Necessity of Future Information in Admission Control

Kuang Xu
Stanford University
Graduate School of Business
Stanford, CA 94305

kuangxu@stanford.edu

We study the necessity of predictive information in a class of queueing admission control problems, where a system manager is allowed to divert incoming jobs up to a fixed rate, in order to minimize the queueing delay experienced by the admitted jobs.

Spencer et al. (2014) show that the system's delay performance can be significantly improved by having access to future information in the form of a lookahead window, during which the times of future arrivals and services are revealed. They prove that, while delay under an optimal online policy diverges to infinity in the heavy-traffic regime, it can stay *bounded* by making use of future information. However, the diversion policies of Spencer et al. (2014) require the length of the lookahead window to grow to infinity at a non-trivial rate in the heavy-traffic regime, and it remained open whether substantial performance improvement could still be achieved with *less* future information.

We resolve this question to a large extent by establishing an asymptotically tight lower bound on how much future information is necessary to achieve superior performance, which matches the upper bound of Spencer et al. (2014) up to a constant multiplicative factor. Our result hence demonstrates that the system's heavy-traffic delay performance is highly sensitive to the amount of future information available. Our proof is based on analyzing certain excursion probabilities of the input sample paths, and exploiting a connection between a policy's diversion decisions and subsequent server idling, which may be of independent interest for related dynamic resource allocation problems.

Key words: admission control, queueing, algorithm, future information, predictive model, heavy-traffic asymptotics

1. Introduction

Recently, there have been substantial interests in developing forecasting systems and predictive models across various application domains, which enable a system manager to obtain (partial) information of *future* inputs, and thus allow for more efficient decision making or resource allocation. Examples of these systems include advanced ordering in supply chains (Fisher and Raman (1996)), appointment booking for elective surgeries (Kim and Horowitz (2002)), and mechanisms for predicting future hospital visits (Wargon et al. (2009), Sun et al. (2009)). Because acquiring accurate predictions can often involve additional infrastructural investments and operational complexities, it is a natural question to ask *how useful* such predictive information can be, in terms of its ability in improving system performance beyond what can be achieved by the more conventional way of *online* decision making, which does not take predictive information into account.

In a recent paper, Spencer et al. (2014) initiated an investigation along this direction in a class of queueing admission control problems, illustrated in Figure 1. An overloaded queue with service rate $1-p$ receives incoming jobs at rate $\lambda \in (1-p, 1)$, and the system manager is allowed to *divert* incoming jobs up to a rate of p, with the objective of minimizing the time-average queueing delay among the admitted jobs. The system manager has access to a *lookahead window* of length W_λ, within which the realizations of future arrivals and service availability are revealed. The online version of the problem, with $W_\lambda = 0$, is a classical queueing model that has been studied in various contexts related to congestion control (Yechiali (1971), Stidham (1985)).

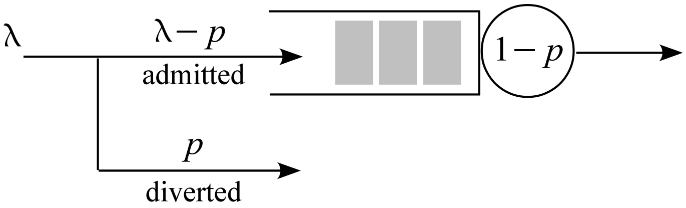

Figure 1 An illustration of the queueing admission control problem.

A main message of Spencer et al. (2014) is that one can drastically reduce queueing delay with a *sufficient* amount of future information. In particular, there exists $c_h > 0$, such that if the length of the lookahead window satisfies

$$W_\lambda \geq c_h \ln \frac{1}{1-\lambda}, \tag{1}$$

then there exists a sequence of diversion policies, so that the resultant delay will stay *bounded* in the heavy-traffic regime of $\lambda \to 1$. In sharp contrast, when no future information is available, the delay under an optimal online policy will diverge to *infinity*, as $\lambda \to 1$.

However, the requirement on the length of the lookahead window, as in Eq. (1), means that the superior delay performance achieved by Spencer et al. (2014) comes at the expense of a non-trivial amount of predictive power. Therefore, it remains to determine whether one could use much less future information and still achieve a significant performance improvement over an optimal online policy. This question is of practical importance, because a larger amount of future information often requires more sophisticated predictive models and computational infrastructures, which can be costly, if not impossible, to build and operate.

The main contribution of the present paper is to provide a negative answer to above question, by showing that there exists a positive constant, c_l, such that if W_λ scales *slower* than $c_l \ln \frac{1}{1-\lambda}$ as

$\lambda \to 1$, then the resulting delay performance can be *no better* than that of an optimal online policy by more than a constant factor. As a by-product of our result, an interesting "conservation law" is established, which suggests that delay and future information are, in some sense, "exchangeable" quantities (see discussions in Section 2.1).

Despite having identical modeling assumptions, our proof techniques are quite different from those employed by Spencer et al. (2014). The core of our arguments hinges upon a relationship between diversions and *future idling* of the server, evaluated over certain subset of input sample paths. This relationship is then used in conjunction with the excursion probabilities of a transition random walk to demonstrate that the system manager *must* maintain a relatively large queue length, when the amount of future information is limited. We believe that this line of arguments is fairly robust to changes in modeling assumptions, and can be generalized, in other dynamic resource allocation problems, to proving lower bounds for the amount of information necessary in achieving desirable performance.

1.1. Organization

The remainder of the paper is organized as follows. In Section 2, we state our main result, Theorem 1, and contrast it with the prior results of Spencer et al. (2014). In the same section, we discuss several implications of the theorem (Section 2.1), as well as connections of our work to the literature (Section 2.2). Section 3 describes the modeling assumptions in more details, and introduces the necessary mathematical formalism. The proof of Theorem 1 is given in Section 4, with an outline of the proof ideas provided at the beginning of the section. We conclude the paper in Section 5 and examine potential directions for future research.

2. Main Result

Review of Prior Results. We begin by informally reviewing the system model in Spencer et al. (2014), which will be described in detail in Section 3. The admission control problem runs in continuous time, and is characterized by three parameters: λ, p, and W_λ. An illustration of the system model is given in Figure 1.

1. Jobs arrives to the system at the rate of λ, where $\lambda \in (0,1)$. There is a single server which processes jobs at the rate of $1-p$, where p is a *fixed* constant in $(0,1)$. It is assumed that the system is operating in the *overload* regime, with $\lambda > 1-p$.
2. Upon each job's arrival, the system manager decides whether the job is to be admitted or diverted. If admitted, the job queues up in an (infinite) buffer until it is processed by the server, and if diverted, it leaves the system immediately. The goal of the system manager is to choose a *diversion policy* that minimizes the time-average queue length induced by the

admitted jobs, subject to the constraint that the infinite-horizon time-average rate of diversion does not exceed p.

We will be primarily interested in the *heavy-traffic regime* of $\lambda \to 1$, where the post-diversion arrival rate approaches the server capacity of $1-p$, assuming that the system manager diverts at the maximum allowable rate of p. Note that by Little's Law, the time-average queue length is equal to the time-average queueing delay multiplied by the post-diversion arrival rate of $\lambda - p$. In the limit of $\lambda \to 1$, the two quantities will differ only by a multiplicative constant of $1-p$. Therefore, from this point on, we will focus on the time-average queue length as the performance metric, with the understanding that an analogous statement will hold for delay as well.

3. The system manager has access to information about the future, which takes the form of a *lookahead window* of length W_λ: at time t, the times of arrivals and service availability within the interval $[t, t+W_\lambda]$ are revealed to the system manager[1]. The case of $W_\lambda = 0$ will be referred to the *online* problem, since the system manager does not have access to any future information.

Denote by $\mathcal{Q}(\pi, \lambda, W_\lambda)$ the time-average queue length under the diversion policy π, given arrival rate λ and a lookahead window of length W_λ. Let $\mathcal{Q}^*(\lambda, W_\lambda)$ be the time-average queue length under an *optimal* diversion policy (assuming such optimal policies exist), with

$$\mathcal{Q}^*(\lambda, W_\lambda) = \min_{\pi} \mathcal{Q}(\pi, \lambda, W_\lambda), \tag{2}$$

It is shown in Spencer et al. (2014) that a finite amount of lookahead into the future is sufficient to yield significant delay improvement over an online policy. In particular, fixing $p \in (0,1)$, they show that the optimal average queue length for an online policy diverges to infinity in the heavy-traffic regime, with

$$\mathcal{Q}^*(\lambda, 0) \sim \log_{\frac{1}{1-p}} \frac{1}{1-\lambda}, \quad \text{as } \lambda \to 1. \tag{3}$$

In sharp contrast, there exists a positive constant c_h, whose value can depend on p, so that if

$$W_\lambda \geq c_h \ln \frac{1}{1-\lambda}, \tag{4}$$

for all λ sufficiently close to 1, then the optimal average queue length converges to a finite constant in the heavy-traffic regime:

$$\mathcal{Q}^*(\lambda, W_\lambda) \to \frac{1-p}{p}, \quad \text{as } \lambda \to 1. \tag{5}$$

[1] Depending on the application, one can think of the lookahead window as being provided by some external oracle, or a predictive model that has access to side information.

A main open question posed by Spencer et al. (2014) is whether significant performance gain over the online policy can still be achieved under much less future information. It is conjectured that if $W_\lambda = o\left(\ln \frac{1}{1-\lambda}\right)$, then the average queue length will necessarily diverge to infinity in the heavy-traffic limit (Conjecture 1, Spencer et al. (2014)). In other words, a sufficient amount of future information may be *essential* in achieving superior delay performance.

Our Result. The main result of this paper confirms, and strengthens, this conjecture of Spencer et al. (2014). We show that if the amount of future information is *insufficient* even by a constant factor, then not only will the delay be infinite in the heavy-traffic regime, but the delay scaling will essentially be *no better* than that of an online policy. Specifically, we have the following theorem.

Theorem 1 **(Necessity of Future Information)**. *Fix* $p \in (0,1)$. *There exist* $c_l > 0$ *and* $\tilde{\lambda} \in (1-p, 1)$*, so that if*

$$W_\lambda \le c_l \ln \frac{1}{1-\lambda}, \quad \forall \lambda \in (\tilde{\lambda}, 1), \tag{6}$$

then[2]

$$\mathcal{Q}^*(\lambda, W_\lambda) = \Theta\left(\ln \frac{1}{1-\lambda}\right), \quad \textit{as } \lambda \to 1. \tag{7}$$

Together with the results of Spencer et al. (2014), Theorem 1 suggests that the performance of the admission control problem *depends critically* on the amount of future information available, and in particular, on how the length of the lookahead window, W_λ, scales relative to the watershed of $\Theta\left(\ln \frac{1}{1-\lambda}\right)$. A graphical illustration of Theorem 1, with a comparison to the results of Spencer et al. (2014), is provided in Figure 2.

The proof of Theorem 1 is given in Section 4. It is worth noting that our proof techniques are quite different from those employed by Spencer et al. (2014). In fact, they are somewhat “dual” to each other: the earlier achievability result (Eq. (5)) was proved by analyzing the distribution of the lengths of busy periods associated with the queue length process (a property in *time*), whereas the core of our arguments relies on the excursion properties of a transient random walk (a property in *space*).

2.1. Implications of Theorem 1

There are several interesting implications of Theorem 1. First, by virtue of being a lower bound for the case where the decision maker is given the exact realizations of future input, Theorem 1 automatically extends to settings where predictions can be noisy or corrupted, as is typically the case in practical applications.

[2] The notation $f(x) = \Theta(g(x))$, as $x \to 1$, represents the statement that, for any sequence $x_n \to 1$, we have $0 < \liminf_{n\to\infty} f(x_n)/g(x_n) \le \limsup_{n\to\infty} f(x_n)/g(x_n) < \infty$.

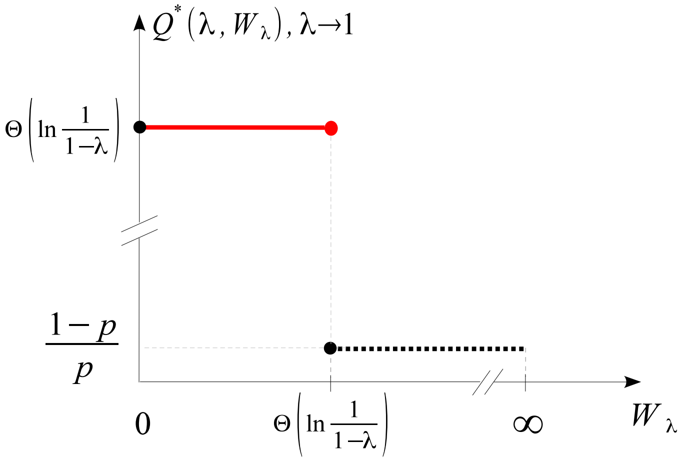

Figure 2 Impact of future information on the effectiveness of admission control, in the heavy-traffic regime of $\lambda \to 1$. The solid red segment corresponds to the regime established by this paper, where $W_\lambda \preccurlyeq c_l \ln \frac{1}{1-\lambda}$ (Theorem 1), and the dotted black segment corresponds to the regime established by Spencer et al. (2014), where $W_\lambda \succcurlyeq c_h \ln \frac{1}{1-\lambda}$ (Eq. (4) and (5) in the current paper). The case of $W_\lambda = 0$ is covered by either paper.

Theorem 1 also implies an interesting "conservation law" between delay and future information: from Eqs. (3) through (7), we see that the sum of $Q^*(\lambda, W_\lambda)$ and W_λ must be of order $\mathbf{\Omega}\left(\ln\frac{1}{1-\lambda}\right)$, as $\lambda \to 1$. In a rough sense, this is because the same type of stochastic discrepancies in the input processes, which necessitate large queueing delays in the heavy-traffic when future information is limited, also determines how much lookahead is required in order to achieve a bounded delay. Even though such conservation seems to suggest that there is no "free lunch" to be had, the ability to understand and make such trade-offs can still be useful, because depending on the application, future information may be significantly less costly than delay, or *vice versa*.

From an operational point of view, although Theorem 1 invalidates the usefulness of future information in certain regimes, it is nevertheless reassuring to know that a simple online policy could do almost as well as any sophisticated prediction-guided policies, even when the amount of predictive information available grows as the traffic intensity increases. Moreover, the theorem does not rule out the possibility of having meaningful prediction-guided policies when future information is limited; it only implies that our search in such scenarios should aim at more moderate, *constant factor* performance improvements over online policies. In fact, numerical results in Xu and Chan (2014) on a similar admission control model suggest that sizable performance gains can still be achievable, even with limited and noisy predictive information.

2.2. Related Work

In terms of modeling assumptions, our setup is identical to that of Spencer et al. (2014), and hence we refer the reader to Spencer et al. (2014) for a review of the model's connections with the literature on classical Markov admission control problems and competitive analysis. The model is also related to a multi-server system with partial resource pooling (cf. Tsitsiklis and Xu (2012)); the reader is referred to Chapter 7 of Xu (2014) for more details. In addition, Xu and Chan (2014) examines the model's relevance in the context of reducing waiting times at emergency departments.

Our result can be viewed as a generalization of the Markov optimal admission control problem that has been studied in the literature (Stidham (1985)), and it is interesting to contrast some of the differences in analytical approaches. Optimal policies in the Markov setting ($W_\lambda = 0$) are known to often admit a *threshold* (or control-limit) form, where a diversion is made only if the current queue length reaches a fixed threshold. To prove the optimality of these policies, one would typically analyze the Bellman equations of the corresponding Markov decision process (MDP) in order to establish a set of monotonicity properties in the policy space, e.g., that the cost-to-go function for a threshold policy would be dominated by policies that divert with non-zero probabilities when the queue is small (c.f. Yechiali (1971)). Successive applications of such monotonicity properties will then narrow the policy space down to only those with a threshold form.

Unfortunately, these arguments employed in the Markov setting do not seem to carry over easily when the lookahead window is taken into account. While our setting can still be cast as an MDP by incorporating the lookahead window into the state space, the structure of the state space is now considerably more complex (and increasingly so, as $W_\lambda \to \infty$), and it is not so clear as to whether any monotonicity property continues to hold. Our proof techniques circumvent this complexity by focusing on the "macroscopic" sample-path characteristics of the system, instead of the more refined details of the Bellman equations. As a trade-off, our analysis is more "coarse" by nature, and it provides neither a characterization of the multiplicative *constant* in the delay scaling, nor a concrete diversion policy that achieves the lower bound of the necessary amount of future information (which, fortunately, has already been given in Spencer et al. (2014)).

Our work is also similar in spirit to the techniques of information relaxation and path-wise optimization for MDPs (Rogers (2007), Brown et al. (2010), Desai et al. (2012)). In this case, one considers an relaxed version of the original MDP, where the decision maker has access to realizations of the future input sample paths. This relaxed problem is often simpler to solve and simulate than the original stochastic optimization problem, and hence can be used, for instance, as a performance benchmark for evaluating heuristic policies. Our work is different from this literature in several aspects. Most notably, we focus on rigorously understanding the stochastic dynamics involved in the relaxed problem with future information, and how performance scales with respect to the length of the lookahead window, as opposed to using the relaxed problem to approximate the performance of an optimal online policy, which is well understood in our setting.

3. Model and Notation

We now present the mathematical formalism and modeling assumptions that will be used throughout the remainder of the paper. An illustration of the system is given in Figure 1.

System Dynamics. The system runs in continuous time, indexed by $t \in \mathbb{R}_+$. There is a *queue* with infinite waiting room, whose length at time t is denoted by $Q(t)$. The input to the system consists of two independent Poisson processes:

1. $\mathcal{A}$, with rate λ, which corresponds to the *arrival* of jobs;
2. $\mathcal{S}$, with rate $1-p$, which corresponds to the generation of *service tokens*.

When an *event* occurs in $\mathcal{A}$ at time t, we say that a job has arrived to the system, and the value of $Q(t)$ is incremented by 1, if the job is "admitted" (see below for the description of admission policies). Similarly, when an event occurs in the process $\mathcal{S}$ at time t, we say that a service token is generated, and the value of $Q(t)$ is decremented by 1, if $Q(t) > 0$, and remains at 0, otherwise.[3]

For our purposes, it is more convenient to work with the sequence $\{(Z_n, R_n) : n \in \mathbb{N}\}$, where

$$Z_n = \text{time of the } n\text{th event in } \mathcal{A} \cup \mathcal{S}, \tag{8}$$

and R_n encodes the type of the nth event, with

$$R_n = \begin{cases} 1, & \text{if the } n\text{th event is in } \mathcal{A} \text{ (arrival)}, \\ -1, & \text{if the } n\text{th event is in } \mathcal{S} \text{ (service token)}. \end{cases} \tag{9}$$

We will let $\{\mathcal{N}(t) : t \in \mathbb{R}_+\}$ be the counting process associated with $\{Z_n\}$, with

$$\mathcal{N}(t) = \sup\{n \in \mathbb{Z}_+ : Z_n \le t\}, \tag{10}$$

and denote by $S(s,t)$ the *difference* between the numbers of arrival and services tokens in the interval $(s,t]$,

$$S(s,t) = \sum_{\mathcal{N}(s)+1 \le n \le \mathcal{N}(t)} R_n. \tag{11}$$

Note that when $\lambda \neq 1-p$ the process $\{S(0,t) : t \in \mathbb{R}_+\}$ is a transient random walk, with

$$\mathbb{E}(S(0,t)) = [\lambda - (1-p)]t. \tag{12}$$

Future Information. The notion of future information is captured by a *lookahead window*. At any time t, the system manager has access to the realization of all events in $\mathcal{A} \cup \mathcal{S}$ in the interval

[3] The generation of a service token at time t can be thought of as the server being able to fetch a new job from the queue at time t. As such, the service token model attributes the randomness in processing times to an external source, which does not depend on the identities of the jobs. It can be shown that, in the online setting, the service token model is equivalent to the more conventional assumption of exponentially distributed job sizes, though such equivalence is generally not true when future information is taken into account. The reader is referred to Page 9 of Spencer et al. (2014), and the references therein, for more details on the service token model.

$[t, t+W_\lambda]$. Throughout, we will denote by W_λ the length of the lookahead window, under arrival rate λ.

Admission Policies. Upon arrival, each job is either *admitted*, in which case it joins the queue, or *diverted*, in which case it disappears from the system immediately. The role of a diversion policy, π, is to output a sequence of diversion decisions for all events, represented by the sequence of indicator variables, $\{H(n): n \in \mathbb{N}\}$, where

$$H(n) = \mathbb{I}\{R_n = 1, \text{ and } \pi \text{ chooses to divert at time } Z_n\}. \tag{13}$$

Given the form of future information, we will require that the diversion policy be $(t+W_\lambda)$-causal, so that the decision made at time t does not depend on any event after time $t+W_\lambda$. A diversion policy is said to be *feasible*, if the resulting time-average rate of diversion is at most p, i.e.,

$$\limsup_{N\to\infty} \frac{\lambda+1-p}{N}\mathbb{E}\left(\sum_{n=1}^{N} H(n)\right) \le p. \tag{14}$$

where the constant $\lambda+1-p$ corresponds to the total rate of events in $\mathcal{A}\cup\mathcal{S}$. The objective of the decision maker is to choose a feasible policy, π, so as to *minimize* the time-average queue length, defined by[4]

$$\mathcal{Q}(\pi,\lambda,W_\lambda) = \limsup_{N\to\infty}\mathbb{E}\left(\frac{1}{N}\sum_{n=1}^{N} Q(Z_n-)\right). \tag{15}$$

3.1. Notation

We will assume that all asymptotic expressions with respect to λ are taken in the limit of $\lambda \to 1$. We will use $f \ll g$ and $f \preccurlyeq g$ to denote $f = o(g)$ and $f = \mathcal{O}(g)$, respectively. We will write $f \preceq g$ to mean that $f(x) \le g(x)$ for all x sufficiently closely to 1, i.e., that there exists $y \in (0,1)$, such that $f(x) \le g(x)$, for all $x \in (y,1)$. The expressions $f \gg$, $\succcurlyeq$ and $\succeq g$ are defined analogously to their respective counterparts. When a statement is made concerning the limit "as $x \to 1$", without specifying the exact sequence with respect to which the limit is taken, it is understood that the statement should hold for any sequence, $\{x_n\}$, with $\lim_{n\to\infty} x_n = 1$. The notation $X \overset{d}{=} Y$ means that the random variables X and Y have the same distribution.

4. Proof of Theorem 1

The remainder of the paper is devoted to the proof of Theorem 1. We begin with a high-level summary of the main steps involved. First, we argue that there exists a stationary optimal policy, which makes decisions only based on the current queue length and the content of the lookahead window. Furthermore, the queue length process under this stationary policy admits a well-defined

[4] Throughout, $f(x-)$ represents the limit $\lim_{y\uparrow x} f(y)$.

steady-state distribution (Section 4.1.1). This stationarity will allow us to simplify the analysis by focusing on the policy's actions over a finite time horizon.

We will prove Theorem 1 by contradiction, where we start by assuming that a small average queue length is indeed achievable under an optimal stationary policy, even with a small lookahead window, and later refute this assumption. Our main arguments are based on the identification of a set of *base sample paths* (Section 4.2), with the property that *any* feasible policy must perform poorly over these sample paths, should the length of the lookahead window be too small. The stationarity property described earlier will then allow us to extend this argument to showing the policy's failure over the infinite time horizon. It is worth noting that the base sample paths are not "typical," in the sense that their occurrences possess only vanishingly small probability, as $\lambda \to 1$. This is because the failures of a policy under a small lookahead window are not caused by the average behavior of the inputs, but rather by some rare excursions of the random walk $S(0,\cdot)$. Though occurring with small probabilities, these excursions are in some sense unforeseeable under a small lookahead window, and their existence forces an optimal policy to be overly restrained in diverting jobs and hence yield a large average queue length.

To carry out the arguments using the base sample paths, we will exploit a key relationship between *diversions* and *server idling*. In particular, we will demonstrate that, without sufficient lookahead, if a constant fraction of the arrivals are diverted during a specific portion of a base sample path, it will inevitably result in excessive idling of the server not far away in the future, even as $\lambda \to 1$. However, such server idling cannot occur in the heavy-traffic limit, since the server must be fully utilized in order to ensure system stability. This reasoning then implies that any policy that makes such diversions must be infeasible, or conversely, that any feasible policy must divert very few arrivals over these segments of the base sample paths (Proposition 1). However, such conservatism comes at a cost, in that it leads to long episodes during which the queue length stays at a high level (Proposition 2). We then argue that the frequent appearances of such "bad" episodes will result in a large average queue length in steady-state, which contradicts with our initial assumption and hence completes the proof of Theorem 1.

4.1. Preliminaries

Without loss of generality, we will consider only the cases where the length of the lookahead window, W_λ, diverges to infinity in the heavy-traffic regime, i.e.,

$$W_\lambda \to \infty, \quad \text{as } \lambda \to 1. \tag{16}$$

To see why this is justified, note that because we can always achieve the same average queue length with a longer lookahead window, the optimal average queue length $\mathcal{Q}^*(\lambda, W_\lambda)$ must be monototicaly non-increasing in W_λ. Therefore, any lower bound we obtain on $\mathcal{Q}^*(\lambda, W_\lambda)$ under the assumption of Eq. (16) also applies to the case where $W_\lambda = \mathcal{O}(1)$. For simplicity of notation, we will drop the dependency on W_λ, and denote by q_λ the optimal average queue length,

$$q_\lambda = \mathcal{Q}^*(\lambda, W_\lambda), \quad \forall \lambda \in (1-p, 1). \tag{17}$$

Main Assumption. We will assume the validity of the following hypothesis throughout the remainder of the proof, which states that it is indeed possible to achieve a small delay as long as W_λ is of order $\mathbf{\Omega}\left(\ln \frac{1}{1-\lambda}\right)$. As will be shown in Section 4.5, invalidating this hypothesis will imply the lower bound in Theorem 1.

Hypothesis 1. *Fix $p \in (0,1)$. Suppose that $W_\lambda \succcurlyeq \ln \frac{1}{1-\lambda}$, as $\lambda \to 1$. Then*

$$q_\lambda \ll \ln \frac{1}{1-\lambda}, \quad \textit{as } n \to \infty. \tag{18}$$

Assuming the validity of Hypothesis 1, it also follows that if $W_\lambda \succcurlyeq \ln \frac{1}{1-\lambda}$, as $\lambda \to 1$, then

$$q_\lambda \ll W_\lambda, \quad \text{as } \lambda \to 1. \tag{19}$$

4.1.1. State Representation and Stationary Policies We show in this section that there always exists an *stationary* optimal policy that depends only on the state, which consists of the current queue length and content of the lookahead window.

Since all diversion decisions are associated with events in $\mathcal{A} \cup \mathcal{S}$, it suffices to specify the nature of future information for the event times, $\{Z_n : n \in \mathbb{N}\}$. At $t = Z_n$, the *content* of the lookahead window is defined to be the vector $F(n) = (F_k(n) : k \in \mathbb{Z}_+)$, where

$$F_k(n) = (Z_{n+k} - Z_n \,,\, R_{n+k}), \quad 0 \le k \le \mathcal{N}(Z_n + W_\lambda) - \mathcal{N}(Z_n). \tag{20}$$

In other words, $F_k(n)$ specifies the time of the kth future event starting from the current time, Z_n, along with its type for all events within the lookahead window of length W_λ. For future events beyond the lookahead window which we have no access to, we simply set the value of $F_k(n)$ to zero:

$$F_k(n) = (0\,,\,0), \quad k > \mathcal{N}(Z_n + W_\lambda) - \mathcal{N}(Z_n). \tag{21}$$

Recall that $Q(t)$ is the queue length at time t. Consider the sequence $\{X(n) : n \in \mathbb{N}\}$, where

$$X(n) = \left(Q(Z_n-), F(n)\right). \tag{22}$$

From this point on, we will refer to $\{X(n) : n \in \mathbb{N}\}$ as the *states* of our system.

Stationary Policies. A diversion policy π is *stationary*, if its diversion decision at time Z_n depends only on the state, $X(n)$, or formally, that

$$\mathbb{P}\left(H(n)=1 \,\middle|\, X(n)\right)=\mathbb{P}\left(H(n)=1 \,\middle|\, \{(Z_k,R_k)\}_{k=1}^{\mathcal{N}(Z_n+W_\lambda)}\right), \quad \text{a.s.} \tag{23}$$

A stationary policy, π, is *stable*, if the evolution of $\{X(n): n\in\mathbb{N}\}$ under π admits a well-defined steady-state distribution, γ, so that steady-state queue length and probability of diversion coincide with the time-average queue length and diversion rate, respectively, given that the initial condition, $X(1)$, is distributed according to γ.

In our admission control problem, because the arrivals and service tokens are generated according to Poisson processes, future evolution of the system starting from $t=Z_n$ is independent conditional on the current state X_n and diversion decision. As such, our problem can be cast as a discrete-time Markov decision process (MDP), with states $\{X_n : n\in\mathbb{N}\}$ and actions that correspond to the probabilities of diversion. Using existing results in the literature (c.f. Hernández-Lerma et al. (2003), Gonzlez-Hernández and Villarreal (2011)), it can be shown that, for MDPs of this kind, there exists an optimal policy that is also stationary and stable. This is summarized in the following lemma, whose proof is given in Appendix A.1.

LEMMA 1. *Fix any $p>0$, $\lambda\in(1-p,1)$, and $W_\lambda>0$. The admission control problem admits a* stable stationary *optimal policy, π, which achieves the minimum time-average queue length among all feasible diversion policies.*

In light of Lemma 1, we will, in the remainder of the proof of Theorem 1, focus on the family of stable stationary policies, which we will refer to simply as *stationary policies.* Given a stationary policy, π, the resultant state sequence $\{X(n): n\in\mathbb{N}\}$ is a stationary Markov chain. Since we are interested in deriving a performance lower bound, we may assume that, at time $t=0$, both the queue length and the content of the lookahead window are initialized according to the steady-state distributions, γ. In particular, we have that

$$\mathbb{E}\left(Q(t)\right)=\mathbb{E}\left(Q(0)\right)=\mathcal{Q}(\pi,\lambda,W_\lambda), \quad t\in\mathbb{R}_+. \tag{24}$$

and, that

$$\mathbb{E}(H(n))=\mathbb{E}(H(1))=\limsup_{N\to\infty}\frac{\mathbb{E}\left(\sum_{n=1}^{N}H(n)\right)}{N}, \quad \forall n\in\mathbb{N}. \tag{25}$$

Define the process $\{L(t): t\in\mathbb{R}_+\}$, where

$$L(t)=\mathbb{I}\left\{Q(t)\le 2q_\lambda\right\}, \quad t\in\mathbb{R}_+. \tag{26}$$

The following lemma will be useful.

Lemma 2. *Fix $p \in (0,1)$. For all $\lambda \in (1-p,1)$, we have that*

$$\mathbb{E}(L(t)) = \mathbb{P}\left(Q(0) \le 2q_\lambda\right) \ge \frac{1}{2}, \quad \forall t \in \mathbb{R}_+, \tag{27}$$

under any optimal stationary policy.

Proof. The result follows from the stationarity of $Q(\cdot)$ and the Markov's inequality:

$$\mathbb{E}(L(t)) = \mathbb{P}\left(Q(t) \le 2q_\lambda\right) = \mathbb{P}\left(Q(0) \le 2q_\lambda\right) = \mathbb{P}\left(Q(0) \le 2\mathbb{E}(Q(0))\right) \ge \frac{1}{2}. \tag{28}$$

Q.E.D.

In the remainder of the proof, we will show that there exists $c_l > 0$ such that if $W_\lambda \preceq c_l \ln \frac{1}{1-\lambda}$, then Eq. (27) cannot be true under any sequence of optimal stationary policies, unless $Q^*(\lambda, W_\lambda) \succcurlyeq \ln \frac{1}{1-\lambda}$. This would invalidate Hypothesis 1, which would in turn prove the lower bound on $\mathcal{Q}^*(\lambda, W_\lambda)$ in Theorem 1.

4.2. Base Sample Paths

We now describe the construction of a set of base sample paths which will serve as the basis of our subsequent analysis. In later sections, we will show that, roughly speaking, the non-negligible chance of occurrence of such sample paths will "force" any feasible policy to be overly conservative in diverting jobs, should W_λ be too small.

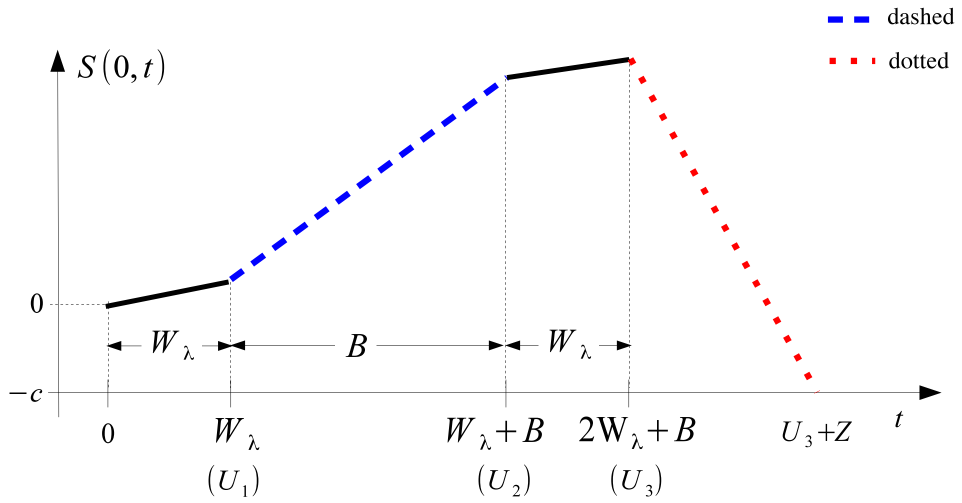

Figure 3 This figure illustrates the "macroscopic" behavior of the base sample paths. The dashed blue segment between W_λ and $W_\lambda + B$ represents a period of sustained upward drift of $S(0,\cdot)$, and the dotted red segment starting at $2W_\lambda + B$ represents a downward drift. The two solid black segments, each with length equal to that of the lookahead window, serve as a "buffer", ensuring that the actions of the diversion policy before the segment are independent from the evolution of $S(0,\cdot)$ afterwards.

Let $B \in \mathbb{R}_+$ be a quantity whose value will be specified in the sequel. We define the following time markers, whose positions relative to each other are illustrated in Figure 3.

$$\begin{aligned} U_1 &= W_\lambda, \\ U_2 &= U_1 + B = W_\lambda + B, \\ U_3 &= U_2 + W_\lambda = 2W_\lambda + B. \end{aligned}$$

The set of base sample paths is defined as the intersection of the events $\mathcal{E}_1$ through $\mathcal{E}_5$, described as follows. Let ϵ, ζ and ϕ be positive constants.

1. Event $\mathcal{E}_1$, parameterized by ϵ and ζ, says that the sample path of $S(0,\cdot)$ stays close to its expected behavior during the interval $(U_1, U_2]$:

$$\mathcal{E}_1 = \{ |S(U_1, t) - [\lambda - (1-p)]t| \leq \epsilon t + \zeta, \text{ for all } t \in (U_1, U_2] \}, \tag{29}$$

   When ϵ is small, this implies that $S(0,\cdot)$ undergoes a consistent *upward* drift during $(U_1, U_2]$. Event $\mathcal{E}_1$ is illustrated by the dashed blue line segment in Figure 3.

2. Event $\mathcal{E}_2$ says that the queue length at $t = 0$ is not too large compared to the optimal average queue length,

$$\mathcal{E}_2 = \{ Q(0) \leq 6q_\lambda \}. \tag{30}$$

3. The events $\mathcal{E}_3$ and $\mathcal{E}_4$ put some restriction on the amount of upward excursion of $S(0,\cdot)$ during the intervals $(0, U_1]$ and $(U_2, U_3]$, respectively,

$$\mathcal{E}_3 = \{ S(0, U_1) \leq 2W_\lambda \}, \tag{31}$$

$$\mathcal{E}_4 = \{ S(U_2, U_3) \leq 2W_\lambda \}, \tag{32}$$

   The main purpose of $\mathcal{E}_3$ and $\mathcal{E}_4$ is to serve as "buffers" to induce certain independence property, which will be useful for subsequent analysis: since the lengths of $(0, U_1]$ and $(U_2, U_3]$ are both equal to that of the lookahead window, the actions of the diversion policy before each interval are *independent* from the evolution of $S(0,\cdot)$ after it. The two events are illustrated by the black line segments in Figure 3.

4. Finally, the event $\mathcal{E}_5$ says that $S(0,\cdot)$ will undergo a substantial *downward* excursion soon after U_3, as is illustrated by the dotted red line segment in Figure 3. Let Z be the stopping time

$$Z = \inf \{ z \in \mathbb{R}_+ : S(U_3, U_3 + z) < -[6q_\lambda + [\lambda - (1-p) - \epsilon]B + \zeta + 4W_\lambda] \}, \tag{33}$$

   and $\mathcal{E}_5$ is defined by putting an upper bound on Z:

$$\mathcal{E}_5 = \{ Z \leq \phi W_\lambda \}. \tag{34}$$

> The right-hand-side of the inequality in the definition of Z was chosen so that, conditional on the joint occurrence of $\mathcal{E}_1$ through $\mathcal{E}_4$, a downward excursion in $S(0,\cdot)$ of such magnitude is guaranteed to *deplete* the queue by time U_3+Z. As will become clearer in the next section, this depletion will help us connect diversions to future idling of the server.

Note that the events $\mathcal{E}_1$, $\mathcal{E}_3$, $\mathcal{E}_4$ and $\mathcal{E}_5$ concern the input sample path $S(0,\cdot)$ only, and are independent of the diversion policies, while $\mathcal{E}_2$ also depends on the choice of diversion policy.

Having described the events that together characterize the base sample paths, we next illustrate some of their statistical properties. The first lemma shows that the events $\mathcal{E}_1$ through $\mathcal{E}_4$ can occur with fairly high probabilities. The proof is given in Appendix A.2.

LEMMA 3. 1. *Fix $\epsilon > 0$. For all $\theta \in (0,1)$, there exists $\zeta > 0$, so that for all $\lambda > 1 - \frac{1}{2}p$,*

$$\inf_{B\geq 0} \mathbb{P}\left(\mathcal{E}_1\right) \geq \theta. \tag{35}$$

2. *Under optimal stationary policies, $\mathbb{P}\left(\mathcal{E}_2\right) = \mathbb{P}(Q(0) \leq 6q_\lambda) \geq \frac{5}{6}$, for all $\lambda \in (1-p,1)$.*
3. $\lim_{\lambda\to 1} \mathbb{P}\left(\mathcal{E}_3\right) = \lim_{\lambda\to 1} \mathbb{P}\left(\mathcal{E}_4\right) = 1.$

The next lemma shows that the event $\mathcal{E}_5$ occurs with a small yet non-negligible probability. The proof is given in Appendix A.3.

LEMMA 4. *Fix $k, \phi, \zeta > 0$, and $\epsilon \in (0, \min\{\zeta, \lambda - (1-p)\})$. Suppose that $B = kW_\lambda$, and $q_\lambda \ll W_\lambda$, as $\lambda \to 1$. There exists $\gamma > 0$, such that*

$$\mathbb{P}\left(\mathcal{E}_5\right) \succcurlyeq \exp\left(-\gamma W_\lambda\right), \quad \textit{as } \lambda \to 1. \tag{36}$$

Finally, the following independence properties among the events will be useful. The proof is given in Appendix A.4.

LEMMA 5. *Fixing a feasible diversion policy, the following holds.*

1. *The events $\mathcal{E}_1, \mathcal{E}_3, \mathcal{E}_4$ and $\mathcal{E}_5$ are mutually independent.*
2. *The event $\mathcal{E}_2$ is independent of $\mathcal{E}_1, \mathcal{E}_4$ and $\mathcal{E}_5$, but not necessarily $\mathcal{E}_3$.*
3. *Denote by Y the number of diversions in the interval $(U_1, U_2]$, i.e.,*

$$Y = \sum_{\mathcal{N}(U_1)+1 \leq n \leq \mathcal{N}(U_2)} H(n). \tag{37}$$

Then Y is independent of $\mathcal{E}_5$.

4.3. From Diversions to Server Idling

The goal of this subsection is to show that, if W_λ is small, then the number of diversions made during the the interval $(U_1, U_2]$, i.e., the random variable Y (Eq. (37)), must also be appropriately small, under any optimal stationary policy. To achieve this, we will exploit a connection between Y and the idling of the server at a later time.

The intuition is perhaps best seen pictorially, as depicted in Figure 3. Conditional on the occurrence of the events $\mathcal{E}_1$ through $\mathcal{E}_5$, and suppose no diversion has been made, the queue length process $Q(t)$ would have "followed" the trajectory depicted in the figure and reached zero by time $U_3 + \phi W_\lambda$. Suppose now that a large number of diversions are made during the interval $(U_1, U_2]$ (dashed line segment in blue), the depletion of the queue implies that there must be an extended period of server idling prior to $U_3 + \phi W_\lambda$. Such idling, if it persists even as $\lambda \to 1$, can be problematic and will be shown to contradict the feasibility of the diversion policy. This in turn implies that the number of diversions in $(U_1, U_2]$ must be small.

The next proposition is the main result of this subsection, which formalizes the above intuition. There is, however, one adjustment: as opposed to conditioning on all five events, which has vanishingly small probability due to the presence of $\mathcal{E}_5$, we will condition only on $\mathcal{E}_1$ and $\mathcal{E}_2$, which occur with high probability. To do so, we will exploit several independence properties among the events, as in Lemma 5, and show that the impact of $S(0,\cdot)$'s downward excursion described by $\mathcal{E}_5$ is unavoidable when W_λ is too small, even without explicitly conditioning on $\mathcal{E}_5$.

Proposition 1. *Fix $k > 0$, and let $B = kW_\lambda$. There exists $c > 0$, so that if*

$$W_\lambda \preceq c \ln \frac{1}{1-\lambda}, \quad \text{as } \lambda \to 1, \tag{38}$$

then for every $\tau > 0$,

$$\lim_{\lambda \to 1} \mathbb{P}\left(Y \geq \tau B \,\middle|\, \mathcal{E}_1 \cap \mathcal{E}_2\right) = 0, \tag{39}$$

under any sequence of optimal stationary policies, where Y is the number of diversions during $(U_1, U_2]$, defined in Eq. (37)*.*

Proof. We say that a service token generated at time t is *wasted*, if there is currently no job in the queue, i.e., $Q(t) = 0$. Let $\{\mathcal{J}(t) : t \in \mathbb{R}_+\}$ be the counting process of wasted service tokens, where

$$\mathcal{J}(t) = \#\text{ of wasted service tokens in } [0, t]. \tag{40}$$

For the sake of contradiction, assume the following is true: if $W_\lambda \succcurlyeq \ln \frac{1}{1-\lambda}$ as $\lambda \to 1$, then there exist $\tau > 0$, and a sequence of optimal stationary policies, $\{\pi_\lambda\}$, under which

$$\liminf_{\lambda \to 1} \mathbb{P}\left(Y \geq \tau B \,\middle|\, \mathcal{E}_1 \cap \mathcal{E}_2\right) = q > 0. \tag{41}$$

The following lemma is a key ingredient to the proof, which says that the number of wasted tokens must be substantial. The proof is based on the intuition explained in the passages above Proposition 1, and is given in Appendix A.5.

Lemma 6. *Fix $k>0$, and let $B=kW_\lambda$. Suppose Eq.* (41) *is true for some sequence of optimal stationary policies, $\{\pi_\lambda\}$. Then there exist $a,\gamma>0$ (whose values can depend on k) such that*

$$\mathbb{E}\left(\mathcal{J}(aW_\lambda)\right) \succsim W_\lambda \exp\left(-\gamma W_\lambda\right), \tag{42}$$

as $\lambda \to 1$, under $\{\pi_\lambda\}$.

Consider an optimal stationary policy. Denote by $\mathcal{H}(t)$ the counting process representing the number of diversions in $[0,t]$, i.e.,

$$\mathcal{H}(t) = \sum_{n=1}^{\mathcal{N}(t)} H(n). \tag{43}$$

By the stationarity of $\{H(n): n\in\mathbb{N}\}$ (Eq. (25)) and definition of $\mathcal{N}(t)$ (Eq. (10)), it is not difficult to show that, for all $t>0$,

$$\begin{aligned}\frac{\mathbb{E}(\mathcal{H}(t))}{t} =& \frac{1}{t}\mathbb{E}\left(\sum_{n=1}^{\mathcal{N}(t)} H(n)\right) = (\lambda+1-p)\mathbb{E}(H(1)) \\ =& \limsup_{N\to\infty} \frac{(\lambda+1-p)\mathbb{E}\left(\sum_{n=1}^{N} H(n)\right)}{N}.\end{aligned} \tag{44}$$

By definition, we have that

$$Q(t) = Q(0) + S(0,t) + \mathcal{J}(t) - \mathcal{H}(t), \quad \forall t>0. \tag{45}$$

Taking expectation on both sides of the above equation, and letting $t=aW_\lambda$, where a is given as in Lemma 6, we have that

$$\begin{aligned}&\frac{\mathbb{E}(\mathcal{H}(aW_\lambda))}{aW_\lambda} - p \\ =& \frac{1}{aW_\lambda}\left(\mathbb{E}\left(S\left(0,aW_\lambda\right)\right) + \mathbb{E}\left(\mathcal{J}(aW_\lambda)\right) + \mathbb{E}(Q(0)) - \mathbb{E}(Q(aW_\lambda))\right) - p \\ \overset{(a)}{=}& [\lambda-(1-p)] - p + \frac{1}{aW_\lambda}\mathbb{E}\left(\mathcal{J}(aW_\lambda)\right) \\ \overset{(b)}{\succsim}& (\lambda-1) + \frac{1}{aW_\lambda}W_\lambda \exp\left(-\gamma W_\lambda\right) \\ \succsim& \exp\left(-\gamma W_\lambda\right) - (1-\lambda),\end{aligned} \tag{46}$$

where γ is given in Lemma 6. Step (a) follows from the fact that $\mathbb{E}(Q(0)) = \mathbb{E}(Q(aW_\lambda))$ by the stationarity of $Q(\cdot)$, and (b) from Eq. (42).

Letting $W_\lambda = c \ln \frac{1}{1-\lambda}$, with $c = 1/2\gamma$, we have that

$$\exp\left(-\gamma W_\lambda\right) \succcurlyeq \sqrt{1-\lambda}, \quad \text{as } \lambda \to 1. \tag{47}$$

Combining Eqs. (46) and (47), we have that

$$\frac{\mathbb{E}(\mathcal{H}(aW_\lambda))}{aW_\lambda} - p \succcurlyeq \sqrt{1-\lambda} - (1-\lambda) \succcurlyeq \sqrt{1-\lambda}, \quad \text{as } \lambda \to 1. \tag{48}$$

In particular, this implies that there exists $\lambda' \in (1-p, 1)$, such that

$$\frac{\mathbb{E}(\mathcal{H}(aW_\lambda))}{aW_\lambda} > p, \quad \forall \lambda \in (\lambda', 1). \tag{49}$$

Since the stationary diversion policies we consider are feasible, we must have that

$$\frac{\mathbb{E}(\mathcal{H}(t))}{t} \overset{(a)}{=} \limsup_{N\to\infty} \frac{(\lambda+1-p)\mathbb{E}\left(\sum_{n=1}^{N} H(n)\right)}{N} \overset{(b)}{\leq} p, \tag{50}$$

for all $\lambda \in (1-p, 1)$, and $t > 0$, where (a) and (b) follow from Eqs. (44) and (14), respectively. This leads to a contradiction with Eq. (49), which invalidates the assumption made in Eq. (41), and hence proves Proposition 1. *Q.E.D.*

4.4. Consequences of Too Few Diversions

Proposition 1 tells us that, under optimal stationary policies, the number of diversions in $(U_1, U_2]$ must be small when W_λ is small. Building on this observation, we now focus on policies that divert "very few" jobs during $(U_1, U_2]$, i.e., with Y scaling sub-linearly with respect to B, and show that they will necessarily lead to a large expected queue length in steady-state. The following proposition is the main result of this subsection.

Proposition 2. *Fix $p \in (0,1)$. There exists $c_l > 0$, so that if*

$$W_\lambda \preceq c_l \ln \frac{1}{1-\lambda}, \quad \textit{as } \lambda \to 1, \tag{51}$$

then

$$\limsup_{\lambda\to 1} \mathbb{E}\left(L(0)\right) \leq \frac{1}{3}. \tag{52}$$

under any sequence of optimal stationary policies.

Proof. We will assume that $B = kW_\lambda$, with $k = 24$, and that $W_\lambda \preceq c_l \ln \frac{1}{1-\lambda}$, where c_l is equal to the constant c in Proposition 1 for the corresponding value of k.

Consider an optimal stationary policy, with a resultant average queue length of q_λ. We will prove the claim by showing that if $\mathcal{E}_1 \cap \mathcal{E}_2$ occurs *and* the number of diversions made in $(U_1, U_2]$ is small (cf. Eq. (39)), then, for a "long time" after U_1, the queue length will stay at a high level (i.e.,

$Q(t) > 2q_\lambda$). Recall that Y is the number of diversions made during the period $(U_1, U_2]$. We have the following inequality, derived from the queueing dynamics:

$$Q(t) \geq Q(U_1) + S(U_1, t) - Y, \quad \forall t \in (U_1, U_2]. \tag{53}$$

By the definition of $\mathcal{E}_1$ (Eq. (29)), Eq. (53), and the fact that $Q(U_1) \geq 0$, we have that

$$\mathbb{P}\left(Q(t) \geq [\lambda - (1-p) - \epsilon]t - \zeta - Y \,\middle|\, \mathcal{E}_1 \cap \mathcal{E}_2\right) = 1, \quad \forall t \in (U_1, U_2]. \tag{54}$$

Let V be the *last* time in $(U_1, U_2]$ when the queue length becomes less than $2q_\lambda$, with

$$V = \sup\left\{t \in [0, B) : Q(U_1 + t) \leq 2q_\lambda\right\}, \quad \text{if } \inf_{t\in[0,B)} Q(U_1 + t) \leq 2q_\lambda, \tag{55}$$

and $V = 0$, otherwise. Applying the definition of V in the context of Eq. (54) yields that

$$\mathbb{P}\left(V \leq \frac{1}{\lambda - (1-p) - \epsilon}\left(2q_\lambda + Y + \zeta + 1\right) \,\middle|\, \mathcal{E}_1 \cap \mathcal{E}_2\right) = 1. \tag{56}$$

Recall from Proposition 1 that, conditional on $\mathcal{E}_1 \cap \mathcal{E}_2$ and assuming $W_\lambda \preceq c_l \ln \frac{1}{1-\lambda}$, Y must be sub-linear in $B = kW_\lambda$. In particular, by Eq. (39), we have that, for all $\tau > 0$,

$$\lim_{\lambda\to 1} \mathbb{P}\left(Y \leq \tau k W_\lambda \,\middle|\, \mathcal{E}_1 \cap \mathcal{E}_2\right) = 1. \tag{57}$$

Combining Eqs. (56) and (57), and the fact that $W_\lambda \to \infty$ as $\lambda \to 1$, we have that, there exists $\upsilon > 0$, such that for all $\tau > 0$,

$$\mathbb{P}\left(V \leq \upsilon q_\lambda + \tau k W_\lambda \,\middle|\, \mathcal{E}_1 \cap \mathcal{E}_2\right) = 1 - \delta(\lambda), \quad \forall \lambda \in (1-p, 1), \tag{58}$$

where $\delta(\cdot)$ is a function with $\lim_{x\to 1}\delta(x) = 0$. In other words, conditional on $\mathcal{E}_1 \cap \mathcal{E}_2$, $Q(t)$ will reach the level of $2q_\lambda$ soon after U_1, with high probability. Using the fact that $V \leq U_2$, Eq. (58) further implies that

$$\mathbb{E}\left(V \,\middle|\, \mathcal{E}_1 \cap \mathcal{E}_2\right) \leq (\upsilon q_\lambda + \tau k W_\lambda)(1 - \delta(\lambda)) + U_2\delta(\lambda) \leq \upsilon q_\lambda + \tau k W_\lambda + U_2\delta(\lambda) \tag{59}$$

Translating this into the value of $\mathbb{E}(V)$, we have that

$$\begin{aligned}
\limsup_{\lambda\to 1} \frac{\mathbb{E}(V)}{U_2} &\leq \limsup_{\lambda\to 1} \frac{1}{U_2}\left(\mathbb{E}(V \mid \mathcal{E}_1 \cap \mathcal{E}_2)\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) + U_2(1 - \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2))\right) \\
&\overset{(a)}{\leq} \limsup_{\lambda\to 1}\left[1 - \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) + \frac{\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2)}{U_2}\left(\upsilon q_\lambda + \tau k W_\lambda + U_2\delta(\lambda)\right)\right] \\
&\overset{(b)}{=} \limsup_{\lambda\to 1}\left(1 - \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) + \frac{kW_\lambda}{U_2}\tau\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2)\right) \\
&\overset{(c)}{=} \limsup_{\lambda\to 1}\left(1 - \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) + \frac{k}{k+1}\tau\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2)\right) \\
&\leq \tau + \limsup_{\lambda\to 1}\left(1 - \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2)\right),
\end{aligned} \tag{60}$$

where step (a) follows from Eq. (59), (b) from the assumptions that $q_\lambda \ll W_\lambda$ and $\lim_{\lambda\to 1}\delta(\lambda)=0$, and (c) from the fact that $U_2 = B + W_\lambda = (k+1)W_\lambda$. We now connect the behavior of $\mathbb{E}(V)$ to that of $\mathbb{E}(L(0)) = \mathbb{P}(Q(0)\le 2q_\lambda)$, as follows. Fixing any $\lambda\in(1-p,1)$, we have that

$$\begin{aligned}\mathbb{E}\left(L(0)\right) &\overset{(a)}{=} \mathbb{E}\left(\frac{1}{U_2}\int_{t=0}^{U_2} L(t)dt\right)\\ &\overset{(b)}{=} \mathbb{E}\left(\frac{1}{U_2}\int_{t=0}^{U1+V} L(t)dt\right)\\ &\overset{(c)}{\le} \mathbb{E}\left(\frac{U_1+V}{U_2}\right)\\ &= \frac{U_1+\mathbb{E}(V)}{U_2}.\end{aligned} \tag{61}$$

where step (a) follows from the stationarity of the process $L(\cdot)$, which in turn follows from the stationarity of $Q(\cdot)$. Step (b) follows from the fact that $L(t)=0$, for all $t\in[U_1+V, U_2]$, which is a consequence of the definition of V in Eq. (55). Step (c) is based on the fact that $L(t)\le 1$, a.s. By Eq. (61), we have that

$$\begin{aligned}\limsup_{\lambda\to 1}\mathbb{E}\left(L(0)\right) &\le \limsup_{\lambda\to 1}\frac{U_1+\mathbb{E}(V)}{U_2}\\ &\overset{(a)}{=} \frac{W_\lambda}{(k+1)W_\lambda} + \limsup_{\lambda\to 1}\frac{\mathbb{E}(V)}{U_2}\\ &\overset{(b)}{\le} \frac{1}{k} + \tau + \limsup_{\lambda\to 1}(1-\mathbb{P}(\mathcal{E}_1\cap\mathcal{E}_2)),\end{aligned} \tag{62}$$

where steps (a) and (b) follow from the fact that $B = kW_\lambda$, and Eq. (60), respectively.

By Claim 3 of Lemma 3, and Claim 1 of Lemma 5, we have that

$$\liminf_{\lambda\to 1}\mathbb{P}\left(\mathcal{E}_1\cap\mathcal{E}_2\right) = \liminf_{\lambda\to 1}\mathbb{P}\left(\mathcal{E}_1\right)\mathbb{P}\left(\mathcal{E}_2\right) \ge \frac{5}{6}\theta, \tag{63}$$

where θ is given in Eq. (35). Set $\tau = k = 24$, and let ζ be sufficiently large so that $\theta \ge 10/9$. We have that

$$\limsup_{\lambda\to 1}(1-\mathbb{P}\left(\mathcal{E}_1\cap\mathcal{E}_2\right)) \le 1 - \frac{5}{6}\cdot\frac{9}{10} = 1/4. \tag{64}$$

From Eq. (61), we have that

$$\limsup_{\lambda\to 1}\mathbb{E}\left(L(0)\right) \le \frac{1}{k} + \tau + (1-\mathbb{P}(\mathcal{E}_1\cap\mathcal{E}_2)) \le \frac{1}{24} + \frac{1}{24} + \frac{1}{4} = \frac{1}{3}, \tag{65}$$

which completes the proof of Proposition 2. Q.E.D.

4.5. Proof of Theorem 1

We now complete the proof of Theorem 1. Assuming the validity of Hypothesis 1, Proposition 2 asserts that there exists $c_l > 0$, so that if $W_\lambda \preceq c_l \ln \frac{1}{1-\lambda}$ as $\lambda \to 1$, we must have that $\limsup_{\lambda\to 1} \mathbb{E}(L(0)) \leq 1/3$ under any sequence of optimal stationary policies. However, this contradicts the requirement that $\mathbb{E}(L(0)) \geq 1/2$, given in Eq. (27), which holds independently of the validity of Hypothesis 1. Therefore, we conclude that Hypothesis 1 must be invalid.

The invalidity of Hypothesis 1 establishes the lower bound in Eq. (7), as follows. The negation of the statement of Hypothesis 1 directly implies that there exists $c_l > 0$, so that if $W_\lambda \preceq c_l \ln \frac{1}{1-\lambda}$, as $\lambda \to 1$, then, for any sequence $\{\lambda_n\}$ in $(1-p, 1)$, with $\lim_{n\to\infty} \lambda_n = 1$, we have that

$$\limsup_{n\to\infty} \frac{\mathcal{Q}^*(\lambda_n, W_{\lambda_n})}{\ln \frac{1}{1-\lambda_n}} > 0. \tag{66}$$

We can further strengthen Eq. (66), and claim that, for any such sequence, we also have that

$$\liminf_{n\to\infty} \frac{\mathcal{Q}^*(\lambda_n, W_{\lambda_n})}{\ln \frac{1}{1-\lambda_n}} > 0. \tag{67}$$

To show Eq. (67), suppose, for the sake of contradiction, that $\liminf_{n\to\infty} \frac{\mathcal{Q}^*(\lambda_n, W_{\lambda_n})}{\ln \frac{1}{1-\lambda_n}} = 0$, for some sequence $\{\lambda_n\}$. This implies that $\{\lambda_n\}$ admits a subsequence, $\{\lambda_{n_k}\}$, such that $\limsup_{k\to\infty} \frac{\mathcal{Q}^*(\lambda_{n_k}, W_{\lambda_n})}{\ln \frac{1}{1-\lambda_{n_k}}} = 0$. The existence of the sequence $\{\lambda_{n_k}\}$ contradicts Eq. (66). This proves Eq. (67), which in turn establishes the lower bound in Eq. (7), i.e., that if $W_\lambda \preceq c_l \ln \frac{1}{1-\lambda}$, as $\lambda \to 1$, then

$$\mathcal{Q}^*(\lambda, W_\lambda) \succcurlyeq \ln\left(\frac{1}{1-\lambda}\right), \quad \text{as } \lambda \to 1. \tag{68}$$

Finally, we show that the lower bound in Eq. (7) is achievable, i.e., that

$$\mathcal{Q}^*(\lambda, W_\lambda) \preccurlyeq \ln\left(\frac{1}{1-\lambda}\right), \quad \text{as } \lambda \to 1, \tag{69}$$

when $W_\lambda \preceq c_l \ln \frac{1}{1-\lambda}$. To this end, we invoke Theorem 7 in Spencer et al. (2014), which shows that a deterministic queue-length-based diversion policy can achieve the scaling of Eq. (69), even when $W_\lambda = 0$.[5] This completes the proof of Theorem 1. Q.E.D.

[5] As is described in Spencer et al. (2014), the scaling in Eq. (69) can be achieved by the following simple threshold policy: divert the arrival if and only if the current queue length is equal to a threshold value x, where x is set to be the smallest value such that the resultant rate of diversion is no more than p. Since the queue length process under this policy is simply a birth-death process truncated at state x, it is easy to verify, via a direct calculation of steady-state probabilities of $Q(t)$, that $q_\lambda \sim \ln \frac{1}{1-\lambda}$, as $\lambda \to 1$.

5. Conclusions and Future Work

In the context of a class of queueing admission control problems, we showed that a non-trivial amount of future information is necessary in order to achieve superior heavy-traffic delay performance compared to an online policy. Theorem 1 also resolves a conjecture posed by Spencer et al. (2014). Our proof exploited certain excursion properties of a transient random walk, which allowed us to connect a policy's diversion decisions to subsequent system idling.

There are several interesting avenues of future research. First, in light of Theorem 1 and the results of Spencer et al. (2014) (Eq. (4)), an immediate question is whether the constants c_h and c_l in the scaling of W_λ coincide. The granularity of our proof technique does not appear to be sufficient to answer this question, which likely demands a finer analysis.

Because our proof relies mostly on the macroscopic properties of the input sample paths, the techniques and resultant insights in this paper seem to be fairly robust and can potentially be generalized to derive lower bounds on the necessary amount of future information for other resource allocation problems. For example, one generalization could be for a setting where the arrival and service token processes are non-Poisson (e.g., renewal or phase-type processes). In this case, we expect similar arguments to work when the process, $S(0,\cdot)$, admits similar excursion properties as in the case of Poisson processes, and does not exhibit substantial long-range correlations (for otherwise, one could potentially obtain more future information by looking into the history of past inputs). Another possibility would be to consider systems with multiple queues, in which case the relevant excursion properties of the input processes would likely be connected to those of random walks in higher dimensions. Yet another variation would be to relax the hard diversion rate constraint, and consider instead the scenario where the system manager is interested in minimizing some combined cost as a function of the delay and diversion rate. However, depending on the cost function, one may need to adjust the performance metric or regime of interest, since the system may not ever have to become critically loaded, simply because the cost structure would encourage a higher rate of diversion as the system load increases.

Finally, at a higher level, while our result focuses on the *quantity* of future information, measured by the length of a lookahead window, there is another important dimension of *quality.* For instance, the observed future input may differ from the actual realizations due to prediction noise, or alternatively, only distributional information of future input is available. Neither our results, nor those of Spencer et al. (2014), deal with the impact of prediction noise, and Xu and Chan (2014) considers only a specific noise model induced by random no-shows. A rigorous understanding of the impact of prediction accuracy in the context of dynamic resource allocation problems could be a promising direction for future research.

6. Acknowledgment

The research reported in this paper was supported in part by NSF grant CMMI-1234062 and a Claude E. Shannon Research Assistantship. The author would like to thank the anonymous reviewers for their detailed and insightful comments.

References

Brown, D. B., J. E. Smith, P. Sun. 2010. Information relaxations and duality in stochastic dynamic programs. *Operations Research* **58**(4) 785–801.

Desai, V. V., V. F. Farias, C. C. Moallemi. 2012. Pathwise optimization for optimal stopping problems. *Management Science* **58**(12) 2292–2308.

Fisher, M., A. Raman. 1996. Reducing the cost of demand uncertainty through accurate response to early sales. *Operations Research* **44**(1) 87–99.

Gonzlez-Hernández, J., C. E. Villarreal. 2011. Optimal policies for constrained average-cost Markov decision processes. *TOP* **19**(1) 107–120.

Hernández-Lerma, O., J. Gonzalez-Hernández, R. R. López-Martinez. 2003. Constrained average cost Markov control processes in Bspaces. *SIAM Journal on Control and Optimization* **42**(2) 442–468.

Kim, S. C., I. Horowitz. 2002. Scheduling hospital services: The efficacy of elective surgery quotas. *Omega* **30** 335–346.

Rogers, L. C. G. 2007. Pathwise stochastic optimal control. *SIAM J. Control Optim.* **46**(3) 1116–1132.

Spencer, J., M. Sudan, K. Xu. 2014. Queuing with future information. *Annals of Applied Probability* **24**(5) 2091–2142.

Stidham, S.Jr. 1985. Optimal control of admission to a queueing system. *IEEE Trans. Automatic Control* **30**(8) 705–713.

Sun, Yan, Bee H Heng, Yian T Seow, Eillyne Seow. 2009. Forecasting daily attendances at an emergency department to aid resource planning. *BMC emergency medicine* **9**(1) 1.

Tsitsiklis, J. N., K. Xu. 2012. On the power of (even a little) resource pooling. *Stochastic Systems* **2**(1) 1–66.

Wargon, M., B. Guidet, T.D. Hoang, G. Hejblum. 2009. A systematic review of models for forecasting the number of emergency department visits. *Emergency Medicine Journal* **26**(6) 395–399.

Xu, K. 2014. On the power of (even a little) flexibility in dynamic resource allocation. Ph.D. thesis, Massachusetts Institute of Technology.

Xu, K., C. W. Chan. 2014. Using future information to reduce waiting times in the emergency department via diversion. *Manuscript* .

Yechiali, U. 1971. On optimal balking rules and toll charges in the $GI/M/1$ queuing process. *Operations Research* **19**(2) 349–370.

Appendix

A. Additional Proofs

A.1. Proof of Lemma 1

Proof. We will formulate our admission control problem as a discrete-time Markov decision process (MDP), and invoke existing results to verify the existence of a stable stationary optimal policy. Recall that state of the system at the nth step is $X_n = (Q(Z_n-), F_n)$, where F_n was defined in Eq. (20). Define $\mathcal{X}$ as the set

$$\mathcal{X} = \mathbb{Z}_+ \times [-1, w]^{\mathbb{N}}. \tag{70}$$

Note that X_n can be represented as an element in $\mathcal{X}$ for all n, because $Q(Z_n-)$ is the queue length just before the nth event and hence belongs to $\mathbb{Z}_+$, and each coordinate of F_n, which either represents the type of an event or an inter-arrival time upper-bounded by W_λ, lies in the interval $[-1, W_\lambda]$. The following topological properties of $\mathcal{X}$ are useful, whose proof is given in Appendix A.6.

Lemma 7. *The following holds.*

1. *$\mathcal{X}$ is Polish, i.e., it is complete and separable.*
2. *Under an appropriate metric, the set $\{x \in \mathcal{X} : x_1 \le a\}$ is compact for all $a \in \mathbb{R}_+$.*

The MDP associated with our admission control problem is defined as follows:

1. The state space is $\mathcal{X}$, defined in Eq. (70).
2. The action space, $\mathcal{L}$, is the closed interval $[0,1]$, and the action at step n, $l_n \in \mathcal{L}$, specifies the probability of diversion, i.e., $l_n = \mathbb{P}(H(n) = 1)$. Denote by $\mathcal{L}(X)$ the set of allowable actions when the system is in state X. Then $\mathcal{L}(X_n) = [0,1]$ if $A_n(0) = 1$, which corresponds to the nth event being an arrival, and $\mathcal{L}(X_n) = \{0\}$ if $S_n(0) = 1$, which corresponds to the nth event being the generation of a service token.
3. The stochastic kernel is the one associated with the Poisson arrival and service token processes, as well as the queueing and diversion dynamics.
4. The nth step is associated with a *penalty*, $f(X_n, l_n)$, which is equal to the queue length, $Q(Z_n-)$. It also incurs a *cost*, $c(X_n, l_n)$, which is equal to the probability of diversion, l_n.
5. The objective is to minimize the time-average penalty, defined in Eq. (15), subject to a constraint on the time-average cost, defined in Eq. (14).

Theorem 3.2 and Lemma 3.5 of Hernández-Lerma et al. (2003) show that an MDP of this kind admits a stable stationary optimal policy, provided that a set of conditions are satisfied, which are given in Section 2 and Assumption 3.1 of Hernández-Lerma et al. (2003). These conditions are met by our MDP, and we highlight a few among them: (1) the state space is Polish (by the first claim of Lemma 7), (2) the set $\{(X,l)\in(\mathcal{X},\mathcal{L}): f(X_n,l_n)\leq a\}$ is compact for all $a\in\mathbb{R}_+$ (by the second claim of Lemma 7), (3) $c(X_n,l_n)$, which in our case is simply equal to l_n, is non-negative and lower semi-continuous in l_n for every state $X_n\in\mathcal{X}$, and (4) the stochastic kernel satisfies a certain weak continuity condition, which essentially requires the distribution of X_n not vary abruptly as a function of the state-action pair (X_n,l_n), and this continuity condition can be verified by using the definitions of Poisson processes and the associated queueing dynamics. This completes the proof of Lemma 1. Q.E.D.

A.2. Proof of Lemma 3

Proof. Recall from Eq. (11) that $S(s,t)$ is defined as the difference between the numbers of arrivals and service tokens in $(s,t]$. Since the arrival and service tokens processes are independent Poisson processes with rate λ and $1-p$, respectively, it is not difficult to verify that

$$S(s,t)\overset{d}{=}\sum_{n=1}^{N_{s,t}}X_n, \tag{71}$$

where $N_{s,t}$ is a Poisson random variable with mean $(\lambda+1-p)(t-s)$, which corresponds to the total number of events in $(t,s]$, and the X_ns are i.i.d., with

$$X_1=\begin{cases}1, & \text{w.p. } \frac{\lambda}{\lambda+1-p},\\ -1, & \text{otherwise,}\end{cases} \tag{72}$$

By Eq. (71), and the fact that $\lim_{B\to\infty}\frac{N_{s,s+B}}{B}=\lambda+1-p$ almost surely, Claim 1 follows from a variation of the standard Functional Law of Large Numbers (FLLN) for the sum of bounded i.i.d. random variables. Claim 3 follows from the Weak Law of Large Numbers applied to the sum of i.i.d. Poisson random variables, and our assumption that $W_\lambda\to\infty$ as $\lambda\to 1$ (Eq. (16)). Finally, Claim 2 follows from the Markov's inequality, in the same way as in Eq. (27), by noting that $\mathbb{E}(Q(0))=q_\lambda$ under an optimal stationary policy. Q.E.D.

A.3. Proof of Lemma 4

Proof. Based on the stationarity of $\mathcal{A}$ and $\mathcal{S}$, and the assumption that $B=kW_\lambda$ and $q_\lambda\ll W_\lambda$, it suffices for us to show, that for any $a,b>0$, there exists $\gamma>0$, such that

$$\mathbb{P}\left(S(0,aW_\lambda)\leq -bW_\lambda\right)\succcurlyeq\exp(-\gamma W_\lambda),\quad \text{as } \lambda\to 1. \tag{73}$$

By definition, the distribution of $S(0,t)$ can be written as

$$S(0,t)\overset{d}{=}A_{\lambda t}-D_{(1-p)t}, \tag{74}$$

where $A_{\lambda t}$ and $D_{(1-p)t}$ are independent Poisson random variables with mean λt and $(1-p)t$, respectively. The following lemma follows from the standard large-deviation principles of Poisson random variables, and its proof is omitted.

Lemma 8. *Let D_x be a Poisson random variable with mean x. Then, for all $c_1 > 0$, there exists $c_2 > 0$, such that*

$$\mathbb{P}(D_x \geq c_1 x) \succsim \exp(-c_2 x), \quad \textit{as } x \to \infty. \tag{75}$$

Combining Lemma 8 and the fact that $W_\lambda \to \infty$ as $\lambda \to 1$, we have that there exists $\gamma > 0$, such that

$$\mathbb{P}\left(D_{(1-p)aW_\lambda} \geq (b+2a)W_\lambda\right) \succsim \exp(-\gamma W_\lambda) \tag{76}$$

as $\lambda \to 1$. We have that

$$\begin{aligned}
&\mathbb{P}(S(0, aW_\lambda) \leq -bW_\lambda) \\
\geq &\mathbb{P}\left(\left\{A_{\lambda a W_\lambda} < 2aW_\lambda\right\} \cap \left\{D_{(1-p)aW_\lambda} \geq (b+2a)W_\lambda\right\}\right) \\
\overset{(a)}{=} &\mathbb{P}\left(A_{\lambda a W_\lambda} < 2aW_\lambda\right)\mathbb{P}\left(D_{(1-p)aW_\lambda} \geq (b+2a)W_\lambda\right) \\
\overset{(b)}{\geq} &\mathbb{P}\left(A_{\lambda a W_\lambda} < 2\lambda aW_\lambda\right)\mathbb{P}\left(D_{(1-p)aW_\lambda} \geq (b+2a)W_\lambda\right) \\
\overset{(c)}{\geq} &\frac{1}{2}\mathbb{P}\left(D_{(1-p)aW_\lambda} \geq (b+2a)W_\lambda\right) \\
\overset{(d)}{\succsim} &\exp(-\gamma W_\lambda),
\end{aligned} \tag{77}$$

as $\lambda \to 1$, where step (a) follows from the independence between $A_{\lambda a W_\lambda}$ and $D_{(1-p)aW_\lambda}$, (b) from the fact that $\lambda < 1$, (c) from the Markov's inequality, and (d) from Eq. (76). This proves Eq. (4), and hence Lemma 4. Q.E.D.

A.4. Proof of Lemma 5

Proof. For Claim 1, observe that each of the event concerns only the behavior of the arrival and service token processes over an interval, and that these intervals are disjoint from each other. Claim 1 follows by noting that both $\mathcal{A}$ and $\mathcal{S}$ are Poisson processes and hence memoryless. For Claim 2, because the policy has access to a lookahead window of length W_λ, the queue length at time t is hence $\mathcal{F}_{t+W_\lambda}$ measurable, where $\mathcal{F}$ is the natural filtration induced by the input processes. The claim follows again from the memoryless property of Poisson processes. Claim 3 follows from the same arguments as for Claim 2. Q.E.D.

A.5. Proof of Lemma 6

Proof. Consider the sequence of optimal stationary policies, $\{\pi_\lambda\}$. Let ϕ be defined as in Eq. (34). Fix $\phi > 0$, and let

$$K = U_3 + \phi W_\lambda \overset{(a)}{=} (k + \phi + 2)W_\lambda, \tag{78}$$

where step (a) follows from the fact that $U_3 = B + 2W_\lambda$ and $B = kW_\lambda$. The main idea for the proof is based on the following observation: conditional on $\cap_{i=1}^{5}\mathcal{E}_i$, the queue length process, $Q(t)$, would have reached zero before time K, even if *no* diversion had been made in $(0,K]$ (illustrated in Figure 3). Therefore, each diversion made in $(U_1,U_2]$ will necessarily lead to a *waste service token* in $(0,K]$, and hence

$$\mathbb{P}\left(\mathcal{J}(K)\geq \tau B \,\middle|\, \cap_{i=1}^{5}\mathcal{E}_i\right)\geq \mathbb{P}\left(Y\geq \tau B \,\middle|\, \cap_{i=1}^{5}\mathcal{E}_i\right). \tag{79}$$

We next give a lower bound on the above probability, as follows:

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{J}(K)\geq \tau B \,\middle|\, \mathcal{E}_1\cap\mathcal{E}_2\right)\\
\geq&\mathbb{P}\left(\mathcal{J}(K)\geq \tau B, \cap_{i=3}^{5}\mathcal{E}_i \,\middle|\, \mathcal{E}_1\cap\mathcal{E}_2\right)\\
=&\mathbb{P}\left(\mathcal{J}(K)\geq \tau B \,\middle|\, \cap_{i=1}^{5}\mathcal{E}_i\right)\mathbb{P}\left(\cap_{i=3}^{5}\mathcal{E}_i \,\middle|\, \mathcal{E}_1\cap\mathcal{E}_2\right)\\
\overset{(a)}{\geq}&\mathbb{P}\left(Y\geq \tau B \,\middle|\, \cap_{i=1}^{5}\mathcal{E}_i\right)\mathbb{P}\left(\cap_{i=3}^{5}\mathcal{E}_i \,\middle|\, \mathcal{E}_1\cap\mathcal{E}_2\right)\\
=&\mathbb{P}\left(Y\geq \tau B, \cap_{i=3}^{5}\mathcal{E}_i \,\middle|\, \mathcal{E}_1\cap\mathcal{E}_2\right)\\
\overset{(b)}{=}&\mathbb{P}\left(\mathcal{E}_5\right)\mathbb{P}\left(Y\geq \tau B, \mathcal{E}_3\cap\mathcal{E}_4 \,\middle|\, \mathcal{E}_1\cap\mathcal{E}_2\right)\\
\geq&\mathbb{P}\left(\mathcal{E}_5\right)\left(\mathbb{P}\left(Y\geq \tau B \,\middle|\, \mathcal{E}_1\cap\mathcal{E}_2\right)+\mathbb{P}\left(\mathcal{E}_3 \,\middle|\, \mathcal{E}_1\cap\mathcal{E}_2\right)+\mathbb{P}\left(\mathcal{E}_4 \,\middle|\, \mathcal{E}_1\cap\mathcal{E}_2\right)-2\right)\\
\geq&\mathbb{P}\left(\mathcal{E}_5\right)\left(\mathbb{P}\left(Y\geq \tau B \,\middle|\, \mathcal{E}_1\cap\mathcal{E}_2\right)+\mathbb{P}\left(\mathcal{E}_3 \,\middle|\, \mathcal{E}_1\cap\mathcal{E}_2\right)+\frac{\mathbb{P}\left(\mathcal{E}_4\right)+\mathbb{P}\left(\mathcal{E}_1\cap\mathcal{E}_2\right)-1}{\mathbb{P}\left(\mathcal{E}_1\cap\mathcal{E}_2\right)}-2\right)\\
\overset{(c)}{=}&\mathbb{P}\left(\mathcal{E}_5\right)\left(\mathbb{P}\left(Y\geq \tau B \,\middle|\, \mathcal{E}_1\cap\mathcal{E}_2\right)+\mathbb{P}\left(\mathcal{E}_3\right)+\frac{\mathbb{P}\left(\mathcal{E}_4\right)-1}{\mathbb{P}\left(\mathcal{E}_1\cap\mathcal{E}_2\right)}-1\right)
\end{aligned} \tag{80}$$

where step (a) follows from Eq. (79), and (b) and (c) from the independence between $\mathcal{E}_5$ and $\mathcal{E}_1\cap\mathcal{E}_2$, and between $\mathcal{E}_3$ and $\mathcal{E}_1\cap\mathcal{E}_2$, respectively (Lemma 5). We have also used the inequality that $\mathbb{P}(A\cap B)\geq \mathbb{P}(A)+\mathbb{P}(B)-1$, for any events A and B.

By Claim 3 of Lemma 3, we have that

$$\lim_{\lambda\to 1}\mathbb{P}\left(\mathcal{E}_3\right)=\lim_{\lambda\to 1}\mathbb{P}\left(\mathcal{E}_4\right)=1. \tag{81}$$

Combining the assumption (Eq. (41))

$$\liminf_{\lambda\to 1}\mathbb{P}\left(Y\geq \tau B \,\middle|\, \mathcal{E}_1\cap\mathcal{E}_2\right)=q>0 \tag{82}$$

with Eqs. (80) and (81), we have that there exists $\tilde{\lambda}\in(0,1)$, such that

$$\begin{aligned}
\mathbb{P}\left(\mathcal{J}(K)\geq \tau B \,\middle|\, \mathcal{E}_1\cap\mathcal{E}_2\right)\geq&\mathbb{P}\left(\mathcal{E}_5\right)\mathbb{P}\left(Y\geq \tau B \,\middle|\, \mathcal{E}_1\cap\mathcal{E}_2\right)\\
\geq&\mathbb{P}\left(\mathcal{E}_5\right)q/2,
\end{aligned} \tag{83}$$

for all $\lambda\in\left(\tilde{\lambda},1\right)$. We have that

$$\mathbb{E}\left(\mathcal{J}(K)\right)\geq\tau B\cdot\mathbb{P}\left(\mathcal{J}(K)\geq\tau B\right)$$

$$
\begin{aligned}
&\geq \tau B \cdot \mathbb{P}\left(\mathcal{J}(K) \geq \tau B, \mathcal{E}_1 \cap \mathcal{E}_2\right) \\
&= \tau B \cdot \mathbb{P}\left(\mathcal{J}(K) \geq \tau B \,\middle|\, \mathcal{E}_1 \cap \mathcal{E}_2\right) \cdot \mathbb{P}\left(\mathcal{E}_1 \cap \mathcal{E}_2\right) \\
&\overset{(a)}{\succsim} B \mathbb{P}\left(\mathcal{E}_5\right) \mathbb{P}\left(\mathcal{E}_1 \cap \mathcal{E}_2\right) \\
&\overset{(b)}{\succsim} B \mathbb{P}\left(\mathcal{E}_5\right) \\
&\overset{(c)}{\succsim} B \exp\left(-\gamma W_\lambda\right),
\end{aligned}
\tag{84}
$$

for some $\gamma > 0$, as $\lambda \to 1$, where step (a) follows from Eq. (83), (b) from Claims 1 and 2 of Lemma 3 and the independence of the events $\mathcal{E}_1$ and $\mathcal{E}_2$ (Claim 1 of Lemma 5), i.e., that

$$
\mathbb{P}\left(\mathcal{E}_1 \cap \mathcal{E}_2\right) = \mathbb{P}\left(\mathcal{E}_1\right) \mathbb{P}\left(\mathcal{E}_2\right) \geq \frac{5}{6}\theta, \tag{85}
$$

and (c) from Lemma 4. This proves Lemma 6, by setting $a = k + \phi + 2$. *Q.E.D.*

A.6. Proof of Lemma 7

Proof. Let $\mathcal{X}_0 = [-1, W_\lambda]^{\mathbb{N}}$. We will show that $\mathcal{X}_0$ is compact under the metric $\|x - y\|_g = \sum_{i=1}^{\infty} 2^{-i} |x_i - y_i|$. If this is true, it is not difficult to show that, for any $a \in \mathbb{R}_+$, the set $\{x \in \mathcal{X} : x_1 \leq a\} = \{0, \ldots, \lfloor a \rfloor\} \times \mathcal{X}_0$ is also compact under $\| \cdot \|_g$, and our second claim follows. Note that a compact metrizable space is Polish, and it is easy to show that $\mathbb{Z}_+$ is Polish under the l_1 norm. Our first claim thus also follows from the compactness of $\mathcal{X}_0$, by observing that the product of two Polish spaces remains Polish.

We now show the compactness of $\mathcal{X}_0$. It suffices to show that any sequence in $\mathcal{X}_0$, $\{x^i\}_{i \in \mathbb{N}}$, admits a sub-sequence that converges to a point in $\mathcal{X}_0$. We will construct such a limiting point coordinate-by-coordinate, as follows. Because x_1^i is an element of the compact interval $[-1, W_\lambda]$ for all $i \in \mathbb{N}$, there exists $y_1 \in [-1, W_\lambda]$ and an increasing sequence, $\{i^{1,j}\}_{j \in \mathbb{N}} \subset \mathbb{N}$, such that $\lim_{j \to \infty} x_1^{i^{1,j}} = y_1$. We now apply the same reasoning for progressively larger values of k: there exist $y_k \in [-1, W_\lambda]$ and $\{i^{k,j}\}_{j \in \mathbb{N}}$ for $k = 2, 3, \ldots$, such that, for every $k \geq 2$, $\{i^{k,j}\}_{j \in \mathbb{N}}$ is a sub-sequence of $\{i^{k-1,j}\}_{j \in \mathbb{N}}$, and

$$
\lim_{j \to \infty} x_k^{i^{k,j}} = y_k. \tag{86}
$$

Fix $k \geq 2$. Because $\{i^{k,j}\}_{j \in \mathbb{N}}$ is a sub-sequence of $\{i^{m,j}\}_{j \in \mathbb{N}}$ for all $m \leq k - 1$, Eq. (86) further implies that

$$
\lim_{j \to \infty} x_m^{i^{k,j}} = y_m, \quad \forall m \in \{1, \ldots, k\}, \tag{87}
$$

or, equivalently, that

$$
\lim_{j \to \infty} \sum_{i=1}^{k} 2^{-m} \left| x_m^{i^{k,j}} - y_m \right| = 0, \quad \forall k \in \mathbb{N}. \tag{88}
$$

Let y be the element of $\mathcal{X}_0$ whose coordinates are defined according to the above procedure. We argue that y is the limiting point for some sub-sequence of $\{x^i\}_{i=1}^{\mathbb{N}}$. For every $k \in \mathbb{N}$, there exists $j(k) \in \mathbb{N}$, such that for all $j \geq j(k)$,

$$\begin{aligned}\left\| y - x^{i^{k,j}} \right\|_g &= \sum_{m=1}^{\infty} 2^{-m} \left| y_m - x_m^{i^{k,j}} \right| \\ &\overset{(a)}{\leq} \left(\sum_{m=1}^{k} 2^{-m} \left| y_m - x_m^{i^{k,j}} \right| \right) + (1+W_\lambda) 2^{-(k-2)} \\ &\overset{(b)}{\leq} \frac{1}{k} + (1+W_\lambda) 2^{-(k-2)}, \end{aligned} \tag{89}$$

where step (a) follows from the fact that $\left| y_m - x_m^{i^{k,j}} \right| \leq 2(W_\lambda + 1)$ for all $m \in \mathbb{N}$, and step (b) from Eq. (88). Define

$$n^k = \max\{i^{m,j(m)} : 1 \leq m \leq k\}, \quad \forall k \in \mathbb{N}. \tag{90}$$

By Eq. (89), we have that

$$\left\| y - x^{n^k} \right\|_g \leq \frac{1}{k} + (1+W_\lambda) 2^{-(k-2)}, \quad \forall k \in \mathbb{N}, \tag{91}$$

Therefore, $\{x^{n^k}\}_{k\in\mathbb{N}}$ is a sub-sequence of $\{x^i\}_{i\in\mathbb{N}}$, and it converges to y as $k \to \infty$ under the metric $\|\cdot\|_g$. This proves that $\mathcal{X}_0$ is compact. Q.E.D.