On Wireless Scheduling Algorithms for Minimizing the Queue-Overflow Probability

V. J. Venkataramanan and Xiaojun Lin, *†

October 4, 2008

Abstract

In this paper, we are interested in wireless scheduling algorithms for the downlink of a single cell that can minimize the queue-overflow probability. Specifically, in a large-deviation setting, we are interested in algorithms that maximize the asymptotic decay-rate of the queue-overflow probability, as the queue-overflow threshold approaches infinity. We first derive an upper bound on the decay-rate of the queue-overflow probability over all scheduling policies. We then focus on a class of scheduling algorithms collectively referred to as the "α-algorithms." For a given $\alpha \geq 1$, the α-algorithm picks the user for service at each time that has the largest product of the transmission rate multiplied by the backlog raised to the power α. We show that when the overflow metric is appropriately modified, the minimum-cost-to-overflow under the α-algorithm can be achieved by a simple linear path, and it can be written as the solution of a vector-optimization problem. Using this structural property, we then show that when α approaches infinity, the α-algorithms asymptotically achieve the largest decay-rate of the queue-overflow probability. Finally, this result enables us to design scheduling algorithms that are both close-to-optimal

*This work has been partially supported by NSF grants CNS-0626703, CNS-0643145, CNS-0721484, CCF-0635202 and a grant from Purdue Research Foundation. An earlier version of this paper has appeared in 45th Annual Allerton Conference on Communication, Control, and Computing, 2007 [1]

†The authors are with Center for Wireless Systems and Applications (CWSA) and School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47906 (email: {vvenkat,linx}@ecn.purdue.edu).

in terms of the asymptotic decay-rate of the overflow probability, and empirically shown to maintain small queue-overflow probabilities over queue-length ranges of practical interest.

1 Introduction

Link scheduling is an important functionality in wireless networks due to both the shared nature of the wireless medium and the variations of the wireless channel over time. In the past, it has been demonstrated that, by carefully choosing the scheduling decision based on the channel state and/or the demand of the users, the system performance can be substantially improved (see the references in [2]). Most studies of scheduling algorithms have focused on optimizing the long-term average throughput of the users. For example, in a typical stability problem [3–5], the goal is to find scheduling algorithms that can stabilize the network at given offered loads, which also ensures that the long-term average service rate is no less than the packet arrival rate at each user. An important result along this direction is the development of the so-called "throughput-optimal" algorithms [3]. An algorithm is called *throughput-optimal* if, at any offered load under which any other algorithm can stabilize the system, this algorithm can stabilize the system as well. Therefore, a throughput-optimal scheduling algorithm is optimal if we only impose stability constraints, i.e., it can stabilize the system over the largest set of offered loads.

While stability (and ensuring that the long-term service rate is no smaller than the arrival rate) is an important first-order metric of success, for many delay-sensitive applications it is far from sufficient. Note that a stability objective ensures that the queue-length (and thus the packet delay) do not increase to infinity. For real-time applications such as voice and video, we often need to ensure a stronger condition that the packet delay can be upper bounded with high probability. In this paper, we impose an alternate constraint on the probability of queue overflow, which is equivalent to a constraint on the delay-violation probability under certain conditions. In other words, we would like to guarantee that the probability of each user's backlog exceeding a given threshold is no greater than a target value.

We are interested in scheduling algorithms that are optimal subject to the above type of queue-overflow constraints. We focus on the downlink of a single cell in a cellular network. The base-station serves multiple users.

Due to interference, the base-station can only serve one user at a time. We assume that perfect channel information is available at the base-station. The question that we attempt to answer is the following: Is there an optimal algorithm in the sense that, at any given offered load, the algorithm can achieve the smallest probability of queue-overflow. Note that if we impose a quality-of-service (QoS) constraint on each user in the form of an upper bound on the queue-overflow probability, then the above optimality condition will also imply that the algorithm can support the largest set of offered loads subject to the QoS constraint.

The above question has well-known to be a difficult one. First, calculating the exact queue-distribution is often mathematically intractable. To make progress, one often has to use some asymptotic techniques, such as heavy-traffic limits [6–8] or large deviations*. To study small queue-overflow probabilities, it is natural to use the large-deviation theory because the probability of the event of interest is very small [11, 12]. In such a large deviation setting, one attempts to compute the asymptotic decay-rate of the queue-overflow probability, as the overflow-threshold approaches infinity. The optimal scheduling algorithm will then correspond to the algorithm that maximizes this decay-rate. Large-deviation theory has been successfully applied to wireline networks (see, e.g., [13–19]) and to wireless scheduling algorithms that only use the channel state to make the scheduling decisions [20–22]. However, when applied to wireless scheduling algorithms that use also the queue-length to make scheduling decisions (e.g., for the throughput-optimal scheduling policies proposed in [3–5]), this approach encounters a significant amount of technical difficulty. Specifically, in order to apply the large-deviation theory to queue-length-based scheduling algorithms, one has to use sample-path large-deviation, and formulate the problem as a multi-dimensional calculus-of-variations (CoV) problem for finding the "most likely path to overflow." The decay-rate of the queue-overflow probability then corresponds to the cost of this path, which is referred to as the "minimum cost to overflow." Unfortunately, for many queue-length-based scheduling algorithms of interest, this multi-dimensional calculus-of-variations problem is very difficult to solve. In the literature, only some restricted cases have been solved: Either restricted problem structures are assumed (e.g., symmetric users and ON-OFF channels [23]), or the size of

*Alternatively, one can focus on providing order-optimal bounds on the expected queue-length/packet delay [9, 10].

the system is very small (only two users) [24].

In a recent work [25]†, the author shows that the "exponential-rule" can maximize the decay-rate of the queue-overflow probability over all scheduling policies. In this paper, we build on the results of our preliminary work in [1], and show a comparable but different result. Specifically, we study a class of scheduling algorithms collectively referred to as the "α-algorithms." For a given $\alpha \geq 1$, the α-algorithm picks the user for service at each time that has the largest product of the transmission rate multiplied by the backlog raised to the power α. We show that when α approaches infinity, the α-algorithms asymptotically achieve the largest decay-rate of the queue-overflow probability. In our preliminary work [1], we establish this result assuming that a sample-path large-deviation principle (LDP) holds for the backlog process. Unfortunately, such a sample-path LDP appears to be difficult to verify. In this paper, we remove this assumption and prove our result using a different approach.

The advantage of working with the α-algorithms instead of the exponential-rule, is that the α-algorithms are scale-invariant (i.e., the outcome of the scheduling decision does not change if all queue-lengths are multiplied by a common factor). Hence, we can use the standard sample-path large-deviation principle (LDP), instead of the refined LDP used in [25] that is more technically-involved. In addition, our results highlight the role that the exponent α plays in determining the asymptotic decay-rate. To circumvent the difficulty of the multi-dimensional calculus-of-variations (CoV) problem, we apply a novel technique introduced in [26]. Specifically, we use a Lyapunov function to map the multi-dimensional CoV problem to a one-dimensional problem, which allows us to bound the minimum-cost-to-overflow by solutions of simple vector-optimization problems. Finally, using the insight of our main result, we design a scheduling algorithm that is both close-to-optimal in terms of the asymptotic decay-rate of the overflow probability, and empirically shown to maintain small queue-overflow probabilities over queue-length ranges of practical interest.

The rest of the paper is organized as follows. We first present the system model and the class of queue-length-based scheduling algorithms (referred to as α-algorithms) in Section 2. This is followed by Section 3 which discusses some of the mathematical preliminaries needed to apply a sample-path LDP. In Section 4, we derive an upper bound on the decay-rate of the

†Note that this work is published after our preliminary results reported in [1].

queue-overflow probability over all scheduling policies. Then in Section 5, we establish a lower bound on the decay-rate of the queue-overflow probability for α-algorithms and show that the bound is tight. In Section 6, we prove the main result that, as the parameter α approaches infinity, this class of scheduling algorithms asymptotically achieve the largest possible decay-rate. In Section 7, we present numerical results, and we discuss how to design a practical algorithm based on the insights from our main results. Then we conclude.

2 The System Model and Assumptions

We consider the downlink of a single cell in which a base-station serves N users. We assume a slotted system, and we assume that the state of the channel at each time slot is chosen *i.i.d* from one of M possible states. Let $C(t)$ denote the state of the channel at time $t = 1, 2, \ldots,$ and let $p_m = \mathbf{P}[C(t) = m], \quad m = 1, 2, \ldots, M$. Let $\boldsymbol{p} = [p_1, ..., p_M]$. We assume that the base-station can serve one user at a time. Let F_m^i denote the service rate for user i when it is picked for service and the channel state is m.

We assume that data for user i arrive as fluid at a constant rate λ_i. Let $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_N]$. Let $Q_i(t)$ denote the backlog of user i at time t, and let $\boldsymbol{Q}(t) = [Q_1(t), \ldots, Q_N(t)]$. In general, the decision of picking which user to serve is a function of the global backlog $\boldsymbol{Q}(t)$ and the channel state $C(t)$. Let $U(t)$ denote the index of the user picked for service at time t. The evolution of the backlog for each user i is then governed by

$$Q_i(t+1) = [Q_i(t) + \lambda_i - \sum_{m=1}^{M} \mathbf{1}_{\{C(t)=m, U(t)=i\}} F_m^i]^+ \tag{1}$$

where $[\cdot]^+$ denotes the projection to $[0, +\infty)$. Note that

$$\sum_{m=1}^{M} \sum_{i=1}^{N} \mathbf{1}_{\{C(t)=m, U(t)=i\}} = 1$$

since only one user can be served at a time.

A particular class of scheduling algorithms that we will focus on are collectively referred to as the "α-algorithms", where α is a parameter that takes values from the set of natural numbers. Given α, the behavior of the algorithm is as follows. When the backlog of the users is $\boldsymbol{Q}(t)$ and the state of the channel is $C(t) = m$, the algorithm chooses to serve the user i for which the product $Q_i^\alpha(t)F_m^i$ is the largest. If there are several users that achieve the largest $Q_i^\alpha(t)F_m^i$ together, one of them is chosen arbitrarily. It is well-known that this class of algorithms are throughput-optimal, i.e. they can stabilize the system at the largest set of offer-loads $\boldsymbol{\lambda}$ [3–5].

Consider the system when it is operated at a given offered load $\boldsymbol{\lambda}$ and is stable under a given scheduling algorithm. Specifically, we assume that there is a positive number $\acute{\epsilon} > 0$ such that $\boldsymbol{\lambda}(1+\acute{\epsilon})$ is in the capacity region of the system. This implies (refer [3]) that there exists $[\hat{\gamma}_m^i] \geq 0$ such that $\sum_{i=1}^N \hat{\gamma}_m^i = 1$ for all $m = 1, ..., M$ and

$$\lambda_i(1+\acute{\epsilon}) \leq \sum_{m=1}^M p_m \hat{\gamma}_m^i F_m^i \text{ for all } i = 1, ..., N. \tag{2}$$

In this paper, we are interested in the probability that the largest backlog exceeds a certain threshold B. i.e.,

$$\mathbf{P}[\max_{1 \leq i \leq N} Q_i(0) \geq B]. \tag{3}$$

Note that the probability in (3) is equivalent to a delay-violation probability when the arrival rates λ_i are constant, because the two types of events are related by (see [23, 27])

$$\mathbf{P}[\text{Delay at link } i \geq d_i] = \mathbf{P}[Q_i(0) \geq \lambda_i d_i].$$

The focus of this paper is in scheduling algorithms that minimize (3).

The problem of calculating the exact probability $\mathbf{P}[\max_{1 \leq i \leq N} Q_i(0) \geq B]$ is often mathematically intractable. In this work, we are interested in using large-deviation theory to compute estimates of this probability. Specifically, we will use the following limits:

$$I_0(\boldsymbol{\lambda}) \triangleq -\liminf_{B \to \infty} \frac{1}{B} \log \mathbf{P}[\max_{1 \leq i \leq N} Q_i(0) \geq B] \tag{4}$$

$$J_0(\boldsymbol{\lambda}) \triangleq -\limsup_{B \to \infty} \frac{1}{B} \log \mathbf{P}[\max_{1 \leq i \leq N} Q_i(0) \geq B]. \tag{5}$$

In essence, $I_0(\boldsymbol{\lambda})$ and $J_0(\boldsymbol{\lambda})$ are upper and lower bounds, respectively, of the decay rate of (3), as the overflow threshold B approaches infinity. In the following sections, we will show that no scheduling algorithm can have a decay-rate larger than a certain value I_{opt} (defined in Section 4), i.e. $I_0(\boldsymbol{\lambda}) \leq I_{\text{opt}}$. Then, we will show that the α-algorithms asymptotically achieve the decay-rate I_{opt}. In other words, for the α-algorithms, $J_0(\boldsymbol{\lambda})$ approaches I_{opt}, as $\alpha \to \infty$.

3 Preliminaries

Since the channel states are *i.i.d.* in time, the following sample-path large-deviation principle (LDP) holds for the channel-state process. Specifically, we define the empirical measure process $\boldsymbol{S}(t) = [S_m(t), m = 1, ..., M]$ as follows,

$$S_m(t) = \int_0^t \mathbf{1}_{\{C(\lfloor \tau \rfloor)=m\}} d\tau,$$

where $\lfloor \tau \rfloor$ represents the largest integer no greater than τ. Then, for any non-negative integer B, define the scaled channel-rate process

$$\boldsymbol{s}^B(t) = \frac{\boldsymbol{S}(Bt)}{B}. \tag{6}$$

It is easy to see that $\boldsymbol{s}^B(\cdot)$ is Lipschitz continuous and hence its derivative exists almost everywhere. For any given $T > 0$, let $\tilde{\Psi}_T$ denote the space of mappings from $[0, T]$ to $\mathbb{R}^M$, equipped with the essential supremum norm [12, p176, p352]. Let $\mathcal{P}_M$ denote the set of probability vectors of dimension M, i.e. $\boldsymbol{\phi} = [\phi_m, m = 1, ..., M] \in \mathcal{P}_M$ implies that $\boldsymbol{\phi} \geq 0$ and $\sum_{m=1}^M \phi_m = 1$. For any $\boldsymbol{\phi} \in \mathcal{P}_M$ define[‡]

$$H(\boldsymbol{\phi}||\boldsymbol{p}) = \sum_{m=1}^M \phi_m \log \frac{\phi_m}{p_m},$$

with the convention that $0 \log 0 = 0$. Then, as $B \to \infty$, it is well-known that the sequence of scaled channel-rate processes $\boldsymbol{s}^B(\cdot)$ on the interval $[0, T]$ satisfies a sample-path large-deviation principle (LDP) with good rate function [12, Mogulskii's Theorem (Thm 5.1.2), p176]:

$$I_s^T(\boldsymbol{s}(\cdot)) = \begin{cases} \int_0^T H(\dot{\boldsymbol{s}}(t)||\boldsymbol{p})dt, & \text{if } s(\cdot) \in \mathcal{AC} \\ \infty & \text{otherwise} \end{cases}$$

[‡]This is commonly known as the relative entropy between $\boldsymbol{\phi}$ and $\boldsymbol{p}$.

where $\mathcal{AC}$ denote the set of absolute continuous functions in $\tilde{\Psi}_T$. This LDP means that, for any set $\tilde{\Gamma}$ of trajectories in $\tilde{\Psi}_T$, the following inequality holds:

$$-\inf_{\boldsymbol{s}(\cdot)\in\tilde{\Gamma}^o} I_s^T(\boldsymbol{s}(\cdot)) \le \liminf_{B\to\infty} \frac{1}{B}\log \mathbf{P}[\boldsymbol{s}^B(\cdot)\in\tilde{\Gamma}]$$
$$\le \limsup_{B\to\infty} \frac{1}{B}\log \mathbf{P}[\boldsymbol{s}^B(\cdot)\in\tilde{\Gamma}] \le -\inf_{\boldsymbol{s}(\cdot)\in\bar{\tilde{\Gamma}}} I_s^T(\boldsymbol{s}(\cdot)), \quad (7)$$

where $\tilde{\Gamma}^o$ and $\bar{\tilde{\Gamma}}$ denote the interior and closure, respectively, of the set $\tilde{\Gamma}$. In addition, if $\tilde{\Gamma}$ is a *continuity* set [12, p5], the two bounds meet and we then have,

$$\lim_{B\to\infty} \frac{1}{B}\log \mathbf{P}[\boldsymbol{s}^B(\cdot)\in\tilde{\Gamma}] = -\inf_{\boldsymbol{s}(\cdot)\in\tilde{\Gamma}} I_s^T(\boldsymbol{s}(\cdot)). \quad (8)$$

Hence, the large-deviation rate-fucnction $I_s^T(\boldsymbol{s}(\cdot))$ characterizes how "rarely" the trajectory $\boldsymbol{s}(\cdot)$ occurs.

Using a similar scaling as $\boldsymbol{s}^B(\cdot)$, define the scaled backlog process

$$\boldsymbol{q}^B(t) = \frac{\boldsymbol{Q}(Bt)}{B}, \text{ for } t = 0, \tfrac{1}{B}, \tfrac{2}{B}, ..., \quad (9)$$

and by linear interpolation otherwise. Hence, for each $\boldsymbol{s}^B(\cdot)$ and a given initial condition $\boldsymbol{q}^B(0)$, we can use (1) to determine the corresponding $\boldsymbol{q}^B(\cdot)$. As $B\to\infty$, we will have a sequence of $\boldsymbol{s}^B(\cdot)$ and $\boldsymbol{q}^B(\cdot)$. It is easy to see that both $\boldsymbol{s}^B(\cdot)$ and $\boldsymbol{q}^B(\cdot)$ are Lipschitz-continuous. Hence, there must exist a subsequence that converges uniformly over the interval $[0,T]$. We use $(\boldsymbol{s}(\cdot),\boldsymbol{q}(\cdot))$ to denote such a limit, and we refer to it as a *fluid sample path*.

In essense, the goal of the rest of the paper is to use the known sample-path LDP of $\boldsymbol{s}^B(\cdot)$ to characterize that of $\boldsymbol{q}^B(\cdot)$ and that of the queue-overflow probability. In [1], we assume that a sample-path LDP also holds for $\boldsymbol{q}^B(\cdot)$. Unfortunately, such an assumption appears to be difficult to verify. Instead, in this paper we will use a different approach to establish the desirable results.

4 An Upper Bound on the Decay-Rate of the Overflow Probability

In this section, we first present an upper bound I_{opt} on $I_0(\boldsymbol{\lambda})$ (defined in (4)) under a given offered load $\boldsymbol{\lambda}$. This value I_{opt} bounds from above the decay-rate for the overflow probability of the stationary backlog process $\boldsymbol{Q}(t)$ over all scheduling policies. For every probability vector $\boldsymbol{\phi} \in \mathcal{P}_M$, define the following optimization problem:

$$\begin{aligned} w(\boldsymbol{\phi}) \triangleq \inf_{[\tilde{\gamma}_m^i]} \quad & \max_{1 \leq i \leq N} [\lambda_i - \sum_{m=1}^{M} \phi_m \tilde{\gamma}_m^i F_m^i]^+ \\ \text{subject to} \quad & \sum_{i=1}^{N} \tilde{\gamma}_m^i = 1 \text{ for all } m = 1, ..., M \\ & \tilde{\gamma}_m^i \geq 0 \text{ for all } i = 1, ..., N \text{ and } m. \end{aligned}$$

Here, $\tilde{\gamma}_m^i$ can be interpreted as some long-term fraction-of-time that user i is served when the channel state is m. Hence, if the channel-rate process is given by $\boldsymbol{s}(t) = \boldsymbol{\phi} t$, then $[\lambda_i - \sum_{m=1}^{M} \phi_m \tilde{\gamma}_m^i F_m^i]^+$ denotes the long-term growth-rate of the backlog of user i. Further, if all queues start empty, then $w(\boldsymbol{\phi})$ is the minimum rate of growth of the backlog of the largest queue.

Next, define I_{opt} as:

$$I_{\text{opt}} \triangleq \inf_{\{\boldsymbol{\phi} \in \mathcal{P}_M \mid w(\boldsymbol{\phi}) > 0\}} \frac{H(\boldsymbol{\phi} || \boldsymbol{p})}{w(\boldsymbol{\phi})}. \tag{10}$$

Given a fixed offered load $\boldsymbol{\lambda}$, assume that the backlog process $\boldsymbol{Q}(\cdot)$ under a given scheduling policy is stationary and ergodic. We will show the following result[§].

Proposition 1 *Under any scheduling policy, the following holds,*

$$\liminf_{B \to \infty} \frac{1}{B} \mathbf{P}(\max_{1 \leq i \leq N} Q_i(0) \geq B) \geq -I_{\text{opt}}. \tag{11}$$

In other words, I_{opt} is an upper bound for the decay-rate of the overflow probability over all scheduling policies. This upper bound, although in a different form, is equal to the one derived in [25].

Towards this end, we first show that the function $w(\cdot)$ provides a lower bound on the backlog of the largest queue, as proved in the following lemma.

Lemma 2 *For any $\epsilon > 0$, there exists $B_0 > 0$ such that for all $B \geq B_0$ and all scaled channel-rate process $\boldsymbol{s}^B(\cdot)$ (with $\boldsymbol{s}^B(0) = 0$), the following holds*

$$\max_{1 \leq i \leq N} q_i^B(T) \geq T w\left(\frac{\boldsymbol{s}^B(T)}{T}\right) - \epsilon, \quad \textit{for all } T > 0.$$

[§]Note Proposition 1 also holds trivially if the system is unstable.

Proof: Note that the queue backlog process is related to the channel-state process by Equation (1). Take the scaling in (6) and (9). Then, given $\boldsymbol{s}^B(\cdot)$, at any time t such that Bt is an integer, we must have,

$$q_i^B(t) \quad \geq \quad \left[\lambda_i t - \int_0^t \sum_{m=1}^{M} \dot{s}_m^B(\tau) \mathbf{1}_{\{U(\lfloor B\tau \rfloor)=i\}} F_m^i d\tau\right]^+ .$$

For any $T > 0$, there must exist a value of t such that Bt is an integer and $|t - T| \leq 1/B$. Hence, for any $\epsilon > 0$, there must exist $B_0 > 0$ such that for all $B \geq B_0$,

$$q_i^B(T) \geq \left[\lambda_i T - \int_0^T \sum_{m=1}^{M} \dot{s}_m^B(\tau) \mathbf{1}_{\{U(\lfloor B\tau \rfloor)=i\}} F_m^i d\tau\right]^+ - \epsilon.$$

Let $\phi_m = s_m^B(T)/T, m = 1, ..., M$. If $\phi_m > 0$, let

$$\tilde{\gamma}_m^i = \frac{1}{s_m^B(T)} \int_0^T \dot{s}_m^B(\tau) \mathbf{1}_{\{U(\lfloor B\tau \rfloor)=i\}} d\tau.$$

Otherwise, let $\tilde{\gamma}_m^1 = 1$ and $\tilde{\gamma}_m^i = 0$ for $i \geq 2$. We then have,

$$q_i^B(T) \geq T\left[\lambda_i - \sum_{m=1}^{M} \phi_m \tilde{\gamma}_m^i F_m^i\right]^+ - \epsilon.$$

Taking the maximum over all $i = 1, ..., N$, we have

$$\max_{1\leq i\leq N} q_i^B(T) \geq T \max_{1\leq i\leq N} [\lambda_i - \sum_{m=1}^{M} \phi_m \tilde{\gamma}_m^i F_m^i]^+ - \epsilon.$$

Finally, since $\sum_{i=1}^N \mathbf{1}_{\{U(\lfloor B\tau \rfloor)=i\}} = 1$, $[\tilde{\gamma}_m^i]$ is a feasible point for the optimization problem $w(\frac{\boldsymbol{s}^B(T)}{T})$. Thus, we obtain the lower bound that $\max_{1\leq i\leq N} q_i^B(T) \geq Tw(\frac{\boldsymbol{s}^B(T)}{T}) - \epsilon$. *Q.E.D.*

In addition, it is easy to show that the value of $w(\boldsymbol{\phi})$ is continuous with respect to $\boldsymbol{\phi}$ as stated below in Lemma 3. Let $\|\boldsymbol{\phi}\|$ denote the Euclidean norm of $\boldsymbol{\phi}$.

Lemma 3 *Let $\boldsymbol{\phi}^1$ and $\boldsymbol{\phi}^2$ be vectors from $\mathcal{P}_M$. The optimization problem $w(\cdot)$ is continuous in the sense that for any $\epsilon > 0$ and $\left\|\boldsymbol{\phi}^1 - \boldsymbol{\phi}^2\right\| < \epsilon$, the following holds,*

$$|w(\boldsymbol{\phi}^1) - w(\boldsymbol{\phi}^2)| \leq \epsilon \sum_{i=1}^{N} \sum_{m=1}^{M} F_m^i.$$

The intuition behind Lemma 3 comes from the fact that the function $\max\limits_{1\leq i\leq N}[\lambda_i - \sum_{m=1}^{M} \phi_m \tilde{\gamma}_m^i F_m^i]^+$ is continuous in $\boldsymbol{\phi}$ for any $[\tilde{\gamma}_m^i]$. The detailed proof is provided in Appendix A. We can now prove Proposition 1.

Proof: (of Proposition 1)

For any $\delta > 0$, we can find $\boldsymbol{\phi}_\delta$ from $\{\boldsymbol{\phi} \in \mathcal{P}_M \mid w(\boldsymbol{\phi}) > 0\}$ such that $\frac{H(\boldsymbol{\phi}_\delta \| \boldsymbol{p})}{w(\boldsymbol{\phi}_\delta)} < I_{\text{opt}} + \delta$. Define $\boldsymbol{s}_\delta(t) \triangleq t\boldsymbol{\phi}_\delta$ for $t \geq 0$. Let ϵ be some positive number and let $T = \frac{1+\epsilon+\epsilon\sum_{m=1}^{M}\sum_{i=1}^{N} F_m^i}{w(\boldsymbol{\phi}_\delta)}$. Let $B_T(\boldsymbol{s}_\delta(\cdot), \epsilon)$ be the set of functions in the space $\tilde{\Psi}^T$ such that $\sup_{t\in[0,T]} \|\boldsymbol{s}(t) - \boldsymbol{s}_\delta(t)\| < \epsilon$. Therefore, for any B, $\boldsymbol{s}^B(\cdot) \in B_T(\boldsymbol{s}_\delta(\cdot), \epsilon)$ implies $\left\|\frac{\boldsymbol{s}^B(T)}{T} - \boldsymbol{\phi}_\delta\right\| < \frac{\epsilon}{T}$. By Lemma 3, this in turn implies that

$$Tw(\frac{\boldsymbol{s}^B(T)}{T}) \geq Tw(\boldsymbol{\phi}_\delta) - \epsilon \sum_{m=1}^{M} \sum_{i=1}^{N} F_m^i. \tag{12}$$

Now, using Lemma 2, for all $B > B_0$ and $\boldsymbol{s}^B(\cdot)$, we have $\max\limits_{1\leq i\leq N} q_i^B(T) \geq Tw(\frac{\boldsymbol{s}^B(T)}{T}) - \epsilon$. Hence, by (12) we conclude that, for all $B > B_0$ and $\boldsymbol{s}^B(\cdot) \in B_T(\boldsymbol{s}_\delta(\cdot), \epsilon)$, we have,

$$\max_{1\leq i\leq N} q_i^B(T) \geq Tw(\boldsymbol{\phi}_\delta) - \epsilon - \epsilon \sum_{m=1}^{M} \sum_{i=1}^{N} F_m^i = 1, \tag{13}$$

where equality holds by the definition of T. Therefore,

$$\begin{aligned}
\mathbf{P}(\max_{1\leq i\leq N} Q_i(0) \geq B) &= \mathbf{P}(\max_{1\leq i\leq N} Q_i(BT) \geq B) \\
&= \mathbf{P}(\max_{1\leq i\leq N} q_i^B(T) \geq 1) \\
&\geq \mathbf{P}(\boldsymbol{s}^B(\cdot) \in B_T(\boldsymbol{s}_\delta(\cdot), \epsilon)).
\end{aligned}$$

By the LDP for $\boldsymbol{s}^B(\cdot)$ (see Inequality (7)), we then have

$$\begin{aligned}
&\liminf_{B\to\infty} \frac{1}{B} \log \mathbf{P}(\max_{1\le i\le N} Q_i(0) \ge B) \\
&\qquad \ge \liminf_{B\to\infty} \frac{1}{B} \log \mathbf{P}[\boldsymbol{s}^B(\cdot) \in B_T(\boldsymbol{s}_\delta(\cdot), \epsilon)] \\
&\qquad \ge - \inf_{\boldsymbol{s}(\cdot)\in B_T(\boldsymbol{s}_\delta(\cdot),\epsilon)} \int_0^T H(\dot{\boldsymbol{s}}(t)||\boldsymbol{p})dt \\
&\qquad \ge - \int_0^T H(\dot{\boldsymbol{s}}_\delta(t)||\boldsymbol{p})dt \\
&\qquad = -TH(\boldsymbol{\phi}_\delta||\boldsymbol{p}) \\
&\qquad \ge -(1+\epsilon+\epsilon\sum_{m=1}^{M}\sum_{i=1}^{N} F_m^i)(I_{\text{opt}}+\delta).
\end{aligned}$$

Since δ and ϵ can be arbitrarily small, we conclude that

$$\liminf_{B\to\infty} \frac{1}{B} \mathbf{P}(\max_{1\le i\le N} Q_i(0) \ge B) \ge -I_{\text{opt}}.$$

Q.E.D.

5 A Lower Bound on the Decay-Rate of the Overflow Probability for α-Algorithms

In this section we will develop a lower-bound on the decay-rate of a modified overflow-probability for the α-algorithms. Then, in the next section, we will use this result to show that, as $\alpha \to \infty$, the α-algorithms asymptotically achieve the maximum decay-rate I_{opt} of the queue-overflow probability.

Throughout this section, we will use the following modified queue-overflow event $\{V_\alpha(\boldsymbol{q}^B(t)) \ge 1\}$, where $V_\alpha(\boldsymbol{q}) \triangleq (\sum_{i=1}^N (q_i)^{\alpha+1})^{\frac{1}{\alpha+1}}$. Note that this overflow-event is different from the queue-overflow event $\{\max_{1\le i\le N} q_i^B(t) \ge 1\}$ that is used in earlier sections. The main intuition is that $V_\alpha(\cdot)$ is the Lyapunov function of the system operated under the α-algorithm. Hence, we can use the technique of [26] to easily characterize the most-likely path to overflow. On the other hand, as $\alpha \to \infty$, the difference between the two overflow-events diminishes, which allows us to bound the overflow probability $\mathbf{P}[\max_{1 \le i \le N} q_i^B(t) \ge 1]$ in Section 6.

5.1 A General Lower Bound

We first provide a lower-bound that relates the decay-rate of the overflow probability to the "minimum-cost-to-overflow" among all fluid sample paths. For ease of exposition, instead of considering the stationary system, we consider a system that starts at time 0 (although the results can also be extended to the stationary system as we will comment later). Specifically, let $\boldsymbol{Q}(0) = 0$. Let $\mathbf{P}_0$ denote the probability measure conditioned on $\boldsymbol{Q}(0) = 0$. For any $T > 0$, let $\hat{\Gamma}_T$ denote the set of fluid sample paths $(\boldsymbol{s}(\cdot), \boldsymbol{q}(\cdot))$ on the interval $[0, T]$ such that $\boldsymbol{q}(0) = 0$ and $V_\alpha(\boldsymbol{q}(T)) \ge 1$. We then have the following lower-bound, which is comparable to Theorem 7.1 of [25] although we do not need to use the refined LDP.

Proposition 4 *Consider $\hat{\Gamma}_T$ as defined earlier. Then, the following holds:*

$$\begin{aligned} & \limsup_{B \to \infty} \frac{1}{B} \log \mathbf{P}_0[V_\alpha(\boldsymbol{q}^B(T)) \ge 1] \\ \le \quad & - \inf_{(\boldsymbol{s}(\cdot), \boldsymbol{q}(\cdot)) \in \hat{\Gamma}_T} \int_0^T H(\dot{\boldsymbol{s}}(t) || \boldsymbol{p}) dt. \end{aligned} \tag{14}$$

Remark: The infimum on the right-hand-side of (14) is often called the "minimum cost to overflow." This result reflects the well-celebrated large-deviation philosophy that "rare events occur in the most likely way." Specifically, Proposition 4 states that the probability of queue overflow is determined mostly by the smallest cost among all fluid sample paths that overflow. This fluid sample path is often referred to as the "most likely path to overflow."

Proof: Fix $T > 0$. Recall that we have set $\boldsymbol{q}^B(0) = 0$ for all B. Let $\tilde{\Gamma}^B$ be the set of channel rate processes $\boldsymbol{s}^B(\cdot)$ such that the corresponding backlog process satisfies $V_\alpha(\boldsymbol{q}^B(T)) \ge 1$. For all $n \ge 1$, we have

$$\limsup_{B \to \infty} \frac{1}{B} \log \mathbf{P}[\boldsymbol{s}^B(\cdot) \in \tilde{\Gamma}^B] \tag{15}$$

$$\le \quad \limsup_{B \to \infty} \frac{1}{B} \log \mathbf{P}[\boldsymbol{s}^B(\cdot) \in \cup_{\acute{B}=n}^{\infty} \tilde{\Gamma}^{\acute{B}}]. \tag{16}$$

By the LDP for $\boldsymbol{s}^B(\cdot)$ (see (7)), we have

$$\limsup_{B\to\infty}\frac{1}{B}\log\mathbf{P}[\boldsymbol{s}^B(\cdot)\in\cup_{\acute{B}=n}^{\infty}\tilde{\Gamma}^{\acute{B}}]\leq$$
$$-\inf_{\boldsymbol{s}(\cdot)\in\overline{\cup_{B=n}^{\infty}\tilde{\Gamma}^B}}\int_0^T H(\dot{\boldsymbol{s}}(t)||\boldsymbol{p})dt.$$

Note that the sequence of sets $\cup_{B=n}^{\infty}\tilde{\Gamma}^B$ is decreasing in n, we therefore have

$$\limsup_{B\to\infty}\frac{1}{B}\log\mathbf{P}[\boldsymbol{s}^B(\cdot)\in\tilde{\Gamma}^B]\leq$$
$$-\lim_{n\to\infty}\inf_{\boldsymbol{s}(\cdot)\in\overline{\cup_{B=n}^{\infty}\tilde{\Gamma}^B}}\int_0^T H(\dot{\boldsymbol{s}}(t)||\boldsymbol{p})dt. \tag{17}$$

It remains to show that the right-hand-side of (17) is no greater than that of (14). For each n, we can find $\boldsymbol{y}_n(\cdot)\in\overline{\cup_{B=n}^{\infty}\tilde{\Gamma}^B}$ such that

$$\int_0^T H(\dot{\boldsymbol{y}}_n(t)||\boldsymbol{p})dt<\inf_{\boldsymbol{s}(\cdot)\in\overline{\cup_{B=n}^{\infty}\tilde{\Gamma}^B}}\int_0^T H(\dot{\boldsymbol{s}}(t)||\boldsymbol{p})dt+\frac{1}{n}. \tag{18}$$

Since $\boldsymbol{y}_n(\cdot)$ is equicontinuous, we can find a subsequence that converges uniformly on $[0,T]$. For ease of exposition, we slightly abuse notation and denote this subsequence by $\boldsymbol{y}_n(\cdot)$. Let $\boldsymbol{y}^*(\cdot)$ denote its limit, i.e., $\lim_{n\to\infty}\boldsymbol{y}_n(\cdot)=\boldsymbol{y}^*(\cdot)$. Since the cost function $\int_0^T H(\cdot||\boldsymbol{p})dt$ is lower semi-continuous, we have

$$\liminf_{n\to\infty}\int_0^T H(\dot{\boldsymbol{y}}_n(t)||\boldsymbol{p})dt\geq\int_0^T H(\dot{\boldsymbol{y}}^*(t)||\boldsymbol{p})dt. \tag{19}$$

For each $\boldsymbol{y}_n(\cdot)$, since it belongs to the closure of $\cup_{B=n}^{\infty}\tilde{\Gamma}^B$, we can find a sequence $\boldsymbol{y}_{n,m}(\cdot)\in\cup_{B=n}^{\infty}\tilde{\Gamma}^B, m=1,2,...$ such that $\boldsymbol{y}_n(\cdot)=\lim_{m\to\infty}\boldsymbol{y}_{n,m}(\cdot)$. Then from all $\boldsymbol{y}_{n,m}(\cdot), n=1,2,...,m=1,2,...,$ we can find another sequence $\boldsymbol{y}_{n,m_n}(\cdot), n=1,2,...$ such that $\lim_{n\to\infty}\boldsymbol{y}_{n,m_n}(\cdot)=\boldsymbol{y}^*(\cdot)$. (For example, we can let $m_1=1$. Then, given m_n, we can choose m_{n+1} such that $\sup_{\{t\in[0,T]\}}\left\|\boldsymbol{y}_{n+1,m_{n+1}}(t)-\boldsymbol{y}_{n+1}(t)\right\|<\frac{\sup_{\{t\in[0,T]\}}\left\|\boldsymbol{y}_{n,m_n}(t)-\boldsymbol{y}_n(t)\right\|}{2}$.) For notational convenience, let $\acute{\boldsymbol{y}}_n(\cdot)$ denote the sequence $\boldsymbol{y}_{n,m_n}(\cdot)$ from here on.

For each n, let $\acute{\boldsymbol{q}}_n(\cdot)$ be the backlog process corresponding to the channel rate process $\acute{\boldsymbol{y}}_n(\cdot)$. By construction, $\acute{\boldsymbol{q}}_n(0)=0$ and $V_\alpha(\acute{\boldsymbol{q}}_n(T))\geq 1$ for all n.

Since the backlog processes are equicontinuous, we can find a subsequence of $(\acute{\boldsymbol{y}}_n, \acute{\boldsymbol{q}}_n)$ such that this subsequence converges to $(\boldsymbol{y}^*(\cdot), \boldsymbol{q}^*(\cdot))$ uniformly over the interval $[0, T]$, where $\boldsymbol{q}^*(\cdot)$ satisfies $\boldsymbol{q}^*(0) = 0$ and $V_\alpha(\boldsymbol{q}^*(T)) \geq 1$. Therefore, $(\boldsymbol{y}^*(\cdot), \boldsymbol{q}^*(\cdot))$ is in $\hat{\Gamma}_T$ and thus

$$\int_0^T H(\dot{\boldsymbol{y}}^*(t)||\boldsymbol{p})dt \quad \geq \inf_{(\boldsymbol{s}(\cdot),\boldsymbol{q}(\cdot))\in\hat{\Gamma}_T} \int_0^T H(\dot{\boldsymbol{s}}(t)||\boldsymbol{p})dt.$$

Combining with (18) and (19), we conclude that

$$\begin{aligned} & \lim_{n\to\infty} \underline{\inf_{\boldsymbol{s}(\cdot)\in\cup_{B=n}^{\infty}\tilde{\Gamma}^B}} \int_0^T H(\dot{\boldsymbol{s}}(t)||\boldsymbol{p})dt \\ \geq \;& \liminf_{n\to\infty} \int_0^T H(\dot{\boldsymbol{y}}_n(t)||\boldsymbol{p})dt \\ \geq \;& \inf_{(\boldsymbol{s}(\cdot),\boldsymbol{q}(\cdot))\in\hat{\Gamma}_T} \int_0^T H(\dot{\boldsymbol{s}}(t)||\boldsymbol{p})dt. \end{aligned}$$

This along with (17) proves the proposition. *Q.E.D.*

5.2 Bounding the Minimum-Cost-to-Overflow Through Lyapunov Functions

Finding the minimum-cost to overflow in (14) is a multi-dimensional calculus-of-variations problem, which is often very difficult [23,24,28]. In this section, we first use the idea of [26] to show another much simpler lower bound (Proposition 6). We will exploit the fact that V_α is a Lyapunov function of the system operated under the α-algorithm. We will then show that this lower bound is indeed equal to the minimum-cost to overflow, and it can be attained by a simple linear trajectory.

We begin with a result that characterizes the relationship between $V_\alpha(\boldsymbol{q}(\cdot))$ and the channel-rate process $\boldsymbol{s}(\cdot)$.

Proposition 5 *Let $(\boldsymbol{s}(\cdot), \boldsymbol{q}(\cdot))$ be any fluid sample path. Except for a set $\mathcal{T}_0$ of measure zero, at any time $t \notin \mathcal{T}_0$ and $\boldsymbol{q}(t) \neq 0$, the drift of the Lyapunov* *function* $V_\alpha(\boldsymbol{q}(t))$ *is given by*

$$\dot{V}_\alpha(\boldsymbol{q}(t)) = \left(\sum_{i=1}^{N}(q_i(t))^{\alpha+1}\right)^{\frac{-\alpha}{\alpha+1}} \left[\sum_{i=1}^{N}(q_i(t))^{\alpha}\lambda_i - \sum_{m=1}^{M}\dot{s}_m(t)\max_{1\le k\le N}((q_k(t))^{\alpha}F_m^k)\right]. \quad (20)$$

The proof is provided in Appendix B.

Remark: An intuitive way to understand Proposition 5 is as follows. From (1), if we take the scaling in (6) and (9) and let $B \to \infty$, we would expect that the limiting fluid sample path will follow an ordinary differential equation as follows: There exists $\tilde{\gamma}_m^i(t), i = 1, ..., N, m = 1, ..., M$ such that

$$\dot{q}_i(t) = \lambda_i - \sum_{m=1}^{M}\dot{s}_m(t)\tilde{\gamma}_m^i(t)F_m^i$$

if $q_i(t) > 0$ or $\lambda_i - \sum_{m=1}^{M}\dot{s}_m(t)\tilde{\gamma}_m^i(t)F_m^i \ge 0$; $\dot{q}_i(t) = 0$, otherwise; and $[\tilde{\gamma}_m^i(t)]$ are non-negative and satisfy

$$\sum_{i=1}^{N}\tilde{\gamma}_m^i(t) = 1 \text{ for all } m = 1, ..., M, \quad (21)$$

$$\tilde{\gamma}_m^i(t) = 0 \text{ whenever } (q_i(t))^{\alpha}F_m^i < \max_{1\le k\le N}(q_k(t))^{\alpha}F_m^k.$$

The variables $\tilde{\gamma}_m^i(t)$ can be viewed as the fraction of time that user i is served when channel state is m, in an infinitesimal interval immediately after t. Then, using the definition of $V_\alpha(\cdot)$, at any time t when $\boldsymbol{q}(t)$ is differentiable, we must have

$$\begin{aligned}\dot{V}_\alpha(\boldsymbol{q}(t)) &= \left(\sum_{i=1}^{N}(q_i(t))^{\alpha+1}\right)^{\frac{-\alpha}{\alpha+1}} \\ &\quad \left[\sum_{i=1}^{N}(q_i(t))^{\alpha}\left(\lambda_i - \sum_{m=1}^{M}\dot{s}_m(t)\tilde{\gamma}_m^i(t)F_m^i\right)\right] \\ &= \left(\sum_{i=1}^{N}(q_i(t))^{\alpha+1}\right)^{\frac{-\alpha}{\alpha+1}} \left[\sum_{i=1}^{N}(q_i(t))^{\alpha}\lambda_i \right. \\ &\quad \left. - \sum_{m=1}^{M}\dot{s}_m(t)\sum_{i=1}^{N}(q_i(t))^{\alpha}\tilde{\gamma}_m^i(t)F_m^i\right].\end{aligned}$$

Using (21), Equation (20) then follows. In Appendix B , we provide the proof of Proposition 5, which makes this argument more precise.

Next, for any $\boldsymbol{\phi} \in \mathcal{P}_M$, let $\boldsymbol{x} = [x_i, i = 1, ..., N]$, and let

$$\begin{aligned} a(\boldsymbol{\phi}) = \max_{\boldsymbol{x} \geq 0} \quad & \left[\sum_{i=1}^{N} x_i^{\alpha} \lambda_i - \sum_{m=1}^{M} \phi_m \max_{1 \leq k \leq N} (x_k^{\alpha} F_m^k) \right] \\ \text{subject to} \quad & \sum_{i=1}^{N} x_i^{\alpha+1} \leq 1. \end{aligned} \tag{22}$$

We will show soon that the Lyapunov drift on the right-hand-side of (20) must be no larger than $a(\dot{\boldsymbol{s}}(t))$. Further, let

$$J_\alpha \triangleq \inf_{\{\boldsymbol{\phi} \in \mathcal{P}_M \mid a(\boldsymbol{\phi}) > 0\}} \frac{H(\boldsymbol{\phi} || \boldsymbol{p})}{a(\boldsymbol{\phi})}. \tag{23}$$

Then intuitively, J_α can be interpreted as a lower bound on unit cost to raise $V_\alpha(\boldsymbol{q}(t))$. In order to overflow, we must raise $V_\alpha(\boldsymbol{q}(t))$ from 0 to 1. Hence, J_α should be a lower bound on the minimum-cost to overflow, which is indeed the case as we show in the following proposition.

Proposition 6 *For any $T > 0$, the following holds,*

$$\limsup_{B \to \infty} \frac{1}{B} \log \mathbf{P}_0[V_\alpha(\boldsymbol{q}^B(T)) \geq 1] \leq -J_\alpha. \tag{24}$$

Remark: Note that the event $V_\alpha(\boldsymbol{q}^B(T)) \geq 1$ is equivalent to $V_\alpha(\boldsymbol{Q}(BT)) \geq B$. As $T \to \infty$, we would expect that the probability $\mathbf{P}_0[V_\alpha(\boldsymbol{q}^B(T)) \geq 1]$ approaches the stationary overflow probability $\mathbf{P}[V_\alpha(\boldsymbol{q}^B(0)) \geq 1]$. Since J_α is independent of T, we would then expect that J_α becomes a lower bound for the decay rate of the stationary overflow probability, i.e.

$$\limsup_{B \to \infty} \frac{1}{B} \log \mathbf{P}[V_\alpha(\boldsymbol{q}^B(0)) \geq 1] \leq -J_\alpha.$$

This convergence can indeed be shown using the so-called Freidlin-Wentzell theory [11, 25]. However, the details are quite technical. Interested readers can refer to Appendix D for the details.

Proof: (of Proposition 6) Fix $T > 0$. Recall the definition of $\hat{\Gamma}_T$ in Section 5.1. For any fluid sample path $(\boldsymbol{s}(\cdot), \boldsymbol{q}(\cdot))$ in $\hat{\Gamma}_T$ (which overflows at time T), we will show that the cost of the path $\int_0^T H(\dot{\boldsymbol{s}}(t)||\boldsymbol{p})dt$ is at least J_α. The result of the proposition then follows from Proposition 4. Towards this end, note that since the backlog process $\boldsymbol{q}(\cdot)$ is Lipschitz-continuous, it is differentiable almost everywhere. According to Proposition 5, for any t such that $t \notin \mathcal{T}_0$ and $\boldsymbol{q}(t) \neq 0$, we must have,

$$\begin{aligned} \dot{V}_\alpha(\boldsymbol{q}(t)) &= \left(\sum_{i=1}^{N}(q_i(t))^{\alpha+1}\right)^{\frac{-\alpha}{\alpha+1}} \left[\sum_{i=1}^{N}(q_i(t))^{\alpha}\lambda_i \right. \\ & \qquad \left. - \sum_{m=1}^{M} \dot{s}_m(t) \max_{1\le k\le N}((q_k(t))^{\alpha} F_m^k)\right] \\ &= \sum_{i=1}^{N} \tilde{q}_i^{\alpha}\lambda_i - \sum_{m=1}^{M} \dot{s}_m(t) \max_{1\le k\le N}(\tilde{q}_k^{\alpha} F_m^k) \end{aligned}$$

where

$$\tilde{q}_i = q_i(t) \left[\sum_{i=1}^{N}(q_i(t))^{\alpha+1}\right]^{-\frac{1}{\alpha+1}}, \text{ i = 1,...,N.}$$

Since $\sum_{i=1}^{N} \tilde{q}_i^{\alpha+1} = 1$, $\tilde{\boldsymbol{q}} = [\tilde{q}_i]$ is a feasible point that satisfies the constraint in (22). We then have

$$\dot{V}_\alpha(\boldsymbol{q}(t)) \le a(\dot{\boldsymbol{s}}(t)).$$

Hence, if $\dot{V}_\alpha(\boldsymbol{q}(t)) > 0$, we must have $a(\dot{\boldsymbol{s}}(t)) > 0$. Then, using the definition of J_α in (23), we have

$$H(\dot{\boldsymbol{s}}(t)||\boldsymbol{p}) \ge J_\alpha \dot{V}_\alpha(\boldsymbol{q}(t)).$$

On the other hand, if $\dot{V}_\alpha(\boldsymbol{q}(t)) \le 0$, the above inequality also holds trivially. Hence, the cost of the path must satisfy

$$\int_0^T H(\dot{\boldsymbol{s}}(t)||\boldsymbol{p})dt \quad \ge \quad J_\alpha \int_0^T \dot{V}_\alpha(\boldsymbol{q}(t))dt.$$

Recall that any fluid sample path in $\hat{\Gamma}_T$ must satisfy $\boldsymbol{q}(0) = 0$ and $V_\alpha(\boldsymbol{q}(T)) \ge 1$. Hence,

$$\int_0^T \dot{V}_\alpha(\boldsymbol{q}(t))dt \ge 1.$$

The result of the proposition then follows. *Q.E.D.*

Remark: We briefly comment on why it is critical to use a Lyapunov function in the above procedure. Although a result similar to Proposition 6 could also be derived if we replace $V_\alpha(\cdot)$ by any function of $\boldsymbol{q}(t)$, such a result is only useful when the lower bound J_α is positive (otherwise the bound is trivial). The fact that $V_\alpha(\cdot)$ is a Lyapunov function is the key to ensure this property. To see this, note that if $\boldsymbol{\phi} = \boldsymbol{p}$, then the drift of the Lyapunov function will be negative for any $\boldsymbol{q}(t)$ (which is required for the stability of the system), implying that the value of $a(\boldsymbol{p}) = 0$. Hence, for the constraint in (23) to be satisfied, $\boldsymbol{\phi}$ must be away from $\boldsymbol{p}$. As a result, the objective function of (23) must be positive. We will see soon that this then implies that the infimum in (23) is also positive.

5.3 The Path-to-Overflow That Attains the Lower Bound J_α

In this subsection, we further simplify J_α, and then show that J_α is equal to the minimum-cost to overflow in (14). We define the following optimization problem. Let $\boldsymbol{y} = [y_1, ..., y_N]$. For any $\boldsymbol{\phi} \in \mathcal{P}_M$, define

$$
\begin{aligned}
w_\alpha(\boldsymbol{\phi}) = \min_{\boldsymbol{y}\geq 0, [\tilde{\gamma}_m^i]\geq 0} \quad & V_\alpha(\boldsymbol{y}) \\
\text{subject to} \quad & y_i = [\lambda_i - \sum_{m=1}^{M} \phi_m \tilde{\gamma}_m^i F_m^i]^+ \text{ for all } i \\
& \sum_{i=1}^{N} \tilde{\gamma}_m^i = 1 \text{ for all } m = 1, ..., M.
\end{aligned}
$$

Note that $w_\alpha(\boldsymbol{\phi})$ is analogous to $w(\boldsymbol{\phi})$ defined in section 4. Again, $\tilde{\gamma}_m^i$ can be interpreted as some long-term fraction-of-time that user i is served when the channel state is m. Hence, if the channel-rate process is given by $\boldsymbol{s}(t) = \boldsymbol{\phi}t$, then y_i denotes the long-term growth-rate of the backlog of user i. Further, if all queues start empty, then $w_\alpha(\boldsymbol{\phi})$ is the minimum rate of growth of $V_\alpha(\boldsymbol{q}(t))$ over all policies. We have the following important lemma.

Lemma 7 *For any $\boldsymbol{\phi} \in \mathcal{P}_M$, the following holds,*

(a)

$$w_\alpha(\boldsymbol{\phi}) = a(\boldsymbol{\phi}).$$

(b) *The optimizer* $\boldsymbol{x}^*$ *for* $a(\boldsymbol{\phi})$ *and the optimizer* $\boldsymbol{y}^*$ *for* $w_\alpha(\boldsymbol{\phi})$ *are both unique and they satisfy* $\boldsymbol{x}^* = \gamma \boldsymbol{y}^*$ *for some* $\gamma > 0$*. Further, if the optimizer* $\boldsymbol{x}^* \neq 0$*, then* $\boldsymbol{x}^*$ *and* $\boldsymbol{y}^*$ *are the only vectors that satisfy the following conditions: there exist* $\mu_m^i \geq 0$ *such that* $\sum_{i=1}^N \mu_m^i = \phi_m$*,* $y_i^* = [\lambda_i - \sum_{m=1}^M \mu_m^i F_m^i]^+$*,* $x_i^* = \gamma y_i^*$ *for some* $\gamma > 0$*,* $\sum_{i=1}^N (x_i^*)^{\alpha+1} \leq 1$*, and*

$$\mu_m^i = 0 \text{ whenever } (x_i^*)^\alpha F_m^i < \max_{1 \leq k \leq N} (x_k^*)^\alpha F_m^k.$$

This lemma is proved by showing that the two problems $a(\boldsymbol{\phi})$ and $w_\alpha(\boldsymbol{\phi})$ can be viewed as dual problems of each other. The details of the proof is provided in Appendix C.

Using part (a) of Lemma 7, we immediately obtain the following.

$$J_\alpha = \inf_{\{\boldsymbol{\phi} \in \mathcal{P}_M \mid w_\alpha(\boldsymbol{\phi}) > 0\}} \frac{H(\boldsymbol{\phi} || \boldsymbol{p})}{w_\alpha(\boldsymbol{\phi})}. \tag{25}$$

Further, according to Proposition 6, the above expression provides a lower bound for the decay-rate of the queue-overflow probability $\mathbf{P}_0[V_\alpha(\boldsymbol{q}_i^B(T)) \geq 1]$ for any $T > 0$. The following lemma shows that J_α is positive, and hence the above bound is non-trivial.

Proposition 8

$$J_\alpha \geq \frac{1}{N^{\frac{1}{\alpha+1}}} I_{\text{opt}}.$$

Proof: Recall that

$$J_\alpha = \inf_{\{\boldsymbol{\phi} \in \mathcal{P}_M \mid w_\alpha(\boldsymbol{\phi}) > 0\}} \frac{H(\boldsymbol{\phi} || \boldsymbol{p})}{w_\alpha(\boldsymbol{\phi})}$$

and

$$I_{\text{opt}} = \inf_{\{\boldsymbol{\phi} \in \mathcal{P}_M \mid w(\boldsymbol{\phi}) > 0\}} \frac{H(\boldsymbol{\phi} || \boldsymbol{p})}{w(\boldsymbol{\phi})}$$

.

For all $\boldsymbol{x} \geq 0$, we have $N^{\frac{1}{\alpha+1}} \max_{1 \leq i \leq N} x_i \geq V_\alpha(\boldsymbol{x})$. Further, since $w(\boldsymbol{\phi})$ and $w_\alpha(\boldsymbol{\phi})$ have the same constraint set, we have $N^{\frac{1}{\alpha+1}} w(\boldsymbol{\phi}) \geq w_\alpha(\boldsymbol{\phi})$ and as a consequence we have

$$\{\boldsymbol{\phi} \mid w_\alpha(\boldsymbol{\phi}) > 0\} \subseteq \{\boldsymbol{\phi} \mid w(\boldsymbol{\phi}) > 0\}. \tag{26}$$

Hence, for any $\boldsymbol{\phi}$ such that $w_\alpha(\boldsymbol{\phi}) > 0$, we have

$$\frac{H(\boldsymbol{\phi}\|\|\boldsymbol{p})}{w_\alpha(\boldsymbol{\phi})} \geq \frac{1}{N^{\frac{1}{\alpha+1}}} \frac{H(\boldsymbol{\phi}\|\|\boldsymbol{p})}{w(\boldsymbol{\phi})}.$$

Taking infimum over the corresponding constraint sets and using (26), we then obtain $J_\alpha \geq \frac{1}{N^{\frac{1}{\alpha+1}}} I_{\text{opt}}$. *Q.E.D.*

Finally, we can show that the lower bound J_α is tight, in the sense that there exists $T > 0$ and a trajectory that overflows at T with cost J_α. We will need the following lemma, which provides a structural property of the fluid sample path when the channel-rate process is linear. Specifically, if the channel-rate process $\boldsymbol{s}(\cdot)$ is linear, then the queue trajectory $\boldsymbol{q}(\cdot)$ must also be linear, and its derivative must solve $w_\alpha(\boldsymbol{\phi})$.

Lemma 9 *Consider a fluid sample path $(\boldsymbol{s}(t), \boldsymbol{q}(t))$ under the α-algorithm. If $\boldsymbol{q}(0) = 0$ and $\boldsymbol{s}(t) = t\boldsymbol{\phi}$ for $t \geq 0$, then the corresponding queue trajectory $\boldsymbol{q}(t)$ must satisfy the following:*

(a) The queue trajectory is linear, i.e., there exists $\tilde{\boldsymbol{y}} = [\tilde{y}_i, i = 1, ..., N] \geq 0$, such that $\boldsymbol{q}(t) = t\tilde{\boldsymbol{y}}$ for all $t \geq 0$.

(b) There must exist $\mu_m^i \geq 0$ such that $\sum_{i=1}^N \mu_m^i = \phi_m$, $\tilde{y}_i = [\lambda_i - \sum_{m=1}^M \mu_m^i F_m^i]^+$ and

$$\mu_m^i = 0 \text{ whenever } \tilde{y}_i^\alpha F_m^i < \max_{1 \leq k \leq N} \tilde{y}_k^\alpha F_m^k.$$

In other words, the queue trajectory $\boldsymbol{q}(t)$ is consistent with the scheduling rule of the α-algorithm.

(c) $\boldsymbol{y}^* = \tilde{\boldsymbol{y}}$ is the unique minimizer of $w_\alpha(\boldsymbol{\phi})$.

Proof: Let

$$\begin{aligned} \Omega(\boldsymbol{\phi}) \quad = \quad & \Big\{ \boldsymbol{\lambda} \mid \text{ there exists } \mu_m^i \geq 0 \text{ such that} \\ & \lambda_i \leq \sum_{m=1}^M \mu_m^i F_m^i \text{ for all } i = 1, ..., N, \\ & \text{and} \sum_{i=1}^N \mu_m^i = \phi_m \text{for all } m = 1, ..., M \Big\}. \end{aligned}$$

Note that if $\boldsymbol{\phi} = \boldsymbol{p}$, then $\Omega(\boldsymbol{\phi})$ corresponds to the capacity region of the system (for stability) [3]. The variables μ_m^i can be viewed as some long-term fraction of time that user i is picked and the channel state is m.

Recall from Proposition 5 that

$$\begin{aligned}\dot{V}_\alpha(\boldsymbol{q}(t)) &= \left(\sum_{i=1}^{N}(q_i(t))^{\alpha+1}\right)^{\frac{-\alpha}{\alpha+1}} \left[\sum_{i=1}^{N}(q_i(t))^{\alpha}\lambda_i \right. \\ &\qquad \left. - \sum_{m=1}^{M}\phi_m \max_{1\le k\le N}(q_k(t))^{\alpha}F_m^k\right].\end{aligned}$$

First, consider the case when $\boldsymbol{\lambda} \in \Omega(\boldsymbol{\phi})$. We will have $\dot{V}_\alpha(\boldsymbol{q}(t)) \le 0$ if $\boldsymbol{q}(t) \ne 0$. Hence, starting from $\boldsymbol{q}(0) = 0$, we must have $V_\alpha(\boldsymbol{q}(t)) = 0$ and $\boldsymbol{q}(t) = 0$ for all $t \ge 0$. Therefore, part (a) holds with $\tilde{y}_i = 0$ for all i. Part (b) then trivially holds. Part (c) follows since the minimizer of $w_\alpha(\boldsymbol{\phi})$ for this case is $\boldsymbol{y}^* = 0$.

On the other hand, if $\boldsymbol{\lambda} \notin \Omega(\boldsymbol{\phi})$, then for all $\boldsymbol{q}(t) \ne 0$, by setting $\hat{q}_i(t) = \frac{q_i(t)}{[\sum_{i=1}^{N}(q_i(t))^{\alpha+1}]^{\frac{1}{\alpha+1}}}$, we have

$$\dot{V}_\alpha(\boldsymbol{q}(t)) = \sum_{i=1}^{N}\hat{q}_i^{\alpha}(t)\lambda_i - \sum_{m=1}^{M}\phi_m \max_{1\le k\le N}\hat{q}_k^{\alpha}(t)F_m^k,$$

and $\left[\sum_{i=1}^{N}\hat{q}_i^{\alpha+1}(t)\right]^{\frac{1}{\alpha+1}} = 1$. We thus have $\dot{V}_\alpha(\boldsymbol{q}(t)) \le a(\boldsymbol{\phi})$ and $V_\alpha(\boldsymbol{q}(t)) \le ta(\boldsymbol{\phi})$ for all $t \ge 0$. This shows that $ta(\boldsymbol{\phi})$ upper bounds the maximum growth of $V_\alpha(\boldsymbol{q}(t))$. On the other hand, let μ_m^i be the average fraction of time in $[0, t]$ that user i is picked and the channel state is m. Then $\sum_{i=1}^{N}\mu_m^i = \phi_m$ for all m, and

$$q_i(t) \ge t[\lambda_i - \sum_{m=1}^{M}\mu_m^i F_m^i]^+.$$

(The inequality is due to the fact that the queue q_i may be empty at some points in this interval). Hence,

$$V_\alpha(\boldsymbol{q}(t)) \ge tw_\alpha(\boldsymbol{\phi}).$$

However, by Lemma 7, $a(\boldsymbol{\phi}) = w_\alpha(\boldsymbol{\phi})$. We thus have

$$V_\alpha(\boldsymbol{q}(t)) = ta(\boldsymbol{\phi}) = tw_\alpha(\boldsymbol{\phi}),$$

i.e. there is only one possible trajectory $V_\alpha(\boldsymbol{q}(t))$ given that $\boldsymbol{s}(t) = t\boldsymbol{\phi}$. Further, we have $V_\alpha(\frac{\boldsymbol{q}(t)}{t}) = w_\alpha(\boldsymbol{\phi})$, i.e., $\frac{\boldsymbol{q}(t)}{t}$ optimizes $w_\alpha(\boldsymbol{\phi})$. Since the optimizer of $w_\alpha(\boldsymbol{\phi})$, denoted by $\tilde{\boldsymbol{y}}$, is unique, we thus have $\boldsymbol{q}(t) = t\tilde{\boldsymbol{y}}$. This shows parts (a) and (c). Part (b) follows from part (b) of Lemma 7.

Q.E.D.

The following result then shows that the lower bound J_α is tight. Recall the definition of $\hat{\Gamma}_T$ in Section 5.1.

Proposition 10 *There exists T and a fluid sample path in $\hat{\Gamma}_T$ whose cost is equal to J_α.*

Proof: Let $\boldsymbol{\phi}^*$ denote the solution to J_α in (25), i.e., $J_\alpha = \frac{H(\boldsymbol{\phi}^*||\boldsymbol{p})}{w_\alpha(\boldsymbol{\phi}^*)}$, and let $w^* = w_\alpha(\boldsymbol{\phi}^*) > 0$. (We can show that such a $\boldsymbol{\phi}^*$ always exists by showing that the infimum in (25) can be taken within a closed subset of the original constraint set.) If we use $\boldsymbol{s}(t) = t\boldsymbol{\phi}^*, \quad t \geq 0$ as the channel-rate process, and let the queue process start from $\boldsymbol{q}(0) = 0$, then $\boldsymbol{q}(\cdot)$ must follow a linear trajectory according to Lemma 9, i.e.,

$$\boldsymbol{q}(t) = t\tilde{\boldsymbol{x}}, \text{ for all } t \geq 0,$$

where $\boldsymbol{y}^* = \tilde{\boldsymbol{x}}$ is the minimizer of $w_\alpha(\boldsymbol{\phi}^*)$.

Let $T = \frac{1}{w_\alpha(\boldsymbol{\phi}^*)}$. Consider such a trajectory over the interval $[0, T]$. Clearly, the cost of this trajectory is equal to J_α. It only remains to show that the trajectory must overflow at T, which is true because

$$V_\alpha(T\tilde{\boldsymbol{x}}) = Tw_\alpha(\boldsymbol{\phi}^*) = 1.$$

Q.E.D.

Hence, we conclude that the minimum-cost to overflow is attained by a simple linear trajectory whose cost is J_α.

6 Asymptotical Optimality of α-algorithms

In this section, we will establish that in the limit as $\alpha \to \infty$, the α-algorithms asymptotically achieve the largest minimum-cost-to-overflow equal to I_{opt}

given in (10). To emphasize the dependence on α, we use $\mathbf{P}_0^\alpha$ to denote the probability distribution conditioned on $\boldsymbol{Q}(0) = 0$ under the α-algorithm (with a particular value of α). We now show the following:

Proposition 11 *For any $T > 0$, the following holds*

$$\lim_{\alpha\to\infty} \limsup_{B\to\infty} \frac{1}{B} \log \mathbf{P}_0^\alpha[\max_{1\le i\le N} q_i^B(T)) \ge 1] \le -I_{\text{opt}}.$$

Proof: Since $\max_{1\le i\le N} q_i(T) \ge 1$ implies $V_\alpha(\boldsymbol{q}(T)) \ge 1$, we must have

$$\mathbf{P}_0^\alpha[\max_{1\le i\le N} q_i^B(T)) \ge 1] \le \mathbf{P}_0^\alpha[V_\alpha(\boldsymbol{q}(T)) \ge 1].$$

Using Proposition 6, for all $T > 0$,

$$\begin{aligned} & \limsup_{B\to\infty} \frac{1}{B} \log \mathbf{P}_0^\alpha[\max_{1\le i\le N} q_i^B(T)) \ge 1] \\ \le \ & \limsup_{B\to\infty} \frac{1}{B} \log \mathbf{P}_0^\alpha[V_\alpha(\boldsymbol{q}(T)) \ge 1] \le -J_\alpha. \end{aligned}$$

From Proposition 8, $\lim_{\alpha\to\infty} J_\alpha \ge I_{\text{opt}}$. The result then follows.

Q.E.D.

Combining Proposition 1 and Proposition 11, we conclude that the α-algorithms asymptotically achieve the largest decay-rate I_{opt} of the queue-overflow probability over all scheduling policies.

6.1 Systems with ON-OFF Channels

Consider the scenario where F_m^i can either take the value 0 or a positive constant C. This scenario corresponds to a wireless system with ON-OFF channels and the ON-rates for all users are the same. In this case, for any $\alpha > 0$,

$$(q_i)^\alpha F_m^i \lesseqqgtr \max_{1\le k\le N} (q_k)^\alpha F_m^k \Leftrightarrow q_i F_m^i \lesseqqgtr \max_{1\le k\le N} q_k F_m^k.$$

Hence, for any $\alpha \ge 1$, the α-algorithms are equivalent to the max-weight algorithm (i.e. with $\alpha = 1$). Using the result in this paper, we immediately reach the following corollary.

Corollary 12 *For the above ON-OFF channel model, the max-weight scheduling algorithm (i.e., $\alpha = 1$) achieves the largest decay-rate I_{opt} of the queue-overflow probability over all scheduling policies.*

Table 1: Link capacities in different states

F_m^i	$m=1$	$m=2$	$m=3$
$i=1$	0	3	5
$i=2$	0	9	0
$i=3$	0	9	1
$i=4$	0	9	1

7 Simulation results

In this section we will provide simulation results to verify the analytical results in earlier sections. We simulate the following system with 4 links (i.e., $N = 4$) and 3 states (i.e., $M = 3$). In each time-slot, one unit of data arrives at each of the links (i.e., $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 1$). The probabilities of each channel state are denoted as p_1, p_2 and p_3, and will be given shortly. The capacity F_m^i of link i in channel state m is given by Table 1. The 95%-confidence intervals are very small, and hence they are not shown in the figures.

We first simulate Case 1 when $p_1 = 0.3$, $p_2 = 0.6$ and $p_3 = 0.1$. In Fig. 1, we plot the value of $\mathbf{P}[\max_{1\le i\le N} Q_i \ge B]$ (in log-scale) against the overflow-threshold B for the α-algorithms, where each curve corresponds to a different value of α. We have also plotted a line with slope equal to I_{opt} given by (10). Recall that I_{opt} is the maximum decay-rate of the queue-overflow probability. We can observe from Fig. 1 that, as the value of α increases, the slopes at the tail of the curves (i.e., for large B) approach I_{opt}. Hence, this confirms our analytical result that, as the value of α increases, the asymptotic decay-rate of the α-algorithms approaches the optimal decay-rate I_{opt}.

We have also simulated the exponential-rule of [25]. At any time t, if the channel state is m, the exponential-rule chooses to serve the link i^* such that

$$i^* = \operatorname*{argmax}_{i=1,\dots N} \quad \exp\left[\frac{Q_i(t)}{1 + \left(\frac{1}{N}\sum_{k=1}^{N} Q_k(t)\right)^{\eta}}\right] F_m^i,$$

where η is a constant parameter in $(0, 1)$. In Fig. 2, we plot $\mathbf{P}[\max_{1\le i\le N} Q_i \ge B]$ against the overflow threshold B for the exponential rule, as the parameter η varies. According to the results of [25], the exponential rule achieves the optimal decay-rate of the queue-overflow probability for any $0 < \eta < 1$. We

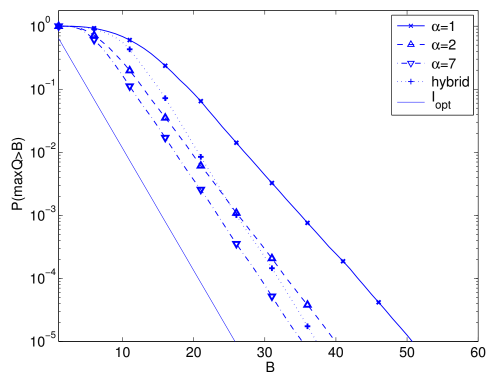

Figure 1: Case 1: Plot of $\mathbf{P}[\max_{1\leq i\leq N} Q_i \geq B]$ versus the overflow-threshold B for the α-algorithms. Each curve corresponds to a different value of α.

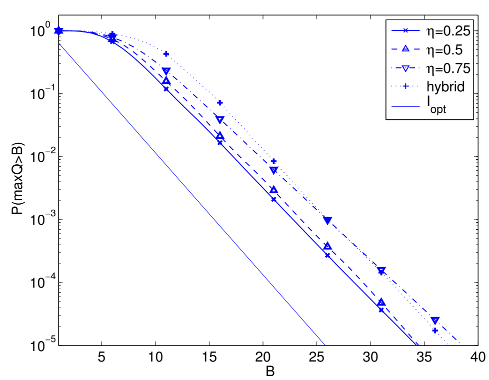

Figure 2: Case 1: Plot of $\mathbf{P}[\max_{1\leq i\leq N} Q_i \geq B]$ versus the overflow threshold B for the exponential-rule. Each curve corresponds to a different value of η.

observe from Fig. 2 that, for $\eta = 0.25$ and $\eta = 0.5$, the slopes at the tail of the curves indeed become parallel to I_{opt} for large B. For $\eta = 0.75$, such convergence has not occurred even for overflow-probability as low as 10^{-5}. Note that one should not conclude from the last curve that the results of [25] are violated: the LDP results of [25] will still kick in eventually, although at a larger value of the overflow-threshold B.

The previous set of simulation results raise some important issues on the applicability of large-deviation results. Specifically, the results in this paper (and in [25]) are large-buffer asymptotes, i.e., they characterize the behavior of the queue only when the overflow-threshold approaches infinity. The results often do not provide much information on what buffer level is large enough for the asymptotic behavior to become dominant. Further, an LDP only specifies the exponential decay-rate. The factor in front of exponential term can still vary substantially. Hence, one needs to be careful when comparing the performance predicted by an LDP with the actual performance of the protocol. This point is best illustrated with Case 2 that we simulated. Here, the probability of each channel state is given by $p_1 = 0.35$, $p_2 = 0.5$ and $p_3 = 0.15$. In Fig. 3, we again plot the value of $\mathbf{P}[\max_{1\le i\le N} Q_i \ge B]$ against the overflow-threshold B for the α-algorithms. We observe from Fig. 3 that, as α increases, the slopes at the tail of the curve indeed approaches I_{opt}. However, for small B the curve in fact shifts to the right, indicating that the actual queue-overflow probability $\mathbf{P}[\max_{1\le i\le N} Q_i \ge B]$ increases as α increases. Such a shift is more evident for smaller value of B. As B increases, for larger values of α the effect of the steeper slopes eventually dominates, and the queue-overflow probability improves as well.

To better understand this behavior, we introduce a state-space plot as in Fig. 6. The x-axis and the y-axis are the length of any two chosen queues (e.g. Q_1 and Q_3 as in Fig. 6). This state space is divided into regions, each of which corresponds to a fixed scheduling decision. For example, in Region 1, Queue 1 is served irrespective of the channel state (this is the case because the length of Queue 1 is much larger than Queue 3). In Region 2, Queue 1 is served in channel state $m = 3$, and Queue 3 is served in channel state $m = 2$. Finally, in Region 3, Queue 3 is served irrespective of the channel state. We refer to these regions as *decision regions*, and their boundary is determined by the scheduling policy. The dots in the figure are the states that have been visited by the system in the simulation (over some given length of time). A similar state space plot for case 2 is shown in Fig. 8.

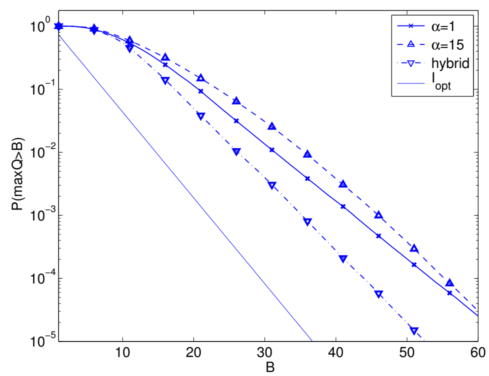

Figure 3: Case 2: Plot of $\mathbf{P}[\max_{1\leq i\leq N} Q_i \geq B]$ versus the overflow-threshold B for the α-algorithm. Each curve corresponds to a different value of α.

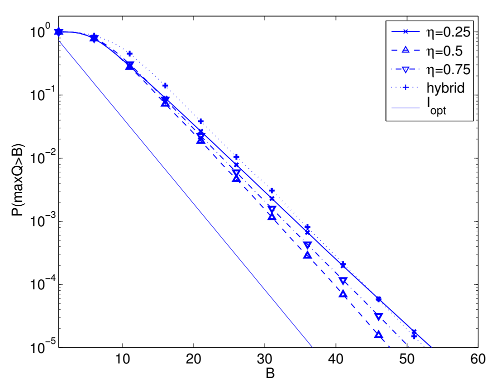

Figure 4: Case 2: Plot of $\mathbf{P}[\max_{1\leq i\leq N} Q_i \geq B]$ versus the overflow threshold B for the exponential-rule. Each curve corresponds to a different value of η.

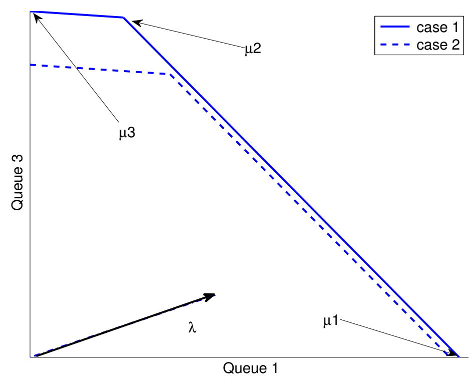

Figure 5: Shape of the capacity region.

Once the probabilities of channel states are given, the capacity region of the system can be determined. For example, Fig. 5 represents the capacity regions of case 1 and 2, projected to the space of Q_1 and Q_3. For this system with two active states, we can draw a correlation between the decision regions (e.g. Fig. 6), and the capacity region (e.g. case 1 in Fig. 5). We will refer to Region 1 and Region 3 as *max-queue* regions, in the sense that the decision is to serve the link with the longest queue, irrespective of the channel state. We refer to Region 2 as the *max-rate* region, in the sense that now the decision is to serve the link with the higher rate, depending on which channel state the system is in. The two max-queue regions can be correlated to the points μ_1 and μ_3 of the capacity region, where one user will be served in all states. The max-rate region can be correlated to the point μ_2 of the capacity region. The significance of this correlation is that region 2 contributes to an enlarged capacity region (i.e., the triangular area $\mu_1\mu_2\mu_3$).

For α-algorithms, as the value of α increases, the boundaries between the decision regions all converge to the diagonal line. This convergence has two implications. First, a larger value of α enlarges the two max-queue regions (see Fig. 7). For example, Point A that was in a max-rate region for small α (see Fig. 6), now moves to the max-queue region (see Fig. 7). Note that at Point A, we have $Q_1 > Q_3$. Hence, as the decision boundaries approach the diagonal line, the algorithm places more emphasis on reducing the largest queue. Intuitively, this helps to improve the decay-rate of the probability that the largest queue overflows.

However, a second effect of increasing α is that the size of the max-rate

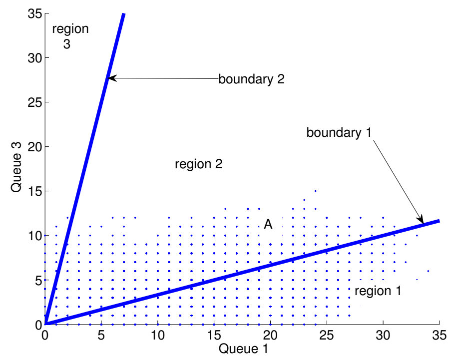

Figure 6: Case 1: Plot of the state space for $\alpha = 1$.

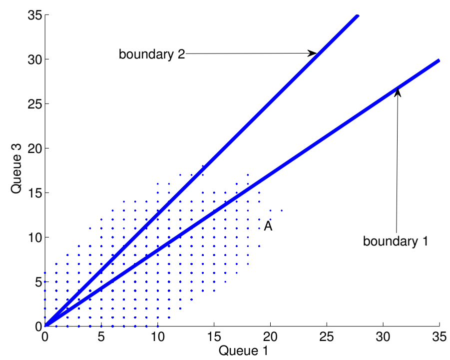

Figure 7: Case 1: Plot of the state space for $\alpha = 7$.

region (i.e., Region 2) is reduced. As a result, for smaller value of queue-length, it becomes less likely that the system state falls into the max-rate region. Recall that the decision rule in the max-rate region contributes to the improved capacity region (i.e., triangular area $\mu_1\mu_2\mu_3$). Hence, with large value of α, the scheduling algorithm is unlikely to take advantage of the increased capacity at small queue-lengths, which leads to a tendency for the queue-length to grow. This phenomenon can be observed by the fact that the dots in Fig. 7 now grows along the two boundary lines. It is even more evident in a similar plot for Case 2 (in Fig. 9). After the queue length increases, eventually the width of Region 2 will be sufficiently large so that the system state is more likely to fall into the max-rate region. Only after that, the effect of LDP starts to kick in, and the decay-rate of the queue-overflow probability starts to improve.

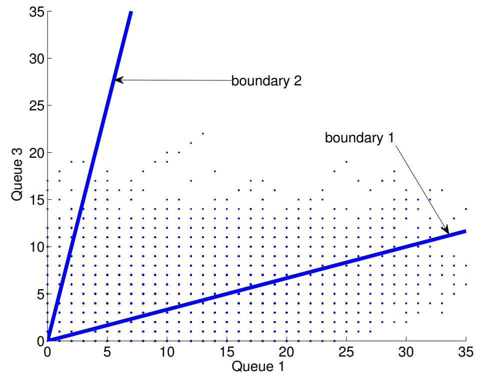

Figure 8: Case 2: Plot of the state space for $\alpha = 1$.

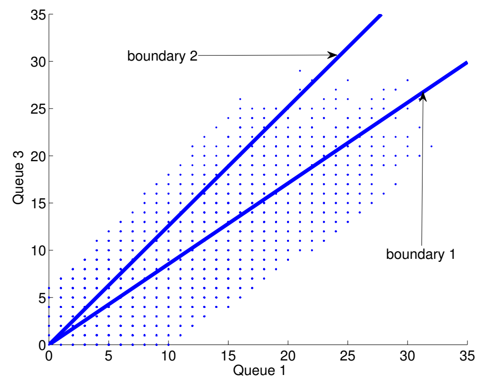

Figure 9: Case 2: Plot of the state space for $\alpha = 7$.

Although the above discussion is restricted to the dynamics of two queues over two active states, we feel that the above two conflicting trends apply to more general cases. Indeed, the understanding of these two trends help us to interpret the results in Fig. 1 and Fig. 3. First, refer to Fig. 6 for Case 1. For small value of α, the queues tend to accumulate around the boundary between Region 1 and Region 2. As α increases, the max-queue region (Region 1) enlarges, which helps to reduce the longer queue and push the state space to the origin (Fig. 7). The conflicting effect due to thinning of the max-rate region is not so strong, and the beneficial effect of large α is manifested. Thus, these plots explain why the performance plot in Fig. 1 improves with increasing α. Now, comparing the capacity region for the two cases (Fig. 5), we find that in case 2, the offered load, $\boldsymbol{\lambda}$, is closer to the line $\mu_1\mu_3$. Hence, the triangular section $\mu_1\mu_2\mu_3$ plays a more significant role in reducing the queue length. We would thus expect the effect of thinning of the max-rate region to be relatively stronger than in the previous case. This is exactly what we observe in Fig. 8 and Fig. 9. At small value of α (Fig. 8), the queues tend to accumulate relatively more in the max-rate region. Now, as α increases, the stronger effect caused by the thinning of the max-rate region forces the queue length to increase (Fig. 9). As a result, at small values of threshold, B, the overflow probability in fact deteriorates.

The above observations motivate us to design a new class of hybrid scheduling policies that have the benefits of both large α (for improving the large-deviation decay-rate of the queue-overflow probability) and small α (for having a large max-rate region, which helps to improve the overflow probability at small queue lengths). Essentially, to have good large-deviation decay-rates of the queue-overflow probability, we need to use a large α so that the decision boundaries become close to parallel to the diagonal line. However, this may lead to poor performance at small queue-lengths due to the thinner max-rate regions. To avoid this, we first use a smaller value of α when the queue-length is small and gradually change to large α when queue increases. More specifically, the hybrid policy works by modifying the weight function. The scheduling policy still picks the user i for service such that it has the largest value of $w_i(\boldsymbol{q})F_m^i$. However, the weight of user i, $w_i(\boldsymbol{q})$, is not equal to q_i^α anymore. Instead, it contains both a term for small α, and a term for large α. Specifically, let us assume that we are interested in transitioning from small α to large α when the queue length is around $B_* = 10$. We tested a hybrid policy that uses a combination of $\alpha = 1$ and $\alpha = 15$. The weight function we used is $w_i(\boldsymbol{q}) = q_i + ([q_i - \frac{K(\boldsymbol{q})}{F_m^i}]^+)^{15}$ where the value $K(\boldsymbol{q})$ will be specified later. For $q_i < \frac{K(\boldsymbol{q})}{F_m^i}$, the weight function is simply q_i. Hence, the behavior of the scheduling algorithm is similar to $\alpha = 1$. For large q_i, the term $(q_i - \frac{K(\boldsymbol{q})}{F_m^i})^{15}$ dominates. Hence, the behavior of the scheduling algorithm switches to that of $\alpha = 15$. The offset $\frac{K(\boldsymbol{q})}{F_m^i}$ is chosen to ensure that the decision boundary does not have sudden jumps. Specifically, the value of $K(\boldsymbol{q})$ is given by

$$K(\boldsymbol{q}) = \min_{1\le i\le N} \left(B_* F_m^i + [B_* - q_i]^+ \max_{1\le k\le N} F_m^k \right). \tag{27}$$

To understand the intuition behind (27), first consider the case when $q_i > B_*$

for all queues. Then, $K(\boldsymbol{q}) = B_* \min_{1\le i\le N} F_m^i$. The offset in this case becomes $(\frac{B_* \min_{1\le i\le N} F_m^i}{F_m^1}, \ldots, B_*, \ldots, \frac{B_* \min_{1\le i\le N} F_m^i}{F_m^N})$ which is exactly the point where the decision boundary of $\alpha = 1$ meets the threshold boundary $\max_{1\le i\le N} q_i = B_*$. However, if we just use $K(\boldsymbol{q}) = B_* \min_{1\le i\le N} F_m^i$, the problem is that the transition to large α occurs too early if not all q_i are greater than B_*. For example, consider channel state $m = 2$. In this case, the offset described above becomes $(B_*, \frac{B_*}{3}, \frac{B_*}{3}, \frac{B_*}{3})$. The projection of this offset value to the space of the queues q_2, q_3 and q_4 is $(\frac{B_*}{3}, \frac{B_*}{3}, \frac{B_*}{3})$. As a result, the transition from $\alpha = 1$ to $\alpha = 15$ would occur too early (at $\frac{B_*}{3}$) for q_2, q_3 or q_4 if q_1 is small. To compensate for this effect, we have introduced the second term in (27). Essentially, if q_1 is small, its channel rates do not play much role in determining the minimum value of (27). In this specific example, if $q_1 = 0$ and $q_2, q_3, q_4 > B_*$, then the offset value is $K(\boldsymbol{q}) = (3B_*, B_*, B_*, B_*)$. Hence, the transition occurs at the desirable values of q_2, q_3 and q_4.

We plot the decision boundaries for this hybrid algorithm in Fig. 10. As we can see, the max-rate region is large even for small queue-lengths. In Fig. 3, we also plotted the performance of the hybrid algorithm. Compare with the curve for $\alpha = 15$, we note that the curve for the hybrid algorithm has moved to the left as desired. Also note that the slope of the curve is close to I_{opt}. Hence, this figure confirms that the hybrid algorithm achieves the benefit of both large α and small α.

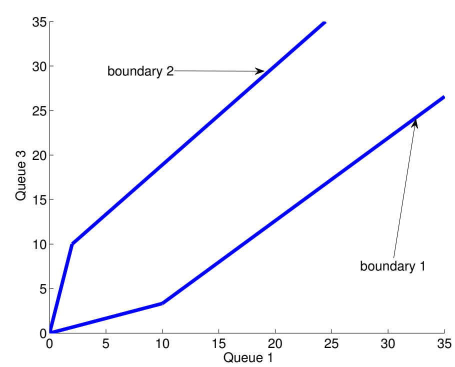

Figure 10: Plot of the decision boundaries for the hybrid algorithm.

We find that the same intuitions seem to also apply for the exponential-

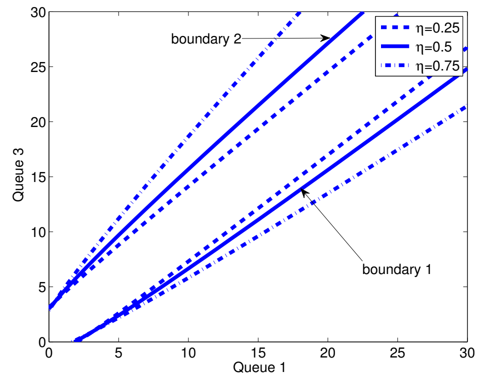

Figure 11: Plot of the decision boundaries for exponential-rule for various values of η.

rule [25]. Recall that Fig. 2 plots the value of $\mathbf{P}[\max_{1\leq i\leq N} Q_i \geq B]$ versus the overflow-threshold B for the exponential-rule when the parameter η varies. A similar figure for Case 2 is given in Fig. 4. To understand why $\eta = 0.5$ seems to produce the best overall performance, we plot the decision boundaries of the exponential-rule in Fig. 11. We can see that, if the value of η is too small, then the max-rate region (between the decision boundaries) is too narrow, which increases the queue-overflow probability at small threshold values. If the value of η is too large, then the max-rate region is big enough. However, the decision boundaries do not become parallel to the diagonal line until the queue-length is very large. Hence, the large-deviation decay-rate kicks in only at a larger queue-length. A medium value of η (around 0.5) seems to achieve a balance between the above two cases, and produces a state-space plot that is similar to our hybrid algorithm (Fig. 10). We have also plotted the performance of the exponential-rule and our hybrid algorithm in Fig. 4. Their performance appears to be quite comparable. Finally, we plot the performance of the hybrid algorithm for case 1 and we find that the hybrid algorithm also performs very well, which indicates that the hybrid algorithm is quite robust and seems to work well in all cases.

8 Conclusion

In this paper, we study wireless scheduling algorithms for the downlink of a single cell that can maximize the asymptotic decay-rate of the queue-overflow probability, as the overflow threshold approaches infinity. Specifically, we focus on the class of "α-algorithms," which pick the user for service at each time that has the largest product of the transmission rate multiplied by the backlog raised to the power α. We show that when α approaches infinity, the α-algorithms asymptotically achieve the largest decay-rate of the queue-overflow probability. A key step in proving this result is to use a Lyapunov function to derive a simple lower bound for the minimum-cost-to-overflow under the α-algorithms. This technique, which is of independent interest, circumvents solving the difficult multi-dimensional calculus-of-variations problem typical in this type of problems. Finally, using the insight from this result, we design scheduling algorithms that are both close-to-optimal in terms of the asymptotic decay-rate of the overflow probability, and empirically shown to maintain small queue-overflow probabilities over queue-length ranges of practical interest. For future work, we plan to extend the results to more general network and channel models.

Acknowledgment

The authors are grateful for the helpful discussions with Prof. Michael Neely, Prof. Sean Meyn and Dr. Alexander Stolyar.

A Proof of Lemma 3

Proof: For every $\delta > 0$, by definition of $w(\boldsymbol{\phi}^1)$ there exists $\tilde{\gamma}_m^{i,\delta} \geq 0$ such that $\sum_{i=1}^N \tilde{\gamma}_m^{i,\delta} = 1$ for all $m = 1, ..., M$, and

$$\max_{1 \leq i \leq N} [\lambda_i - \sum_{m=1}^M \phi_m^1 \tilde{\gamma}_m^{i,\delta} F_m^i]^+ < w(\boldsymbol{\phi}^1) + \delta.$$

Since by assumption $|\phi_m^1 - \phi_m^2| < \epsilon$ for all $m = 1, ..., M$, we have, for all $i = 1, ..., N$,

$$\begin{aligned} & (\lambda_i - \sum_{m=1}^M \phi_m^1 \tilde{\gamma}_m^{i,\delta} F_m^i) \\ \geq \;& (\lambda_i - \sum_{m=1}^M \phi_m^2 \tilde{\gamma}_m^{i,\delta} F_m^i) - \epsilon \sum_{m=1}^M F_m^i. \end{aligned}$$

Taking projection on both sides to the positive real axis, and using the fact that $(x-y)^+ \geq x^+ - y$ for non-negative y, we have

$$\begin{aligned}
& w(\boldsymbol{\phi}^1) + \delta \\
\geq\ & \max_{1\leq i\leq N} [\lambda_i - \sum_{m=1}^{M} \phi_m^1 \tilde{\gamma}_m^{i,\delta} F_m^i]^+ \\
\geq\ & \max_{1\leq i\leq N} [\lambda_i - \sum_{m=1}^{M} \phi_m^2 \tilde{\gamma}_m^{i,\delta} F_m^i]^+ - \epsilon \sum_{i=1}^{N}\sum_{m=1}^{M} F_m^i \\
\geq\ & w(\boldsymbol{\phi}^2) - \epsilon \sum_{i=1}^{N}\sum_{m=1}^{M} F_m^i.
\end{aligned}$$

Since δ can be made arbitrarily small, the result then follows. *Q.E.D.*

B Proof of Proposition 5

Proof: Let $\mathcal{T}_0$ denote the set of time t where either $\boldsymbol{s}(\cdot)$ or $\boldsymbol{q}(\cdot)$ is not differentiable. Since $\boldsymbol{s}(\cdot)$ and $\boldsymbol{q}(\cdot)$ are both Lipschitz-continuous, $\mathcal{T}_0$ is of measure zero. Further, since $(\boldsymbol{s}(\cdot), \boldsymbol{q}(\cdot))$ is a fluid-sample-path, there exists a sequence $(\boldsymbol{s}^B(\cdot), \boldsymbol{q}^B(\cdot))$ that converges to $(\boldsymbol{s}(\cdot), \boldsymbol{q}(\cdot))$ uniformly over compact intervals. Fix any $t \notin \mathcal{T}_0$ and $\boldsymbol{q}(t) \neq 0$. We consider the summation

$$\sum_{i=1}^{N} (q_i(t))^\alpha (q_i^B(t+\delta) - q_i^B(t))$$

for a small $\delta > 0$. For each $i = 1, ..., N$, if $q_i(t) > 0$, by the continuity of $q_i(\cdot)$, we can find $\delta_1^i > 0$ such that $q_i(s) \geq \frac{3}{4} q_i(t)$ for $s \in [t, t+\delta_1^i]$. Further, since the convergence to $\boldsymbol{q}(\cdot)$ is uniform over the interval $[t, t+\delta_1]$, there exists $B_1^i > 0$ such that for all $B \geq B_1^i$ and $s \in [t, t+\delta_1^i]$, we have

$$q_i^B(s) \geq \frac{q_i(t)}{2} > 0.$$

Take $\bar{F}$ such that $F_m^i \leq \bar{F}$ for all m and i. We then have, for all $B \geq \max\{B_1^i, \frac{2\bar{F}}{q_i(t)}\}$ and $s \in [t, t+\delta_1]$, the unscaled queue length must satisfy

$$Q_i(\lfloor Bs \rfloor) \geq \bar{F}.$$

Hence, the projection operation in (1) is not needed. By applying the scaling operation (6) and (9), we have, for all $B \geq \max\{B_1^i, \frac{2\bar{F}}{q_i(t)}\}$ and $\delta < \delta_1^i$,

$$
\begin{aligned}
q_i^B(t+\delta) &= q_i^B(t) + \lambda_i \delta \\
&\quad - \frac{1}{B} \sum_{y=\lfloor Bt \rfloor}^{\lfloor B(t+\delta) \rfloor - 1} \sum_{m=1}^{M} \mathbf{1}_{\{C(y)=m\}} \mathbf{1}_{\{U(y)=i\}} F_m^i + O(\frac{1}{B}),
\end{aligned}
$$

where the additional $O(1/B)$ term is to account for the fact that $\boldsymbol{q}^B(t+\delta)$ is interpolated when $B(t+\delta)$ is not an integer. We then have,

$$
\begin{aligned}
& (q_i(t))^\alpha (q_i^B(t+\delta) - q_i^B(t)) \\
= \; & (q_i(t))^\alpha \\
& \times \left[\lambda_i \delta - \frac{1}{B} \sum_{y=\lfloor Bt \rfloor}^{\lfloor B(t+\delta) \rfloor - 1} \sum_{m=1}^{M} \mathbf{1}_{\{C(y)=m\}} \mathbf{1}_{\{U(y)=i\}} F_m^i \right] \\
& \quad + O(1/B).
\end{aligned}
$$

If $q_i(t) = 0$, then the above equation also holds trivially. Let

$$
B_1 = \max \left\{ \max_{i: q_i(t) > 0} B_1^i, \max_{i: q_i(t) > 0} \frac{2\bar{F}}{q_i(t)} \right\}
$$

and $\delta_1 = \min_{i: q_i(t) > 0} \delta_1^i$. Hence, summing over all $i = 1, ..., N$, we have, for all $B \geq B_1$ and $\delta < \delta_1$,

$$
\begin{aligned}
& \sum_{i=1}^{N} (q_i(t))^\alpha (q_i^B(t+\delta) - q_i^B(t)) \\
= \; & \sum_{i=1}^{N} (q_i(t))^\alpha \lambda_i \delta - \frac{1}{B} \sum_{y=\lfloor Bt \rfloor}^{\lfloor B(t+\delta) \rfloor - 1} \sum_{m=1}^{M} \mathbf{1}_{\{C(y)=m\}} \\
& \quad \times \left[\sum_{i=1}^{N} \mathbf{1}_{\{U(y)=i\}} (q_i(t))^\alpha F_m^i \right] + O(\frac{1}{B}). \qquad (28)
\end{aligned}
$$

Next, consider the second summation and consider the term corresponding to each channel state m. For those users i such that $(q_i(t))^\alpha F_m^i < \max_{1 \leq k \leq N} (q_k(t))^\alpha F_m^k$, since $\boldsymbol{q}(\cdot)$ is Lipschitz-continuous and the convergence of $\{\boldsymbol{q}^B(\cdot)\}$ to $\boldsymbol{q}(\cdot)$ is uniform over compact intervals, there exists a small enough $\delta_2 > 0$ and a large enough $B_2 > 0$ such that for all $B \geq B_2$ and $s \in [t, t+\delta_2]$, we have

$$(q_i^B(s))^\alpha F_m^i < \max_{1\leq k\leq N} (q_k^B(s))^\alpha F_m^k.$$

Hence, over the time interval $[t, t+\delta_2]$ the user i will not be selected in the channel state m, i.e.,

$$\mathbf{1}_{\{U(\lfloor Bs \rfloor)=i\}} = 0 \text{ for } s \in [t, t+\delta_1]. \tag{29}$$

We then have, for all $B \geq \max\{B_1, B_2\}$ and $\delta < \min\{\delta_1, \delta_2\}$,

$$\begin{aligned}
&\frac{1}{B} \sum_{\lfloor y=Bt \rfloor}^{\lfloor B(t+\delta)\rfloor -1} \sum_{m=1}^{M} \mathbf{1}_{\{C(y)=m\}} \\
&\qquad \times \left[\sum_{i=1}^{N} \mathbf{1}_{\{U(y)=i\}} (q_i(t))^\alpha F_m^i) \right] \\
= &\sum_{m=1}^{M} \left[\frac{1}{B} \sum_{\lfloor y=Bt \rfloor}^{\lfloor B(t+\delta)\rfloor -1} \mathbf{1}_{\{C(y)=m\}} \right] \max_{1\leq k\leq N} (q_k(t))^\alpha F_m^k) \\
= &\sum_{m=1}^{M} (s_m^B(t+\delta) - s_m^B(t)) \max_{1\leq k\leq N} (q_k(t))^\alpha F_m^k) + O(\frac{1}{B}).
\end{aligned}$$

Substituting into (28), and letting $B \to \infty$, we then have

$$\begin{aligned}
\sum_{i=1}^{N} (q_i(t))^\alpha (q_i(t+\delta) - q_i(t)) = \sum_{i=1}^{N} (q_i(t))^\alpha \lambda_i \delta \\
- \sum_{m=1}^{M} (s_m(t+\delta) - s_m(t)) \max_{1\leq k\leq N} (q_k(t))^\alpha F_m^k).
\end{aligned} \tag{30}$$

Finally, note that since $t \notin \mathcal{T}_0$, $\boldsymbol{q}(\cdot)$ is differentiable at t. We then have,

$$\begin{aligned}
&V_\alpha(\boldsymbol{q}(t+\delta)) - V_\alpha(\boldsymbol{q}(t)) \\
= &\left[\sum_{i=1}^{N} (q_i(t))^{\alpha+1} \right]^{\frac{-\alpha}{\alpha+1}} \sum_{i=1}^{N} (q_i(t))^\alpha [q_i(t+\delta) - q_i(t)] \\
&\qquad + o(\delta).
\end{aligned}$$

Dividing both sides by δ, using (30) and taking the limit as δ goes to 0, the result then follows.

Q.E.D.

C Proof of Lemma 7

Proof: We first show that $a(\boldsymbol{\phi})$ and $w_\alpha(\boldsymbol{\phi})$ are dual problem of each other. Letting $\xi_i = x_i^\alpha, i = 1, ..., N$ and $\boldsymbol{\xi} = [\xi_i]$, the problem $a(\boldsymbol{\phi})$ can be rewritten as

$$\begin{aligned} a(\boldsymbol{\phi}) = \max_{\boldsymbol{\xi} \geq 0} \quad & \left[\sum_{i=1}^{N} \xi_i \lambda_i - \sum_{m=1}^{M} \phi_m \max_{1 \leq i \leq N} \xi_i F_m^i \right] \\ \text{subject to} \quad & \sum_{i=1}^{N} \xi_i^{\frac{\alpha+1}{\alpha}} \leq 1. \end{aligned}$$

Introducing the variable $\eta_m \geq \max_{1 \leq i \leq N} \xi_i F_m^i$, the problem $a(\boldsymbol{\phi})$ can be further rewritten as

$$\begin{aligned} a(\boldsymbol{\phi}) = \max_{\boldsymbol{\xi} \geq 0, \boldsymbol{\eta}} \quad & \left[\sum_{i=1}^{N} \xi_i \lambda_i - \sum_{m=1}^{M} \phi_m \eta_m \right] \\ \text{subject to} \quad & \sum_{i=1}^{N} \xi_i^{\frac{\alpha+1}{\alpha}} \leq 1 \\ & \eta_m \geq \xi_i F_m^i \text{ for all } i, m. \end{aligned}$$

This is a convex optimization problem. Introducing the Lagrange multiplier $\mu_m^i \geq 0$ for each of the constraints $\eta_m \geq \xi_i F_m^i$, we obtain the Lagrangian

$$\begin{aligned} & L(\boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\eta}) \\ = \; & \left[\sum_{i=1}^{N} \xi_i \lambda_i - \sum_{m=1}^{M} \phi_m \eta_m \right] + \sum_{m=1}^{M} \sum_{i=1}^{N} \mu_m^i [\eta_m - \xi_i F_m^i] \\ = \; & \sum_{i=1}^{N} \xi_i (\lambda_i - \sum_{m=1}^{M} \mu_m^i F_m^i) - \sum_{m=1}^{M} \eta_m \left(\phi_m - \sum_{i=1}^{N} \mu_m^i \right). \end{aligned}$$

The dual objective function is then given by

$$\begin{aligned} D(\boldsymbol{\mu}) = \max_{\boldsymbol{\xi}\geq 0,\boldsymbol{\eta}} \quad & L(\boldsymbol{\xi},\boldsymbol{\mu},\boldsymbol{\eta}) \\ \text{subject to} \quad & \sum_{i=1}^{N} \xi_i^{\frac{\alpha+1}{\alpha}} \leq 1. \end{aligned}$$

Note that if $\sum_{i=1}^{N} \mu_m^i \neq \phi_m$, then $D(\boldsymbol{\mu}) = +\infty$ since we can set $|\eta_m|$ arbitrarily large. Otherwise, if $\sum_{i=1}^{N} \mu_m^i = \phi_m$ for all m, we then have,

$$\begin{aligned} D(\boldsymbol{\mu}) = \max_{\boldsymbol{\xi}\geq 0} \quad & \sum_{i=1}^{N} \xi_i(\lambda_i - \sum_{m=1}^{M} \mu_m^i F_m^i) \\ \text{subject to} \quad & \sum_{i=1}^{N} \xi_i^{\frac{\alpha+1}{\alpha}} \leq 1. \end{aligned} \tag{31}$$

Clearly, for those i such that $\lambda_i < \mu_m^i F_m^i$, the optimal solution for $D(\boldsymbol{\mu})$ is $\xi_i = 0$. Let $\mathcal{I}$ denote the set of i such that $\lambda_i - \mu_m^i F_m^i \geq 0$. If $\mathcal{I}$ is an empty set, then $D(\boldsymbol{\mu}) = 0$. If $\mathcal{I}$ is not empty, we can use Holder's inequality that, for any positive p and q such that $1/p + 1/q = 1$, the following holds,

$$\sum_{i=1}^{N} a_i b_i \leq [\sum_{i=1}^{N} a_i^p]^{1/p} [\sum_{i=1}^{N} b_i^q]^{1/q},$$

where equality holds if and only if there is a constant γ such that $a_i^p = \gamma b_i^q$ for all i. Hence, for all $\boldsymbol{\xi}$ such that the constraint (31) is satisfied, we have

$$\begin{aligned} & \sum_{i\in\mathcal{I}} \xi_i(\lambda_i - \sum_{m=1}^{M} \mu_m^i F_m^i) \\ = \quad & \sum_{i=1}^{N} \xi_i[\lambda_i - \sum_{m=1}^{M} \mu_m^i F_m^i]^+ \\ \leq \quad & \left[\sum_{i=1}^{N} \xi_i^{\frac{\alpha+1}{\alpha}}\right]^{\frac{\alpha}{\alpha+1}} \left[\sum_{i=1}^{N} ([\lambda_i - \sum_{m=1}^{M} \mu_m^i F_m^i]^+)^{\alpha+1}\right]^{\frac{1}{\alpha+1}} \\ \leq \quad & \left[\sum_{i=1}^{N} ([\lambda_i - \sum_{m=1}^{M} \mu_m^i F_m^i]^+)^{\alpha+1}\right]^{\frac{1}{\alpha+1}}, \end{aligned}$$

where equality holds if and only if

$$\sum_{i=1}^{N} \xi_i^{\frac{\alpha+1}{\alpha}} = 1, \tag{32}$$

and for some constant $\gamma > 0$,

$$\xi_i^{\frac{\alpha+1}{\alpha}} = \gamma^{\alpha+1}([\lambda_i - \sum_{m=1}^{M} \mu_m^i F_m^i]^+)^{\alpha+1}, \text{ for } i = 1, \ldots, N,$$

or, equivalently,

$$\xi_i^{\frac{1}{\alpha}} = \gamma[\lambda_i - \sum_{m=1}^{M} \mu_m^i F_m^i]^+, \text{ for } i = 1, \ldots, N. \tag{33}$$

Such a vector $\boldsymbol{\xi}$ clearly exists when $\mathcal{I}$ is not empty. Hence, if $\sum_{i=1}^{N} \mu_m^i = \phi_m$ for all m, we have

$$D(\boldsymbol{\mu}) = \left[\sum_{i=1}^{N}([\lambda_i - \sum_{m=1}^{M} \mu_m^i F_m^i]^+)^{\alpha+1}\right]^{\frac{1}{\alpha+1}},$$

which is true even when $\mathcal{I}$ is empty. We can therefore conclude that the dual problem is

$$\begin{aligned} \min_{\boldsymbol{\mu}\geq 0} D(\boldsymbol{\mu}) = \min_{\boldsymbol{y}\geq 0, \boldsymbol{\mu}\geq 0} \quad & \left(\sum_{i=1}^{N} y_i^{\alpha+1}\right)^{\frac{1}{\alpha+1}} \\ \text{subject to} \quad & y_i = \left[\lambda_i - \sum_{m=1}^{M} \mu_m^i F_m^i\right]^+ \\ & \sum_{i=1}^{N} \mu_m^i = \phi_m \text{ for all } m. \end{aligned}$$

This is exactly the problem $w_\alpha(\boldsymbol{\phi})$. Hence, strong duality implies that $a(\boldsymbol{\phi}) = w_\alpha(\boldsymbol{\phi})$.

The optimizer $\boldsymbol{y}$ of $w_\alpha(\boldsymbol{\phi})$ must be unique since the objective function in $w_\alpha(\boldsymbol{\phi})$ is strictly convex in $\boldsymbol{y}$. Using the complementary slackness condition, for any optimizer $\boldsymbol{\xi}$ and $\boldsymbol{\mu}$, we must have

$$\begin{aligned}
&\mu_m^i \geq 0, \quad \sum_{i=1}^N \mu_m^i = \phi_m, \\
&\xi_i^{\frac{1}{\alpha}} = \gamma[\lambda_i - \sum_{m=1}^M \mu_m^i F_m^i]^+ \\
&\mu_m^i = 0 \text{ if } \xi_i F_m^i < \max_{1\leq k\leq N} \xi_k F_m^k \\
&\sum_{i=1}^N \xi_i^{\frac{\alpha+1}{\alpha}} = 1 \text{ whenever } \boldsymbol{\xi} \neq 0 \text{ by (32).}
\end{aligned}$$

Since $\xi_i = x_i^\alpha$ and $y_i = [\lambda_i - \sum_{m=1}^M \mu_m^i F_m^i]^+$, we must have $\boldsymbol{x} = \gamma \boldsymbol{y}$. Further, if $\boldsymbol{x} \neq 0$, then since $\boldsymbol{y}$ is unique and $\sum_{i=1}^N x_i^{\alpha+1} = 1$, $\boldsymbol{x}$ must also be unique. The above set of equations are then exactly the condition in part (b) of the lemma. Conversely, any $\boldsymbol{\xi}$ and $\boldsymbol{\mu}$ (or, equivalently, $\boldsymbol{x}$ and $\boldsymbol{\mu}$) that satisfy the condition must correspond to the maximizer of $a(\boldsymbol{\phi})$ and $w_\alpha(\boldsymbol{\phi})$, respectively. Since the optimizers of $a(\boldsymbol{\phi})$ and $w_\alpha(\boldsymbol{\phi})$ are both unique, there is at most one $\boldsymbol{x}$ that satisfies the set of conditions in part (b) of the lemma. *Q.E.D.*

D Bounding the Decay-rate of the Stationary Overflow Probability: Supporting Results

In this section and the next, we will extend the result of Proposition 6 and show that J_α is also a lower-bound of the decay-rate of the stationary overflow probability, i.e.,

$$\limsup_{B\to\infty} \frac{1}{B} \log \mathbf{P}[V_\alpha(\boldsymbol{q}^B(0)) \geq 1] \leq -J_\alpha. \tag{34}$$

We will first establish some supporting results.

D.1 A More General Lower Bound

We start from a stronger version of Proposition 4, the proof of which is also fairly similar. For any given $T > 0$, let Ψ_T denote the space of mappings from

$[0, T]$ to $\mathbb{R}^N$, equipped with the essential supremum norm [12, p176, p352]. For any $\boldsymbol{x} \in \mathbb{R}^N$, let $\mathbf{P}_{\boldsymbol{x}}$ denote the distribution conditioned on $\boldsymbol{q}(0) = \boldsymbol{x}$.

Proposition 13 *Let $\mathcal{X}$ denote a closed set in $\mathbb{R}^N$. Let Γ_T be a closed subset in Ψ_T such that any $\boldsymbol{x}(\cdot) \in \Gamma_T$ satisfies $\boldsymbol{x}(0) \in \mathcal{X}$. Let $\hat{\Gamma}_T$ denote the set of fluid sample paths $(\boldsymbol{s}(\cdot), \boldsymbol{q}(\cdot))$ on the interval $[0, T]$ such that $\boldsymbol{q}(\cdot) \in \Gamma_T$. Then the following holds,*

$$\begin{aligned} & \limsup_{B \to \infty} \frac{1}{B} \log \left(\sup_{\{\boldsymbol{x} \in \mathcal{X}\}} \mathbf{P}_{\boldsymbol{x}}[\boldsymbol{q}^B(\cdot) \in \Gamma_T] \right) \\ \leq \quad & - \inf_{(\boldsymbol{s}(\cdot), \boldsymbol{q}(\cdot)) \in \hat{\Gamma}_T} \int_0^T H(\dot{\boldsymbol{s}}(t) || \boldsymbol{p}) dt. \end{aligned} \tag{35}$$

Proof: Let $\tilde{\Gamma}^B$ be the set of $\boldsymbol{s}^B(\cdot)$ on $[0, T]$ such that there exists $\boldsymbol{x} \in \mathcal{X}$ with which $\boldsymbol{s}^B(\cdot)$ drives a backlog process $\boldsymbol{q}^B(\cdot)$ that starts from $\boldsymbol{q}^B(0) = \boldsymbol{x}$ and satisfies $\boldsymbol{q}^B(\cdot) \in \Gamma_T$. Then, we have the following

$$\begin{aligned} & \limsup_{B \to \infty} \frac{1}{B} \log \left(\sup_{\{\boldsymbol{x} \in \mathcal{X}\}} \mathbf{P}_{\boldsymbol{x}}[\boldsymbol{q}^B(\cdot) \in \Gamma_T] \right) \\ \leq \quad & \limsup_{B \to \infty} \frac{1}{B} \log \left(\mathbf{P}[\boldsymbol{s}^B(\cdot) \in \tilde{\Gamma}^B] \right). \end{aligned}$$

The rest of the proof is identical to that of Proposition 4 starting from inequality (15). The only difference is that, in the last step, the sequence $(\acute{\boldsymbol{y}}_n(\cdot), \acute{\boldsymbol{q}}_n(\cdot))$ converges to $(\boldsymbol{y}^*(\cdot), \boldsymbol{q}^*(\cdot))$ where $\boldsymbol{q}^*(\cdot) \in \Gamma_T$ (because $\acute{\boldsymbol{q}}_n(\cdot) \in \Gamma_T$ and the set Γ_T is closed under the essential supremum norm.) *Q.E.D.*

D.2 Additional Bounds for the Lyapunov Drift

We will also need the following bounds on the Lyapunov drift. First, we establish a result for the Lyapunov drift of fluid sample paths. Without loss of generality, assume that $\lambda_i > 0$ for all $i = 1, ..., N$. Further, for stability, we assume that for some $\acute{\epsilon} > 0$, $\boldsymbol{\lambda}(1 + \acute{\epsilon})$ is inside the capacity region. Recall that this implies (2).

Proposition 14 *Let $(\boldsymbol{s}(\cdot), \boldsymbol{q}(\cdot))$ be any fluid sample path. Except for a set $\mathcal{T}_0$ of measure zero, at any time $t \notin \mathcal{T}_0$ and $\boldsymbol{q}(t) \neq 0$, the following bounds on the drift of the Lyapunov function $V_\alpha(\boldsymbol{q}(t))$ holds:*

(a)

$$\dot{V}_\alpha(\boldsymbol{q}(t)) \leq (\sum_{i=1}^{N} \lambda_i^{\alpha+1})^{\frac{1}{\alpha+1}}.$$

(b) For any $\epsilon > 0$, if $|\dot{s}_m(t) - p_m| < \epsilon$ for $m = 1, \ldots, M$, then

$$\dot{V}_\alpha(\boldsymbol{q}(t)) \leq -\acute{\epsilon} \frac{\min\limits_{1 \leq i \leq N} \lambda_i}{N^{\frac{\alpha}{\alpha+1}}} + \epsilon \sum_{m=1}^{M} \left[\sum_{i=1}^{N} (F_m^i)^{\alpha+1} \right]^{\frac{1}{\alpha+1}}. \tag{36}$$

Proof: We first show Part (a). As in the proof of Proposition 5, let $\mathcal{T}_0$ denote the set of time t where either $\boldsymbol{s}(\cdot)$ or $\boldsymbol{q}(\cdot)$ is not differentiable. Fix any $t \notin \mathcal{T}_0$ such that $\boldsymbol{q}(t) \neq 0$. By Proposition 5, we have

$$\begin{aligned} \dot{V}_\alpha(\boldsymbol{q}(t)) &= \left[\sum_{i=1}^{N} (q_i(t))^{\alpha+1} \right]^{\frac{-\alpha}{\alpha+1}} \left[\sum_{i=1}^{N} (q_i(t))^\alpha \lambda_i \right. \\ &\qquad \left. - \sum_{m=1}^{M} \dot{s}_m(t) \max_{1 \leq k \leq N} (q_k(t))^\alpha F_m^k \right] \\ &\leq \left[\sum_{i=1}^{N} (q_i(t))^{\alpha+1} \right]^{\frac{-\alpha}{\alpha+1}} \left[\sum_{i=1}^{N} (q_i(t))^\alpha \lambda_i \right]. \end{aligned} \tag{37}$$

By Holder's inequality, we know

$$\sum_{i=1}^{N} x_i^\alpha y_i \leq (\sum_{i=1}^{N} x_i^{\alpha+1})^{\frac{\alpha}{\alpha+1}} (\sum_{i=1}^{N} y_i^{\alpha+1})^{\frac{1}{\alpha+1}}.$$

Applying this to the second term in (37), we obtain the result in part (a).

To show part (b), using (2), we have

$$(1+\acute{\epsilon}) \sum_{i=1}^{N} (q_i(t))^\alpha \lambda_i \leq \sum_{m=1}^{M} p_m \sum_{i=1}^{N} (q_i(t))^\alpha \hat{\gamma}_m^i F_m^i \tag{38}$$

$$\leq \sum_{m=1}^{M} p_m \max_{1 \leq i \leq N} \hat{q}_i^\alpha F_m^i. \tag{39}$$

Since $|\dot{s}_m(t) - p_m| < \epsilon$, for each m, we have $\dot{s}_m(t) \geq p_m - \epsilon$. Hence, using Proposition 5 again,

$$\begin{aligned}
\dot{V}_\alpha(\boldsymbol{q}(t)) &\leq \left[\sum_{i=1}^{N}(q_i(t))^{\alpha+1}\right]^{\frac{-\alpha}{\alpha+1}} \left[\sum_{i=1}^{N}(q_i(t))^{\alpha}\lambda_i \right. \\
&\qquad \left. - \sum_{m=1}^{M}(p_m - \epsilon) \max_{1\leq k\leq N}(q_k(t))^{\alpha}F_m^k\right] \\
&\leq \left[\sum_{i=1}^{N}(q_i(t))^{\alpha+1}\right]^{\frac{-\alpha}{\alpha+1}} \left[-\acute{\epsilon}\sum_{i=1}^{N}(q_i(t))^{\alpha}\lambda_i \right. \\
&\qquad \left. + \epsilon\sum_{m=1}^{M}\sum_{i=1}^{N}(q_i(t))^{\alpha}F_m^i\right] \\
&\leq -\acute{\epsilon}\left[\sum_{i=1}^{N}(q_i(t))^{\alpha+1}\right]^{\frac{-\alpha}{\alpha+1}} \left[\sum_{i=1}^{N}(q_i(t))^{\alpha}\lambda_i\right] \\
&\qquad +\epsilon\left[\sum_{i=1}^{N}(q_i(t))^{\alpha+1}\right]^{\frac{-\alpha}{\alpha+1}} \left[\sum_{m=1}^{M}\sum_{i=1}^{N}(q_i(t))^{\alpha}F_m^i\right].
\end{aligned}$$

We will simplify the above expression. For the first term, we use the following two inequalities

$$\left[\sum_{i=1}^{N}(q_i(t))^{\alpha+1}\right]^{\frac{\alpha}{\alpha+1}} \leq N^{\frac{\alpha}{\alpha+1}} \max_{1\leq i\leq N}(q_i(t))^{\alpha}$$

and

$$\max_{1\leq i\leq N}(q_i(t))^{\alpha} \leq \frac{\sum_{i=1}^{N}(q_i(t))^{\alpha}\lambda_i}{\min\limits_{1\leq i\leq N}\lambda_i}.$$

We then have,

$$-\acute{\epsilon}\left[\sum_{i=1}^{N}(q_i(t))^{\alpha+1}\right]^{\frac{-\alpha}{\alpha+1}} \left[\sum_{i=1}^{N}(q_i(t))^{\alpha}\lambda_i\right] \leq \frac{-\acute{\epsilon}\min\limits_{1\leq i\leq N}\lambda_i}{N^{\frac{\alpha}{\alpha+1}}}.$$

For the second term, we use Holder's inequality again,

$$\left[\sum_{i=1}^{N}(q_i(t))^{\alpha+1}\right]^{\frac{\alpha}{\alpha+1}} \left[\sum_{i=1}^{N}(F_m^i)^{\alpha+1}\right]^{\frac{1}{\alpha+1}} \geq \sum_{i=1}^{N}(q_i(t))^{\alpha}F_m^i.$$

Summing over the index m on both sides and rearranging the terms, we then have,

$$\begin{aligned}
& \epsilon\left[\sum_{i=1}^{N}(q_i(t))^{\alpha+1}\right]^{\frac{-\alpha}{\alpha+1}}\left[\sum_{m=1}^{M}\sum_{i=1}^{N}(q_i(t))^{\alpha}F_m^i\right] \\
\leq \quad & \epsilon\sum_{m=1}^{M}\left[\sum_{i=1}^{N}(F_m^i)^{\alpha+1}\right]^{\frac{1}{\alpha+1}}.
\end{aligned}$$

The result then follows. *Q.E.D.*

We next establish a result for the drift of the Lyapunov function of the backlog process in discrete time. We will need the following Lemma.

Lemma 15 *Consider two N-dimensional vectors* $\boldsymbol{x} \geq 0$ *and* $\boldsymbol{\Delta x}$ *such that* $\boldsymbol{x} + \boldsymbol{\Delta x} \geq 0$. *Assume that there exist* $L_i, i = 1, ..., N$ *such that* $|\Delta x_i| < L_i$ *for all* i *and* $V_\alpha(\boldsymbol{x}) > \sum_{i=1}^{N} L_i$. *The following holds,*

$$V_\alpha(\boldsymbol{x}+\boldsymbol{\Delta x}) \leq V_\alpha(\boldsymbol{x}) + \frac{\sum_{i=1}^{N}(x_i)^{\alpha}\Delta x_i}{(V_\alpha(\boldsymbol{x}))^{\alpha}} + \alpha\frac{\sum_{i=1}^{N}(\Delta x_i)^2}{V_\alpha(\boldsymbol{x}) - \sum_{i=1}^{N} L_i}.$$

Proof: Consider the function

$$\begin{aligned}
y(t) &= V_\alpha(\boldsymbol{x} + t\boldsymbol{\Delta x}) \\
&= \left[\sum_{i=1}^{N}(x_i + t\Delta x_i)^{\alpha+1}\right]^{\frac{1}{\alpha+1}}.
\end{aligned}$$

Using the Mean-Value-Theorem, we have

$$y(1) = y(0) + \frac{dy(0)}{dt} + \frac{1}{2}\frac{d^2y(\acute{t})}{dt^2} \text{ for some } \acute{t} \in [0,1]. \tag{40}$$

The derivatives of $y(\cdot)$ are

$$\frac{dy(t)}{dt} = \frac{\sum_{i=1}^{N}(x_i + t\Delta x_i)^{\alpha}\Delta x_i}{(V_\alpha(\boldsymbol{x} + t\boldsymbol{\Delta x}))^{\alpha}} \tag{41}$$

$$\begin{aligned}
\frac{d^2y(t)}{dt^2} &= \alpha\frac{\sum_{i=1}^{N}(\Delta x_i)^2(x_i + t\Delta x_i)^{\alpha-1}}{(V_\alpha(\boldsymbol{x} + t\boldsymbol{\Delta x}))^{\alpha}} \\
&\quad -\alpha\frac{\sum_{i=1}^{N}(\Delta x_i)^2(x_i + t\Delta x_i)^{2\alpha}}{(V_\alpha(\boldsymbol{x} + t\boldsymbol{\Delta x}))^{2\alpha+1}}.
\end{aligned} \tag{42}$$

Since $x_i + t\Delta x_i \le V_\alpha(\boldsymbol{x} + \boldsymbol{\Delta x})$, we have from (42) that

$$\frac{d^2 y(t)}{dt^2} \le \alpha \frac{\sum_{i=1}^{N} (\Delta x_i)^2}{V_\alpha(\boldsymbol{x} + t\boldsymbol{\Delta x})}.$$

Further, by the convexity of $V_\alpha(\cdot)$ and $y(\cdot)$, we have

$$\begin{aligned} V_\alpha(\boldsymbol{x} + t\boldsymbol{\Delta x}) &= y(t) \\ &\ge y(0) + t\frac{dy(0)}{dt} \\ &= V_\alpha(\boldsymbol{x}) + \frac{t\sum_{i=1}^{N}(x_i)^\alpha \Delta x_i}{(V_\alpha(\boldsymbol{x}))^\alpha}. \end{aligned}$$

Using the fact that $x_i \le V_\alpha(\boldsymbol{x})$ and Δx_i is bounded below by $-L_i$, we have $V_\alpha(\boldsymbol{x} + t\boldsymbol{\Delta x}) \ge V_\alpha(\boldsymbol{x}) - \sum_{i=1}^{N} L_i$ for $t \in [0, 1]$. This implies that

$$\frac{d^2 y(t)}{dt^2} \le \alpha \frac{\sum_{i=1}^{N} (\Delta x_i)^2}{V_\alpha(\boldsymbol{x}) - \sum_{i=1}^{N} L_i}.$$

This inequality, along with (41) and (40), proves the lemma. *Q.E.D.*

Using this lemma, we can then provide a bound on the drift of the Lyapunov function in discrete time. Define $M_i = \max\{\lambda_i, \max_{1\le m\le M} F_m^i\}$. Define $\mathcal{F}_{\frac{n}{B}}$ as the σ-field generated by random variables $(\boldsymbol{q}^B(\frac{1}{B}), \boldsymbol{q}^B(\frac{2}{B}), \ldots, \boldsymbol{q}^B(\frac{n}{B}))$.

Corollary 16 *Assume that $V_\alpha(\boldsymbol{q}^B(\frac{n-1}{B})) \ge \delta > 0$. There exists $B_0 > 0$ (which may depend on δ) such that for all $B \ge B_0$, the following holds,*

$$\begin{aligned} &\mathbf{E}\left[V_\alpha(\boldsymbol{q}^B(\frac{n}{B})) - V_\alpha(\boldsymbol{q}^B(\frac{n-1}{B}))|\mathcal{F}_{\frac{n-1}{B}}\right] \\ \le\ & \frac{-\acute{\epsilon}\min_{1\le i\le N}\lambda_i}{BN^{\frac{\alpha}{\alpha+1}}} + \frac{\sum_{i=1}^{N}(\frac{M_i}{B})^{\alpha+1}}{\delta^\alpha} + \alpha\frac{\sum_{i=1}^{N}(\frac{M_i}{B})^2}{\delta - \sum_{i=1}^{N}\frac{M_i}{B}}. \end{aligned}$$

Proof: Let $\boldsymbol{x} = \boldsymbol{q}^B(\frac{n-1}{B})$ and $\boldsymbol{x} + \boldsymbol{\Delta x} = \boldsymbol{q}^B(\frac{n}{B})$. Then from (1), we have $|\Delta x_i| = |q_i^B(\frac{n}{B}) - q_i^B(\frac{n-1}{B})| \le \frac{M_i}{B}$. Let $B_0 = \frac{2\sum_{i=1}^{N} M_i}{\delta}$. Then, for all $B \ge B_0$, we have $\sum_{i=1}^N \frac{M_i}{B} \leq \frac{\delta}{2} < V_\alpha(\boldsymbol{q}^B(\frac{n-1}{B}))$. Therefore, we can apply Lemma 15 by taking $L_i = \frac{M_i}{B}$, and we have,

$$\begin{aligned} & V_\alpha(\boldsymbol{q}^B(\frac{n}{B})) - V_\alpha(\boldsymbol{q}^B(\frac{n-1}{B})) \qquad (43) \\ \leq \quad & \frac{\sum_{i=1}^N (q_i^B(\frac{n-1}{B}))^\alpha \Delta x_i}{V_\alpha(\boldsymbol{q}^B(\frac{n-1}{B}))^\alpha} + \alpha \frac{\sum_{i=1}^N (\Delta x_i)^2}{V_\alpha(\boldsymbol{q}^B(\frac{n-1}{B})) - \sum_{i=1}^N \frac{M_i}{B}}. \end{aligned}$$

The last term can be bounded by

$$\alpha \frac{\sum_{i=1}^N (\Delta x_i)^2}{V_\alpha(\boldsymbol{q}^B(\frac{n-1}{B})) - \sum_{i=1}^N \frac{M_i}{B}} \leq \alpha \frac{\sum_{i=1}^N (\frac{M_i}{B})^2}{\delta - \sum_{i=1}^N \frac{M_i}{B}}. \qquad (44)$$

It remains to bound the conditional expectation of the first term on the right-hand-side of (43). Note that when $q_i^B(\frac{n-1}{B}) > \frac{M_i}{B}$, we have

$$\Delta x_i = \frac{\lambda_i - \sum_{m=1}^M \mathbf{1}_{\{C(n-1)=m\}} \mathbf{1}_{\{U(n-1)=i\}} F_m^i}{B}.$$

When $q_i^B(\frac{n-1}{B}) \leq \frac{M_i}{B}$, we have

$$\Delta x_i \leq \frac{\lambda_i - \sum_{m=1}^M \mathbf{1}_{\{C(n-1)=m\}} \mathbf{1}_{\{U(n-1)=i\}} F_m^i}{B} + \frac{M_i}{B}.$$

Combining the two cases, we have,

$$\begin{aligned} & \sum_{i=1}^N (q_i^B(\frac{n-1}{B}))^\alpha \Delta x_i \\ \leq \quad & \sum_{i=1}^N (\frac{M_i}{B})^{\alpha+1} + \sum_{i=1}^N (q_i^B(\frac{n-1}{B}))^\alpha \frac{\lambda_i}{B} \\ & - \sum_{i=1}^N (q_i^B(\frac{n-1}{B}))^\alpha \frac{\sum_{m=1}^M \mathbf{1}_{\{C(n-1)=m, U(n-1)=i\}} F_m^i}{B}. \end{aligned}$$

Taking expectation over the channel state m, we have,

$$\begin{aligned}
&\mathbf{E}\left[\sum_{i=1}^{N}(q_i^B(\frac{n-1}{B}))^\alpha \Delta x_i|\mathcal{F}_{\frac{n-1}{B}}\right] \\
\leq\ & \sum_{i=1}^{N}(\frac{M_i}{B})^{\alpha+1}+\sum_{i=1}^{N}(q_i^B(\frac{n-1}{B}))^\alpha\frac{\lambda_i}{B} \\
& -\frac{\sum_{m=1}^{M} p_m \max_{1\leq k\leq N}(q_k^B(\frac{n-1}{B}))^\alpha F_m^k}{B},
\end{aligned}$$

where we have used the fact that the scheduling algorithm chooses to serve the link i with the largest value of $(q_i^B(\frac{n-1}{B}))^\alpha F_m^i$.

Now, since $\boldsymbol{\lambda}(1+\acute{\epsilon})$ is in the capacity region, using (38), we must have,

$$\begin{aligned}
&\mathbf{E}\left[\sum_{i=1}^{N}(q_i^B(\frac{n-1}{B}))^\alpha \Delta x_i|\mathcal{F}_{\frac{n-1}{B}}\right] \\
\leq\ & -\acute{\epsilon}\sum_{i=1}^{N}(q_i^B(\frac{n-1}{B}))^\alpha\frac{\lambda_i}{B}+\sum_{i=1}^{N}(\frac{M_i}{B})^{\alpha+1}.
\end{aligned}$$

Substiting this inequality and (44) to (43), we have

$$\begin{aligned}
&\mathbf{E}\left[V_\alpha(\boldsymbol{q}^B(\frac{n}{B}))-V_\alpha(\boldsymbol{q}^B(\frac{n-1}{B}))|\mathcal{F}_{\frac{n-1}{B}}\right] \\
\leq\ & -\frac{\acute{\epsilon}\sum_{i=1}^{N}(q_i^B(\frac{n-1}{B}))^\alpha\frac{\lambda_i}{B}}{V_\alpha(\boldsymbol{q}^B(\frac{n-1}{B}))^\alpha}+\frac{\sum_{i=1}^{N}(\frac{M_i}{B})^{\alpha+1}}{\delta^\alpha} \\
& +\alpha\frac{\sum_{i=1}^{N}(\frac{M_i}{B})^2}{\delta-\sum_{i=1}^{N}\frac{M_i}{B}}.
\end{aligned}$$

The result then follows by noting that

$$\begin{aligned}
\sum_{i=1}^{N}(q_i^B(\frac{n-1}{B}))^\alpha &\geq \max_{1\leq i\leq N}(q_i^B(\frac{n-1}{B}))^\alpha \\
&\geq \frac{1}{N^{\frac{\alpha}{\alpha+1}}}V_\alpha(\boldsymbol{q}^B(\frac{n-1}{B}))^\alpha.
\end{aligned}$$

Q.E.D.

E Bounding the Decay-rate of the Stationary Overflow Probability: The Proof using Freidlin-Wentzell Theory

We next prove (34) using Freidlin-Wentzell theory [11,25]. Let $\mathbf{P}^B(\cdot)$ denote the stationary probability distribution of $\boldsymbol{q}^B(t)$. Since $\boldsymbol{\lambda}(1+\acute{\epsilon})$ is in the capacity region, the backlog process is a stable Markov process. We define the following stopping times. Fix any $\rho \in (0,1)$. Choose δ, ϵ and C such that $0 < \delta < \epsilon < \rho < 1 < C$. For each B, consider the following sequence of stopping times. (For a similar construction, see section 8.4 in [25].)

$$
\begin{aligned}
\beta_0^B &= \frac{\lceil \inf\{t \geq 0 | V_\alpha(\boldsymbol{q}^B(t)) \leq \delta\} B \rceil}{B} \\
\eta_j^B &= \frac{\lceil \inf\{t \geq \beta_j^B | V_\alpha(\boldsymbol{q}^B(t)) \geq \epsilon\} B \rceil}{B}, \; j = 1, 2, \ldots \\
\beta_j^B &= \frac{\lceil \inf\{t \geq \eta_{j-1}^B | V_\alpha(\boldsymbol{q}^B(t)) \leq \delta\} B \rceil}{B}, \; j = 2, 3 \ldots
\end{aligned}
$$

Consider the Markov chain $\hat{\boldsymbol{q}}^B$ obtained by sampling $\boldsymbol{q}^B(t)$ at the stopping times η_j^B, i.e., $\hat{\boldsymbol{q}}^B(j) = \boldsymbol{q}^B(\eta_j^B)$, $j = 1, 2, \ldots$ Let $\hat{\mathbf{P}}^B$ denote the stationary distribution of this chain and let θ^B denote its state space. Note that when B is sufficiently large, we must have $V_\alpha(\boldsymbol{x}) \leq \rho$ for every $\boldsymbol{x} \in \theta^B$.

The following equation provides a way to compute the stationary distribution of $\boldsymbol{q}^B(\cdot)$. Let $\mathbf{P}_{\boldsymbol{x}}$ denote the distribution conditioned on $\boldsymbol{q}^B(0) = \boldsymbol{x}$, and let $\mathbf{E}_{\boldsymbol{x}}$ denote the expectation taken with repect to $\mathbf{P}_{\boldsymbol{x}}$. Then, the following holds (see [19, Lemma 10.1]),

$$
\begin{aligned}
&\mathbf{P}^B(V_\alpha(\boldsymbol{q}^B(0)) \geq 1) \\
= \; & \frac{\int_{\theta^B} \hat{\mathbf{P}}^B(d\boldsymbol{x}) \mathbf{E}_{\boldsymbol{x}}(\int_0^{\eta_1^B} \mathbf{1}_{\{V_\alpha(\boldsymbol{q}^B(t)) \geq 1\}} dt)}{\int_{\theta^B} \hat{\mathbf{P}}^B(d\boldsymbol{x}) \mathbf{E}_{\boldsymbol{x}}(\eta_1^B)}. \qquad (45)
\end{aligned}
$$

Recall that we are interested in the asymptotic decay-rate of the left hand side. Hence, we can use (45) to bound the limit $\limsup_{B\to\infty} \frac{1}{B} \log(\mathbf{P}^B(V_\alpha(\boldsymbol{q}^B(0)) > 1))$ as follows. First, we will show that the denominator in (45) is bounded from below. Then, we will show that the numerator is bounded from above and we will estimate the asymptotics of this bound.

E.1 Bounding the Denominator of (45)

We first show that the denominator in (45) is bounded from below. Consider the following identity:

$$\begin{aligned} & V_\alpha(\boldsymbol{q}^B(\eta_1^B)) - V_\alpha(\boldsymbol{q}^B(\beta_1^B)) \\ = & \sum_{n=1}^{(\eta_1^B-\beta_1^B)B} \left[V_\alpha(\boldsymbol{q}^B(\beta_1^B + \frac{n}{B})) - V_\alpha(\boldsymbol{q}^B(\beta_1^B + \frac{n-1}{B})) \right]. \end{aligned} \tag{46}$$

By (1), we have that

$$q_i^B(\beta_1^B + \frac{n}{B}) \leq q_i^B(\beta_1^B + \frac{n-1}{B}) + \frac{\lambda_i}{B}.$$

Therefore, by the monotonicity and convexity of the Lyapunov function, we have

$$\begin{aligned} & V_\alpha(\boldsymbol{q}^B(\beta_1^B + \frac{n}{B})) - V_\alpha(\boldsymbol{q}^B(\beta_1^B + \frac{n-1}{B})) \\ \leq & V_\alpha(\boldsymbol{q}^B(\beta_1^B + \frac{n-1}{B}) + \frac{\boldsymbol{\lambda}}{B}) - V_\alpha(\boldsymbol{q}^B(\beta_1^B + \frac{n-1}{B})) \\ \leq & \frac{1}{B} V_\alpha(\boldsymbol{\lambda}). \end{aligned}$$

Substituting into (46), we have,

$$\begin{aligned} & V_\alpha(\boldsymbol{q}^B(\eta_1^B)) - V_\alpha(\boldsymbol{q}^B(\beta_1^B)) \\ \leq & \sum_{n=1}^{(\eta_1^B-\beta_1^B)B} \frac{1}{B} V_\alpha(\boldsymbol{\lambda}) \\ = & (\eta_1^B - \beta_1^B) V_\alpha(\boldsymbol{\lambda}). \end{aligned}$$

Now, for sufficiently large B, we must have

$$V_\alpha(\boldsymbol{q}^B(\eta_1^B)) \geq \epsilon - \frac{\epsilon - \delta}{4},$$

and

$$V_\alpha(\boldsymbol{q}^B(\beta_1^B)) \leq \delta + \frac{\epsilon - \delta}{4}.$$

We thus have,

$$\eta_1^B - \beta_1^B \geq \frac{V_\alpha(\boldsymbol{q}^B(\eta_1^B)) - V_\alpha(\boldsymbol{q}^B(\beta_1^B))}{V_\alpha(\boldsymbol{\lambda})} \geq \frac{\epsilon - \delta}{2V_\alpha(\boldsymbol{\lambda})}.$$

Hence, the denominator of (45) is bounded from below, i.e.,

$$\mathbf{E}_{\boldsymbol{x}}(\eta_1^B) \geq \mathbf{E}_{\boldsymbol{x}}(\eta_1^B - \beta_1^B) \geq \frac{\epsilon - \delta}{2V_\alpha(\boldsymbol{\lambda})}.$$

E.2 Bounding the Numerator of (45)

Next, we bound the numerator of (45) as follows. We will use the following additional stopping time in our analysis.

$$\eta^{B,\uparrow} \triangleq \frac{\lceil \inf \{t \geq 0 | V_\alpha(\boldsymbol{q}^B(t)) \geq 1\} B \rceil}{B}.$$

We then have,

$$\begin{aligned} & \mathbf{E}_{\boldsymbol{x}}(\int_0^{\eta_1^B} \mathbf{1}_{\{V_\alpha(\boldsymbol{q}^B(t)) \geq 1\}} dt) \\ \leq\ & \mathbf{E}_{\boldsymbol{x}}(\mathbf{1}_{\{\eta^{B,\uparrow} \leq \beta_1^B\}}(\beta_1^B - \eta^{B,\uparrow})) \\ =\ & \mathbf{E}_{\boldsymbol{x}}(\beta_1^B - \eta^{B,\uparrow} | \eta^{B,\uparrow} \leq \beta_1^B) \mathbf{P}_{\boldsymbol{x}}(\eta^{B,\uparrow} \leq \beta_1^B) \\ =\ & \mathbf{E}_{\boldsymbol{x}}[\mathbf{E}_{\boldsymbol{q}(\eta^{B,\uparrow})}(\beta_1^B) | \eta^{B,\uparrow} \leq \beta_1^B] \mathbf{P}_{\boldsymbol{x}}(\eta^{B,\uparrow} \leq \beta_1^B). \end{aligned}$$

For sufficiently large B, we must have $\boldsymbol{q}(\eta^{B,\uparrow}) \leq C$. Hence,

$$\begin{aligned} & \mathbf{E}_{\boldsymbol{x}}(\int_0^{\eta_1^B} \mathbf{1}_{\{V_\alpha(\boldsymbol{q}^B(t)) \geq 1\}} dt) \\ \leq\ & \left[\sup_{\{\boldsymbol{y} | V_\alpha(\boldsymbol{y}) \leq C\}} \mathbf{E}_{\boldsymbol{y}}(\beta_1^B) \right] \mathbf{P}_{\boldsymbol{x}}(\eta^{B,\uparrow} \leq \beta_1^B). \end{aligned}$$

For any $T > 0$, we can further bound the above probability by

$$\begin{aligned} & \mathbf{E}_{\boldsymbol{x}}(\int_0^{\eta_1^B} \mathbf{1}_{\{V_\alpha(\boldsymbol{q}^B(t)) \geq 1\}} dt) \\ \leq\ & \left[\sup_{\{\boldsymbol{y} | V_\alpha(\boldsymbol{y}) \leq C\}} \mathbf{E}_{\boldsymbol{y}}(\beta_1^B) \right] \left[\mathbf{P}_{\boldsymbol{x}}(\eta^{B,\uparrow} \leq T) \right. \\ & \left. + \mathbf{P}_{\boldsymbol{x}}(\beta_1^B \geq T) \right]. \end{aligned}$$

Therefore, for sufficiently large B, by combining the bound on the denominator and the bound on the numerator, we can bound (45) as follows:

$$\begin{aligned}
&\mathbf{P}^B(V_\alpha(\boldsymbol{q}^B(0)) \geq 1) \\
\leq \quad & \frac{\sup_{\{\boldsymbol{y} \mid V_\alpha(\boldsymbol{y}) \leq C\}} \mathbf{E}_{\boldsymbol{y}}(\beta_1^B)}{\frac{\epsilon-\delta}{2V_\alpha(\boldsymbol{\lambda})}} \\
& \times \left[\mathbf{P}_{\boldsymbol{x}}(\eta^{B,\uparrow} \leq T) + \mathbf{P}_{\boldsymbol{x}}(\beta_1^B \geq T)\right]. \qquad (47)
\end{aligned}$$

Next, we show that $\sup_{\{\boldsymbol{y}|V_\alpha(\boldsymbol{y})\leq C\}} \mathbf{E}_{\boldsymbol{y}}(\beta_1^B)$ is bounded from above and hence does not influence the asymptotic decay-rate of $\mathbf{P}^B(V_\alpha(\boldsymbol{q}^B(0)) \geq 1)$. For $\tilde{M} > 0$ such that $B\tilde{M}$ is an integer, define the truncated stopping time

$$\beta_1^{B,\tilde{M}} = \min\{\tilde{M}, \beta_1^B\}.$$

By Dynkin's formula (see [29, Thm 11.3.1]), we have

$$\begin{aligned}
&\mathbf{E}_{\boldsymbol{y}}(V_\alpha(\boldsymbol{q}^B(\beta_1^{B,\tilde{M}}))) = \mathbf{E}_{\boldsymbol{y}}(V_\alpha(\boldsymbol{q}^B(0))) \\
&+\mathbf{E}_{\boldsymbol{y}}\left(\sum_{n=1}^{B\beta_1^{B,\tilde{M}}} \mathbf{E}\left[V_\alpha(\boldsymbol{q}^B(\frac{n}{B})) - V_\alpha(\boldsymbol{q}^B(\frac{n-1}{B}))|\mathcal{F}_{\frac{n-1}{B}}\right]\right). \qquad (48)
\end{aligned}$$

By definition, we have $\mathbf{E}_{\boldsymbol{y}}(V_\alpha(\boldsymbol{q}^B(0))) = V_\alpha(\boldsymbol{y})$. Further, by our construction, we have $\boldsymbol{q}^B(\frac{n-1}{B}) \geq \delta$ for $n = 1$ to $B\beta_1^{B,\tilde{M}}$. Therefore, by Corollary 16, we have, for all sufficiently large B,

$$\begin{aligned}
&\mathbf{E}\left[V_\alpha(\boldsymbol{q}^B(\frac{n}{B})) - V_\alpha(\boldsymbol{q}^B(\frac{n-1}{B}))|\mathcal{F}_{\frac{n-1}{B}}\right] \\
\leq \quad & \frac{-\acute{\epsilon} \min\limits_{1\leq i\leq N} \lambda_i}{BN^{\frac{\alpha}{\alpha+1}}} + \frac{\sum_{i=1}^N (\frac{M_i}{B})^{\alpha+1}}{\delta^\alpha} + \alpha\frac{\sum_{i=1}^N (\frac{M_i}{B})^2}{\delta - \sum_{i=1}^N \frac{M_i}{B}}.
\end{aligned}$$

By choosing large B, we can make the second and third terms on the right-hand-side to be smaller than $\frac{\acute{\epsilon} \min_{1\leq i\leq N} \lambda_i}{2BN^{\frac{\alpha}{\alpha+1}}}$. We then have, for all sufficiently large B,

$$\mathbf{E}\left[V_\alpha(\boldsymbol{q}^B(\frac{n}{B})) - V_\alpha(\boldsymbol{q}^B(\frac{n-1}{B}))|\mathcal{F}_{\frac{n-1}{B}}\right] \leq \frac{-\acute{\epsilon} \min\limits_{1\leq i\leq N} \lambda_i}{2BN^{\frac{\alpha}{\alpha+1}}}.$$

Substituting into (48) and using $\mathbf{E}_{\boldsymbol{y}}(V_\alpha(\boldsymbol{q}^B(\beta_1^{B,\tilde{M}}))) \geq 0$, we obtain

$$\mathbf{E}_{\boldsymbol{y}}(\beta_1^{B,\tilde{M}}) \leq \frac{V_\alpha(\boldsymbol{y})}{\frac{\epsilon \min_{1\leq i\leq N} \lambda_i}{2N^{\frac{\alpha}{\alpha+1}}}}.$$

Note that this is true for all $\tilde{M}$. Letting $\tilde{M} \to \infty$, and using the Monotone Convergence Theorem [30, p208], we then have

$$\mathbf{E}_{\boldsymbol{y}}(\beta_1^B) \leq \frac{2CN^{\frac{\alpha}{\alpha+1}}}{\epsilon \min_{1\leq i\leq N} \lambda_i},$$

for all $\boldsymbol{y}$ such that $V_\alpha(\boldsymbol{y}) \leq C$. Hence, we conclude that $\sup_{\{\boldsymbol{y}|V_\alpha(\boldsymbol{y})\leq C\}} \mathbf{E}_{\boldsymbol{y}}(\beta_1^B)$ is bounded from above by a constant.

Substituting the above bound into (47), we can then conclude that

$$\begin{aligned}
&\limsup_{B\to\infty} \frac{1}{B} \log(\mathbf{P}^B(V_\alpha(\boldsymbol{q}^B(0)) > 1))) \\
\leq\ & \max\left\{ \limsup_{B\to\infty} \frac{1}{B} \log(\sup_{\{\boldsymbol{x}\ |\ V_\alpha(\boldsymbol{x})\leq\rho\}} \mathbf{P}_{\boldsymbol{x}}(\eta^{B,\uparrow} \leq T)), \right. \\
&\left. \qquad \limsup_{B\to\infty} \frac{1}{B} \log(\sup_{\{\boldsymbol{x}\ |\ V_\alpha(\boldsymbol{x})\leq\rho\}} \mathbf{P}_{\boldsymbol{x}}(\beta_1^B \geq T)) \right\}.
\end{aligned} \tag{49}$$

In the rest of the section, we will bound the two terms on the right-hand-side. Specifically, we will show that

$$\limsup_{B\to\infty} \frac{1}{B} \log(\sup_{\{\boldsymbol{x}\ |\ V_\alpha(\boldsymbol{x})\leq\rho\}} \mathbf{P}_{\boldsymbol{x}}(\eta^{B,\uparrow} \leq T)) \leq -(1-\rho)J_\alpha. \tag{50}$$

Further, by choosing large T, we can make

$$\limsup_{B\to\infty} \frac{1}{B} \log(\sup_{\{\boldsymbol{x}\ |\ V_\alpha(\boldsymbol{x})\leq\rho\}} \mathbf{P}_{\boldsymbol{x}}(\beta_1^B \geq T)) < -(1-\rho)J_\alpha. \tag{51}$$

The result in (34) then follows by letting $\rho \to 0$.

E.3 Asymptotics for $\sup_{\{\boldsymbol{x} \mid V_\alpha(\boldsymbol{x}) \le \rho\}} \mathbf{P}_{\boldsymbol{x}}(\eta^{B,\uparrow} \le T)$

We first show (50). Let $\mathcal{X} = \{\boldsymbol{x} \mid V_\alpha(\boldsymbol{x}) \le \rho\}$ and let

$$\begin{aligned}\Gamma_{\le\rho} \;\triangleq\; & \{\boldsymbol{x}(\cdot) \mid V_\alpha(\boldsymbol{x}(0)) \le \rho \text{ and} \\ & \quad V_\alpha(\boldsymbol{x}(t)) \ge 1 \text{ for some } t \in (0,T]\}.\end{aligned}$$

Further, let $\hat{\Gamma}_{\le\rho}$ be the set of fluid sample paths $(\boldsymbol{s}(\cdot), \boldsymbol{q}(\cdot))$ such that $\boldsymbol{q}(\cdot)$ is in $\Gamma_{\le\rho}$. By Proposition 13, we have

$$\begin{aligned}& \limsup_{B\to\infty} \frac{1}{B} \log(\sup_{\{\boldsymbol{x}\in\mathcal{X}\}} \mathbf{P}_{\boldsymbol{x}}(\boldsymbol{q}^B(\cdot) \in \Gamma_{\le\rho})) \\ \le & -\inf_{(\boldsymbol{s}(\cdot),\boldsymbol{q}(\cdot))\in\hat{\Gamma}_{\le\rho}} \int_0^T H(\dot{\boldsymbol{s}}(t)||\boldsymbol{p})dt. \end{aligned} \tag{52}$$

It remains to show that for any fluid sample path $(\boldsymbol{s}(\cdot), \boldsymbol{q}(\cdot))$ in $\hat{\Gamma}_{\le\rho}$, its cost $\int_0^T H(\dot{\boldsymbol{s}}(t)||\boldsymbol{p})dt$ must be no smaller than $(1-\rho)J_\alpha$. The proof technique is similar to that of Proposition 6. Specifically, for any $(\boldsymbol{s}(\cdot), \boldsymbol{q}(\cdot)) \in \hat{\Gamma}_{\le\rho}$, let T' by the first time that $V_\alpha(\boldsymbol{q}(t) \ge 1$. Without loss of generality, assume that $\boldsymbol{q}(t) \ne 0$ for $t \in (0,T')$. Using similar techniques as in the proof of Proposition 6, we can show that for all $t \in (0,T')$ where $\boldsymbol{q}(\cdot)$ is differentiable, we have,

$$\dot{V}_\alpha(\boldsymbol{q}(t)) \le a(\dot{\boldsymbol{s}}(t)).$$

Hence, if $\dot{V}_\alpha(\boldsymbol{q}(t)) > 0$, we must have $a(\dot{\boldsymbol{s}}(t)) > 0$. Then, using the definition of J_α in (23), we have

$$H(\dot{\boldsymbol{s}}(t)||\boldsymbol{p}) \ge J_\alpha \dot{V}_\alpha(\boldsymbol{q}(t)).$$

On the other hand, if $\dot{V}_\alpha(\boldsymbol{q}(t)) \le 0$, the above inequality also holds trivially. Hence, the cost of the path must satisfy

$$\begin{aligned}& \int_0^T H(\dot{\boldsymbol{s}}(t)||\boldsymbol{p})dt \\ \ge & \int_0^{T'} H(\dot{\boldsymbol{s}}(t)||\boldsymbol{p})dt \\ \ge & \; J_\alpha \int_0^{T'} \dot{V}_\alpha(\boldsymbol{q}(t))dt.\end{aligned}$$

Recall that any fluid sample path in $\hat{\Gamma}_{\le\rho}$ must satisfy $\boldsymbol{q}(0) \le \rho$. Therefore,

$$\int_0^{T'} \dot{V}_\alpha(\boldsymbol{q}(t)) \ge (1-\rho).$$

Hence, by (52), we conclude that

$$\begin{aligned} & \limsup_{B\to\infty} \frac{1}{B} \log(\sup_{\{\boldsymbol{x} \mid V_\alpha(\boldsymbol{x}) \le \rho\}} \mathbf{P}_{\boldsymbol{x}}(\eta^{B,\uparrow} \le T)) \\ \le & -\inf_{(\boldsymbol{s}(\cdot),\boldsymbol{q}(\cdot))\in\hat{\Gamma}_{\le\rho}} \int_0^T H(\dot{\boldsymbol{s}}(t)||\boldsymbol{p})dt \\ \le & -(1-\rho)J_\alpha. \end{aligned}$$

E.4 Asymptotics for $\sup_{\{\boldsymbol{x} \mid V_\alpha(\boldsymbol{x})\le\rho\}} \mathbf{P}_{\boldsymbol{x}}(\beta_1^B \ge T)$

We next show that (51) holds if we choose a large T. Let $\mathcal{X} = \{\boldsymbol{x} \mid V_\alpha(\boldsymbol{x}) \le \rho\}$ and redefine

$$\begin{aligned} \Gamma_{\le\rho} \triangleq & \{\boldsymbol{x}(\cdot) \mid V_\alpha(\boldsymbol{x}(0)) \le \rho \text{ and} \\ & V_\alpha(\boldsymbol{x}(t)) > \delta \text{ for all } t\in[0,T-1]\}. \end{aligned}$$

Further, redefine $\hat{\Gamma}_{\le\rho}$ to be the set of fluid sample paths $(\boldsymbol{s}(\cdot),\boldsymbol{q}(\cdot))$ such that $\boldsymbol{q}(\cdot)$ is in $\Gamma_{\le\rho}$. Using Proposition 13 again, we have

$$\begin{aligned} & \limsup_{B\to\infty} \frac{1}{B} \log(\sup_{\{\boldsymbol{x}\in\mathcal{X}\}} \mathbf{P}_{\boldsymbol{x}}(\boldsymbol{q}^B(\cdot)\in\Gamma_{\le\rho})) \\ \le & -\inf_{(\boldsymbol{s}(\cdot),\boldsymbol{q}(\cdot))\in\hat{\Gamma}_{\le\rho}} \int_0^T H(\dot{\boldsymbol{s}}(t)||\boldsymbol{p})dt. \end{aligned}$$

By definition, for any $(\boldsymbol{s}(\cdot),\boldsymbol{q}(\cdot)) \in \hat{\Gamma}_{\le\rho}$, we have $V_\alpha(\boldsymbol{q}(t)) \ge \delta$ for $t\in[0,T-1]$. Hence,

$$\begin{aligned} \delta & \le V_\alpha(\boldsymbol{q}(0)) + \int_0^{T-1} \dot{V}_\alpha(\boldsymbol{q}(t))dt \\ & \le \rho + \int_0^{T-1} \dot{V}_\alpha(\boldsymbol{q}(t))dt. \end{aligned}$$

Now, let

$$\epsilon = \acute{\epsilon} \frac{\min_{1\le i\le N}\lambda_i}{2N^{\frac{\alpha}{\alpha+1}} \sum_{m=1}^M (\sum_{i=1}^N (F_m^i)^{\alpha+1})^{\frac{1}{\alpha+1}}}.$$

Applying Proposition 14, we obtain,

$$\begin{aligned}\delta \;\le\; & \rho + \int_0^{T-1} \mathbf{1}_{\{|\dot{\boldsymbol{s}}(t)-\boldsymbol{p}|<\epsilon\}} \left(\frac{-\acute{\epsilon} \min_{1\le i\le N} \lambda_i}{2N^{\alpha/(\alpha+1)}} \right) dt \\ & + \int_0^{T-1} \mathbf{1}_{\{|\dot{\boldsymbol{s}}(t)-\boldsymbol{p}|\ge\epsilon\}} \left[\left(\sum_{i=1}^{N} \lambda_i^{\alpha+1} \right)^{1/(\alpha+1)} \right] dt.\end{aligned}$$

From the above inequality, we then obtain the following bound

$$\begin{aligned}& \int_0^{T-1} \mathbf{1}_{\{|\dot{\boldsymbol{s}}(t)-\boldsymbol{p}|\ge\epsilon\}} dt \\ \ge \;& \frac{(T-1)\frac{\acute{\epsilon}\min_{1\le i\le N}\lambda_i}{2N^{\alpha/(\alpha+1)}} + \delta - \rho}{\frac{\acute{\epsilon}\min_{1\le i\le N}\lambda_i}{2N^{\alpha/(\alpha+1)}} + (\sum_{i=1}^{N}\lambda_i^{\alpha+1})^{1/(\alpha+1)}}. \end{aligned} \tag{53}$$

Now, note that,

$$\begin{aligned}& \int_0^{T} H(\dot{\boldsymbol{s}}(t)||\boldsymbol{p}) dt \\ \ge \;& \left[\min_{|\dot{\boldsymbol{s}}(t)-\boldsymbol{p}|\ge\epsilon} H(\dot{\boldsymbol{s}}(t)||\boldsymbol{p}) \right] \int_0^{T-1} \mathbf{1}_{\{|\dot{\boldsymbol{s}}(t)-\boldsymbol{p}|\ge\epsilon\}} dt.\end{aligned}$$

By (53), the quantity on the right-hand-side can be made arbitrarily large by increasing T. Inequality (51) then holds.

Finally, to complete the proof of (34), note that by substituting (50) and (51) to (49), we have shown that, for any $\rho \in (0,1)$,

$$\limsup_{B\to\infty} \frac{1}{B} \log(\mathbf{P}^B(V_\alpha(\boldsymbol{q}^B(0)) \ge 1)) \le -(1-\rho)J_\alpha.$$

Since ρ can be made arbitrarily small, by letting $\rho \to 0$, the result (34) then follows.

References

[1] V. J. Venkataramanan and X. Lin, "Structural Properties of LDP for Queue-Length Based Wireless Scheduling Algorithms," in *45th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, September 2007.

[2] X. Lin, N. B. Shroff, and R. Srikant, "A Tutorial on Cross-Layer Optimization in Wireless Networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, August 2006.

[3] L. Tassiulas and A. Ephremides, "Stability Properties of Constrained Queueing Systems and Scheduling Policies for Maximum Throughput in Multihop Radio Networks," *IEEE Transactions on Automatic Control*, vol. 37, no. 12, pp. 1936–1948, December 1992.

[4] M. J. Neely, E. Modiano, and C. E. Rohrs, "Dynamic Power Allocation and Routing for Time Varying Wireless Networks," in *Proceedings of IEEE INFOCOM*, San Francisco, April 2003.

[5] A. Eryilmaz, R. Srikant, and J. Perkins, "Stable Scheduling Policies for Fading Wireless Channels," *IEEE/ACM Transactions on Networking*, vol. 13, no. 2, pp. 411–424, April 2005.

[6] A. L. Stolyar, "MaxWeight Scheduling in a Generalized Switch: State Space Collapse and Workload Minimization in Heavy Traffic," *Annals of Applied Probability*, vol. 14, no. 1, pp. 1–53, 2004.

[7] D. Shah and D. Wischik, "Optimal Scheduling Algorithms for Input-Queued Switches," in *Proceedings of IEEE INFOCOM*, Barcelona, Spain, April 2006.

[8] S. P. Meyn, "Stability and Asymptotic Optimality of Generalized MaxWeight Policies," *SIAM J. Control and Optimization*, to appear.

[9] M. J. Neely, "Order Optimal Delay for Opportunistic Scheduling in Multi-User Wireless Uplinks and Downlinks," *IEEE/ACM Transactions on Networking*, 2008.

[10] ——, "Delay Analysis for Maximal Scheduling in Wireless Networks with Bursty Traffic," in *Proceedings of IEEE INFOCOM*, Phoenix, AZ, April 2008.

[11] A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis: Queues, Communications, and Computing.* London: Chapman & Hall, 1995.

[12] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York: Springer-Verlag, 1998.

[13] A. I. Elwalid and D. Mitra, "Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High Speed Networks," *IEEE/ACM Transactions on Networking*, vol. 1, no. 3, pp. 329–343, June 1993.

[14] G. Kesidis, J. Walrand, and C.-S. Chang, "Effective Bandwidth for Multiclass Markov Fluid and other ATM Sources," *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, pp. 424–428, Aug. 1993.

[15] C.-S. Chang and J. A. Thomas, "Effective Bandwidth in High-Speed Digital Networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1091–1114, Aug. 1995.

[16] F. P. Kelly, "Effective Bandwidth in Multiclass Queues," *Queueing Systems*, vol. 9, pp. 5–16, 1991.

[17] W. Whitt, "Tail Probabilities with Statistical Multiplexing and Effective Bandwidth for Multi-class Queues," *Telecommunication Syst.*, vol. 2, pp. 71–107, 1993.

[18] A. L. Stolyar and K. Ramanan, "Largest Weighted Delay First Scheduling: Large Deviations and Optimality," *Annals of Applied Probability*, vol. 11, no. 1, pp. 1–48, 2001.

[19] A. L. Stolyar, "Control of End-to-End Delay Tails in a Multiclass Networks: LWDF Discipline Optimality," *Annals of Applied Probability*, vol. 13, no. 3, pp. 1151–1206, 2003.

[20] D. Wu and R. Negi, "Effective Capacity: A Wireless Link Model for Support of Quality of Service," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 630–643, July 2003.

[21] ——, "Downlink Scheduling in a Cellular Network for Quality of Service Assurance," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 5, pp. 1547–1557, September 2004.

[22] ——, "Utilizing Multiuser Diversity for Efficient Support of Quality of Service over a Fading Channel," *IEEE Transactions on Vehicular Technology*, vol. 54, no. 3, pp. 1198–1206, May 2005.

[23] L. Ying, R. Srikant, A. Eryilmaz, and G. E. Dullerud, "A Large Deviations Analysis of Scheduling in Wireless Networks," *IEEE Transactions on Information Theory*, vol. 52, no. 11, November 2006.

[24] S. Shakkottai, "Effective Capacity and QoS for Wireless Scheduling," *IEEE Transactions on Automatic Control*, vol. 53, no. 3, April 2008.

[25] A. L. Stolyar, "Large Deviations of Queues Sharing a Randomly Time-varying Server," *Queueing Systems*, vol. To appear, 2008.

[26] X. Lin, "On Characterizing the Delay Performance of Wireless Scheduling Algorithms," in *44th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, September 2006.

[27] A. Eryilmaz and R. Srikant, "Scheduling with Quality of Service Constraints over Rayleigh Fading Channels," in *Proceedings of the IEEE Conference on Decision and Control*, December 2004.

[28] D. Bertsimas, I. C. Paschalidis, and J. N. Tsitsiklis, "Asymptotic Buffer Overflow Probabilities in Multiclass Multiplexers: An Optimal Control Approach," *IEEE Transactions on Automatic Control*, vol. 43, no. 3, pp. 315–335, March 1998.

[29] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*. New York: Springer, 1993.

[30] P. Billingsley, *Probability and Measure*. New York: John Wiley & Son, 1979.