

Stochastic Optimization for Markov Modulated Networks with Application to Delay Constrained Wireless Scheduling

Michael J. Neely

Abstract—We consider a wireless system with a small number of delay constrained users and a larger number of users without delay constraints. We develop a scheduling algorithm that reacts to time varying channels and maximizes throughput (to within a desired proximity), stabilizes all queues, and satisfies the delay constraints. The problem is solved by reducing the constrained optimization to a set of weighted stochastic shortest path problems, which act as natural generalizations of max-weight policies to Markov modulated networks. We also present performance bounds when the shortest path problems are solved inexactly, and discuss the additional complexity as compared to systems without delay constraints. The solution technique is general and applies to other constrained stochastic network optimization problems.

I. INTRODUCTION

This paper considers delay-aware scheduling in a multi-user wireless uplink or downlink with K delay-constrained users and N delay-unconstrained users, each with different transmission channels. The system operates in slotted time with normalized slots $t \in \{0, 1, 2, \dots\}$. Every slot, a random number of new packets arrive from each user. Packets are queued for eventual transmission, and every slot a scheduler looks at the queue backlog and the current channel states and chooses one channel to serve. The number of packets transmitted over that channel depends on its current channel state. The goal is to stabilize all queues, satisfy average delay constraints for the delay-constrained users, and drop as few packets as possible.

Without the delay constraints, this problem is a classical *opportunistic scheduling* problem, and can be solved with efficient max-weight algorithms based on Lyapunov drift and Lyapunov optimization (see [2] and references therein). The delay constraints make the problem a much more complex *Markov Decision Problem* (MDP). While general methods for solving MDPs exist (see, for example, [3][4]), they typically suffer from a curse of dimensionality. Specifically, the number of queue state vectors grows geometrically in the number of queues. This creates non-polynomial implementation complexity for offline approaches such as linear programming [3][4], and non-polynomial complexity and/or learning time for online or quasi online/offline approaches such as Q -learning and temporal differences for dynamic

programs [5][6][7], or stochastic approximation with 2-timescale arguments for constrained MDPs [8][9][10].

We do not solve this fundamental curse of dimensionality. Rather, we avoid this difficulty by focusing on the special structure that arises in a wireless network with a *relatively small number of delay-constrained users* (say, $K \leq 5$), but with an arbitrarily large number of users without delay constraints (so that N can be large). This is an important scenario, particularly in cases when the number of “best effort” users in a network is much larger than the number of delay-constrained users. We develop a solution that, on each slot, requires a computation that has a complexity that depends geometrically in K , but only polynomially in N . Further, the resulting convergence times and delays are fully polynomial in the total number of queues $K + N$. Our solution uses a concept of *forced renewals* that introduces a deviation from optimality that can be made arbitrarily small with a corresponding polynomial tradeoff in convergence time. Finally, we show that a simple Robbins-Monro approximation can be used, without knowledge of the channel or traffic statistics, and yields similar performance. Our methods are general and can be applied to other MDPs for queueing networks with similar structure.

Related prior work on delay optimality for multi-user opportunistic scheduling under special symmetric assumptions is developed in [11][12][13], and single-queue delay optimization problems are treated in [14][15][10][9] using dynamic programming and Markov Decision theory. Approximate dynamic programming algorithms are applied to multi-queue switches in [16] and shown to perform well in simulation. Optimal asymptotic energy-delay tradeoffs are developed for single queue systems in [17], and optimal energy-delay and utility-delay tradeoffs for multi-queue systems are treated in [18][19]. The algorithms of [18][19] have very low complexity and provably converge quickly even for large networks, although the tradeoff-optimal delay guarantees they achieve do not necessarily optimize the coefficient multiplier in the delay expression.

Our approach in the present paper treats the MDP problem associated with delay constraints using Lyapunov drift and Lyapunov optimization theory [2]. We extend the max-weight principles for stochastic network optimization to treat *Markov-modulated networks*, where the network costs depend on both the control actions taken and the current state (such as the queue state) the system is in. For each cost constraint we define a *virtual queue*. This is similar to the Lagrange multiplier approaches used in the related works [10][9] that treat power minimization for single-queue

Michael J. Neely is with the Electrical Engineering department at the University of Southern California, Los Angeles, CA. (web: <http://www-rcf.usc.edu/~mjneely>).

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.

wireless links with an average delay constraint (see also [8]). However, we treat multi-queue networks, and we use a different analytical approach that emphasizes stochastic shortest paths over variable length frames. Because of this, our approach can be used in conjunction with a variety of existing online techniques for solving shortest path problems (see, for example, [5][7]). Further, we provide bounds on the performance degradation when the shortest path problems are solved inexactly. Hence, our work can be used with recent *approximate dynamic programming* approaches that approximate the value function of the shortest path problem with a simpler function [20][16][5].

II. NETWORK MODEL

Consider a stochastic queueing network that operates in slotted time $t \in \{0, 1, 2, \dots\}$. Let \mathcal{N} represent a finite set of queues to be stabilized, and let $\mathbf{Q}(t) = (Q_n(t))_{n \in \mathcal{N}}$ denote the vector of queue backlogs on slot t . Each queue $Q_n(t)$ has an infinite buffer and has dynamics:

$$Q_n(t+1) = \max[Q_n(t) - \mu_n(t), 0] + R_n(t) \quad (1)$$

where $\mu_n(t)$ and $R_n(t)$ are the service rate and new arrivals, respectively, for queue n on slot t . Let $\boldsymbol{\mu}(t)$ and $\mathbf{R}(t)$ be the vector of service rates and arrival variables with entries $n \in \mathcal{N}$. On each slot t , the vectors $\boldsymbol{\mu}(t)$ and $\mathbf{R}(t)$ are determined as functions of a *random event* $\Omega(t)$, a *state variable* $z(t)$, and a *control action* $I(t)$:

$$\boldsymbol{\mu}(t) \triangleq \hat{\boldsymbol{\mu}}(I(t), \Omega(t), z(t)) \quad , \quad \mathbf{R}(t) \triangleq \hat{\mathbf{R}}(I(t), \Omega(t), z(t))$$

Specifically, the control action $I(t)$ is made every slot with knowledge of $z(t)$ and $\Omega(t)$ (and also $\mathbf{Q}(t)$), and is constrained to take values in an abstract set $\mathcal{I}_{\Omega(t), z(t)}$ that has arbitrary cardinality and that possibly depends on $\Omega(t)$ and $z(t)$. The random event $\Omega(t)$ takes values in a set with arbitrary cardinality, and represents a collection of network parameters (such as channel states) that can randomly change from slot to slot. We assume that $\Omega(t)$ is i.i.d. over slots with some fixed (but potentially unknown) distribution that does not depend on the current state or the past network control actions. The state variable $z(t)$ takes values in a set \mathcal{Z} with arbitrary cardinality, and represents a controlled Markov chain related to the network (this will be used to represent delay-constrained queues in the next subsection). The probabilistic transitions of $z(t)$ depend on $\Omega(t)$ and on the control decision $I(t)$ through a general Markov relation:

$$z(t) \xrightarrow{I(t), \Omega(t)} z(t+1)$$

When \mathcal{Z} is finite or countably infinite, for all $y, z \in \mathcal{Z}$ we define the transition probability matrix $P_{yz}(I, \Omega)$:

$$P_{yz}(I, \Omega) \triangleq Pr[z(t+1) = z \mid z(t) = y, I(t) = I, \Omega(t) = \Omega]$$

The state space \mathcal{Z} is assumed to contain a state 0 that is accessible from any state $z \in \mathcal{Z}$. In the next sub-section, we impose an additional *ϕ -forced renewal assumption* that requires the probability of returning to the 0 state to be at least $\phi > 0$ for any policy (discussed in more detail in Section II-A).

For each slot t we have a collection of general *network penalties* $x_m(t)$ for $m \in \{0, 1, \dots, M\}$ for some positive integer M . These are defined by *penalty functions* $\hat{x}_m(\cdot)$ that represent different types of costs incurred when a control action $I(t)$ is taken under a given $\Omega(t)$ and $z(t)$:

$$x_m(t) \triangleq \hat{x}_m(I(t), \Omega(t), z(t))$$

The penalty functions are arbitrary and possibly negative (negative penalties can represent rewards).¹ However, for convenience we assume that the functions $\hat{x}_m(\cdot)$, $\hat{R}_n(\cdot)$, $\hat{\mu}_n(\cdot)$ are upper and lower bounded by finite constants. In particular we assume $x_0^{min} \leq \hat{x}_0(\cdot) \leq x_0^{max}$ for some constants x_0^{min} , x_0^{max} .

For each penalty $m \in \{0, 1, \dots, M\}$, each queue $Q_n(t)$ for $n \in \mathcal{N}$, and for a given control policy that makes decisions $I(t)$ over time, we define the following time averages:

$$\bar{x}_m \triangleq \limsup_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E} \{ \hat{x}_m(I(\tau), \Omega(\tau), z(\tau)) \}$$

$$\bar{Q}_n \triangleq \limsup_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E} \{ Q_n(\tau) \}$$

We say that a queue $Q_n(t)$ is *stable* if $\bar{Q}_n < \infty$.² We now state the stochastic optimization problem of interest.

Stochastic Optimization Problem: Determine a control policy that solves:

$$\text{Minimize:} \quad \bar{x}_0 \quad (2)$$

$$\text{Subject to:} \quad \bar{x}_m \leq x_m^{av} \text{ for all } m \in \{1, \dots, M\} \quad (3)$$

$$\bar{Q}_n < \infty \text{ for all } n \in \mathcal{N} \quad (4)$$

where each constant x_m^{av} represents a desired constraint on the time average of penalty $x_m(t)$.

This is similar to the stochastic network optimization problems of [2][21], with the exception that the penalty function now includes the state variable $z(t) \in \mathcal{Z}$, which has control-dependent transitions. We say that the problem is a *stochastic feasibility problem* if we desire only to satisfy the time average constraints (3)-(4), without regard for minimization of \bar{x}_0 . The problem (2)-(4) is generalized in our technical report [1] to treat optimization of convex functions of time averages, similar to the objectives considered without the Markov modulated state variable $z(t)$ in [22][2][23][24].

A. The Forced Renewal Assumption

To ensure that the $z(t)$ state variable ‘‘renews’’ itself regularly by returning to the 0 state, we consider the following simple (and sub-optimal) mechanism. Let $\Omega(t) \triangleq [\omega(t); \phi(t)]$, where $\omega(t)$ is the actual random network event (taking values in an abstract set \mathcal{W} with arbitrary cardinality), and $\phi(t)$ is an independent Bernoulli 0/1 variable that is i.i.d. over slots with $Pr[\phi(t) = 1] = \phi$, for some small but positive *forced renewal probability* $\phi > 0$. If $\phi(t) = 1$, the

¹We also assume that linear functionals of $\hat{R}_n(\cdot)$, $\hat{\mu}_n(\cdot)$, $\hat{x}_m(\cdot)$ have well defined maximizers $I \in \mathcal{I}_{\Omega(t), z(t)}$ for each $\Omega(t)$, $z(t)$.

²This is called *strongly stable* as it implies finite average queue backlog.

system experiences a *forced renewal event* which ensures that $z(t+1) = 0$. The value of $\phi(t)$ is known to the network controller at the beginning of slot t , although if $\phi(t) = 1$ the renewal itself only occurs at the end of slot t when the next state $z(t+1)$ is forced to 0. In this way, the control action taken during a slot t in which $\phi(t) = 1$ still affects the functions $\hat{\mu}(I(t), \Omega(t), z(t))$, $\hat{R}(I(t), \Omega(t), z(t))$, $\hat{x}_m(I(t), \Omega(t), z(t))$, and these functions possibly have different values when $\phi(t) = 0$ versus $\phi(t) = 1$.

This forced renewal structure implicitly assumes that the system can physically reset the state variable $z(t)$ to zero on any slot. Further, even if the system has this physical capability, it is generally sub-optimal to force such renewals with probability ϕ every slot. However, for many systems of interest (such as the network defined in the next subsection), if ϕ is small then the optimal performance over systems constrained by this ϕ -forced renewal assumption is close to the optimal performance for systems without forced renewals. Throughout this paper, we define optimality in terms of systems with ϕ -forced renewals, with the understanding that ϕ is a small but positive value.

B. Wireless Systems with Delay Constraints

The above framework can treat many systems, including wireless systems with opportunistic scheduling. For example, suppose $\omega(t) = [\mathbf{A}(t); \mathbf{S}(t)]$ represents an i.i.d. vector of new arrivals $\mathbf{A}(t)$ and observed channel states $\mathbf{S}(t)$ for all queues. Let $\mathcal{N} \triangleq \{1, \dots, N\}$ denote a set of delay-unconstrained users, and let $\mathcal{K} \triangleq \{1, \dots, K\}$ represent a set of delay-constrained users. All packets have fixed lengths, and we let $\mathbf{Q}(t) = (Q_n(t))_{n \in \mathcal{N}}$ and $\mathbf{Z}(t) = (Z_k(t))_{k \in \mathcal{K}}$ be the vector of integer queue lengths for all delay-unconstrained and delay-constrained users, respectively, on slot t . Suppose the queue for each delay-constrained user has finite size B_{max} , and let \mathcal{Z} represent the state space for $\mathbf{Z}(t)$, which has a finite size of $(B_{max} + 1)^K$. Let $z(t)$ represent $\mathbf{Z}(t)$.

Forced renewals occur according to the i.i.d. Bernoulli process $\phi(t)$ with forced renewal probability $\phi > 0$. If a forced renewal occurs on slot t (so that $\phi(t) = 1$), all data in all delay-constrained queues $k \in \mathcal{K}$ is dropped at the end of the slot, so that $z(t+1) = 0$ (where the state $0 \in \mathcal{Z}$ represents the vector of all zeros). The data in the delay-unconstrained queues is not dropped. The maximum drop rate in a queue $k \in \mathcal{K}$ due to such forced renewals is at most $(B_{max} + \lambda_k)\phi$ drops/slot, where λ_k is the rate of new arrivals to queue k . The value $(B_{max} + \lambda_k)\phi$ can be made arbitrarily small with a small choice of ϕ .

Control actions $I(t)$ determine transmission decisions $\mu_n(t)$ (such as selecting a single queue from all options, based on the channel state), as well as *packet drop decisions* that respect the finite buffer size of the delay-constrained queues. Specifically, for each finite buffer queue $k \in \mathcal{K}$, decision variables $R_k(t)$ and $D_k(t)$ respectively represent the amount of new arrivals added on slot t and new packets dropped on slot t , where $R_k(t) + D_k(t) = A_k(t)$. The update

equation for delay-constrained queues is then given by:

$$Z_k(t+1) = \begin{cases} \max[Z_k(t) - \mu_k(t), 0] + R_k(t) & \text{if } \phi(t) = 0 \\ 0 & \text{if } \phi(t) = 1 \end{cases} \quad (5)$$

Given the $\Omega(t) = [\phi(t); \omega(t)]$, $z(t)$, and $I(t)$ values, the next-state is deterministically known, so that the transition probabilities $P_{zy}(I, \Omega)$ in this example are either 0 or 1.

C. Example Penalties for Average Congestion and Delay

To enforce a constraint on the average congestion in queue $Z_k(t)$ (for a given $k \in \mathcal{K}$), we can define a penalty function:

$$\hat{x}_k(I(t), \Omega(t), z(t)) = Z_k(t)$$

This penalty function does not use the $I(t)$ or $\Omega(t)$ arguments, and uses the fact that $Z_k(t)$ is a component of the $z(t)$ state variable. Enforcing the constraint $\bar{x}_k \leq x_k^{av}$ ensures that average queue congestion is no more than x_k^{av} .

To enforce a constraint on the time average rate of packet drops in a delay-constrained queue $k \in \mathcal{K}$, we can define a penalty function of the form:

$$\hat{x}_k(I(t), \Omega(t), z(t)) = \begin{cases} A_k(t) - R_k(t) & , \text{if } \phi(t) = 0 \\ A_k(t) + Z_k(t) - \tilde{\mu}_k(t) & , \text{if } \phi(t) = 1 \end{cases}$$

where $\tilde{\mu}_k(t) \triangleq \min[\mu_k(t), Z_k(t)]$ and represents the number of packets served in queue k on slot t . In this case, the penalty is equal to the exact amount of packet drops in queue k on slot t , so that ensuring $\bar{x}_k \leq x_k^{av}$ enforces a constraint on the time average rate of packet drops.

Finally, to enforce a constraint that the average delay of (non-dropped) packets in a queue $k \in \mathcal{K}$ is less than or equal to some desired bound W_k^{av} (where W_k^{av} is a given constant), we can use a penalty function of the form:

$$\hat{x}_k(I(t), \Omega(t), z(t)) = Z_k(t) - \tilde{\mu}_k(t)W_k^{av}$$

and enforce the constraint $\bar{x}_k \leq 0$. Assuming time average limits are well defined, this ensures that

$$\bar{Z}_k - \tilde{\lambda}_k W_k^{av} \leq 0 \quad (6)$$

where $\tilde{\lambda}_k$ is the time average rate of actual packets served in queue k (and is also the time average rate of non-dropped packets that are admitted). By Little's Theorem [25][3], we know that $\bar{Z}_k = \tilde{\lambda}_k \bar{W}_k$, where \bar{W}_k is the average delay of non-dropped packets in queue k . Using this with (6), we deduce that $\bar{W}_k \leq W_k^{av}$ (assuming that $\lambda_k > 0$).

D. Slackness Assumptions

Consider the general stochastic queueing network model, and let $\mathcal{M} \triangleq \{1, \dots, M\}$ represent the set of penalties involved in the feasibility constraints (3)-(4).

Definition 1: A control policy $I(t)$ is a (z, Ω) -only policy if it satisfies the ϕ -forced renewal assumption, and if it makes stationary and randomized decisions $I(t) \in \mathcal{I}_{\Omega(t), z(t)}$ for each slot t based only on the current $\Omega(t)$ and $z(t)$.

Assume that $z(0) = 0$, and define *renewal events* as times $\{t_g\}_{g=0}^{\infty}$ when forced renewals ensure that $z(t_g) = 0$ (including also the time $t_0 = 0$). Define the *renewal interval*

as the duration of slots between renewal events (including the first and not including the second), and note that the expected duration is equal to $1/\phi$. Under any particular (z, Ω) -only policy $I^*(t)$, the system has independent and identically distributed behavior on each renewal interval. By basic renewal theory, all time average penalties have well defined limits that are exactly equal to the expected sum penalty over a renewal interval divided by the expected duration of the renewal interval [25]. Starting at time 0 and defining T as the time of the next renewal event, we have:

$$\bar{x}_m^* = \frac{\mathbb{E} \left\{ \sum_{\tau=0}^{T-1} \hat{x}_m(I^*(\tau), \Omega(\tau), z^*(\tau)) \right\}}{\mathbb{E} \{T\}} \quad \forall m \in \mathcal{M}$$

$$\bar{\mu}_n^* - \bar{r}_n^* = \frac{\mathbb{E} \left\{ \sum_{\tau=0}^{T-1} \hat{d}_n(I^*(\tau), \Omega(\tau), z^*(\tau)) \right\}}{\mathbb{E} \{T\}} \quad \forall n \in \mathcal{N}$$

where $\mathbb{E} \{T\} = 1/\phi$, and where $\hat{d}_n(I, \Omega, z)$ is defined:

$$\hat{d}_n(I, \Omega, z) \triangleq \hat{\mu}_n(I, \Omega, z) - \hat{R}_n(I, \Omega, z)$$

Suppose there exists a (z, Ω) -only policy $I^*(t)$ that satisfies the feasibility constraints (3)-(4). Let $z^*(t)$ represent the resulting network state variable under this policy, and let \bar{x}_m^* , $\bar{\mu}_n^*$, \bar{r}_n^* respectively represent the time average of penalty $x_m(t)$, transmission $\mu_n(t)$, and admission $R_n(t)$, under policy $I^*(t)$. Because queue stability requires the time average arrival rate to be less than or equal to the time average service rate, it is easy to show that (3)-(4) imply:

$$\bar{x}_m^* \leq x_m^{av} \quad \text{for all } m \in \mathcal{M} \quad (7)$$

$$\bar{\mu}_n^* - \bar{r}_n^* \geq 0 \quad \text{for all } n \in \mathcal{N} \quad (8)$$

Let x_0^{opt} represent the infimum value of \bar{x}_0 over all (z, Ω) -only policies that satisfy (7)-(8). We shall measure optimality of our algorithm designs with respect to x_0^{opt} . This is typically non-restrictive. For example, if \mathcal{Z} has a finite state space, it can be shown that the infimum of \bar{x}_0 over *all policies* that satisfy (3)-(4) is equal to x_0^{opt} .³

In addition to assuming that the feasibility constraints (7)-(8) can be satisfied by a (z, Ω) -only policy, we make the following two mild assumptions. The first is made only for convenience, and states that the infimum value x_0^{opt} can be achieved by a particular (z, Ω) -only policy.⁴ The second is a *slackness assumption* that is a stochastic analogue of a *Slater condition* for static optimization problems [26].

³This can be shown by well known optimality of stationary randomized policies for MDP problems over finite state spaces [4] and for queue stability problems [21], although the formal proof is omitted for brevity.

⁴Our main theorem, Theorem 1, can be proven without Assumption 1 by taking a limit over policies that approach x_0^{opt} .

Assumption 1: (Optimization) There exists a (z, Ω) -only policy $I^*(t)$ that satisfies:

$$\frac{\mathbb{E} \left\{ \sum_{\tau=0}^{T-1} x_0^*(\tau) \right\}}{\mathbb{E} \{T\}} = x_0^{opt} \quad (9)$$

$$\frac{\mathbb{E} \left\{ \sum_{\tau=0}^{T-1} x_m^*(\tau) \right\}}{\mathbb{E} \{T\}} \leq x_m^{av} \quad \forall m \in \mathcal{M} \quad (10)$$

$$\frac{\mathbb{E} \left\{ \sum_{\tau=0}^{T-1} d_n^*(\tau) \right\}}{\mathbb{E} \{T\}} \geq 0 \quad \forall n \in \mathcal{N} \quad (11)$$

where T is the size of the first renewal interval, $z^*(\tau)$ is the network state at time τ under policy $I^*(t)$, and for notational simplicity we have defined:

$$x_m^*(\tau) \triangleq \hat{x}_m(I^*(\tau), \Omega(\tau), z^*(\tau))$$

$$d_n^*(\tau) \triangleq \hat{d}_n(I^*(\tau), \Omega(\tau), z^*(\tau))$$

Assumption 2: (Slackness of Feasibility) There exists a value $\epsilon > 0$ such that the constraints of (7)-(8) can be met with ϵ slackness. Specifically, there is a (z, Ω) -only policy $I^*(t)$ (not necessarily the same policy as in Assumption 1) that satisfies the following for all $m \in \mathcal{M}$ and $n \in \mathcal{N}$:

$$\frac{\mathbb{E} \left\{ \sum_{\tau=0}^{T-1} x_m^*(\tau) \right\}}{\mathbb{E} \{T\}} \leq x_m^{av} - \epsilon \quad (12)$$

$$\frac{\mathbb{E} \left\{ \sum_{\tau=0}^{T-1} d_n^*(\tau) \right\}}{\mathbb{E} \{T\}} \geq \epsilon \quad (13)$$

III. THE DYNAMIC CONTROL ALGORITHM

To solve the stochastic feasibility and stochastic optimization problems for our queueing network, we extend the framework of [2] to a case of variable length frames. Specifically, for each time average penalty constraint (3), parameterized by $m \in \mathcal{M} \triangleq \{1, \dots, M\}$, we define a *virtual queue* $Y_m(t)$ that is initialized to zero and has dynamic update equation:

$$Y_m(t+1) = \max[Y_m(t) - x_m^{av} + x_m(t), 0] \quad (14)$$

where $x_m(t) = \hat{x}_m(I(t), \Omega(t), z(t))$ is the penalty incurred on slot t by a particular choice of the control decision $I(t)$ (under the observed $\Omega(t)$ and $z(t)$). The intuition is that if the virtual queue $Y_m(t)$ is stable, then the time average rate of the “input process” $x_m(t)$ is less than or equal to the “service rate” x_m^{av} [21]. This turns the time average constraint into a simple queue stability problem.⁵

A. Lyapunov Drift

Define $\mathbf{Y}(t)$ as a vector of all virtual queues $Y_m(t)$ for $m \in \mathcal{M}$, and define $\Theta(t) \triangleq [\mathbf{Y}(t); \mathbf{Q}(t)]$ as the combined queue vector. Assume all queues are initially empty, so that $\Theta(0) = \mathbf{0}$. Define the following Lyapunov function:

$$L(\Theta(t)) \triangleq \frac{1}{2} \sum_{n \in \mathcal{N}} Q_n(t)^2 + \frac{1}{2} \sum_{m \in \mathcal{M}} Y_m(t)^2$$

⁵Note that $Y_m(t)$ can be viewed as a “generalized” queue, as the “service rate” x_m^{av} can be negative, as can the $x_m(t)$ value.

Suppose time t_g is a renewal event, and let T be the random time until the next renewal event. Define the *variable-slot conditional Lyapunov drift* $\Delta_T(\Theta(t_g))$ as follows:⁶

$$\Delta_T(\Theta(t_g)) \triangleq \mathbb{E} \{L(\Theta(t_g + T)) - L(\Theta(t_g)) \mid \Theta(t_g), z(t_g) = 0\} \quad (15)$$

The expectation in the drift definition above is with respect to the random renewal interval duration T , the random events that can take place over this interval, and the possibly random control actions $I(t)$ that are made during this interval. The explicit conditioning on $z(t_g) = 0$ in (15) will be suppressed in the remainder of this paper, as this conditioning is implied given that t_g is a renewal time.

It is important to note the following subtlety: While the actual drift is viewed over renewal times, the queue states $Q(t_g)$ at renewal events are not necessarily identical, and the implemented policy itself may not be stationary. Thus, actual system behavior is not necessarily i.i.d. over different renewal intervals. However, these times act as convenient “time-stamps” over which to analytically compare the Lyapunov drift of the implemented policy with the corresponding drifts of the (z, Ω) -only policies of Assumptions 1 and 2.

Lemma 1: (Lyapunov Drift) Under any network control policy for choosing $I(t)$ over time, the variable-slot conditional Lyapunov drift satisfies the following at any renewal time t_g and any $\Theta(t_g)$:

$$\Delta_T(\Theta(t_g)) \leq B + D(\Theta(t_g)) \quad (16)$$

where $D(\Theta(t_g))$ is defined:

$$D(\Theta(t_g)) \triangleq - \sum_{n \in \mathcal{N}} Q_n(t_g) \mathbb{E} \left\{ \sum_{\tau=0}^{T-1} d_n(t_g + \tau) \mid \Theta(t_g) \right\} - \sum_{m \in \mathcal{M}} Y_m(t_g) \mathbb{E} \left\{ T x_m^{av} - \sum_{\tau=0}^{T-1} x_m(t_g + \tau) \mid \Theta(t_g) \right\} \quad (17)$$

where we recall that:

$$d_n(t) \triangleq \hat{d}_n(I(t), \Omega(t), z(t)), \quad x_m(t) \triangleq \hat{x}_m(I(t), \Omega(t), z(t))$$

and where B is a finite constant defined:

$$B \triangleq \frac{\sigma^2(2 - \phi)}{2\phi^2} \quad (18)$$

where σ^2 is a constant that satisfies the following for all t :

$$\sigma^2 \geq \sum_{n \in \mathcal{N}} [\mu_n(t)^2 + R_n(t)^2] + \sum_{m \in \mathcal{M}} (x_m(t) - x_m^{av})^2$$

Note that σ^2 is finite because $x_m(t)$, $\mu_n(t)$, and $R_n(t)$ are uniformly bounded.

Proof: (Lemma 1) The proof follows by squaring the queue update equations (1) and (14) and using a multi-slot drift analysis, and is omitted for brevity (see [1]). ■

Let $V \geq 0$ be a non-negative parameter that we shall use to affect proximity to the optimal solution (with a tradeoff in convergence times and average queue congestion, as shown below). For pure feasibility problems, we set $V = 0$. As in [2][21] for the case of single-slot problems, our strategy in

⁶Note that proper notation for the drift should be $\Delta_T(\Theta(t_g), t_g)$, as the drift may result from a non-stationary policy and hence can depend on the starting time t_g , although we use the simpler notation $\Delta_T(\Theta(t_g))$ as a formal representation of the right hand side of (15).

this variable slot scenario is, upon every renewal event, to take control actions that minimize the following “drift-plus-penalty” expression:

$$D(\Theta(t_g)) + V \mathbb{E} \left\{ \sum_{\tau=0}^{T-1} x_0(t_g + \tau) \mid \Theta(t_g) \right\} \quad (19)$$

where T is the random time until the next renewal event. We call the policy that implements the solution to (19) every slot the *weighted stochastic shortest path policy*.

Given the queue backlogs $\Theta(t_g)$ at the start of the renewal time t_g , the expression (19) represents a sum of random drift and penalty terms (which depend on control actions) over the course of a renewal interval. Hence, controlling the system to minimize this sum amounts to solving a *weighted stochastic shortest path problem* over the renewal interval (see [5] for a treatment of the theory of stochastic shortest path problems). This generalizes the well known max-weight policies of [2][21][27]. Indeed, in [2][21][27] there is no $z(t)$ state and so “renewals” occur every slot and the shortest path problem reduces to a simple greedy control action that minimizes a weighted drift-plus-penalty term over one slot. In this generalization, the queue backlogs still act as weights, but the solution of the stochastic shortest path problem is not greedy and requires consideration of how an action at time t affects $z(t)$ in future slots $t \geq t_g$. Here we have defined renewals according to forced renewal events. Our work [1] considers other types of renewals, including renewals that are defined by any visitation to the $z(t) = 0$ state (including forced and unforced visitations), which is useful for feasibility problems as the renewal interval duration can be smaller than that of forced renewals.

B. Performance Theorem

For a given parameter $V \geq 0$, suppose we *approximately* implement the weighted stochastic shortest path rule (19). Specifically, suppose there are constants $C \geq 0$ and $\delta \geq 0$ such that on every renewal interval we observe the queue states $\Theta(t_g)$ and take actions that satisfy the following approximation:

$$D(\Theta(t_g)) + V \mathbb{E} \left\{ \sum_{\tau=0}^{T-1} x_0(t_g + \tau) \mid \Theta(t_g) \right\} \leq D^{ssp}(\Theta(t_g)) + V \mathbb{E} \left\{ \sum_{\tau=0}^{T-1} x_0^{ssp}(t_g + \tau) \mid \Theta(t_g) \right\} + C + \delta \sum_{n \in \mathcal{N}} Q_n(t_g) + \delta \sum_{m \in \mathcal{M}} Y_m(t_g) + V\delta \quad (20)$$

where $x_0(t)$ represents the penalty that is incurred by the implemented policy, $x_0^{ssp}(t)$ is the penalty that would be incurred under the stochastic shortest path solution to (19), and T is the renewal frame size (which is unaffected by control decisions and satisfies $\mathbb{E}\{T\} = 1/\phi$). Note that if the exact stochastic shortest path solution to (19) is used every renewal interval, we have $C = \delta = 0$.

Theorem 1: Suppose Assumptions 1 and 2 hold for a given $\epsilon > 0$. Fix a parameter $V > 0$. If there are constants

$C \geq 0$ and $\delta \geq 0$ such that (20) is satisfied for every renewal interval, and if δ is small enough so that $\epsilon > \phi\delta$, then $\bar{Q}_n < \infty$ and $\bar{Y}_m < \infty$ for all $n \in \mathcal{N}$ and $m \in \mathcal{M}$ (and consequently feasibility constraints (3)-(4) are satisfied). Furthermore, for renewal times t_g (for $g \in \{0, 1, 2, \dots\}$) and for any positive integer G we have:⁷

$$\frac{1}{G} \sum_{g=0}^{G-1} \left[\sum_{n \in \mathcal{N}} \mathbb{E}\{Q_n(t_g)\} + \sum_{m \in \mathcal{M}} \mathbb{E}\{Y_m(t_g)\} \right] \leq \frac{(B+C)\phi + V(\phi\delta + x_0^{max} - x_0^{min})}{\epsilon - \phi\delta} + \frac{\phi\mathbb{E}\{L(\Theta(0))\}}{G(\epsilon - \phi\delta)} \quad (21)$$

Finally, the time average penalty satisfies:

$$\limsup_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{x_0(\tau)\} \leq x_0^{opt} + \frac{(B+C)\phi}{V} + \phi\delta[1 + (x_0^{max} - x_0^{opt})/\epsilon] \quad (22)$$

and the right-hand side is also a bound on the average penalty over G renewal intervals divided by the average duration of G renewal intervals, for any positive G .

Note from (22) and (21) that the time average of $x_0(t)$ can be made arbitrarily close to (or below) $x_0^{opt} + \phi\delta[1 + (x_0^{max} - x_0^{opt})/\epsilon]$ as V is increased, with a tradeoff in average queue size that is linear in V . The value δ determines how close this performance is to the optimal value x_0^{opt} . In the case $\delta = 0$ (which holds, for example, if our approximation to the stochastic shortest path problem differs from the optimal solution only by a constant C that is independent of queue length), then the V parameter affects a $[O(1/V); O(V)]$ performance-delay tradeoff, as in [2], so that distance to optimality is $O(1/V)$ and hence can be made arbitrarily small, at the expense of an increase in the average backlog of the queues that is linear in V . This average backlog of the queues $Q(t)$ directly affects their average delay (via Little's Theorem), while the average backlog of the virtual queues $Y_m(t)$ affects the average convergence time required to achieve the performance guarantees (see also [28]). The theorem holds for possibly non-zero initial conditions $\Theta(0)$, and can often be extended using *place holder backlog* from [29] to improve the backlog of $Q_n(t)$ and $Y_m(t)$.

Proof: Let t_g be a renewal time, and let T be the renewal duration. From (20) and (16) we have:

$$\begin{aligned} \Delta_T(\Theta(t_g)) + V\mathbb{E}\left\{\sum_{\tau=0}^{T-1} x_0(t_g + \tau) \mid \Theta(t_g)\right\} &\leq B + C \\ + D^{ssp}(\Theta(t_g)) + V\mathbb{E}\left\{\sum_{\tau=0}^{T-1} x_0^{ssp}(t_g + \tau) \mid \Theta(t_g)\right\} & \\ + \delta \sum_{n \in \mathcal{N}} Q_n(t_g) + \delta \sum_{m \in \mathcal{M}} Y_m(t_g) + V\delta &(23) \end{aligned}$$

By definition of the weighted stochastic shortest path policy, we have:

$$\begin{aligned} D^{ssp}(\Theta(t_g)) + V\mathbb{E}\left\{\sum_{\tau=0}^{T-1} x_0^{ssp}(t_g + \tau) \mid \Theta(t_g)\right\} &\leq \\ D^*(\Theta(t_g)) + V\mathbb{E}\left\{\sum_{\tau=0}^{T-1} x_0^*(t_g + \tau) \mid \Theta(t_g)\right\} &(24) \end{aligned}$$

⁷Note from (18) that $B = O(1/\phi^2)$ and so $B\phi = O(1/\phi)$.

where $D^*(\Theta(t_g))$ and $x_0^*(t_g + \tau)$ correspond to any other control policy $I^*(t)$ that could be implemented over the renewal interval. Using (24) in (23) together with the definition of $D^*(\Theta(t_g))$ given in (17) yields:

$$\begin{aligned} \Delta_T(\Theta(t_g)) + V\mathbb{E}\left\{\sum_{\tau=0}^{T-1} x_0(t_g + \tau) \mid \Theta(t_g)\right\} &\leq B + C \\ + V\delta + V\mathbb{E}\left\{\sum_{\tau=0}^{T-1} x_0^*(t_g + \tau) \mid \Theta(t_g)\right\} & \\ - \sum_{n \in \mathcal{N}} Q_n(t_g)\mathbb{E}\left\{\sum_{\tau=0}^{T-1} d_n^*(t_g + \tau) - \delta \mid \Theta(t_g)\right\} & \\ - \sum_{m \in \mathcal{M}} Y_m(t_g)\mathbb{E}\left\{-\delta + \sum_{\tau=0}^{T-1} [x_m^{av} - x_m^*(t_g + \tau)] \mid \Theta(t_g)\right\} &(25) \end{aligned}$$

where we have used the following notation:

$$\begin{aligned} x_m^*(\tau) &= \hat{x}_m(I^*(\tau), \Omega(\tau), z^*(\tau)) \\ d_n^*(\tau) &= \hat{d}_n(I^*(\tau), \Omega(\tau), z^*(\tau)) \end{aligned}$$

Now choose $I^*(t)$ as the policy of Assumption 2 that yields (12)-(13). Plugging (12)-(13) directly into (25) and using the fact that $\mathbb{E}\{T\} = 1/\phi$ yields:

$$\begin{aligned} \Delta_T(\Theta(t_g)) + V\mathbb{E}\left\{\sum_{\tau=0}^{T-1} x_0(t_g + \tau) \mid \Theta(t_g)\right\} &\leq B + C \\ + V\delta + V\mathbb{E}\left\{\sum_{\tau=0}^{T-1} x_0^*(t_g + \tau) \mid \Theta(t_g)\right\} & \\ - \sum_{n \in \mathcal{N}} Q_n(t_g)\left(\frac{\epsilon}{\phi} - \delta\right) - \sum_{m \in \mathcal{M}} Y_m(t_g)\left(\frac{\epsilon}{\phi} - \delta\right) &(26) \end{aligned}$$

Using the bounds x_0^{min} and x_0^{max} in the above inequality and rearranging terms yields:

$$\begin{aligned} \Delta_T(\Theta(t_g)) + \sum_{n \in \mathcal{N}} Q_n(t_g)\left(\frac{\epsilon}{\phi} - \delta\right) + \sum_{m \in \mathcal{M}} Y_m(t_g)\left(\frac{\epsilon}{\phi} - \delta\right) & \\ \leq B + C + V(\delta + (x_0^{max} - x_0^{min})/\phi) & \end{aligned}$$

Taking expectations, summing over all renewal events t_g (for $g \in \{0, 1, 2, \dots, G-1\}$ and $t_0 = 0$) and using telescoping sums and non-negativity of $L(\cdot)$ (as in [2]) yields:

$$\begin{aligned} \frac{1}{G} \sum_{g=0}^{G-1} \left[\sum_{n \in \mathcal{N}} \mathbb{E}\{Q_n(t_g)\} + \sum_{m \in \mathcal{M}} \mathbb{E}\{Y_m(t_g)\} \right] &\leq \\ \frac{B + C + V(\delta + (x_0^{max} - x_0^{min})/\phi)}{\epsilon/\phi - \delta} + \frac{\mathbb{E}\{L(\Theta(0))\}}{G(\epsilon/\phi - \delta)} & \end{aligned}$$

This proves (21). This inequality also implies all queues are strongly stable [1], and hence by the stability theory in [21] we know that all constraints (3)-(4) are satisfied.

To prove (22) consider again the drift inequality (25), but now plug in the following (z, Ω) -only policy $I^*(t)$: Define probability $\theta \triangleq \delta\phi/\epsilon$. This is a valid probability because $\epsilon > \phi\delta$ by assumption. At each time t_g that marks the beginning of a renewal, independently flip a biased coin with probabilities θ and $1 - \theta$, and carry out one of the two following policies for the full duration of the renewal interval:

- With probability θ : Use the stationary randomized policy from Assumption 2 for the duration of the renewal interval, which yields (12)-(13).
- With probability $1 - \theta$: Use the stationary randomized policy from Assumption 1 for the duration of the renewal interval, which yields (9)-(11).

With this policy $I^*(t)$, from (9) and $\mathbb{E}\{T\} = 1/\phi$ we have:

$$\mathbb{E}\left\{\sum_{\tau=0}^{T-1} x_0^*(t_g + \tau)\right\} \leq \frac{\theta x_0^{max} + (1 - \theta)x_0^{opt}}{\phi} \quad (27)$$

We also have from (12)-(13) and (10)-(11):

$$\begin{aligned} \mathbb{E}\left\{\sum_{\tau=0}^{T-1} x_m^*(t_g + \tau)\right\} &\leq \frac{\theta(x_m^{av} - \epsilon) + (1 - \theta)x_m^{av}}{\phi} \\ \mathbb{E}\left\{\sum_{\tau=0}^{T-1} d_n^*(t_g + \tau)\right\} &\geq \frac{\theta\epsilon}{\phi} \end{aligned} \quad (28)$$

Plugging (27)-(29) into (25) and using the definition of $\theta = \delta\phi/\epsilon$ yields:

$$\begin{aligned} \Delta_T(\Theta(t_g)) + V\mathbb{E}\left\{\sum_{\tau=0}^{T-1} x_0(t_g + \tau) \mid \Theta(t_g)\right\} &\leq B + C \\ &+ V\delta + V\frac{1}{\phi}[\theta x_0^{max} + (1 - \theta)x_0^{opt}] \end{aligned}$$

The above holds for all times t_g that mark the beginning of renewal intervals. Defining $T = T_g$ (the duration of the g th renewal interval) and taking expectations of the above inequality yields:

$$\begin{aligned} \mathbb{E}\left\{L(\Theta(t_g + T_g)) - L(\Theta(t_g)) + V\sum_{\tau=0}^{T_g-1} x_0(t_g + \tau)\right\} &\leq \\ B + C + V\delta + \frac{V\delta(x_0^{max} - x_0^{opt})}{\epsilon} + \frac{Vx_0^{opt}}{\phi} \end{aligned}$$

Summing over $g \in \{0, \dots, G-1\}$, dividing by VG/ϕ , and using the fact that $L(\Theta(t_g)) \geq 0$ yields for any positive integer G :

$$\begin{aligned} \frac{\mathbb{E}\left\{\sum_{\tau=0}^{t_G-1} x_0(\tau)\right\}}{G/\phi} &\leq x_0^{opt} + \frac{\phi\mathbb{E}\{L(\Theta(0))\}}{VG} \\ &\frac{(B+C)\phi}{V} + \delta\phi + \frac{\delta\phi}{\epsilon}(x_0^{max} - x_0^{opt}) \end{aligned} \quad (30)$$

Because renewal intervals $\{T_g\}_{g=1}^{\infty}$ are i.i.d. geometric random variables with $\mathbb{E}\{T_g\} = 1/\phi$, we have by the law of large numbers that $t_G/G \rightarrow 1/\phi$ with probability 1. Using this, in [1] we show that the lim sup of the left hand side of (30) as $G \rightarrow \infty$ is the same as the left hand side of (22). ■

IV. SOLVING THE SHORTEST PATH PROBLEM

Consider now the stochastic shortest path problem given by expression (19). Here we describe its solution under the assumption that the state space \mathcal{Z} is finite. Without loss of generality, assume we start at time 0 and have (possibly non-zero) backlogs $\Theta = \Theta(0)$. Let T be the renewal interval size. For every step $\tau \in \{0, \dots, T-1\}$, define

$c_{\Theta}(I(\tau), \Omega(\tau), z(\tau))$ as the incurred cost assuming that the queue state at the beginning of the renewal is $\Theta(0)$:

$$\begin{aligned} c_{\Theta}(I(\tau), \Omega(\tau), z(\tau)) &\triangleq - \sum_{n \in \mathcal{N}} Q_n(0) \hat{d}_n(I(\tau), \Omega(\tau), z(\tau)) \\ &- \sum_{m \in \mathcal{M}} Y_m(0) [x_m^{av} - \hat{x}_m(I(\tau), \Omega(\tau), z(\tau))] \\ &+ V\hat{x}_0(I(\tau), \Omega(\tau), z(\tau)) \end{aligned}$$

Let $I^{ssp}(\tau)$ denote the optimal control action on slot τ for solving the stochastic shortest path problem, given that the controller first observes $\Omega(\tau)$ and $z(\tau)$. Define $\mathcal{Z}_r \triangleq \mathcal{Z} \cup \{r\}$, where we have added a new state “ r ” to represent the renewal state, which is the termination state of the stochastic shortest path problem. Appropriately adjust the probability transition matrix $P = (P_{zy}(I, \Omega))$ to account for this new state [5]. Define $\mathbf{J} = (J_z)_{z \in \mathcal{Z}_r}$ as a vector of optimal costs, where J_z is the minimum expected sum cost to the renewal state given that we start in state z , and $J_r = 0$. By basic dynamic programming theory [5], the optimal control action on each slot τ (given $\Omega(\tau)$ and $z(\tau)$) is:

$$I(\tau) = \arg \min_{I \in \mathcal{I}_{\Omega(\tau), z(\tau)}} [c_{\Theta}(I, \Omega(\tau), z(\tau)) + \sum_{y \in \mathcal{Z}_r} P_{z(\tau), y}(I, \Omega(\tau)) J_y] \quad (31)$$

This policy is easily implemented provided that the J_z values are known. It is well known that the \mathbf{J} vector satisfies the following vector dynamic programming equation:

$$\begin{aligned} \mathbf{J} &= \phi \mathbb{E} \left\{ \min_{I \in \mathcal{I}_{[\omega(t), 1], z}} c_{\Theta}^{(1)}(I, \omega(t)) \right\} + \\ (1 - \phi) \mathbb{E} \left\{ \min_{I \in \mathcal{I}_{[\omega(t), 0], z}} [c_{\Theta}^{(0)}(I, \omega(t)) + P^{(0)}(I, \omega(t)) \mathbf{J}] \right\} \end{aligned} \quad (32)$$

where we have used an entry-wise min (possibly with different I vectors being used for minimizing each entry $z \in \mathcal{Z}$). Thus, the notation $I \in \mathcal{I}_{\Omega(t), z}$ emphasizes that for a given $z \in \mathcal{Z}$, the control action I is chosen from the set $\mathcal{I}_{\Omega(t), z}$. Further, $c_{\Theta}^{(i)}(I, \omega(t))$ is defined as a vector with entries $c_{\Theta}^{(i)}(I, \Omega(t)) = c_{\Theta}(I, [\omega(t), i], z)$ that corresponds to $\phi(t) = i$ (for $i \in \{0, 1\}$), and $P^{(0)}(I, \omega(t)) = (P_{zy}(I, [\omega(t), 0]))$ is the probability transition matrix under $\Omega(t) = [\omega(t), 0]$ and control action I . The expectation in (32) is over the distribution of the i.i.d. process $\omega(t)$. We assume that the probability transition matrix $P^{(0)}(I, \omega(t))$ is known (it is a known 0/1 matrix in the delay-constrained example of Section II-B). We next show how to compute an approximation of \mathbf{J} based on random samples of $\omega(t)$ and using a classic Robbins-Monro iteration.

A. Estimation Through Random i.i.d. Samples

Suppose we have an infinite sequence of random variables arranged in batches with batch size L , with ω_{bi} denoting the i th sample of batch b . All random variables are i.i.d. with probability distribution the same as $\omega(t)$, and all are independent of the queue state Θ that is used for this stochastic shortest path problem. Consider the following two

mappings Ψ and $\tilde{\Psi}$ from a \mathbf{J} vector to another \mathbf{J} vector:

$$\Psi \mathbf{J} \triangleq \phi \mathbb{E} \left\{ \min_{I \in \mathcal{I}_{[\omega(t), 1], \mathcal{Z}}} \mathbf{c}_{\Theta}^{(1)}(I, \omega(t)) \right\} + (1 - \phi) \mathbb{E} \left\{ \min_{I \in \mathcal{I}_{[\omega(t), 0], \mathcal{Z}}} \left[\mathbf{c}_{\Theta}^{(0)}(I, \omega(t)) + P^{(0)}(I, \omega(t)) \mathbf{J} \right] \right\} \quad (33)$$

$$\tilde{\Psi} \mathbf{J} \triangleq \phi \frac{1}{L} \sum_{i=1}^L \min_{I \in \mathcal{I}_{[\omega_{bi}, 1], \mathcal{Z}}} \mathbf{c}_{\Theta}^{(1)}(I, \omega_{bi}) + (1 - \phi) \frac{1}{L} \sum_{i=1}^L \min_{I \in \mathcal{I}_{[\omega_{bi}, 0], \mathcal{Z}}} \left[\mathbf{c}_{\Theta}^{(0)}(I, \omega_{bi}) + P^{(0)}(I, \omega_{bi}) \mathbf{J} \right] \quad (34)$$

where the min is entrywise over each vector entry. The expectation in (33) is implicitly conditioned on a given Θ vector, and is with respect to the random $\omega(t)$ event that is independent of Θ . The mapping Ψ cannot be implemented without computing the expectation, whereas the mapping $\tilde{\Psi}$ can be implemented as a ‘‘simulation’’ over the L random samples ω_{bi} (assuming such samples can be generated or obtained). Note however that the expected value of $\tilde{\Psi} \mathbf{J}$ is exactly equal to $\Psi \mathbf{J}$. Thus, given an initial vector \mathbf{J}_b for use for step b (with some initial guess for \mathbf{J}_0 , such as $\mathbf{J}_0 = \mathbf{0}$), we can write $\tilde{\Psi} \mathbf{J}_b = \Psi \mathbf{J}_b + \boldsymbol{\eta}_b$, where $\boldsymbol{\eta}_b$ is a zero-mean random vector with $\mathbb{E} \{ \boldsymbol{\eta}_b | \mathbf{J}_b \} = \mathbf{0}$. For $b \in \{0, 1, 2, \dots\}$ we have the Robbins-Monro iteration:

$$\mathbf{J}_{b+1} = \gamma \tilde{\Psi} \mathbf{J}_b + (1 - \gamma) \mathbf{J}_b \quad (35)$$

where γ is a value such that $0 < \gamma < 1$, chosen to be suitably small to provide an accurate approximation, as specified in [1]. We can show that the number of independent samples and Robbins-Monro iterations required to satisfy the approximation guarantee (20) for a given $\delta > 0$ is *polynomial* in $N + K$ and in $1/\delta$ and $1/\phi$ (see [1]). However, each iteration (35) must be done for every entry of the \mathbf{J} vector, the size of which is equal to the cardinality of the set \mathcal{Z} (which is geometric in the number of delay-constrained queues K but independent of the number of delay-unconstrained queues N). This illustrates that we can solve problems with a very large number of delay-unconstrained queues, provided that the number of delay-constrained queues is small. It remains to show how such i.i.d. samples ω_{bi} can be found in such a way that they are *also* independent of the queue states $\Theta(t_g)$ that start renewals. This is done using $\omega(\tau)$ values observed on selected previous slots $\tau \leq t$. Analytically, this is justified by a *delayed queue analysis* given in [1], which shows that using out-of-date queue backlog does not limit optimality.

REFERENCES

- [1] M. J. Neely. Stochastic optimization for markov modulated networks with application to delay constrained wireless scheduling. *ArXiv Technical Report, arXiv:0905.4757v1*, May 2009.
- [2] 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.
- [3] S. Ross. *Introduction to Probability Models*. Academic Press, 8th edition, Dec. 2002.
- [4] E. Altman. *Constrained Markov Decision Processes*. Boca Raton, FL, Chapman and Hall/CRC Press, 1999.
- [5] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, Mass, 1996.
- [6] J. Abounadi, D. Bertsekas, and V. S. Borkar. Learning algorithms for markov decision processes with average cost. *SIAM Journal on Control and Optimization*, vol. 20, pp. 681-698, 2001.
- [7] S. Meyn. *Control Techniques for Complex Networks*. Cambridge University Press, 2008.
- [8] 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*, Maui, Hawaii, Dec. 2003.
- [9] 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, May 2007.
- [10] N. Salodkar, A. Borkar, 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.
- [11] L. Tassiulas and A. Ephremides. Dynamic server allocation to parallel queues with randomly varying connectivity. *IEEE Transactions on Information Theory*, vol. 39, pp. 466-478, March 1993.
- [12] E. M. Yeh. *Multiaccess and Fading in Communication Networks*. PhD thesis, Massachusetts Institute of Technology, Laboratory for Information and Decision Systems (LIDS), 2001.
- [13] A. Ganti, E. Modiano, and J. N. Tsitsiklis. Optimal transmission scheduling in symmetric communication models with intermittent connectivity. *IEEE Transactions on Information Theory*, vol. 53, no. 3, March 2007.
- [14] A. Fu, E. Modiano, and J. Tsitsiklis. Optimal energy allocation for delay-constrained data transmission over a time-varying channel. *Proc. IEEE INFOCOM*, 2003.
- [15] M. Goyal, A. Kumar, and V. Sharma. Power constrained and delay optimal policies for scheduling transmission over a fading channel. *Proc. IEEE INFOCOM*, April 2003.
- [16] C. C. Moallemi, S. Kumar, and B. Van Roy. Approximate and data-driven dynamic programming for queuing networks. Submitted for publication, 2008.
- [17] R. Berry and R. Gallager. Communication over fading channels with delay constraints. *IEEE Transactions on Information Theory*, vol. 48, no. 5, pp. 1135-1149, May 2002.
- [18] M. J. Neely. Optimal energy and delay tradeoffs for multi-user wireless downlinks. *IEEE Transactions on Information Theory*, vol. 53, no. 9, pp. 3095-3113, Sept. 2007.
- [19] M. J. Neely. Super-fast delay tradeoffs for utility optimal fair scheduling in wireless networks. *IEEE Journal on Selected Areas in Communications, Special Issue on Nonlinear Optimization of Communication Systems*, vol. 24, no. 8, pp. 1489-1501, Aug. 2006.
- [20] W. B. Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. John Wiley & Sons, 2007.
- [21] 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.
- [22] M. J. Neely, E. Modiano, and C. Li. Fairness and optimal stochastic control for heterogeneous networks. *Proc. IEEE INFOCOM*, March 2005.
- [23] A. Stolyar. Maximizing queueing network utility subject to stability: Greedy primal-dual algorithm. *Queueing Systems*, vol. 50, pp. 401-457, 2005.
- [24] A. Stolyar. Greedy primal-dual algorithm for dynamic resource allocation in complex networks. *Queueing Systems*, vol. 54, pp. 203-220, 2006.
- [25] R. Gallager. *Discrete Stochastic Processes*. Kluwer Academic Publishers, Boston, 1996.
- [26] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1995.
- [27] 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.
- [28] M. J. Neely. Distributed and secure computation of convex programs over a network of connected processors. *DCDIS Conf., Guelph, Ontario, July 2005*.
- [29] M. J. Neely and R. Urgaonkar. Opportunism, backpressure, and stochastic optimization with the wireless broadcast advantage. *Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, Oct. 2008*.