}_{[t_0+1,t-1]}\right] &\leq \sum_{\ell=t_0+1}^{t-1} \Pr\left[F(\widehat{\tau}_{N_\ell}) - \widehat{F}_{N_\ell}(\widehat{\tau}_{N_\ell}) < -\frac{1}{2(d+1)})\right] \\ &\leq \sum_{\ell=t_0+1}^{t-1} \Pr\left[\sup_{x\in\mathbb{R}} \left|\widehat{F}_{N_\ell}(x) - F(x)\right| > \frac{1}{2(d+1)}\right] \\ &\leq \sum_{\ell=t_0+1}^{t-1} 2\exp\left(-\left(\left\lfloor\frac{\ell-1}{d+1}\right\rfloor\right)\frac{1}{2(d+1)^2}\right) \\ &\leq \frac{\delta}{2}.\end{aligned}$$

where the third inequality follows by Lemma 4.5 combined with $N_\ell \geq \lfloor (\ell-1)/(d+1) \rfloor$, by Fact 4.6. The last inequality follows by our choice of t_0.

Upper-bounding term (B). The first step to upper-bound term (B) above is to compare through a deterministic charging argument the number of samples collected within rounds $t_0 + 1$ to $t - 1$ by $\mathcal{L}$ with that of $\mathcal{A}'$.

Let us fix any sequence of realized rewards X_ℓ for $\ell \in [1, t-1]$, which satisfies the event $\mathcal{G}_{[t_0+1,t-1]}$ (recall that for fixed realizations the trajectory of $\mathcal{L}$ is deterministic). We argue that in any such realization, it holds $N_{[t_0+1,t-1]} \leq N'_{[t_0+1,t-1]}$, namely, algorithm $\mathcal{L}$ collects at least as many samples as $\mathcal{A}'$.

Let ω_ℓ (resp., ω'_ℓ) be the availability state of algorithm $\mathcal{L}$ (resp. $\mathcal{A}'$) at round t, namely, the number of rounds until the resource becomes available again. Recall that, by definition of $\mathcal{A}'$, the two algorithms meet at round $t_0 + 1$. Further, we call some round ℓ a *meeting point* if the two algorithms meet at state $\omega_\ell = \omega'_\ell = 0$. Then, it suffices to show that between two consecutive meeting points ℓ_1 and ℓ_2, the number of samples collected by $\mathcal{L}$ cannot be less than that of $\mathcal{A}'$.

Suppose the two algorithms reach a meeting point ℓ_1. First, we note that as long as none of the algorithms collects a reward, the difference between the numbers of observed samples does not change. Second, by definition of $\mathcal{G}_{[t_0+1,t-1]}$, the threshold of $\mathcal{L}$ cannot be smaller than that of $\mathcal{A}'$ for any $\ell \in [t_0+1, t-1]$ and, hence, if $\mathcal{L}$ collects a reward at time ℓ_1, so must $\mathcal{A}'$. Now, suppose at some meeting point $\ell_1 \in [t_0+1, t-1]$ algorithm $\mathcal{A}'$ collects a reward and $\mathcal{L}$ does not. Then, if $\mathcal{L}$ does not collect any reward until the next meeting point (where $\mathcal{A}'$ returns to state 0), then $\mathcal{L}$ has trivially observed more rewards than $\mathcal{A}'$ between the two meeting points. In the opposite case, let ν be the number of rounds after ℓ_1 where $\mathcal{L}$ observes rewards while $\mathcal{A}'$ is blocked. This creates an excess of ν in the samples collected by $\mathcal{L}$ compared to $\mathcal{A}'$. Now, notice that in order for this excess to decrease, it has to be that $\mathcal{A}'$ stays at state 0, while $\mathcal{L}$ is blocked. However, every time the excess decreases by 1, algorithm $\mathcal{L}$ comes one step closer to also being available (thus leading to a new meeting point). By the above argument, it is easy to verify that the excess in number of samples of $\mathcal{L}$ against $\mathcal{A}'$ can never become negative.

By the above analysis, for every fixed sample path where $\mathcal{G}_{[t_0+1,t-1]}$ holds, we have that $N_{[t_0+1,t-1]} \leq N'_{[t_0+1,t-1]}$ and, thus

$$\Pr\left[N_{[t_0+1,t-1]} < \underline{N}(t), \mathcal{G}_{[t_0+1,t-1]}\right] \leq \Pr\left[N'_{[t_0+1,t-1]} < \underline{N}(t), \mathcal{G}_{[t_0+1,t-1]}\right] \leq \Pr\left[N'_{[t_0+1,t-1]} < \underline{N}(t)\right].$$

Now, by applying Lemma 4.7 for $\delta/2$ (which allows $\mathcal{A}'$ to start from an arbitrary state) for $t - (t_0+1) + 1 = t - t_0$ rounds, we get that $\Pr\left[N'_{[t_0+1,t-1]} < \underline{N}(t)\right] \leq \frac{\delta}{2}$, which concludes the proof. □

Putting it all together. Given the result of compensated coupling in Lemma 4.2, together with Lemmas 4.3 and 4.8, the proof of Theorem 4.1 is immediate:

Proof of Theorem 4.1. By Lemma 4.2, it suffices to bound $\Pr\left[X_t \in [\widehat{\tau}_{N_t}, \tau) \cup [\tau, \widehat{\tau}_{N_t})\right]$ for each time t. Now, by Lemma 4.8, and taking $t_0 = 2(d+1)^3 \log(4n/\delta)$, if we have that $t \geq t_1 := (t_0+1) + \frac{2e^3 d \log(2/\delta)}{\rho'(d)}$, then with probability at least $1-\delta$, $N_t \geq \rho'(d)/2 \cdot (t - t_1)$. Combining this insight with Lemmas 4.2 and 4.3, it follows that, setting $m_t = \rho'(d)/2 \cdot (t - t_1)$,

$$\begin{aligned}
\mathrm{Regret}_{\rho(d)}(n) &\leq t_1 + \sum_{t>t_1}^{n} \left(\Pr\left[X_t \in [\widehat{\tau}_{N_t}, \tau) \cup [\tau, \widehat{\tau}_{N_t}) \text{ and } N_t \geq m_t\right] + \Pr\left[N_t < m_t\right]\right) \\
&\leq t_1 + n \cdot \delta + \sum_{t>t_1}^{n} \left(\epsilon_{m_t} + m_t^{-1} + \delta_{m_t}\right) \\
&\leq t_1 + n \cdot \delta + \sum_{s=1}^{n-t_1} \left(\sqrt{\frac{2(\log(\rho'(d)^2/2 \cdot s^2) + \log(s(s+1)))}{\rho'(d)s}} + \frac{2}{\rho'(d)s} + \frac{4}{\rho'(d)^2 s^2} \right) \\
&\leq t_1 + n \cdot \delta + \frac{2}{\sqrt{\rho'(d)}} \sqrt{2\left(\log(\rho'(d)^2/2 \cdot n^2) + \log(n(n+1))\right) \cdot n} \\
&\quad + \frac{2}{\rho'(d)}(1 + \log(n)) + \frac{2\pi^2}{3\rho'(d)^2}.
\end{aligned}$$

Therefore, recalling our choices of t_0 and t_1, and choosing $\delta = 1/n$, we conclude that

$$\begin{aligned}\text{Regret}_{\rho(d)}(n) &\leq 4\sqrt{\frac{\log(2n^2)}{\rho'(d)}\cdot n} + 2(d+1)^3\left(1+\frac{1}{(d+1)^3\rho'(d)}\right) + \frac{e^3}{(d+1)^2\rho'(d)}\log(4n^2) \\ &\quad + 2\left(1+\frac{1}{\rho'(d)}\right) + \frac{2\pi^2}{3\rho'(d)^2} \\ &\lesssim \sqrt{n\cdot\log(n)} + d^3\log(n),\end{aligned}$$

as claimed. □

5 UNCONDITIONAL HARDNESS AND REGRET LOWER BOUND

In this section, we study the tightness of the upper bounds presented in Sections 3 and 4. Specifically, we first show that the policy we provide for the Bayesian setting achieves (asymptotically) the optimal competitive guarantee. Then, we show that the $\rho(d)$-regret upper bound we provide for learning this optimal policy has an optimal dependence in the time horizon up to poly-logarithmic factors.

5.1 Unconditional hardness

Our upper bound on the asymptotic competitive guarantee is based on the following construction:

Hard Environment: For any fixed $\epsilon > 0$ and delay $d \geq 1$, we consider a discrete reward distribution, such that $X = 1$ ("small") with probability $1-\epsilon$, and $X = X_{\max} = 1 + \frac{1}{d\cdot\epsilon}$ ("large") with probability ϵ. The time horizon is n.

The following result implies that our Bayesian policy of Section 3 achieves the best possible competitive guarantee (asymptotically). Interestingly, this result has been proven in the setting of contextual blocking bandits [5], which is related to our setting, when the distribution is discrete and has a small finite support. For completeness, we provide a proof of this result in Appendix C.

Theorem 5.1 (Unconditional hardness, [5]). *For any $\epsilon > 0$ and delay $d \geq 1$, there cannot exist a $(\rho(d)+\epsilon)$-competitive algorithm for the asymptotic case of our problem, where $\rho(d) = \frac{d+1}{2d+1}$.*

5.2 Regret lower bound

We now turn our attention to the regret against an optimal gambler's policy. We are able to show the following lower bound on the regret guarantee.

Theorem 5.2 (Regret lower bound). *For any learning policy and any $d \geq 1$, there exists an environment with delay d such that the regret of that policy is at least $\Omega\left(\sqrt{n}/d^{3/2}\right)$.*

For any delay $d \geq 1$, time horizon $n \geq 1$, and parameter $\epsilon \in (0, 1/(d+1))$, the proof of Theorem 5.2 relies on the construction of the following two environments:

Environment $\mathcal{E}_1$ (resp., $\mathcal{E}_2$): We consider a discrete reward distribution, such that $X = 1$ ("small") with probability $1 - 1/(d+1) + \epsilon$ (resp., $1 - 1/(d+1) - \epsilon$), and $X = X_{\max} = \frac{2d+1}{d}$ ("large") with probability $\frac{1}{d+1} - \epsilon$ (resp., $\frac{1}{d+1} + \epsilon$).[2]

It is not hard to see that, at any round t, and for our chosen environments, any policy can be viewed as choosing to play one of the following two natural "*strategies*," which apply both in environments $\mathcal{E}_1$ and $\mathcal{E}_2$ (we make this point clear in the following paragraph):

[2]One might notice that the regret upper bound in Theorem 4.1 assumed all rewards were on the interval $[0,1]$, while here, the rewards are on the interval $[1,3]$ (since $(2d+1)/d \leq 3$). This discrepancy is no issue, however, since is straightforward to show that Theorem 4.1 continues to hold (up to constant factors) as long as the rewards are upper bounded by a constant.

Strategy $\mathcal{S}_1$ (resp., $\mathcal{S}_2$): If the resource is available at time t, accept any reward (resp., accept only the reward $X_{\max} = \frac{2d+1}{d}$). Otherwise, skip the round.

One key difficulty in proving Theorem 5.2 is that, unlike in standard bandit lower bounds, our policies observe each reward *before* making the decision of whether or not to accept it. However, a crucial insight is that, for the environments $\mathcal{E}_1$ and $\mathcal{E}_2$ considered in the lower bound construction, we may assume w.l.o.g. that *any* policy decides before observing the reward whether to play strategy $\mathcal{S}_1$ or $\mathcal{S}_2$ at each round. While this assumption is not generally true, it holds for our choice of $\mathcal{E}_1$ and $\mathcal{E}_2$. In order to see that, we first note that, for the purpose of lower bounding the regret, we can assume that the player knows a priori the support of the reward distribution (which is common in both environments). Further, since the large reward $X = X_{\max} = \frac{2d+1}{d}$ is always collected by both $\mathcal{S}_1$ and $\mathcal{S}_2$ (if the resource is available), any algorithm can be characterized by the probability of playing according to strategy $\mathcal{S}_1$ or $\mathcal{S}_2$ at each round. Thus, since this decision produces a different outcome only when the subsequent reward is small (that is, $X = 1$), any algorithm can simulate this decision before the observing the next reward.

Collected rewards. It is convenient to consider the *asymptotic expected time-averaged reward* collected by an (asymptotically) *optimal*[3] policy, which we denote as:

$$\mathbb{E}\left[\overline{\mathrm{ALG}}^{*}_{\infty}\right] = \lim_{n\to\infty} \sup_{\tau\in\mathbb{R}} \frac{1}{n} \sum_{t\in[n]} \mathbb{E}\left[X_t \cdot \mathbb{1}\left\{X_t > \tau \text{ and } \mathrm{free}(t)\right\}\right]. \tag{7}$$

Let us denote by $\mathrm{Regret}_{\rho(d)}(n;\mathcal{E}_i)$ the regret of a policy under environment $\mathcal{E}_i$. Note that, as a consequence of Lemmas 3.2 and 3.4, and since the rewards offered at each round are at most a constant, the regret in environment $\mathcal{E}_i$ can be lower bounded under this notation as:

$$\mathrm{Regret}_{\rho(d)}(n;\mathcal{E}_i) \geq n \cdot \mathbb{E}\left[\overline{\mathrm{ALG}}^{*}_{\infty}\right] - \mathbb{E}\left[\sum_{t\in[n]} X_t \cdot \mathbb{1}\left\{\mathrm{ALG \text{ collects } } X_t\right\}\right] - O(d). \tag{8}$$

In the next Lemma, we characterize the asymptotically-optimal policy for each of our environments.

Lemma 5.3. *The asymptotically-optimal policy for environment $\mathcal{E}_1$ (resp., $\mathcal{E}_2$) is to play strategy $\mathcal{S}_1$ (resp., $\mathcal{S}_2$) at* every *time step. Further, the asymptotic expected time-averaged reward collected by these policies is*

$$\mathop{\mathbb{E}}_{\mathcal{E}_1}\left[\overline{\mathrm{ALG}}^{*}_{\infty}\right] = \frac{1-\epsilon}{d} \qquad \textit{and} \qquad \mathop{\mathbb{E}}_{\mathcal{E}_2}\left[\overline{\mathrm{ALG}}^{*}_{\infty}\right] = \frac{1+\epsilon\cdot\gamma(d,\epsilon)}{d},$$

where $\gamma(d,\epsilon) := \frac{1+d}{1+d\left(\frac{1}{d+1}+\epsilon\right)}$. *Note that* $\gamma(d,\epsilon) > 1$ *for every* $d \geq 1$ *and* $\epsilon \in (-\frac{1}{d+1}, \frac{1}{d+1})$.

Time-aggregated suboptimality gaps. Another main difficulty in proving Theorem 5.2 is that, unlike in standard bandit lower bound arguments (e.g., [7]), the instantaneous regret can be *negative*. In particular, whenever an algorithm collects the large reward, the optimal policy's expected instantaneous reward is *smaller* than this value.

However, the key insight is that even though a policy might *instantaneously* be better than the expected optimal policy, the resource becomes blocked for the next d time steps and, thus, the policy becomes unable to collect (or even observe) the associated rewards. Therefore, instead of considering what the optimal policy collects *instantaneously*, we consider what it collects *cumulatively* over this $d+1$-time window, so that the regret over this interval is no longer negative.

[3]Note that, as a consequence of Lemma 3.2, *there exists* an asymptotically optimal *threshold-based* policy. Hence, we may restrict our attention to *threshold-based* policies, which accept a reward only it is greater than some threshold τ.

Lemma 5.4. *Let $A_t \in \{\mathcal{S}_1, \mathcal{S}_2\}$ be the strategy chosen by an algorithm $\mathcal{A}$ at time t, and denote by $T_{\mathcal{S}_i}(n) = \sum_{t=1}^{n} \mathbb{1}\{A_t = \mathcal{S}_i \text{ and } \text{free}_{\mathcal{A}}(t)\}$ the number of times over a time horizon n where strategy $\mathcal{S}_i$ is played while the resource is not blocked. Then, the regret in environment $\mathcal{E}_i$ can be lower bounded as*

$$\text{Regret}_{\rho(d)}(n; \mathcal{E}_i) \geq \Delta_{\mathcal{S}_1}^{\mathcal{E}_i} \mathop{\mathbb{E}}_{\mathcal{E}_i}\left[T_{\mathcal{S}_1}(n)\right] + \Delta_{\mathcal{S}_2}^{\mathcal{E}_i} \mathop{\mathbb{E}}_{\mathcal{E}_i}\left[T_{\mathcal{S}_2}(n)\right] - O(d),$$

where $\Delta_{\mathcal{S}_j}^{\mathcal{E}_i}$ is the time-aggregated suboptimality gap *corresponding to the regret incurred by the algorithm for playing strategy $\mathcal{S}_j$ in environment $\mathcal{E}_i$, where:*

$$\Delta_{\mathcal{S}_1}^{\mathcal{E}_i} = (d+1) \cdot \mathop{\mathbb{E}}_{\mathcal{E}_i}\left[\overline{\text{ALG}}_{\infty}^{*}\right] - \mathop{\mathbb{E}}_{\mathcal{E}_i}[X],$$
$$\Delta_{\mathcal{S}_2}^{\mathcal{E}_i} = \mathop{\mathbb{E}}_{\mathcal{E}_i}\left[\overline{\text{ALG}}_{\infty}^{*}\right] \cdot \Pr_{\mathcal{E}_i}[X = 1] + \left((d+1) \mathop{\mathbb{E}}_{\mathcal{E}_i}\left[\overline{\text{ALG}}_{\infty}^{*}\right] - X_{\max}\right) \cdot \Pr_{\mathcal{E}_i}[X = X_{\max}].$$

In particular, $\Delta_{\mathcal{S}_1}^{\mathcal{E}_1} = 0 = \Delta_{\mathcal{S}_2}^{\mathcal{E}_2}$, and

$$\Delta_{\mathcal{S}_2}^{\mathcal{E}_1} = \frac{\epsilon \cdot d + \epsilon^2 \cdot (d+1)}{1+d} = \Theta(\epsilon) \quad \text{and} \quad \Delta_{\mathcal{S}_1}^{\mathcal{E}_2} = \frac{\epsilon \cdot d - \epsilon^2 \cdot (d+1)}{1 + d\left(\frac{1}{d+1} + \epsilon\right)} = \Theta(d \cdot \epsilon).$$

With this regret decomposition in place, we are now ready to prove Theorem 5.2.

Proof of Theorem 5.2. Using the time-aggregation insight from Lemma 5.4, together with the fact that the algorithm *deterministically* collects $n/(d+1)$ samples over n time steps (i.e., $T_{\mathcal{S}_1}(n) + T_{\mathcal{S}_2}(n) \geq n/(d+1)$), we may obtain a regret lower bound in a similar manner as in the stochastic bandit setting (see, e.g., [34, Theorem 15.1]). In particular,

$$\begin{aligned}
&\text{Regret}_{\rho(d)}(n; \mathcal{E}_1) + \text{Regret}_{\rho(d)}(n; \mathcal{E}_2) \\
&\geq \Delta_{\mathcal{S}_2}^{\mathcal{E}_1} \mathop{\mathbb{E}}_{\mathcal{E}_1}\left[T_{\mathcal{S}_2}(n)\right] + \Delta_{\mathcal{S}_1}^{\mathcal{E}_2} \mathop{\mathbb{E}}_{\mathcal{E}_2}\left[T_{\mathcal{S}_1}(n)\right] - O(d) \\
&\geq \Delta_{\mathcal{S}_2}^{\mathcal{E}_1} \mathop{\mathbb{E}}_{\mathcal{E}_1}\left[T_{\mathcal{S}_2}(n)\mathbb{1}\left\{T_{\mathcal{S}_2}(n) \geq \frac{n}{2(d+1)}\right\}\right] + \Delta_{\mathcal{S}_1}^{\mathcal{E}_2} \mathop{\mathbb{E}}_{\mathcal{E}_2}\left[T_{\mathcal{S}_1}(n)\mathbb{1}\left\{T_{\mathcal{S}_2}(n) < \frac{n}{2(d+1)}\right\}\right] - O(d) \\
&\geq \frac{\min\{\Delta_{\mathcal{S}_2}^{\mathcal{E}_1}, \Delta_{\mathcal{S}_1}^{\mathcal{E}_2}\} n}{2(d+1)} \left(\Pr_{\mathcal{E}_1}\left[T_{\mathcal{S}_2}(n) \geq \frac{n}{2(d+1)}\right] + \Pr_{\mathcal{E}_2}\left[T_{\mathcal{S}_2}(n) < \frac{n}{2(d+1)}\right]\right) - O(d) \\
&\geq \frac{\min\{\Delta_{\mathcal{S}_2}^{\mathcal{E}_1}, \Delta_{\mathcal{S}_1}^{\mathcal{E}_2}\} n}{4(d+1)} \exp\left(-D_{\text{KL}}\left(\Pr_{\mathcal{E}_1}[\cdot] \parallel \Pr_{\mathcal{E}_2}[\cdot]\right)\right) - O(d),
\end{aligned}$$

where the first inequality follows by the fact that $\Delta_{\mathcal{S}_i}^{\mathcal{E}_i} = 0$ for $i \in \{1, 2\}$, and the second by the fact that $T_{\mathcal{S}_1} + T_{\mathcal{S}_2} \geq n/(d+1)$, since the resource is available for at least $n/(d+1)$ rounds. Finally, the third inequality follows by the Bretagnolle-Huber inequality [6].

Let $H_n = (A_1, X_1, R_1, \ldots, A_n, X_n, R_n)$ be the sequence of action-observed sample-collected rewards produced by an n-round interaction between the learning policy and the environment. Denote p (resp., p') as the Radon-Nikodym derivative of $\Pr_{\mathcal{E}_1}[\cdot]$ (resp., $\Pr_{\mathcal{E}_2}[\cdot]$). Under this notation, we

have that

$$\begin{aligned} D_{\mathrm{KL}}\left(\Pr_{\mathcal{E}_1}[\cdot] \parallel \Pr_{\mathcal{E}_2}[\cdot]\right) &= \mathop{\mathbb{E}}_{\mathcal{E}_1}\left[\log\left(\frac{p(H_n)}{p'(H_n)}\right)\right] \\ &= \sum_{t=1}^{n} \mathop{\mathbb{E}}_{\mathcal{E}_1}\left[\log\left(\frac{p(A_t \mid H_{t-1})p(X_t \mid H_{t-1}, A_t)p(R_t \mid H_{t-1}, A_t, X_t)}{p'(A_t \mid H_{t-1})p'(X_t \mid H_{t-1}, A_t)p'(R_t \mid H_{t-1}, A_t, X_t)}\right)\right] \\ &= \sum_{t=1}^{n} \mathop{\mathbb{E}}_{\mathcal{E}_1}\left[\log\left(\frac{p(X_t \mid H_{t-1}, A_t)}{p'(X_t \mid H_{t-1}, A_t)}\right)\right] \\ &\leq n \cdot D_{\mathrm{KL}}\left(p(X) \parallel p'(X)\right), \end{aligned}$$

where the first equality follows by definition of KL-divergence and the second by Bayes' rule. The third equality follows by noting that (i) w.l.o.g., the policy decides on an action (or a distribution over actions) before observing the reward and (ii) the collected reward is a deterministic function of the history and current action and sample. The fourth inequality follows since the sample X_t is either (a) not observed in *either* environment (an occurrence which is *completely* determined by the hisory H_{t-1}), or (b) drawn independently from the environment's distribution.

Since the KL-divergence is upper-bounded by the χ^2-distance, we get that

$$D_{\mathrm{KL}}\left(p(X) \parallel p'(X)\right) \leq \chi^2\left(p(X) \parallel p'(X)\right) = \frac{4\epsilon^2}{\left(\frac{1}{d+1}+\epsilon\right)\left(1-\frac{1}{d+1}-\epsilon\right)} \leq 16 \cdot (d+1) \cdot \epsilon^2,$$

where the final inequality holds assuming that $0 < \epsilon < \frac{1}{2(d+1)}$. Finally, by combining the above bounds, we have that

$$\mathrm{Regret}_{\rho(d)}(n; \mathcal{E}_1) + \mathrm{Regret}_{\rho(d)}(n; \mathcal{E}_2) \geq \frac{\min\{\Delta_{\mathcal{S}_2}^{\mathcal{E}_1}, \Delta_{\mathcal{S}_1}^{\mathcal{E}_2}\} \cdot n}{4(d+1)} \cdot \exp\left(-16(d+1) \cdot \epsilon^2 \cdot n\right) - O(d).$$

Noting that $\epsilon < \frac{1}{2(d+1)}$, the expressions from Lemma 5.4 imply that $\min\{\Delta_{\mathcal{S}_2}^{(1)}, \Delta_{\mathcal{S}_1}^{(2)}\} \geq 2\cdot\epsilon/7$. Therefore, taking $\epsilon = \frac{1}{4\sqrt{(d+1)n}}$ (which satifies $\epsilon \leq \frac{1}{2(d+1)}$, since w.l.o.g., $n \geq d \geq \frac{d+1}{4}$ for $d \geq 1$), we then have that

$$\mathrm{Regret}_{\rho(d)}(n; \mathcal{E}_1) + \mathrm{Regret}_{\rho(d)}(n; \mathcal{E}_2) \geq \frac{e^{-1}\sqrt{n}}{56(d+1)^{3/2}} - O(d).$$

Hence, there must exist an environment with regret at least $\Omega\left(\sqrt{n}/d^{3/2}\right)$. □

CONCLUSION AND FURTHER DIRECTIONS

We introduced and studied a practical variant of the IID prophet inequality problem where, once a reward is collected by the decision-maker, she loses her ability to collect *or observe* any reward for a fixed number of rounds. For the Bayesian case of our problem, we designed a simple threshold-based prophet inequality and proved its asymptotic optimality. For the learning case, we showed that the empirical estimate of the threshold from the Bayesian setting has a sufficiently high accuracy, and that using this estimate as a threshold achieves sublinear regret. Moreover, by introducing a notion of *time-aggregated suboptimality gaps*, we were able to reduce a specific instance of our learning problem to that of a two-armed bandit problem, allowing us to prove a lower bound on regret which matches our upper bound up to poly-logarithmic factors.

Our work leaves behind a number of interesting open questions. One natural extension of our model is that of *stochastic delay* of known distribution (possibly correlated with the rewards). We remark, however, that in the case where the delay realization is not observed by the gambler before taking an action, there is no policy with non-trivial competitive guarantee against a prophet which

knows all reward and delay realizations a priori. Another natural extension is to *multiple* (identical or not) reusable resources. For the identical case, while the problem can still be modeled as an LP (similarly to (MP)), it is unclear to us if its optimal solution can still motivate the construction of an efficient (or even threshold-based) algorithm. Finally, it would be interesting to show if there is any advantage a learning policy might receive by observing samples at *every* round, independently of the resource availability. Notice that since, by Lemma 4.8, our policy collects $\Omega(t)$ samples at time t with high probability, any such difference should come from the *deterministic* guarantee on the number of samples available. Identifying the actual dependence on the delay parameter is by itself an interesting direction.

ACKNOWLEDGEMENTS

This research is supported in part by NSF Grants 1826320, 2019844 and 2112471, ONR Grant N00014-19-1-2566, and the Wireless Networking and Communications Group (WNCG) Industrial Affiliates Program.

REFERENCES

[1] Melika Abolhassani, Soheil Ehsani, Hossein Esfandiari, MohammadTaghi Hajiaghayi, Robert Kleinberg, and Brendan Lucier. 2017. Beating 1-1/e for ordered prophets. In *Proceedings of the 49th Annual ACM SIGACT Symp. on Theory of Computing*. 61–71.

[2] Saeed Alaei. 2014. Bayesian combinatorial auctions: Expanding single buyer mechanisms to many buyers. *SIAM J. Comput.* 43, 2 (2014), 930–972.

[3] Saeed Alaei, MohammadTaghi Hajiaghayi, and Vahid Liaghat. 2012. Online Prophet-Inequality Matching with Applications to Ad Allocation. In *Proceedings of the 13th ACM Conf. on Electronic Commerce* (Valencia, Spain) *(EC '12)*. ACM, NY, NY, USA, 18–35.

[4] Pablo D Azar, Robert Kleinberg, and S Matthew Weinberg. 2014. Prophet inequalities with limited information. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*. SIAM, 1358–1377.

[5] Soumya Basu, Orestis Papadigenopoulos, Constantine Caramanis, and Sanjay Shakkottai. 2021. Contextual blocking bandits. In *Int'l Conf. on Artificial Intelligence and Statistics*. PMLR, 271–279.

[6] Jean Bretagnolle and Catherine Huber. 1979. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 47, 2 (1979), 119–137.

[7] Sébastien Bubeck and Nicolo Cesa-Bianchi. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721* (2012).

[8] Constantine Caramanis, Paul Dütting, Matthew Faw, Federico Fusco, Philip Lazos, Stefano Leonardi, Orestis Papadigenopoulos, Emmanouil Pountourakis, and Rebecca Reiffenhäuser. 2022. Single-Sample Prophet Inequalities via Greedy-Ordered Selection. In *Proceedings of the 2022 Annual ACM-SIAM Symp. on Discrete Algorithms (SODA)*. SIAM, 1298–1325.

[9] Shuchi Chawla, Jason D. Hartline, David L. Malec, and Balasubramanian Sivan. 2010. Multi-Parameter Mechanism Design and Sequential Posted Pricing. In *Proceedings of the Forty-Second ACM Symp. on Theory of Computing* (Cambridge, Massachusetts, USA) *(STOC '10)*. ACM, NY, NY, USA, 311–320.

[10] Xinyun Chen, Yunan Liu, and Guiyu Hong. 2020. An online learning approach to dynamic pricing and capacity sizing in service systems. arXiv:2009.02911 [math.PR]

[11] Yuan Shih Chow, Herbert Ellis Robbins, and David Siegmund. 1971. Great expectations: The theory of optimal stopping.

[12] José Correa, Paul Dütting, Felix Fischer, and Kevin Schewior. 2019. Prophet Inequalities for I.I.D. Random Variables from an Unknown Distribution. In *Proceedings of the 2019 ACM Conf. on Economics and Computation* (Phoenix, AZ, USA) *(EC '19)*. ACM, NY, NY, USA, 3–17.

[13] José Correa, Patricio Foncea, Ruben Hoeksma, Tim Oosterwijk, and Tjark Vredeveld. 2017. Posted price mechanisms for a random stream of customers. In *Proceedings of the 2017 ACM Conf. on Economics and Computation*. 169–186.

[14] Jose Correa, Patricio Foncea, Ruben Hoeksma, Tim Oosterwijk, and Tjark Vredeveld. 2019. Recent Developments in Prophet Inequalities. *SIGecom Exch.* 17, 1 (May 2019), 61–70.

[15] José R. Correa, Andrés Cristi, Boris Epstein, and José A. Soto. 2020. The Two-Sided Game of Googol and Sample-Based Prophet Inequalities. In *Proceedings of the 2020 ACM-SIAM Symp. on Discrete Algorithms, SODA 2020, Salt Lake City, UT, USA, January 5-8, 2020*. 2066–2081.

[16] John P. Dickerson, Karthik A. Sankararaman, Aravind Srinivasan, and Pan Xu. 2021. Allocation Problems in Ride-Sharing Platforms: Online Matching with Offline Reusable Resources. *ACM Trans. Econ. Comput.* 9, 3, Article 13 (jun

2021), 17 pages.
[17] Devdatt P Dubhashi and Alessandro Panconesi. 2009. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press.
[18] Paul Dutting, Michal Feldman, Thomas Kesselheim, and Brendan Lucier. 2020. Prophet inequalities made easy: Stochastic optimization by pricing nonstochastic inputs. *SIAM J. Comput.* 49, 3 (2020), 540–582.
[19] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. 1956. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics* (1956), 642–669.
[20] Tomer Ezra, Michal Feldman, Nick Gravin, and Zhihao Gavin Tang. 2020. Online Stochastic Max-Weight Matching: Prophet Inequality for Vertex and Edge Arrival Models. In *Proceedings of the 21st ACM Conf. on Economics and Computation* (Virtual Event, Hungary) *(EC '20)*. ACM, NY, NY, USA, 769–787.
[21] Moran Feldman, Ola Svensson, and Rico Zenklusen. 2016. Online Contention Resolution Schemes. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symp. on Discrete Algorithms* (Arlington, Virginia) *(SODA '16)*. Society for Industrial and Applied Mathematics, USA, 1014–1033.
[22] Amos Fiat, Ilia Gorelik, Haim Kaplan, and Slava Novgorodov. 2015. The temp secretary problem. In *Algorithms-ESA 2015*. Springer, 631–642.
[23] Xiao-Yue Gong, Vineet Goyal, Garud N. Iyengar, David Simchi-Levi, Rajan Udwani, and Shuangyu Wang. 0. Online Assortment Optimization with Reusable Resources. *Management Science* 0, 0 (0), null.
[24] Vineet Goyal, Garud Iyengar, and Rajan Udwani. 2021. Asymptotically Optimal Competitive Ratio for Online Allocation of Reusable Resources. arXiv:2002.02430 [cs.DS]
[25] Nikolai Gravin and Hongao Wang. 2019. Prophet Inequality for Bipartite Matching: Merits of Being Simple and Non Adaptive. In *Proceedings of the 2019 ACM Conf. on Economics and Computation* (Phoenix, AZ, USA) *(EC '19)*. ACM, NY, NY, USA, 93–109.
[26] Mohammad Taghi Hajiaghayi, Robert Kleinberg, and Tuomas Sandholm. 2007. Automated online mechanism design and prophet inequalities. In *AAAI*, Vol. 7. 58–65.
[27] Theodore P Hill and Robert P Kertz. 1982. Comparisons of stop rule and supremum expectations of iid random variables. *The Annals of Probability* 10, 2 (1982), 336–345.
[28] Thomas Jaksch, Ronald Ortner, and Peter Auer. 2010. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research* 11, 4 (2010).
[29] Thomas Kesselheim and Andreas Tönnis. 2016. Think eternally: Improved algorithms for the temp secretary problem and extensions. *arXiv preprint arXiv:1606.06926* (2016).
[30] Robert Kleinberg and Seth Matthew Weinberg. 2012. Matroid prophet inequalities. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*. 123–136.
[31] Ulrich Krengel and Louis Sucheston. 1977. Semiamarts and finite values. *Bull. Amer. Math. Soc.* 83 (1977), 745–747.
[32] Ulrich Krengel and Louis Sucheston. 1978. On semiamarts, amarts, and processes with finite value. *Probability on Banach Spaces* (01 1978), 197–266.
[33] Kailasam Lakshmanan, Ronald Ortner, and Daniil Ryabko. 2015. Improved regret bounds for undiscounted continuous reinforcement learning. In *Int'l Conf. on Machine Learning*. PMLR, 524–532.
[34] Tor Lattimore and Csaba Szepesvári. 2020. *Bandit Algorithms*. Cambridge University Press. https://doi.org/10.1017/9781108571401
[35] Retsef Levi and Ana Radovanović. 2010. Provably Near-Optimal LP-Based Policies for Revenue Management in Systems with Reusable Resources. *Operations Research* 58, 2 (2010), 503–507.
[36] David A Levin and Yuval Peres. 2017. *Markov chains and mixing times*. Vol. 107. American Mathematical Soc.
[37] Brendan Lucier. 2017. An Economic View of Prophet Inequalities. *SIGecom Exch.* 16, 1 (Sept. 2017), 24–47.
[38] Ronald Ortner and Daniil Ryabko. 2013. Online regret bounds for undiscounted continuous reinforcement learning. *arXiv preprint arXiv:1302.2550* (2013).
[39] Jian QIAN, Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. 2019. Exploration Bonus for Regret Minimization in Discrete and Continuous Average Reward MDPs. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc.
[40] Aviad Rubinstein. 2016. Beyond Matroids: Secretary Problem and Prophet Inequality with General Constraints *(STOC '16)*. ACM, NY, NY, USA, 324–332.
[41] Aviad Rubinstein and Sahil Singla. 2017. Combinatorial Prophet Inequalities. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symp. on Discrete Algorithms* (Barcelona, Spain) *(SODA '17)*. Society for Industrial and Applied Mathematics, USA, 1671–1687.
[42] Aviad Rubinstein, Jack Z. Wang, and S. Matthew Weinberg. 2020. Optimal Single-Choice Prophet Inequalities from Samples. In *11th Innovations in Theoretical Computer Science Conf., ITCS 2020, January 12-14, 2020, Seattle, Washington, USA (LIPIcs, Vol. 151)*. 60:1–60:10.

[43] Ester Samuel-Cahn. 1984. Comparison of Threshold Stop Rules and Maximum for Independent Nonnegative Random Variables. *Annals of Probability* 12 (1984), 1213–1216.
[44] Albert N Shiryaev. 2007. *Optimal stopping rules*. Vol. 8. Springer Science & Business Media.
[45] Aristide Tossou, Debabrota Basu, and Christos Dimitrakakis. 2019. Near-optimal optimistic reinforcement learning using empirical bernstein inequalities. *arXiv preprint arXiv:1905.12425* (2019).
[46] Alberto Vera and Siddhartha Banerjee. 2019. The bayesian prophet: A low-regret framework for online decision making. *ACM SIGMETRICS Performance Evaluation Review* 47, 1 (2019), 81–82.

A CONSTRUCTING AN OPTIMAL POLICY IN THE BAYESIAN SETTING: OMITTED PROOFS

Lemma 3.2. *Let* MP* *be an optimal solution of* (MP) *and* OPT *be the reward collected by the prophet. Then, it is the case that*

$$\text{MP}^* \geq \left(1 - \frac{d+1}{n+d+1}\right) \cdot \mathbb{E}\left[\text{OPT}\right].$$

Proof. Let $\hat{q}(x) = \frac{1}{n}\sum_{t=1}^{n} f(x, X_t \in \text{OPT})$, where by OPT we denote both the total reward and the set of rewards collected by the prophet, and f denotes the joint density of a reward $X_t \sim \mathcal{D}$ and the outcome of $\mathbb{1}\{X_t \in \text{OPT}\}$. For the expected prophet's reward, we have:

$$\begin{aligned}\mathbb{E}\left[\text{OPT}\right] &= \sum_{t\in[n]} \int_{x=0}^{\infty} x \cdot f(x, X_t \in \text{OPT}) dx \\ &= n \cdot \int_{x=0}^{\infty} x \cdot \frac{1}{n} \sum_{t\in[n]} f(x, X_t \in \text{OPT}) dx = n \cdot \int_{x=0}^{\infty} x \cdot \hat{q}(x) dx.\end{aligned}$$

For the second set of constraints of (MP) and for any x, we have

$$0 \leq \hat{q}(x) = \frac{1}{n}\sum_{t=1}^{n} f(x, X_t \in \text{OPT}) \leq \frac{1}{n}\sum_{t=1}^{n} f(x, X_t \in \text{OPT}) + f(x, X_t \notin \text{OPT}) = f(x).$$

Finally, for the first set of constraints, we have

$$\int_{x=0}^{\infty} \hat{q}(x) dx = \frac{1}{n}\sum_{t=1}^{n} \int_{x=0}^{\infty} f(x, X_t \in \text{OPT}) dx \leq \frac{1}{n}\left\lceil \frac{n}{d+1} \right\rceil \leq \frac{1}{d+1}\left(1 + \frac{d+1}{n}\right),$$

where we use the fact that for time horizon n and delay d, at most $\left\lceil \frac{n}{d+1} \right\rceil$ elements can be collected.

By the above analysis, it follows immediately that the solution $\tilde{q}(x) = \left(1 + \frac{d+1}{n}\right)^{-1} \hat{q}(x)$ for each x is a feasible solution to (MP) with objective value $\left(1 - \frac{d+1}{n+d+1}\right) \cdot \mathbb{E}\left[\text{OPT}\right]$. This concludes the proof. □

Lemma 3.3. *For any continuous distribution* $\mathcal{D}$ *and* $\tau = F^{-1}(1 - \frac{1}{d+1})$*, the optimal solution of* (MP) *equals* $n \cdot \mathbb{E}\left[X \cdot \mathbb{1}\{X \geq \tau\}\right]$.

Proof. Let λ_x (resp., r_x) be the dual variable corresponding to constraint $q(x) \leq f(x)$ (resp., $q(x) \geq 0$) of (MP) for each $x \geq 0$. We denote by κ the dual variable corresponding to constraint $\int_{x=0}^{\infty} q(x) dx \leq \frac{1}{d+1}$.

The following formulation is the dual of (MP):

$$\begin{aligned} \textbf{minimize:} \quad & \int_{x=0}^{\infty} \lambda_x \cdot f(x)dx + \frac{\kappa}{d+1} && \text{(DMP)} \\ \textbf{s.t.:} \quad & \kappa + \lambda_x - r_x \;\leq\; n \cdot x && \forall x \geq 0 \\ & \lambda_x, r_x \;\geq\; 0 && \forall x \geq 0 \\ & \kappa \;\geq\; 0. \end{aligned}$$

Let $\tau = F^{-1}\left(1 - \frac{1}{d+1}\right)$, where F is the c.d.f. of $\mathcal{D}$. Note that by continuity of $\mathcal{D}$, we have that $\int_{x=\tau}^{\infty} f(x)dx = \frac{1}{d+1}$. We consider the following assignment: (i) We set $q(x) = f(x)$ for all $x \geq \tau$, and $q(x) = 0$, otherwise. (ii) We set $\lambda_x = n \cdot x - n \cdot \tau$ for $x \geq \tau$ and $\lambda_x = 0$, otherwise. (iii) Similarly, we set $r_x = 0$ for $x \geq \tau$ and $r_x = n \cdot \tau - n \cdot x$, otherwise. Finally, (iv) we set $\kappa = n \cdot \tau$.

For the above assignment, the primal objective becomes

$$n \cdot \int_{x=0}^{\infty} x \cdot q(x)dx = n \cdot \int_{x=\tau}^{\infty} x \cdot f(x)dx = n \cdot \mathbb{E}\left[X \cdot \mathbb{1}\left\{X \geq \tau\right\}\right].$$

Similarly, the dual objective becomes

$$\begin{aligned} \int_{x=0}^{\infty} \lambda_x \cdot f(x)dx + \frac{\kappa}{d+1} &= n \cdot \int_{x=\tau}^{\infty} (x - \tau) \cdot f(x)dx + n \cdot \frac{\tau}{d+1} \\ &= n \cdot \int_{x=\tau}^{\infty} x \cdot f(x)dx - n \cdot \tau \cdot \int_{x=\tau}^{\infty} f(x)dx - n \cdot \frac{\tau}{d+1} \\ &= n \cdot \mathbb{E}\left[X \cdot \mathbb{1}\left\{X \geq \tau\right\}\right], \end{aligned}$$

where in the last equality we use the fact that $\int_{x=\tau}^{\infty} f(x)dx = \frac{1}{d+1}$, by definition of τ.

Now that strong duality is established, it suffices to verify the Karush–Kuhn–Tucker (KKT) optimality conditions. The associated Lagrangian is defined as

$$\begin{aligned} L(q, \lambda, \kappa, r) = -n \cdot \int_{x=0}^{\infty} x \cdot q(x)dx + \int_{x=0}^{\infty} \lambda_x \left(q(x) - f(x)\right) dx \\ +\kappa \left(\int_{x=0}^{\infty} q(x)dx - \frac{1}{d+1}\right) - \int_{x=0}^{\infty} r_x \cdot q(x)dx. \end{aligned}$$

It is easy to verify that the above assignment satisfies primal and dual feasibility. Further, for any $x \geq 0$, it can be verified that

$$\frac{\partial L(q, \lambda, \kappa, r)}{\partial q(x)} = nx - \lambda_x - \kappa - r_x = 0.$$

In order to verify the complementary slackness conditions, for each $x \geq 0$, we have

$$\lambda_x \cdot \left(q(x) - f(x)\right) = 0,$$

since $\lambda_x = 0$ if and only if $q(x) < f(x)$. Similarly, for each $x \geq 0$, we have $r_x \cdot q(x) = 0$, since $r_x = 0$ for any $x \geq \tau$, where $q(x) = f(x)$.

Finally, we have

$$\kappa \cdot \left(\int_{x=0}^{\infty} q(x)dx - \frac{1}{d+1}\right) = \kappa \cdot \left(\int_{x=\tau}^{\infty} f(x)dx - \frac{1}{d+1}\right) = 0,$$

since $\int_{x=\tau}^{\infty} f(x)dx = \frac{1}{d+1}$. □

Hence, it follows that $\widehat{F}_s(\widehat{\tau}_s) \leq 1 - \frac{1}{d+1} + \frac{1}{s}$. Further, by construction, $\widehat{F}_s(\widehat{\tau}_s) \geq 1 - \frac{1}{d+1}$. Combining these two bounds proves the claim.

B DESIGNING A REGRET-MINIMIZING POLICY FOR THE LEARNING SETTING: OMITTED PROOFS

Lemma 4.7. *Let N'_t denote the number of samples available at the beginning of round t to the policy which uses as threshold $\tau' = F^{-1}(1 - {}^{1.5}/_{d+1})$, initialized in an arbitrary availability state $S'_0 \in \{0, \dots, d\}$. Then, with probability at least $1 - \delta$,*

$$N'_t \geq \left(\rho'(d) - e^3 \cdot d \sqrt{\frac{\log({}^1/_\delta)}{2(t-1)}} \right) \cdot (t-1) - e^3 \cdot d,$$

where $\rho'(d) = {}^{(d+1)}/_{2.5d+1}$.

Proof. Let us denote by $\mathcal{A}'$ the fixed-threshold policy with threshold $\tau' = F^{-1}(1 - {}^{1.5}/_{d+1})$. We show this result by proving that N'_t satisfies the averaged Lipschitz condition, which implies that we can apply the Azuma-Hoeffding inequality [17, Corollary 5.20].

To begin, let $S'_t \in \{0, \dots, d\}$ denote the availability state of $\mathcal{A}'$ at the end of round t, where $S'_0 \in \{0, \dots, d\}$ is arbitrary. To begin, let us observe that, by the same arguments as used in Lemma 3.4,

$$\mathbb{E}\left[N'_t\right] = \sum_{s=1}^{t-1} \mathbb{E}\left[\mathbb{1}\left\{S'_{s-1} = 0\right\}\right] \geq \rho'(d) \cdot (t-1) - e^3 \cdot d, \tag{9}$$

where $\rho'(d) = {}^{(d+1)}/_{(2.5d+1)}$. We aim to prove that N'_t satisfies the averaged Lipschitz condition with parameter $e^3 \cdot d$, that is, for any time $s \in \{0, \dots, t-1\}$, and for any states $\omega, \omega' \in \{0, \dots, d\}$, we will show that

$$\begin{aligned}
&\left|\mathbb{E}\left[N_t \mid S'_0, \dots, S'_{s-1}, S'_s = \omega\right] - \mathbb{E}\left[N_t \mid S'_0, \dots, S'_{s-1}, S'_s = \omega'\right]\right| \\
&= \left|\sum_{i=s}^{t} \mathbb{E}\left[\mathbb{1}\left\{S'_i = 0\right\} \mid S'_0, \dots, S'_{s-1}, S'_s = \omega\right] - \mathbb{E}\left[\mathbb{1}\left\{S'_i = 0\right\} \mid S'_0, \dots, S'_{s-1}, S'_s = \omega'\right]\right| \\
&\leq e^3 \cdot d.
\end{aligned}$$

Let us denote $\widetilde{S}'_i$ as the process which follows S'_i for the first $s-1$ time steps, and is at state ω' at time s, and evolves according to the analogous coupling as was considered in the proof of Lemma 3.4. Then under this notation, and as a result of Lemma 3.4, we have that

$$\begin{aligned}
&\left|\mathbb{E}\left[N_t \mid S'_0, \dots, S'_{s-1}, S'_s = \omega\right] - \mathbb{E}\left[N_t \mid S'_0, \dots, S'_{s-1}, S'_s = \omega'\right]\right| \\
&= \left|\sum_{i=s}^{t} \mathbb{E}\left[\mathbb{1}\left\{S'_i = 0\right\} - \mathbb{1}\left\{\widetilde{S}'_i = 0\right\} \mid S'_0, \dots, S'_{s-1}, S'_s = \omega, \widetilde{S}'_s = \omega'\right]\right| \\
&\leq \sum_{i=s}^{t} \mathbb{E}\left[\left|\mathbb{1}\left\{S'_i = 0\right\} - \mathbb{1}\left\{\widetilde{S}'_i = 0\right\}\right| \mid S'_0, \dots, S'_{s-1}, S'_s = \omega, \widetilde{S}'_s = \omega'\right] \\
&\leq \sum_{i=s}^{t} \mathbb{E}\left[\mathbb{1}\left\{S'_i \neq \widetilde{S}'_i\right\} \mid S'_0, \dots, S'_{s-1}, S'_s = \omega, \widetilde{S}'_s = \omega'\right] \\
&\leq e^3 \cdot d,
\end{aligned}$$

where the first inequality follows by the triangle inequality and Jensen's inequality, the second follows since if one of S'_i or $\widetilde{S}'_i$ is not 0, then $S'_i \neq \widetilde{S}'_i$, and third follows by the same arguments as in Lemma 3.4.

Hence, by applying (9) together with Azuma-Hoeffding [17, Corollry 5.20], we conclude that

$$\Pr\left[N'_t < \rho'(d)\cdot(t-1) - e^3\cdot d\left(1+\sqrt{\frac{(t-1)\log(1/\delta)}{2}}\right)\right]$$
$$\leq \Pr\left[N'_t < \mathbb{E}\left[N'_t\right] - e^3\cdot d\sqrt{\frac{(t-1)\log(1/\delta)}{2}}\right] \leq \delta,$$

as claimed. □

C UNCONDITIONAL HARDNESS AND REGRET LOWER BOUND: OMITTED PROOFS

Theorem 5.1 (Unconditional hardness, [5]). *For any $\epsilon > 0$ and delay $d \geq 1$, there cannot exist a $(\rho(d)+\epsilon)$-competitive algorithm for the asymptotic case of our problem, where $\rho(d) = \frac{d+1}{2d+1}$.*

Proof. For any fixed $\epsilon > 0$ and delay d, we consider a discrete reward distribution, such that $X = 1$ with probability $1-\epsilon$, and $X = X_{\max} = 1+\frac{1}{d\epsilon}$ with probability ϵ. The time horizon is assumed to be infinite, thus we focus our attention on maximizing the expected average reward.

It is easy to see that any optimal policy can either (i) collect any observed reward, if the resource is available, or (ii) collect the reward only if $X_t = X_{\max}$. Clearly, since $X_{\max} > 1$, no optimal policy skips the reward $X_{\max}$, if it is possible to collect it. For the policy of case (i), we can easily verify that the average expected reward is equal to $\frac{1}{d+1}(\epsilon X_{\max} + 1 - \epsilon)$. For the policy of case (ii), notice that when the resource is available, a reward is collected with probability ϵ (that is, if $X_t = X_{\max}$). By analyzing the underlying Markov Reward Process, it is easy to see that the expected average reward of the second policy is exactly equal to $\epsilon\cdot X_{\max}\cdot\frac{1}{1+d\epsilon}$, where $\frac{1}{1+d\epsilon}$ is the probability that the resource is available. For $X_{\max} = 1+\frac{1}{d\epsilon}$, we can see that both policies have exactly the same average reward, which is equal to $\mathbb{E}[\mathsf{ALG}] = 1/d$. Thus, the expected average reward that can be collected by any gambler in the above instance is exactly $1/d$.

We now turn our attention to the prophet's expected average reward. Given that the analysis of $\mathbb{E}[\mathsf{OPT}]$ is hard, we instead lower bound the prophet's expected reward by considering an approximate prophet that computes a possibly suboptimal solution given knowledge of the reward realizations. For $k \in \mathbb{Z}$, we divide the time horizon into blocks of $k(d+1)$ consecutive time steps. At each block, the approximate prophet operates as follows: (a) if all the rewards of the block are 1, the prophet greedily collects any reward within the block starting from the first time step. On the other hand, (b) if there exists a reward $X_{\max}$ within the first $(k-1)(d+1)+1$ time steps of the block, the prophet collects only this reward within the block. For simplicity, we assume that in any other case the prophet collects no reward from the block. We remark that in the above algorithm, the resource is always available at the beginning of each new block. In this setting, the expected average reward of the approximate prophet is

$$\mathbb{E}[\mathsf{OPT}'] = \frac{1}{k(d+1)}\left((1-\epsilon)^{k(d+1)}\cdot k + ((k-1)(d+1)+1)\,\epsilon(1-\epsilon)^{(k-1)(d-1)}X_{\max}\right)$$
$$= \frac{(1-\epsilon)^{k(d+1)}}{d+1} + \left(1-\frac{1}{k}+\frac{1}{k(d+1)}\right)(1-\epsilon)^{(k-1)(d-1)}\left(\epsilon+\frac{1}{d}\right).$$

Now, by setting $k = \frac{1}{d}\lceil o(\epsilon^{-1})\rceil$ and taking the limit for $\epsilon \to 0$, the expected reward collected by the approximate prophet becomes $\mathbb{E}[\mathsf{OPT}'] = 1/(d+1) + 1/d$. Therefore, the competitive ratio of this instance can be upper-bounded by

$$\frac{\mathbb{E}[\mathsf{ALG}]}{\mathbb{E}[\mathsf{OPT}]} \leq \frac{\mathbb{E}[\mathsf{ALG}]}{\mathbb{E}[\mathsf{OPT}']} = \frac{1/d}{1/(d+1)+1/d} = \frac{d+1}{2d+1} = \rho(d).$$

□

Lemma 5.3. *The asymptotically-optimal policy for environment $\mathcal{E}_1$ (resp., $\mathcal{E}_2$) is to play strategy $\mathcal{S}_1$ (resp., $\mathcal{S}_2$) at* every *time step. Further, the asymptotic expected time-averaged reward collected by these policies is*

$$\mathop{\mathbb{E}}_{\mathcal{E}_1}\left[\overline{\mathrm{ALG}}_\infty^*\right] = \frac{1-\epsilon}{d} \qquad \textit{and} \qquad \mathop{\mathbb{E}}_{\mathcal{E}_2}\left[\overline{\mathrm{ALG}}_\infty^*\right] = \frac{1+\epsilon\cdot\gamma(d,\epsilon)}{d},$$

where $\gamma(d,\epsilon) := \frac{1+d}{1+d\left(\frac{1}{d+1}+\epsilon\right)}$. Note that $\gamma(d,\epsilon) > 1$ for every $d \geq 1$ and $\epsilon \in (-\frac{1}{d+1}, \frac{1}{d+1})$.

Proof. We can easily observe that, when the policy is initialized in the *available* state, by always playing under strategy $\mathcal{S}_1$ it collects a reward *exactly* once every $d+1$ time steps. Thus, at any time $t = k(d+1)$ for some integer $k \geq 0$, in environment $\mathcal{E}_1$, by playing under strategy $\mathcal{S}_1$ the algorithm collects

$$\mathop{\mathbb{E}}_{\mathcal{E}_1}\left[X_{k(d+1)}\right] = 1\cdot\left(1-\frac{1}{d+1}+\epsilon\right) + X_{\max}\cdot\left(\frac{1}{d+1}-\epsilon\right) = \frac{(d+1)(1-\epsilon)}{d}.$$

Similarly, when the policy plays in environment $\mathcal{E}_1$ under strategy $\mathcal{S}_2$, *once it reaches stationarity*[4], at each time t it collects

$$X_{\max}\cdot\Pr_{\mathcal{E}_1}\left[X_t = X_{\max}\right]\cdot\pi_{\mathcal{E}_1}(0) = \frac{1-\epsilon\cdot\gamma(d,-\epsilon)}{d},$$

where $\pi_{\mathcal{E}_1}(\cdot)$ is the stationary distribution of always playing under $\mathcal{S}_2$ in environment $\mathcal{E}_1$, and $\gamma(d,-\epsilon) = \frac{1+d}{1+d\left(\frac{1}{d+1}-\epsilon\right)} > 1$ (since $\epsilon \in (0, 1/d+1)$).

Therefore, noting that the rewards collected in environment $\mathcal{E}_2$ have *exactly* the same expression after replacing ϵ with $-\epsilon$, the (asymptotically) optimal policy for $\mathcal{E}_1$ (resp., $\mathcal{E}_2$) is to play $\mathcal{S}_1$ (resp., $\mathcal{S}_2$) at *every* time step. Using the above expressions, we thus have that

$$\mathop{\mathbb{E}}_{\mathcal{E}_1}\left[\overline{\mathrm{ALG}}_\infty^*\right] = \frac{1-\epsilon}{d} \qquad \text{and} \qquad \mathop{\mathbb{E}}_{\mathcal{E}_2}\left[\overline{\mathrm{ALG}}_\infty^*\right] = \frac{1+\epsilon\cdot\gamma(d,\epsilon)}{d},$$

as claimed. □

Lemma 5.4. *Let $A_t \in \{\mathcal{S}_1, \mathcal{S}_2\}$ be the strategy chosen by an algorithm $\mathcal{A}$ at time t, and denote by $T_{\mathcal{S}_i}(n) = \sum_{t=1}^n \mathbb{1}\{A_t = \mathcal{S}_i \textit{ and } \mathsf{free}_{\mathcal{A}}(t)\}$ the number of times over a time horizon n where strategy $\mathcal{S}_i$ is played while the resource is not blocked. Then, the regret in environment $\mathcal{E}_i$ can be lower bounded as*

$$\mathrm{Regret}_{\rho(d)}(n;\mathcal{E}_i) \geq \Delta_{\mathcal{S}_1}^{\mathcal{E}_i}\mathop{\mathbb{E}}_{\mathcal{E}_i}\left[T_{\mathcal{S}_1}(n)\right] + \Delta_{\mathcal{S}_2}^{\mathcal{E}_i}\mathop{\mathbb{E}}_{\mathcal{E}_i}\left[T_{\mathcal{S}_2}(n)\right] - O(d),$$

where $\Delta_{\mathcal{S}_j}^{\mathcal{E}_i}$ is the time-aggregated suboptimality gap *corresponding to the regret incurred by the algorithm for playing strategy $\mathcal{S}_j$ in environment $\mathcal{E}_i$, where:*

$$\Delta_{\mathcal{S}_1}^{\mathcal{E}_i} = (d+1)\cdot\mathop{\mathbb{E}}_{\mathcal{E}_i}\left[\overline{\mathrm{ALG}}_\infty^*\right] - \mathop{\mathbb{E}}_{\mathcal{E}_i}\left[X\right],$$

$$\Delta_{\mathcal{S}_2}^{\mathcal{E}_i} = \mathop{\mathbb{E}}_{\mathcal{E}_i}\left[\overline{\mathrm{ALG}}_\infty^*\right]\cdot\Pr_{\mathcal{E}_i}\left[X=1\right] + \left((d+1)\mathop{\mathbb{E}}_{\mathcal{E}_i}\left[\overline{\mathrm{ALG}}_\infty^*\right] - X_{\max}\right)\cdot\Pr_{\mathcal{E}_i}\left[X = X_{\max}\right].$$

[4]Note that, by Lemma 3.4, and since $X_{\max} \leq 3$, the expected reward collected by always playing under strategy $\mathcal{S}_2$ in stationarity and initialized as deterministically available, over any time horizon, differs at most by $O(d)$. Thus, when considering the asymptotic time-averaged reward collected by this policy, it suffices to assume it is stationary.

In particular, $\Delta_{\mathcal{S}_1}^{\mathcal{E}_1} = 0 = \Delta_{\mathcal{S}_2}^{\mathcal{E}_2}$, *and*

$$\Delta_{\mathcal{S}_2}^{\mathcal{E}_1} = \frac{\epsilon \cdot d + \epsilon^2 \cdot (d+1)}{1+d} = \Theta(\epsilon) \quad \textit{and} \quad \Delta_{\mathcal{S}_1}^{\mathcal{E}_2} = \frac{\epsilon \cdot d - \epsilon^2 \cdot (d+1)}{1 + d\left(\frac{1}{d+1} + \epsilon\right)} = \Theta(d \cdot \epsilon).$$

Proof. Recalling the *lower bound* for the regret from (8), in order to prove the claimed bound, it suffices to show that

$$n \cdot \mathop{\mathbb{E}}_{\mathcal{E}_i}\left[\overline{\mathrm{ALG}}_\infty^*\right] - \sum_{t\in[n]} \mathbb{E}\left[X_t \cdot \mathbb{1}\left\{\text{ALG collects } X_t\right\}\right] = \Delta_{\mathcal{S}_1}^{\mathcal{E}_i} \mathop{\mathbb{E}}_{\mathcal{E}_i}\left[T_{\mathcal{S}_1}(n)\right] + \Delta_{\mathcal{S}_2}^{\mathcal{E}_i} \mathop{\mathbb{E}}_{\mathcal{E}_i}\left[T_{\mathcal{S}_2}(n)\right] - O(d).$$

Notice that, without loss of generality, at every round t, the algorithm can decide on a strategy $A_t \in \{\mathcal{S}_1, \mathcal{S}_2\}$ *regardless* of whether the resource is available or not. Hence, we may decompose the inner terms of the LHS as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{E}_i}\left[\overline{\mathrm{ALG}}_\infty^*\right] - \mathop{\mathbb{E}}_{\mathcal{E}_i}\left[X_t \cdot \mathbb{1}\left\{\text{ALG collects } X_t\right\}\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{E}_i}\left[\overline{\mathrm{ALG}}_\infty^*\right] \cdot \mathbb{E}\left[\mathbb{1}\left\{A_t = \mathcal{S}_1\right\}\right] - \mathop{\mathbb{E}}_{\mathcal{E}_i}\left[X_t\right] \mathop{\mathbb{E}}_{\mathcal{E}_i}\left[\mathbb{1}\left\{A_t = \mathcal{S}_1, \mathsf{free}(t)\right\}\right] \\
&\quad + \mathop{\mathbb{E}}_{\mathcal{E}_i}\left[\overline{\mathrm{ALG}}_\infty^*\right] \cdot \mathop{\mathbb{E}}_{\mathcal{E}_i}\left[\mathbb{1}\left\{A_t = \mathcal{S}_2, X_t = 1\right\}\right] - 0 \cdot \mathop{\mathbb{E}}_{\mathcal{E}_i}\left[\mathbb{1}\left\{A_t = \mathcal{S}_2, X_t = 1, \mathsf{free}(t)\right\}\right] \\
&\quad + \mathop{\mathbb{E}}_{\mathcal{E}_i}\left[\overline{\mathrm{ALG}}_\infty^*\right] \cdot \mathop{\mathbb{E}}_{\mathcal{E}_i}\left[\mathbb{1}\left\{A_t = \mathcal{S}_2, X_t = X_{\max}\right\}\right] - X_{\max} \mathop{\mathbb{E}}_{\mathcal{E}_i}\left[\mathbb{1}\left\{A_t = \mathcal{S}_2, X_t = X_{\max}, \mathsf{free}(t)\right\}\right].
\end{aligned}$$

Now, at this point, one might be concerned that some of the terms could be *negative*. For example, *at the time* when the algorithm plays $\mathcal{S}_2$ and collects $X_{\max}$, the *asymptotically-optimal* policy only collects $\mathbb{E}[\overline{\mathrm{ALG}}_\infty^*] < X_{\max}$. However, observe that in that case the algorithm *cannot* collect any reward for the next d time steps. Meanwhile, the asymptotically-optimal policy collects $\mathbb{E}[\overline{\mathrm{ALG}}_\infty^*]$ at *every* time step. Applying this observation to the first and third terms above, and assuming that $t \le n - d$, we may *rewrite* the above in the following manner:

$$\begin{aligned}
&\mathbb{E}\left[\overline{\mathrm{ALG}}_\infty^*\right] - \mathbb{E}\left[X_t \cdot \mathbb{1}\left\{\text{ALG collects } X_t\right\}\right] \\
&= \left((d+1)\mathbb{E}\left[\overline{\mathrm{ALG}}_\infty^*\right] - \mathbb{E}\left[X_t\right]\right) \mathbb{E}\left[\mathbb{1}\left\{A_t = \mathcal{S}_1, \mathsf{free}(t)\right\}\right] \\
&\quad + \left(\mathbb{E}\left[\overline{\mathrm{ALG}}_\infty^*\right] - 0\right) \mathbb{E}\left[\mathbb{1}\left\{A_t = \mathcal{S}_2, X_t = 1, \mathsf{free}(t)\right\}\right] \\
&\quad + \left((d+1)\mathbb{E}\left[\overline{\mathrm{ALG}}_\infty^*\right] - X_{\max}\right) \mathbb{E}\left[\mathbb{1}\left\{A_t = \mathcal{S}_2, X_t = X_{\max}, \mathsf{free}(t)\right\}\right].
\end{aligned}$$

Recalling that the choice of A_t, as well as the availability of the resource at time t is *independent* of X_t, and summing the above expression over all $t \in [n]$, we obtain

$$\sum_{t\in[n]} \mathbb{E}\left[\overline{\mathrm{ALG}}_\infty^*\right] - \mathbb{E}\left[X_t \cdot \mathbb{1}\left\{\text{ALG collects } X_t\right\}\right] \quad = \Delta_{\mathcal{S}_1}^{\mathcal{E}_i} \mathbb{E}\left[T_{\mathcal{S}_1}(n)\right] + \Delta_{\mathcal{S}_2}^{\mathcal{E}_i} \mathbb{E}\left[T_{\mathcal{S}_2}(n)\right] - O(d),$$

where the additive loss of $O(d)$ corresponds to the final $O(d)$ rounds for which the time-aggregation arguments above may not (necessarily) apply. Finally, observe that, using the expressions for $\mathbb{E}[\overline{\mathrm{ALG}}_\infty^*]$ from Lemma 5.3, it is easy to verify that $\Delta_{\mathcal{S}_1}^{\mathcal{E}_1} = 0 = \Delta_{\mathcal{S}_2}^{\mathcal{E}_2}$, and that

$$\Delta_{\mathcal{S}_2}^{\mathcal{E}_1} = \frac{\epsilon \cdot d + \epsilon^2 \cdot (d+1)}{1+d} \quad \text{and} \quad \Delta_{\mathcal{S}_1}^{\mathcal{E}_2} = \frac{\epsilon \cdot d - \epsilon^2 \cdot (d+1)}{1 + d\left(\frac{1}{d+1} + \epsilon\right)}$$

□

Received February 2022; revised March 2022; accepted April 2022