Journal of Machine Learning Research 22 (2021) 1-44 Submitted 7/20; Revised 5/21; Published 9/21

Fast Learning for Renewal Optimization in Online Task Scheduling

Michael J. Neely MJNEELY@USC.EDU
Department of Electrical Engineering
University of Southern California
Los Angeles, CA, 90089-2565, USA

Editor: Shie Mannor

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 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.

Keywords: stochastic processes, dynamic control, opportunistic scheduling, cloud computing

1. 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 $k \in \{0, 1, 2, \ldots\}$ shall be called *frame* k. Fix m as a positive integer. At the start of each frame $k \in \{0, 1, 2, \ldots\}$ 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

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

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 $k \in \{0, 1, 2, \ldots\}$ 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 infinite horizon reward per unit time is

$$\lim_{K\to\infty} \frac{\sum_{k=0}^{K-1} R[k]}{\sum_{k=0}^{K-1} T[k]}$$

temporarily assuming the limit exists. Let θ^* denote the optimal average reward per unit time. The goal is to develop an algorithm for making decisions over frames that yields an average reward per unit time that converges to θ^* as quickly as possible. Long term optimality is defined by all possible algorithms, including algorithms that know the probability distribution $F_S(s)$. However, convergence time to optimality is considered for algorithms that have no prior 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.

1.1 Discussion

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, (Gallager, 1996)):

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

To avoid divide-by-zero issues, it is assumed there is a known constant $T_{min} > 0$ such that $\mathbb{E}\left[T[0]\right] \geq T_{min}$, regardless of the decision that is made for task 0.[1]

It can be shown that the optimal reward per unit time can be achieved over the class of stationary and randomized algorithms (see (Neely, 2010)). In principle, one could perform an *offline computation* to find the best stationary and randomized algorithm. For example, suppose there is a finite set Ω_S of task types, and let $\pi(s)$ denote the probability mass function for task types:

$$\pi(s) = P[S[k] = s] \quad \forall s \in \Omega_S$$

Suppose for each $s \in \Omega_S$ the set of decision options $\mathcal{D}(s)$ is finite. A stationary and randomized algorithm observes the task type $S[k]$ for each new task k and then randomly

1. This assumption that T_{min} exists holds trivially in the special case when there is a positive lower bound on the smallest possible frame size, such as when all frames are at least one unit of time.

chooses a decision $(T[k], R[t]) \in \mathcal{D}(S[k])$ according to conditional probabilities

$$p((t,r)|s) = P[(T[k], R[k]) = (t,r)|S[k] = s]$$

Offline computation to maximize the ratio of expectations in the right-hand-side of (1) chooses conditional probabilities $p((t,r)|s)$ to solve

$$\text{Maximize:} \quad \frac{\sum_{s\in\Omega_S}\sum_{(t,r)\in\mathcal{D}(s)}\pi(s)p((t,r)|s)r}{\sum_{s\in\Omega_S}\sum_{(t,r)\in\mathcal{D}(s)}\pi(s)p((t,r)|s)t} \tag{2}$$

$$\text{Subject to:} \quad p((t,r)|s) \geq 0 \quad \forall s \in \Omega_S, \forall (t,r) \in \mathcal{D}(s) \tag{3}$$

$$\sum_{(t,r)\in\mathcal{D}(s)} p((t,r)|s) = 1 \quad \forall s \in \Omega_S \tag{4}$$

This is a linear fractional program and the optimal objective value in (2) is θ^*. It is not practical to solve (2)-(4) because: (i) We do not know the probabilities $\pi(s)$; (ii) The set Ω_S of task types can be very large, possibly containing more elements than there are atoms in the universe, and so the problem can be intractable even if probabilities $\pi(s)$ were somehow known for all $s \in \Omega_S$.

The proposed algorithm of this paper is not a stationary and randomized algorithm, 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. We show that it quickly converges to the same optimal reward per unit time, as specified in (2)-(4), with a decision complexity for each task k that is independent of the size of the set Ω_S (this set can even be infinite).

1.2 Comparison to problems with unknown states

Our problem formulation is an *opportunistic scheduling problem* where a task type $S[k]$ is observed for each task k and the decision options and rewards are fully known based on $S[k]$. This problem is important because the task type information $S[k]$ can be used to inform the decision of how to execute task k. This problem structure is different from online convex optimization problems and multi-arm bandit problems. In those problems, the decision set is the same for all steps k, but the reward vectors or reward functions are random.

For example, while the decision set $\mathcal{D}(S[k])$ for our problem depends on the observed state $S[k]$, one can imagine a *different problem formulation* with a decision set $\mathcal{D}$ that is always the same and with a random state $S[k]$ that is unknown. In this case, we can define $d[k] \in \mathcal{D}$ as the particular decision choice on step k, and the decision $d[k]$ would produce a random vector $(T[k], R[k])$ with a distribution that depends on $d[k]$. This changes the problem structure to a multi-arm bandit problem where the "arms" are the decisions $d \in \mathcal{D}$. The optimal reward per unit time for this new problem is typically decreased because, fundamentally, the same optimality point cannot be achieved when the system state $S[k]$ is unknown. The new optimality point no longer depends on the full probability distribution for $S[k]$ as needed in (2)-(4). Rather, the new optimality point depends only on the mean vectors $(t(d), r(d))$ associated with each arm $d \in \mathcal{D}$, where

$$(t(d), r(d)) = \mathbb{E}\left[(T[k], R[k])|d[k] = d\right]$$

If the mean vectors $(t(d), r(d))$ were fully known, we would always choose the arm $d \in \mathcal{D}$ that maximizes $r(d)/t(d)$. Further, estimation of $(t(d), r(d))$ could be done easily by *exploring* each arm $d \in \mathcal{D}$ according to various bandit techniques such as the resource-constrained techniques in (Badanidiyuru et al., 2018)(Agrawal and Devanur, 2014)(Xia et al., 2015).

The opportunistic scheduling problem of our paper has a different structure and cannot be solved by bandit techniques. Even though the state $S[t]$ is fully known, it is not clear how to utilize this extra information to achieve the (correspondingly larger) optimal reward per unit time. One might guess that choosing the option $(T[k], R[k]) \in \mathcal{D}(S[k])$ to greedily maximize $R[k]/T[k]$ on every task k is optimal. This is not true. A simple counter-example is provided in Section 2. Generally, solving (2)-(4) is not trivial. The state probabilities $\pi(s)$ must somehow be incorporated into the solution without knowledge of these probabilities. When there are an overwhelmingly large number of these probabilities, it is not even possible to estimate them.

Opportunistic scheduling problems are also fundamentally different from deterministic optimization and online convex optimization problems. For example, in (Neely, 2019), an opportunistic scheduling problem for wireless networks with fixed time slots is shown to have an optimal convergence time of $\Theta(\log(t)/t)$ for any smooth and concave utility function, regardless of whether or not the function is strongly concave. This is in stark contrast to deterministic convex optimization and online convex optimization for which strong convexity/concavity is known to significantly improve performance, and for which convergence tradeoffs are different (see (Nesterov, 2004) for bounds on certain types of deterministic convex minimization problems, and (Zinkevich, 2003) (Hazan et al., Dec. 2007) (Hazan and Kale, 2014) for online convex optimization tradeoffs).

The current paper is different from the opportunistic scheduling work in (Neely, 2019) because, as we show, the variable frame lengths create different convergence properties. It turns out that strong convexity shall arise as an important feature here, but in a different context.

1.3 Relation to finite horizon problems

This paper defines optimality in terms of the infinite horizon limit in (1). The linear fractional program in (2)-(4) can be viewed as searching over the set of all expectation vectors $(\mathbb{E}[T], \mathbb{E}[R])$ that can be achieved to find an optimal $(\mathbb{E}[T^*], \mathbb{E}[R^*])$ that maximizes the ratio $\mathbb{E}[R^*]/\mathbb{E}[T^*]$. Define this maximum ratio as

$$\theta^* = \frac{\mathbb{E}[R^*]}{\mathbb{E}[T^*]}$$

Intuitively, algorithms that are designed to achieve an infinite horizon limit of θ^* will also do a good job of maximizing the total accumulated reward over a fixed but long time horizon t_{total}. An alternative (and more complex) problem formulation might fix a finite

time horizon t_{total} and then seek to maximize total expected reward over this time:

$$\begin{aligned} \text{Maximize:} \quad & \mathbb{E}\left[\sum_{k=0}^{K} R[k]\right] \\ \text{Subject to:} \quad & \sum_{k=0}^{K} T[k] \leq t_{total} \end{aligned}$$

where K is a random number of tasks that can be completed up to the time limit, and the $(T[k], R[k])$ values for each task k are chosen from the set $\mathcal{D}(S[k])$. This can be viewed as an opportunistic scheduling version of a stochastic knapsack problem. This problem is more complex than the linear fractional program (2)-(4) because it would require an additional state variable that represents the remaining time until the deadline. However, when t_{total} is large, the problem can be closely approximated by the following problem of choosing a real number $m > 0$ and choosing an expectation vector $(\mathbb{E}[T], \mathbb{E}[R])$ to solve

$$\text{Maximize:} \quad m\mathbb{E}[R] \tag{5}$$

$$\text{Subject to:} \quad m\mathbb{E}[T] \leq t_{total} \tag{6}$$

Intuitively, problem (5)-(6) relates to using a stationary and randomized policy that yields i.i.d. vectors $\{(T[k], R[k])\}_{k=0}^{\infty}$ with expectations $(\mathbb{E}[T], \mathbb{E}[R])$. The value m represents a real-valued relaxation of an integer number of tasks that can be performed until the *expected* task size exceeds the time limit t_{total}. This type of approximation was used for the different context of bandit problems in (Badanidiyuru et al., 2018). When t_{total} is large, the approximation error can be shown to be small by bounding the "overshoot" associated with performing one more task before time expires (see stopping time and renewal theorems in (Gut, 2009)(Asmussen, 2003)).

It is easy to see that the approximate problem (5)-(6) is exactly solved by using expectations $(\mathbb{E}[T^*], \mathbb{E}[R^*])$ that achieve the maximum ratio θ^*, and by using $m^* = t_{total}/\mathbb{E}[T^*]$. Indeed, this particular solution satisfies the constraint (6) with equality and achieves an objective value of

$$m^*\mathbb{E}[R^*] = \theta^* t_{total}$$

No other values for $(m, \mathbb{E}[T], \mathbb{E}[R])$ that satisfy the constraints can have an objective value higher than $\theta^* t_{total}$ because

$$\begin{aligned} m\mathbb{E}[R] &= \left(\frac{\mathbb{E}[R]}{\mathbb{E}[T]}\right) m\mathbb{E}[T] \\ &\overset{(a)}{\leq} \theta^* m\mathbb{E}[T] \\ &\overset{(b)}{\leq} \theta^* t_{total} \end{aligned}$$

where (a) holds by definition of θ^* as the maximum ratio of expectations; (b) holds by (6). Overall, this discussion emphasizes that the infinite horizon problem considered in this paper can also be used as an efficient method to treat fixed but large time horizons.

1.4 Prior work

Optimization of renewal systems is related to linear fractional programming (see, for example, (Schaible, 1983)(Boyd and Vandenberghe, 2004)). An offline method for optimizing Markov renewal systems via linear fractional programming is in (Fox, 1966). Online methods for renewal optimization are developed in (Neely, 2013)(Neely, 2010), which treat systems with additional time-average constraints. The work (Neely, 2013) 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 (Neely, 2013) requires knowledge of the probability distribution for $S[k]$. An approximate implementation is also given in (Neely, 2013) that uses a bisection procedure that does not require knowledge of the probability distribution. This method is applied to treat data center scheduling in (Wei and Neely, 2017), asynchronous renewal timelines in (Wei and Neely, 2018), power-aware computing in (Neely, 2012), and has connections to the delay-optimal queueing work in (Li and Neely, 2014). An alternative Robbins-Monro technique is used in (Neely, 2010) 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 (Robbins and Monro, 1951). The work (Robbins and Monro, 1951) treats a problem of finding a root θ 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

$$M(x) = \mathbb{E}\left[Y[k]|X[k] = x\right] \quad \forall x \in \mathbb{R} \tag{7}$$

The estimated root is then updated via the iteration:

$$X[k+1] = X[k] - \eta[k]Y[k] \tag{8}$$

where $\{\eta[k]\}_{k=0}^{\infty}$ is a sequence of positive stepsizes. Under certain assumptions, the work (Robbins and Monro, 1951) shows that $X[k]$ converges in probability to the root θ. The technique of (Robbins and Monro, 1951) 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 (Nemirovski and Yudin, 1983)(Polyak and Juditsky, 1992)(Nemirovski et al., 2009)(Kushner and Yin, 2003)(Borkar, 2008). Modifications of Robbins-Monro type iterations are considered in (Toulis et al., 2020)(Toulis et al., 2014), improvements for binary data are in (Joseph, 2004), and applications to Bayesian inference is explored in (Mandt et al., 2017). 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 (Nemirovski and Yudin, 1983) show these convergence rates cannot be improved. However, an example in (Nemirovski et al., 2009) 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 (Nemirovski et al., 2009).

The current paper uses an iteration similar to (8). 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

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

which is different from minimizing the single expectation that defines $M(x)$ in (7). Furthermore, the expectations $\mathbb{E}\left[R[k]\right]$ and $\mathbb{E}\left[T[k]\right]$ 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 type methods is made in (Neely, 2013) where it is shown that the optimal time-average reward θ^* 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 θ^* using an *auxiliary variable* $\theta[k]$ on each frame k. However, it is not enough to simply find a value θ that is close to θ^*. That is because our goal is to obtain an online algorithm with a time-average (starting from frame 0) that converges to θ^* 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 (Nemirovski and Yudin, 1983).

The focus on convergence time in this paper is conceptually similar to analysis of regret in multi-armed bandit systems, see, for example, scalar-based bandit problems in (Bubeck and Cesa-Bianchi, 2012)(P. Auer et al., 1995)(Auer, 2002), vector-based bandits in (Badanidiyuru et al., 2018)(Agrawal and Devanur, 2014)(Xia et al., 2015), and a recent renewal-based bandit formulation in (Cayci et al., 2019). Such problems have an "exploration versus exploitation" structure that is different from the structure of the current paper. In (Cayci et al., 2019), 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.

1.5 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 θ^* 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 θ^* is faster than the rate at which the online averages converge to θ^*. Specifically, we show:

- Under the general system structure, the proposed algorithm has an optimality gap that decays like $O(1/\sqrt{k})$.
- Under a special "strongly concave" 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 average 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 θ^* 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 a 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 (Hazan and Kale, 2014) and a recent mapping technique in (Neely, 2020a).
- 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 ϵ of optimal. This result is similar in spirit to a known square-root converse result for pseudo-regret in multi-arm bandit problems in (Bubeck and Cesa-Bianchi, 2012). However, the square root law does not arise for the same reason as in (Bubeck and Cesa-Bianchi, 2012). Indeed, the two problems are structurally different and use different proofs.

2. Examples

2.1 Cloud computing with two choices

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:

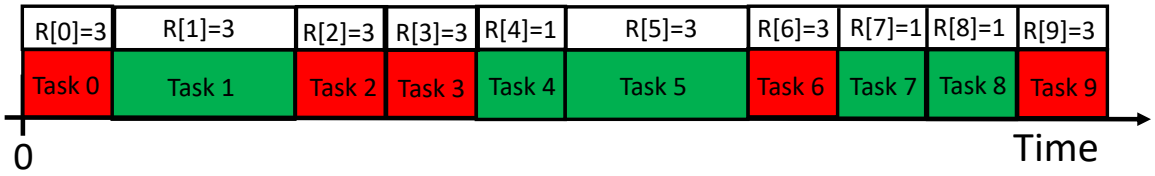

Figure 1: A sequence of back-to-back tasks of different types: Red tasks ($S[k] = 0$) have $(T[k], R[k]) = (1, 3)$. Green tasks ($S[k] = 1$) have $(T[k], R[k]) \in \{(1, 1), (2, 3)\}$.

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 6 shows that no causal algorithm that does not have a-priori knowledge of p can optimize this system faster than a square root law.

2.2 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 $f : \mathbb{R}^2 \to \mathbb{R}$ 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 $T[k] \in [T_{min}, T_{max}]$.

2.3 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 λ). Let $(T_j[k], R_j[k])$ represent the time and profit characteristics for each project $j \in \{1, \ldots, N[k]\}$. The joint distribution for

these random parameters is arbitrary and unknown. At the start of frame k, Alice chooses which single project $j \in \{1, \ldots, N[k]\}$ 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:[2]

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

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?

3. Formulation

As described in the introduction, the task type sequence $\{S[k]\}_{k=0}^{\infty}$ is i.i.d. over frames. Every frame $k \in \{0, 1, 2, \ldots\}$ 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 $S[k]$.

3.1 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*.[3] The probability law can be different for each frame and can depend on observations from previous frames.

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 $L : \Omega_S \to [0, \infty)$ and $U : \Omega_S \to [0, \infty)$ that satisfy

$$\inf_{(t,r)\in\mathcal{D}(s)} \{t\} \geq L(s) \quad \forall s \in \Omega_S \tag{9}$$

$$\sup_{(t,r)\in\mathcal{D}(s)} \{t^2 + r^2\} \leq U(s) \quad \forall s \in \Omega_S \tag{10}$$

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

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

$$\mathbb{E}\left[L(S[0])\right] > 0 \tag{11}$$
$$\mathbb{E}\left[U(S[0])\right] < \infty \tag{12}$$

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 (9)-(12) it follows that regardless of the decisions we have

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

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

$$T_{min} > 0 \tag{13}$$
$$T_{min} \leq \mathbb{E}\left[T[k]\right] \leq T_{max} \tag{14}$$
$$R_{min} \leq \mathbb{E}\left[R[k]\right] \leq R_{max} \tag{15}$$
$$\mathbb{E}\left[T[k]^2\right] \leq C_1 \tag{16}$$
$$\mathbb{E}\left[R[k]^2\right] \leq C_2 \tag{17}$$

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

3.2 Optimization goal

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

$$\text{Maximize:} \quad \liminf_{K\to\infty} \frac{\sum_{k=0}^{K-1} \mathbb{E}\left[R[k]\right]}{\sum_{k=0}^{K-1} \mathbb{E}\left[T[k]\right]} \tag{18}$$
$$\text{Subject to:} \quad (T[k], R[k]) \in \mathcal{D}(S[k]) \quad \forall k \in \{0, 1, 2, \ldots\} \tag{19}$$

The objective in (18) considers a ratio of expectations, similar to the renewal optimization problems considered in (Neely, 2013)(Neely, 2010). Define θ^* as the supremum value of the objective function (18) over all algorithms that satisfy the constraints (19). The value θ^* 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 θ^* 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 $F_S(s)$).

3.3 Characterizing optimality

This subsection summarizes key facts from (Neely, 2013). Let $\mathcal{A} \subseteq [T_{min}, T_{max}] \times [R_{min}, R_{max}]$ be the set of all 1-shot expectations $(\mathbb{E}\left[T[0]\right], \mathbb{E}\left[R[0]\right])$ 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

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

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

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

Let $\overline{\mathcal{A}}$ denote the closure of set $\mathcal{A}$. Recall that all points $(t, r) \in \overline{\mathcal{A}}$ have $t \geq T_{min} > 0$. It can be shown that the supremum objective θ^* for the problem (18)-(19) is achievable by a particular (possibly non-stationary) algorithm for making decisions over frames and satisfies:

$$\theta^* = \sup_{(t,r)\in\overline{\mathcal{A}}}\left\{\frac{r}{t}\right\} \tag{20}$$

The supremum on the right-hand-side of (20) is achievable because r/t is a continuous function over the compact set $\overline{\mathcal{A}}$. In particular, there exists a (possibly non-unique) optimal point $(t^*, r^*) \in \overline{\mathcal{A}}$ such that

$$\theta^* = \frac{r^*}{t^*} \tag{21}$$

Lemma 1 *(From (Neely, 2013)) Let θ^* be the optimal ratio in* (20)*. Then*

$$\sup_{(t,r)\in\overline{\mathcal{A}}}\{r - \theta^* t\} = 0 \tag{22}$$

Proof A proof is provided for completeness. Fix $(t, r) \in \overline{\mathcal{A}}$ and recall that $t > 0$. Then

$$r - \theta^* t = \left(-\theta^* + \frac{r}{t}\right) t \leq \left(-\theta^* + \sup_{(t',r')\in\overline{\mathcal{A}}}\left\{\frac{r'}{t'}\right\}\right) t = 0$$

where the final equality uses the definition of θ^* in (20). This holds for all $(t, r) \in \overline{\mathcal{A}}$ and so

$$\sup_{(t,r)\in\overline{\mathcal{A}}}\{r - \theta^* t\} \leq 0$$

To prove the reverse inequality, since $(t^*, r^*) \in \overline{A}$ we have

$$\sup_{(t,r)\in\overline{\mathcal{A}}}\{r - \theta^* t\} \geq r^* - \theta^* t^* = 0$$

where the final equality holds by (21). ■

4. 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 θ_{min} and θ_{max} be finite and known values that satisfy:

$$\theta_{min} \leq \theta^* \leq \theta_{max} \tag{23}$$

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

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

where we recall that the R_{min} and R_{max} values can be negative. Values of θ_{min} and θ_{max} that more tightly bracket θ^* 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 θ^* as follows: Initialize $\theta[0] \in [\theta_{min}, \theta_{max}]$ as an arbitrary deterministic value. On each frame $k \in \{0, 1, 2, \ldots\}$ do:

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

$$\text{Maximize:} \quad R[k] - \theta[k]T[k] \tag{24}$$
$$\text{Subject to:} \quad (T[k], R[k]) \in \mathcal{D}(S[k]) \tag{25}$$

  breaking ties arbitrarily.[4] There is at least one maximizer because the set $\mathcal{D}(S[k])$ is compact.

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

$$\theta[k+1] = [\theta[k] + \eta[k](R[k] - \theta[k]T[k])]_{\theta_{min}}^{\theta_{max}} \tag{26}$$

  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 (26) is inspired by the classic Robbins-Monro iteration in (8). The main complexity of the algorithm is the selection of $(T[k], R[k])$ in (24)-(25) to maximize a linear function over the 2-dimensional decision set $\mathcal{D}(S[k])$. Recall that the decision set is only assumed to be compact (it can be finite or infinite, convex or nonconvex). If $\mathcal{D}(S[k])$ contains 100 points, then we simply consider all 100 points and choose the best one. Simulations in Section 8 treat cases where the decision set contains a finite number of points, and also cases where it contains an infinite curve of points.

The objective in the maximization step (24) is similar to the $r - \theta^* t$ expression from Lemma 1, with the exception that the vector $(t, r) \in \overline{\mathcal{A}}$, which represents an achievable expectation $(\mathbb{E}[T[k]], \mathbb{E}[R[k]])$, is replaced by the actual realization $(T[k], R[k])$ that is chosen

4. 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.

for task k. However, since $\theta[k]$ is independent of $S[k]$, choosing the realization $(T[k], R[k]) \in \mathcal{D}(S[k])$ to solve (24)-(25) results in a conditional expectation $\mathbb{E}\left[(T[k], R[k])|\theta[k]\right]$ that maximizes $r - \theta[k]t$ over all $(t, r) \in \overline{\mathcal{A}}$. Another difference is that the value of θ^* is unknown, so we use the estimated value $\theta[k]$ in (24). The estimate $\theta[k]$ is updated according to (26), which increases or decreases $\theta[k]$ depending on whether $R[k] - \theta[k]T[k]$ is positive or negative. This is because, as shown in (Neely, 2013), the value of $\sup_{(t,r)\in\overline{\mathcal{A}}}\{r - \theta t\}$ yields a sign that indicates whether $\theta > \theta^*$, $\theta = \theta^*$, or $\theta < \theta^*$.

4.1 Technical lemma

The following lemma holds because the $\theta[k]$ parameter in the algorithm (24)-(26) depends only on $\{S[0], \ldots, S[k-1]\}$ and does not depend on the task type $S[k]$.

Lemma 2 *Consider algorithm* (24)-(26) *with any initial constant* $\theta[0] \in [\theta_{min}, \theta_{max}]$ *and any positive stepsizes* $\{\eta[k]\}_{k=0}^{\infty}$. *For each* $k \in \{0, 1, 2, \ldots\}$ *and for any versions of* $\mathbb{E}\left[T[k]|\theta[k]\right]$ *and* $\mathbb{E}\left[R[k]|\theta[k]\right]$, *the following holds with probability 1:*[5]

$$\left(\mathbb{E}\left[T[k]|\theta[k]\right], \mathbb{E}\left[R[k]|\theta[k]\right]\right) \in \overline{\mathcal{A}} \tag{27}$$

That is, (27) *holds for* almost all realizations of $\theta[k]$. *Further, there are versions of* $\mathbb{E}\left[T[k]|\theta[k]\right]$ *and* $\mathbb{E}\left[R[k]|\theta[k]\right]$ *under which* (27) *holds* surely for all realizations *of* $\theta[k]$.

Proof This is a special case of a more general result developed in (Neely, 2020c). ■

Intuitively, the above lemma holds because $\{S[k]\}_{k=0}^{\infty}$ is i.i.d. over frames. Since $\theta[k]$ depends only on $\{S[0], \ldots, S[k-1]\}$, 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 $\mathbb{E}\left[T[k]|\theta[k]\right]$ and $\mathbb{E}\left[R[k]|\theta[k]\right]$ shall represent any particular versions of the conditional expectations.

4.2 Analysis of the $(T[k], R[k])$ decision

The following lemmas use θ^* as the optimal ratio in (20) and assume (t^*, r^*) is a vector in $\overline{\mathcal{A}}$ that satisfies $\theta^* = r^*/t^*$. Lemma 3 connects to Lemma 1 by bounding the deviation from 0 for $\mathbb{E}\left[R[k] - \theta[k]T[k]|\theta[k]\right]$ when $\theta[k] \neq \theta^*$.

Lemma 3 *Consider the algorithm* (24)-(26) *with any initial constant* $\theta[0] \in [\theta_{min}, \theta_{max}]$ *and any positive stepsizes* $\{\eta[k]\}_{k=0}^{\infty}$. *For each* $k \in \{0, 1, 2, \ldots\}$ *we have for* almost all realizations *of* $\theta[k]$*:*

$$\mathbb{E}\left[R[k] - \theta[k]T[k]|\theta[k]\right] \geq t^*(\theta^* - \theta[k]) \tag{28}$$

and

$$\mathbb{E}\left[R[k] - \theta^*T[k]|\theta[k]\right] \geq (\theta[k] - \theta^*)\mathbb{E}\left[(T[k] - t^*)|\theta[k]\right] \tag{29}$$

That is, the probability that random variable $\theta[k]$ *does not satisfy both* (28) *and* (29) *is* 0.

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

Proof Fix $k \in \{0, 1, 2, \ldots\}$. 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):[6]

$$\mathbb{E}\left[R[k] - \theta[k]T[k]|\theta[k]\right] \geq \mathbb{E}\left[R^*[k]|\theta[k]\right] - \theta[k]\mathbb{E}\left[T^*[k]|\theta[k]\right] \tag{30}$$

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 $(\mathbb{E}\left[T[0]\right], \mathbb{E}\left[R[0]\right]) = (t, r)$. 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

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

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

$$\mathbb{E}\left[T^*[k]|\theta[k]\right] = t$$
$$\mathbb{E}\left[R^*[k]|\theta[k]\right] = r$$

Substituting these identities into (30) gives (with probability 1):

$$\mathbb{E}\left[R[k] - \theta[k]T[k]|\theta[k]\right] \geq r - \theta[k]t \tag{31}$$

Let (t^*, r^*) be a vector in $\overline{\mathcal{A}}$ 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 (31) 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 $i \in \{1, 2, 3, \ldots\}$ and so:[7]

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

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

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

where the final equality uses $\theta^* = r^*/t^*$. This proves (28). Adding $(\theta[k] - \theta^*)\mathbb{E}\left[T[k]|\theta[k]\right]$ to both sides proves (29). ■

6. 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 $\mathbb{E}\left[X - Y|Z\right] \geq 0$ with probability 1.
7. Recall that if $\{F_i\}_{i=1}^{\infty}$ is an infinite sequence of events that satisfy $P[F_i] = 1$ for all $i \in \{1, 2, 3, \ldots\}$ then $P[\cap_{i=1}^{\infty} F_i] = 1$.

Lemma 4 *Consider the algorithm (24)-(26) 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:*

$$\left|\theta^* - \frac{\sum_{k=0}^{K-1}\mathbb{E}\left[R[k]\right]}{\sum_{k=0}^{K-1}\mathbb{E}\left[T[k]\right]}\right| \leq \frac{1}{KT_{min}}\sum_{k=0}^{K-1}\mathbb{E}\left[|\theta[k]-\theta^*|\cdot|T[k]-t^*|\right] \tag{32}$$

where T_{min} satisfies (13)-(14).

Proof Fix $k \in \{0, 1, 2, \ldots\}$. We have $(\mathbb{E}\left[T[k]\right], \mathbb{E}\left[R[k]\right]) \in \overline{\mathcal{A}}$ and so

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

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

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

Rearranging terms yields

$$\theta^* \geq \frac{\sum_{k=0}^{K-1}\mathbb{E}\left[R[k]\right]}{\sum_{k=0}^{K-1}\mathbb{E}\left[T[k]\right]} \tag{33}$$

On the other hand, we can take expectations of (29) and use the law of iterated expectations to obtain

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

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

$$\begin{aligned}\frac{\sum_{k=0}^{K-1}\mathbb{E}\left[R[k]\right]}{\sum_{k=0}^{K-1}\mathbb{E}\left[T[k]\right]} &\geq \theta^* + \frac{1}{\sum_{k=0}^{K-1}\mathbb{E}\left[T[k]\right]}\sum_{k=0}^{K-1}\mathbb{E}\left[(\theta[k]-\theta^*)(T[k]-t^*)\right] \\ &\geq \theta^* - \frac{1}{KT_{min}}\sum_{k=0}^{K-1}\mathbb{E}\left[|\theta[k]-\theta^*|\cdot|T[k]-t^*|\right] \end{aligned} \tag{34}$$

Combining (33) and (34) proves the lemma. ■

To make the right-hand-side of (32) close to 0, it is desirable to make the $\theta[k]$ parameter close to θ^*.

4.3 Analysis of the update rule

Consider the algorithm (24)-(26) with any initial constant $\theta[0] \in [\theta_{min}, \theta_{max}]$ and any positive stepsizes. Define b as a constant that satisfies

$$\frac{1}{2}\mathbb{E}\left[(R[k] - \theta[k]T[k])^2\right] \leq b \quad \forall k \in \{0, 1, 2, \ldots\} \tag{35}$$

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 (13)-(17). The following lemma is similar in spirit to the analysis of Robbins-Monro iterations for different systems in (Robbins and Monro, 1951)(Nemirovski et al., 2009). Here, the size of T_{min} determines a weight on an important negative quadratic term.

Lemma 5 *Under algorithm* (24)-(26) *with any initial constant* $\theta[0] \in [\theta_{min}, \theta_{max}]$ *and any (deterministic) positive stepsizes* $\{\eta[k]\}_{k=0}^{\infty}$*, we have for all frames* $k \in \{0, 1, 2, \ldots\}$

$$\frac{1}{2}\mathbb{E}\left[(\theta[k+1]-\theta^*)^2\right] \leq \left(\frac{1}{2} - T_{min}\eta[k]\right)\mathbb{E}\left[(\theta[k]-\theta^*)^2\right] + \eta[k]^2 b \tag{36}$$

where the constant b *satisfies* (35)*; the constant* T_{min} *satisfies* (13)-(14)*; the constant* θ^* *is defined by* (20)*.*

Proof Fix $k \in \{0, 1, 2, \ldots\}$. 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 (26) we have

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

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

$$\frac{1}{2}\mathbb{E}\left[(\theta[k+1]-\theta^*)^2\right] \leq \frac{1}{2}\mathbb{E}\left[(\theta[k]-\theta^*)^2\right] + \eta[k]^2 b + \eta[k]\mathbb{E}\left[(\theta[k]-\theta^*)(R[k]-\theta[k]T[k])\right] \tag{37}$$

To complete the proof it suffices to provide the following bound on the final term of (37):

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

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

$$(\theta[k]-\theta^*)\mathbb{E}\left[R[k]-\theta[k]T[k]|\theta[k]\right] \leq -T_{min}(\theta[k]-\theta^*)^2 \tag{38}$$

To show (38) we consider two cases.

- Case 1: Suppose $\theta[k] < \theta^*$. By (28) we have for almost all $\theta[k]$ for which $\theta[k] - \theta^* < 0$:

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

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

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

  where the final inequality holds because $t^* \geq T_{min}$. Thus, (38) holds in this Case 1.

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

$$\begin{aligned}
&(\theta[k]-\theta^*)\mathbb{E}\left[R[k]-\theta[k]T[k]|\theta[k]\right] \\
&= (\theta[k]-\theta^*)\mathbb{E}\left[R[k]-\theta^*T[k]|\theta[k]\right] - (\theta[k]-\theta^*)^2\mathbb{E}\left[T[k]|\theta[k]\right] \\
&\leq (\theta[k]-\theta^*)\mathbb{E}\left[R[k]-\theta^*T[k]|\theta[k]\right] - T_{min}(\theta[k]-\theta^*)^2 \qquad (39)
\end{aligned}$$

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

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

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

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

Substituting this into (39) shows that (38) holds in this Case 2.

■

4.4 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 θ^*. However, the following lemma uses a decreasing stepsize to achieve faster convergence. 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 (Nemirovski et al., 2009): Work in (Nemirovski et al., 2009) 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* (24)-(26) *with any initial constant* $\theta[0] \in [\theta_{min}, \theta_{max}]$ *and with stepsizes*

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

we have

$$\mathbb{E}\left[(\theta[k] - \theta^*)^2\right] \leq \frac{2b}{kT_{min}^2} \quad \forall k \in \{1, 2, 3, \ldots\} \tag{40}$$

where the constant b *satisfies* (35)*; the constant* T_{min} *satisfies* (13)-(14)*; the constant* θ^* *is defined by* (20)*.*

Proof The proof uses an induction argument inspired by the analysis in (Bubeck, 2015) for a different class of problems (namely, Frank-Wolfe methods for deterministic convex optimization). For simplicity define

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

It suffices to show $z_k \leq \frac{b}{kT_{min}^2}$ for all $k \in \{1, 2, 3, \ldots\}$. From (36) we have

$$z_{k+1} \leq (1 - 2T_{min}\eta[k])z_k + \eta[k]^2 b \quad \forall k \in \{0, 1, 2, \ldots\} \tag{41}$$

Applying the above inequalty 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 $k \in \{1, 2, 3, \ldots\}$ (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 (41):

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

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 $k \geq 0$.

■

4.5 Online performance theorem

Theorem 1 *(General performance) Under the algorithm* (24)-(26) *with initial constant* $\theta[0] \in [\theta_{min}, \theta_{max}]$ *and stepsizes* $\eta[k] = \frac{1}{(k+2)T_{min}}$ *for* $k \in \{0, 1, 2, \ldots\}$ *we have*

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

where C_1 *satisfies* (16) *and* b *satisfies* (35)*. Hence, deviation from the optimal ratio* θ^* *decays like* $O(1/\sqrt{K})$*.*

Proof From (32) we have for all integers $K \geq 2$:

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

where (a) follows by the Cauchy-Schwarz inequality; (b) holds by (16) and $(T[k] - t^*)^2 \leq T[k]^2 + (t^*)^2 \leq T[k]^2 + C_1$; (c) holds by (40); (d) holds because $\sum_{k=1}^{K-1} \frac{1}{\sqrt{k}} \leq 1 + \int_1^{K-1} \frac{1}{\sqrt{t}} dt$. ■

4.6 Strongly concave curvature

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

$$r \leq r^* + \theta^*(t - t^*) - \frac{c}{2}(t - t^*)^2 \quad \forall (t, r) \in \overline{\mathcal{A}} \tag{42}$$

Theorem 2 *(Performance with strongly concave curvature) Assume* $\overline{\mathcal{A}}$ *has the strongly concave curvature property (with parameter* c*) specified in* (42)*. Under the algorithm* (24)-(26) *with initial constant* $\theta[0] \in [\theta_{min}, \theta_{max}]$ *and stepsize* $\eta[k] = \frac{1}{(k+2)T_{min}}$ *for* $k \in$

$\{0, 1, 2, \ldots\}$ *we have*

$$\left|\theta^* - \frac{\sum_{k=0}^{K-1} \mathbb{E}\left[R[k]\right]}{\sum_{k=0}^{K-1} \mathbb{E}\left[T[k]\right]}\right| \leq \frac{2(\theta[0]-\theta^*)^2 + \frac{4b}{T_{min}^2}(1+\log(K-1))}{KcT_{min}} \quad \forall K \in \{2,3,4,\ldots\} \quad (43)$$

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

Proof Fix $k \in \{0, 1, 2, \ldots\}$. For almost all $\theta[k]$ we have (by Lemma 2):

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

So by (42) we have with probability 1:

$$\mathbb{E}\left[R[k]|\theta[k]\right] \leq r^* + \theta^*(\mathbb{E}\left[T[k]|\theta[k]\right] - t^*) - \frac{c}{2}(\mathbb{E}\left[T[k]|\theta[k]\right] - t^*)^2 \quad (44)$$

From (29) we have for all real numbers $\beta > 0$ (with prob 1):

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

where (a) uses the fact $ab \geq -\frac{a^2+b^2}{2}$ for all real numbers a, b; (b) uses (44); (c) uses $r^* - \theta^* t^* = 0$. Choose $\beta > 0$ so that $\beta^2/c = 1/2$. Then

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

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

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

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

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

where (a) holds by (40); (b) holds because $\sum_{k=1}^{K-1} 1/k \leq 1 + \int_1^{K-1}(1/t)dt$. Rearranging terms gives

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

where the final inequality holds because $\mathbb{E}\left[T[k]\right] \geq T_{min}$ for all k. On the other hand (33) implies that the ratio on the left-hand-side is less than or equal to θ^*. This proves the result. ■

4.7 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 strong concavity 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.

5. 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 (Hazan and Kale, 2014). 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 (Neely, 2020a).

5.1 System

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

$$P[S[k]=1] = q \quad ; \quad P[S[k]=0] = 1-q \tag{45}$$

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

$$\mathcal{D}(S[k]) = \begin{cases} (1,1) & \text{, if } S[k]=0 \\ \{(x, 2-(2-x)^2) \in \mathbb{R}^2 : x \in [1,2]\} & \text{, if } S[k]=1 \end{cases} \tag{46}$$

The decision structure (46) 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*.[8]

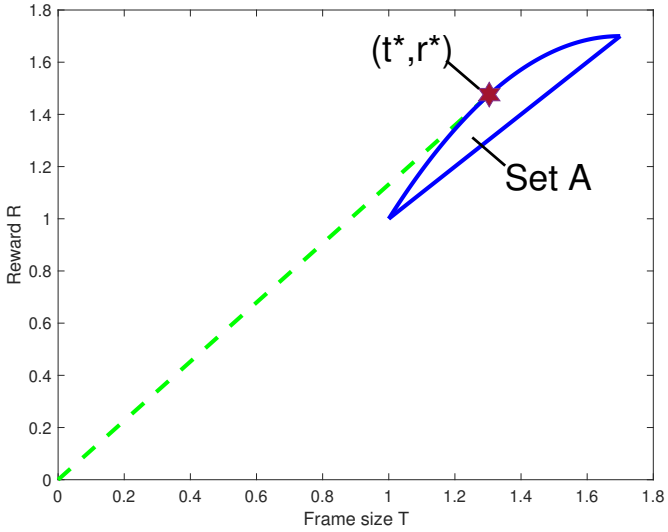

Figure 2: The set $\mathcal{A}$ with a strongly concave upper boundary for the example of Section 5.1 with $q = 0.7$.

Lemma 7 *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:*

$$t^* = \sqrt{1+q} \tag{47}$$

$$r^* = \frac{2(q+1)\left(-1+\sqrt{1+q}\right)}{q} \tag{48}$$

$$\theta^* = \frac{r^*}{t^*} = 2 - \frac{2}{q}\left(-1+\sqrt{1+q}\right) \tag{49}$$

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

$$r \leq \theta^* t - \frac{1}{q}(t - t^*)^2 \quad \forall (t, r) \in \mathcal{A} \tag{50}$$

8. The strongly concave decision curve gives rise to a set $\overline{\mathcal{A}}$ 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 $\overline{\mathcal{A}}$ has a strongly concave upper boundary.

Proof This holds by basic analysis of the set $\mathcal{A}$ in Fig. 2 and details are in the technical report (Neely, 2020c). ■

5.2 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 $k \in \{0, 1, 2, \ldots\}$, define $H[k]$ as the system history up to but not including frame k:

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

where $H[0]$ is formally defined as 0 (since there is no history before frame 0). On each frame k, a general causal and measurable algorithm makes decisions as a deterministic function of $(H[k], U)$ where U is uniformly distributed in $[0, 1)$ and is independent of $\{S[k]\}_{k=0}^{\infty}$. The variable U represents an external source of randomness that can inform randomized decisions. Let $\{f_k\}_{k=0}^{\infty}$ be a sequence of Borel-measurable functions of the following form:

$$f_0 : [0, 1) \to \mathcal{D}(1)$$
$$f_k : \{0, 1\}^k \times [0, 1) \to \mathcal{D}(1) \quad \forall k \in \{1, 2, 3, \ldots\}$$

These functions $\{f_k\}_{k=0}^{\infty}$ establish control decisions on each frame $k \in \{0, 1, 2, \ldots\}$:

$$(T[k], R[k]) = \begin{cases} (1, 1) & , \text{ if } S[k] = 0 \\ f_k(H[k], U) & , \text{ if } S[k] = 1 \end{cases} \tag{51}$$

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 $k \in \{0, 1, 2, \ldots\}$ 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.[9] 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 8 *Fix $q \in [1/4, 3/4]$ and consider any causal and measurable decisions $(T[k], R[k]) \in \mathcal{D}(S[k])$ for $k \in \{0, 1, 2, \ldots\}$. For each $k \in \{0, 1, 2, \ldots\}$ we have:*

$$\mathbb{E}[(T[k], R[k])|H[k]] \in \mathcal{A} \tag{52}$$

Furthermore,

$$\mathbb{E}[R[k]|H[k]] \leq \theta^* \mathbb{E}[T[k]|H[k]] - \frac{1}{q}(\mathbb{E}[T[k]|H[k]] - t^*)^2 \tag{53}$$

where $t^* = \sqrt{1+q}$ and θ^* satisfies (49).

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

Proof The fact (52) is similar to Lemma 2 and can be proven by noting $\overline{\mathcal{A}} = \mathcal{A}$ and $S[k]$ is independent of $(U, H[k])$ (see (Neely, 2020c) for details). Substituting (52) into (50) proves (53). ■

5.3 The Bernoulli estimation theorem from (Hazan and Kale, 2014)

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

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

and for each $k \in \{1,2,3,\ldots\}$ the function $g_k(s_0,\ldots,s_{k-1})$ maps a binary-valued sequence $(s_0,\ldots,s_{k-1})$ 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:

$$G[k] = g_k(S[0], S[1], \ldots, S[k-1]) \quad \forall k \in \{1,2,3,\ldots\} \tag{54}$$

The following theorem from (Hazan and Kale, 2014) provides a lower bound on the mean square error for any sequence of estimation functions.[10]

Theorem 3 *(Bernoulli Estimation (Hazan and Kale, 2014)) 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*

$$\sum_{k=1}^{K} \mathbb{E}\left[(q - G[k])^2\right] \geq \Omega(\log(K)) \tag{55}$$

where the random variables $G[k]$ are defined in (54) for each $k \in \{1,2,\ldots\}$.

5.4 Completing the converse

Lemma 9 *Define $\phi : [1/4, 3/4] \to \mathbb{R}$ by*

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

Then

a) ϕ is continuous and strictly decreasing with minimim and maximum values

$$\phi_{min} = \phi(3/4) \approx 0.430501$$
$$\phi_{max} = \phi(1/4) \approx 0.472136$$

10. The result in (Hazan and Kale, 2014) also applies to randomized estimation functions. This paper shall only need the result for deterministic estimation functions.

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 ϕ *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

$$|\phi(q) - \phi(u)| \geq \beta |q - u| \quad \forall u, q \in [1/4, 3/4] \tag{56}$$

Proof Parts (a) and (c) follow by basic analysis on the function ϕ. 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 with $u < x < q$ such that

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

and so

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

■

Theorem 4 *Let* $\{f_k\}_{k=0}^{\infty}$ *be any sequence of decision functions that define a causal and measurable algorithm according to* (51). *There is a parameter* $q \in [1/4, 3/4]$ *such that using these decision functions with an i.i.d. Bernoulli-q process* $\{S[k]\}_{k=0}^{\infty}$ *and with an independent uniform random variable* U*, the resulting (causal and measurable) decisions* $(T[k], R[k]) \in \mathcal{D}(S[k])$ *satisfy*

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

where θ^* *is the optimal ratio in* (49). *In particular, no algorithm can have error that decays faster than* $\Omega(\log(K)/K)$.

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

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

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

$$G[k] = g_k(H[k]) \quad \forall k \in \{1, 2, 3, \ldots\} \tag{57}$$

and so

$$G[k] = \phi^{-1}\left([z_k(H[k]) - 1]_{\phi_{min}}^{\phi_{max}}\right) \tag{58}$$

Observe that for $k \in \{1, 2, 3, \ldots\}$, 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 $H[k] = (S[0], \ldots, S[k-1])$ 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 ϕ^{-1}. Hence, if we can express 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 $k \in \{1, 2, 3, \ldots\}$. We have for each $h_k \in \{0, 1\}^k$:

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

where (a) holds by (53); (b) holds by (51); (c) holds by definition of ϕ and because $t^* = \sqrt{1+q}$ (recall (47)); (d) holds because the distance between $z_k(H[k]) - 1$ 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 (58); (f) holds by (56); (g) holds because $q \geq 1/4$. Taking expectations of both sides with respect to the random $H[k]$ gives

$$\mathbb{E}\left[R[k]\right] \leq \theta^*\mathbb{E}\left[T[k]\right] - \frac{\beta^2}{4}\mathbb{E}\left[(G[k] - q)^2\right] \quad \forall k \in \{1, 2, 3, \ldots\}$$

Summing gives

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

where the final inequality holds by Lemma 1. Dividing by $\sum_{k=0}^{K-1}\mathbb{E}\left[T[k]\right]$ and noting that this is less than or equal to $2K$ (since $T[k] \leq 2$ always, see (46)) yields

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

where the final inequality holds by application of the Bernoulli estimation bound (55). ■

6. 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 2 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 ϵ 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:

$$\theta^*_{(1/2-\epsilon)} = 2 + 2\epsilon \tag{59}$$

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

$$\theta^*_{(1/2+\epsilon)} = \frac{6}{3+2\epsilon} \tag{60}$$

The structure of considering only two possible hypotheses that are difficult to discern is similar in spirit to the converse result of (Bubeck and Cesa-Bianchi, 2012) 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 $\beta[k]$, which is 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 $S[0], \ldots, S[k-1]$:

$$\beta[0] = \hat{\beta}_0(U) \tag{61}$$

$$\beta[k] = \hat{\beta}_k(U, S[0], \ldots, S[k-1]) \quad \forall k \in \{1, 2, 3, \ldots\} \tag{62}$$

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

$$\begin{aligned} \mathbb{E}\left[R[k]|\beta[k]\right] &= (1-p)(3) + p(1+2\beta[k]) \\ \mathbb{E}\left[T[k]|\beta[k]\right] &= (1-p)(1) + p(1+\beta[k]) \\ \mathbb{E}\left[R[k] - \theta^* T[k]|\beta[k]\right] &= (1-p)(3) + p(1+2\beta[k]) - \theta^*(1-p+p(1+\beta[k])) \end{aligned} \tag{63}$$

Theorem 5 *Fix δ such that $0 < \delta \leq 1/256$ and let $\epsilon = 64\delta$. Consider any general causal algorithm of the type (61)-(62). Fix $K \in \{1, \ldots, \lfloor \frac{3}{2^{19}\delta^2} \rfloor\}$. If for the case $p = \frac{1}{2} - \epsilon$ the algorithm satisfies*

$$\frac{\sum_{k=0}^{K} \mathbb{E}\left[R[k]\right]}{\sum_{k=0}^{K} \mathbb{E}\left[T[k]\right]} > \theta^*_{(1/2-\epsilon)} - \delta$$

then for the case $p = \frac{1}{2} + \epsilon$ the algorithm must satisfy

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

The proof is developed in the next subsections by fixing $\delta \in (0, 1/256]$, defining $\epsilon = 64\delta$, and fixing a particular causal algorithm of the type described in this section.

6.1 Case $H_{(1/2-\epsilon)}$

Fix a particular causal algorithm of the type (61)-(62). Substituting (59) into (63) and doing the basic but tedious arithmetic gives

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

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

$$\mathbb{E}_{(1/2-\epsilon)}\left[R[k] - \theta^*_{(1/2-\epsilon)} T[k]\right] \leq -\frac{\epsilon}{4} P[\beta[k] > 1/2] \tag{64}$$

6.2 Case $H_{(1/2+\epsilon)}$

Substituting (60) into (63)

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

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

$$\mathbb{E}_{(1/2+\epsilon)}\left[R[k]-\theta^*_{(1/2+\epsilon)}T[k]\right] \leq -\frac{\epsilon}{4}P[\beta[k]\leq 1/2] \tag{65}$$

6.3 Bernoulli estimation for mean absolute error

For each $k\in\{1,2,3,\ldots\}$ 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

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

The sequence of functions $\{A_k(\cdot)\}_{k=1}^{\infty}$ 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 $S[0],\ldots,S[k-1]$.

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 (Neely, 2020b) and is a modified version of a lemma for mean squared error developed in (Hazan and Kale, 2014):

Lemma 10 *Suppose p and q are in the interval $[1/4,3/4]$. For any positive integer k and any Borel-measurable function $\hat{A}_k(\cdot)$ of the structure defined above, we have*

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

Proof See Lemma 6 in (Neely, 2020b). ■

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

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

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

$$\mathbb{E}_{(1/2-\epsilon)}\left[|A[k]_{1/2-\epsilon} - (1/2-\epsilon)|\right] = 2\epsilon P[\beta[k] > 1/2] \tag{66}$$

$$\mathbb{E}_{(1/2+\epsilon)}\left[|A[k]_{1/2+\epsilon} - (1/2+\epsilon)|\right] = 2\epsilon P[\beta[k] \leq 1/2] \tag{67}$$

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

$$\begin{aligned}
&\mathbb{E}_{(1/2-\epsilon)}\left[R[k] - \theta^*_{(1/2-\epsilon)}T[k]\right] + \mathbb{E}_{(1/2+\epsilon)}\left[R[k] - \theta^*_{(1/2+\epsilon)}T[k]\right] \\
&\overset{(a)}{\leq} \frac{-\epsilon}{4}\left[P[\beta[k] > 1/2] + P[\beta[k] \leq 1/2]\right] \\
&\overset{(b)}{=} \frac{-1}{8}\left(\mathbb{E}_{(1/2-\epsilon)}\left[|A[k]_{1/2-\epsilon} - (1/2-\epsilon)|\right] + \mathbb{E}_{(1/2+\epsilon)}\left[|A[k]_{1/2+\epsilon} - (1/2+\epsilon)|\right]\right) \\
&\overset{(c)}{\leq} \frac{-\epsilon}{16}
\end{aligned} \tag{68}$$

where (a) holds by (64)-(65); (b) holds by (66)-(67); (c) holds by application of Lemma 10 for $p = 1/2-\epsilon$ and $q = 1/2+\epsilon$ (indeed the conditions of Lemma 10 hold because $\epsilon \in (0, 1/4]$, p and q are in $[1/4, 3/4]$, and $|p-q| = 2\epsilon$). The condition $2\epsilon \leq \frac{\sqrt{3}}{4\sqrt{2k}}$ is equivalent to

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

Fix $K \in \{1,2,\ldots,\lfloor \frac{3}{128\epsilon^2} \rfloor\}$ (this corresponds to K in the interval specified by Theorem 5 with $\epsilon = 64\delta$). Summing (68) over $k \in \{0,\ldots,K\}$ gives

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

where (a) neglects the nonpositive $k = 0$ term (recall Lemma 1). It follows that either

$$\sum_{k=0}^{K} \mathbb{E}_{(1/2-\epsilon)}\left[R[k] - \theta^*_{(1/2-\epsilon)}T[k]\right] \leq -\frac{K\epsilon}{32} \tag{69}$$

or

$$\sum_{k=0}^{K} \mathbb{E}_{(1/2+\epsilon)}\left[R[k] - \theta^*_{(1/2+\epsilon)}T[k]\right] \leq -\frac{K\epsilon}{32} \tag{70}$$

Assume (69) holds: Then assuming $p = 1/2 - \epsilon$ and rearranging terms in (69) yields (using notation $\mathbb{E}[\cdot]$ and θ^* instead of $\mathbb{E}_{(1/2-\epsilon)}[\cdot]$ and $\theta^*_{(1/2-\epsilon)}$ for simplicity):

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

where (a) holds because this example has $\mathbb{E}[T[k]] \leq 2$ for all k; (b) holds because $\epsilon = 64\delta$. Therefore, if (69) holds, then running the algorithm in the case when the true parameter is $p = 1/2 - \epsilon$ means that

$$\frac{\sum_{k=0}^{K}\mathbb{E}[R[k]]}{\sum_{k=0}^{K}\mathbb{E}[T[k]]} \leq \theta^* - \delta \tag{71}$$

On the other hand, if (69) fails then (70) must hold and by the same argument it follows that (71) is true for the case $p = 1/2 + \epsilon$. So for any particular causal algorithm, (71) must hold for either the case $p = 1/2 - \epsilon$ or the case $p = 1/2 + \epsilon$. This proves Theorem 5.

6.4 Discussion

Recall that Theorem 1 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 8). 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.

7. Probability 1 convergence

This section considers the general system described in Section 3.1 (not necessarily having the strong concavity property used in Section 4.6). Recall that $\{S[k]\}_{k=0}^{\infty}$ is i.i.d. with some distribution $F_S(s) = P[S[k] \leq s]$ for $s \in \mathbb{R}^m$. We show:

- (Theorem 6) No algorithm can have a sample path time-average that exceeds θ^*.
- (Theorem 7) The algorithm (24)-(26) with the stepsize rule $\eta[k] = \frac{1}{(k+2)T_{min}}$ ensures $\theta[k] \to \theta^*$ with probability 1.
- (Theorem 8) The algorithm (24)-(26) with stepsize rule $\eta[k] = \frac{1}{(k+2)T_{min}}$ ensures

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

7.1 Preliminary lemma

Lemma 11 *Assume $\{X_i\}_{i=1}^{\infty}$ is a sequence of random variables that satisfies*

$$\sum_{i=1}^{\infty} \frac{\mathbb{E}\left[X_i^2\right]}{i^2} < \infty \tag{72}$$

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

$$a \leq \mathbb{E}\left[X_i | X_0, \ldots, X_{i-1}\right] \leq b \quad \forall i \in \{1, 2, 3, \ldots\} \tag{73}$$

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

$$a \leq \liminf_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} X_i \leq \limsup_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} X_i \leq b \tag{74}$$

Proof See (Neely, 2020c). ■

7.2 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 $\{S[k]\}_{k=0}^{\infty}$. 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. A general *causal and measurable* algorithm makes decisions on each frame $k \in \{0, 1, 2, \ldots\}$ by

$$(T[k], R[k]) = f_k(U, S[0], S[1], \ldots, S[k]) \tag{75}$$

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

$$f_k(u, s_0, \ldots, s_{k-1}, s_k) \in \mathcal{D}(s_k) \quad \forall s_k \in \Omega_S \tag{76}$$

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

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

where θ^* is from (20)*. Furthermore*

$$T_{min} \leq \limsup_{K\to\infty} \frac{1}{K}\sum_{k=0}^{K-1} T[k] \leq T_{max} \quad \textit{(with prob 1)} \tag{77}$$

Proof See (Neely, 2020c). ■

Theorem 7 *Under algorithm* (24)-(26) *with stepsize* $\eta[k] = \frac{1}{(k+2)T_{min}}$ *for all* $k \in \{0, 1, 2, \ldots\}$ *we have*

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

where θ^* *is from* (20).

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

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

where the final inequality holds by (40). 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

$$\lim_{i\to\infty} \theta[i^2] = \theta^* \quad \text{(with prob 1)} \tag{78}$$

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 (78) yields

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

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

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

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

$$\sum_{k=1}^{\infty} P[G_k > \epsilon] < \infty \tag{79}$$

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

$$\begin{aligned} P[G_k > \epsilon] &\leq \frac{\mathbb{E}\left[G_k\right]}{\epsilon} \\ &= \frac{1}{\epsilon}\mathbb{E}\left[\max_{i\in\{k^2,\ldots,(k+1)^2-1\}} \left\{(\theta[i] - \theta[k^2])^2\right\}\right] \end{aligned} \tag{80}$$

Let V be a finite constant that satisfies

$$\mathbb{E}\left[(R[i]-\theta[i]T[i])^2\right] \leq V \quad \forall i \in \{0,1,2,\ldots\} \tag{81}$$

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 (26) we have for all $j \in \{0,1,2,\ldots\}$

$$\begin{aligned} |\theta[j+1]-\theta[j]| &\overset{(a)}{=} \left|[\theta[j]+\eta[j](R[j]-\theta[j]T[j])]_{\theta_{min}}^{\theta_{max}} - [\theta[j]]_{\theta_{min}}^{\theta_{max}}\right| \\ &\overset{(b)}{\leq} |\eta[j](R[j]-\theta[j]T[j])| \end{aligned} \tag{82}$$

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 $i \in \{k^2, \ldots, (k+1)^2-1\}$:

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

where (a) holds by (82); (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 $i \in \{k^2, \ldots, (k+1)^2-1\}$ gives

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

Taking an expectation and using the Cauchy-Schwarz inequality and (81) gives

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

Substituting this into (80) gives

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

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

Theorem 8 *Under algorithm* (24)-(26) *with stepsize* $\eta[k] = \frac{1}{(k+2)T_{min}}$ *for all* $k \in \{0, 1, 2, \ldots\}$ *we have*

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

where θ^* *is from* (20).

Proof We have by the rule (24)-(25):

$$R[k] - \theta[k]T[k] \geq R^*[k] - \theta[k]T^*[k] \quad \forall k \in \{0, 1, 2, \ldots\} \tag{83}$$

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:

$$\lim_{K\to\infty} \frac{1}{K} \sum_{k=0}^{K-1} (T^*[k], R^*[k]) = (t, r) \quad \text{(with prob 1)} \tag{84}$$

Rearranging terms in (83) gives the following for all $k \in \{0, 1, 2, \ldots\}$:

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

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

$$\begin{aligned}
&\frac{1}{K} \sum_{k=0}^{K-1} R[k] - \theta^* \frac{1}{K} \sum_{k=0}^{K-1} T[k] \\
&\geq \frac{1}{K} \sum_{k=0}^{K-1} R^*[k] - \theta^* \frac{1}{K} \sum_{k=0}^{K-1} T^*[k] - \frac{1}{K} \sum_{k=0}^{K-1} (\theta[k] - \theta^*)(T[k] - T^*[k]) \\
&\geq \frac{1}{K} \sum_{k=0}^{K-1} R^*[k] - \theta^* \frac{1}{K} \sum_{k=0}^{K-1} T^*[k] - \frac{1}{K} \sum_{k=0}^{K-1} |\theta[k] - \theta^*|(T[k] + T^*[k])
\end{aligned} \tag{85}$$

Claim: We have with probability 1:

$$\limsup_{K\to\infty} \frac{1}{K} \sum_{k=0}^{K-1} |\theta[k] - \theta^*|(T[k] + T^*[k]) = 0 \tag{86}$$

We postpone the proof of (86). Taking a lim inf of (85) and substituting (86) and (84) gives, with probability 1:

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

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 $(t^*, r^*) \in \overline{\mathcal{A}}$ and using the fact that $\theta^* = r^*/t^*$ gives

$$\liminf_{K\to\infty} \left[\frac{1}{K}\sum_{k=0}^{K-1} R[k] - \theta^* \frac{1}{K}\sum_{k=0}^{K-1} T[k]\right] \geq 0 \quad \text{(with prob 1)} \tag{87}$$

This, together with (77), proves that with probability 1:

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

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

It remains to prove (86) of the Claim. We know $\theta[k] \to \theta^*$ with probability 1. Fix $\epsilon > 0$. With probability 1 we know $|\theta[k] - \theta^*| \leq \epsilon$ for all sufficiently large k and so

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

Observe that both $T[k]$ and $T^*[k]$ are from causal and measurable algorithms and so they both satisfy (77). Thus

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

This holds for all $\epsilon > 0$ and so (86) follows. ■

8. Simulation

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

8.1 System 1: Selecting one of multiple projects

Consider the project selection problem of Section 2.3. 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

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

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 $j \in \{1, \ldots, i\}$ are generated

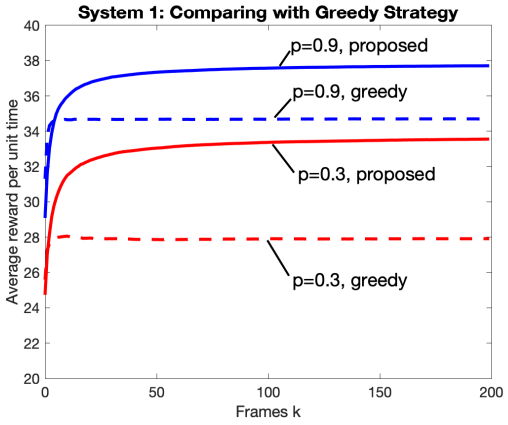

Figure 3: Comparing the proposed algorithm with the greedy strategy for two different values of p. Data is averaged over 5000 independent sample paths.

independently with $T_j \sim Uniform([1, 10])$ and $R_j = A_j T_j$ where $A_j \sim Unif([0, 50])$. 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 θ^* analytically so the dashed horizontal lines in Fig. 4 are estimated values of θ^* 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 θ-empirical heuristic algorithm of (Neely, 2013) (see Section VI.b of (Neely, 2013)). The θ-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 time-average reward seen up to frame k. In (Neely, 2013) and (Neely, 2010) it is shown that *if this algorithm converges* then it converges to the optimal θ^*, but no proof of convergence and no convergence time results are known. The simulation shows 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.

8.2 System 2: A curve of choices

Now consider the system described in Section 5.1, where θ^* 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:

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

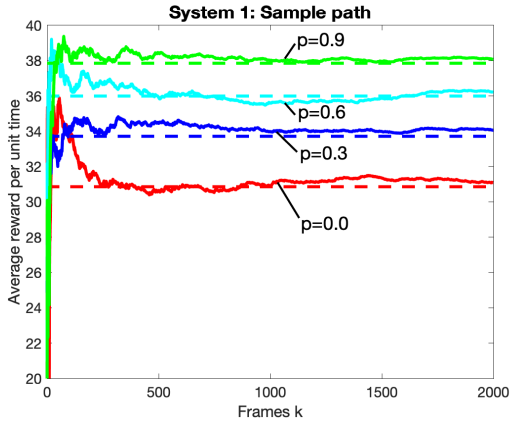

Figure 4: Simulation of a single sample path for System 2: $\sum_{i=0}^{k-1} R[i]/\sum_{i=0}^{k-1} T[i]$ versus $k \in \{0, \ldots, 2000\}$ for four different values of p. Dashed horizontal lines are obtained by averaging the final value at time 2000 over 5000 independent sample paths.

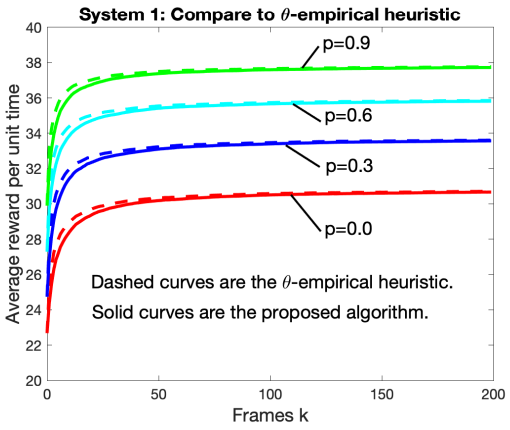

Figure 5: Comparing the proposed algorithm to the θ-empirical heuristic of (Neely, 2013). The simulation averages over 5000 independent sample paths and plots $\mathbb{E}[\sum_{i=0}^{k-1} R[i]/\sum_{i=0}^{k-1} T[i]]$ versus $k \in \{0, \ldots, 200\}$ for four different values of p.

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 θ^* values in (49). Fig. 6 shows how close one sample path time-average comes to θ^* 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 θ^*, which is consistent with the probability 1 sample path convergence proven in Section 7. Fig. 7 shows the expected performance (with expectations computed by averaging over 5000 independent sample paths). Fig. 7 plots

$\mathbb{E}\left[\sum_{i=0}^{k-1} R[i]/\sum_{i=0}^{k-1} T[i]\right]$. It was observed that plots of $\sum_{i=0}^{k-1}\mathbb{E}\left[R[i]\right]/\sum_{i=0}^{k-1}\mathbb{E}\left[T[i]\right]$ over the same number of frames looked similar (those plots are omitted for brevity).

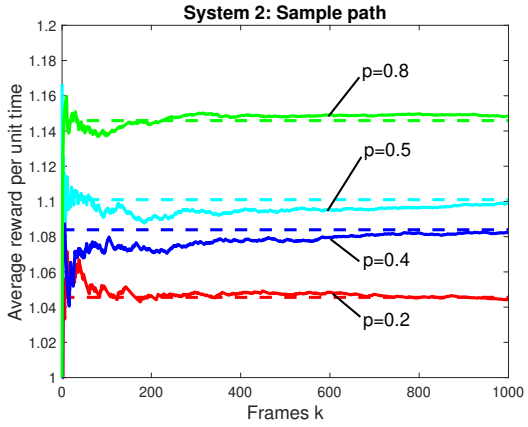

Figure 6: Simulation of a single sample path for System 2: $\sum_{i=0}^{k-1} R[i]/\sum_{i=0}^{k-1} T[i]$ versus $k \in \{0,\ldots,1000\}$ for four different Bernoulli probabilities p. Dashed horizontal lines are the optimal values.

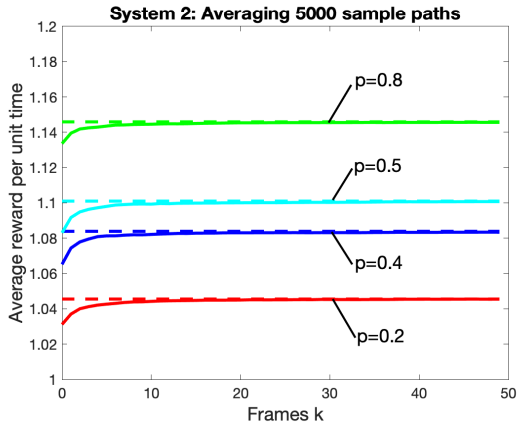

Figure 7: Simulation averaging over 5000 sample paths for System 2: $\mathbb{E}\left[\sum_{i=0}^{k-1} R[i]/\sum_{i=0}^{k-1} T[i]\right]$ versus $k \in \{0,\ldots,50\}$ for four different Bernoulli probabilities p. Dashed horizontal lines are the optimal values.

8.3 System 3: Two choices

Consider the system of Section 2, which is the same system for which a square-root converse result was proven in Section 6. 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 θ^* values computed analytically in Section 6. Fig. 9 plots smoother curves that are averaged over 5000 independent sample paths. The curves in Fig. 9 are plotted over the smaller timeline $k \in \{0, \ldots, 50\}$ 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.[11] Indeed, simulations were considered with $p = 0.5 \pm \delta$ for various small δ 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 6.4.

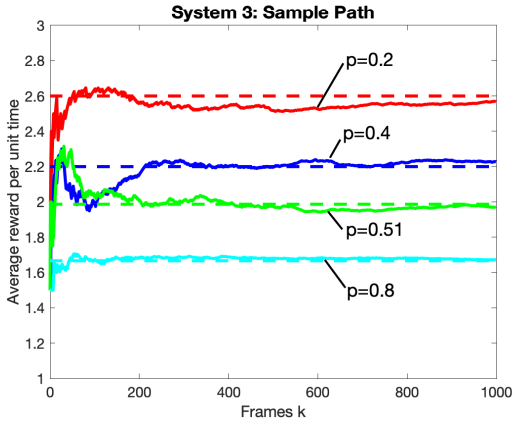

Figure 8: Simulation of a single sample path for System 3: $\sum_{i=0}^{k-1} R[i] / \sum_{i=0}^{k-1} T[i]$ versus $k \in \{0, \ldots, 1000\}$ for four different Bernoulli probabilities p. Dashed horizontal lines are the optimal values.

9. 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 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

11. 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 6.4.

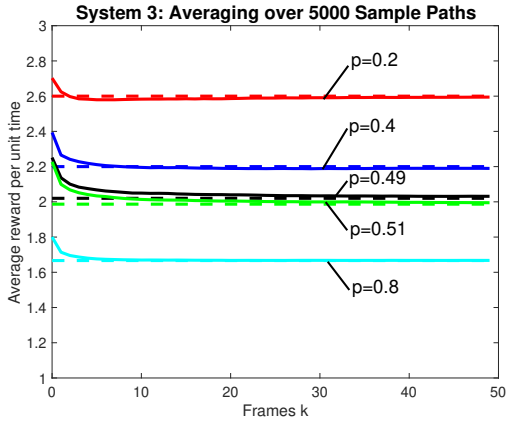

Figure 9: Simulation averaging over 5000 sample paths for System 3: $\mathbb{E}\left[\sum_{i=0}^{k-1} R[i]/\sum_{i=0}^{k-1} T[i]\right]$ versus $k \in \{0, \ldots, 50\}$ for five different Bernoulli probabilities p. Dashed horizontal lines are the optimal values.

terms of expectations achieved by the algorithm. The algorithm was also shown to have sample paths that converge to optimality with probability 1.

Acknowledgements

This work was supported by grant NSF CCF-1718477.

References

S. Agrawal and N. R. Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14, page 989–1006, New York, NY, USA, 2014. Association for Computing Machinery. doi: 10.1145/2600057.2602844.

S. Asmussen. *Applied Probability and Queues, Second Edition*. New York: Spring-Verlag, 2003.

P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

A. Badanidiyuru, R. Kleinberg, and A. Slivkins. Bandits with knapsacks. *J. ACM*, 65(3), March 2018. ISSN 0004-5411. doi: 10.1145/3164539.

V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Springer, 2008.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231——357, 2015.

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.

S. Cayci, A. Eryilmaz, and R. Srikant. Learning to control renewal processes with bandit feedback. *Proc. ACM SIGMETRICS*, July 2019.

B. Fox. Markov renewal programming by linear fractional programming. *Siam J. Appl. Math*, vol. 14, no. 6, Nov. 1966.

R. Gallager. *Discrete Stochastic Processes*. Kluwer Academic Publishers, Boston, 1996.

A. Gut. *Stopped Random Walks*. Springer-Verlag, New York Inc., 2009.

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.

E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, vol. 69, no. 2, pp. 169-192, Dec. 2007.

V. R. Joseph. Efficient Robbins-Monro procedure for binary data. *Biometrika*, 91(2):461–470, June 2004.

H. J. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003.

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.

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.

M. J. Neely. *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Morgan & Claypool, 2010.

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.

M. J. Neely. Dynamic optimization and learning for renewal systems. *IEEE Transactions on Automatic Control*, vol. 58, no. 1, pp. 32-46, Jan. 2013.

M. J. Neely. Convergence and adaptation for utility optimal opportunistic scheduling. *IEEE/ACM Transactions on Networking*, 27(3):904–917, June 2019.

M. J. Neely. A converse result on convergence time for opportunistic wireless scheduling. *Proc. IEEE INFOCOM*, 2020a.

M. J. Neely. A converse result on convergence time for opportunistic wireless scheduling. *ArXiv technical report, arXiv:2001.01031v4*, March 2020b.

M. J. Neely. Fast learning for renewal optimization in online task scheduling. *ArXiv technical report, arXiv:2007.09532v1*, July 2020c.

A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience Series in Discrete Mathematics, John Wiley, 1983.

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.

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, Boston, 2004.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

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.

H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.

S. Schaible. Fractional programming. *Zeitschrift fur Operations Research*, vol. 27, no. 1, pp. 39-54, Dec. 1983.

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.

P. Toulis, T. Horel, and E. M. Airoldi. The proximal Robbins-Monro method. *arXiv:1510.00967v4*, Feb. 2020.

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.

X. Wei and M. J. Neely. Asynchronous optimization over weakly coupled renewal systems. *Stochastic Systems*, 8(3), Sept. 2018.

Y. Xia, W. Ding, X. Zhang, N. Yu, and T. Qin. Budgeted bandit problems with continuous random costs. In *ACML*, 2015.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. *Proc. 20th International Conference on Machine Learning (ICML)*, 2003.