Dynamic Optimization and Learning for Renewal Systems

Michael J. Neely

Abstract— **We consider the problem of optimizing time averages in systems with independent and identically distributed behavior over renewal frames. This includes scheduling and task processing to maximize utility in stochastic networks with variable length scheduling modes. Every frame, a new policy is implemented that affects the frame size and that creates a vector of attributes. An algorithm is developed for choosing policies on each frame in order to maximize a concave function of the time average attribute vector, subject to additional time average constraints. The algorithm is based on Lyapunov optimization concepts and involves minimizing a "drift-plus-penalty" ratio over each frame. The algorithm can learn efficient behavior without a-priori statistical knowledge by sampling from the past. Our framework is applicable to a large class of problems, including Markov decision problems.**

I. Introduction

Consider a stochastic system that regularly experiences times when the system state is refreshed, called *renewal times*. The goal is to develop a control algorithm that maximizes the time average of a reward process associated with the system, subject to time average constraints on a collection of penalty processes. The renewal-reward theorem is a simple and elegant technique for computing time averages in such systems (see, for example, [1][2]). However, the renewal-reward theorem requires random events to be independent and identically distributed (i.i.d.) over each renewal frame. While this i.i.d. assumption may hold if a single control law is implemented repeatedly, it is often difficult to choose in advance a single control law that optimizes the system subject to the desired constraints. This paper investigates the situation where the control policies used may differ from frame to frame, and are designed to dynamically solve the problem of interest.

This renewal problem arises in many different applications. One application of interest is a *task processing network*. For example, consider a network of wireless devices that repeatedly collaborate to accomplish tasks (such as reporting sensor data to a destination, or performing distributed computation on data). Tasks are performed one after the other, and for each task we must decide what modes of operation and communication to use, possibly allowing some nodes of the network to remain idle to save power. It is then important to make decisions that maximize the time average utility associated with task processing, subject to time average power constraints at each node. Alternatively, one may want to minimize time average power, subject to constraints on utility and on the "left-over" communication rates available for data that is not associated with the task processing.

This paper develops a general framework for solving such problems. To do so, we extend the theory of Lyapunov optimization from [3]. Specifically, work in [3] considers discrete time queueing networks and develops a simple *drift-plus-penalty* rule for making optimal decisions. These decisions are made in a greedy manner every slot based only on the observed traffic and channel conditions for that slot, without requiring a-priori knowledge of the underlying probability distribution. However, the work in [3] assumes all slots have fixed length, the random network condition is observed at the beginning of each slot and does not change over the slot, and this condition is not influenced by control actions. The general renewal problem treated in the current paper is more complex because each frame may have a different length and may contain a sequence of random events. The frame length and the random event sequence may depend on the control decisions made over the course of the frame. Rather than making a single decision every slot, every frame we must specify a *policy*, being a contingency plan for making decisions over the course of the frame in reaction to the resulting system events.

This paper solves the general problem with a conceptually simple technique that chooses a policy to minimize a *drift-plus-penalty ratio* every frame. We first develop algorithms for minimizing the time average of a penalty process subject to a collection of time average constraints. We then consider maximization of a concave function of a vector of time average attributes subject to similar constraints. This utility maximization problem is challenging because of the variable frame length. We overcome this challenge with a novel transformation together with a variation of Jensen's inequality.

While this paper focuses on task processing applications, we note that our renewal framework can also handle *Markov decision problems*. Specifically, suppose the system operates according to either a continuous or discrete time Markov chain with control-dependent transition probabilities. If the chain has a recurrent state, then renewals can be defined as re-visitations to this state, and the same drift-plus-penalty ratio technique can be applied. However, the drift-plus-penalty ratio may be difficult to optimize for Markov decision problems with high dimension (see also [4]).

Prior work on learning algorithms for Markov decision problems is in [5], and related work in [6][7][8][9] considers learning for optimization of energy and delay in queueing systems. The works [5]-[9] use stochastic approximation theory and two-timescale convergence analysis. The Lagrange multiplier updates in [5]-[9] are analogous to the *virtual queue updates* we use in this paper. However, the Lyapunov optimization framework we use is different and does not require a two-

Michael J. Neely is with the Electrical Engineering department at the University of Southern California, Los Angeles, CA. This material is supported in part by one or more of the following: the DARPA IT-MANET program grant W911NF-07-0028, the NSF Career grant CCF-0747525, and continuing through participation in the Network Science Collaborative Technology Alliance sponsored by the U.S. Army Research Laboratory.

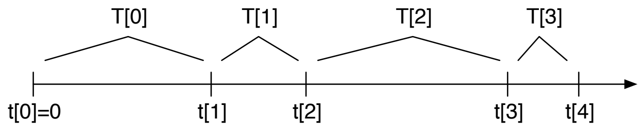

Fig. 1. A timeline illustrating renewal frames for the system.

timescale approach. It also provides more explicit bounds on convergence times and deviations from optimality, and allows a broader class of problems such as task processing problems.

The Lyapunov optimization technique that we use in this paper is based on our previous work in [3][10][11][12] that develops the drift-plus-penalty method for stochastic network optimization, including opportunistic scheduling for throughput-utility maximization [3][10][12] and average power minimization [11] (see also [13]). Alternative "fluid-based" stochastic optimization techniques for queueing networks are developed in [14][15][16][17], and dual and primal-dual algorithms for systems without queues, based on tracking a corresponding static optimization problem, are in [18][19][20]. Our current paper considers the more complex renewal problem, and leverages ideas in [4][21], where [4] considers a frame-based Lyapunov framework for Markov decision problems involving network delay, and [21] develops a ratio rule for utility optimization in wireless systems with variable length frames and time-correlated channels.

Recent work in [22] considers a task processing system where multiple wireless "reporting nodes" select data formats (e.g., "voice" or "video") in which to deliver sensed information. The work [22] also uses a renewal structure. However, it assumes a single random event occurs at the beginning of each renewal frame, and the event and frame size are not influenced by control actions. More general problems can be treated using the theory developed in the current paper.

II. Renewal System Model

Consider a system that operates over *renewal frames*. Specifically, consider the timeline of non-negative real times $t \geq 0$, and suppose this timeline is segmented into successive frames of duration $\{T[0], T[1], T[2], \ldots\}$, as shown in Fig. 1. Define $t[0] = 0$, and for each positive integer r define $t[r]$ as the *rth renewal time*:

$$t[r] \overset{\triangle}{=} \textstyle\sum_{i=0}^{r-1} T[i]$$

The interval of all times t such that $t[r] \leq t < t[r+1]$ is defined as the *rth renewal frame*, defined for each $r \in \{0, 1, 2, \ldots\}$.

At the beginning of each renewal frame r, the controller selects a *policy* $\pi[r]$ from an abstract policy space $\mathcal{P}$, and implements the policy over the duration of the frame. There may be random events that arise over the renewal frame (with distributions that are possibly dependent on the policy), and the policy specifies a contingency plan for reacting to these events. The policy incurs a vector of *penalties* $\boldsymbol{y}[r] = (y_0[r], y_1[r], \ldots, y_L[r])$ and *attributes* $\boldsymbol{x}[r] = (x_1[r], \ldots, x_M[r])$ for some integers $L \geq 0$, $M \geq 0$ (where $L = 0$ corresponds to problems without $\boldsymbol{y}[r]$ penalties, and $M = 0$ corresponds to problems without $\boldsymbol{x}[r]$ attributes). The policy may also affect the renewal frame duration $T[r]$. Formally, the values $T[r]$, $y_l[r]$, $x_m[r]$ are determined by *random functions* $\hat{T}(\cdot)$, $\hat{y}_l(\cdot)$, $\hat{x}_m(\cdot)$ of the policy $\pi[r]$:

$$\begin{aligned} T[r] &\overset{\triangle}{=} \hat{T}(\pi[r]) && (1) \\ y_l[r] &\overset{\triangle}{=} \hat{y}_l(\pi[r]) \ \forall l \in \{0, 1, \ldots, L\} && (2) \\ x_m[r] &\overset{\triangle}{=} \hat{x}_m(\pi[r]) \ \forall m \in \{1, \ldots, M\} && (3) \end{aligned}$$

We assume the values of $[\hat{T}(\pi[r]), (\hat{y}_l(\pi[r])), (\hat{x}_m(\pi[r]))]$ for frame r are conditionally independent of events in previous frames given the particular policy $\pi = \pi[r]$, and are identically distributed over all frames that use the same policy π.

Consider now a particular control algorithm that chooses policies $\pi[r] \in \mathcal{P}$ every frame r according to some well defined (possibly probabilistic) rule, and define the following frame-average expectations, defined for integers $R > 0$:

$$\overline{T}[R] \overset{\triangle}{=} \frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E}\left\{T[r]\right\} \ , \ \overline{y}_l[R] \overset{\triangle}{=} \frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E}\left\{y_l[r]\right\} \qquad (4)$$

where we recall that $T[r]$, $y_l[r]$, $x_m[r]$ depend on the policy $\pi[r]$ by (1)-(3). Define $\overline{x}_m[R]$ similarly, and define the infinite horizon frame-average expectations $\overline{T}$, $\overline{y}_l$, $\overline{x}_m$ by:

$$(\overline{T}, \overline{y}_l, \overline{x}_m) = \lim_{R \to \infty} (\overline{T}[R], \overline{y}_l[R], \overline{x}_m[R])$$

where we temporarily assume the limits are well defined.

A. Optimization Objective

The first type of problem we consider uses only penalties $\boldsymbol{y}[r]$: We must choose a policy $\pi[r] \in \mathcal{P}$ every frame r to minimize the ratio $\overline{y}_0/\overline{T}$ subject to constraints on $\overline{y}_l/\overline{T}$:

$$\begin{aligned} &\text{Minimize:} && \overline{y}_0/\overline{T} && (5) \\ &\text{Subject to:} && \overline{y}_l/\overline{T} \leq c_l \ \forall l \in \{1, \ldots, L\} && (6) \\ & && \pi[r] \in \mathcal{P} \ \forall r \in \{0, 1, 2, \ldots\} && (7) \end{aligned}$$

where c_l for $l \in \{1, \ldots, L\}$ are a given collection of real-valued (possibly negative) constants.

The motivation for looking at the ratio $\overline{y}_l/\overline{T}$ is that it defines the *time average penalty associated with the* $y_l[r]$ *process*. To see this, suppose the following limits converge to constants y_l^{av} and T^{av} with probability 1:

$$\lim_{R \to \infty} \frac{1}{R} \sum_{r=0}^{R-1} y_l[r] = y_l^{av} \ , \ \lim_{R \to \infty} \frac{1}{R} \sum_{r=0}^{R-1} T[r] = T^{av} \ \ (w.p.1)$$

Under very mild conditions, the existence of the limits y_l^{av} and T^{av} implies the frame-average expectations also have well defined limits, with $\overline{y}_l = y_l^{av}$ and $\overline{T} = T^{av}$. This holds, for example, whenever $y_l[r]$ and $T[r]$ are deterministically bounded by finite constants, or when more general conditions hold that allow the Lebesgue dominated convergence theorem to be applied [23]. Then the time average penalty per unit time associated with $y_l[r]$ (sampled only at renewal times for simplicity) satisfies with probability 1:

$$\lim_{R \to \infty} \frac{\sum_{r=0}^{R-1} y_l[r]}{\sum_{r=0}^{R-1} T[r]} = \lim_{R \to \infty} \frac{\frac{1}{R}\sum_{r=0}^{R-1} y_l[r]}{\frac{1}{R}\sum_{r=0}^{R-1} T[r]} = \frac{\overline{y}_l}{\overline{T}}$$

Therefore, the value $\overline{y}_l/\overline{T}$ indeed represents the limiting penalty per unit time associated with the process $y_l[r]$.

The problem (5)-(7) seeks only to minimize a time average subject to time average constraints. The second problem we consider, more general than the first, seeks to maximize a *concave and entrywise non-decreasing function* $\phi(\boldsymbol{\gamma})$ of the time average attribute vector ratio $\overline{\boldsymbol{x}}/\overline{T}$, where $\overline{\boldsymbol{x}} = (\overline{x}_1, \ldots, \overline{x}_M)$:

$$\begin{aligned} \text{Maximize:} \quad & \phi(\overline{\boldsymbol{x}}/\overline{T}) && (8) \\ \text{Subject to:} \quad & \overline{y}_l/\overline{T} \leq c_l \;\; \forall l \in \{1, \ldots, L\} && (9) \\ & \pi[r] \in \mathcal{P} \;\; \forall r \in \{0, 1, 2, \ldots\} && (10) \end{aligned}$$

where $\phi(\boldsymbol{\gamma})$ is a given concave and entrywise non-decreasing utility function defined over $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_M) \in \mathbb{R}^M$.

B. Boundedness Assumptions

We assume $x_m[r]$, $T[r]$, and $y_0[r]$ have bounded conditional expectations, regardless of the policy. That is, there are finite constants x_m^{min}, x_m^{max}, T^{min}, T^{max}, y_0^{min}, y_0^{max} such that for all $\pi[r] \in \mathcal{P}$ and all $m \in \{1, \ldots, M\}$ we have:

$$\begin{gathered} y_0^{min} \leq \mathbb{E}\left\{\hat{y}_0(\pi[r])|\pi[r]\right\} \leq y_0^{max} \\ 0 < T^{min} \leq \mathbb{E}\left\{\hat{T}(\pi[r])|\pi[r]\right\} \leq T^{max} \\ x_m^{min} \leq \mathbb{E}\left\{\hat{x}_m(\pi[r])|\pi[r]\right\} \leq x_m^{max} \end{gathered}$$

Define γ_m^{min} and γ_m^{max} by:

$$\begin{aligned} \gamma_m^{min} & \triangleq \min[x_m^{min}/T^{min}, x_m^{min}/T^{max}] \\ \gamma_m^{max} & \triangleq \max[x_m^{max}/T^{max}, x_m^{max}/T^{max}] \end{aligned}$$

Define the hyper-rectangle $\mathcal{R}$ by:

$$\mathcal{R} \triangleq \{\boldsymbol{\gamma} \in \mathbb{R}^M | \gamma_m^{min} \leq \gamma_m \leq \gamma_m^{max} \;\; \forall m \in \{1, \ldots, M\}\} \quad (11)$$

Then for any algorithm that chooses policies $\pi[r] \in \mathcal{P}$ for all frames r, it is not difficult to show that $\overline{x}_m[R]/\overline{T}[R] \in \mathcal{R}$ for all $R \in \{1, 2, 3, \ldots\}$, where $\overline{T}[R]$, $\overline{x}_m[R]$, $\overline{T}[R]$ are frame average expectations over the first R frames, as defined by (4).

Finally, we assume the conditional second moments of $T[r]$, $x_m[r]$, and $y_l[r]$ (for $l \neq 0$) are finite, regardless of the policy. That is, there is a finite constant σ_1 such that for all $\pi[r] \in \mathcal{P}$:

$$\begin{aligned} \mathbb{E}\left\{\hat{T}(\pi[r])^2|\pi[r]\right\} & \leq \sigma_1 \\ \mathbb{E}\left\{\hat{y}_l(\pi[r])^2|\pi[r]\right\} & \leq \sigma_1 \;\; \forall l \in \{1, \ldots, L\} \\ \mathbb{E}\left\{\hat{x}_m(\pi[r])^2|\pi[r]\right\} & \leq \sigma_1 \;\; \forall m \in \{1, \ldots, M\} \end{aligned}$$

C. Optimality of i.i.d. Algorithms

We now state the problem (5)-(7) more precisely, using $\limsup$s which do not require existence of a well defined limit:

$$\begin{aligned} \text{Minimize:} \quad & \limsup_{R\rightarrow\infty} \tfrac{\overline{y}_0[R]}{\overline{T}[R]} && (12) \\ \text{Subject to:} \quad & \limsup_{R\rightarrow\infty} \tfrac{\overline{y}_l[R]}{\overline{T}[R]} \leq c_l \;\; \forall l \in \{1, \ldots, L\} && (13) \\ & \pi[r] \in \mathcal{P} \;\; \forall r \in \{0, 1, 2, \ldots\} && (14) \end{aligned}$$

Assume that the constraints (13)-(14) are feasible, and define $ratio^{opt}$ as the infimum ratio in (12) over all algorithms that satisfy these constraints.

Define an *i.i.d. algorithm* as one that, at the beginning of each new frame $r \in \{0, 1, 2, \ldots\}$, chooses a policy $\pi[r]$ by independently and probabilistically selecting $\pi \in \mathcal{P}$ according to some distribution that is the same for all frames r. Let $\pi^*[r]$ represent such an i.i.d. algorithm. Then the random variables $\{\hat{T}(\pi^*[r])\}_{r=0}^{\infty}$ are independent and identically distributed (i.i.d.) over frames, as are $\{\hat{y}_l(\pi^*[r])\}_{r=0}^{\infty}$. Thus, by the law of large numbers, these have well defined time averages $\overline{T}^*$ and $\overline{y}_l^*$ with probability 1, where the averages are equal to the expectations over one frame.

Lemma 1: (Optimality over i.i.d. algorithms) If the constraints (13)-(14) are feasible, then for any $\delta > 0$, there exists an i.i.d. algorithm $\pi^*[r]$ that satisfies:

$$\mathbb{E}\left\{\hat{y}_0(\pi^*[r])\right\} \leq \mathbb{E}\left\{\hat{T}(\pi^*[r])\right\}(ratio^{opt} + \delta) \quad (15)$$

$$\mathbb{E}\left\{\hat{y}_l(\pi^*[r])\right\} \leq \mathbb{E}\left\{\hat{T}(\pi^*[r])\right\}(c_l + \delta) \;\; \forall l \in \{1, \ldots, L\} \quad (16)$$

Proof: The proof is similar to results in [11][13], and is omitted for brevity. □

III. Optimizing Time Averages

Here we develop an algorithm to treat the problem (5)-(7). To treat the constraints $\overline{y}_l/\overline{T} \leq c_l$, which are equivalent to the constraints $\overline{y}_l \leq c_l\overline{T}$, we define *virtual queues* $Z_l[r]$ for $l \in \{1, \ldots, L\}$, with finite initial condition and with update equation:

$$Z_l[r+1] = \max[Z_l[r] + y_l[r] - c_lT[r], 0] \forall l \in \{1, \ldots, L\} \quad (17)$$

The intuition is that if we can *stabilize* the queue $Z_l[r]$, then the time average of the "service process" $c_lT[r]$ is greater than or equal to the time average of the "arrival process" $y_l[r]$ (see also [11] for application to *virtual power queues* for meeting time average power constraints).

Let $\boldsymbol{Z}[r] = (Z_1[r], \ldots, Z_L[r])$ be the vector of virtual queues, and define the following *quadratic Lyapunov function* $L(\boldsymbol{Z}[r])$:

$$L(\boldsymbol{Z}[r]) \triangleq \tfrac{1}{2}\textstyle\sum_{l=1}^{L} Z_l[r]^2$$

The value $L(\boldsymbol{Z}[r])$ is a scalar measure of the size of the queue backlogs. The intuition is that if we can take actions that consistently push this value down, then queues can be stabilized. Define the *frame-based conditional Lyapunov drift* $\Delta(\boldsymbol{Z}[r])$ by:

$$\Delta(\boldsymbol{Z}[r]) \triangleq \mathbb{E}\left\{L(\boldsymbol{Z}[r+1]) - L(\boldsymbol{Z}[r])|\boldsymbol{Z}[r]\right\}$$

Lemma 2: Under any control decision for choosing $\pi[r] \in \mathcal{P}$, we have for all r and all possible $\boldsymbol{Z}[r]$:

$$\Delta(\boldsymbol{Z}[r]) \leq B + \sum_{l=1}^{L} Z_l[r]\mathbb{E}\left\{y_l[r] - c_lT[r]|\boldsymbol{Z}[r]\right\} \quad (18)$$

where B is a constant that satisfies for all r and all possible $\boldsymbol{Z}[r]$:

$$B \geq \frac{1}{2}\sum_{l=1}^{L} \mathbb{E}\left\{(y_l[r] - c_lT[r])^2|\boldsymbol{Z}[r]\right\} \quad (19)$$

Such a constant B exists by the boundedness assumptions in Section II-B.

Proof: Squaring (17) yields:

$$\begin{aligned} Z_l[r+1]^2 &\leq (Z_l[r] + y_l[r] - c_lT[r])^2 \\ &= Z_l[r]^2 + (y_l[r] - c_lT[r])^2 \\ &\quad +2Z_l[r](y_l[r] - c_lT[r]) \end{aligned}$$

Taking conditional expectations, dividing by 2, and summing over $l \in \{1, \ldots, L\}$ yields the result. □

A. The Drift-Plus-Penalty Ratio Algorithm

Our *Drift-Plus-Penalty Ratio Algorithm* is designed to minimize a sum of the variables on the right-hand-side of the drift bound (18) and a penalty term, divided by an expected frame size, as in [21]. The penalty term uses a non-negative constant V that will be shown to affect a performance tradeoff:

- (Policy Selection) Every frame $r \in \{0, 1, 2, \ldots\}$, observe the virtual queues $\boldsymbol{Z}[r]$ and choose a policy $\pi[r] \in \mathcal{P}$ to minimize the following expression:

$$\frac{\mathbb{E}\left\{V\hat{y}_0(\pi[r]) + \sum_{l=1}^{L} Z_l[r]\hat{y}_l(\pi[r])|\boldsymbol{Z}[r]\right\}}{\mathbb{E}\left\{\hat{T}(\pi[r])|\boldsymbol{Z}[r]\right\}} \quad (20)$$

- (Queue Update) Observe the resulting $\boldsymbol{y}[r]$ and $T[r]$ values, and update virtual queues $Z_l[r]$ by (17).

Details on minimizing (20) are given in Section V. Rather than assuming we achieve the exact infimum of (20) over all policies $\pi[r] \in \mathcal{P}$, it is useful to allow our decisions to come within an additive constant C of the infimum.

Definition 1: A policy $\pi[r]$ is a *C-additive approximation* for the problem (20) if for a given constant $C \geq 0$ we have:

$$\frac{\mathbb{E}\left\{V\hat{y}_0(\pi[r]) + \sum_{l=1}^{L} Z_l[r]\hat{y}_l(\pi[r])|\boldsymbol{Z}[r]\right\}}{\mathbb{E}\left\{\hat{T}(\pi[r])|\boldsymbol{Z}[r]\right\}} \leq$$

$$C + \inf_{\pi \in \mathcal{P}}\left[\frac{\mathbb{E}\left\{V\hat{y}_0(\pi) + \sum_{l=1}^{L} Z_l[r]\hat{y}_l(\pi)|\boldsymbol{Z}[r]\right\}}{\mathbb{E}\left\{\hat{T}(\pi)|\boldsymbol{Z}[r]\right\}}\right]$$

In Section V-B it is shown that the infimum of (20) over $\pi \in \mathcal{P}$ is the same as the infimum over the extended class of probabilistically mixed strategies that choose a *random* $\pi \in \mathcal{P}$ according to some distribution (exactly what i.i.d. policies do every frame). Thus, if policy $\pi[r]$ is a C-additive approximation, then:

$$\mathbb{E}\left\{V\hat{y}_0(\pi[r]) + \textstyle\sum_{l=1}^{L} Z_l[r]\hat{y}_l(\pi[r])|\boldsymbol{Z}[r]\right\} \leq$$

$$\mathbb{E}\left\{\hat{T}(\pi[r])|\boldsymbol{Z}[r]\right\}\left[C + \tfrac{\mathbb{E}\left\{V\hat{y}_0(\pi^*[r]) + \sum_{l=1}^{L} Z_l[r]\hat{y}_l(\pi^*[r])\right\}}{\mathbb{E}\left\{\hat{T}(\pi^*[r])\right\}}\right] \quad (21)$$

where $\pi^*[r]$ is any i.i.d. algorithm. Note that conditional expectations given $\boldsymbol{Z}[r]$ are the same as unconditional expectations under i.i.d. algorithms, because their decisions are independent of system history.

Theorem 1: (Algorithm Performance) Assume the constraints of problem (12)-(14) are feasible. Fix constants $C \geq 0$, $V \geq 0$, and assume the above algorithm is implemented using any C-additive approximation every frame r for the minimization in (20). Assume initial conditions satisfy $\mathbb{E}\{L(\boldsymbol{Z}[0])\} < \infty$. Then:

a) For all $l \in \{1, \ldots, L\}$ we have:

$$\limsup_{R\to\infty} \overline{y}_l[R]/\overline{T}[R] \leq c_l \;\; \forall l \in \{1, \ldots, L\} \quad (22)$$

$$\limsup_{R\to\infty} \frac{\sum_{r=0}^{R-1} y_l[r]}{\sum_{r=0}^{R-1} T[r]} \leq c_l \;\; (w.p.1) \quad (23)$$

where "w.p.1" stands for "with probability 1."

b) For all integers $R > 0$ we have:

$$\frac{\overline{y}_0[R]}{\overline{T}[R]} \leq ratio^{opt} + \frac{(B/\overline{T}[R] + C)}{V} + \frac{\mathbb{E}\{L(\boldsymbol{Z}[0])\}}{VR\overline{T}[R]} \quad (24)$$

and hence:

$$\limsup_{R\to\infty} \overline{y}_0[R]/\overline{T}[R] \leq ratio^{opt} + (B/T^{min} + C)/V \quad (25)$$

where B is defined in (19), and $ratio^{opt}$ is the optimal solution to (12)-(14).

Thus, the algorithm satisfies all constraints, and the value of V can be chosen appropriately large to make $(B/T^{min} + C)/V$ arbitrarily small, ensuring that the time average penalty is arbitrarily close to its optimal value $ratio^{opt}$. The tradeoff in choosing a large value of V comes in the size of the $Z_l[r]$ queues and the number of frames required for $\mathbb{E}\{Z_l[R]\}/R$ to approach zero (which affects convergence time of the algorithm, see (33) in the proof). In particular, it can be shown from (30) that there are constants F_1, F_2 such that (see [24]):

$$\frac{\mathbb{E}\{Z_l[R]\}}{R} \leq \sqrt{\frac{F_1 + VF_2}{R} + \frac{\sum_{l=1}^{L}\mathbb{E}\{Z_l[0]^2\}}{R^2}}$$

Proof: (Theorem 1) Consider any frame $r \in \{0, 1, 2, \ldots\}$. Combining (18) and (21) yields:

$$\Delta(\boldsymbol{Z}[r]) + V\mathbb{E}\{\hat{y}_0(\pi[r])|\boldsymbol{Z}[r]\} \leq B +$$

$$\mathbb{E}\left\{\hat{T}(\pi[r])|\boldsymbol{Z}[r]\right\}\left[C + \tfrac{\mathbb{E}\left\{V\hat{y}_0(\pi^*[r]) + \sum_{l=1}^{L} Z_l[r]\hat{y}_l(\pi^*[r])\right\}}{\mathbb{E}\left\{\hat{T}(\pi^*[r])\right\}}\right]$$

$$-\textstyle\sum_{l=1}^{L} Z_l[r]c_l\mathbb{E}\left\{\hat{T}(\pi[r])|\boldsymbol{Z}[r]\right\} \quad (26)$$

In the above inequality, $\pi[r]$ represents the C-additive approximate decision actually made, and $\pi^*[r]$ is from any alternative i.i.d. algorithm. Fixing any $\delta > 0$, plugging the i.i.d. algorithm $\pi^*[r]$ from (15)-(16) into the right-hand-side of (26), and letting $\delta \to 0$ yields:

$$\begin{aligned} &\Delta(\boldsymbol{Z}[r]) + V\mathbb{E}\{\hat{y}_0(\pi[r])|\boldsymbol{Z}[r]\} \leq B \\ &+\mathbb{E}\left\{\hat{T}(\pi[r])|\boldsymbol{Z}[r]\right\}[C + Vratio^{opt}] \end{aligned} \quad (27)$$

Taking expectations of the above yields:

$$\begin{aligned} &\mathbb{E}\{L(\boldsymbol{Z}[r+1])\} - \mathbb{E}\{L(\boldsymbol{Z}[r])\} + V\mathbb{E}\{\hat{y}_0(\pi[r])\} \leq \\ &\qquad B + \mathbb{E}\left\{\hat{T}(\pi[r])\right\}[C + Vratio^{opt}] \end{aligned} \quad (28)$$

Summing the above over $r \in \{0, \ldots, R-1\}$ for some integer $R > 0$ and dividing by R yields:

$$\begin{aligned} &\frac{\mathbb{E}\{L(\boldsymbol{Z}[R])\} - \mathbb{E}\{L(\boldsymbol{Z}[0])\}}{R} + V\overline{y}_0[R] \leq \\ &\qquad B + \overline{T}[R][C + Vratio^{opt}] \end{aligned} \quad (29)$$

Rearranging terms in the above and using the fact that $\mathbb{E}\{L(\boldsymbol{Z}[R])\} \geq 0$ yields the result of part (b).

To prove part (a), from (27) there is a constant F such that:

$$\Delta(\boldsymbol{Z}[r]) \leq F \tag{30}$$

Thus, the drift of a quadratic Lyapunov function is bounded by a constant. Further, the second moments of per-frame changes in $Z_l[r]$ are bounded because of the second moment assumptions on $y_l[r]$ and $T[r]$. It follows that (see [24]):

$$\lim_{R\to\infty} \mathbb{E}\left\{Z_l[R]\right\}/R = 0 \tag{31}$$

$$\lim_{R\to\infty} Z_l[R]/R = 0 \;\; (w.p.1) \tag{32}$$

Now from the queue update (17) we have for any frame r:

$$Z_l[r+1] \geq Z_l[r] + y_l[r] - c_lT[r]$$

Summing the above over $r \in \{0, \ldots, R-1\}$ for some integer $R > 0$ yields:

$$Z_l[R] - Z_l[0] \geq \textstyle\sum_{r=0}^{R-1}[y_l[r] - c_lT[r]]$$

Taking expectations, dividing by R, and using $\mathbb{E}\left\{Z_l[0]\right\} \geq 0$ yields for all integers $R > 0$:

$$\frac{\mathbb{E}\left\{Z_l[R]\right\}}{R} \geq \overline{y}_l[R] - c_l\overline{T}[R]$$

Thus:

$$\frac{\overline{y}_l[R]}{\overline{T}[R]} \leq c_l + \frac{\mathbb{E}\left\{Z_l[R]\right\}}{R\overline{T}[R]} \leq c_l + \frac{\mathbb{E}\left\{Z_l[R]\right\}}{RT^{min}} \tag{33}$$

Taking limits of the above and using (31) proves (22). A similar argument uses (32) to prove (23). □

Under a mild "Slater-type" assumption that ensures the constraints (13) are achievable with "ϵ-slackness," the queues $Z_l[R]$ can be shown to be *strongly stable*, in the sense that the time average expectation is bounded by $O(V)$. If further mild *fourth moment boundedness assumptions* hold for $y_l[r]$ and $T[r]$ then the same bound (25) can be shown to hold for pure time averages with probability 1 [24].

IV. Utility Optimization

Consider now the problem (8)-(10), which seeks to maximize $\phi(\overline{\boldsymbol{x}}/\overline{T})$ subject to $\overline{y}_l/\overline{T} \leq c_l$ for all $l \in \{1, \ldots, L\}$. We transform this problem of maximizing a function of a time average ratio into a problem of the type (5)-(7). The following variation on Jensen's inequality is crucial in this transformation:

Lemma 3: (Variation on Jensen's Inequality) Let $\phi(\boldsymbol{\gamma})$ be any continuous and concave function defined over $\boldsymbol{\gamma} \in \mathcal{R}$ for some closed and bounded hyper-rectangle $\mathcal{R}$. Let $(T[r], \boldsymbol{\gamma}[r])$ be a sequence of arbitrarily correlated random vectors for $r \in \{0, 1, 2, \ldots\}$. Assume that $T[r] > 0$, $\boldsymbol{\gamma}[r] \in \mathcal{R}$ for all r, and:

$$0 < T^{min} \leq \mathbb{E}\left\{T[r]\right\} \leq T^{max} < \infty \;\; \forall r \in \{0, 1, 2, \ldots\}$$

Then for any $R > 0$:

$$\frac{\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left\{T[r]\phi(\boldsymbol{\gamma}[r])\right\}}{\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left\{T[r]\right\}} \leq \phi\left(\frac{\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left\{T[r]\boldsymbol{\gamma}[r]\right\}}{\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left\{T[r]\right\}}\right)$$

Furthermore, assuming that the limits $\overline{T\phi(\boldsymbol{\gamma})}$ and $\overline{T\boldsymbol{\gamma}}$ defined below exist, we have:

$$\overline{T\phi(\boldsymbol{\gamma})}/\overline{T} \leq \phi(\overline{T\boldsymbol{\gamma}}/\overline{T}) \tag{34}$$

where:

$$\overline{T\phi(\boldsymbol{\gamma})} \triangleq \lim_{R\to\infty}\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left\{T[r]\phi(\boldsymbol{\gamma}[r])\right\}$$

$$\overline{T\boldsymbol{\gamma}} \triangleq \lim_{R\to\infty}\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left\{T[r]\boldsymbol{\gamma}[r]\right\}$$

Proof: See [13]. □

Now define an auxiliary vector $\boldsymbol{\gamma}[r] = (\gamma_1[r], \ldots, \gamma_M[r])$, to be chosen in the set $\mathcal{R}$ defined in (11) on every frame r.

Lemma 4: (Equivalent Transformation) The problem (8)-(10) is equivalent to the following transformed problem:

$$\begin{aligned} \text{Maximize:} & \quad \overline{T\phi(\boldsymbol{\gamma})}/\overline{T} & (35)\\ \text{Subject to:} & \quad \overline{x}_m \geq \overline{T\gamma_m} \;\; \forall m \in \{1, \ldots, M\} & (36)\\ & \quad \overline{y}_l/\overline{T} \leq c_l \;\; \forall l \in \{1, \ldots, L\} & (37)\\ & \quad \boldsymbol{\gamma}[r] \in \mathcal{R} \;\; \forall r \in \{0, 1, 2, \ldots\} & (38)\\ & \quad \pi[r] \in \mathcal{P} \;\; \forall r \in \{0, 1, 2, \ldots\} & (39) \end{aligned}$$

Proof: We briefly sketch the proof: Let $\pi^*[r]$, $\boldsymbol{\gamma}^*[r]$ be a policy that optimally solves the above transformed problem, and assume for simplicity it yields well defined time averages $\overline{T}^*$, $\overline{y}_l^*$, $\overline{x}_m^*$, $\overline{T^*\phi(\boldsymbol{\gamma}^*)}$, $\overline{T^*\boldsymbol{\gamma}^*}$, and optimal utility $util^* = \overline{T^*\phi(\boldsymbol{\gamma}^*)}/\overline{T}^*$. Then the policy $\pi^*[r]$ also satisfies all constraints of problem (8)-(10), and yields:

$$\phi(\overline{\boldsymbol{x}}^*/\overline{T}^*) \geq \phi(\overline{T^*\boldsymbol{\gamma}^*}/\overline{T^*}) \geq \overline{T^*\phi(\boldsymbol{\gamma}^*)}/\overline{T}^* \triangleq util^*$$

where the first inequality above holds by (36) and the entry-wise non-decreasing property of $\phi(\boldsymbol{\gamma})$, and the second holds by (34). Thus, the optimal utility of problem (8)-(10) is greater than or equal to that of the transformed problem. A similar argument shows it is also less than or equal to the optimal utility of the transformed problem. □

The transformed problem (35)-(39) has the structure of the problem (5)-(7) if we define $y_0[r] \triangleq -T[r]\phi(\boldsymbol{\gamma}[r])$, write the constraints (36) as $\overline{T\gamma_m - x_m} \leq 0$, and define policy decision $\pi'[r] \triangleq (\pi[r], \boldsymbol{\gamma}[r]) \in \mathcal{P} \times \mathcal{R}$. The resulting algorithm is thus the same as that given in Section III-A, and for this context it is given as follows: For the constraints (37), use the same virtual queues $Z_l[r]$ defined in (17). For the constraints (36), define virtual queues $G_m[r]$ for $m \in \{1, \ldots, M\}$ by:

$$G_m[r+1] = \max[G_m[r] + T[r]\gamma_m[r] - x_m[r], 0] \tag{40}$$

Define $\boldsymbol{G}[r] \triangleq (G_1[r], \ldots, G_M[r])$. The drift-plus-penalty ratio to minimize every frame r is then:

$$\frac{\mathbb{E}\left\{-V\hat{T}(\pi[r])\phi(\boldsymbol{\gamma}[r]) + \sum_{l=1}^{L}Z_l[r]\hat{y}_l(\pi[r])|\boldsymbol{Z}[r]\right\}}{\mathbb{E}\left\{\hat{T}(\pi[r])|\boldsymbol{Z}[r]\right\}} + \frac{\mathbb{E}\left\{\sum_{m=1}^{M}G_m[r][\hat{T}(\pi[r])\gamma_m[r] - \hat{x}_m(\pi[r])]|\boldsymbol{Z}[r]\right\}}{\mathbb{E}\left\{\hat{T}(\pi[r])|\boldsymbol{Z}[r]\right\}}$$

It is easy to see that the above can be minimized by separately choosing $\boldsymbol{\gamma}[r] \in \mathcal{R}$ and $\pi[r] \in \mathcal{P}$ to minimize their respective terms, and that $\hat{T}(\pi[r])$ cancels out of the auxiliary variable decisions. The resulting algorithm is thus to observe $\boldsymbol{Z}[r]$ and $\boldsymbol{G}[r]$ every frame $r \in \{0, 1, 2, \ldots\}$ and perform the following:

- (Auxiliary Variables) Choose $\boldsymbol{\gamma}[r] \in \mathcal{R}$ to maximize:
$$V\phi(\boldsymbol{\gamma}[r]) - \sum_{m=1}^{M} G_m[r]\gamma_m[r]$$
- (Policy Selection) Choose $\pi[r] \in \mathcal{P}$ to minimize:
$$\frac{\mathbb{E}\left\{\sum_{l=1}^{L} Z_l[r]\hat{y}_l(\pi[r]) - \sum_{m=1}^{M} G_m[r]\hat{x}_m(\pi[r])|\boldsymbol{Z}[r]\right\}}{\mathbb{E}\left\{\hat{T}(\pi[r])|\boldsymbol{Z}[r]\right\}}$$
- (Virtual Queue Update) Update $\boldsymbol{Z}[r]$ by (17) and $\boldsymbol{G}[r]$ by (40).

The auxiliary variable update is a simple deterministic maximization of a concave function over a hyper-rectangle, and can be separated into M optimizations of single-variable concave functions over an interval if the utility function has the form $\phi(\boldsymbol{\gamma}) = \sum_{m=1}^{M}\phi_m(\gamma_m)$. The policy selection step is again an optimization of a ratio of expectations and can be done as described in Section V.

V. Optimizing the Ratio of Expectations

Here we show how to minimize the ratio of expectations given in (20) (and also in the policy selection stage of the previous section). These problems can be written more generally as choosing a policy $\pi[r] \in \mathcal{P}$ to minimize the ratio:
$$\frac{\mathbb{E}\left\{a(\pi)\right\}}{\mathbb{E}\left\{b(\pi)\right\}}$$
where $a(\pi), b(\pi)$ are random functions of $\pi \in \mathcal{P}$, and $b(\pi)$ is strictly positive with $T^{max} \geq \mathbb{E}\left\{b(\pi)|\pi\right\} \geq T^{min} > 0$ for all $\pi \in \mathcal{P}$. The function $b(\pi)$ is equal to $\hat{T}(\pi)$. The function $a(\pi)$ depends on $\boldsymbol{Z}[r]$, and the above expectations are implicitly conditioned on $\boldsymbol{Z}[r]$, although we suppress this notation for simplicity. Define θ^* as the optimal ratio:
$$\theta^* \triangleq \inf_{\pi\in\mathcal{P}}\left[\frac{\mathbb{E}\left\{a(\pi)\right\}}{\mathbb{E}\left\{b(\pi)\right\}}\right]$$
If the expectation $\mathbb{E}\left\{b(\pi)\right\}$ is the same for all $\pi \in \mathcal{P}$ (such as when the frame size is independent of the policy), then θ^* is obtained by infimizing the numerator $\mathbb{E}\left\{a(\pi)\right\}$. This is typically easier (often involving learning for *stochastic shortest path* computations [25][4]). Otherwise, the following simple lemma is useful.

Lemma 5: We have:
$$\inf_{\pi\in\mathcal{P}}\mathbb{E}\left\{a(\pi) - \theta^* b(\pi)\right\} = 0 \tag{41}$$
Further, for any real number θ, we have:
$$\inf_{\pi\in\mathcal{P}}\mathbb{E}\left\{a(\pi) - \theta b(\pi)\right\} < 0 \quad \text{if } \theta > \theta^* \tag{42}$$
$$\inf_{\pi\in\mathcal{P}}\mathbb{E}\left\{a(\pi) - \theta b(\pi)\right\} > 0 \quad \text{if } \theta < \theta^* \tag{43}$$
Proof: See [13]. □

A. The Bisection Algorithm

Lemma 5 immediately leads to the following simple bisection algorithm: Suppose we have upper and lower bounds θ_{min} and θ_{max}, so that we know $\theta_{min} \leq \theta^* \leq \theta_{max}$. Then we can define $\theta = (\theta_{min} + \theta_{max})/2$, and compute the value of $\inf_{\pi\in\mathcal{P}}\mathbb{E}\left\{a(\pi) - \theta b(\pi)\right\}$. If the result is 0, then $\theta = \theta^*$. If positive, then $\theta < \theta^*$, and otherwise $\theta > \theta^*$. We can then refine our upper and lower bounds. This leads to a simple iterative algorithm where the distance between the upper and lower bounds decreases by a factor of 2 on each iteration. It thus approaches the optimal θ^* value exponentially fast. Each step of the iteration involves minimizing an expectation, rather than a ratio of expectations.

B. Optimizing over Pure Policies

Note that for any set of policies $\mathcal{S}$, Lemma 5 implies that $\inf_{\pi\in\mathcal{S}}\mathbb{E}\left\{a(\pi) - \theta b(\pi)\right\} = 0$ if and only if $\theta = \inf_{\pi\in\mathcal{S}}\mathbb{E}\left\{a(\pi)\right\}/\mathbb{E}\left\{b(\pi)\right\}$. Now suppose we have a set of policies $\mathcal{P}^{pure}$ that we call *pure policies*, and that the policy space $\mathcal{P}$ consists of all pure policies as well as all "mixtures" (or convex combinations) of pure policies, being policies that choose a pure policy in $\mathcal{P}^{pure}$ with some particular probability distribution. More generally, define Ω as the set of all vectors $(\mathbb{E}\left\{a(\pi)\right\}, \mathbb{E}\left\{b(\pi)\right\})$ achievable over $\pi \in \mathcal{P}^{pure}$, and suppose the set of all $(\mathbb{E}\left\{a(\pi)\right\}, \mathbb{E}\left\{b(\pi)\right\})$ achievable over $\pi \in \mathcal{P}$ is equal to the convex hull of Ω. Recall that θ^* is the infimum ratio of $\mathbb{E}\left\{a(\pi)\right\}/\mathbb{E}\left\{b(\pi)\right\}$ over $\pi \in \mathcal{P}$. Then:
$$\begin{aligned} 0 = \inf_{\pi\in\mathcal{P}}\mathbb{E}\left\{a(\pi) - \theta^* b(\pi)\right\} &= \inf_{(a,b)\in Conv(\Omega)}[a - \theta^* b] \\ &= \inf_{(a,b)\in\Omega}[a - \theta^* b] \\ &= \inf_{\pi\in\mathcal{P}^{pure}}\mathbb{E}\left\{a(\pi) - \theta^* b(\pi)\right\} \end{aligned}$$
where the third inequality holds because the infimum of a linear function over the convex hull of a set is equal to the infimum over the set itself. It follows that θ^* is also the infimum ratio of $\mathbb{E}\left\{a(\pi)\right\}/\mathbb{E}\left\{b(\pi)\right\}$ over $\pi \in \mathcal{P}^{pure}$.

This means that to achieve the infimum ratio over policies $\pi \in \mathcal{P}$, it suffices to restrict our search to pure policies.

C. Optimizing with Initial Information

Suppose at the beginning of each frame, we observe a vector $\boldsymbol{\eta}[r]$ of *initial information* that can affect the penalties and frame size. Suppose that $\{\boldsymbol{\eta}[r]\}_{r=0}^{\infty}$ is i.i.d. over frames. Each policy $\pi \in \mathcal{P}$ first observes $\boldsymbol{\eta}[r]$ and then chooses a *sub-policy* $\pi' \in \mathcal{P}_{\boldsymbol{\eta}[r]}$, where $\mathcal{P}_{\boldsymbol{\eta}[r]}$ is a space that possibly depends on $\boldsymbol{\eta}[r]$. To minimize $\mathbb{E}\left\{a(\pi)\right\}$, it suffices to observe $\boldsymbol{\eta}[r]$ and choose $\pi' \in \mathcal{P}_{\boldsymbol{\eta}[r]}$ to minimize the conditional expectation $\mathbb{E}\left\{a(\pi')|\boldsymbol{\eta}[r]\right\}$. However, this is not necessarily true for minimizing the ratio $\mathbb{E}\left\{a(\pi)\right\}/\mathbb{E}\left\{b(\pi)\right\}$.

A correct approach is the following: If θ^* is known, we can simply choose $\pi' \in \mathcal{P}_{\boldsymbol{\eta}[r]}$ to minimize:
$$\mathbb{E}\left\{a(\pi') - \theta^* b(\pi')|\boldsymbol{\eta}[r]\right\}$$
If θ^* is unknown, we can carry out the bisection routine. Let θ be the midpoint in the current iteration. We must compute:
$$\inf_{\pi\in\mathcal{P}}\mathbb{E}\left\{a(\pi) - \theta b(\pi)\right\} = \mathbb{E}\left\{\inf_{\pi'\in\mathcal{P}_{\boldsymbol{\eta}[r]}}\mathbb{E}\left\{a(\pi') - \theta b(\pi')|\boldsymbol{\eta}[r]\right\}\right\} \tag{44}$$
The infimizing decision π' can be made by observing $\boldsymbol{\eta}[r]$, without requiring knowledge of its probability distribution. However, the value in (44) cannot be computed without knowledge of this distribution. Instead, suppose we have W

i.i.d. samples $\{\boldsymbol{\eta}_w\}_{w=1}^{W}$. We can then approximate the value in (44) by the function $val(\theta)$ defined below:

$$val(\theta) \triangleq \frac{1}{W}\sum_{w=1}^{W} \inf_{\pi' \in \mathcal{P}_{\eta_w}} \mathbb{E}\left\{a(\pi') - \theta b(\pi')|\boldsymbol{\eta}_w\right\} \tag{45}$$

By the law of large numbers, $val(\theta)$ approaches the exact value of (44) with a large choice of W. The bisection routine can be carried out using the $val(\theta)$ approximation, being sure to use the same samples at each step of the iteration (but different samples on each frame r). Note that $val(\theta)$ is non-increasing in θ, so the bisection will converge provided that it is initialized so that $val(\theta_{min}) \geq 0$ and $val(\theta_{max}) \leq 0$. If we cannot independently generate W samples, we use the W past observed values of $\boldsymbol{\eta}[r]$ from previous frames. There is a subtle issue here, as these past values have influenced system performance and are thus correlated with the current $a(\pi)$ and $b(\pi)$ functions. However, a *delayed queue argument* similar to that given in [26] shows these past values can still be used.

D. Alternative Formulation

Note that constraints of the form $\overline{y}_l \leq 0$ are equivalent to $\overline{y}_l/\overline{T} \leq c_l$ in the special case $c_l = 0$, and thus can be handled using the framework of this paper. Now consider the following problem structure:

$$\begin{array}{rl} \text{Minimize:} & \overline{y}_0 \\ \text{Subject to:} & \overline{y}_l/\overline{T} \leq c_l \ \ \forall l \in \{1,\ldots,L\} \\ & \pi[r] \in \mathcal{P} \ \ \forall r \in \{0,1,2,\ldots\} \end{array}$$

Such a problem has a different structure than the problem (5)-(7), and is *easier to solve* as it does not require a ratio of expectations. It can be solved using the same virtual queues $Z_l[r]$ in (17), but every frame r observing $\boldsymbol{Z}[r]$ and selecting a policy $\pi[r] \in \mathcal{P}$ to minimize the following expression:

$$\mathbb{E}\{V\hat{y}_0(\pi[r]) + \textstyle\sum_{l=1}^{L} Z_l[r][\hat{y}_l(\pi[r]) - c_l\hat{T}(\pi[r])]|\boldsymbol{Z}[r]\}$$

Analysis is omitted for brevity (see Exercise 7.3 in [13]).

E. Alternative Algorithm

The following is an alternative algorithm for the original problem (5)-(7) that does not require a ratio minimization (and hence does not require a bisection step): Use the same virtual queues $Z_l[r]$ in (17). Define $\theta[0] = 0$, and define $\theta[R]$ for $R \in \{1,2,3,\ldots\}$ by:

$$\theta[R] \triangleq \textstyle\sum_{r=0}^{R-1} y_0[r] / \sum_{r=0}^{R-1} T[r] \tag{46}$$

Every frame r, observe $\boldsymbol{Z}[r]$ and $\theta[r]$ and select a policy $\pi[r] \in \mathcal{P}$ to minimize the following expression:

$$\begin{array}{c} \mathbb{E}\{V[\hat{y}_0(\pi[r]) - \theta[r]\hat{T}(\pi[r])]|\boldsymbol{Z}[r],\theta[r]\} \\ +\mathbb{E}\{\sum_{l=1}^{L} Z_l[r][\hat{y}_l(\pi[r]) - c_l\hat{T}(\pi[r])]|\boldsymbol{Z}[r],\theta[r]\} \end{array} \tag{47}$$

It is shown in Exercise 7.5 of [13] that all constraints are met, and that if $\theta[r]$ converges to a constant with probability 1, then with probability 1:

$$\lim_{R\to\infty} \textstyle\sum_{r=0}^{R-1} y_0[r] / \sum_{r=0}^{R-1} T[r] \leq ratio^{opt} + O(1/V)$$

The disadvantage is that the convergence time is not as clear as that given in part (b) of Theorem 1. Further, use of the time average (46) makes it difficult to adapt to changes in system parameters, so that it may be better to approximate (46) with a moving average or an exponentially decaying average.

VI. Simulations for a Task Processing Network

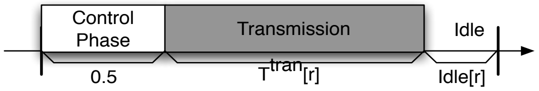

Fig. 2. An illustration of the 3 phases of a renewal frame $r \in \{0,1,2,\ldots\}$.

Here we provide a simple task processing example. An infinite sequence of tasks must be processed one at a time with the help of a network of 5 wireless devices. This applies, for example, in scenarios similar to [22] where each new task represents an event that is sensed by the wireless devices (each at different sensing qualities [27]), and we must select which device reports the event information. The renewal structure is shown in Fig. 2. At the beginning of each new task r, a period of 0.5 time units is expended to communicate control information about the task. Each of the 5 devices expends 0.5 units of energy in this control phase. At the end of this phase, the network controller obtains a vector $\boldsymbol{\eta}[r]$ of parameters for task r. The vector $\boldsymbol{\eta}[r]$ has the form:

$$\boldsymbol{\eta}[r] = [(qual_1[r], T_1^{tran}[r]), \ldots, (qual_5[r], T_5^{tran}[r])]$$

where for each $l \in \{1,\ldots,5\}$, $qual_l[r]$ is a real number representing the *information quality* if device l is chosen to process task r, and $T_l^{tran}[r]$ is the *transmission time* required for device l to transmit the corresponding information to a receiving station. The controller must choose one of the 5 devices to process the task, and must also choose the amount of *idle time* at the end of the frame (chosen within the interval $[0,5]$), so that the policy decision $\pi[r]$ has the form:

$$\pi[r] = (l[r], Idle[r]) \in \{1,2,3,4,5\} \times \{I \in \mathbb{R}|0 \leq I \leq 5\}$$

Define P^{tran} as the power expenditure associated with wireless transmission. The chosen device $l[r]$ expends $P^{tran} \times T_{l[r]}^{tran}$ units of energy in the transmit phase, while all other devices $l \neq l[r]$ expend no energy in this phase. None of the devices expend energy in the idle phase, which helps to limit the average power expenditure in the system.

The goal is to maximize the *quality of information (q.o.i) per unit time* subject to an average power constraint of 0.25 at each device. Define $\hat{y}_0(\pi[r])$ as -1 times the q.o.i. obtained for task r, $\hat{y}_l(\pi[r])$ as the energy expended by device l on task r, and $\hat{T}(\pi[r])$ as the frame duration for task r:

$$\begin{array}{rcl} \hat{y}_0(\pi[r]) & \triangleq & -qual_{l[r]}[r] \\ \hat{y}_l(\pi[r]) & \triangleq & 0.5 + P^{tran}T_l^{tran}[r]1_{\{l[r]=l\}} \ \forall l \in \{1,\ldots,5\} \\ \hat{T}(\pi[r]) & \triangleq & 0.5 + T_{l[r]}^{tran}[r] + Idle[r] \end{array}$$

where $1_{\{l[r]=l\}}$ is an indicator function that is 1 if $l[r] = l$ and 0 else. The problem is then to minimize $\overline{y}_0/\overline{T}$ subject to $\overline{y}_l/\overline{T} \leq 0.25$ for all $l \in \{1,\ldots,5\}$.

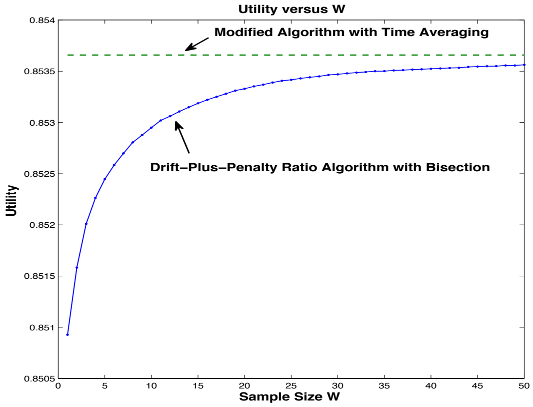

Fig. 3. Utility for the drift-plus-penalty ratio algorithm (with bisection) and the time-averaged alternative.

We simulate the drift-plus-penalty ratio algorithm for 10^6 frames, using the bisection method with W past samples of $\boldsymbol{\eta}[r]$ as in (45) of Section V-C. We use $P^{tran} = 1.0$. The vectors $\{\boldsymbol{\eta}[r]\}_{r=0}^{\infty}$ are assumed to be i.i.d. with independently chosen components, where $T_l^{tran}[r]$ is uniformly distributed in $[0.5, 2.5]$ for all l, and $qual_l[r]$ is uniformly distributed in $[0, l]$ for $l \in \{1, 2, 3, 4, 5\}$ (so that device 5 tends to have the highest quality, while device 1 tends to have the lowest). We initialize $\theta_{min} = -5V$, $\theta_{max} \triangleq \sum_{l=1}^{5} Z_l[r]3$. Each step of the bisection computes $val(\theta)$ according to a simple deterministic optimization, and the bisection routine is run for each frame until $\theta_{max} - \theta_{min} < 0.001$. Using $V = 100$, the resulting q.o.i per unit time is plotted in Fig. 3. This increases to its optimal value as W is increased. However, in this example, W does not need to be very large for accurate results: Even $W = 1$ produces a value that is near optimal (note that the y-axis in Fig. 3 distinguishes utility only in the 3rd significant digit).

All average power constraints are met in all simulations (for each W). Results for $W = 10$ are: $q.o.i/\overline{T} = 0.852950$, $\overline{T} = 3.180275$, $\overline{Idle} = 1.421260$, $\overline{y}_0 = -2.712615$, and:

$$\overline{y}_1/\overline{T} = 0.182335 \leq 0.25$$

$$\overline{y}_2/\overline{T} = 0.249547 \leq 0.25 \ , \ \overline{y}_3/\overline{T} = 0.250018 \leq 0.25$$

$$\overline{y}_4/\overline{T} = 0.250032 \leq 0.25 \ , \ \overline{y}_5/\overline{T} = 0.250046 \leq 0.25$$

It can be seen that devices $\{2, \ldots, 5\}$ are utilized to their maximum power constraints because these tend to give the highest quality, while average power for device 1 is slack.

The alternative algorithm of Section V-E, which does not require a bisection routine and amounts to a simple deterministic optimization for (47) every frame, achieves similar time average power expenditures to the above. It also achieves utility as shown in Fig. 3, being the constant that does not depend on W (as no sampling from the past is needed). Its utility is slightly larger than that of the bisection algorithm, and is approached by the bisection algorithm as W increases. It appears that this algorithm is simpler and yields "automatic learning" by using the time average value $\theta[r]$, but it might have trouble adapting if system parameters change.

References

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

[2] S. Ross. *Introduction to Probability Models*. Academic Press, 8th edition, Dec. 2002.

[3] L. Georgiadis, M. J. Neely, and L. Tassiulas. Resource allocation and cross-layer control in wireless networks. *Foundations and Trends in Networking*, vol. 1, no. 1, pp. 1-149, 2006.

[4] M. J. Neely. Stochastic optimization for markov modulated networks with application to delay constrained wireless scheduling. *Proc. IEEE Conf. on Decision and Control (CDC)*, Shanghai, China, Dec. 2009.

[5] F. J. Vázquez Abad and V. Krishnamurthy. Policy gradient stochastic approximation algorithms for adaptive control of constrained time varying markov decision processes. *Proc. IEEE Conf. on Decision and Control*, Dec. 2003.

[6] D. V. Djonin and V. Krishnamurthy. q-learning algorithms for constrained markov decision processes with randomized monotone policies: Application to mimo transmission control. *IEEE Transactions on Signal Processing*, vol. 55, no. 5, pp. 2170-2181, May 2007.

[7] N. Salodkar, A. Bhorkar, A. Karandikar, and V. S. Borkar. An on-line learning algorithm for energy efficient delay constrained scheduling over a fading channel. *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 4, pp. 732-742, May 2008.

[8] F. Fu and M. van der Schaar. A systematic framework for dynamically optimizing multi-user video transmission. *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 3, pp. 308-320, April 2010.

[9] F. Fu and M. van der Schaar. Decomposition principles and online learning in cross-layer optimization for delay-sensitive applications. *IEEE Trans. Signal Processing*, vol. 58, no. 3, pp. 1401-1415, March 2010.

[10] M. J. Neely. *Dynamic Power Allocation and Routing for Satellite and Wireless Networks with Time Varying Channels*. PhD thesis, Massachusetts Institute of Technology, LIDS, 2003.

[11] M. J. Neely. Energy optimal control for time varying wireless networks. *IEEE Transactions on Information Theory*, vol. 52, no. 7, pp. 2915-2934, July 2006.

[12] M. J. Neely, E. Modiano, and C. Li. Fairness and optimal stochastic control for heterogeneous networks. *Proc. IEEE INFOCOM*, March 2005.

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

[14] A. Eryilmaz and R. Srikant. Fair resource allocation in wireless networks using queue-length-based scheduling and congestion control. *IEEE/ACM Transactions on Networking*, vol. 15, no. 6, pp. 1333-1344, Dec. 2007.

[15] A. Stolyar. Maximizing queueing network utility subject to stability: Greedy primal-dual algorithm. *Queueing Systems*, vol. 50, no. 4, pp. 401-457, 2005.

[16] A. Stolyar. Greedy primal-dual algorithm for dynamic resource allocation in complex networks. *Queueing Systems*, vol. 54, no. 3, pp. 203-220, 2006.

[17] Q. Li and R. Negi. Scheduling in wireless networks under uncertainties: A greedy primal-dual approach. *Arxiv Technical Report: arXiv:1001:2050v2*, June 2010.

[18] X. Lin and N. B. Shroff. Joint rate control and scheduling in multihop wireless networks. *Proc. of 43rd IEEE Conf. on Decision and Control, Paradise Island, Bahamas*, Dec. 2004.

[19] R. Agrawal and V. Subramanian. Optimality of certain channel aware scheduling policies. *Proc. 40th Annual Allerton Conference on Communication , Control, and Computing, Monticello, IL*, Oct. 2002.

[20] H. Kushner and P. Whiting. Asymptotic properties of proportional-fair sharing algorithms. *Proc. of 40th Annual Allerton Conf. on Communication, Control, and Computing*, 2002.

[21] C.-P. Li and M. J. Neely. Network utility maximization over partially observable markovian channels. *Arxiv Technical Report: arXiv:1008.3421v1*, Aug. 2010.

[22] B. Liu, P. Terlecky, A. Bar-Noy, R. Govindan, and M. J. Neely. Optimizing information credibility in social swarming applications. *ArXiv technical report, arXiv:1009:6006*, Sept. 2010.

[23] D. Williams. *Probability with Martingales*. Cambridge Mathematical Textbooks, Cambridge University Press, 1991.

[24] M. J. Neely. Queue stability and probability 1 convergence via lyapunov optimization. *Arxiv Technical Report*, Oct. 2010.

[25] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, Mass, 1996.

[26] M. J. Neely. Max weight learning algorithms with application to scheduling in unknown environments. *arXiv:0902.0630v1*, Feb. 2009.

[27] C. Bisdikian, L. M. Kaplan, M. B. Srivastava, D. J. Thornley, D. Verma, and R. I. Young. Building principles for a quality of information specification for sensor information. *12th Int'l Conf. on Information Fusion (Fusion '09), Seattle, WA*, July 2009.