

# Fast Learning for Renewal Optimization in Online Task Scheduling

Michael J. Neely

University of Southern California

<https://viterbi-web.usc.edu/~mjneely/>

## Abstract

This paper considers online optimization of a renewal-reward system. A controller performs a sequence of tasks back-to-back. Each task has a random vector of parameters, called the *task type vector*, that affects the task processing options and also affects the resulting reward and time duration of the task. The probability distribution for the task type vector is unknown and the controller must learn to make efficient decisions so that time average reward converges to optimality. Prior work on such renewal optimization problems leaves open the question of optimal convergence time. This paper develops an algorithm with an optimality gap that decays like  $O(1/\sqrt{k})$ , where  $k$  is the number of tasks processed. The same algorithm is shown to have faster  $O(\log(k)/k)$  performance when the system satisfies a strong concavity property. The proposed algorithm uses an auxiliary variable that is updated according to a classic Robbins-Monro iteration. It makes online scheduling decisions at the start of each renewal frame based on this variable and on the observed task type. A matching converse is obtained for the strongly concave case by constructing an example system for which all algorithms have performance at best  $\Omega(\log(k)/k)$ . A matching  $\Omega(1/\sqrt{k})$  converse is also shown for the general case without strong concavity.

## I. INTRODUCTION

Consider a system where a controller performs a sequence of tasks over time (see Fig. 1). The tasks are performed back-to-back so that when task  $k$  ends the task  $k + 1$  immediately begins. The interval of time over which the system performs task  $k \in \{0, 1, 2, \dots\}$  shall be called *frame  $k$* . Fix  $m$  as a positive integer. At the start of each frame  $k \in \{0, 1, 2, \dots\}$  the controller observes a vector  $S[k] \in \mathbb{R}^m$  that determines the *task type*. Components of  $S[k]$  may include parameters that determine the characteristics of task  $k$ . Assume that  $\{S[k]\}_{k=0}^{\infty}$  is independent and identically distributed (i.i.d.) over frames with distribution function

$$F_S(s) = P[S[k] \leq s] \quad \forall s \in \mathbb{R}^m$$

where vector inequality is taken entrywise. The distribution function  $F_S(s)$  is not necessarily known to the controller. After observing  $S[k]$ , the controller chooses to operate in one of various *task processing modes* for the duration of frame  $k$ . The available modes can depend on  $S[k]$ . The  $S[k]$  value and the particular mode that is chosen together determine the *task duration*  $T[k]$  and the *task reward*  $R[k]$  for frame  $k$ . For example,  $R[k]$  can be the monetary profit earned by completing the task on frame  $k$ . In a network scheduling scenario,  $R[k]$  can be the total amount of data transmitted on frame  $k$ . Alternatively, we can have  $R[k] = -P[k]$  where  $P[k]$  is a power cost incurred on frame  $k$ .

Every frame  $k \in \{0, 1, 2, \dots\}$  the controller first observes  $S[k]$  and then chooses a *decision vector*:

$$(T[k], R[k]) \in \mathcal{D}(S[k])$$

where  $\mathcal{D}(S[k])$  is the set of all possible decision vectors available for the task type  $S[k]$  (considering all processing modes). The total reward per unit time, averaged over the first  $K$  tasks, is

$$\frac{\sum_{k=0}^{K-1} R[k]}{\sum_{k=0}^{K-1} T[k]} \quad (1)$$

The goal is to develop an algorithm for making decisions over frames that ensures the limiting average reward per unit time converges to optimality at the fastest possible rate. Long term optimality is defined by all possible algorithms, including algorithms that have knowledge of the probability distribution  $F_S(s)$ . However, convergence time to optimality is considered for algorithms that have no a-priori knowledge of  $F_S(s)$ . For fast convergence, algorithms must quickly learn whatever aspects of the distribution are relevant for making intelligent control decisions that maximize average reward.

This problem is called a *renewal optimization problem* because the system state renews itself on each new frame (when a new  $S[k]$  is observed). For example, consider a stationary and randomized algorithm that, on every frame  $k$ , chooses  $(T[k], R[k]) \in \mathcal{D}(S[k])$  independently of the past using a fixed conditional probability distribution given the observed  $S[k]$ . Then  $\{R[k]\}_{k=0}^{\infty}$  and  $\{T[k]\}_{k=0}^{\infty}$  are i.i.d. and standard renewal theory implies (see, for example, [1]):

$$\lim_{K \rightarrow \infty} \frac{\sum_{k=0}^{K-1} R[k]}{\sum_{k=0}^{K-1} T[k]} = \frac{\mathbb{E}[R[0]]}{\mathbb{E}[T[0]]} \quad \text{with prob 1}$$

In principle one could design the best stationary and randomized algorithm *offline* by choosing, for each possible realization of  $S[k]$ , a conditional probability distribution for  $(T[k], R[k])$  given  $S[k]$ , to maximize the following ratio of expectations:

$$\frac{\mathbb{E}[R[k]]}{\mathbb{E}[T[k]]} = \frac{\mathbb{E}[\mathbb{E}[R[k]|S[k]]]}{\mathbb{E}[\mathbb{E}[T[k]|S[k]]]}$$

This offline design would require knowledge of the probability distribution for  $S[k]$ . The proposed algorithm of this paper is not stationary, and so  $\{R[k]\}_{k=0}^{\infty}$  and  $\{T[k]\}_{k=0}^{\infty}$  are not i.i.d. sequences. The proposed algorithm operates online with no a-priori knowledge of the distribution for  $S[k]$ . It must adapt its decisions by learning from the past.

#### A. Prior work

Optimization of renewal systems is related to linear fractional programming (see, for example, [2][3]). An offline method for optimizing Markov renewal systems via linear fractional programming is in [4]. Online methods for renewal optimization are developed in [5][6], which treat systems with additional time average constraints. The work [5] develops a drift-plus-penalty ratio rule for making decisions that are shown, over time, to satisfy the constraints and achieve a time averaged reward that is arbitrarily close to optimal. The algorithm in [5] requires knowledge of the probability distribution for  $S[k]$ . An approximate implementation is also given in [5] that uses a bisection procedure that does not require knowledge of the probability distribution. This method is applied to treat data center scheduling in [7], asynchronous renewal timelines in [8], power-aware computing in [9], and has connections to the delay-optimal queueing work in [10]. An alternative Robbins-Monro technique is used in [6] and shown to perform well in simulation, but its convergence time is not analyzed. This prior work leaves open the question of optimal convergence time in a renewal system where there is no a-priori knowledge of the task type distribution  $S[k]$ . That question is resolved in the current paper.

The algorithm of the current paper is closely related to the classical Robbins-Monro iteration [11]. The work [11] treats a problem of finding a root  $\theta$  to an equation  $M(x) = 0$  in the case when a nondecreasing function  $M : \mathbb{R} \rightarrow \mathbb{R}$  is unknown and can only be indirectly evaluated. Specifically, on each iteration  $k$  we hand a value  $X[k]$  to an oracle, where  $X[k]$  represents our best guess of the root at time  $k$ , and the oracle returns a random variable  $Y[k]$  whose expectation is equal to  $M(X[k])$ . That is

$$M(x) = \mathbb{E}[Y[k]|X[k] = x] \quad \forall x \in \mathbb{R} \quad (2)$$

The estimated root is then updated via the iteration:

$$X[k+1] = X[k] - \eta[k]Y[k] \quad (3)$$

where  $\{\eta[k]\}_{k=0}^{\infty}$  is a sequence of positive stepsizes. Under certain assumptions, the work [11] shows that  $X[k]$  converges in probability to the root  $\theta$ . The technique of [11] inspired the field of *stochastic approximation* and has been extended to treat minimization of convex functions  $M : \mathbb{R}^n \rightarrow \mathbb{R}$  when an oracle returns stochastic gradients or subgradients [12][13][14][15][16]. Modifications of Robbins-Monro type iterations are considered in [17][18], improvements for binary data are in [19], and applications to Bayesian inference is explored in [20]. If  $M(x)$  is convex then, under a carefully chosen sequence of stepsizes, the optimality gap decays like  $O(1/\sqrt{k})$ , where  $k$  is the number of iterations, while if  $M(x)$  is *strongly* convex then the optimality gap decays like  $O(1/k)$ . Converse results in [12] show these convergence rates cannot be improved. However, an example in [14] shows convergence is sensitive to choosing the stepsize based on knowledge of the strong convexity parameter. When this parameter is over-estimated the convergence can be as slow as  $\Omega(1/k^{1/5})$ . This is used to motivate robust methods in [14].

The current paper uses an iteration similar to (3). However, renewal optimization problems have a different structure from the stochastic approximation problems described in the previous paragraph. For example, optimality for the current paper is related to maximizing a ratio of expectations

$$\frac{\mathbb{E}[R[k]]}{\mathbb{E}[T[k]]}$$

which is different from minimizing the single expectation that defines  $M(x)$  in (2). Furthermore, the expectations  $\mathbb{E}[R[k]]$  and  $\mathbb{E}[T[k]]$  are not determined by a single input parameter  $x$  but by a family of conditional distributions for choosing  $(T[k], R[k])$  given  $S[k]$ . The structure of the randomness is also different: In stochastic approximation we choose a input vector  $X[k]$  and an oracle gives us a noisy version of  $M(X[k])$ , whereas in the current paper the system gives us a random “input” state  $S[k]$  from which we choose a (fully known) output  $(T[k], R[k]) \in \mathcal{D}(S[k])$ . The renewal problem is *online* and all decisions (starting from frame 0) are important in creating time averages that are close to optimal. This is different from problems that seek a single vector  $x$  and do not care what decisions are made in the past as long as they lead to a good eventual choice of  $x$ . Finally, there are no convexity assumptions on the sets  $\mathcal{D}(s)$ . Nevertheless, a connection to Robbins-Monro techniques is made in [5] where it is shown that the optimal time average reward  $\theta^*$  is the root of the following  $M(\theta)$  function

$$M(\theta) = \mathbb{E} \left[ \sup_{(T[k], R[k]) \in \mathcal{D}(S[k])} \{R[k] - \theta T[k]\} \right]$$

where the expectation is with respect to the random  $S[k]$  (that has an unknown distribution).

The current paper uses this observation to motivate a Robbins-Monro style iteration for finding the root  $\theta^*$  using an *auxiliary variable*  $\theta[k]$  on each frame  $k$ . However, it is not enough to simply find a value  $\theta$  that is close to  $\theta^*$ . That is because our goal is to obtain an online algorithm with a time average (starting from frame 0) that converges to  $\theta^*$  as quickly as possible. Fast learning is crucial because early mistakes are included in the time average that we are trying to optimize. The fundamental convergence times for renewal systems that are established in this paper are different from the fundamental convergence times for stochastic approximation in [12].

The focus on convergence time in this paper is conceptually similar to analysis of regret in multi-armed bandit systems, see for example [21][22][23] and a recent renewal-based bandit formulation in [24]. Such problems have an “exploration versus exploitation” structure that is different from the structure of the current paper. In [24], one of multiple “bandit-style” arms is pulled, each arm giving a reward after a random renewal-time. The goal is to learn the best arm to pull. In contrast, rather than choosing an unknown arm and receiving a reward, our system receives a random task state  $S[k]$  and chooses one of multiple (known) control actions  $(T[k], R[k]) \in \mathcal{D}(S[k])$  in a way that most quickly optimizes the system.

## B. Our contributions

This paper develops an online algorithm for making decisions  $(T[k], R[k]) \in \mathcal{D}(S[k])$  on each frame  $k$ , without knowing the distribution for  $S[k]$ , that ensures time average reward converges to the optimal value  $\theta^*$  with probability 1. The algorithm uses an auxiliary variable  $\theta[k]$  that is updated according to a Robbins-Monro type iteration. However, the rate at which  $\theta[k]$  converges to  $\theta^*$  is faster than the rate at which the online averages converge to  $\theta^*$ . Specifically, we show:

- Under general system structure, the proposed algorithm has an optimality gap that decays like  $O(1/\sqrt{k})$ .
- Under special “strong concavity” system structure, the proposed algorithm ensures an improved  $O(\log(k)/k)$  performance. The algorithm is robust in the sense that it does not require knowledge of the strong concavity parameter. However, it turns out that knowledge of a minimum frame size parameter  $T_{min}$  is crucial for selecting the stepsize. Fortunately  $T_{min}$  is easily known in most practical systems.
- Regardless of whether or not the system exhibits strong concavity, the mean squared error between  $\theta[k]$  and  $\theta^*$  decays like  $O(1/k)$ . It is remarkable that convergence of this auxiliary variable is independent of strong concavity and is fundamentally different from convergence of the online time averages (which do depend on strong concavity structure).
- We present a matching  $\Omega(\log(k)/k)$  converse result for systems with strongly concave structure. This is done by constructing an example system for which no algorithm can achieve faster convergence. The proof utilizes a Bernoulli estimation theorem from [25] and a recent mapping technique in [26].
- We also present a matching  $\Omega(1/\sqrt{k})$  converse for an example system without the strongly concave property. The proof has a different structure from the strongly concave case: It shows that for a given  $\epsilon > 0$  and for any algorithm operating on the system, there is a probability parameter for  $S[k]$  under which the algorithm needs at least  $\Omega(1/\epsilon^2)$  frames to ensure performance is within  $\epsilon$  of optimal. This result is similar in spirit to a known square-root converse result for pseudo-regret in multi-arm bandit problems in [21]. However, the square root law does not arise for the same reason as in [21]. Indeed, the two problems are structurally different and use different proofs.

## II. EXAMPLES

### A. Cloud computing with two choices

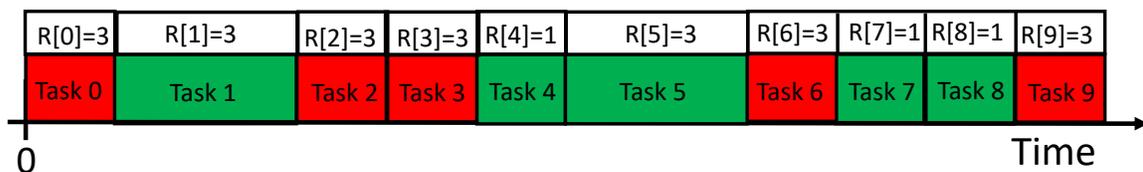


Fig. 1. A sequence of back-to-back tasks of different types: Red tasks ( $S[k] = 0$ ) have  $(T[k], R[k]) = (1, 3)$ . Green tasks ( $S[k] = 1$ ) have  $(T[k], R[k]) \in \{(1, 1), (2, 3)\}$ .

Suppose a cloud computing device performs back-to-back tasks. Each task is one of two types, shown as red or green in Fig. 1. Let  $S[k] \in \{0, 1\}$  be the type of task  $k$  and assume  $\{S[k]\}_{k=0}^{\infty}$  is i.i.d. with  $P[S[k] = 1] = p$ . Red tasks always take 1 unit of time and earn 3 units of revenue. Green tasks earn revenue depending on the *quality* of the processing, and there are only two processing modes available. The decision sets are described in the table below:

Task type	color	Decision set
$S[k] = 0$	red	$\mathcal{D}(0) = \{(1, 3)\}$
$S[k] = 1$	green	$\mathcal{D}(1) = \{(2, 3), (1, 1)\}$

If  $S[k] = 1$  we can choose either high quality processing (which takes 2 units of time and brings revenue 3) or low quality processing (which takes 1 unit of time and brings revenue 1). Which should we choose? Since  $\frac{3}{2} > \frac{1}{1}$ , a naive guess is that it is always optimal to choose high quality. This is not true: It depends on the value of  $p$ . This is because Type 0 tasks are more valuable than Type 1 tasks, so it may be better to quickly get a Type 1 task over with in hopes that the next task is Type 0. It can be shown that it is best to always choose low quality if  $0 \leq p < 1/2$ , and to always choose high quality if  $1/2 \leq p \leq 1$ . If  $p$  is unknown, an intuitively good online control algorithm is to form a running estimate  $\hat{p}[k] = \frac{1}{k} \sum_{j=0}^{k-1} 1_{\{S[j]=1\}}$  and choose high quality whenever  $\hat{p}[k] \geq 1/2$ . Unfortunately this method can make mistakes if  $p \approx 1/2$ . Section [VI](#) shows that no causal algorithm that does not have a-priori knowledge of  $p$  can optimize this system faster than a square root law.

### B. Infinitely many choices

Now suppose the time  $T[k]$  spent on each task can be chosen as any real number in the interval  $[T_{min}, T_{max}]$ , where  $T_{min}$  and  $T_{max}$  are some given positive constants. Suppose the corresponding revenue is:

$$R[k] = A[k]f(T[k]B[k], C[k])$$

where  $f: \mathbb{R}^2 \rightarrow \mathbb{R}$  is some function (possibly discontinuous and nonconvex) that is nondecreasing in the first coordinate. Thus, revenue can be larger if more time is spent on the task. The peculiarities of each task are described by the random vector  $S[k] = (A[k], B[k], C[k])$ , which has an unknown joint cumulative distribution  $F_{A,B,C}(a, b, c)$ . Every frame  $k$  the controller observes  $S[k]$  and chooses  $T[k] \in [T_{min}, T_{max}]$ .

### C. Project selection

Suppose Alice works on one project at a time. On each frame  $k$ , Alice receives a random number  $N[k]$  of new potential projects, each with a different profit and time commitment. Suppose  $N[k]$  has some unknown probability mass function (for example,  $N[k]$  might be a Poisson random variable with some unknown parameter  $\lambda$ ). Let  $(T_j[k], R_j[k])$  represent the time and profit characteristics for each project  $j \in \{1, \dots, N[k]\}$ . The joint distribution for these random parameters is arbitrary and unknown. At the start of frame  $k$ , Alice chooses which single project  $j \in \{1, \dots, N[k]\}$  to work on. She can also choose to work on nothing for one time unit, which yields  $(T[k], R[k]) = (1, 0)$ . This is useful if she believes that none of the current project options are desirable. For example, if there is only one project option and it requires a large time commitment but brings only a small profit, she might reject this project in favor of waiting for a new batch of options. If  $N[k] = 0$  then there are no options and so  $(T[k], R[k]) = (1, 0)$ .

The task type  $S[k]$  formally contains all  $N[k]$  and  $(T_j[k], R_j[k])$  parameters, and the decision set is:<sup>1</sup>

$$\mathcal{D}(S[k]) = \{(1, 0), (T_1[k], R_1[k]), \dots, (T_{N[k]}[k], R_{N[k]}[k])\}$$

How do we know if a particular option, say  $(10.7, 1.2)$ , is good or bad? Can we learn to make good project selection decisions? How fast can we learn?

## III. FORMULATION

As described in the introduction, the task type sequence  $\{S[k]\}_{k=0}^{\infty}$  is i.i.d. over frames. Every frame  $k \in \{0, 1, 2, \dots\}$  the controller observes  $S[k]$  and chooses  $(T[k], R[k]) \in \mathcal{D}(S[k])$ , where  $\mathcal{D}(S[k])$  is a known set of options for  $(T[k], R[k])$  that are available under  $S[k]$ .

### A. Structural assumptions

Assume  $\{S[k]\}_{k=0}^{\infty}$  is an i.i.d. sequence of random vectors that take values in a set  $\Omega_S \subseteq \mathbb{R}^m$ . For each  $s \in \Omega_S$  the decision set  $\mathcal{D}(s)$  is assumed to be a compact subset of  $\mathbb{R}^2$  that satisfies

$$\mathcal{D}(s) \subseteq (0, \infty) \times (-\infty, \infty)$$

This ensures all vectors  $(t, r) \in \mathcal{D}(s)$  have  $t > 0$  (so frame sizes are positive). It is assumed that on each slot  $k$  the decisions  $(T[k], R[k]) \in \mathcal{D}(S[k])$  are made according to some probability law that ensures  $T[k]$  and  $R[k]$  are both random variables, meaning they are *probabilistically measurable*.<sup>2</sup> The probability law can be different for each frame and can depend on observations from previous frames.

<sup>1</sup>There is no loss of generality in assuming  $S[k] \in \mathbb{R}^m$  for some positive integer  $m$ . That is because we can formally pack an infinite sequence of task parameters into a single real number by organizing the digits of its infinite decimal expansion.

<sup>2</sup>The assumption that  $T[k]$  and  $R[k]$  are probabilistically measurable is mild and precludes using the Axiom of Choice to make nonmeasurable decisions.

It is useful to ensure bounds on the first and second moments of  $T[k]$  and  $R[k]$  that hold regardless of the decisions. For this we assume existence of lower bound and upper bound functions  $L : \Omega_S \rightarrow [0, \infty)$  and  $U : \Omega_S \rightarrow [0, \infty)$  that satisfy

$$\inf_{(t,r) \in \mathcal{D}(s)} \{t\} \geq L(s) \quad \forall s \in \Omega_S \quad (4)$$

$$\sup_{(t,r) \in \mathcal{D}(s)} \{t^2 + r^2\} \leq U(s) \quad \forall s \in \Omega_S \quad (5)$$

and for which  $L(S[0])$  and  $U(S[0])$  are random variables with expectations that satisfy

$$\mathbb{E}[L(S[0])] > 0 \quad (6)$$

$$\mathbb{E}[U(S[0])] < \infty \quad (7)$$

An example when such bounding functions exist is when  $\mathcal{D}(s) \subseteq [t_1, t_2] \times [r_1, r_2]$  for all  $s \in \Omega_S$ , where  $t_1, t_2$  and  $r_1, r_2$  are some constants that satisfy  $0 < t_1 \leq t_2$  and  $r_1 \leq r_2$ . In that case we have constant functions  $L(s) = t_1$  and  $U(s) = t_2^2 + \max\{r_1^2, r_2^2\}$  for all  $s \in \Omega_S$ .

From (4)-(7) it follows that regardless of the decisions we have

$$\begin{aligned} T[k] &\geq L(S[k]) \quad \forall k \in \{0, 1, 2, \dots\} \\ T[k]^2 + R[k]^2 &\leq U(S[k]) \quad \forall k \in \{0, 1, 2, \dots\} \end{aligned}$$

and there are constants  $T_{min}, T_{max}, R_{min}, R_{max}, C_1, C_2$  such that for all  $k \in \{0, 1, 2, \dots\}$ :

$$T_{min} > 0 \quad (8)$$

$$T_{min} \leq \mathbb{E}[T[k]] \leq T_{max} \quad (9)$$

$$R_{min} \leq \mathbb{E}[R[k]] \leq R_{max} \quad (10)$$

$$\mathbb{E}[T[k]^2] \leq C_1 \quad (11)$$

$$\mathbb{E}[R[k]^2] \leq C_2 \quad (12)$$

It is assumed the controller knows the values of  $T_{min}, T_{max}, R_{min}, R_{max}$ .

## B. Optimization goal

The goal is to choose decision vectors over time to solve:

$$\text{Maximize: } \liminf_{K \rightarrow \infty} \frac{\sum_{k=0}^{K-1} \mathbb{E}[R[k]]}{\sum_{k=0}^{K-1} \mathbb{E}[T[k]]} \quad (13)$$

$$\text{Subject to: } (T[k], R[k]) \in \mathcal{D}(S[k]) \quad \forall k \in \{0, 1, 2, \dots\} \quad (14)$$

The objective in (13) considers a ratio of expectations, similar to the renewal optimization problems considered in [5][6]. Define  $\theta^*$  as the supremum value of the objective function (13) over all algorithms that satisfy the constraints (14). The value  $\theta^*$  considers all algorithms that result in measurable decision vectors, including algorithms that know, starting from time 0, the full distribution function  $F_S(s)$  and all future values  $\{S[k]\}_{k=0}^{\infty}$ . The key result of this paper is to establish the fundamental *convergence time* required to approach a value close to  $\theta^*$  under the more practical class of algorithms that are *causal* (so they have no knowledge of the future) and *statistics unaware* (so they have no a-priori knowledge of the distribution  $F_S(s)$ ).

## C. Characterizing optimality

This subsection summarizes key facts from [5]. Let  $\mathcal{A} \subseteq [T_{min}, T_{max}] \times [R_{min}, R_{max}]$  be the set of all 1-shot expectations ( $\mathbb{E}[T[0]], \mathbb{E}[R[0]]$ ) achievable on frame 0, considering all possible conditional probability distributions for choosing  $(T[0], R[0]) \in \mathcal{D}(S[0])$  given the observed task type  $S[0]$ . It can be shown that  $\mathcal{A}$  is a convex set. Because  $\{S[k]\}_{k=0}^{\infty}$  is i.i.d. over frames, the set of expectations achievable on slot 0 is the same as the set of expectations achievable on any slot  $k$ . Thus, under any algorithm for making probabilistically measurable decisions over frames we have

$$(\mathbb{E}[T[k]], \mathbb{E}[R[k]]) \in \mathcal{A} \quad \forall k \in \{0, 1, 2, \dots\}$$

and since a convex combination of vectors in the convex set  $\mathcal{A}$  must also be in  $\mathcal{A}$ , we have

$$\frac{1}{K} \sum_{k=0}^{K-1} (\mathbb{E}[T[k]], \mathbb{E}[R[k]]) \in \mathcal{A} \quad \forall K \in \{1, 2, 3, \dots\}$$

Let  $\bar{\mathcal{A}}$  denote the closure of set  $\mathcal{A}$ . Recall that all points  $(t, r) \in \bar{\mathcal{A}}$  have  $t \geq T_{min} > 0$ . It can be shown that the supremum objective  $\theta^*$  for problem (13)-(14) is achievable by a particular (possibly non-stationary) algorithm for making decisions over frames and satisfies:

$$\theta^* = \sup_{(t,r) \in \bar{\mathcal{A}}} \left\{ \frac{r}{t} \right\} \quad (15)$$

The supremum on the right-hand-side of (15) is achievable because  $r/t$  is a continuous function over the compact set  $\bar{\mathcal{A}}$ . In particular, there exists a (possibly non-unique) optimal point  $(t^*, r^*) \in \bar{\mathcal{A}}$  such that

$$\theta^* = \frac{r^*}{t^*}$$

*Lemma 1:* (From [5]) Let  $\theta^*$  be the optimal ratio in (15). Then

$$\sup_{(t,r) \in \bar{\mathcal{A}}} \{r - \theta^* t\} = 0 \quad (16)$$

#### IV. AN ITERATIVE ALGORITHM

Here we develop an algorithm that is causal and that does not know the distribution function  $F_S(s)$ . Assume the values  $T_{min}, T_{max}, R_{min}, R_{max}$  are known to the controller. Let  $\theta_{min}$  and  $\theta_{max}$  be finite and known values that satisfy:

$$\theta_{min} \leq \theta^* \leq \theta_{max} \quad (17)$$

For example, the following values for  $\theta_{min}$  and  $\theta_{max}$  can be used:

$$\theta_{min} = \min \left\{ \frac{R_{min}}{T_{min}}, \frac{R_{min}}{T_{max}} \right\}$$

$$\theta_{max} = \max \left\{ \frac{R_{max}}{T_{min}}, \frac{R_{max}}{T_{max}} \right\}$$

Values of  $\theta_{min}$  and  $\theta_{max}$  that more tightly bracket  $\theta^*$  can be used if the structure of the system allows such tighter values to be known. The following algorithm introduces a sequence of *estimates*  $\{\theta[k]\}_{k=0}^{\infty}$  of  $\theta^*$  as follows: Initialize  $\theta[0] \in [\theta_{min}, \theta_{max}]$  as an arbitrary deterministic value. On each frame  $k \in \{0, 1, 2, \dots\}$  do:

- Observe  $S[k] \in \Omega_S$  and the current  $\theta[k]$  value. Choose  $(T[k], R[k])$  to solve:

$$\text{Maximize: } R[k] - \theta[k]T[k] \quad (18)$$

$$\text{Subject to: } (T[k], R[k]) \in \mathcal{D}(S[k]) \quad (19)$$

breaking ties arbitrarily.<sup>3</sup> There is at least one maximizer because the set  $\mathcal{D}(S[k])$  is compact.

- Update  $\theta[k]$  via the iteration:

$$\theta[k+1] = [\theta[k] + \eta[k](R[k] - \theta[k]T[k])]_{\theta_{min}}^{\theta_{max}} \quad (20)$$

where  $\{\eta[k]\}_{k=0}^{\infty}$  is a deterministic sequence of *step sizes* to be chosen later;  $[x]_{\theta_{min}}^{\theta_{max}}$  denotes a projection of real number  $x$  onto the interval  $[\theta_{min}, \theta_{max}]$ .

The update equation (20) is inspired by the classic Robbins-Monro iteration in (3).

##### A. Technical lemma

*Lemma 2:* Consider algorithm (18)-(20) with any  $\theta[0] \in [\theta_{min}, \theta_{max}]$  and any positive stepsizes  $\{\eta[k]\}_{k=0}^{\infty}$ . For each  $k \in \{0, 1, 2, \dots\}$  and for any versions of  $\mathbb{E}[T[k]|\theta[k]]$  and  $\mathbb{E}[R[k]|\theta[k]]$ , the following holds with probability 1:<sup>4</sup>

$$(\mathbb{E}[T[k]|\theta[k]], \mathbb{E}[R[k]|\theta[k]]) \in \bar{\mathcal{A}} \quad (21)$$

That is, (21) holds for *almost all realizations* of  $\theta[k]$ . Further, there are versions of  $\mathbb{E}[T[k]|\theta[k]]$  and  $\mathbb{E}[R[k]|\theta[k]]$  under which (21) holds *surely for all realizations* of  $\theta[k]$ .

*Proof:* This is a special case of a more general result developed in the Appendix.  $\square$

Intuitively, the above lemma holds because  $\{S[k]\}_{k=0}^{\infty}$  is i.i.d. over frames. Since  $\theta[k]$  depends only on  $\{S[0], \dots, S[k-1]\}$ , the random variables  $\theta[k]$  and  $S[k]$  are independent. Since the decision set  $\mathcal{D}(S[k])$  depends only on  $S[k]$ , knowing  $\theta[k]$  does not change the set of expectations that can be achieved on frame  $k$ . For the rest of this paper  $\mathbb{E}[T[k]|\theta[k]]$  and  $\mathbb{E}[R[k]|\theta[k]]$  shall represent any particular versions of the conditional expectations.

<sup>3</sup>As a minor detail, we note that the tiebreaking rule must conform to a particular probability law that yields probabilistically measurable  $(T[k], R[k])$  variables. An example tiebreaking rule is to choose, among all vectors  $(t, r) \in \mathcal{D}(S[k])$  that tie, the vector with the smallest  $t$ -coordinate.

<sup>4</sup>Recall that if  $X$  and  $Y$  are two random variables with  $\mathbb{E}[|X|] < \infty$  then: (i) There can be multiple versions of the conditional expectation  $\mathbb{E}[X|Y]$ ; (ii) Each version is a random variable that is a deterministic function of  $Y$ ; (iii) Any two versions  $\phi(Y)$  and  $\psi(Y)$  satisfy  $P[\phi(Y) = \psi(Y)] = 1$ .

### B. Analysis of the $(T[k], R[k])$ decision

The following lemmas use  $\theta^*$  as the optimal ratio in (15) and assume  $(t^*, r^*)$  is a vector in  $\bar{\mathcal{A}}$  that satisfies  $\theta^* = r^*/t^*$ .

*Lemma 3:* Consider the algorithm (18)-(20) with any  $\theta[0] \in [\theta_{min}, \theta_{max}]$  and any positive stepsizes  $\{\eta[k]\}_{k=0}^{\infty}$ . For each  $k \in \{0, 1, 2, \dots\}$  we have for almost all realizations of  $\theta[k]$ :

$$\mathbb{E}[R[k] - \theta[k]T[k]|\theta[k]] \geq t^*(\theta^* - \theta[k]) \quad (22)$$

and

$$\mathbb{E}[R[k] - \theta^*T[k]|\theta[k]] \geq (\theta[k] - \theta^*)\mathbb{E}[(T[k] - t^*)|\theta[k]] \quad (23)$$

That is, the probability that random variable  $\theta[k]$  does not satisfy both (22) and (23) is 0.

*Proof:* Fix  $k \in \{0, 1, 2, \dots\}$ . The decision vector  $(T[k], R[k])$  is chosen by observing  $\theta[k]$  and  $S[k]$  and selecting the vector in  $\mathcal{D}(S[k])$  that maximizes  $R[k] - \theta[k]T[k]$ , and so

$$R[k] - \theta[k]T[k] \geq R^*[k] - \theta[k]T^*[k]$$

where  $(T^*[k], R^*[k])$  is any other (potentially randomized) vector in  $\mathcal{D}(S[k])$ . Taking conditional expectations gives (with probability 1)<sup>5</sup>

$$\mathbb{E}[R[k] - \theta[k]T[k]|\theta[k]] \geq \mathbb{E}[R^*[k]|\theta[k]] - \theta[k]\mathbb{E}[T^*[k]|\theta[k]] \quad (24)$$

Note that  $\theta[k]$  depends on history in the system that occurred before frame  $k$ , and in particular  $\theta[k]$  is independent of  $S[k]$ . Fix  $(t, r) \in \mathcal{A}$ . By definition of  $\mathcal{A}$ , there is a conditional distribution for choosing  $(T[0], R[0]) \in \mathcal{D}(S[0])$ , given the observed  $S[0]$ , such that  $(\mathbb{E}[T[0]], \mathbb{E}[R[0]]) = (t, r)$ . Since  $S[k]$  is independent of  $\theta[k]$  and has the same distribution as  $S[0]$ , we can use the same conditional distribution to produce a random vector  $(T^*[k], R^*[k]) \in \mathcal{D}(S[k])$  that is independent of  $\theta[k]$  such that

$$(\mathbb{E}[T^*[k]], \mathbb{E}[R^*[k]]) = (t, r)$$

and since  $(T^*[k], R^*[k])$  is independent of  $\theta[k]$  we have (with probability 1):

$$\begin{aligned} \mathbb{E}[T^*[k]|\theta[k]] &= t \\ \mathbb{E}[R^*[k]|\theta[k]] &= r \end{aligned}$$

Substituting these identities into (24) gives (with probability 1):

$$\mathbb{E}[R[k] - \theta[k]T[k]|\theta[k]] \geq r - \theta[k]t \quad (25)$$

Let  $(t^*, r^*)$  be a vector in  $\bar{\mathcal{A}}$  that satisfies  $\theta^* = r^*/t^*$ . Since  $(t^*, r^*)$  is in the closure of the set  $\mathcal{A}$ , there is a sequence of points  $\{(t_i, r_i)\}_{i=1}^{\infty}$  in  $\mathcal{A}$  that converge to the value  $(t^*, r^*)$ . Since (25) was shown to hold (with probability 1) for an arbitrary  $(t, r) \in \mathcal{A}$ , with probability 1 it holds simultaneously for all of the (countably many)  $(t_i, r_i) \in \mathcal{A}$  for  $i \in \{1, 2, 3, \dots\}$  and so<sup>6</sup>

$$\mathbb{E}[R[k] - \theta[k]T[k]|\theta[k]] \geq r_i - \theta[k]t_i \quad \forall i \in \{1, 2, 3, \dots\}$$

Taking a limit as  $i \rightarrow \infty$  yields (with probability 1):

$$\begin{aligned} \mathbb{E}[R[k] - \theta[k]T[k]|\theta[k]] &\geq r^* - \theta[k]t^* \\ &= t^*(\theta^* - \theta[k]) \end{aligned}$$

where the final equality uses  $\theta^* = r^*/t^*$ . This proves (22). Adding  $(\theta[k] - \theta^*)\mathbb{E}[T[k]|\theta[k]]$  to both sides proves (23).  $\square$

*Lemma 4:* Consider the algorithm (18)-(20) with any  $\theta[0] \in [\theta_{min}, \theta_{max}]$  and any positive stepsizes  $\{\eta[k]\}_{k=0}^{\infty}$ . We have for all positive integers  $K$ :

$$\left| \theta^* - \frac{\sum_{k=0}^{K-1} \mathbb{E}[R[k]]}{\sum_{k=0}^{K-1} \mathbb{E}[T[k]]} \right| \leq \frac{1}{KT_{min}} \sum_{k=0}^{K-1} \mathbb{E}[|\theta[k] - \theta^*| \cdot |T[k] - t^*|] \quad (26)$$

where  $T_{min}$  satisfies (8)-(9).

*Proof:* Fix  $k \in \{0, 1, 2, \dots\}$ . We have  $(\mathbb{E}[T[k]], \mathbb{E}[R[k]]) \in \bar{\mathcal{A}}$  and so

$$\mathbb{E}[R[k] - \theta^*\mathbb{E}[T[k]]] \leq \sup_{(t,r) \in \bar{\mathcal{A}}} [r - \theta^*t] = 0$$

<sup>5</sup>This uses the measure theory fact that if  $X$  and  $Y$  are random variables with finite expectations that satisfy  $X - Y \geq 0$  surely, and if  $Z$  is another random variable, then  $\mathbb{E}[X - Y|Z] \geq 0$  with probability 1.

<sup>6</sup>Recall that if  $\{F_i\}_{i=1}^{\infty}$  is an infinite sequence of events that satisfy  $P[F_i] = 1$  for all  $i \in \{1, 2, 3, \dots\}$  then  $P[\cap_{i=1}^{\infty} F_i] = 1$ .

where the final equality holds by Lemma 1. Summing over  $k \in \{0, \dots, K-1\}$  gives

$$\sum_{k=0}^{K-1} \mathbb{E} [R[k] - \theta^* T[k]] \leq 0$$

Rearranging terms yields

$$\theta^* \geq \frac{\sum_{k=0}^{K-1} \mathbb{E} [R[k]]}{\sum_{k=0}^{K-1} \mathbb{E} [T[k]]} \quad (27)$$

On the other hand, we can take expectations of (23) and use the law of iterated expectations to obtain

$$\mathbb{E} [R[k] - \theta^* T[k]] \geq \mathbb{E} [(\theta[k] - \theta^*)(T[k] - t^*)]$$

Summing this over  $k \in \{0, \dots, K-1\}$  and rearranging terms yields

$$\begin{aligned} \frac{\sum_{k=0}^{K-1} \mathbb{E} [R[k]]}{\sum_{k=0}^{K-1} \mathbb{E} [T[k]]} &\geq \theta^* + \frac{1}{\sum_{k=0}^{K-1} \mathbb{E} [T[k]]} \sum_{k=0}^{K-1} \mathbb{E} [(\theta[k] - \theta^*)(T[k] - t^*)] \\ &\geq \theta^* - \frac{1}{KT_{min}} \sum_{k=0}^{K-1} \mathbb{E} [|\theta[k] - \theta^*| \cdot |T[k] - t^*|] \end{aligned} \quad (28)$$

Combining (27) and (28) proves the lemma.  $\square$

To make the right-hand-side of (26) close to 0, it is desirable to make the  $\theta[k]$  parameter close to  $\theta^*$ .

### C. Analysis of the update rule

Consider the algorithm (18)-(20) with any  $\theta[0] \in [\theta_{min}, \theta_{max}]$  and any positive stepsizes. Define  $b$  as a constant that satisfies

$$\frac{1}{2} \mathbb{E} [(R[k] - \theta[k]T[k])^2] \leq b \quad \forall k \in \{0, 1, 2, \dots\} \quad (29)$$

Such a constant  $b$  exists because  $\theta[k] \in [\theta_{min}, \theta_{max}]$  and the  $R[k]$  and  $T[k]$  variables satisfy the first and second moment bounds (8)-(12). The following lemma is similar in spirit to the analysis of Robbins-Monro iterations for different systems in [11][14].

*Lemma 5:* Under algorithm (18)-(20) with any  $\theta[0] \in [\theta_{min}, \theta_{max}]$  and any positive stepsizes  $\{\eta[k]\}_{k=0}^{\infty}$ , we have for all frames  $k \in \{0, 1, 2, \dots\}$

$$\frac{1}{2} \mathbb{E} [(\theta[k+1] - \theta^*)^2] \leq \left( \frac{1}{2} - T_{min} \eta[k] \right) \mathbb{E} [(\theta[k] - \theta^*)^2] + \eta[k]^2 b \quad (30)$$

where the constant  $b$  satisfies (29); the constant  $T_{min}$  satisfies (8)-(9); the constant  $\theta^*$  is defined by (15).

*Proof:* Fix  $k \in \{0, 1, 2, \dots\}$ . For simplicity of notation define  $z[k] = \theta[k] + \eta[k](R[k] - \theta[k]T[k])$ . Recall that  $[x]_{\theta_{min}}^{\theta_{max}}$  denotes a projection of the real number  $x$  onto the interval  $[\theta_{min}, \theta_{max}]$ . By (20) we have

$$\begin{aligned} (\theta[k+1] - \theta^*)^2 &= ([z[k]]_{\theta_{min}}^{\theta_{max}} - \theta^*)^2 \\ &\stackrel{(a)}{=} ([z[k]]_{\theta_{min}}^{\theta_{max}} - [\theta^*]_{\theta_{min}}^{\theta_{max}})^2 \\ &\stackrel{(b)}{\leq} (z[k] - \theta^*)^2 \end{aligned}$$

where (a) uses the fact that  $\theta^* \in [\theta_{min}, \theta_{max}]$ ; (b) uses the fact that the distance between the projections of two real numbers onto a closed interval is less than or equal to the distance between those real numbers. Thus

$$\begin{aligned} \frac{1}{2} (\theta[k+1] - \theta^*)^2 &\leq \frac{1}{2} (\theta[k] - \theta^* + \eta[k](R[k] - \theta[k]T[k]))^2 \\ &= \frac{1}{2} (\theta[k] - \theta^*)^2 + \frac{\eta[k]^2}{2} (R[k] - \theta[k]T[k])^2 + \eta[k](\theta[k] - \theta^*)(R[k] - \theta[k]T[k]) \end{aligned}$$

Taking expectations gives

$$\frac{1}{2} \mathbb{E} [(\theta[k+1] - \theta^*)^2] \leq \frac{1}{2} \mathbb{E} [(\theta[k] - \theta^*)^2] + \eta[k]^2 b + \eta[k] \mathbb{E} [(\theta[k] - \theta^*)(R[k] - \theta[k]T[k])] \quad (31)$$

To complete the proof it suffices to provide the following bound on the final term of (31):

$$\eta[k] \mathbb{E} [(\theta[k] - \theta^*)(R[k] - \theta[k]T[k])] \leq -\eta[k] T_{min} \mathbb{E} [(\theta[k] - \theta^*)^2]$$

To do this, it suffices to show the following conditional expectation holds for almost all realizations of  $\theta[k]$ , that is with probability 1:

$$(\theta[k] - \theta^*)\mathbb{E}[R[k] - \theta[k]T[k]|\theta[k]] \leq -T_{min}(\theta[k] - \theta^*)^2 \quad (32)$$

To show (32) we consider two cases.

- Case 1: Suppose  $\theta[k] - \theta^* < 0$ . By (22) we have for almost all  $\theta[k]$  for which  $\theta[k] - \theta^* < 0$ :

$$\mathbb{E}[R[k] - \theta[k]T[k]|\theta[k]] \geq t^*(\theta^* - \theta[k])$$

Multiplying both sides by the (negative) value  $\theta[k] - \theta^*$  flips the inequality to yield

$$\begin{aligned} (\theta[k] - \theta^*)\mathbb{E}[R[k] - \theta[k]T[k]|\theta[k]] &\leq -t^*(\theta^* - \theta[k])^2 \\ &\leq -T_{min}(\theta^* - \theta[k])^2 \end{aligned}$$

where the final inequality holds because  $t^* \geq T_{min}$ . Thus, (32) holds in this Case 1.

- Case 2: Suppose  $\theta[k] - \theta^* \geq 0$ . We have

$$(\theta[k] - \theta^*)(R[k] - \theta[k]T[k]) = (\theta[k] - \theta^*)(R[k] - \theta^*T[k]) - T[k](\theta[k] - \theta^*)^2$$

Taking conditional expectations given  $\theta[k]$  gives, for almost all  $\theta[k]$  that satisfy  $\theta[k] - \theta^* \geq 0$ :

$$\begin{aligned} (\theta[k] - \theta^*)\mathbb{E}[R[k] - \theta[k]T[k]|\theta[k]] &= (\theta[k] - \theta^*)\mathbb{E}[R[k] - \theta^*T[k]|\theta[k]] - (\theta[k] - \theta^*)^2\mathbb{E}[T[k]|\theta[k]] \\ &\leq (\theta[k] - \theta^*)\mathbb{E}[R[k] - \theta^*T[k]|\theta[k]] - T_{min}(\theta[k] - \theta^*)^2 \end{aligned} \quad (33)$$

where the final inequality holds because  $\mathbb{E}[T[k]|\theta[k]] \geq T_{min}$  (recall Lemma 2). However, by Lemma 2 we know the conditional expectations  $(\mathbb{E}[T[k]|\theta[k]], \mathbb{E}[R[k]|\theta[k]])$  are in the set  $\bar{\mathcal{A}}$  (with probability 1) and so (with probability 1):

$$\begin{aligned} \mathbb{E}[R[k]|\theta[k]] - \theta^*\mathbb{E}[T[k]|\theta[k]] &\leq \sup_{(t,r) \in \bar{\mathcal{A}}} [r - \theta^*t] \\ &= 0 \end{aligned}$$

where the final equality holds by Lemma 1. Multiplying the above inequality by the nonnegative value  $(\theta[k] - \theta^*)$  does not flip the inequality and we obtain

$$(\theta[k] - \theta^*)\mathbb{E}[R[k] - \theta[k]T[k]|\theta[k]] \leq 0$$

Substituting this into (33) shows that (32) holds in this Case 2. □

#### D. Decreasing stepsize

The previous lemma can be used with a constant stepsize  $\eta[k] = \epsilon$  to prove a bound on the mean squared error between  $\theta[k]$  and  $\theta^*$ . However, the following lemma uses a decreasing stepsize to achieve a faster convergence. The proof uses an induction argument inspired by the analysis in [27] for a different type of problem (deterministic Frank-Wolfe methods for convex optimization). Due to the renewal optimization structure of the current problem, the stepsize used here must be sized carefully with respect to the  $T_{min}$  parameter. This required care in choosing the stepsize is analogous to the discussion on stepsize for a different class of systems in [14]: Work in [14] shows how, for a class of systems with strongly convex properties, a fast  $O(1/k)$  convergence rate can be degraded into a slow  $O(1/k^{1/5})$  rate if the stepsize parameter is not carefully sized according to a strong convexity parameter (which may be difficult to know in practice). That example is used to motivate alternative robust approaches for the systems studied there. The lemma below applies to a different kind of system and does not require any type of strong convexity/concavity. It uses a value  $T_{min}$  for the stepsize selection, but  $T_{min}$  is easy to know in practice. For example, if all frame sizes are at least 1 unit of time, we can use  $T_{min} = 1$ .

*Lemma 6:* Under the algorithm (18)-(20) with any  $\theta[0] \in [\theta_{min}, \theta_{max}]$  and with stepsizes

$$\eta[k] = \frac{1}{(k+2)T_{min}} \quad \forall k \in \{0, 1, 2, \dots\}$$

we have

$$\mathbb{E}[(\theta[k] - \theta^*)^2] \leq \frac{2b}{kT_{min}^2} \quad \forall k \in \{1, 2, 3, \dots\} \quad (34)$$

where the constant  $b$  satisfies (29); the constant  $T_{min}$  satisfies (8)-(9); the constant  $\theta^*$  is defined by (15).

*Proof:* For simplicity define

$$z_k = \frac{1}{2}\mathbb{E}[(\theta[k] - \theta^*)^2] \quad \forall k \in \{0, 1, 2, \dots\}$$

It suffices to show  $z_k \leq \frac{b}{kT_{min}^2}$  for all  $k \in \{1, 2, 3, \dots\}$ . From (30) we have

$$z_{k+1} \leq (1 - 2T_{min}\eta[k])z_k + \eta[k]^2b \quad \forall k \in \{0, 1, 2, \dots\} \quad (35)$$

Applying the above inequality at  $k = 0$  and using  $\eta[0] = 1/(2T_{min})$  gives

$$z_1 \leq \frac{b}{4T_{min}^2}$$

We now use induction with the base case  $k = 1$ . Suppose that  $z_k \leq b/(kT_{min}^2)$  for some  $k \in \{1, 2, 3, \dots\}$  (it holds for  $k = 1$  by the above inequality, since  $b/(4T_{min}^2) \leq b/T_{min}^2$ ). We show the same holds for  $k + 1$ . We have from (35):

$$\begin{aligned} z_{k+1} &\leq (1 - 2T_{min}\eta[k])z_k + \eta[k]^2b \\ &\stackrel{(a)}{=} \left(\frac{k}{k+2}\right)z_k + \frac{b}{(k+2)^2T_{min}^2} \\ &\stackrel{(b)}{\leq} \left(\frac{k}{k+2}\right)\frac{b}{kT_{min}^2} + \frac{b}{(k+2)^2T_{min}^2} \\ &= \frac{b(k+3)}{(k+2)^2T_{min}^2} \\ &\stackrel{(c)}{\leq} \frac{b}{(k+1)T_{min}^2} \end{aligned}$$

where (a) holds because  $\eta[k] = \frac{1}{(k+2)T_{min}}$ ; (b) holds by the induction assumption  $z_k \leq \frac{b}{kT_{min}^2}$ ; (c) holds because  $(k+3)/(k+2)^2 \leq 1/(k+1)$  for all  $k \geq 0$ .  $\square$

### E. Online performance theorem

**Theorem 1: (General performance)** Under the algorithm (18)-(20) with  $\theta[0] \in [\theta_{min}, \theta_{max}]$  and stepsizes  $\eta[k] = \frac{1}{(k+2)T_{min}}$  for  $k \in \{0, 1, 2, \dots\}$  we have

$$\left| \theta^* - \frac{\sum_{k=0}^{K-1} \mathbb{E}[R[k]]}{\sum_{k=0}^{K-1} \mathbb{E}[T[k]]} \right| \leq \frac{\sqrt{2C_1}}{KT_{min}} \left[ |\theta[0] - \theta^*| + \frac{-\sqrt{2b} + \sqrt{8b(K-1)}}{T_{min}} \right] \quad \forall K \in \{2, 3, 4, \dots\}$$

where  $C_1$  satisfies (11) and  $b$  satisfies (29). Hence, deviation from the optimal ratio  $\theta^*$  decays like  $O(1/\sqrt{K})$ .

*Proof:* From (26) we have for all integers  $K \geq 2$ :

$$\begin{aligned} \left| \theta^* - \frac{\sum_{k=0}^{K-1} \mathbb{E}[R[k]]}{\sum_{k=0}^{K-1} \mathbb{E}[T[k]]} \right| &\leq \frac{1}{KT_{min}} \sum_{k=0}^{K-1} \mathbb{E}[|\theta[k] - \theta^*| \cdot |T[k] - t^*|] \\ &\stackrel{(a)}{\leq} \frac{1}{KT_{min}} \sum_{k=0}^{K-1} \sqrt{\mathbb{E}[(\theta[k] - \theta^*)^2] \mathbb{E}[(T[k] - t^*)^2]} \\ &\stackrel{(b)}{\leq} \frac{\sqrt{2C_1}}{KT_{min}} \left[ \sqrt{\mathbb{E}[(\theta[0] - \theta^*)^2]} + \sum_{k=1}^{K-1} \sqrt{\mathbb{E}[(\theta[k] - \theta^*)^2]} \right] \\ &\stackrel{(c)}{\leq} \frac{\sqrt{2C_1}}{KT_{min}} \left[ \sqrt{\mathbb{E}[(\theta[0] - \theta^*)^2]} + \sum_{k=1}^{K-1} \sqrt{\frac{2b}{kT_{min}^2}} \right] \\ &\stackrel{(d)}{\leq} \frac{\sqrt{2C_1}}{KT_{min}} \left[ \sqrt{\mathbb{E}[(\theta[0] - \theta^*)^2]} + \frac{-\sqrt{2b} + \sqrt{8b(K-1)}}{T_{min}} \right] \end{aligned}$$

where (a) follows by the Cauchy-Schwarz inequality; (b) holds by (11) and  $(T[k] - t^*)^2 \leq T[k]^2 + (t^*)^2 \leq T[k]^2 + C_1$ ; (c) holds by (34); (d) holds because  $\sum_{k=1}^{K-1} \frac{1}{\sqrt{k}} \leq 1 + \int_1^{K-1} \frac{1}{\sqrt{t}} dt$ .  $\square$

### F. Strongly concave curvature

This section proves that the algorithm achieves a faster convergence rate in the special case when the set  $\bar{\mathcal{A}}$  has a strongly concave property. Specifically, suppose the set  $\bar{\mathcal{A}}$  has a strongly concave upper boundary about the optimality point  $(t^*, r^*)$ , so that for some  $c > 0$  we have (see Fig. 2):

$$r \leq r^* + \theta^*(t - t^*) - \frac{c}{2}(t - t^*)^2 \quad \forall (t, r) \in \bar{\mathcal{A}} \quad (36)$$

*Theorem 2:* (Performance with strongly concave curvature) Assume  $\bar{\mathcal{A}}$  has the strongly concave curvature property specified in (36). Under the algorithm (18)-(20) with  $\theta[0] \in [\theta_{min}, \theta_{max}]$  and stepsize  $\eta[k] = \frac{1}{(k+2)T_{min}}$  for  $k \in \{0, 1, 2, \dots\}$  we have

$$\left| \theta^* - \frac{\sum_{k=0}^{K-1} \mathbb{E}[R[k]]}{\sum_{k=0}^{K-1} \mathbb{E}[T[k]]} \right| \leq \frac{2(\theta[0] - \theta^*)^2 + \frac{4b}{T_{min}^2}(1 + \log(K-1))}{KcT_{min}} \quad \forall K \in \{2, 3, 4, \dots\} \quad (37)$$

and so deviation from the optimal ratio  $\theta^*$  decays like  $O(\log(K)/K)$ .

*Proof:* For almost all  $\theta[k]$  we have (by Lemma 2):

$$(\mathbb{E}[T[k]|\theta[k]], \mathbb{E}[R[k]|\theta[k]]) \in \bar{\mathcal{A}}$$

And so by (36) we have with probability 1:

$$\mathbb{E}[R[k]|\theta[k]] \leq r^* + \theta^*(\mathbb{E}[T[k]|\theta[k]] - t^*) - \frac{c}{2}(\mathbb{E}[T[k]|\theta[k]] - t^*)^2 \quad (38)$$

From (23) we have for all real numbers  $\beta > 0$  (with prob 1):

$$\begin{aligned} \mathbb{E}[R[k] - \theta^*T[k]|\theta[k]] &\geq (\theta[k] - \theta^*)(\mathbb{E}[T[k]|\theta[k]] - t^*) \\ &= \left(\frac{1}{\beta}(\theta[k] - \theta^*)\right) \beta(\mathbb{E}[T[k]|\theta[k]] - t^*) \\ &\stackrel{(a)}{\geq} -\frac{(\theta[k] - \theta^*)^2}{2\beta^2} - \frac{\beta^2(\mathbb{E}[T[k]|\theta[k]] - t^*)^2}{2} \\ &\stackrel{(b)}{\geq} -\frac{(\theta[k] - \theta^*)^2}{2\beta^2} - \frac{\beta^2}{c}[r^* + \theta^*(\mathbb{E}[T[k]|\theta[k]] - t^*) - \mathbb{E}[R[k]|\theta[k]]] \\ &\stackrel{(c)}{=} -\frac{(\theta[k] - \theta^*)^2}{2\beta^2} + \frac{\beta^2}{c}\mathbb{E}[R[k] - \theta^*T[k]|\theta[k]] \end{aligned}$$

where (a) uses the fact  $ab \geq -\frac{a^2+b^2}{2}$  for all real numbers  $a, b$ ; (b) uses (38); (c) uses  $r^* - \theta^*t^* = 0$ . Choose  $\beta > 0$  so that  $\beta^2/c = 1/2$ . Then

$$\mathbb{E}[R[k] - \theta^*T[k]|\theta[k]] \geq -\frac{(\theta[k] - \theta^*)^2}{\beta^2} = -2\frac{(\theta[k] - \theta^*)^2}{c}$$

where the final equality uses  $\beta^2 = c/2$ . Taking expectations of the above and using the law of iterated expectations gives

$$\mathbb{E}[R[k] - \theta^*T[k]] \geq -2\frac{\mathbb{E}[(\theta[k] - \theta^*)^2]}{c}$$

Fix integer  $K \geq 2$ . Summing the above inequality over  $k \in \{0, \dots, K-1\}$  gives

$$\begin{aligned} \sum_{k=0}^{K-1} \mathbb{E}[R[k]] - \theta^* \sum_{k=0}^{K-1} \mathbb{E}[T[k]] &\geq -2\frac{\mathbb{E}[(\theta[0] - \theta^*)^2]}{c} - 2\sum_{k=1}^{K-1} \frac{\mathbb{E}[(\theta[k] - \theta^*)^2]}{c} \\ &\stackrel{(a)}{\geq} -2\frac{\mathbb{E}[(\theta[0] - \theta^*)^2]}{c} - 2\sum_{k=1}^{K-1} \frac{2b}{kcT_{min}^2} \\ &\stackrel{(b)}{\geq} -2\frac{\mathbb{E}[(\theta[0] - \theta^*)^2]}{c} - \frac{4b}{cT_{min}^2}(1 + \log(K-1)) \end{aligned}$$

where (a) holds by (34); (b) holds because  $\sum_{k=1}^{K-1} 1/k \leq 1 + \int_1^{K-1} (1/t)dt$ . Rearranging terms gives

$$\begin{aligned} \frac{\sum_{k=0}^{K-1} \mathbb{E}[R[k]]}{\sum_{k=0}^{K-1} \mathbb{E}[T[k]]} &\geq \theta^* - \frac{2\mathbb{E}[(\theta[0] - \theta^*)^2] + \frac{4b}{T_{min}^2}(1 + \log(K-1))}{c \sum_{k=0}^{K-1} \mathbb{E}[T[k]]} \\ &\geq \theta^* - \frac{2\mathbb{E}[(\theta[0] - \theta^*)^2] + \frac{4b}{T_{min}^2}(1 + \log(K-1))}{KcT_{min}} \end{aligned}$$

where the final inequality holds because  $\mathbb{E}[T[k]] \geq T_{min}$  for all  $k$ . On the other hand (27) implies that the ratio on the left-hand-side is less than or equal to  $\theta^*$ . This proves the result.  $\square$

### G. Discussion

Theorem 1 shows the optimality gap decays like  $O(1/\sqrt{K})$  for general systems. Theorem 2 shows that for systems with a strongly concave property, the optimality gap decays much faster according to  $O(\log(K)/K)$ . This improved convergence speed does not require any changes in the algorithm itself. Indeed, the algorithm does not need to know whether or not the system has the strongly concave property. If the system *does* have the strongly concave property, the algorithm automatically yields faster convergence without having to know the strong concavity parameter  $c$ .

## V. MATCHING CONVERSE FOR STRONGLY CONCAVE STRUCTURE

This section constructs a particular system (with a strongly concave curvature) for which all causal algorithms that do not have a-priori knowledge of the probability distribution  $F_S(s)$  have optimality gaps that decay no faster than  $\Omega(\log(K)/K)$ . This matches the  $O(\log(K)/K)$  achievability result of Theorem 2 and shows that this convergence rate is optimal over the class of systems with strongly concave curvature. For the proof, we construct a nontrivial mapping from the sequential decision problem to a related estimation problem. This allows use of the Bernoulli estimation theorem of [25]. This mapping technique is conceptually similar to the method recently used to prove a converse result for a different class of systems with unit timeslots in [26].

### A. System

Suppose the task type process is an i.i.d. Bernoulli process  $\{S[k]\}_{k=0}^{\infty}$  with

$$P[S[k] = 1] = q \quad ; \quad P[S[k] = 0] = 1 - q \quad (39)$$

where  $q$  is an unknown probability. For technical reasons, we assume throughout that  $q \in [1/4, 3/4]$ . Every frame  $k \in \{0, 1, 2, \dots\}$  the controller observes  $S[k]$  and then chooses  $(T[k], R[k]) \in \mathcal{D}(S[k])$ , where

$$\mathcal{D}(S[k]) = \begin{cases} (1, 1) & , \text{ if } S[k] = 0 \\ \{(x, 2 - (2 - x)^2) \in \mathbb{R}^2 : x \in [1, 2]\} & , \text{ if } S[k] = 1 \end{cases} \quad (40)$$

The decision structure (40) defines a system with *inflexible tasks* (type 0) and *flexible tasks* (type 1). Indeed, if  $S[k] = 0$  then the controller must choose  $(T[k], R[k]) = (1, 1)$ . However, if  $S[k] = 1$  then the controller can choose  $(T[k], R[k])$  as any point on the curve  $(x, 2 - (2 - x)^2)$  for  $x \in [1, 2]$ . A higher reward is obtained for larger values of  $x$ , but with diminishing returns (see Fig. 2). This particular curve is chosen as a representative example with *strongly concave curvature*<sup>7</sup>.

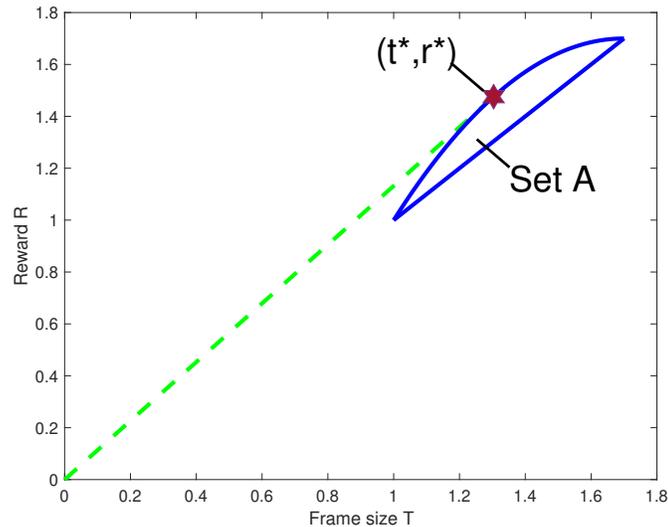


Fig. 2. The set  $\mathcal{A}$  with a strongly concave upper boundary for the example of Section V-A with  $q = 0.7$ .

*Lemma 7:* Fix  $q \in [1/4, 3/4]$ . Let  $\mathcal{A}$  be the set of one-shot expectations for the example problem (39)-(40). Define the sets

$$\begin{aligned} \mathcal{Z} &= \left\{ (1 - q + qx, 1 + q - q(2 - x)^2) \in \mathbb{R}^2 : x \in [1, 2] \right\} \\ \mathcal{W} &= \left\{ (t, r) \in \mathbb{R}^2 : t \in [1, 1 + q], t \leq r \leq (2t - 1) - \frac{1}{q}(t - 1)^2 \right\} \end{aligned}$$

<sup>7</sup>The strongly concave decision curve gives rise to a set  $\bar{\mathcal{A}}$  with a strongly concave upper boundary. The  $\Omega(\log(K)/K)$  converse result in this section can be extended to more general systems for which  $\bar{\mathcal{A}}$  has a strongly concave upper boundary.

Then

$$\mathcal{A} = \text{Conv}(\mathcal{Z}) = \mathcal{W}$$

where  $\text{Conv}(\mathcal{Z})$  denotes the convex hull of the set  $\mathcal{Z}$ . In particular,  $\mathcal{A}$  is closed and so  $\overline{\mathcal{A}} = \mathcal{A}$ .

*Proof:* (Part 1): We show  $\text{Conv}(\mathcal{Z}) \subseteq \mathcal{A}$ . Fix  $z \in \mathcal{Z}$ . Then there is an  $x \in [1, 2]$  such that

$$z = (1 - q + qx, 1 + q - q(2 - x)^2)$$

The policy that chooses  $(T[0], R[0]) = (1, 1)$  when  $S[0] = 0$  and  $(T[0], R[0]) = (x, 2 - (2 - x)^2)$  when  $S[0] = 1$  yields:

$$\begin{aligned} \mathbb{E}[(T[0], R[0])] &= (1 - q) \cdot (1, 1) + q \cdot (x, 2 - (2 - x)^2) \\ &= (1 - q + qx, 1 + q - q(2 - x)^2) \\ &= z \end{aligned}$$

and so the vector  $z$  is achievable as an expectation on slot 0, meaning  $z \in \mathcal{A}$ . This holds for all  $z \in \mathcal{Z}$  and so

$$\mathcal{Z} \subseteq \mathcal{A} \implies \text{Conv}(\mathcal{Z}) \subseteq \text{Conv}(\mathcal{A}) = \mathcal{A}$$

where the final equality holds because the set  $\mathcal{A}$  is convex and so it is equal to its convex hull.

(Part 2): We show  $\mathcal{A} \subseteq \mathcal{W}$ . Fix  $a \in \mathcal{A}$ . By definition of  $\mathcal{A}$  as the region of one-shot expectations, there is a policy that yields decisions  $(T[0], R[0]) \in \mathcal{D}(S[0])$  such that:

$$(\mathbb{E}[T[0]], \mathbb{E}[R[0]]) = a$$

To prove  $a \in \mathcal{W}$  it suffices to show that

$$1 \leq \mathbb{E}[T[0]] \leq 1 + q \tag{41}$$

$$\mathbb{E}[T[0]] \leq \mathbb{E}[R[0]] \leq (2\mathbb{E}[T[0]] - 1) - \frac{1}{q}(\mathbb{E}[T[0]] - 1)^2 \tag{42}$$

Regardless of the value of  $S[0]$  we have from the structure of  $\mathcal{D}(S[0])$  in (40):

$$1 \leq T[0] \leq 1 + 1_{\{S[0]=1\}}$$

where  $1_{\{S[0]=1\}}$  is an indicator function that is 1 if  $S[0] = 1$  and 0 else. Taking expectations of the above inequality yields (41). Again regardless of  $S[0]$ , we have

$$T[0] \leq R[0] = (2T[0] - 1) - (T[0] - 1)^2 \tag{43}$$

Indeed, if  $S[0] = 0$  then  $T[0] = R[0] = 1$  and so (43) holds, while if  $S[0] = 1$  then (43) again holds because by (40) we have

$$\begin{aligned} R[0] &= 2 - (2 - T[0])^2 \\ &\geq T[0] \end{aligned}$$

where the final inequality holds because  $T[0] \in [1, 2]$ . Taking expectations of (43) gives

$$\begin{aligned} \mathbb{E}[T[0]] &\leq \mathbb{E}[R[0]] \\ &= (2\mathbb{E}[T[0]] - 1) - \mathbb{E}[(T[0] - 1)^2] \\ &\stackrel{(a)}{=} (2\mathbb{E}[T[0]] - 1) - q\mathbb{E}[(T[0] - 1)^2 | S[0] = 1] \\ &\stackrel{(b)}{\leq} (2\mathbb{E}[T[0]] - 1) - q(\mathbb{E}[T[0] - 1 | S[0] = 1])^2 \\ &= (2\mathbb{E}[T[0]] - 1) - \frac{1}{q}(q\mathbb{E}[T[0] - 1 | S[0] = 1])^2 \\ &\stackrel{(c)}{=} (2\mathbb{E}[T[0]] - 1) - \frac{1}{q}(\mathbb{E}[T[0] - 1])^2 \end{aligned}$$

where (a) holds because  $T[0] = 1$  whenever  $S[0] = 0$ ; (b) holds by Jensen's inequality applied to the conditional expectation; (c) holds because  $\mathbb{E}[T[0] - 1] = q\mathbb{E}[T[0] - 1 | S[0] = 1]$  (since  $T[0] - 1 = 0$  if  $S[0] = 0$ ). This establishes (42) and thus proves  $\mathcal{A} \subseteq \mathcal{W}$ .

(Part 3): We show  $\mathcal{W} \subseteq \text{Conv}(\mathcal{Z})$ . This can be shown by noting  $\mathcal{W}$  is the convex hull of the following set of points

$$\left\{ \left( t, (2t - 1) - \frac{1}{q}(t - 1)^2 \right) : t \in [1, 1 + q] \right\}$$

and these points coincide with the set  $\mathcal{Z}$ . □

*Lemma 8:* The point  $(t^*, r^*) \in \mathcal{A}$  that maximizes  $r/t$  over all points  $(t, r) \in \mathcal{A}$  is achieved on the upper boundary of  $\mathcal{A}$  (see Fig. 2) and satisfies:

$$t^* = \sqrt{1+q} \quad (44)$$

$$r^* = \frac{2(q+1)(-1+\sqrt{1+q})}{q} \quad (45)$$

$$\theta^* = \frac{r^*}{t^*} = 2 - \frac{2}{q}(-1 + \sqrt{1+q}) \quad (46)$$

Further, by strong concavity of the upper boundary of set  $\mathcal{A}$  we have

$$r \leq \theta^* t - \frac{1}{q}(t - t^*)^2 \quad \forall (t, r) \in \mathcal{A} \quad (47)$$

*Proof:* By the previous lemma, the upper boundary of  $\mathcal{A}$  consists of points  $(t, r)$  such that

$$r = (2t - 1) - \frac{1}{q}(t - 1)^2 \quad , t \in [1, 1+q] \quad (48)$$

It is not difficult to show the maximizer of  $r/t$  over all  $(t, r) \in \mathcal{A}$  is on this upper boundary and is given by the point  $(t^*, r^*)$  specified in (44)-(46) (details omitted for brevity). The final inequality (47) hold because the upper boundary curve (48) can be reparameterized by

$$r = \theta^* t - \frac{1}{q}(t - t^*)^2, t \in [1, 1+q]$$

which is a quadratic that matches the squared term and constant term in the quadratic (48) and also matches at the point  $(t^*, r^*)$ .  $\square$

### B. Causal and measurable algorithms

Our converse result considers algorithms that choose  $(T[k], R[k]) \in \mathcal{D}(S[k])$  every frame  $k$  in a way that is *causal* (so the algorithm has no knowledge of the future) and *probabilistically measurable* (so probability distributions and expectations are well defined). For each  $k \in \{0, 1, 2, \dots\}$ , define  $H[k]$  as the system history up to but not including frame  $k$ :

$$H[k] = (S[0], S[1], \dots, S[k-1]) \quad \forall k \in \{1, 2, 3, \dots\}$$

where  $H[0]$  is formally defined as NULL (since there is no history before frame 0). On frame  $k$ , a general algorithm has no choice if it observes  $S[k] = 0$  (since it must use  $(T[k], R[k]) = (1, 1)$  in that case) but must select  $(T[k], R[k]) \in \mathcal{D}(1)$  if it observes  $S[k] = 1$ . A general causal and measurable algorithm must make this (potentially randomized) decision based only on the history  $H[k]$  observed on frames  $\{0, 1, \dots, k-1\}$ .

It suffices to assume that on frame  $k$ , the algorithm makes decisions as a deterministic function of  $(H[k], U)$  where  $U$  is uniformly distributed in  $[0, 1)$  and is independent of  $\{S[k]\}_{k=0}^\infty$ . Let  $\{f_k\}_{k=0}^\infty$  be a sequence of functions of the following form:

$$f_k : \{0, 1\}^k \times [0, 1] \rightarrow \mathcal{D}(1) \quad \forall k \in \{1, 2, 3, \dots\}$$

where  $\{0, 1\}^0$  is formally defined as NULL so that  $f_0(u)$  depends only on  $u \in [0, 1]$ . These functions  $\{f_k\}_{k=0}^\infty$  define control decisions of a general causal and measurable algorithm on each frame  $k \in \{0, 1, 2, \dots\}$ :

$$(T[k], R[k]) = \begin{cases} (1, 1) & , \text{ if } S[k] = 0 \\ f_k(H[k], U) & , \text{ if } S[k] = 1 \end{cases} \quad (49)$$

The functions  $\{f_k\}_{k=0}^\infty$  are assumed to be *probabilistically measurable* on the sigma algebra  $\sigma(U; \{S[k]\}_{k=0}^\infty)$ . Since  $\mathcal{D}(1)$  is a bounded set, the corresponding expectations  $(\mathbb{E}[T[k]], \mathbb{E}[R[k]])$  are well defined and finite, as are the conditional expectations given  $(H[k], U)$ . We say that decisions  $(T[k], R[k]) \in \mathcal{D}(S[k])$  for all  $k \in \{0, 1, 2, \dots\}$  are *causal and measurable* if they come from a sequence of deterministic functions  $\{f_k\}_{k=0}^\infty$  with the above structure.

This algorithm structure is not restrictive: From the single random variable  $U$ , we can construct an infinite sequence of i.i.d. uniformly distributed random variables  $\{U_i\}_{i=1}^\infty$ , where each  $U_i$  is a deterministic function of  $U$ .<sup>8</sup> This allows the controller to make as many calls to a random number generator as needed (assuming an at-most countably infinite number of such calls).

*Lemma 9:* Fix  $q \in [1/4, 3/4]$  and consider any causal and measurable decisions  $(T[k], R[k]) \in \mathcal{D}(S[k])$  for  $k \in \{0, 1, 2, \dots\}$ . For each  $k \in \{0, 1, 2, \dots\}$  we have:

$$\mathbb{E}[(T[k], R[k]) | H[k]] \in \mathcal{A} \quad (50)$$

Furthermore,

$$\mathbb{E}[R[k] | H[k]] \leq \theta^* \mathbb{E}[T[k] | H[k]] - \frac{1}{q}(\mathbb{E}[T[k] | H[k]] - t^*)^2 \quad (51)$$

<sup>8</sup>This can be done by writing  $U = \sum_{i=1}^\infty X_i 10^{-i}$  in the unique base-10 expansion where  $X_i \in \{0, \dots, 9\}$  is the  $i$ th digit of the expansion and  $\{X_i\}_{i=1}^\infty$  does not have an infinite tail of 9s, and defining  $U_n$  for each  $n \in \{1, 2, 3, \dots\}$  by  $U_n = \sum_{i=1}^\infty X_{g(n,i)} 10^{-i}$  where  $g : \mathbb{N}^2 \rightarrow \mathbb{N}$  is any bijection.

where  $t^* = \sqrt{1+q}$  and  $\theta^*$  satisfies (46).

*Proof:* The fact (50) is similar to Lemma 2 and is formally proven by noting that  $\bar{\mathcal{A}} = \mathcal{A}$  and  $S[k]$  is independent of  $(U, H[k])$ , and then applying the general theorem in the Appendix (details omitted for brevity). Substituting (50) into (47) proves (51).  $\square$

### C. The Bernoulli estimation theorem from [25]

Let  $\{g_k\}_{k=1}^{\infty}$  be an infinite sequence of deterministic functions of the form:

$$g_k : \{0, 1\}^k \rightarrow [0, 1] \quad \forall k \in \{1, 2, 3, \dots\}$$

and for each  $k \in \{1, 2, 3, \dots\}$  the function  $g_k(s_0, \dots, s_{k-1})$  maps a binary-valued sequence  $(s_0, \dots, s_{k-1})$  to a real number in the interval  $[0, 1]$ . Let  $\{S[k]\}_{k=0}^{\infty}$  be an i.i.d. sequence of Bernoulli random variables with parameter  $q = P[S[k] = 1]$ . Assume that  $q$  is an unknown parameter in the interval  $[1/4, 3/4]$ . The functions  $g_k$  shall be called *estimation functions* because they can map the first  $k$  observations of the Bernoulli random variables to a (deterministic) estimate  $G[k]$  of the  $q$  parameter via:

$$G[k] = g_k(S[0], S[1], \dots, S[k-1]) \quad \forall k \in \{1, 2, 3, \dots\} \quad (52)$$

The following theorem from [25] provides a lower bound on the mean square error for any sequence of estimation functions<sup>9</sup>

*Theorem 3:* (Bernoulli Estimation [25]) For any sequence of estimation functions  $\{g_k\}_{k=0}^{\infty}$  as defined above, there exists a probability  $q \in [1/4, 3/4]$  such that if  $\{S[k]\}_{k=0}^{\infty}$  is an i.i.d. Bernoulli sequence with parameter  $q$ , then for all positive integers  $K$  we have

$$\sum_{k=1}^K \mathbb{E} [(q - G[k])^2] \geq \Omega(\log(K)) \quad (53)$$

where the random variables  $G[k]$  are defined in (52) for each  $k \in \{1, 2, \dots\}$ .

### D. Completing the converse

*Lemma 10:* Define  $\phi : [1/4, 3/4] \rightarrow \mathbb{R}$  by

$$\phi(q) = \frac{-1 + \sqrt{1+q}}{q}$$

Then

a)  $\phi$  is continuous and strictly decreasing with minimum and maximum values

$$\phi_{min} = \phi(3/4) \approx 0.430501$$

$$\phi_{max} = \phi(1/4) \approx 0.472136$$

b) There is a continuous inverse function  $\phi^{-1} : [\phi_{min}, \phi_{max}] \rightarrow [1/4, 3/4]$  that satisfies

$$\phi(\phi^{-1}(y)) = y \quad \forall y \in [\phi_{min}, \phi_{max}]$$

c) The derivative of  $\phi$  exists for all  $q \in (1/4, 3/4)$  and

$$|\phi'(q)| \geq \beta \quad \forall q \in (1/4, 3/4)$$

where  $\beta = |\phi'(3/4)| \approx 0.0700485$ .

d) We have

$$|\phi(q) - \phi(u)| \geq \beta|q - u| \quad \forall u, q \in [1/4, 3/4] \quad (54)$$

*Proof:* Parts (a) and (c) follow by basic analysis on the function  $\phi$ . Part (b) follows immediately from (a). To prove (d), without loss of generality assume  $1/4 \leq u < q \leq 3/4$ . By the mean value theorem, there is a point  $x \in (u, q)$  such that

$$\frac{\phi(q) - \phi(u)}{q - u} = \phi'(x)$$

and so

$$\left| \frac{\phi(q) - \phi(u)}{q - u} \right| = |\phi'(x)| \geq \beta$$

$\square$

*Theorem 4:* Let  $\{f_k\}_{k=0}^{\infty}$  be any sequence of decision functions that define a causal and measurable algorithm according to (49). There is a parameter  $q \in [1/4, 3/4]$  such that using these decision functions with an i.i.d. Bernoulli- $q$  process  $\{S[k]\}_{k=0}^{\infty}$

<sup>9</sup>The result in [25] also applies to randomized estimation functions. This paper shall only need the result for deterministic estimation functions.

and with an independent uniform random variable  $U$ , the resulting (causal and measurable) decisions  $(T[k], R[k]) \in \mathcal{D}(S[k])$  satisfy

$$\frac{\sum_{k=0}^{K-1} \mathbb{E}[R[k]]}{\sum_{k=0}^{K-1} \mathbb{E}[T[k]]} \leq \theta^* - \Omega\left(\frac{\log(K)}{K}\right)$$

where  $\theta^*$  is the optimal ratio in (46). In particular, no algorithm can have error that decays faster than  $\Omega(\log(K)/K)$ .

*Proof:* Define for each  $k \in \{0, 1, 2, \dots\}$  and each  $h_k \in \{0, 1\}^k$ :

$$z_k(h_k) = \mathbb{E}[T[k] | S[k] = 1, H[k] = h_k]$$

Define  $g_k : \{0, 1\}^k \rightarrow [0, 1]$  by

$$g_k(h_k) = \phi^{-1}\left([z_k(h_k) - 1]_{\phi_{min}}^{\phi_{max}}\right)$$

Since  $\phi^{-1} : [\phi_{min}, \phi_{max}] \rightarrow [1/4, 3/4]$ , these  $\{g_k\}$  functions can be viewed as estimation functions. Define

$$G[k] = g_k(H[k]) \quad \forall k \in \{1, 2, 3, \dots\} \quad (55)$$

and so

$$G[k] = \phi^{-1}\left([z_k(H[k]) - 1]_{\phi_{min}}^{\phi_{max}}\right) \quad (56)$$

Observe that for  $k \in \{1, 2, 3, \dots\}$ , the  $g_k$  functions can be viewed as estimation functions for the parameter  $q$ , with  $G[k]$  the corresponding estimator based on the past  $k$  observations, because they have the general structure specified by the Bernoulli estimation theorem (Theorem 3). Indeed  $G[k]$  maps  $H[k] = (S[0], \dots, S[k-1])$  to the unit interval  $[0, 1]$ , and this map is measurable because it is a composition with the measurable  $z_k(h)$  function, the continuous projection to the interval  $[\phi_{min}, \phi_{max}]$ , and the continuous inverse function  $\phi^{-1}$ . Hence, if we can expression the ratio of expectations in question as a sum of the mean squared error between  $G[k]$  and  $q$ , we can apply Theorem 3.

Fix  $k \in \{0, 1, 2, \dots\}$ . We have for each  $h_k \in \{0, 1\}^k$ :

$$\begin{aligned} \mathbb{E}[R[k] | H[k]] &\stackrel{(a)}{\leq} \theta^* \mathbb{E}[T[k] | H[k]] - \frac{1}{q} (\mathbb{E}[T[k] | H[k]] - t^*)^2 \\ &\stackrel{(b)}{=} \theta^* \mathbb{E}[T[k] | H[k]] - \frac{1}{q} ((1-q) + qz_k(H[k]) - t^*)^2 \\ &= \theta^* \mathbb{E}[T[k] | H[k]] - q \left( z_k(H[k]) - 1 - \frac{-1+t^*}{q} \right)^2 \\ &\stackrel{(c)}{=} \theta^* \mathbb{E}[T[k] | H[k]] - q (z_k(H[k]) - 1 - \phi(q))^2 \\ &\stackrel{(d)}{\leq} \theta^* \mathbb{E}[T[k] | H[k]] - q \left( [z_k(H[k]) - 1]_{\phi_{min}}^{\phi_{max}} - \phi(q) \right)^2 \\ &= \theta^* \mathbb{E}[T[k] | H[k]] - q \left( \phi(\phi^{-1}([z_k(H[k]) - 1]_{\phi_{min}}^{\phi_{max}})) - \phi(q) \right)^2 \\ &\stackrel{(e)}{=} \theta^* \mathbb{E}[T[k] | H[k]] - q (\phi(G[k]) - \phi(q))^2 \\ &\stackrel{(f)}{\leq} \theta^* \mathbb{E}[T[k] | H[k]] - q\beta^2 (G[k] - q)^2 \\ &\stackrel{(g)}{\leq} \theta^* \mathbb{E}[T[k] | H[k]] - (1/4)\beta^2 (G[k] - q)^2 \end{aligned}$$

where (a) holds by (51); (b) holds by (49); (c) holds by definition of  $\phi$  and because  $t^* = \sqrt{1+q}$  (recall (44)); (d) holds because the distance between  $(z[k] - 1)$  and  $\phi(q)$  is greater than or equal to the distance between their projections onto the interval  $[\phi_{min}, \phi_{max}]$  (and the fact that  $\phi(q) \in [\phi_{min}, \phi_{max}]$ ); (e) holds by (56); (f) holds by (54); (g) holds because  $q \geq 1/4$ .

Taking expectations of both sides with respect to the random  $H[k]$  gives

$$\mathbb{E}[R[k]] \leq \theta^* \mathbb{E}[T[k]] - \frac{\beta^2}{4} \mathbb{E}[(G[k] - q)^2]$$

Summing gives

$$\begin{aligned} \sum_{k=0}^{K-1} \mathbb{E}[R[k]] &\leq \theta^* \sum_{k=0}^{K-1} \mathbb{E}[T[k]] - \frac{\beta^2}{4} \sum_{k=0}^{K-1} \mathbb{E}[(G[k] - q)^2] \\ &\leq \theta^* \sum_{k=0}^{K-1} \mathbb{E}[T[k]] - \frac{\beta^2}{4} \sum_{k=1}^{K-1} \mathbb{E}[(G[k] - q)^2] \end{aligned}$$

Dividing by the expression  $\sum_{k=0}^{K-1} \mathbb{E}[T[k]]$  and noting that this expression is less than or equal to  $2K$  (since  $T[k] \leq 2$  always, see (40)) yields

$$\begin{aligned} \frac{\sum_{k=0}^{K-1} \mathbb{E}[R[k]]}{\sum_{k=0}^{K-1} \mathbb{E}[T[k]]} &\leq \theta^* - \frac{\beta^2}{8K} \sum_{k=1}^{K-1} \mathbb{E}[(G[k] - q)^2] \\ &\leq \theta^* - \Omega(\log(K)/K) \end{aligned}$$

where the final inequality holds by application of the Bernoulli estimation bound (53).  $\square$

## VI. MATCHING SQUARE ROOT CONVERSE

This section presents a square-root converse result for systems without the strongly concave structure. We consider the example system of Section III with task type process  $\{S[k]\}_{k=0}^{\infty}$  that is i.i.d. Bernoulli with  $P[S[k] = 1] = p$ , where  $p$  is an unknown parameter. The nature of this converse is different from the one in the previous section and we shall consider only two possible values of  $p$ : Fix  $\epsilon$  such that  $0 < \epsilon \leq 1/4$ . Consider the two possible hypotheses:

- Hypothesis  $H_{(1/2-\epsilon)}$ :  $p = 1/2 - \epsilon$ . Under this hypothesis it can be shown that:

$$\theta_{(1/2-\epsilon)}^* = 2 + 2\epsilon \quad (57)$$

- Hypothesis  $H_{(1/2+\epsilon)}$ :  $p = 1/2 + \epsilon$ . Under this hypothesis it can be shown that:

$$\theta_{(1/2+\epsilon)}^* = \frac{6}{3 + 2\epsilon} \quad (58)$$

The structure of considering only two possible hypotheses that are difficult to discern is similar in spirit to the converse result of [21] for multi-armed bandit systems, where a square root law is also developed. However, the square root arises for a different reason here. Indeed, the system treated in this paper has a different structure and requires a different proof.

Fix  $U$  uniform over  $[0, 1)$  and assume  $U$  is independent of  $\{S[k]\}_{k=0}^{\infty}$ . Consider a general causal algorithm that has no knowledge of  $p$  and that makes decisions as follows: On each frame  $k$  it chooses a conditional probability  $\beta[k]$ , being the conditional probability of choosing high quality given that  $S[k] = 1$ , as some deterministic and measurable function  $\hat{\beta}_k(\cdot)$  of  $U$  and  $S[0], \dots, S[k-1]$ :

$$\beta[0] = \hat{\beta}_0(U) \quad (59)$$

$$\beta[k] = \hat{\beta}_k(U, S[0], \dots, S[k-1]) \quad \forall k \in \{1, 2, 3, \dots\} \quad (60)$$

For each frame  $k$ , given  $\beta[k]$ , we have

$$(T[k], R[k]) = \begin{cases} (1, 3) & \text{with prob } 1 - p \\ (2, 3) & \text{with prob } p\beta[k] \\ (1, 1) & \text{with prob } p(1 - \beta[k]) \end{cases}$$

and so

$$\begin{aligned} \mathbb{E}[R[k]|\beta[k]] &= (1-p)(3) + p(1 + 2\beta[k]) \\ \mathbb{E}[T[k]|\beta[k]] &= (1-p)(1) + p(1 + \beta[k]) \\ \mathbb{E}[R[k] - \theta^*T[k]|\beta[k]] &= (1-p)(3) + p(1 + 2\beta[k]) - \theta^*(1 - p + p(1 + \beta[k])) \end{aligned} \quad (61)$$

*Theorem 5:* Fix  $\delta$  such that  $0 < \delta \leq 1/256$ . Consider any general causal algorithm of the type (59)-(60). Fix  $K \in \{1, \dots, \lfloor \frac{3}{2^{19}\delta^2} \rfloor\}$ . If for the case  $p = \frac{1}{2} - 64\delta$  the algorithm satisfies:

$$\frac{\sum_{k=0}^K \mathbb{E}[R[k]]}{\sum_{k=0}^K \mathbb{E}[T[k]]} > \theta^* - \delta$$

then the algorithm fails to achieve this for the case  $p = \frac{1}{2} + 64\delta$ .

The proof is developed in the next subsections by defining  $\epsilon = 64\delta$  and by fixing a particular causal algorithm of the type described in this section.

### A. Case $H_{(1/2-\epsilon)}$

Fix  $\epsilon = 64\delta$ . Fix a particular causal algorithm of the type (59)-(60). Substituting (57) into (61) and doing the basic but tedious arithmetic gives

$$\begin{aligned} \mathbb{E}_{(1/2-\epsilon)} \left[ R[k] - \theta_{(1/2-\epsilon)}^* T[k] | \beta[k] \right] &= -\beta\epsilon + 2\beta\epsilon^2 \\ &\stackrel{(a)}{\leq} -\frac{\beta[k]}{2}\epsilon \\ &\stackrel{(b)}{\leq} -\frac{\epsilon}{4} 1_{\{\beta[k] > 1/2\}} \end{aligned}$$

where where  $\mathbb{E}_{(1/2-\epsilon)} [\cdot]$  denotes an expectation under the assumption that the true Bernoulli parameter is  $p = 1/2 - \epsilon$ ; (a) holds because  $0 < \epsilon \leq 1/4$ ; (b) uses the indicator function  $1_{\{\beta[k] > 1/2\}}$  that is 1 if  $\beta[k] > 1/2$  and 0 else. Taking expectations and using the law of iterated expectations gives

$$\mathbb{E}_{(1/2-\epsilon)} \left[ R[k] - \theta_{(1/2-\epsilon)}^* T[k] \right] \leq -\frac{\epsilon}{4} P[\beta[k] > 1/2] \quad (62)$$

### B. Case $H_{(1/2+\epsilon)}$

Fix  $\epsilon = 64\delta$ . Substituting (58) into (61)

$$\begin{aligned} \mathbb{E}_{(1/2+\epsilon)} \left[ R[k] - \theta_{(1/2+\epsilon)}^* T[k] | \beta[k] \right] &= -(1 - \beta[k]) \frac{2\epsilon + 4\epsilon^2}{3 + 2\epsilon} \\ &\stackrel{(a)}{\leq} -(1 - \beta[k]) \frac{\epsilon}{2} \\ &\stackrel{(b)}{\leq} -\frac{\epsilon}{4} 1_{\{\beta[k] \leq 1/2\}} \end{aligned}$$

where  $\mathbb{E}_{(1/2+\epsilon)} [\cdot]$  denotes an expectation under the assumption that the true Bernoulli parameter is  $p = 1/2 + \epsilon$ ; (a) holds because  $(2 + 4\epsilon)/(3 + 2\epsilon) \geq 1/2$  whenever  $0 < \epsilon < 1/4$ ; (b) uses the indicator function  $1_{\{\beta[k] \leq 1/2\}}$  that is 1 if  $\beta[k] \leq 1/2$  and 0 else. Taking expectations gives

$$\mathbb{E}_{(1/2+\epsilon)} \left[ R[k] - \theta_{(1/2+\epsilon)}^* T[k] \right] \leq -\frac{\epsilon}{4} P[\beta[k] \leq 1/2] \quad (63)$$

### C. Bernoulli estimation for mean absolute error

For each  $k \in \{1, 2, 3, \dots\}$  let  $\hat{A}_k$  be Borel-measurable functions of the type:

$$\hat{A}_k : [0, 1) \times \{0, 1\}^k \rightarrow [0, 1]$$

and define

$$A[k] = \hat{A}_k(U, S[0], S[1], \dots, S[k-1])$$

The sequence of functions  $\{\hat{A}_k(\cdot)\}_{k=1}^{\infty}$  shall be called *estimation functions* because they represent any way of estimating the Bernoulli parameter  $p$  based only on  $U$  and the first  $k$  observations  $S[0], \dots, S[k-1]$ .

Let  $A[k]_p$  be the resulting (random) estimation of the Bernoulli parameter given that the true parameter is  $p$ . Let  $\mathbb{E}_p [|A[k]_p - p|]$  denote the expected mean absolute error given the true parameter is  $p$ . The following Bernoulli estimation lemma for mean absolute error is from [28] and is a modified version of a lemma for mean squared error developed in [25]:

*Lemma 11:* For any sequence of Borel-measurable functions  $\hat{A}_k(\cdot)$  of the structure defined above, we have

$$\mathbb{E}_p [|A[k]_p - p|] + \mathbb{E}_q [|A[k]_q - q|] \geq \frac{|p - q|}{4} \quad \text{whenever } |p - q| \leq \frac{\sqrt{3}}{4\sqrt{2k}}$$

*Proof:* See Lemma 6 in [28]. □

To use the above lemma, define the following estimator functions for each  $k \in \{1, 2, 3, \dots\}$ :

$$\hat{A}_k(u, s_0, \dots, s_{k-1}) = \begin{cases} 1/2 - \epsilon & \text{if } \hat{\beta}_k(u, s_0, \dots, s_{k-1}) \leq 1/2 \\ 1/2 + \epsilon & \text{if } \hat{\beta}_k(u, s_0, \dots, s_{k-1}) > 1/2 \end{cases}$$

where  $(u, s_0, \dots, s_{k-1}) \in [0, 1) \times \{0, 1\}^k$ . Then

$$\mathbb{E}_{(1/2-\epsilon)} [|A[k]_{1/2-\epsilon} - (1/2 - \epsilon)|] = 2\epsilon P[\beta[k] > 1/2] \quad (64)$$

$$\mathbb{E}_{(1/2+\epsilon)} [|A[k]_{1/2+\epsilon} - (1/2 + \epsilon)|] = 2\epsilon P[\beta[k] \leq 1/2] \quad (65)$$

Thus for each  $k \in \{1, 2, 3, \dots\}$  such that  $2\epsilon \leq \frac{\sqrt{3}}{4\sqrt{2k}}$  we have

$$\begin{aligned}
& \mathbb{E}_{(1/2-\epsilon)} \left[ R[k] - \theta_{(1/2-\epsilon)}^* T[k] \right] + \mathbb{E}_{(1/2+\epsilon)} \left[ R[k] - \theta_{(1/2+\epsilon)}^* T[k] \right] \\
& \stackrel{(a)}{\leq} \frac{-\epsilon}{4} [P[\beta[k] > 1/2] + P[\beta[k] \leq 1/2]] \\
& \stackrel{(b)}{=} \frac{-1}{8} (\mathbb{E}_{(1/2-\epsilon)} [|A[k]_{1/2-\epsilon} - (1/2 - \epsilon)|] + \mathbb{E}_{(1/2+\epsilon)} [|A[k]_{1/2+\epsilon} - (1/2 + \epsilon)|]) \\
& \stackrel{(c)}{\leq} \frac{-\epsilon}{16}
\end{aligned} \tag{66}$$

where (a) holds by (62)-(63); (b) holds by (64)-(65); (c) holds by appciation of Lemma 11 for  $p = 1/2 - \epsilon$  and  $q = 1/2 + \epsilon$ . The condition  $2\epsilon \leq \frac{\sqrt{3}}{4\sqrt{2k}}$  is equivalent to

$$k \leq \frac{3}{128\epsilon^2}$$

Fix  $K \in \{1, 2, \dots, \lfloor \frac{3}{128\epsilon^2} \rfloor\}$  (this corresponds to  $K$  in the interval specified by Theorem 5 with  $\epsilon = 64\delta$ ). Summing (66) over  $k \in \{0, \dots, K\}$  gives

$$\begin{aligned}
& \sum_{k=0}^K \mathbb{E}_{(1/2-\epsilon)} \left[ R[k] - \theta_{(1/2-\epsilon)}^* T[k] \right] + \sum_{k=0}^K \mathbb{E}_{(1/2+\epsilon)} \left[ R[k] - \theta_{(1/2+\epsilon)}^* T[k] \right] \\
& \stackrel{(a)}{\leq} \sum_{k=1}^K \frac{-\epsilon}{16} \\
& = -\frac{K\epsilon}{16}
\end{aligned}$$

where (a) neglects the nonpositive  $k = 0$  term (recall Lemma 11). It follows that either

$$\sum_{k=0}^K \mathbb{E}_{(1/2-\epsilon)} \left[ R[k] - \theta_{(1/2-\epsilon)}^* T[k] \right] \leq -\frac{K\epsilon}{32} \tag{67}$$

or

$$\sum_{k=0}^K \mathbb{E}_{(1/2+\epsilon)} \left[ R[k] - \theta_{(1/2+\epsilon)}^* T[k] \right] \leq -\frac{K\epsilon}{32} \tag{68}$$

Assume (67) holds: Then assuming  $p = 1/2 - \epsilon$  and rearranging terms in (67) yields (using notation  $\mathbb{E}[\cdot]$  instead of  $\mathbb{E}_{(1/2-\epsilon)}[\cdot]$  for simplicity):

$$\begin{aligned}
\frac{\sum_{k=0}^K \mathbb{E} [R[k]]}{\sum_{k=0}^K \mathbb{E} [T[k]]} & \leq \theta^* - \frac{K\epsilon}{32 \sum_{k=0}^K \mathbb{E} [T[k]]} \\
& \stackrel{(a)}{\leq} \theta^* - \frac{\epsilon}{64} \\
& \stackrel{(b)}{=} \theta^* - \delta
\end{aligned}$$

where (a) holds because this example has  $\mathbb{E} [T[k]] \leq 2$  for all  $k$ ; (b) holds because  $\epsilon = 64\delta$ . Therefore, if (67) holds, then running the algorithm in the case when the true parameter is  $p = 1/2 - \epsilon$  means that

$$\frac{\sum_{k=0}^K \mathbb{E} [R[k]]}{\sum_{k=0}^K \mathbb{E} [T[k]]} \leq \theta^* - \delta \tag{69}$$

On the other hand, if (67) fails then (68) must hold and by the same argument it follows that (69) is true for the case  $p = 1/2 + \epsilon$ . So for any particular causal algorithm, (69) must hold for either the case  $p = 1/2 - \epsilon$  or the case  $p = 1/2 + \epsilon$ . This proves Theorem 5.

#### D. Discussion

Recall that Theorem 11 shows that the proposed algorithm of this paper achieves an optimality gap of  $O(1/\sqrt{k})$  for general renewal optimization systems (including those without a strongly concave structure). The square root converse result of Theorem 5 is theoretically important because it matches this achievability result and demonstrates that the  $O(1/\sqrt{k})$  behavior cannot generally be improved. From the above result, one would expect a computer simulation of this system to converge more slowly when  $p \approx 1/2$ . Surprisingly, simulations show the algorithm converges quickly even for such cases (see Section VIII). This

may be due to the very small coefficient  $\frac{3}{2^{19}}$  obtained in Theorem 5. One interpretation is that, while Theorem 5 proves that no mathematical analysis can improve convergence for general systems beyond a square root law, the small constant coefficient suggests that performance is not significantly degraded for practical scenarios and timescales.

## VII. PROBABILITY 1 CONVERGENCE

This section considers the general system described in Section III-A (not necessarily having the strongly concave property used in Section IV-F). Recall that  $\{S[k]\}_{k=0}^{\infty}$  is i.i.d. with some distribution  $F_S(s) = P[S[k] \leq s]$  for  $s \in \mathbb{R}^m$ . The three theorems in this section show:

- (Theorem 6) No algorithm can have a sample path time average that exceeds  $\theta^*$ .
- (Theorem 7) The algorithm (18)-(20) with the stepsize rule  $\eta[k] = \frac{1}{(k+2)T_{min}}$  ensures  $\theta[k] \rightarrow \theta^*$  with probability 1.
- (Theorem 8) The algorithm (18)-(20) with stepsize rule  $\eta[k] = \frac{1}{(k+2)T_{min}}$  ensures

$$\lim_{K \rightarrow \infty} \frac{\sum_{k=0}^{K-1} R[k]}{\sum_{k=0}^{K-1} T[k]} = \theta^* \quad (\text{with prob 1})$$

### A. Preliminary lemma

Consider a probability experiment with sample space  $\Omega$ , sigma-algebra of events  $\mathcal{F}$ , and probability measure  $P : \mathcal{F} \rightarrow \mathbb{R}$ . For each  $i \in \{1, 2, 3, \dots\}$  let  $X_i : \Omega \rightarrow \mathbb{R}$  be a random variable.

*Lemma 12:* Assume the sequence of random variables  $\{X_i\}_{i=1}^{\infty}$  satisfies

$$\sum_{i=1}^{\infty} \frac{\mathbb{E}[X_i^2]}{i^2} < \infty \quad (70)$$

and there are finite constants  $a, b$  such that with probability 1 we have

$$a \leq \mathbb{E}[X_i | X_0, \dots, X_{i-1}] \leq b \quad \forall i \in \{1, 2, 3, \dots\} \quad (71)$$

where  $X_0$  is defined to be 0. Then with probability 1 we have

$$a \leq \liminf_{k \rightarrow \infty} \frac{1}{k} \sum_{i=1}^k X_i \leq \limsup_{k \rightarrow \infty} \frac{1}{k} \sum_{i=1}^k X_i \leq b \quad (72)$$

*Proof:* Define  $Y_0 = 0$  and for each  $i \in \{1, 2, 3, \dots\}$  define

$$Y_i = X_i - \mathbb{E}[X_i | X_0, \dots, X_{i-1}] \quad (73)$$

By the law of iterated expectations we have

$$\mathbb{E}[Y_i | Y_0, \dots, Y_{i-1}] = 0 \quad \forall i \in \{0, 1, 2, \dots\} \quad (74)$$

Further, since  $(x - y)^2 \leq 2x^2 + 2y^2$  for all real numbers  $x, y$ , we have by (73) that for all positive integers  $i$

$$Y_i^2 \leq 2X_i^2 + 2\mathbb{E}[X_i | X_0, \dots, X_{i-1}]^2$$

so for all positive integers  $i$

$$\begin{aligned} \mathbb{E}[Y_i^2] &\leq 2\mathbb{E}[X_i^2] + 2\mathbb{E}[\mathbb{E}[X_i | X_0, \dots, X_{i-1}]^2] \\ &\stackrel{(a)}{\leq} 2\mathbb{E}[X_i^2] + 2\mathbb{E}[\mathbb{E}[X_i^2 | X_0, \dots, X_{i-1}]] \\ &\stackrel{(b)}{=} 4\mathbb{E}[X_i^2] \end{aligned}$$

where (a) holds by Jensen's inequality; (b) holds by the law of iterated expectations. Since (70) holds it follows that

$$\sum_{i=1}^{\infty} \frac{\mathbb{E}[Y_i^2]}{i^2} < \infty \quad (75)$$

By a theorem in [29], the conditions (74) and (75) imply

$$\lim_{k \rightarrow \infty} \frac{1}{k} \sum_{i=1}^k Y_i = 0 \quad (\text{with prob 1}) \quad (76)$$

On the other hand, by (71) we have for all positive integers  $k$  that with probability 1:

$$a \leq \frac{1}{k} \sum_{i=1}^k \mathbb{E} [X_i | X_0, \dots, X_{i-1}] \leq b$$

Substituting the definition of  $Y_i$  into the above inequality gives for all positive integers  $k$  (with probability 1):

$$a \leq \frac{1}{k} \sum_{i=1}^k X_i - \frac{1}{k} \sum_{i=1}^k Y_i \leq b$$

Taking a lim sup of the above inequality and using (76) gives (with probability 1):

$$a \leq \limsup_{k \rightarrow \infty} \frac{1}{k} \sum_{i=1}^k X_i \leq b$$

and a similar result holds when taking the lim inf. This proves (72).  $\square$

### B. Sample path convergence

Recall that  $\{S[k]\}_{k=0}^{\infty}$  are i.i.d. vectors in  $\mathbb{R}^m$ . Let  $U$  be uniformly distributed over  $[0, 1)$  and independent of  $\{S[k]\}_{k=0}^{\infty}$ . Define

$$\begin{aligned} W[0] &= 0 \\ W[k] &= (S[0], \dots, S[k-1]) \quad \forall k \in \{1, 2, 3, \dots\} \end{aligned}$$

The variable  $U$  is used as an external source of randomness to enable potentially randomized decisions. As described in the previous section, there is no loss of generality in using a single variable  $U$ : From  $U$  we can create an infinite sequence of i.i.d. uniformly distributed variables, all deterministic functions of  $U$ . A general *causal and measurable* algorithm makes decisions on each frame  $k \in \{0, 1, 2, \dots\}$  by

$$(T[k], R[k]) = f_k(U, W[k], S[k]) \quad (77)$$

where for each  $k$ ,  $f_k(u, w, s)$  is a Borel-measurable function that satisfies

$$f_k(u, w, s) \in \mathcal{D}(s) \quad \forall s \in \Omega_S \quad (78)$$

*Theorem 6:* Under any causal and measurable algorithm we have

$$\limsup_{K \rightarrow \infty} \frac{\sum_{k=0}^{K-1} R[k]}{\sum_{k=0}^{K-1} T[k]} \leq \theta^* \quad (\text{with prob 1})$$

where  $\theta^*$  is from (15). Furthermore

$$T_{min} \leq \limsup_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} T[k] \leq T_{max} \quad (\text{with prob 1}) \quad (79)$$

*Proof:* Let  $U$  be the uniform random variable and for each  $k$  let  $(T[k], R[k])$  be the decisions from some causal and measurable algorithm (with structure defined by (77)-(78)). Fix  $k \in \{0, 1, 2, \dots\}$ . Since  $S[k]$  is independent of  $(U, W[k])$  we have from Theorem 9 in the Appendix:

$$\mathbb{E}[(T[k], R[k]) | U, W[k]] \in \bar{\mathcal{A}} \quad (\text{with prob 1}) \quad (80)$$

Furthermore from Theorem 9 and (11)-(12) we have

$$\mathbb{E} [T[k]^2 | U, W[k]] \leq C_1 \quad (\text{with prob 1}) \quad (81)$$

$$\mathbb{E} [R[k]^2 | U, W[k]] \leq C_2 \quad (\text{with prob 1}) \quad (82)$$

Define

$$X[k] = R[k] - \theta^* T[k] \quad (83)$$

**Claim 1:** There is a constant  $V$  such that for all  $k \in \{0, 1, 2, \dots\}$  we have

$$\mathbb{E} [T[k]^2] \leq V$$

$$\mathbb{E} [R[k]^2] \leq V$$

$$\mathbb{E} [X[k]^2] \leq V$$

**Proof of Claim 1:** Fix  $k \in \{0, 1, 2, \dots\}$ . Taking expectations of (81) and using the law of iterated expectations gives

$$\mathbb{E}[T[k]^2] \leq C_1 \quad (84)$$

Similarly, taking expectations of (82) gives

$$\mathbb{E}[R[k]^2] \leq C_2 \quad (85)$$

We have from the definition of  $X[k]$ :

$$\begin{aligned} X[k]^2 &= (R[k] - \theta[k]T[k])^2 \\ &\stackrel{(a)}{\leq} 2R[k]^2 + 2\theta[k]^2 T[k]^2 \\ &\stackrel{(b)}{\leq} 2R[k]^2 + 2\max\{\theta_{min}^2, \theta_{max}^2\} T[k]^2 \end{aligned}$$

where (a) holds because  $(x + y)^2 \leq 2x^2 + 2y^2$  for any real numbers  $x, y$ ; (b) holds because  $\theta[k] \in [\theta_{min}, \theta_{max}]$ . Taking expectations of the above and using (84)-(85) gives

$$\mathbb{E}[X[k]^2] \leq 2C_2 + 2\max\{\theta_{min}^2, \theta_{max}^2\} C_1 \quad (86)$$

Defining  $V$  as the maximum of the bounds in (84),(85),(86) proves Claim 1.

**Claim 2:** For all  $k \in \{0, 1, 2, \dots\}$  we have

$$\begin{aligned} T_{min} &\leq \mathbb{E}[T[k]|T[0], \dots, T[k-1]] \leq T_{max} \quad (\text{with prob 1}) \\ \mathbb{E}[X[k]|X[0], \dots, X[k-1]] &\leq 0 \quad (\text{with prob 1}) \end{aligned}$$

**Proof of Claim 2:** By (9) and Lemma 1 we know that every  $(t, r) \in \bar{\mathcal{A}}$  satisfies  $t \geq T_{min}$  and  $r/t \leq \theta^*$ . In view of (80), it holds that

$$\begin{aligned} T_{min} &\leq \mathbb{E}[T[k]|U, W[k]] \leq T_{max} \quad (\text{with prob 1}) \\ \mathbb{E}[X[k]|U, W[k]] &\leq 0 \quad (\text{with prob 1}) \end{aligned}$$

Further

$$\begin{aligned} \sigma(U, W[k]) &= \sigma(U, S[0], S[1], \dots, S[k-1]) \\ &\supseteq \sigma(T[0], T[1], \dots, T[k-1]) \end{aligned}$$

where the final inclusion holds because  $(T[0], T[1], \dots, T[k-1])$  is a deterministic function of  $U, S[0], S[1], \dots, S[k-1]$ . Thus, for any random variable  $J$  (specifically using either  $J = T[k]$  or  $J = X[k]$ ) we have

$$\mathbb{E}[J|T[0], \dots, T[k-1]] = \mathbb{E}\left[\underbrace{\mathbb{E}[J|U, W[k]]}_{inner} | T[0], \dots, T[k-1]\right]$$

so that if the inner expectation satisfies a lower and/or upper bound with probability 1, so does the outer expectation. This proves Claim 2.

Claims 1 and 2 ensure that  $T[k]$  and  $X[k]$  satisfy the requirements of Lemma 12, and so

$$T_{min} \leq \liminf_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} T[k] \leq T_{max} \quad (\text{with prob 1}) \quad (87)$$

$$\limsup_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} X[k] \leq 0 \quad (\text{with prob 1}) \quad (88)$$

The inequality (87) proves (79). Substituting the definition of  $X[k]$  (from (83)) into (88) yields (with prob 1)

$$\limsup_{K \rightarrow \infty} \left[ \left( \frac{1}{K} \sum_{k=0}^{K-1} R[k] \right) - \theta^* \left( \frac{1}{K} \sum_{k=0}^{K-1} T[k] \right) \right] \leq 0$$

and from (87) this implies the result.  $\square$

**Theorem 7:** Under algorithm (18)-(20) with stepsize  $\eta[k] = \frac{1}{(k+2)T_{min}}$  for all  $k \in \{0, 1, 2, \dots\}$  we have

$$\lim_{k \rightarrow \infty} \theta[k] = \theta^* \quad (\text{with prob 1})$$

where  $\theta^*$  is from (15).

*Proof:* Fix  $\epsilon > 0$ . By the Markov/Chebyshev inequality we have

$$P[|\theta[k] - \theta^*| \geq \epsilon] \leq \frac{\mathbb{E}[(\theta[k] - \theta^*)^2]}{\epsilon^2} \leq \frac{2b}{k\epsilon^2 T_{min}^2} \quad \forall k \in \{1, 2, 3, \dots\}$$

where the final inequality holds by (34). It follows that

$$\sum_{i=1}^{\infty} P[|\theta[i^2] - \theta^*| \geq \epsilon] \leq \frac{2b}{\epsilon^2 T_{min}^2} \sum_{i=1}^{\infty} \frac{1}{i^2} < \infty$$

And so the Borel-Cantelli theorem ensures that  $\{|\theta[i^2] - \theta^*| \geq \epsilon\}$  happens for an at most finite number of indices  $i$  with probability 1. Since  $\epsilon > 0$  was arbitrary, this implies

$$\lim_{i \rightarrow \infty} \theta[i^2] = \theta^* \quad (\text{with prob 1}) \quad (89)$$

Every positive integer  $k$  must be between two perfect squares  $n_k^2$  and  $(n_k + 1)^2$ :

$$n_k^2 \leq k < (n_k + 1)^2$$

Then for all positive integers  $k$  we have

$$|\theta[k] - \theta^*| \leq |\theta[k] - \theta[n_k^2]| + |\theta[n_k^2] - \theta^*|$$

Taking a lim sup of the above inequality and using (89) yields

$$\limsup_{k \rightarrow \infty} |\theta[k] - \theta^*| \leq \limsup_{k \rightarrow \infty} |\theta[k] - \theta[n_k^2]| \quad (\text{with prob 1})$$

It suffices to show the right-hand-side of the above inequality is 0 with probability 1. For each positive integer  $k$  define

$$G_k = \max_{i \in \{k^2, \dots, (k+1)^2 - 1\}} \{(\theta[i] - \theta[k^2])^2\}$$

Fix  $\epsilon > 0$ . It suffices to show that, with probability 1,  $\{G_k > \epsilon\}$  occurs for at most finitely many positive integers  $k$ . To show this, by the Borel-Cantelli theorem it suffices to show

$$\sum_{k=1}^{\infty} P[G_k > \epsilon] < \infty \quad (90)$$

For each positive integer  $k$  we have by the Markov inequality:

$$\begin{aligned} P[G_k > \epsilon] &\leq \frac{\mathbb{E}[G_k]}{\epsilon} \\ &= \frac{1}{\epsilon} \mathbb{E} \left[ \max_{i \in \{k^2, \dots, (k+1)^2 - 1\}} \{(\theta[i] - \theta[k^2])^2\} \right] \end{aligned} \quad (91)$$

Let  $V$  be a finite constant that satisfies

$$\mathbb{E}[(R[i] - \theta[i]T[i])^2] \leq V \quad \forall i \in \{0, 1, 2, \dots\} \quad (92)$$

Such a value  $V$  exists because  $\theta[i] \in [\theta_{min}, \theta_{max}]$  always, and second moments of  $R[i]$  and  $T[i]$  are uniformly bounded for all  $i$ . Observe by the update procedure for  $\theta[j]$  in (20) we have for all  $j \in \{0, 1, 2, \dots\}$

$$\begin{aligned} |\theta[j+1] - \theta[j]| &\stackrel{(a)}{\leq} \left| [\theta[j] + \eta[j](R[j] - \theta[j]T[j])]_{\theta_{min}}^{\theta_{max}} - [\theta[j]]_{\theta_{min}}^{\theta_{max}} \right| \\ &\stackrel{(b)}{\leq} |\eta[j](R[j] - \theta[j]T[j])| \end{aligned} \quad (93)$$

where (a) uses  $\theta[j] \in [\theta_{min}, \theta_{max}]$ ; (b) uses the fact that the distance between the projections of two numbers onto a closed interval is less than or equal to the distance between the numbers. For all positive integers  $k$  and all  $i \in \{k^2, \dots, (k+1)^2 - 1\}$ :

$$\begin{aligned} (\theta[i] - \theta[k^2])^2 &= \left( \sum_{j=k^2}^{i-1} (\theta[j+1] - \theta[j]) \right)^2 \\ &\leq \left( \sum_{j=k^2}^{i-1} |\theta[j+1] - \theta[j]| \right)^2 \\ &\stackrel{(a)}{\leq} \left( \sum_{j=k^2}^{i-1} |\eta[j](R[j] - \theta[j]T[j])| \right)^2 \\ &\stackrel{(b)}{\leq} \frac{1}{(k^2+2)^2 T_{min}^2} \sum_{j=k^2}^{(k+1)^2-1} \sum_{r=k^2}^{(k+1)^2-1} |R[j] - \theta[j]T[j]| \cdot |R[r] - \theta[r]T[r]| \end{aligned}$$

where (a) holds by (93); (b) holds because  $\eta[j] \leq \frac{1}{(k^2+2)T_{min}}$  for all  $j \geq k^2$ . Taking the maximum of the above inequality over all  $i \in \{k^2, \dots, (k+1)^2 - 1\}$  gives

$$\max_{i \in \{k^2, \dots, (k+1)^2 - 1\}} \{(\theta[i] - \theta[k^2])^2\} \leq \frac{1}{(k^2+2)^2 T_{min}^2} \sum_{j=k^2}^{(k+1)^2-1} \sum_{r=k^2}^{(k+1)^2-1} |R[j] - \theta[j]T[j]| \cdot |R[r] - \theta[r]T[r]|$$

Taking an expectation and using the Cauchy-Schwarz inequality and (92) gives

$$\mathbb{E} \left[ \max_{i \in \{k^2, \dots, (k+1)^2 - 1\}} \{(\theta[i] - \theta[k^2])^2\} \right] \leq \frac{V((k+1)^2 - k^2)^2}{(k^2+2)^2 T_{min}^2} = \frac{V(2k+1)^2}{(k^2+2)^2 T_{min}^2}$$

Substituting this into (91) gives

$$P[G_k > \epsilon] \leq \frac{1}{\epsilon} \frac{V(2k+1)^2}{(k^2+2)^2 T_{min}^2}$$

which decays like  $O(1/k^2)$  and so  $\sum_{k=1}^{\infty} P[G_k > \epsilon] < \infty$ . □

*Theorem 8:* Under algorithm (18)-(20) with stepsize  $\eta[k] = \frac{1}{(k+2)T_{min}}$  for all  $k \in \{0, 1, 2, \dots\}$  we have

$$\lim_{K \rightarrow \infty} \frac{\sum_{k=0}^{K-1} R[k]}{\sum_{k=0}^{K-1} T[k]} = \theta^* \quad (\text{with prob 1})$$

where  $\theta^*$  is from (15).

*Proof:* We have by the rule (18)-(19):

$$R[k] - \theta[k]T[k] \geq R^*[k] - \theta[k]T^*[k] \quad \forall k \in \{0, 1, 2, \dots\} \quad (94)$$

where  $(T[k], R[k]) \in \mathcal{D}(S[k])$  is the decision made by the algorithm and  $(T^*[k], R^*[k]) \in \mathcal{D}(S[k])$  is any alternative decision. Fix  $(t, r) \in \mathcal{A}$ , so that  $(t, r)$  can be achieved as an expectation of  $(T^*[k], R^*[k])$  on frame  $k$  under some particular decision rule. Let  $(T^*[k], R^*[k]) \in \mathcal{D}(S[k])$  be the decision that is made based purely on observing  $S[k]$ , independent of the past, and that satisfies  $\mathbb{E}[(T^*[k], R^*[k])] = (t, r)$ . Then  $\{(T^*[k], R^*[k])\}_{k=0}^{\infty}$  is a sequence of i.i.d. vectors and so by the law of large numbers:

$$\lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} (T^*[k], R^*[k]) = (t, r) \quad (\text{with prob 1}) \quad (95)$$

Rearranging terms in (94) gives the following for all  $k \in \{0, 1, 2, \dots\}$ :

$$R[k] - \theta^* T[k] \geq R^*[k] - \theta^* T^*[k] + (\theta[k] - \theta^*)(T[k] - T^*[k])$$

Summing over  $k \in \{0, \dots, K-1\}$  and dividing by  $K$  gives

$$\begin{aligned} &\frac{1}{K} \sum_{k=0}^{K-1} R[k] - \theta^* \frac{1}{K} \sum_{k=0}^{K-1} T[k] \\ &\geq \frac{1}{K} \sum_{k=0}^{K-1} R^*[k] - \theta^* \frac{1}{K} \sum_{k=0}^{K-1} T^*[k] - \frac{1}{K} \sum_{k=0}^{K-1} (\theta[k] - \theta^*)(T[k] - T^*[k]) \\ &\geq \frac{1}{K} \sum_{k=0}^{K-1} R^*[k] - \theta^* \frac{1}{K} \sum_{k=0}^{K-1} T^*[k] - \frac{1}{K} \sum_{k=0}^{K-1} |\theta[k] - \theta^*| (T[k] + T^*[k]) \end{aligned} \quad (96)$$

**Claim:** We have with probability 1 we have:

$$\limsup_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} |\theta[k] - \theta^*| (T[k] + T^*[k]) = 0 \quad (97)$$

We postpone the proof of (97). Taking a lim inf of (96) and substituting (97) and (95) gives, with probability 1:

$$\liminf_{K \rightarrow \infty} \left[ \frac{1}{K} \sum_{k=0}^{K-1} R[k] - \theta^* \frac{1}{K} \sum_{k=0}^{K-1} T[k] \right] \geq r - \theta^* t \quad (\text{with prob 1})$$

This holds for all  $(t, r) \in \mathcal{A}$ . Taking a limit over a countably infinite sequence of points  $(t_i, r_i) \in \mathcal{A}$  that approach the point  $(t^*, r^*) \in \bar{\mathcal{A}}$  and using the fact that  $\theta^* = r^*/t^*$  gives

$$\liminf_{K \rightarrow \infty} \left[ \frac{1}{K} \sum_{k=0}^{K-1} R[k] - \theta^* \frac{1}{K} \sum_{k=0}^{K-1} T[k] \right] \geq 0 \quad (\text{with prob 1}) \quad (98)$$

This, together with (79), proves that with probability 1:

$$\liminf_{K \rightarrow \infty} \frac{\sum_{k=0}^{K-1} R[k]}{\sum_{k=0}^{K-1} T[k]} \geq \theta^*$$

On the other hand, Theorem 6 ensures the lim sup is less than or equal to  $\theta^*$  with probability 1. Thus, the limit is exactly  $\theta^*$  (with prob 1).

It remains to prove (97) of the Claim. We know  $\theta[k] \rightarrow \theta^*$  with probability 1. Fix  $\epsilon > 0$ . With probability 1 we know  $|\theta[k] - \theta^*| \leq \epsilon$  for all sufficiently large  $k$  and so

$$\limsup_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} |\theta[k] - \theta^*| (T[k] + T^*[k]) \leq \epsilon \limsup_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} (T[k] + T^*[k]) \quad (\text{with prob 1})$$

Observe that both  $T[k]$  and  $T^*[k]$  are from causal and measurable algorithms and so they both satisfy (79). Thus

$$\limsup_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} |\theta[k] - \theta^*| (T[k] + T^*[k]) \leq \epsilon 2T_{max} \quad (\text{with prob 1})$$

This holds for all  $\epsilon > 0$  and so (97) follows.  $\square$

## VIII. SIMULATION

This section presents simulations of the proposed algorithm under the initial condition  $\theta[0] = \theta_{min}$  and stepsize  $\eta[k] = \frac{1}{(k+2)T_{min}}$ .

### A. System 1: Selecting one of multiple projects

Consider the project selection problem of Section II-C. On each frame  $k$  we receive  $N[k]$  new potential projects, where  $N[k] \in \{0, 1, 2, 3\}$  with  $P[N[k] = i] = p_i$  and

$$p_0 = 0.1, p_1 = 0.9 - p, p_2 = p/2, p_3 = p/2$$

where  $p \in [0, 0.9]$  is a parameter varied in the simulations (larger values of  $p$  yield stochastically more projects). The decision set for each frame  $k$  is

$$\mathcal{D}(S[k]) = \{(1, 0), (T_1, R_1), \dots, (T_{N[k]}, R_{N[k]})\}$$

where the  $(1, 0)$  option corresponds to working on no project for 1 time unit (and receiving no profit). Given that  $N[k] = i$ , the vectors  $(T_j, R_j)$  for  $j \in \{1, \dots, i\}$  are generated independently with  $T_j \sim \text{Uniform}([1, 10])$  and  $R_j = A_j T_j$  where  $A_j \sim \text{Uniform}([0, 50])$ . The proposed algorithm uses  $[\theta_{min}, \theta_{max}] = [0, 50]$  and  $T_{min} = 1$ .

Fig. 3 compares the proposed algorithm to the greedy strategy that chooses the project  $j$  that maximizes the ratio  $R_j/T_j$ . The proposed algorithm has significant gains. Fig. 4 explores convergence of the time average for one sample path of the proposed algorithm over 2000 frames for various parameter values  $p$ . It is difficult to calculate  $\theta^*$  analytically so the dashed horizontal lines in Fig. 4 are estimated values of  $\theta^*$  obtained from 5000 independent runs. It is interesting to note that, considering only frames for which  $N[k] \geq 1$ , the proposed algorithm chooses to reject all offered projects a significant fraction of time: For parameters  $p \in \{0, 0.3, 0.6, 0.9\}$  the conditional rejection probabilities (given  $N[k] \geq 1$ ) were 0.46, 0.40, 0.33, 0.25, respectively.

Fig. 5 compares the proposed algorithm to the  $\theta$ -empirical heuristic algorithm of [5] (see Section VI.b of [5]). The  $\theta$ -empirical heuristic has a similar structure that maximizes  $R[k] - \theta[k]T[k]$  over all  $(T[k], R[k])$  choices, but sets  $\theta[k]$  to the empirical

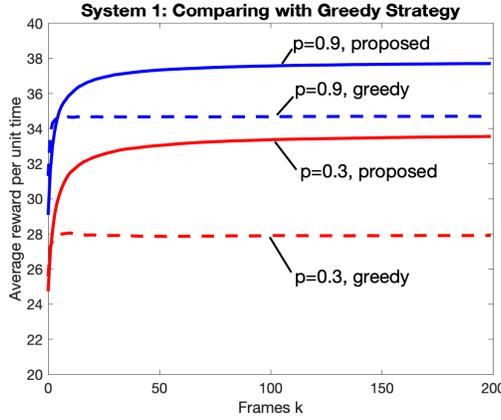


Fig. 3. Comparing the proposed algorithm with the greedy strategy for two different values of  $p$ . Data is averaged over 5000 independent sample paths.

time average reward seen up to frame  $k$ . In [5] and [6] it is shown that *if this algorithm converges* then it converges to the optimal  $\theta^*$ , but no proof of convergence and no convergence time results are known. From simulation it yields very similar results, and even (slightly) faster convergence, in comparison to the proposed algorithm (compare the dashed curves to the solid curves of the same color in Fig. 5). The advantage of the proposed algorithm is that it comes with a proof of convergence along with convergence time guarantees.

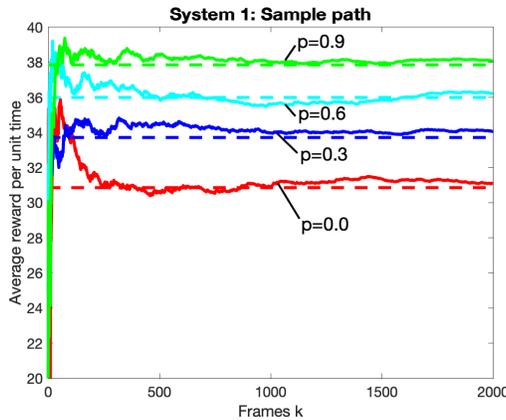


Fig. 4. Simulation of a single sample path for System 2:  $\sum_{i=0}^{k-1} R[i] / \sum_{i=0}^{k-1} T[i]$  versus  $k \in \{0, \dots, 2000\}$  for four different values of  $p$ . Dashed horizontal lines are obtained by averaging the final value at time 2000 over 5000 independent sample paths.

### B. System 2: A curve of choices

Now consider the system described in Section V-A, where  $\theta^*$  is analytically known. The proposed algorithm uses  $[\theta_{min}, \theta_{max}] = [1, 2]$  and  $T_{min} = 1$ . If  $S[k] = 1$  the algorithm chooses from a curve of choices:

$$(T[k], R[k]) = (x[k], 2 - (2 - x[k])^2)$$

to maximize  $T[k] - \theta[k]R[k]$  subject to  $1 \leq x[k] \leq 2$ , which has solution  $x[k] = \left[2 - \frac{\theta[k]}{2}\right]_1^2$ . Results for four different parameter values of  $p = P[S[k] = 1]$  are in Figs. 6 and 7. The dashed horizontal lines are the analytically optimal  $\theta^*$  values in (46). Fig. 6 shows how close one sample path time average comes to  $\theta^*$  after 1000 frames. When the simulation time is extended it was observed that all four sample paths settled into near-constant values that were indistinguishable from  $\theta^*$ , which is consistent with the probability 1 sample path convergence proven in Section VII. Fig. 7 shows the expected performance (with expectations computed by averaging over 5000 independent sample paths). Fig. 7 plots  $\mathbb{E} \left[ \frac{\sum_{i=0}^{k-1} R[i]}{\sum_{i=0}^{k-1} T[i]} \right]$ . It was observed that plots of  $\frac{\sum_{i=0}^{k-1} \mathbb{E}[R[i]]}{\sum_{i=0}^{k-1} \mathbb{E}[T[i]]}$  over the same number of frames looked similar (those plots are omitted for brevity).

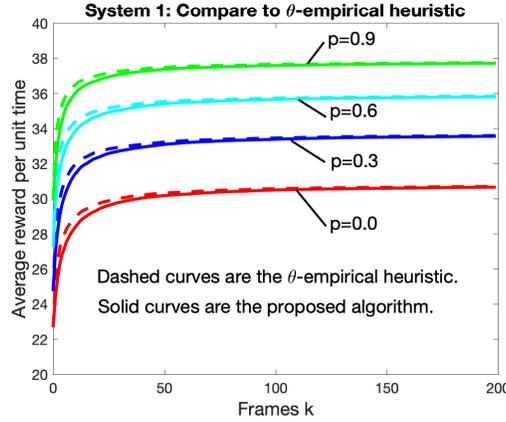


Fig. 5. Comparing the proposed algorithm to the  $\theta$ -empirical heuristic of [5]. The simulation averages over 5000 independent sample paths and plots  $\mathbb{E}[\sum_{i=0}^{k-1} R[i] / \sum_{i=0}^{k-1} T[i]]$  versus  $k \in \{0, \dots, 200\}$  for four different values of  $p$ .

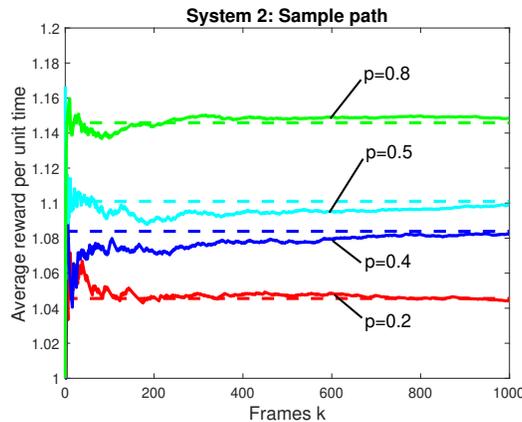


Fig. 6. Simulation of a single sample path for System 2:  $\sum_{i=0}^{k-1} R[i] / \sum_{i=0}^{k-1} T[i]$  versus  $k \in \{0, \dots, 1000\}$  for four different Bernoulli probabilities  $p$ . Dashed horizontal lines are the optimal values.

### C. System 3: Two choices

Consider the system of Section III, which is the same system for which a square-root converse result was proven in Section VI. When  $S[k] = 1$  there are only two choices:  $(T[k], R[k]) \in \{(1, 1), (2, 3)\}$ . We use  $[\theta_{min}, \theta_{max}] = [1, 3]$ ,  $T_{min} = 1$ . Fig. 8 plots data from a single sample path run over 1000 frames for four different values of  $p = P[S[k] = 1]$ . The dashed horizontal lines are the exact  $\theta^*$  values computed analytically in Section VI. Fig. 9 plots smoother curves that are averaged over 5000 independent sample paths. The curves in Fig. 9 are plotted over the smaller timeline  $k \in \{0, \dots, 50\}$  to show convergence of the expected value. It can be seen that the algorithm converges quickly to optimality for all  $p$  choices. There was no significant behavioral difference observed when  $p \approx 0.5$ , even though the  $p = 0.5$  threshold played a crucial role in the square root converse result.<sup>10</sup> Indeed, simulations were considered with  $p = 0.5 \pm \delta$  for various small  $\delta$  values including  $\delta = 0$  (those curves fell in between the  $p = 0.49$  and  $p = 0.51$  curves of Fig. 9 but the data is omitted for clarity of the plots and for brevity). Our hypothesis about why the  $p = 0.5$  threshold was not more noticeable in the simulations is discussed in Section VI-D.

## IX. CONCLUSION

This paper develops an online algorithm for making decisions in a renewal system. The algorithm is shown to converge to the optimal time average reward with the fastest possible asymptotic convergence time. The algorithm adjusts an auxiliary variable according to a Robbins-Monro iteration. It also makes online decisions on each frame that are informed by the current value of this variable. When the system has a strongly concave structure the algorithm is shown to have an optimality gap of  $O(\log(k)/k)$ . A matching converse result shows this gap is the best possible in the strongly concave scenario. In general

<sup>10</sup>When one stares at Fig. 9 long enough, one might become convinced of a *very slight convergence time increase* for the green and black curves, representing data when  $p \approx 0.5$ , in comparison to the curves when  $p$  is far from 0.5. However, this difference is minor. The author expected to see a bigger behavioral difference about  $p = 0.5$ , but that did not occur. See discussion in Section VI-D.

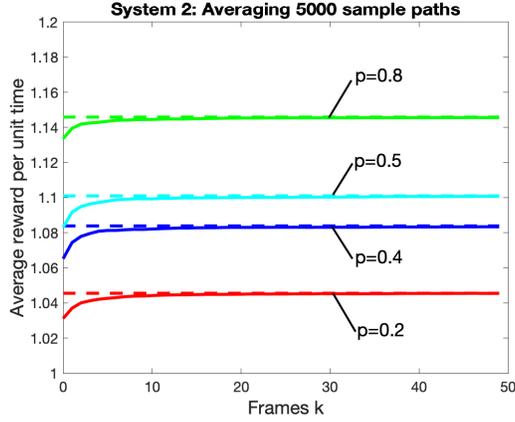


Fig. 7. Simulation averaging over 5000 sample paths for System 2:  $\mathbb{E} \left[ \frac{\sum_{i=0}^{k-1} R[i]}{\sum_{i=0}^{k-1} T[i]} \right]$  versus  $k \in \{0, \dots, 50\}$  for four different Bernoulli probabilities  $p$ . Dashed horizontal lines are the optimal values.

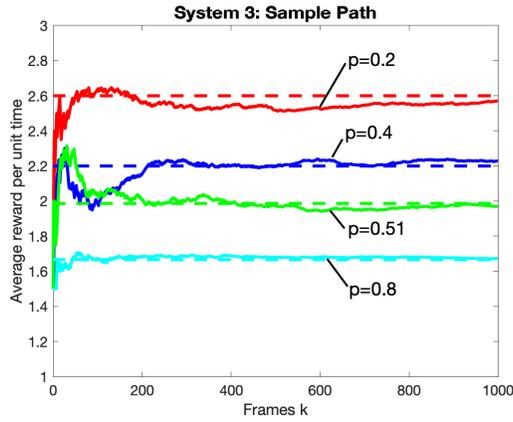


Fig. 8. Simulation of a single sample path for System 3:  $\frac{\sum_{i=0}^{k-1} R[i]}{\sum_{i=0}^{k-1} T[i]}$  versus  $k \in \{0, \dots, 1000\}$  for four different Bernoulli probabilities  $p$ . Dashed horizontal lines are the optimal values.

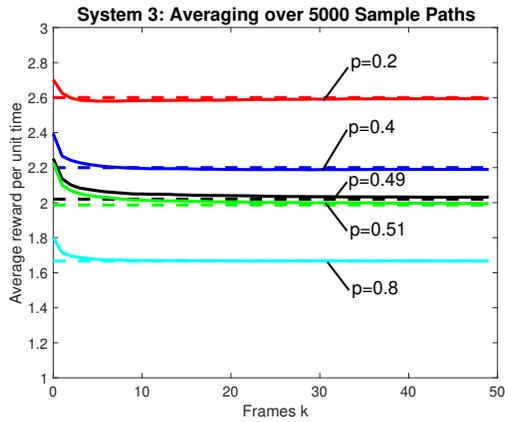


Fig. 9. Simulation averaging over 5000 sample paths for System 3:  $\mathbb{E} \left[ \frac{\sum_{i=0}^{k-1} R[i]}{\sum_{i=0}^{k-1} T[i]} \right]$  versus  $k \in \{0, \dots, 50\}$  for five different Bernoulli probabilities  $p$ . Dashed horizontal lines are the optimal values.

conditions (without strong concavity) the algorithm was shown to have an optimality gap of  $O(1/\sqrt{k})$  and a matching converse was also demonstrated. The convergence results are presented in terms of expectations achieved by the algorithm. The algorithm was also shown to have sample paths that converge to optimality with probability 1.

#### APPENDIX

Consider a probability experiment with sample space  $\Omega$ , sigma-algebra of events  $\mathcal{F}$ , and probability measure  $P : \mathcal{F} \rightarrow \mathbb{R}$ . Fix  $m, v$  as positive integers and consider random vectors

$$\begin{aligned} U &: \Omega \rightarrow [0, 1) \\ W &: \Omega \rightarrow \mathbb{R}^v \\ S &: \Omega \rightarrow \mathbb{R}^m \end{aligned}$$

Assume that

- $U$  is uniformly distributed over  $[0, 1)$ .
- $S$  is independent of  $(U, W)$ .

Define  $\Omega_S$  and  $\Omega_W$  by

$$\begin{aligned} \Omega_S &= \{S(\omega) : \omega \in \Omega\} \\ \Omega_W &= \{W(\omega) : \omega \in \Omega\} \end{aligned}$$

For each  $s \in \Omega_S$  let  $D(s) \subseteq \mathbb{R}^n$  be a general set. We consider random vectors  $X : \Omega \rightarrow \mathbb{R}^n$  of the form

$$X = f(U, W, S)$$

where  $f$  is a function that has the following properties, called *f-requirements*:

- $f : \mathbb{R} \times \mathbb{R}^v \times \mathbb{R}^m \rightarrow \mathbb{R}^n$  is Borel-measurable.
- $f(u, w, s) \in D(s)$  for all  $u \in [0, 1)$ ,  $w \in \Omega_W$ ,  $s \in \Omega_S$ .

Intuitively, we want to observe  $S$  and choose a random vector  $X \in D(S)$ , where the vector  $X$  is informed by observations of  $(U, W)$ . The random vector  $W$  represents any other parameters of the system that can be observed that are distinct from  $S$ . The random variable  $U$  represents a source of external randomness that can be used to choose  $X$ .

It is assumed throughout that there exists at least one function  $f$  that satisfies the *f-requirements*. We want any random vector  $X$  that is constructed this way to have a finite and bounded expectation. For this, we additionally assume there is a nonnegative function  $M : \Omega_S \rightarrow [0, \infty)$  such that

$$\left( \sup_{x \in D(s)} \|x\| \right) \leq M(s) \quad \forall s \in \Omega_S \quad (99)$$

where  $\|x\| = \sqrt{\sum_{i=1}^n x_i^2}$ , and such that  $M(S)$  is a random variable with finite expectation  $\mathbb{E}[M(S)]$ . Let  $A \subseteq \mathbb{R}^n$  be the set of all expectations  $\mathbb{E}[X] = (\mathbb{E}[X_1], \dots, \mathbb{E}[X_n])$  that can be achieved in this way, considering all possible functions  $f$  that satisfy the *f-requirements*. Let  $\bar{A}$  be the closure of  $A$ . It can be shown that  $\bar{A}$  is nonempty, compact, and convex.

*Theorem 9:* Let  $X = f(U, W, S)$  where  $f$  satisfies the *f-requirements*. Let  $Z$  be a random variable that is a version of  $\mathbb{E}[X|U, W]$ . Then

$$P[Z \in \bar{A}] = 1 \quad (100)$$

Further, there exists a random variable  $\tilde{Z}$  that is surely in the set  $\bar{A}$  and that is also a version of  $\mathbb{E}[X|U, W]$ .

*Proof:* Let  $Z$  be a version of  $\mathbb{E}[X|U, W]$ . Then

$$\sigma(Z) \subseteq \sigma((U, W)) \quad (101)$$

where  $\sigma(Z)$  is the sigma-algebra generated by the random vector  $Z$ , and  $\sigma((U, W))$  is the sigma-algebra generated by the random vector  $(U, W)$ . The set  $\bar{A}$  is closed, hence Borel-measurable, so  $P[Z \in \bar{A}]$  is well defined. Suppose (100) fails, so that

$$P[Z \in \bar{A}] < 1 \quad (102)$$

We shall reach a contradiction. Since  $\bar{A}$  is compact and convex it is *constructible*, meaning it is the *countable* intersection of closed half-spaces (see Proposition 7.5.6 in [30]):

$$\bar{A} = \bigcap_{j=1}^{\infty} \{x \in \mathbb{R}^n : a_j^\top x \leq b_j\} \quad (103)$$

for some  $a_j \in \mathbb{R}^k$  and  $b_j \in \mathbb{R}$  for all  $j \in \{1, 2, 3, \dots\}$ . Thus

$$P[Z \in \bar{A}] = P[\bigcap_{j=1}^{\infty} \{a_j^\top Z \leq b_j\}]$$

and so from (102):

$$P[\cap_{j=1}^{\infty} \{a_j^\top Z \leq b_j\}] < 1 \quad (104)$$

The above inequality implies it is *impossible* to have  $P[a_j^\top Z \leq b_j] = 1$  for all  $j \in \{1, 2, 3, \dots\}$  (since the intersection of a countably infinite number of probability-1 events is again a probability-1 event). So there is at least one positive integer  $k$  such that

$$P[a_k^\top Z \leq b_k] < 1 \quad (105)$$

For simplicity of notation define  $a = a_k$  and  $b = b_k$ . With this notation (105) becomes  $P[a^\top Z \leq b] < 1$  and so

$$P[a^\top Z > b] > 0 \quad (106)$$

Define  $B \subseteq \Omega$  as the following event:

$$B = \{a^\top Z > b\} \quad (107)$$

By construction of  $B$  we have

$$B \in \sigma(Z) \subseteq \sigma((U, W))$$

where the final inclusion holds by (101). By (106) we know

$$P[B] > 0 \quad (108)$$

Since  $Z$  is a version of  $\mathbb{E}[X|U, W]$ , we know that  $a^\top Z$  is a version of  $\mathbb{E}[a^\top X|U, W]$ . Since  $B \in \sigma((U, W))$  we have by the defining property of a conditional expectation of  $a^\top X$  given  $(U, W)$ :

$$\mathbb{E}[a^\top X 1_B] = \mathbb{E}[a^\top Z 1_B] \quad (109)$$

where  $1_B$  is an indicator function for event  $B$ .

**Claim:**  $\mathbb{E}[a^\top Z 1_B] > bP[B]$ .

**Proof of Claim:** By definition of event  $B$  in (107) we have

$$\{(a^\top Z - b)1_B > 0\} \iff B$$

and so

$$P[(a^\top Z - b)1_B > 0] = P[B] > 0$$

where the final inequality holds by (108). Since the nonnegative random variable  $(a^\top Z - b)1_B$  has a nonzero probability of being positive, it must have positive expectation. This proves the claim.

Substituting the result of the claim into (109) gives:

$$\mathbb{E}[a^\top X 1_B] > bP[B]$$

Dividing by the positive number  $P[B]$  gives

$$a^\top \left( \frac{\mathbb{E}[X 1_B]}{P[B]} \right) > b$$

Since  $P[B] > 0$ , the above inequality is equivalent to

$$a^\top \mathbb{E}[X|B] > b \quad (110)$$

We now construct another random variable  $\tilde{X} = \tilde{f}(U, W, S)$  for some other function  $\tilde{f}$  that meets the  $f$ -requirements. For any such  $\tilde{X}$  we have by definition of the set  $A$ :

$$\mathbb{E}[\tilde{X}] \in A \subseteq \bar{A} \subseteq \{x \in \mathbb{R}^n : a^\top x \leq b\}$$

where the final inclusion is from (103), and this directly implies

$$a^\top \mathbb{E}[\tilde{X}] \leq b \quad (111)$$

We can reach a contradiction between (111) and (110) if the expectation of  $\mathbb{E}[\tilde{X}]$  can be chosen to match  $\mathbb{E}[X|B]$ .

This can be done as follows: Since  $P[B] > 0$  we can define

$$F_{U,W|B}(u, w) = P[U \leq u, W \leq w|B] \quad \forall (u, w) \in \mathbb{R} \times \mathbb{R}^v$$

where vector inequality  $W \leq w$  is taken entrywise. Map the uniformly distributed variable  $U$  through a deterministic and Borel-measurable function  $\beta: \mathbb{R} \rightarrow \mathbb{R}^{v+1}$  to obtain a random vector

$$(\alpha(U), \beta(U)) = (\alpha(U), \beta_1(U), \dots, \beta_v(U))$$

that has distribution  $F_{U,W|B}(u, w)$  for all  $u \in \mathbb{R}$  and  $w \in \mathbb{R}^v$ . Define the function  $\tilde{f}$  by

$$\tilde{f}(u, w, s) = f(\alpha(u), \beta(u), s) \quad \forall u \in \mathbb{R}, w \in \mathbb{R}^v, s \in \mathbb{R}^m$$

Since  $f$  is Borel-measurable and  $\beta$  is Borel-measurable, we know  $\tilde{f}$  is Borel-measurable. Further, for all  $(u, w, s) \in [0, 1) \times \Omega_W \times \Omega_S$  we have

$$\tilde{f}(u, w, s) = f(\alpha(u), \beta(u), s) \in D(s)$$

where the final inclusion holds because  $f$  satisfies the second  $f$ -requirement. Thus,  $\tilde{f}$  satisfies both  $f$ -requirements.

The random variable  $\tilde{X}$  is defined by

$$\tilde{X} = \tilde{f}(U, W, S) = f(\alpha(U), \beta(U), S) \quad (112)$$

Since  $S$  is independent of  $(U, W)$  and hence independent of  $(\alpha(U), \beta(U))$ , the random vector  $(\alpha(U), \beta(U), S)$  has the same distribution as the conditional distribution of  $(U, W, S)$  given that event  $B$  occurs. Thus

$$\mathbb{E}[f(\alpha(U), \beta(U), S)] = \mathbb{E}[f(U, W, S)|B]$$

Multiplying both sides of the above equality by  $a^\top$  and substituting  $\tilde{X} = f(\alpha(U), \beta(U), S)$  gives

$$a^\top \mathbb{E}[\tilde{X}] = a^\top \mathbb{E}[f(U, W, S)|B] > b$$

where the final inequality holds by (I10). This contradicts (I11). This completes the proof of  $P[Z \in \bar{A}] = 1$ .

Finally, to construct another version of  $\mathbb{E}[X|U, W]$  that is *surely* in the set  $\bar{A}$  (rather than just *almost surely*), fix any vector  $a \in \bar{A}$  and define  $\tilde{Z}$  by

$$\tilde{Z} = \begin{cases} Z & \text{if } Z \in \bar{A} \\ a & \text{else} \end{cases}$$

Then  $\tilde{Z} \in \bar{A}$  surely, and  $\sigma(\tilde{Z}) \subseteq \sigma(Z) \subseteq \sigma((U, W))$ . But  $\tilde{Z}$  differs from  $Z$  only on a set of measure zero, and so  $\tilde{Z}$  is also a version of  $\mathbb{E}[X|U, W]$ .  $\square$

Lemma 2 from Section IV-A is a special case of Theorem 9: Fix  $k \in \{0, 1, 2, \dots\}$ , define  $S = S[k]$ ,  $W = \theta[k]$ ,  $(X_1, X_2) = (T[k], R[k])$ . The algorithm makes decisions for  $(X_1, X_2)$  as a function of  $W$  and  $S$  (without using a source of randomness  $U$ ).

## REFERENCES

- [1] R. Gallager. *Discrete Stochastic Processes*. Kluwer Academic Publishers, Boston, 1996.
- [2] S. Schaible. Fractional programming. *Zeitschrift für Operations Research*, vol. 27, no. 1, pp. 39-54, Dec. 1983.
- [3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [4] B. Fox. Markov renewal programming by linear fractional programming. *Siam J. Appl. Math.*, vol. 14, no. 6, Nov. 1966.
- [5] M. J. Neely. Dynamic optimization and learning for renewal systems. *IEEE Transactions on Automatic Control*, vol. 58, no. 1, pp. 32-46, Jan. 2013.
- [6] M. J. Neely. *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Morgan & Claypool, 2010.
- [7] X. Wei and M. J. Neely. Data center server provision: Distributed asynchronous control for coupled renewal systems. *IEEE/ACM Transactions on Networking*, 25(5), Aug. 2017.
- [8] X. Wei and M. J. Neely. Asynchronous optimization over weakly coupled renewal systems. *Stochastic Systems*, 8(3), Sept. 2018.
- [9] M. J. Neely. Low power dynamic scheduling for computing systems. In F. R. Yu, X. Zhang, and V. C. M. Leung, editors, *Green Communications and Networking*, pages pp. 219–259. CRC Press, 2012.
- [10] C. Li and M. J. Neely. Solving convex optimization with side constraints in a multi-class queue by adaptive  $c\mu$  rule. *Queueing Systems*, vol. 77, pp. 331-372, 2014.
- [11] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.
- [12] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience Series in Discrete Mathematics, John Wiley, 1983.
- [13] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.
- [14] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.
- [15] H. J. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003.
- [16] V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Springer, 2008.
- [17] P. Toulis, T. Horel, and E. M. Airoldi. The proximal Robbins-Monro method. *arXiv:1510.00967v4*, Feb. 2020.
- [18] P. Toulis, E. Airoldi, and J. Rennie. Statistical analysis of stochastic gradient methods for generalized linear models. *Proc. 31st International Conference on Machine Learning*, 32(2):667–675, 2014.
- [19] V. R. Joseph. Efficient Robbins-Monro procedure for binary data. *Biometrika*, 91(2):461–470, June 2004.
- [20] S. Mandt, M. D. Hoffman, and D. M. Blei. Stochastic gradient descent as approximate Bayesian inference. *Journal of Machine Learning Research*, 18:1–35, 2017.
- [21] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. In *Foundations and Trends in Machine Learning*, pages 1–122, January 2012.
- [22] P. Auer, N. Cesa-Bianchi, and Y. Freund and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. *36th Annual Symposium on Foundations of Computer Science*, pp. 322-331, Nov. 1995.
- [23] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.
- [24] S. Cayci, A. Eryilmaz, and R. Srikant. Learning to control renewal processes with bandit feedback. *Proc. ACM SIGMETRICS*, July 2019.

- [25] E. Hazan and S. Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, vol. 15 (July):pp. 2489–2512, 2014.
- [26] M. J. Neely. A converse result on convergence time for opportunistic wireless scheduling. *Proc. IEEE INFOCOM*, 2020.
- [27] S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231—357, 2015.
- [28] M. J. Neely. A converse result on convergence time for opportunistic wireless scheduling. *ArXiv technical report, arXiv:2001.01031v4*, March 2020.
- [29] Y. S. Chow. On a strong law of large numbers for martingales. *Ann. Math Statist*, vol. 38, no. 2, 1967.
- [30] J. M. Borwein and J. D. Vanderwerff. *Convex Functions: Constructions, Characterizations and Counterexamples*. Cambridge University Press, 2010.