Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Asymptotically Optimal Appointment Schedules

Mor Armony, Rami Atar, Harsha Honnappa

To cite this article:
Mor Armony, Rami Atar, Harsha Honnappa (2019) Asymptotically Optimal Appointment Schedules. Mathematics of Operations Research 44(4):1345-1380. https://doi.org/10.1287/moor.2018.0973

http://pubsonline.informs.org/journal/moor

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 44, No. 4, November 2019, pp. 1345–1380
ISSN 0364-765X (print), ISSN 1526-5471 (online)

Asymptotically Optimal Appointment Schedules

Mor Armony,[a] Rami Atar,[b] Harsha Honnappa[c]

[a] Stern School of Business, New York University, New York, New York 10012; [b] Viterbi Faculty of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 3200003, Israel; [c] School of Industrial Engineering, Purdue University, West Lafayette, Indiana 47907

Contact: marmony@stern.nyu.edu, http://orcid.org/0000-0001-5970-3753 (MA); atar@ee.technion.ac.il, http://orcid.org/0000-0001-8341-5049 (RA); honnappa@purdue.edu (HH)

Received: September 8, 2017
Revised: July 24, 2018
Accepted: August 5, 2018
Published Online in Articles in Advance: August 6, 2019

MSC2000 Subject Classification: Primary: 68M20, 90B36; secondary: 60K25, 90C15
OR/MS Subject Classification: Primary: queues: optimization; secondary: production/scheduling: sequencing stochastic, programming: stochastic

https://doi.org/10.1287/moor.2018.0973

Abstract. We consider the problem of scheduling appointments for a finite customer population to a service facility with customer no-shows to minimize the sum of customer waiting time and server overtime costs. Because appointments need to be scheduled ahead of time, we refer to this problem as an optimization problem rather than a dynamic control one. We study this optimization problem in fluid and diffusion scales and identify asymptotically optimal schedules in both scales. In fluid scale, we show that it is optimal to schedule appointments so that the system is in critical load; thus, heavy-traffic conditions are obtained as a result of *optimization* rather than as an *assumption*. In diffusion scale, we solve this optimization problem in the large horizon limit. Our explicit stationary solution of the corresponding Brownian optimization problem translates the customer delay versus server overtime trade-off to a trade-off between the state of a reflected Brownian motion in the half-line and its local time at zero. Motivated by work on competitive ratios, we also consider a reference model in which an oracle provides the decision maker with the complete randomness information. The difference between the values of the scheduling problem for the two models, to which we refer as the *stochasticity gap* (SG), quantifies the degree to which it is harder to design a schedule under uncertainty than when the stochastic primitives (i.e., the no-shows and service times) are known in advance. In the fluid scale, the SG converges to zero, but in the diffusion scale, it converges to a positive constant that we compute.

Funding: The research of R. Atar was supported in part by the Israel Science Foundation [Grant 1184/16]. The research of H. Honnappa was supported in part by the National Science Foundation Division of Civil, Mechanical & Manufacturing Innovation [Grant CMMI-1636069].

Keywords: queues • appointment scheduling • stochastic optimization • asymptotic optimality • fluid approximation • diffusion optimization • stochasticity gap

1. Introduction

We study the problem of determining an optimal appointment schedule for a finite number of customers at a service system that only accepts arrivals in a finite time horizon but renders service to all arriving customers. Broadly, the objective is to assign deterministic arrival epochs to a finite population such that the server is optimally utilized while the cumulative delay experienced by the customers is minimized. The optimization problem is stochastic in nature not only because of the randomness in service times, but also because some arrivals do not show up. As opposed to typical stochastic control problems that appear in the literature, the schedule needs to be determined *off-line*, ahead of time, with no access to the realization of the stochastic primitives over time. Thus, in standard queueing terminology, the problem under consideration is about selecting arrival times for a given number of customers into a $\cdot/G/1$ queue over a finite time horizon so as to minimize delay and server utilization costs and in presence of customer no-shows.

Our research is motivated by systems such as outpatient clinics that render service to a finite number of patients during a working day (7 a.m. to 4 p.m., for instance). No patients are accepted for service if they arrive after the end of the horizon, but all patients that do arrive are rendered service. No-shows in outpatient care is a problem most clinics struggle with regularly. According to Cayirli et al. [9], no-show rates may be up to 60%, depending on the clinic-specific characteristics. Patient overbooking has been proposed as an effective strategy to handle clinic underutilization resulting from patient no-show (LaGanga and Lawrence [23]). At the same time, overbooking may lead to clinic overcrowding that will intensify patient waiting time and doctor overtime. Our paper aims at determining effective appointment schedules that will minimize these wait times and overtimes.

We model the service system as a single-server queue with an infinite buffer. We assume that the service times are generally distributed, independent and identically distributed (IID), nonnegative random variables

(RVs) with a finite second moment and that the server is nonidling and operates according to a first-come, first-served (FCFS) discipline. The entire finite population needs to be provided with appointments, and these appointments are allowed to span the entire time horizon but not beyond. The actual arrival process is thinned with probability p because of no-shows from this deterministic appointment schedule.

Our optimization problem's goal is to determine an appointment schedule to minimize the objective of a weighted sum of the expected cumulative wait time of all the customers that arrive in the (finite) arrival horizon and the expected overage time, defined as the amount of time it takes for the server to clear out the backlog after the end of the horizon. The appointment-scheduling problem is reducible to a stochastic bin-packing problem, making it NP-hard, and is generally solved using various heuristics (see Gupta and Denton [14]).

Here, we introduce two large population-limit regimes that lead to simpler optimization problems and yield exact solutions. We operate in a large population-limit framework that reveals the fundamental complicating factors in the optimization problem. Our scaling regimes let the population size tend to infinity while simultaneously accelerating the service rate in proportion to the population size. In the fluid regime, the service time cost is scaled by the inverse of the population size. Customers are assigned arrival epochs according to a sequence of arbitrary schedules. In the limit, the fluid regime washes out the stochastic variation and captures the "mean" or first-order effects in the queue performance. We posit a variational fluid optimization problem (FOP) and solve it in Proposition 1 under an "overload" condition that the aggregate available fluid service is less than the expected aggregate fluid arrivals, and thus, the overtime cost is nonzero. The optimal cumulative fluid schedule function matches the cumulative service completions in the arrival horizon and schedules the remaining fluid at the end of the horizon. This result shows that the heavy-traffic, critical-load condition emerges as a consequence of optimization as opposed to a postulated assumption. In Theorems 1 and 2, we prove that the value of the fluid optimization problem is asymptotically achievable by a carefully constructed sequence of simple finite population schedules.

Although the FOP results in a simple and intuitive schedule, it only considers first-order effects and washes away stochastic fluctuations. In reality, we expect the inherent stochasticity of the system to have a significant impact on system performance and, thus, on the design of the schedule. Indeed, one might ask whether considering second-order terms would shed more light on the optimal appointment schedule. Specifically, it is of interest to see how fluctuations of order of the square root of the population size about the fluid solution impact the schedule. To formalize this question, we posit a Brownian optimization problem (BOP), assuming the same overload condition as in the fluid scale. The BOP is stated in terms of an equation driven by a Brownian motion and a control, obtained as a formal diffusion limit of the queue length in heavy traffic. This is in line with the fact that the fluid-optimal schedule enforces criticality within the appointment horizon. The BOP is not a dynamic control problem, but one in which the control trajectory has to be planned ahead of time zero. This makes insights, structure, and tools from dynamic programming irrelevant. Although we are unable to obtain an explicit solution of the BOP when set on a fixed time horizon, we do derive an explicit solution to it in the large time-horizon limit. This limiting procedure allows for a long-run average argument in a way that does not trivialize the end-of-horizon effect. In Proposition 3, we prove that the value of the BOP in the large horizon limit can be achieved by a reflected Brownian motion (RBM) that has a constant negative drift. Equivalently, in this limit, the BOP is solved by a stationary RBM. We identify the optimal drift coefficient in Lemma 6. In Theorems 3 and 4, we prove that the value of the BOP is also asymptotically achievable by a carefully constructed sequence of finite population schedules.

There is, of course, a "price" to be paid for having to schedule appointments at time zero without any stochastic information revealed ahead of time. We quantify this by introducing the notion of a *stochasticity gap* (SG), defined as the difference between the appropriately scaled finite population value and the value of the "complete information" (CI) problem. In the CI problem, an oracle reveals all stochastic primitives (or future events) to the optimizer at time zero. The CI problem is not completely trivial, but much easier than the original one. The CI and the fluid-optimal schedules are similar in that they both schedule appointments such that customers arrive at (precisely or approximately, respectively) the time when they are ready to be served, and any excess jobs are scheduled at the end of the horizon. Indeed, Proposition 2 shows that in the fluid scale the asymptotic SG is zero. On the other hand, in diffusion scaling we calculate the SG and show in Proposition 4 that it is strictly positive.

To summarize, our main contributions are as follows:

i. Under an overload condition, we identify explicitly computable asymptotically optimal (AO) schedules in the fluid and diffusion regimes. Our proposed schedules are simple in that they set interarrival times to be deterministic and stationary up to the end of the horizon. They schedule the remainder of the arrivals to show

up at precisely the end-of-horizon time. Ours constitute the first analytical results for the appointment-scheduling problem with no-shows in the large population limit.

ii. A critical load condition, which sets the ground for a heavy traffic analysis at the diffusion scale, emerges because it is optimal to operate at criticality rather than as an assumption.

iii. The essence of the optimization problem is that it must be carried out without the randomness being revealed. We analyze the SG as a means of quantifying the cost associated with this uncertainty.

The rest of the paper is organized as follows. We conclude this section with a brief overview of the relevant literature and notational convention. Section 2 provides the problem formulation and a summary of the main results. In Sections 3 and 4, we solve the problem under fluid and diffusion scaling, respectively. We compute the corresponding SG in these sections as well. Section 5 contains a numerical study and its analysis. We conclude with final remarks and a discussion of future research directions in Section 6.

We now review some of the relevant results in the field. There is a vast literature on appointment scheduling in healthcare, which we will not attempt to summarize here; we direct the reader to the comprehensive reviews in Cayirli and Veral [8], Gupta and Denton [14], and Hall [15]. We note two that are particularly relevant to our study. A scheduling problem close to ours has been studied in Zacharias and Pinedo [26]. The problem they consider is that of determining an optimal schedule for heterogeneous patients in the presence of no shows with the objective of minimizing waiting cost plus idling and overtime cost. Our model may be considered a special case of their model in that our patients are homogeneous, and our idling cost is zero. However, their model considers fixed appointment slots, and ours allows for appointment times to be a result of the optimization problem. For a finite and fixed patient population, the authors characterize structural properties of an optimal schedule. Explicit solutions are given for some special cases and are studied numerically for the more general problem. The authors observe that optimal solutions tend to front-load (more overbooking toward the beginning of the day). The numerical solutions of Zacharias and Pinedo [26] also show that some appointment slots should be overbooked but not all, with only up to a couple of patients scheduled per slot.

A second paper that is relevant is Hassin and Mendel [17], which considers a similar problem definition. However, the model there assumes that service times are exponentially distributed, and there is no fixed horizon in which the finite number of arrivals must be scheduled. The cost function again trades off the expected cumulative waiting time of the customers that show up against the expected "server time" beyond the last scheduled arrival epoch. The authors provide extensive numerical analysis of the finite-population scheduling problem. In particular, they numerically compute the optimal schedule, which shows that overbooking is possible for the first few and last few arrival epochs, and arrivals "in the middle" are almost uniformly spread out. Furthermore, they also contrast the value of their problem against that of an oracle problem akin to our CI problem and note the fact that value of the latter is significantly lower.

Also relevant is Benjaafar and Jouini [3], which considers finite population scheduled arrival models with both no-shows and tardiness. Under the assumption of exponentially distributed service times, the authors derive exact expressions for various performance metrics. More recently, Kim et al. [20] develop a high-fidelity simulation model in a data-driven manner using arrival and appointment information from a single clinic and observe that randomness in the number of scheduled patients, unscheduled arrivals, and no-shows contribute to the stochasticity in the traffic pattern.

The second relevant stream of literature is on the asymptotics of scheduled arrivals (Araman and Glynn [1]) and transitory queueing (Honnappa et al. [18, 19]). This stream considers limits of processes that are generated from a queue with an arrival process that is originated from a finite population. Closest to our model is section 4.3 of Honnappa et al. [19], in which the authors consider scheduled arrivals with epoch uncertainty. Indeed, the asymptotic scaling and limiting regimes are similar to ours. The two main differences are that (i) the appointment times are given and are assumed a priori to be equally spaced and (ii) all customers are assumed to show up but they may be nonpunctual.

Perhaps the closest paper to ours is Kuiper et al. [22], in which the authors consider a similar limiting regime as ours and are also concerned with optimizing the scheduled appointments. Like Hassin and Mendel [17], the paper focuses on the infinite-horizon problem and restricts itself to the set of deterministic and stationary schedules; however, no-shows are not considered. Its objective is to minimize the sum of customer waiting cost plus server idling cost. A scheduling policy based on a diffusion approximation is proposed without formally establishing asymptotic optimality. In contrast, we establish asymptotic optimality of our proposed schedule and, by way of doing that, establish that, indeed, asymptotic optimality is achieved by optimizing only within stationary policies.

1.1. Notation

Let $\mathcal{D}[0,\infty)$ be the space of functions $f:[0,\infty)\to\mathbb{R}$ that are right continuous with left limits, equipped with the Skorokhod J_1 topology. Let $\mathcal{D}^+[0,\infty)\subset\mathcal{D}[0,\infty)$ be the subset of nonnegative, nondecreasing functions. For a sequence $\{X_n\}$, X, of RVs, $X_n\Rightarrow X$ as $n\to\infty$ denotes convergence in law. For a sequence $\{X_n\}$, X, of stochastic processes with sample paths in $\mathcal{D}[0,\infty)$, $X_n\Rightarrow X$ denotes convergence in law in the J_1 topology. In this paper, all statements involving convergence of processes $X_n\Rightarrow X$ are to processes X that have almost surely (a.s.) continuous sample paths; thus, these convergences can equivalently be understood as convergence in law in the *uniformly on compacts* (u.o.c.) topology. Let $(\cdot)^+:=\max\{\cdot,0\}$. For an event A, $\mathbf{1}_A$ is the corresponding indicator function. For $f:\mathbb{R}_+\to\mathbb{R}$, $T>0$, we denote $\|f\|_T=\sup_{t\in[0,T]}|f(t)|$. A one-dimensional Brownian motion (BM) with drift m and diffusion coefficient σ, starting from zero, is referred to as an (m,σ) BM. The letter c denotes a positive constant whose value is immaterial and may change from line to line.

2. Problem Setting

2.1. Model

Consider a single-server queue with an infinite waiting room. A finite number of jobs arrive at the queue over a finite time horizon and are served on a FCFS schedule. Jobs are given appointments at fixed times during the day (not necessarily uniformly spaced), and we assume that jobs that do turn up do so precisely at the appointment time; that is, we assume punctual arrivals but allow no-shows. We also assume that the service times are IID with finite second moments.

The RVs and stochastic processes are defined on a probability space $(\Omega,\mathcal{F},\mathbb{P})$, and the symbol $\mathbb{E}$ denotes expectation with respect to $\mathbb{P}$. The number of requested appointments (or population size) is denoted as N. Let $H>0$ denote the operating time horizon. A *schedule* is any deterministic, nondecreasing sequence $\{T_i, i=1,\dots,N\}$ taking values in $[0,H]$. It represents scheduled arrival epochs. The collection of all schedules is denoted by $\mathcal{T}$. We denote by $E(t)$ the cumulative number of scheduled arrivals by time t; that is, $E(t)=\sum_{i=1}^N \mathbf{1}_{\{T_i\le t\}}$. The function E is referred to as the *scheduling* function. Let $\{\xi_i, i\in\mathbb{N}\}$ be IID Bernoulli RVs with mean $p\in(0,1]$. They are used as a model for actual arrivals, namely $\xi_i=1$ if and only if the ith scheduled job shows up. With these elements, the *cumulative arrival* process, $A\in\mathcal{D}^+[0,H]$, is given by $A(t)=\sum_{i=1}^N \xi_i\mathbf{1}_{\{T_i\le t\}}$. Note that, with $\Xi(k)=\sum_{i=1}^k\xi_i$, $k\in\mathbb{Z}_+$, one can express this relation as

$$A(t)=\sum_{i=1}^{E(t)}\xi_i=\Xi\circ E(t). \tag{1}$$

Let $\{v_i\}$ be an IID sequence of nonnegative RVs with mean μ^{-1} and squared coefficient of variation $C_S^2\in(0,\infty)$. Assume that this sequence and the sequence $\{\xi_i\}$ are mutually independent. These RVs are used to model service times in the following way. Let $\{v_i, i\in\mathbb{N}\}$ be the service time of the ith served job. Let $S(t)=\max\{m|\sum_{i=1}^m v_i\le t\}$. Then $S(t)$ is the cumulative number of service completions by the time the server is busy for t units of time.

Let Q denote the *number-in-system* process. Then the *cumulative busyness* process is given by $B(\cdot)=\int_0^{\cdot}\mathbf{1}_{\{Q(s)>0\}}ds$. A simple balance equation for Q is

$$Q=A-S\circ B. \tag{2}$$

2.2. Cost and Optimization Problem

Two primary performance measures of interest to us are the *overall waiting time*, or *makespan*, and the *overage time*. The former is defined as

$$W=\int_0^\infty (Q(s)-1)^+ds, \tag{3}$$

and represents the sum, over all arriving jobs, of the job's waiting time in the queue (not counting the time of service). The overage time is the amount of time after the end of the horizon $[0,H]$ it takes to complete the last arrival in $[0,H]$. If we denote by τ the time when the last arrival in $[0,H]$ departs from the server, then the overage time is given by

$$O=(\tau-H)^+. \tag{4}$$

Note that τ can be expressed in terms of the processes introduced earlier as

$$\tau = \inf\{t : S \circ B(t) \geq \Xi(N)\}, \tag{5}$$

for $\Xi(N)$ is the total number of arrivals, and $S \circ B(t)$ is the number of departures by time t.

The goal of the system operator is to schedule the N jobs so as to minimize a combination of the expected overall waiting time experienced by the jobs (makespan) and the expected overage time. Thus we consider the *finite population optimal-schedule problem (FPOP)* defined by considering the cost

$$J(\{T_i\}) = c_w \mathbb{E}[W] + c_o \mathbb{E}[O], \qquad \{T_i\} \in \mathcal{T}, \tag{6}$$

and the value

$$V = \inf_{\{T_i\} \in \mathcal{T}} J(\{T_i\}). \tag{7}$$

Here, c_w and c_o are nonnegative constants. In general, solving the FPOP is formidable. Rather than solve it directly, we solve fluid- and diffusion-scale problems and prove the existence of asymptotically optimal finite population schedules that approach the fluid- and diffusion-optimal solutions in the large population limit.

The main objective of this work is to study the asymptotics of problems (6) and (7). Before we turn to it, we comment on another, related problem setting. Suppose that all the information on the stochastic data are known to the decision maker when the decision maker selects the schedule at time zero. In this version of the problem, which we refer to as the CI problem, the selection of schedule $\{T_i\}$ may depend on the stochastic data $(\{\xi_i\}, \{\nu_i\})$. We do not consider it as a practically motivated setting by itself because in applications we have in mind these stochastic ingredients are not known in advance. However, it is useful to regard it as a reference model and to relate it to our main problems (6) and (7). We also observe that the CI problem represents an *information relaxation* of the actual problem, and the solution to the CI problem yields what can be considered the expected value of perfect information in stochastic programming and dynamic programming (Avriel and Williams [2], Dempster [11], Birge [5], Brown et al. [7], Brown and Haugh [6]).

For a precise formulation, let Σ denote the sigma field generated by the collection $(\{\xi_i\}, \{\nu_i\})$. An Σ-measurable RV $\{\Theta_i\}$ taking values in $\mathcal{T}$ is called a *CI schedule*. Let $\mathcal{T}^{\mathrm{CI}}$ denote the collection of all CI schedules (note that schedules in $\mathcal{T}^{\mathrm{CI}}$ are not allowed to change the *order* of the scheduled jobs). Then, analogously to (6), we let

$$J^{\mathrm{CI}}(\{\Theta_i\}) = c_w \mathbb{E}[W] + c_o \mathbb{E}[O], \qquad \{\Theta_i\} \in \mathcal{T}^{\mathrm{CI}}, \tag{8}$$

$$V^{\mathrm{CI}} = \inf_{\{\Theta_i\} \in \mathcal{T}^{\mathrm{CI}}} J^{\mathrm{CI}}(\{\Theta_i\}), \tag{9}$$

where W and O correspond to the selection $\{\Theta_i\}$. Clearly, one always has $V \geq V^{\mathrm{CI}}$. We refer to the difference

$$\gamma = V - V^{\mathrm{CI}}$$

as the stochasticity gap as it quantifies the gap between performance with and without knowing the stochastic ingredients.

Problem (8)–(9) is much easier than our main problem and is in fact, fully solvable. We devote Section 2.5 to present its solution.

2.3. Large-Population Asymptotic Framework

Because problem (6)–(7) is prohibitively difficult to solve exactly, we instead take a large-population asymptotic approach in which we consider a sequence of systems in which the number of scheduled jobs grows large and the cost is scaled to make the problem tractable. Specifically, we consider a sequence of systems indexed by $n \in \mathbb{N}$ in which the population size N_n satisfies $N_n = \lceil \alpha n \rceil$, where $\alpha > 0$ is a fixed parameter. A schedule in the nth system is a nondecreasing sequence $\{T_{i,n}, i = 1, \ldots, N_n\}$ taking values in $[0, H]$ with the corresponding scheduling function defined as $E_n(t) = \sum_{i=1}^{N_n} \mathbf{1}_{\{T_{i,n} \leq t\}}$, and $A_n(t) = \sum_{i=1}^{E_n(t)} \xi_i = \Xi \circ E_n(t)$. The collection of all schedules for the nth system is denoted by $\mathcal{T}_n$. Let $S_n(t) := S(nt) = \max\{m | \sum_{i=1}^{m} \nu_i \leq nt\} = \max\{m | \sum_{i=1}^{m} \nu_{i,n} \leq t\}$, where $\nu_{i,n} = n^{-1}\nu_i$. In parallel to (2), the resulting number-in-system process for the nth system may be expressed as

$$Q_n = A_n - S_n \circ B_n, \tag{10}$$

with $B_n(\cdot) = \int_0^{\cdot} \mathbf{1}_{\{Q_n(s) > 0\}} ds$.

Under the large-population scaling, population is assumed to grow linearly with n. One may interpret the scaling of the various processes a couple of different ways. Under one interpretation, time is scaled by n, the scheduling time horizon becomes $[0, nH]$, and service times are of order $\mathcal{O}(1)$. A second interpretation is that the scheduling time horizon remains as $[0, H]$, but the service times are scaled by $1/n$. That is, when the population grows, the server speeds up at a rate that is proportional to the population size. Although these two interpretations are mathematically equivalent, we provide intuition throughput the paper that is consistent with the second interpretation.

Paralleling (3)–(5), we have that the makespan, the overage time, and the departure time of the last arrival are respectively defined as

$$W_n = \int_0^\infty (Q_n(s) - 1)^+ ds, \tag{11}$$

$$O_n = (\tau_n - H)^+, \tag{12}$$

and

$$\tau_n = \inf\{t : S_n \circ B_n(t) \geq \Xi(N_n)\}. \tag{13}$$

Paralleling (6) and (7), the *large population optimal-schedule problem (LPOP)* defined by considering the cost

$$J_n(\{T_{i,n}\}) = c_{w,n}\mathbb{E}[W_n] + c_{o,n}\mathbb{E}[O_n], \qquad \{T_{i,n}\} \in \mathcal{T}_n, \tag{14}$$

where $c_{w,n}$ and $c_{o,n}$ are appropriately scaled constants and the value

$$V_n = \inf_{\{T_{i,n}\}\in\mathcal{T}_n} J_n(\{T_{i,n}\}). \tag{15}$$

Similarly, for the complete information case, we have

$$J_n^{\mathrm{CI}}(\{\Theta_{i,n}\}) = c_{w,n}\mathbb{E}[W_n] + c_{o,n}\mathbb{E}[O_n], \qquad \{\Theta_{i,n}\} \in \mathcal{T}^{\mathrm{CI}}, \tag{16}$$

and

$$V_n^{\mathrm{CI}} = \inf_{\{\Theta_{i,n}\}\in\mathcal{T}_n^{\mathrm{CI}}} J_n^{\mathrm{CI}}(\{\Theta_{i,n}\}). \tag{17}$$

Finally, let

$$\gamma_n = V_n - V_n^{\mathrm{CI}}.$$

To see what is the appropriate scaling for the cost coefficients of $J_n(\cdot)$, note that the leading (first-order) term in the expression for the number of jobs that the server can handle in the interval $[0, H]$ is $n\mu H$ (recall that we assume that in the nth system the server works at a rate $n\mu$). Similarly, the leading term in the total number of jobs that arrive into the system in $[0, H]$ is $p\alpha n$. To capture the case in which the server incurs a nonnegligible overage cost, we assume that the system is overloaded. That is, we assume that $p\alpha n > n\mu H$, or equivalently,

$$p\alpha > \mu H. \tag{18}$$

Thus, regardless of the schedule, at time H, the number of jobs present in the system is of order n. Because the service rate is also of order n, it takes a constant (order one) time to handle these jobs. That is, the overage time is $\mathcal{O}(1)$. Along the same lines, notice that the number of arriving jobs is of order n, and their individual waiting time is of order one. Thus, the total waiting time (makespan) is $\mathcal{O}(n)$. This suggests that, to get a meaningful cost function $J_n(\cdot)$, the cost parameter $c_{w,n}$ should be scaled by n^{-1}, and the cost parameter $c_{o,n}$ should remain a constant. Thus, we assume for the rest of the paper that

$$c_{w,n} = n^{-1}c_w, \quad c_{o,n} = c_o. \tag{19}$$

We study this asymptotic problem under two scalings. The first is a fluid scaling (see Section 3) in which only first-order deterministic effects are accounted for. The second is a diffusion scaling (see Section 4) under which a refinement of the fluid solution is considered to account for stochastic second-order terms.

2.4. Main Results

Our paper focuses on solving the LPOP asymptotically under the large population limiting regime. Our first-order analysis uses fluid scaling and captures the deterministic elements of the system, not accounting for

stochasticity. This type of analysis allows us to identify a simple near-optimal scheduling rule and gives us useful insights about the original finite-population problem. Our second-order analysis incorporates the stochastic elements back into the model and offers a solution that is asymptotically optimal under diffusion scaling for large time horizon H. In both scaling regimes, we also identify the SG defined as the appropriately scaled difference between our proposed solution and the solution under complete information.

2.4.1. Fluid Scale. Assume that the cost parameters are of the form postulated in (19), and recall that this form was selected in such a way that the two additive components of $J_n(\cdot)$ in (14) are both of order $\mathcal{O}(1)$. In the fluid scaling, the stochastic variation is "washed out," ensuring that the stochastic optimization in (14) is approximated by a variational problem in the large population asymptotic. This allows us to focus on the $\mathcal{O}(1)$ terms in the optimization.

Consider

$$E_n^f(t) = \begin{cases} 1 + \left\lfloor \frac{n\mu t}{p} \right\rfloor & t < H, \\ N_n & t = H, \end{cases} \tag{20}$$

and its corresponding schedule

$$T_{i,n}^f = \min\left\{ \frac{p}{n\mu}(i-1), H \right\}, \quad i = 1, \ldots, N_n, \quad n \in \mathbb{N}. \tag{21}$$

Then, we establish that $\{T_{i,n}^f\}$ is asymptotically optimal in the fluid scale. Specifically, we show that

$$\lim_{n\to\infty} J_n\left(\{T_{i,n}^f\}\right) = \lim_{n\to\infty} V_n =: \bar{V}.$$

The schedule $\{T_{i,n}^f\}$ satisfies the following properties:

- The appointment times up to time H are at intervals of equal duration of $\frac{p}{n\mu}$ time units.
- All patients who do not get appointments before time H are scheduled to arrive at that time.
- The arrival rate during $[0, H)$ is equal to the service rate of $n\mu$. Thus, it is asymptotically optimal to operate the system at a critically loaded heavy-traffic regime. Note that heavy traffic is obtained here as a *result* of optimality and not as an *assumption.*
- The critically loaded regime implies that, in the fluid scale, the server idle time is negligible compared with the overage time when all customers have been served and that no customers wait during $[0, H)$.

In terms of SG, our results show that in the fluid scaling this gap vanishes in the limit. Specifically, we show that

$$\lim_{n\to\infty} \gamma_n = \lim_{n\to\infty} (V_n - V_n^{\mathrm{CI}}) = 0. \tag{22}$$

This result implies that, at the fluid scaling, knowing whether customers will show or not and their actual service time is only marginally beneficial to the system manager. In particular, knowing these quantities, on average, is sufficient at the fluid level.

2.4.2. Diffusion Scale. In diffusion scaling, we are interested in fluctuations about the fluid solution that are of order $\mathcal{O}(1/\sqrt{n})$. Specifically, we focus on the centered and scaled cost function

$$\hat{J}_n\left(\{T_{i,n}\}\right) = \sqrt{n}\left(J_n\left(\{T_{i,n}\}\right) - \bar{V}\right), \tag{23}$$

and its corresponding centered and scaled value function

$$\hat{V}_n = \inf_{\{T_{i,n}\} \in \mathcal{T}_n} \hat{J}_n\left(\{T_{i,n}\}\right).$$

Our diffusion scale results are for a large time horizon. To state these results, we add the time horizon H as a subscript to all relevant quantities. Consider the following scheduling function

$$E_{n,H}^d(t) = \begin{cases} 1 + \left\lfloor \frac{nt}{p}\left(\mu + \frac{\beta^*}{\sqrt{n}}\right) \right\rfloor & t < H, \\ N_n & t = H, \end{cases} \tag{24}$$

where $\beta^* = -\sqrt{\frac{c_w\left(p(1-p)+\mu^3\sigma^2\right)}{2\left(c_w(2\bar{\tau}-H)+c_o/\mu\right)}}$ with

$$\bar{\tau} = \frac{p\alpha}{\mu}. \tag{25}$$

The corresponding schedule is

$$T^d_{i,n,H} = \min\left\{\frac{p}{n\left(\mu+\beta^*/\sqrt{n}\right)}(i-1), H\right\}, \quad i = 1,\dots,N_n, \quad n \in \mathbb{N}. \tag{26}$$

Then, we establish that $\{T^d_{i,n,H}\}$ is asymptotically optimal in the diffusion scale. More precisely, under the assumption that the service times ν_i possess a $3+\varepsilon$ moment, we show that

$$\lim_{H\to\infty}\limsup_{n\to\infty}\frac{1}{H}\hat{J}_{n,H}\left(\{T^d_{i,n,H}\}\right) = \lim_{H\to\infty}\liminf_{n\to\infty}\frac{1}{H}\hat{V}_{n,H}.$$

The schedule $\{T^d_{i,n,H}\}$ satisfies the following properties:

- The appointment times up to time H are at intervals of equal duration of $\frac{p}{n\left(\mu+\beta^*/\sqrt{n}\right)}$ time units. This interval duration is slightly *longer* than that of the fluid schedule and deviates from it by a term of the order of $\mathbb{O}(1/\sqrt{n})$.
- For general negative β, if the appointment times are at integer multiples of $\frac{p}{n(\mu+\beta/\sqrt{n})}$ then, in the scaling limit, the queue length process converges to an RBM on the half line with a constant negative drift β.
- The constant β^* is obtained by considering this RBM and minimizing a cost with two additive terms: one proportional to $|\beta|$ and representing server idleness cost and the other inversely proportional to $|\beta|$ and representing the holding cost.

In terms of SG, it turns out that, although in the fluid scale the SG is negligible, it is strictly positive in the diffusion scale. Specifically, let $\hat{\gamma}_{n,H} = \sqrt{n}\left(V_{n,H} - V^{CI}_{n,H}\right)$. Then, we show that

$$\lim_{H\to\infty}\liminf_{n\to\infty}\frac{1}{H}\hat{\gamma}_{n,H} = \lim_{H\to\infty}\liminf_{n\to\infty}\frac{1}{H}\hat{V}_{n,H} =: \hat{V}^* > 0,$$

where $\hat{V}^*$ is a constant.

2.5. Exact Analysis of the CI Problem

In the CI problem, all stochastic data are known to the decision maker. It, thus, can be treated, for each realization of these data, as a deterministic allocation problem. The number of show-ups is given by $\Xi(N)$, and the total amount of work associated with them is $\tau^* := V(\Xi(N)) = \inf\{t : S(t) \geq \Xi(N)\}$. Hence, τ^* is a lower bound on τ for any CI schedule. It is easy to see that there is no gain by allowing the system to be empty (thus, the server idle) for some time prior to the time of completion of all jobs, τ. Indeed, if there is any interval $[a,b)$ on which the system is empty and there is still a job (or more) that will show up at time b, then advancing all jobs scheduled at b or later by $b-a$ units of time (so that the job originally scheduled at b arrives at a, etc.) does not affect the waiting time of any of the jobs arriving at times $t \geq a$, and it can only decrease the overage time. Thus, it suffices to consider only allocations for which $B(t) = t$ (equivalently, $Q(t) \geq 1$) for all $t < \tau$. Moreover, for all such allocations, clearly $\tau = \tau^*$. As a result, (2) gives

$$Q = A - S \text{ on } [0,\tau^*]. \tag{27}$$

The sample path of $A(t)$, $t \in [0,\infty)$, can be any member of $\bar{D}[0,\infty)$ that is integer valued and satisfies

$$A(t) = \Xi(N) \text{ for all } t \geq H. \tag{28}$$

Now, in view of (27), the requirement $Q(t) \geq 1$ alluded to implies

$$A(t) \geq 1 + S(t) \text{ for all } t < \tau^*. \tag{29}$$

Among all paths satisfying (28) and (29), there is one that is pointwise minimal, namely

$$A^*(t) = \begin{cases} 1+S(t), & t < H, \\ \Xi(N), & t \geq H. \end{cases}$$

It corresponds to allocating the jobs in such a way that there are no waiting customers during $[0, H)$, and in the event that $\tau^* \geq H$, the remaining customers that are $\Xi(N) - (1 + S(H-))$ in number are all scheduled at the very last moment, H. By (3) and (27), $W = \int_0^{\tau^*} (A(s) - S(s) - 1)^+ ds$. Thus, W is monotone in A in the following sense: if $A(t) \geq \tilde{A}(t)$ for all $t \in [0, \tau^*]$, then $W \geq \tilde{W}^*$. We conclude that A^* minimizes W, and because we have already mentioned that it minimizes τ, it also minimizes their weighted sum in pathwise sense. Consequently, this CI schedule minimizes the cost (8).

Finally, we can also compute this cost. On the event $\tau^* < H$, $W^* = 0$, when $\tau^* \geq H$,

$$W^* = \int_H^{\tau^*} (\Xi(N) - S(s) - 1)^+ ds = \int_H^{\tau^*} (\Xi(N) - S(s) - 1) ds.$$

We, thus, obtain

$$\begin{aligned} V^{\mathrm{CI}} &= c_w \mathbb{E}[W^*] + c_o \mathbb{E}[(\tau^* - H)^+] \\ &= c_w \mathbb{E}\left[\int_H^{H \vee \tau^*} (\Xi(N) - S(s) - 1) ds \right] + c_o \mathbb{E}[(\tau^* - H)^+]. \end{aligned} \tag{30}$$

3. Large Population Asymptotics: Fluid Scale

This section studies our appointment-scheduling model in fluid scale. By exploiting the stochastic regularity that emerges in this scaling limit, we identify a deterministic, first-order approximation to the FPOP that governs the limit behavior. We find an optimal solution to this limiting problem and show that the cost associated with this solution is asymptotically achievable in the fluid-scale limit.

We start by stating and solving a formal fluid problem. Later, we show that the optimal value of this fluid problem constitutes a lower bound on the fluid-scaled FPOP. Subsequently, we show that this value also constitutes an upper bound on the fluid-scaled FPOP shown by identifying a sequence of simple policies for the FPOP that asymptotically achieve this value. Thus, we establish that this sequence of policies is asymptotically optimal in the fluid scale.

3.1. Fluid Model

Let $\{E_n\}$ be an arbitrary sequence of scheduling functions. Following (1), the fluid-scaled cumulative arrival process is defined as

$$\bar{A}_n = \frac{1}{n} \Xi \circ n \bar{E}_n,$$

where $\bar{E}_n = \frac{1}{n} E_n$ is the fluid-scaled schedule. The functional law of large number (FLLN) implies that, as $n \to \infty$,

$$\frac{1}{n} \Xi(\lfloor ne \rfloor) \Rightarrow pe, \tag{31}$$

where $e : \mathbb{R}_+ \to \mathbb{R}_+$ is the identity map; note that here the convergence is u.o.c. even though the prelimit processes are assumed to exist in the Skorokhod J_1 topology because the limit process is continuous. This is the case in the remainder of the discussion unless noted otherwise. Throughout this section, we use the notation ε_n for a generic sequence of stochastic processes that converge to the zero process in probability as $n \to \infty$ as well as for a generic sequence of RVs that converges to zero in probability. It follows from (31) that

$$\bar{A}_n = n^{-1} \Xi(n \bar{E}_n) = p \bar{E}_n + \varepsilon_n. \tag{32}$$

Also, the FLLN for renewal processes (Chen and Yao [10, chapter 5]) implies that, as $n \to \infty$,

$$\bar{S}_n := \frac{1}{n} S_n \Rightarrow \mu e. \tag{33}$$

As a result, $\bar{S}_n(B_n(t)) = \mu B_n(t) + \varepsilon_n$. In view of these identities, the fluid-scaled queue length process $\bar{Q}_n := n^{-1} Q_n$ is given by

$$\bar{Q}_n = \bar{A}_n - \bar{S}_n \circ B_n = p \bar{E}_n - \mu e + \mu(e - B_n) + \varepsilon_n.$$

The idleness process $I_n := e - B_n$ is nondecreasing and is flat on excursions of $\bar{Q}_n$ away from zero, and thus, $(\bar{Q}_n, \mu I_n)$ forms a solution to the Skorokhod problem with data $p\bar{E}_n - \mu e + \varepsilon_n$. That is,

$$\left(\bar{Q}_n, \mu I_n\right) = \left(\Gamma_1(p\bar{E}_n - \mu e + \varepsilon_n), \Gamma_2\left(p\bar{E}_n - \mu e + \varepsilon_n\right)\right), \tag{34}$$

with $\Gamma_1(x) = x + \Gamma_2(x)$ and $\Gamma_2(x)(t) = \sup_{0\le s\le t}(-x(s))^+$.

Recall the fluid-scaled makespan RV $\bar{W}_n := n^{-1}W_n = n^{-1}\int_0^\infty (Q_n(t)-1)^+ dt$, and the overage time $O_n = (\tau_n - H)^+$, where $\tau_n = \inf\{t>0 : S_n(B_n(t)) \ge \Xi(N_n)\}$. From (32)–(34), by formally removing the error terms, we derive a fluid model as follows. Let $\mathcal{L} = \{\lambda \in \mathcal{D}^+[0,\infty) : \lambda(t) = \alpha \text{ for } t \ge H\}$. Given $\lambda \in \mathcal{L}$, let $q = \Gamma_1(p\lambda - \mu e) = p\lambda - \mu e + \eta$, where $\eta(t) := \Gamma_2(p\lambda - \mu e)(t)$ is the correction term (or Skorokhod term). These are fluid models for the arrival process and queue length, respectively. Let also $\bar{B}(t) := t - \mu^{-1}\eta(t)$ stand for cumulative busyness, and let $\tau = \inf\{t > 0 : \mu\bar{B}(t) \ge p\alpha\}$, $\bar{W} = \int_0^\infty q(t)dt$, and $\bar{O} = (\tau - H)^+$ denote the fluid models for the termination time, wait, and overage time, respectively. An FOP is formulated by letting

$$\bar{J}(\lambda) = c_w\bar{W} + c_o\bar{O}, \tag{35}$$

and

$$\bar{V} := \inf_{\lambda\in\mathcal{L}} \bar{J}(\lambda). \tag{36}$$

In Section 3.2, we show that there exists a $\lambda^* \in \mathcal{L}$ that attains the minimum in (36). In Section 3.3, we show that the FOP value $\bar{V}$ is an asymptotic lower bound on the fluid scale cost $J_n(\{T_i\})$ under an arbitrary sequence of schedules. In Section 3.4, we construct a bespoke sequence of finite population schedules that asymptotically *achieves* $\bar{V}$, thus proving asymptotic optimality in fluid scale.

3.2. Fluid Optimal Schedule

Under our overload assumption (18), it is straightforward to see that our fluid model satisfies, for any $\lambda \in \mathcal{L}$, $0 < \tau - H = \mu^{-1}q(H)$. An optimal control should minimize the trade-off between the fluid overage time and fluid makespan. The main result of this section in Proposition 1 identifies such an optimal control.

Proposition 1. *The optimal value of the FOP is given by*

$$\bar{V} = c_w\frac{(p\alpha - \mu H)^2}{2\mu} + c_o\frac{p\alpha - \mu H}{\mu},$$

and a fluid optimal control $\lambda^* \in \mathcal{L}$ that obtains this value is

$$\lambda^*(t) = \begin{cases} p^{-1}\mu t, & t \in [0, H) \\ \alpha, & t \in [H, \infty). \end{cases} \tag{37}$$

Proof. Let $\lambda \in \mathcal{L}$. The identity $q = \Gamma_1(p\lambda - \mu e)$ and the fact that $\lambda(H) = \alpha$ imply $q(H) \ge p\lambda(H) - \mu H = p\alpha - \mu H$. As a consequence, for $t \in [H, \tau)$,

$$q(t) \ge q(H) - (t-H)\mu \ge p\alpha - \mu H - (t-H)\mu. \tag{38}$$

Consequently,

$$\tau - H \ge \mu^{-1}(p\alpha - \mu H). \tag{39}$$

The lower bounds (38) and (39) translate easily into lower bounds on $\bar{W}$ and $\bar{O}$. Namely,

$$\bar{W} = \int_0^\infty q(t)dt \ge \int_H^\tau q(t)dt \ge \int_H^{\mu^{-1}p\alpha} [p\alpha - \mu H - (t-H)\mu]dt = \frac{(p\alpha - \mu H)^2}{2\mu}, \tag{40}$$

where the second inequality uses both (38) and (39). Moreover, because, by (39), $\tau - H > 0$, we have

$$\bar{O} = \tau - H \ge \mu^{-1}(p\alpha - \mu H). \tag{41}$$

The lower bounds (40) and (41) on $\bar{W}$ and $\bar{O}$, respectively, are valid for arbitrary $\lambda \in \mathcal{L}$. Moreover, by direct calculation, they are both achieved by λ^*. This completes the proof. Q.E.D.

Some remarks on Proposition 1 are warranted. First, and most importantly, under (18), the optimal schedule ensures that the queue length is zero in $[0,H)$ (note that, by the very definition of τ, q and positive in $[H,\tau)$ and zero on $[\tau,\infty)$). This is an intuitively satisfying result in the sense that, given the system operator's goal of minimizing a positive combination of makespan and overage time, it would make most sense to fully utilize the available capacity but *not* overload the system. Thus, the optimal schedule matches the arrival "rate" with the effective service rate $p^{-1}\mu$ in the interval $[0,H)$ and schedules the remainder (of $\alpha - p^{-1}\mu H$ fluid units) at H. In other words, the heavy-traffic condition, valid throughout the interval $[0,H)$, emerges as a *consequence* of optimization. This result is in stark contrast to most queueing control problems in which the heavy-traffic condition is assumed at the outset. Notice too that the fluid-optimal solution parallels the sample pathwise solution in the complete information problem. In the absence of stochastic variation in the arrivals and service, it is clearly possible to optimally arrange the appointments such that there is no waiting for any of the jobs that do turn up and schedule the remainder at the end of the horizon.

3.3. Lower Bound on the Fluid Scale Cost

Recall that the fluid-scaled makespan $\bar{W}_n$ is defined as $\bar{W}_n = n^{-1}\int_0^\infty (Q_n(s)-1)^+ds$. Now, fix $0<K<\infty$ (to be determined later) and note that

$$\bar{W}_n \geq n^{-1}\int_0^K (Q_n(s)-1)^+ds \geq \int_0^K \bar{Q}_n(s)ds - Kn^{-1}. \tag{42}$$

By (34), there exists a sequence of processes $\{\varepsilon_n\}$ that converges to the zero process in probability as $n\to\infty$, such that $\bar{Q}_n = \Gamma_1(p\bar{E}_n - \mu e + \varepsilon_n)$. By the Lipschitz continuity of the Skorokhod map, it follows that $\bar{Q}_n \geq \Gamma_1(p\bar{E}_n - \mu e) + \varepsilon_n$. Substituting this into (42), we observe that

$$\bar{W}_n \geq \int_0^K \Gamma_1(p\bar{E}_n - \mu e)(s)ds + \varepsilon_n - Kn^{-1},$$

clearly implying that

$$\bar{W}_n \geq \inf_{\lambda\in\mathcal{D}^+[0,\infty)} \int_0^K \Gamma_1(p\lambda - \mu e)(t)dt + \varepsilon_n - Kn^{-1}. \tag{43}$$

From the definition of τ_n (13), it follows that $\tau_n \geq \inf\{t : \mu B_n(t) + \varepsilon_n \geq \alpha p\}$, where we have used the fact that $\bar{S}_n(B_n(t)) = \mu B_n(t) + \varepsilon_n$ and $\bar{A}_n(H) = \alpha p + \varepsilon_n'$; note that we use ε_n to represent the difference between the two mean-zero error sequences. Because we have $t \geq B_n(t)$, it follows that $\tau_n \geq \inf\{t : \mu t + \varepsilon_n \geq \alpha p\}$, implying that $\tau_n \geq \alpha p\mu^{-1} + \varepsilon_n = \bar{\tau} + \varepsilon_n$. It follows that

$$(\tau_n - H)^+ \geq (\bar{\tau} - H + \varepsilon_n)^+. \tag{44}$$

Now, consider

$$\tilde{V} := c_w \inf_{\lambda\in\mathcal{L}} \left\{ \int_0^\infty \Gamma_1(p\lambda - \mu e)(t)dt \right\} + c_o(\bar{\tau} - H)^+.$$

As the next lemma shows, $\tilde{V}$ equals the FOP value and can be achieved by the optimal schedule in Proposition 1.

Lemma 1.

1. $\bar{V} = \tilde{V}$.
2. *The upper limit* ∞ *in the integral can be replaced by any* K *sufficiently large.*
3. *The minimum is attained by* λ^* *defined in* (37).

Proof. Consider $\tilde{V}$ first and recall that, for a fixed λ, $\tau = \inf\{t>0 : \mu(t - I(t)) = \alpha p\}$. If $\tau > \bar{\tau}$, it automatically follows that $I(\bar{\tau}) > 0$. This implies that the makespan cost $c_w\int_0^\infty (p\lambda(t) - \mu(t - I(t)))dt$ is not optimal. To see this, note that the makespan cost can be lower bounded by choosing λ' such that $\lambda'(t) = \mu t$ for all $t\in[0,\tau]$, and $\tau = \bar{\tau}$ in this case. Thus, any optimal solution should be such that $I(t) = 0$ up to $\tau = \bar{\tau}$, in which case we minimize

$$\int_0^{\bar{\tau}} q(t)dt = \int_0^{\bar{\tau}} p\lambda(t)dt - \int_0^{\bar{\tau}} \mu t dt,$$

where only the first term on the right-hand side (RHS) is controlled. The minimizing schedule λ that satisfies the constraint that $\lambda(t) = \alpha, t \geq H$ is

$$\lambda(t) = \begin{cases} p^{-1}\mu t, & t \in [0, H) \\ \alpha, & t \in [H, \infty). \end{cases} \quad \text{Q.E.D.}$$

Now we show that Lemma 1, together with (43) and (44), implies that the FOP value lower bounds the fluid-scaled cost.

Theorem 1. *The fluid-scaled cost of an arbitrary sequence of schedules, $\{T_i\}$, is asymptotically lower bounded by the fluid optimal value $\bar{V}$. That is,* $\liminf_{n\to\infty} \bar{J}_n(\{T_i\}) \geq \bar{V}$.

Proof. Let $j_n := c_w \bar{W}_n + c_o O_n$ represent the random cost incurred by following schedule $\{T_i\}$. Because the constant K was arbitrary, we can set it to be greater than $\bar{\tau}$. Lemma 1, together with (43) and (44), implies that

$$j_n \geq (\bar{V} + \varepsilon_n - Kn^{-1}) \vee 0.$$

Observe that $j_n \Rightarrow \bar{V}$ as $n \to \infty$. Because $j_n \geq 0$, Fatou's lemma (Ethier and Kurtz [13]) implies that

$$\liminf_{n\to\infty} \mathbb{E}[j_n] \geq \mathbb{E}[\liminf_{n\to\infty} j_n] = \bar{V}. \quad \text{Q.E.D.}$$

3.4. Upper Bound on the Fluid Scale Cost

We now construct a sequence of scheduling policies whose fluid-scaled cost is asymptotically upper-bounded by the FOP value $\bar{V}$. Given the lower bound result in the previous subsection, this sequence is, thus, asymptotically optimal in the fluid limit.

Recall the fluid-optimal schedule

$$\lambda^*(t) = \begin{cases} \mu p^{-1} t, & t \in [0, H) \\ \alpha, & t \in [H, \infty). \end{cases}$$

Consider the following sequence of scheduling functions indexed by n

$$E_n^f(t) := \begin{cases} 1 + \left\lfloor \dfrac{n\mu t}{p} \right\rfloor, & t < H, \\ N_n, & t = H, \end{cases}$$

and its corresponding schedule

$$T_{i,n}^f = \min\left\{ \frac{p}{n\mu}(i-1), H \right\}, \quad i = 1, \ldots, N_n, \quad n \in \mathbb{N}. \tag{45}$$

The scheduling function E_n^f is interpreted as follows: for each n, customers are scheduled to arrive one at a time at uniformly spaced intervals of length $p(n\mu)^{-1}$ up to time H with the leftover $N_n - (1 + \lfloor n\mu H/p \rfloor)$ customers who are scheduled to arrive at time H.

The main result of this section establishes the fact that the expected fluid-scaled cost $J_n(\{T_{i,n}^f\})$ converges to the fluid-optimal value as well.

Theorem 2. *Suppose that the schedule is $\{T_{i,n}^f\}$ for each n. Then*

$$\limsup_{n\to\infty} J_n(\{T_{i,n}^f\}) \leq \bar{V}. \tag{46}$$

Lemma 2. *The finite population schedule satisfies $\bar{E}_n^f \to \lambda^*$ uniformly on compacts as $n \to \infty$.*

Now, let (Q_n, I_n) represent the queue length and the idleness processes when appointments are scheduled per E_n^f. FLLNs for the arrival and service processes and Lemma 2 together imply the following result:

Lemma 3. *The following hold with respect to the schedule $T_{i,n}^f$ of (45):*

i. *The fluid-scaled queue length and idleness processes satisfy an FLLN: $(\bar{Q}_n, I_n) \Rightarrow (q^*, \iota^*)$ as $n \to \infty$, where*

$$q^*(t) = \begin{cases} 0, & t \in [0, H) \\ \alpha p - \mu t, & t \in [H, \bar{\tau}] \\ 0, & t \in (\bar{\tau}, \infty), \end{cases}$$

and

$$\iota^*(t) = \begin{cases} 0, & t \in [0, \bar{\tau}] \\ t - \bar{\tau}, & t \in (\bar{\tau}, \infty), \end{cases}$$

where $\bar{\tau}$ *is given in* (25).

ii. *The fluid-scaled makespan and overage time RVs satisfy* $(\bar{W}_n, O_n) \Rightarrow (\bar{W}^*, \bar{O}^*)$ *as* $n \to \infty$, *where* $\bar{W}^* = \frac{(p\alpha - \mu H)^2}{2\mu}$ *and* $\bar{O}^* = (\bar{\tau} - H)$.

Proof of Theorem 2. Let $J_n^R(\{T_{i,n}^f\}) = c_w \bar{W}_n + c_o O_n$ be the random cost. We start by noting that the convergence result in Lemma 3 implies that the random cost $J_n^R(\{T_{i,n}^f\})$ weakly converges to $\bar{J}(\lambda^*) = \bar{V}$ as $n \to \infty$. This implies that $J_n(\{T_{i,n}^f\}) = \mathbb{E}[J_n^R(\{T_{i,n}^f\})]$ will converge to $\bar{V}$ provided that $J_n^R(\{T_{i,n}^f\})$ is uniformly integrable. The remainder of this proof is dedicated to proving this claim.

We prove that $J_n^R(\{T_{i,n}^f\})$ is uniformly integrable by showing that $\mathbb{E}\,|J_n^R(\{T_{i,n}^f\})|^2 \le C < \infty$ for all $n \in \mathbb{N}$. Consider the sequence $\{O_n\}$ first. Note that it suffices to consider the case in which $\tau_n > H$. The number of jobs waiting in the queue at the end of the arrival horizon H is $N_n - D_n(H) > 0$, where $D_n(H)$ is the number of departures in $[0, H]$. Because there are no more arrivals after time H, it can be seen that

$$\tau_n - H \le \sum_{i=D_n(H)+1}^{N_n} v_{i,n} \le \sum_{i=1}^{N_n} v_{i,n}. \tag{47}$$

Now, let $\Upsilon(m) := \sum_{i=1}^{m} v_i$ and $\Upsilon_n(m) := \sum_{i=1}^{m} v_{i,n}$. Then Minkowski's inequality implies that $(\mathbb{E}|\Upsilon(N_n)|^2)^{1/2} \le N_n(\mathbb{E}\cdot|\Upsilon(1)|^2)^{1/2}$. Therefore, we obtain

$$\left(E\,|\Upsilon_n(N_n)|^2\right)^{1/2} \le \frac{N_n}{n}\left(E\,|\Upsilon(1)|^2\right)^{1/2} \le \alpha\left(\mathbb{E}\,|\Upsilon(1)|^2\right)^{1/2}, \tag{48}$$

where the last inequality follows from the fact that $N_n/n \le \alpha$ by definition. Equation (47) and this bound imply that

$$\mathbb{E}|(\tau_n - H)^+|^2 \le \mathbb{E}|\tau_n - H|^2 \le \alpha \mathbb{E}\,|\Upsilon(1)|^2 < \infty,$$

where the finiteness of the second moment is by assumption. Because the bound is independent of n, it follows that $O_n = (\tau_n - H)^+$ are uniformly integrable.

Now consider the fluid-scaled makespan. Using the fact that the queue drains out and remains empty after τ_n, it follows that

$$\bar{W}_n = n^{-1}\int_0^H (Q_n(t) - 1)^+ dt + n^{-1}\int_H^{H \vee \tau_n} (Q_n(t) - 1)^+ dt.$$

Note that the first term on the right-hand side of the inequality is bounded above by $n^{-1}N_n H \le \alpha H$. Thus, it suffices to consider the second term when $\tau_n > H$. As there are $N_n - D_n(H)$ jobs waiting for service at the end of the horizon, it follows that

$$\begin{aligned} n^{-1}\int_H^{\tau_n} Q_n(t)dt &\le n^{-1}\left\{(N_n - D_n(H))v_{D_n(H)+1,n} + (N_n - D_n(H) - 1)v_{D_n(H)+2,n} + \cdots + v_{N_n}^n\right\} \\ &= \frac{1}{n}\sum_{i=1}^{N_n - D_n(H)} (N_n - D_n(H) + 1 - i)v_{D_n(H)+i,n} \\ &= \frac{N_n - D_n(H)}{n}\sum_{i=1}^{N_n - D_n(H)} v_{D_n(H)+i,n} - \frac{1}{n}\sum_{i=1}^{N_n - D_n(H)} (i-1)v_{D_n(H)+i,n} \\ &\le \frac{N_n}{n}\sum_{i=1}^{N_n} v_{i,n}, \end{aligned}$$

where the last inequality follows from the fact that $n^{-1}\sum_{i=1}^{N_n - D_n(H)}(i-1)v_{D_n(H)+i,n} \ge 0$ and $D_n(H) \ge 0$ for all $n \ge 1$. Using the bound in (48), we have

$$n^{-1}\left(\mathbb{E}\left(\int_H^{\tau_n} Q_n(t)dt\right)^2\right)^{1/2} \le \alpha\left(\mathbb{E}|\Upsilon(1)|^2\right)^{1/2}.$$

Thus, it follows that $\mathbb{E}|\bar{W}_n|^2$ is uniformly bounded for all $n \in \mathbb{N}$, implying that $\{\bar{W}_n\}$ is uniformly integrable. Finally, because $\{J_n^R(\{T_{i,n}^f\}), n \geq 1\}$ is a sequence of RVs that are each linear combinations of uniformly integrable RVs, it is uniformly integrable as well. Q.E.D.

Thus, Theorem 2 shows that the family of finite population schedules $\{E_n^f\}$ is asymptotically optimal in the sense that the FOP value can be achieved in the large population limit.

Proof of Lemma 2. Fix $t \in [0, H)$ and $n \geq 1$. By definition, it follows that

$$\left|\frac{E_n^f(t)}{n} - \lambda^*(t)\right| \leq \frac{1}{n} + \frac{1}{n}\left(\frac{n\mu t}{p} - \left\lfloor \frac{n\mu t}{p} \right\rfloor\right) \leq \frac{2}{n}.$$

On the other hand, fix $t \geq H$ and observe that

$$\left|\frac{E_n^f(t)}{n} - \alpha\right| \leq \left|\frac{N_n}{n} - \alpha\right| \leq \frac{2}{n}. \tag{49}$$

These two bounds are independent of t, proving the lemma. Q.E.D.

Proof of Lemma 3. Part (i) follows by Lemma 2 and (34), upon applying the continuous mapping theorem, and by noting that indeed $q^* = \Gamma_1(p\lambda^* - \mu e)$ and $\iota^* = \Gamma_2(p\lambda^* - \mu e)/\mu$.

For part (ii), we first establish that $O_n \Rightarrow \bar{O}^*$ or, equivalently, that $\tau_n \Rightarrow \bar{\tau} = p\alpha/\mu$. We first argue that τ_n is equal to the total work that enters in the entire horizon plus the idleness by time τ_n. That is,

$$\tau_n = \sum_{i=1}^{\Xi(N_n)} v_{i,n} + I_n(\tau_n). \tag{50}$$

To see this, note that by (13) τ_n satisfies

$$\Xi(N_n) = S_n \circ B_n(\tau_n) = S_n \circ (\tau_n - I_n(\tau_n)),$$

and (50) follows by taking the inverse of S_n on both sides. Note that by the FLLN the first term in the expression for τ_n in (50) converges to $p\alpha/\mu$. It is left to show that $I_n(\tau_n) \Rightarrow 0$. To show the latter, note that if $\tau_n > H$ then the server is nonidling on the interval $[H, \tau_n)$. Thus, $I_n(\tau_n) \leq I_n(H)$ regardless of whether $\tau_n > H$ or not. But by part (i) of this lemma we have that $I_n(H) \Rightarrow \iota^*(H) = 0$.

Second, normalizing by n in (11) and using the fact that $Q_n(t) = 0$ for $t \geq \tau_n$, we have

$$\bar{W}_n = \int_0^{\tau_n} (\bar{Q}_n(s) - n^{-1})^+ ds.$$

We have already shown that $\tau_n \Rightarrow \bar{\tau}$ and that $\bar{Q}_n \Rightarrow q^*$, and because these two limits are deterministic, joint convergence also holds. Thus, by the continuous mapping theorem, $\bar{W}_n \Rightarrow \int_0^{\bar{\tau}} q^*(s)ds = \int_H^{\bar{\tau}} q^*(s)ds = \frac{(p\alpha - \mu H)^2}{2\mu}$. Q.E.D.

3.5. The Stochasticity Gap at the Fluid Scale

The first-order deterministic FOP is solved by scheduling demand to match the available capacity. In addition, the previous two sections have shown that the fluid-scale cost is bounded by the FOP value and proposed an asymptotically optimal schedule for the LPOP. At the same time, for the CI problem, we identified, in Section 2.5, that a Σ_n-measurable schedule (30) that optimizes this problem; this schedule allocates appointments such that there are no waiting customers during $[0, H)$, and the server is never idle. Clearly, for the LPOP, there is a cost to be paid for scheduling appointments without a priori knowledge of the randomness. The parallels between the CI and FOP optimal schedule and Theorem 1 suggest that there may be a gap between the LPOP value (V_n) and the value of the CI problem (V_n^{CI}). We quantify this stochasticity gap by showing that $\gamma_n := V_n - V_n^{CI} \geq 0$ decreases to zero as $n \to \infty$. Recall from Proposition 1 that $\bar{V} = c_w \frac{(p\alpha - \mu H)^2}{2\mu} + c_o \frac{p\alpha - \mu H}{\mu}$ is the value of the FOP. The following is the main result of this section.

Proposition 2. *The SG in the fluid limit is zero. That is, $\gamma_n \to 0$ as $n \to \infty$.*

The proof of Proposition 2 follows as a consequence of the following lemmas.

Lemma 4. *The optimal overage time in the CI problem satisfies*

$$\mathbb{E}[(\tau_n^* - H)^+] \to (\bar{\tau} - H) = \frac{p\alpha - \mu H}{\mu} \quad \text{as } n \to \infty, \tag{51}$$

where recall from (25) *that* $\bar{\tau} = p\alpha/\mu$.

Now, let $X_n(t) := n^{-1}(\Xi(N_n) - S_n(t) - 1)^+$ and $x(t) = (\alpha p - \mu t)^+$ for all $t \geq 0$. We prove that the expectation of the integral $\int_H^{H \vee \tau_n^*} X_n(t)dt$ converges as $n \to \infty$.

Lemma 5. *The optimal expected makespan of the CI problem satisfies*

$$\mathbb{E}\left[\int_H^{H \vee \tau_n^*} X_n(t)dt\right] \to \int_H^{\bar{\tau}} x(t)dt = \frac{(p\alpha - \mu H)^2}{2\mu} \text{ as } n \to \infty. \tag{52}$$

Proof of Proposition 2. Note that $\gamma_n = V_n - V_n^{CI} = (V_n - \bar{V}) + (\bar{V} - V_n^{CI})$. Then, Lemmas 4 and 5 imply that $\bar{V} - V_n^{CI} \to 0$ as $n \to \infty$. Theorem 2 implies that $\limsup_{n\to\infty}(V_n - \bar{V}) \leq \limsup_{n\to\infty}(\bar{J}_n(\{T_{i,n}^f\}) - \bar{V}) \leq 0$. On the other hand, Theorem 1 implies that $\liminf_{n\to\infty}(V_n - \bar{V}) \geq 0$. Thus, $(V_n - \bar{V}) \to 0$ as $n \to \infty$. Q.E.D.

Proof of Lemma 4. By definition, $\tau_n^* := \Upsilon_n(\Xi(N_n))$, where recall that $\Upsilon_n(m) := \sum_{i=1}^m \nu_{i,n}$. It is straightforward to deduce that $(\tau_n^* - H)^+ \Rightarrow (\bar{\tau} - H)$ as $n \to \infty$ as a consequence of the FLLN. Because $\Xi(N_n) \leq N_n$ and $\Upsilon_n(\cdot) \geq 0$ for all $n \geq 1$, it follows that $(\tau_n^* - H)^+ \leq (\Upsilon_n(N_n) - H)^+$. On the other hand, following (48), we have $\mathbb{E}\,|(\Upsilon_n(N_n) - H)^+|^2 \leq 2\alpha^2\mathbb{E}\,|\Upsilon(1)^2| + 2H^2 < \infty$, where recall that $\Upsilon(m) := \sum_{i=1}^m \nu_i$, implying that $\mathbb{E}\,|(\tau_n^* - H)^+|^2 < \infty$. Therefore, $(\tau_n^* - H)^+$ is uniformly integrable, implying (51). Q.E.D.

Proof of Lemma 5. We first prove that the optimal makespan converges in probability to the limit on the right-hand side. It suffices to assume that $\tau_n^* > H$ because we know from the proof of Lemma 4 that $\tau_n^* \Rightarrow \bar{\tau}$ as $n \to \infty$ and that $\bar{\tau} > H$ (by the overload assumption (18)). Thus, consider

$$\left|\int_H^{\tau_n^*} X_n(t)dt - \int_H^{\bar{\tau}} x(t)dt\right| \leq \left|\int_H^{\tau_n^*} X_n(t)dt - \int_H^{\bar{\tau}} X_n(t)dt\right| + \left|\int_H^{\bar{\tau}} X_n(t)dt - \int_H^{\bar{\tau}} x(t)dt\right|. \tag{53}$$

Consider the first term on the RHS and observe that

$$\begin{aligned}\left|\int_H^{\tau_n^*} X_n(t)dt - \int_H^{\bar{\tau}} X_n(t)dt\right| &= \left|\int_{\tau_n^*}^{\bar{\tau}} X_n(t)dt\mathbf{1}_{\{\tau_n^* \leq \bar{\tau}\}} + \int_{\bar{\tau}}^{\tau_n^*} X_n(t)dt\mathbf{1}_{\{\tau_n^* > \bar{\tau}\}}\right| \\ &\leq \alpha|\tau_n^* - \bar{\tau}|,\end{aligned}$$

where the last inequality follows from the fact that $X_n(t) \leq \alpha$ for all $t \in [0, \infty)$. Therefore,

$$\left|\int_H^{\tau_n^*} X_n(t)dt - \int_H^{\bar{\tau}} X_n(t)dt\right| \Rightarrow 0$$

as $n \to \infty$. Next, consider the second term on the RHS. Using the facts that $X_n(t) \Rightarrow x(t)$ as $n \to \infty$ pointwise and $|X_n(t) - x(t)| \leq \alpha$ for all $t \in [0, \infty)$, the bounded convergence theorem implies that $\int_H^{\bar{\tau}} |X_n(t) - x(t)|dt \Rightarrow 0$ as $n \to \infty$. Thus, it follows that $\int_H^{H \vee \tau_n^*} X_n(t)dt \Rightarrow \int_H^{\bar{\tau}} x(t)dt$ as $n \to \infty$. Finally, observe that

$$\left|\int_H^{H \vee \tau_n^*} X_n(t)dt\right|^2 \leq \alpha^2|(\tau_n^* - H)^+|^2,$$

and from the analysis in Lemma 4, it follows that

$$\mathbb{E}\left|\int_H^{H \vee \tau_n^*} X_n(t)dt\right|^2 < \infty.$$

Therefore, the sequence of integrals are uniformly integrable, implying (52). Q.E.D.

4. Large-Population Optima: Diffusion Scale

Some of the results in this section require a strengthening of the second moment condition of the service time.

Assumption 1. *The service times* ν_i *possess a finite* $3 + \varepsilon$ *moment; that is,* $\mathbb{E}[\nu_1^{3+\varepsilon}] < \infty$ *for some* $\varepsilon > 0$.

4.1. Model Equations and BOP Derivation

As captured by the fact that the FOP is deterministic, the inherent stochasticity in the FPOP degenerates in fluid scale. A more realistic setting should capture the effect of the stochastic variation introduced by the no-shows and the random service times. In this section, we consider the scheduling problem at diffusion scale that incorporates second-order effects.

Our goal in this subsection is to write down equations for the various quantities of interest related to the diffusion-scale problem and then use these equations to propose a BOP.

By the fluid scale analysis, the rescaled asymptotically optimal schedule $E_n^f(t)$ converges to $\lambda^*(t) = (p^{-1}\mu t) \wedge \alpha$. Under our assumption (18), $p^{-1}\mu H < \alpha$, and so the function λ^* has a jump of size $\lambda^H := \alpha - p^{-1}\mu H$ at H. In general, we denote with a superscript H quantities that correspond to the allocation at the singular time point H.

Denote

$$\hat{E}_n(t) = \frac{E_n(t) - np^{-1}\mu t}{\sqrt{n}}, \qquad t \in [0, H).$$

Recall that $E_n(H) = N_n = \lceil \alpha n \rceil$. If we let $E_n^H = E_n(H) - E_n(H-)$, then

$$\hat{E}_n^H := \frac{E_n^H - n\lambda^H}{\sqrt{n}} = \frac{N_n - E_n(H-) - n\alpha + np^{-1}\mu H}{\sqrt{n}} = -\hat{E}_n(H-) + \frac{\lceil n\alpha \rceil - n\alpha}{\sqrt{n}}. \tag{54}$$

Next, recall $\Xi(k) = \sum_{i=1}^k \xi_i$, $k \in \mathbb{Z}_+$. Let

$$\hat{\Xi}_n(t) = \frac{\sum_{i=1}^{\lceil nt \rceil}(\xi_i - p)}{\sqrt{n}} = \frac{\Xi(\lceil nt \rceil) - p\lceil nt \rceil}{\sqrt{n}}, \qquad t \in \mathbb{R}_+. \tag{55}$$

Now, $A_n(t) = \Xi \circ E_n(t)$ by (1). Let us consider this process on $[0, H)$ separately from its jump at H. For $t \in [0, H)$, use the previous notation to write

$$\begin{aligned}\hat{A}_n(t) &:= \frac{A_n(t) - n\mu t}{\sqrt{n}} = \frac{\Xi(E_n(t)) - pE_n(t)}{\sqrt{n}} + p\frac{E_n(t) - np^{-1}\mu t}{\sqrt{n}} \\ &= \frac{\Xi(\lceil n\bar{E}_n(t) \rceil) - p\lceil n\bar{E}_n(t) \rceil}{\sqrt{n}} + p\frac{E_n(t) - np^{-1}\mu t}{\sqrt{n}} = \hat{\Xi}_n(\bar{E}_n(t)) + p\hat{E}_n(t), \quad t \in [0, H).\end{aligned} \tag{56}$$

The number of show-ups at time H is $A_n^H := A_n(H) - A_n(H-) = \Xi(N_n) - \Xi(E_n(H-))$. Hence,

$$\begin{aligned}\hat{A}_n^H &:= \frac{A_n^H - np\lambda^H}{\sqrt{n}} \\ &= \frac{\Xi(N_n) - np\alpha}{\sqrt{n}} - \frac{\Xi(E_n(H-)) - pE_n(H-)}{\sqrt{n}} - p\frac{E_n(H-) - np^{-1}\mu H}{\sqrt{n}} \\ &= \hat{\Xi}_n(\alpha) - \hat{\Xi}_n(\bar{E}_n(H-)) - p\hat{E}_n(H-).\end{aligned} \tag{57}$$

Next, we let

$$\hat{S}_n(t) = \frac{S_n(t) - n\mu t}{\sqrt{n}}, \qquad t \geq 0.$$

We define the diffusion-scale queue length, for $t \in [0, H)$ only, as

$$\hat{Q}_n(t) = \frac{Q_n(t)}{\sqrt{n}}, \qquad t \in [0, H).$$

It is possible to consider the diffusion-scale queue length for $t \geq H$ by first centering about $q^*(t)$ (as defined in Lemma 3) and then rescaling, but to avoid confusion, we do not extend the process $\hat{Q}^n$ beyond the interval $[0, H)$. We denote $q_n = n^{-1/2}(Q_n(H) - nq^*(H))$.

By (2), letting $I_n(t) = t - B_n(t)$ denote the cumulative idleness process,

$$\begin{aligned}\hat{Q}_n(t) &= n^{-1/2}(A_n(t) - S_n \circ B_n(t)) \\ &= \hat{A}_n(t) - \hat{S}_n(B_n(t)) + n^{1/2}\mu I_n(t) \\ &= \hat{\Xi}_n(\bar{E}_n(t)) + p\hat{E}_n(t) - \hat{S}_n(B_n(t)) + n^{1/2}\mu I_n(t), \qquad t \in [0,H), \quad (58) \\ &= \Gamma_1[p\hat{E}_n + X_n](t), \qquad t \in [0,H), \quad (59)\end{aligned}$$

where (56) is used, and one denotes

$$X_n(t) = \hat{\Xi}_n(\bar{E}_n(t)) - \hat{S}_n(B_n(t)).$$

The queue length dynamics for the BOP are later derived from this relation.

As for q_n, we can write it, using (57) and (58), as

$$\begin{aligned}q_n &= \hat{Q}_n(H-) + \hat{A}_n^H \\ &= \Gamma_1[p\hat{E}_n + X_n](H-) + \hat{\Xi}_n(\alpha) - \hat{\Xi}_n(\bar{E}_n(H-)) - p\hat{E}_n(H-). \quad (60)\end{aligned}$$

Next, we develop equations for the two ingredients of the cost, namely the overtime $[\tau_n - H]^+$ and the makespan, suitably normalized at the diffusion scale.

For $t \geq H$, (2) is still valid, but $A_n(t)$ is simply given by $A_n(H)$ because there are no arrivals after time H. Moreover, the server is busy continuously on $[H, \tau_n)$ on the event $\tau_n > H$. Hence, $B_n(t) = B_n(H) + t - H$ for $t \in [H, \tau_n)$. Clearly, $Q_n(H) = Q_n(H-) + A_n^H$. For $t \in [H, \tau_n]$, the queue length is given by

$$\begin{aligned}Q_n(t) &= Q_n(H) - D_n(t) + D_n(H) = Q_n(H) - S_n(B_n(t)) + S_n(B_n(H)) \\ &= np\lambda^H + [Q_n(H-) + A_n^H - np\lambda^H] - n\mu(t-H) - [S_n(B_n(t)) - n\mu(B_n(t))] \\ &\quad + [S_n(B_n(H)) - n\mu B_n(H)].\end{aligned}$$

Dividing by $\sqrt{n}$ for $t \in [H, \tau_n]$,

$$n^{-1/2}Q_n(t) = n^{1/2}p\lambda^H + q_n - n^{1/2}\mu(t-H) - \hat{S}_n(B_n(t)) + \hat{S}_n(B_n(H)). \quad (61)$$

We have for τ_n the equation $Q_n(\tau_n) = 0$ and for $\bar{\tau}$, $q^*(\bar{\tau}) = 0$, where we recall that, for $t > H$, $q^*(t) = (p\lambda^H - \mu(t-H)) \vee 0 = (p\alpha - \mu H - \mu(t-H)) \vee 0 = (p\alpha - \mu t) \vee 0$. Hence, $p\lambda^H - \mu(\bar{\tau} - H) = 0$. Using these two relations in (61) gives

$$0 = n^{1/2}[-\mu(\tau_n - \bar{\tau})] + q_n - \hat{S}_n(B_n(\tau_n)) + \hat{S}_n(B_n(H)).$$

If we set

$$\hat{\tau}_n = n^{1/2}(\tau_n - \bar{\tau}), \quad (62)$$

then

$$\mu\hat{\tau}_n = q_n - \hat{S}_n(B_n(\tau_n)) + \hat{S}_n(B_n(H)). \quad (63)$$

Equation (63) is used to propose a formal limit of $\hat{\tau}_n$ and, later, to analyze rigorously the weak limit thereof.

Next, the FOP quantity for the makespan is $\bar{W} = \int_H^{\bar{\tau}}(p\lambda^H - \mu(t-H))dt$, where $\bar{\tau} = H + p\mu^{-1}\lambda^H$. Thus,

$$\begin{aligned}\hat{W}_n &:= n^{1/2}(\bar{W}_n - \bar{W}) \\ &= n^{-1/2}W_n - n^{1/2}\bar{W} \\ &= \hat{W}_n(1) + \hat{W}_n(2), \quad (64)\end{aligned}$$

$$\hat{W}_n(1) := \int_0^{H\wedge\tau_n} \hat{Q}_n(t)dt, \quad (65)$$

$$\hat{W}_n(2) := n^{-1/2}\int_{H\wedge\tau_n}^{\tau_n} Q_n(t)dt - n^{1/2}\int_H^{\bar{\tau}}(p\lambda^H - \mu(t-H))dt. \quad (66)$$

On the event $\tau_n < H$, the first term in the expression for $\hat{W}_n(2)$ is zero, and we obtain $\hat{W}_n(2) = -n^{1/2}\bar{W}$. Next, consider the event $\tau_n \geq H$. Then

$$\hat{W}_n(2) = \int_H^{\tau_n} [n^{-1/2}Q_n(t) - n^{1/2}(p\lambda^H - \mu(t-H))]dt + n^{1/2}\int_{\bar{\tau}}^{\tau_n}(p\lambda^H - \mu(t-H))dt. \tag{67}$$

A use of (61) and the computed value of $\bar{\tau}$ gives

$$\hat{W}_n(2) = \int_H^{\tau_n} [q_n - \hat{S}_n(B_n(t)) + \hat{S}_n(B_n(H))]dt + p\alpha n^{1/2}(\tau_n - \bar{\tau}) - \frac{\mu}{2}n^{1/2}(\tau_n - \bar{\tau})^2. \tag{68}$$

To derive the BOP, note first that the functional central limit theorem (FCLT) applies to the processes $\hat{\Xi}_n$ and $\hat{S}_n$. That is, let $X^{(1)}$ and $X^{(2)}$ be mutually independent, one-dimensional BMs with zero drift and diffusion coefficient $(p(1-p))^{1/2}$ and $\mu^{1/2}C_S$, respectively. Then by the FCLT, $(\hat{\Xi}_n, \hat{S}_n) \Rightarrow (X^{(1)}, X^{(2)})$ (Billingsley [4, section 17]).

We take *formal* limits in Equation (58). Denote by Q, U, and L limits of the processes $\hat{Q}_n$, $p\hat{E}_n$, and $n^{1/2}\mu I_n$, respectively, on the time interval $[0, H)$. Denote $\tilde{\mu} = p^{-1}\mu$, and *approximate* $\bar{E}_n(t)$ as $\tilde{\mu}t$, and $B_n(t)$ as t, $t \in [0, H)$. Then, we expect the following relationship to hold in the limit

$$\hat{Q}(t) = U(t) + X^{(1)}(\tilde{\mu}t) - X^{(2)}(t) + L(t), \qquad t \in [0, H). \tag{69}$$

Moreover,

$$\hat{Q}(t) \geq 0, \ t \in [0, H), \text{ and } \int_{[0,H)} \hat{Q}(t)dL(t) = 0. \tag{70}$$

From (57), letting A^H be a weak limit of $\hat{A}_n^H$, we have $A^H = X^{(1)}(\alpha) - X^{(1)}(\tilde{\mu}H) - U(H-)$. To obtain an expression for $\hat{\tau}$, a weak limit of $\hat{\tau}_n$, use (63) to write

$$\hat{\tau} = \mu^{-1}(\hat{Q}(H-) + A^H - X^{(2)}(\bar{\tau}) + X^{(2)}(H)) = \mu^{-1}(X^{(1)}(\alpha) - X^{(2)}(\bar{\tau}) + L(H-)).$$

With $\hat{W}$, $\hat{W}(1)$, and $\hat{W}(2)$ representing limits of $\hat{W}_n$, $\hat{W}_n(1)$, and $\hat{W}_n(2)$, respectively, we have, from (64), $\mathbb{E}\hat{W}(1) = \mathbb{E}\int_0^H \hat{Q}(t)dt$ (because τ_n converges to $\bar{\tau} \geq H$), and by (68), taking into account that $n^{1/2}(\tau_n - \bar{\tau})^2 = n^{-1/2}\hat{\tau}_n^2$ and that $\hat{\tau}_n$ weakly converge (recall that this derivation is a formal step),

$$\mathbb{E}\hat{W}(2) = (\bar{\tau} - H)\mathbb{E}(\hat{Q}(H-) + A^H) + p\alpha\mathbb{E}\hat{\tau} = (\bar{\tau} - H + \mu^{-1}p\alpha)\mathbb{E}L(H-) = (2\mu^{-1}p\alpha - H)\mathbb{E}L(H-).$$

The cost is, thus, given by

$$\begin{aligned}\hat{J}(U) &= c_w\mathbb{E}[\hat{W}(1) + \hat{W}(2)] + c_o\mathbb{E}[\hat{\tau}] \\ &= c_w\mathbb{E}\int_0^H \hat{Q}(t)dt + c_w(2\mu^{-1}p\alpha - H)\mathbb{E}[L(H-)] + c_o\mu^{-1}\mathbb{E}[L(H-)] \\ &= c_w\mathbb{E}\int_0^H \hat{Q}(t)dt + \tilde{c}_o\mathbb{E}[L(H-)],\end{aligned} \tag{71}$$

where $\tilde{c}_o = c_w(2\mu^{-1}p\alpha - H) + c_o\mu^{-1}$. Finally, we can simplify (69) by considering X, a BM with drift zero and diffusion coefficient $\sigma = (\tilde{\mu}p(1-p) + \mu C_S^2)^{1/2} = \mu^{1/2}(1 - p + C_S^2)^{1/2}$ in place of the two BM terms and write

$$\hat{Q}(t) = U(t) + X(t) + L(t), \qquad t \in [0, H). \tag{72}$$

We, thus, let $\mathcal{U}$ denote the collection of right continuous with left limits (RCLL) functions $u : [0, H) \to \mathbb{R}$ and note that given $u \in \mathcal{U}$, (70) and (72) uniquely define Q^U and L^U in terms of X.

We can now state the BOP of interest as a problem involving (70)–(72) with value given by

$$\hat{V} = \inf_{U \in \mathcal{U}} \hat{J}(U). \tag{73}$$

We stress that the BOP has been obtained by means of formal limits. Its justification as a problem that is rigorously related to the prelimit constitutes the main results of this section.

Remark 1. Instead of $\mathcal{U}$ being RCLL functions defined on $[0, H)$, we can work with $\mathcal{U}_H$, the set of RCLL functions on $[0, H]$, and replace the term $\mathbb{E}[L(H-)]$ by $\mathbb{E}[L(H)]$ in (71). This does not change the value $\hat{V}$ because, at optimality, U must be continuous at H. This is because having a jump $L(H) - L(H-) > 0$ can only increase the cost $\hat{J}$ as compared

with having $L(H) = L(H-)$ (the jump cannot be negative because L is nondecreasing). Throughout what follows, we work with $\mathcal{U}_H$ instead of $\mathcal{U}$ and also with this slightly modified definition of $\hat{J}(U)$.

Remark 2. We can present the optimization problem in a way that the cost is more explicit and moreover makes it easy to see that it is a convex optimization problem (Mitter [24]). The pair of Equations (70) and (72) is related to the Skorokhod problem on the half line. Namely, $Q = \Gamma_1[U + X]$. Thus,

$$\hat{Q}(t) = U(t) + X(t) - \inf_{s\in[0,t]}[(U(s) + X(s)) \wedge 0], \qquad L(t) = \hat{Q}(t) - U(t) - X(t). \tag{74}$$

We can, therefore, write $\hat{J}$ as

$$\hat{J}(U) = c_w \mathbb{E}\int_0^H \Gamma_1[U + X](t)dt + \tilde{c}_o \mathbb{E}[\Gamma_1[U + X](H) - U(H) - X(H)].$$

4.2. Large Time Solution of the BOP

In this subsection, we analyze the BOP at the large H limit. Note carefully that assumption (18) puts a restriction on H, namely $p^{-1}\mu H < \alpha$. Thus, H cannot be taken arbitrarily large without modifying (μ, p, α). In our treatment, μ and p remain fixed, and α and H grow so that the assumption remains valid. However, this issue is not significant in this subsection in which we only work with the BOP itself because the parameter α does not show up in it. It does become relevant in later sections.

Recall that X is a $(0, \sigma)$ BM, $\mathcal{U}_H$ denotes the collection of RCLL functions $[0, H] \to \mathbb{R}$, and for $U \in \mathcal{U}_H$ let $L = L^U$ and $Q = Q^U$ be defined as

$$L_t = \sup_{s\in[0,t]}(-X(s) - U(s))^+, \qquad \hat{Q}(t) = X(t) + U(t) + L(t), \qquad t \in [0, H].$$

Let also

$$\hat{J}_H(U) = \frac{c_w}{H}\mathbb{E}\int_0^H \hat{Q}(t)dt + \frac{\tilde{c}_o}{H}\mathbb{E}[\hat{Q}(H) - U(H)], \qquad U \in \mathcal{U}_H, \tag{75}$$

$$\hat{V}_H = \inf_{U\in\mathcal{U}_H} \hat{J}_H(U).$$

Denote by $\mathcal{U}_H^{\text{lin}}$ the collection of linear functions $U(t) = \beta t$, $t \in [0, H]$ for some $\beta \in \mathbb{R}$. Note that the process Q corresponding to such a control is a reflected BM with drift β and diffusion coefficient σ.

For $\beta < 0$, let $m_{\text{RBM}(\beta)}(dx) = -\frac{2\beta}{\sigma^2}e^{2\beta x/\sigma^2}dx$, $x \in [0, \infty)$. This probability measure on $[0, \infty)$ is the stationary distribution of RBM with drift $\beta < 0$ and diffusion coefficient σ. Let

$$V^* = \inf_{\beta<0}\left[c_w \int x\, m_{\text{RBM}(\beta)}(dx) - \tilde{c}_o\beta\right]. \tag{76}$$

We next establish that for the large horizon BOP it is sufficient to consider control functions in $\mathcal{U}_H^{\text{lin}}$.

Proposition 3. *One has*

$$\lim_{H\to\infty} \hat{V}_H = \lim_{H\to\infty} \inf_{U\in\mathcal{U}_H^{\text{lin}}} \hat{J}_H(U) \tag{77}$$

$$= V^* = \sigma\sqrt{2c_w\tilde{c}_o}. \tag{78}$$

Moreover, $\beta^* = -\sigma\sqrt{c_w/(2\tilde{c}_o)}$ is optimal for both the expressions in (76) and (77); that is, with $U^*(t) = \beta^* t$, $\lim_{H\to\infty} \hat{J}_H(U^*) = V^*$ and $c_w \int x\, m_{\text{RBM}(\beta^*)}(dx) - \tilde{c}_o\beta^* = V^*$.

The proof is based on several lemmas. The first is concerned with large time behavior of RBM and computes V^*.

Lemma 6.

1. *For each $\beta \in \mathbb{R}$, let Q_t^β be a (β, σ) RBM starting at the origin. Then, for $\beta < 0$ and $t \geq 0$, $\mathbb{E}[Q_t^\beta] \leq \int x m_{\text{RBM}(\beta)}(dx)$. Moreover,*

$$\lim_{t\to\infty} \inf_{\beta\in\mathbb{R}}[c_w\mathbb{E}[Q_t^\beta] - \tilde{c}_o\beta] = V^*.$$

2. *The infimum in (76) is attained at* $\beta^* = -\sigma\sqrt{c_w/(2\tilde{c}_o)}$ *and* $V^* = \sigma\sqrt{2c_w\tilde{c}_o}$.

The following lemma shows that one can focus on controls under which $\mathbb{E}[Q_H]$ is sublinear in H. Stated precisely, we show

Lemma 7. *For every* $\varepsilon > 0$, *there exists* H_0 *such that for* $H > H_0$ *and* U *for which* $\mathbb{E}[Q_H^U] \geq \varepsilon H$ *one can find* $\tilde{U}$ *with* $\mathbb{E}[Q_H^{\tilde{U}}] < \varepsilon H$ *and* $\hat{J}_H(\tilde{U}) \leq \hat{J}_H(U) + e_1(H)$. *Here,* $e_1(H)$ *does not depend on* U *and converges to zero as* $H \to \infty$.

This lemma confirms the following intuition. Because the driving BM X has mean zero, a nearly optimal U will dictate that the process $Q(t)$ remains $O(1)$ as time t gets large rather then let it grow linearly in t so as to avoid penalty associated with Q. Hence, policies under which one has $\mathbb{E}[Q_H] \geq \varepsilon H$ can be improved.

The following lemma argues that one may focus on controls that are constant on initial and terminal intervals. More precisely, given $z, k, H \in (0, \infty)$, $2z < H$, define the class of controls $\mathcal{U}^{\#}(z, k, H)$ as the collection of members $U \in \mathcal{U}_H$ that satisfy $U_t = k$ for $t \in [0, z)$ and $U_t = U_H$ for $t \in [H - z, H]$.

Lemma 8. *Given* z, k, H *and* $U \in \mathcal{U}_H$, *let* $U^{\#} \in \mathcal{U}^{\#}(z, k, H)$ *be defined as*

$$U_t^{\#} = \begin{cases} k, & t \in [0, z), \\ U_t, & t \in [z, H - z), \\ U_H, & t \in [H - z, H]. \end{cases}$$

Fix $\varepsilon > 0$, *let* $H_0 = H_0(\varepsilon)$ *be as in Lemma 7, and consider* $H > H_0$ *and* $U \in \mathcal{U}_H$ *for which* $\mathbb{E}[Q_H] \leq \varepsilon H$. *Then*

$$\hat{J}_H(U^{\#}) \leq \hat{J}_H(U) + e_2(z, k, H), \tag{79}$$

where

$$\limsup_H e_2(z, k, H) \leq c_w \mathbb{E}[r] + c_w \varepsilon z, \tag{80}$$
$$r = r(k, z, X) = \sup_{s \in [0,z)} (-k - X_s)^+.$$

The intuition behind this lemma is that, if the time horizon H is large, one could modify the control U on intervals of fixed length (namely, $[0, z]$ and $[H - z, H]$) with little overall effect.

The following lemma relates the large time behavior of $\hat{J}_H(U)$ for U as in Lemma 8 to the expression V^*. Its proof uses the special structure of controls from Lemma 8 that are constant in the initial and terminal parts of the time horizon.

Lemma 9. *One has*

$$\liminf_{H \to \infty} \inf_{U \in \mathcal{U}^{\#}(z,k,H)} \hat{J}_H(U) \geq V^* - e_3(z),$$

where $e_3(z) \to 0$ *as* $z \to \infty$.

Proof of Proposition 3. Using Lemmas 7–9, for any $\varepsilon > 0$, $z > 0$ and $k > 0$,

$$\begin{aligned} \liminf_{H \to \infty} \hat{V}_H &\geq \liminf_{H \to \infty} \inf_{U \in \mathcal{U}_H : E Q_H^U < \varepsilon H} \hat{J}_H(U) \\ &\geq \liminf_{H \to \infty} \inf_{U \in \mathcal{U}^{\#}(z,k,H) : E Q_H^U < \varepsilon H} \hat{J}_H(U) - \limsup_{H \to \infty} e_2(z, k, H) \\ &\geq V^* - \limsup_{H \to \infty} e_2(z, k, H) - e_3(z), \end{aligned}$$

for e_2 and e_3 as in these lemmas. Thus,

$$\liminf_{H \to \infty} \hat{V}_H \geq V^* - c_w \mathbb{E}\left[\sup_{s \in [0,z)} (-k - X_s)^+ \right] - c_w \varepsilon z - e_3(z).$$

We refer to the last three terms on the RHS as the first, second, and third error terms in the order at which they appear. We first take $\varepsilon \to 0$ so that the second error term vanishes. Then we take $k \to \infty$ to have the first error term vanish as a direct consequence of $\sup_{s \in [0,z)} (-k - X_s)^+ \leq (-k + \|X\|_z)^+$ and $\mathbb{E}[\|X\|_z] < \infty$. Finally, we take $z \to \infty$, and the third term vanishes. We have, thus, shown that $\liminf_{H \to \infty} \hat{V}_H \geq V^*$.

For a matching upper bound, for any $\beta<0$, letting $U(t)=U^\beta(t)=\beta t$, we have with the notation of Lemma 6

$$\hat{J}_H(U^\beta)=\frac{c_w}{H}\mathbb{E}\int_0^H Q_t^\beta dt+\frac{\tilde{c}_o}{H}\mathbb{E}(Q_H^\beta-\beta H)\le c_w\int x m_{\mathrm{RBM}(\beta)}(dx)-\tilde{c}_o\beta+\frac{\tilde{c}_o}{H}\int x m_{\mathrm{RBM}(\beta)}(dx),$$

where we used (75) and the domination stated in Lemma 6(1). As a result,

$$\begin{aligned}\hat{V}_H&\le \hat{J}_N(U^{\beta^*})\\&\le c_w\int x m_{\mathrm{RBM}(\beta^*)}(dx)-\tilde{c}_o\beta^*+\frac{\tilde{c}_o}{H}\int x m_{\mathrm{RBM}(\beta^*)}(dx)\\&=V^*+\frac{\tilde{c}_o}{H}\int x m_{\mathrm{RBM}(\beta^*)}(dx),\end{aligned}$$

where Lemma 6(2) is used. This shows $\limsup_{H\to\infty}\hat{V}_H\le V^*$. We conclude that $\lim_{H\to\infty}\hat{V}_H=V^*$. Q.E.D.

Proof of Lemma 6. For each $t\ge 0$, the cumulative distribution function (CDF) of Q_t^β is given by

$$P(Q_t^\beta\le y)=\Phi\Big(\frac{y-\beta t}{\sigma t^{1/2}}\Big)-e^{2\beta y/\sigma^2}\Phi\Big(\frac{-y-\beta t}{\sigma t^{1/2}}\Big),\qquad y\ge 0,$$

where Φ is the standard normal CDF (Harrison [16]). For fixed $\beta<0$, the limit distribution as $t\to\infty$ is exponential with mean $\sigma^2/(2|\beta|)$. Moreover, it can be directly checked that the CDF is monotone decreasing in t. Hence, by monotone convergence, the expectation $\mathbb{E}[Q_t^\beta]$ converges as $t\to\infty$ to $\sigma^2/(2|\beta|)$, provided $\beta<0$. Moreover, by the aforementioned monotonicity of the CDF, $\mathbb{E}[Q_t^\beta]$ is monotone increasing in t and is, therefore, bounded above by the latter constant, which proves the first assertion in Lemma 6(1). For $\beta\ge 0$, $P(Q_t^\beta\le y)\to 0$ for all y; hence, $\mathbb{E}[Q_t^\beta]\to\infty$. Thus, denoting $F(t,\beta)=c_w\mathbb{E}[Q_t^\beta]-\tilde{c}_o\beta$ and $F(\beta)=c_w\sigma^2/(2|\beta|)+\tilde{c}_o\beta$ for $\beta<0$, $F(\beta)=\infty$ for $\beta\ge 0$, we have the pointwise convergence $\lim_{t\to\infty}F(t,\beta)=F(\beta)$.

Our goal now is to show

$$\lim_{t\to\infty}\inf_{\beta\in\mathbb{R}}F(t,\beta)=\inf_{\beta<0}F(\beta).\tag{81}$$

We achieve this in three steps. First, we show that $\inf_{\beta\ge 0}F(t,\beta)\to\infty$ as $t\to\infty$. Then we argue that there exist $-\infty<a<-1<b<0$ such that, for all large t, $\inf_{\beta\in\mathbb{R}}F(t,\beta)=\inf_{\beta\in[a,b]}F(t,\beta)$. Then we are in a position to use Dini's theorem to argue that the order of the t-limit and the β-infimum can be interchanged. We use an additional monotonicity property. It can be readily checked by the CDF formula that $\beta\to P(Q_t^\beta\le y)$ is monotone decreasing (for each t and y). Hence, $\beta\to\mathbb{E}[Q_t^\beta]$ is monotone increasing (for each t).

For the first step alluded to, the pointwise convergence $EQ_t^0\to\infty$ as $t\to\infty$ can be used to deduce $\inf_{\beta\in[0,1]}[c_wEQ_t^\beta-\tilde{c}_o\beta]\to\infty$ because for $\beta\in[0,1]$ and all t we have $c_wEQ_t^\beta-\tilde{c}_o\beta\ge c_wEQ_t^0-\tilde{c}_o\to\infty$ as $t\to\infty$. To show $\inf_{\beta\in(1,\infty)}[c_wEQ_t^\beta-\tilde{c}_o\beta]\to\infty$, we argue as follows. By the formula for Q_t^β, we have the lower bound $Q_t^\beta\ge X_t+\beta t$. Hence, for all $\beta>1$, $c_wQ_t^\beta-\tilde{c}_o\beta\ge c_wX_t+\beta(c_wt-\tilde{c}_o)$. Hence, for $t>\tilde{c}_o/c_w$, $c_wQ_t^\beta-\tilde{c}_o\beta\ge c_wX_t+(c_wt-\tilde{c}_o)$. Taking expectation gives $\inf_{\beta\in(1,\infty)}[c_wEQ_t^\beta-\tilde{c}_o\beta]\ge c_wt-\tilde{c}_o\to\infty$ as $t\to\infty$.

For the next step, fix $u>0$ and let $-\infty<a<-1<b<0$ such that $\tilde{c}_o|a|>u$, $\tilde{c}_o|b|<1$ and $F(b)>u$ (note that $F(0-)=\infty$). Then, for any $\beta<a$, $F(t,\beta)\ge\tilde{c}_o|\beta|>u$. Next, consider $\beta\in(b,0)$. The pointwise convergence of $F(t,b)$ to $F(b)$ implies that, for some t_0 and all $t\ge t_0$, $F(t,b)>u-1$, and because $\tilde{c}_o|b|<1$, $c_wEQ_t^b>u-2$. Using again the monotonicity in β, $F(t,\beta)\ge c_wEQ_t^\beta>u-2$ for $\beta\in(b,0)$.

Because for fixed $\beta<0$ the limit $\lim_{t\to\infty}F(t,\beta)$ is finite, and the constant u is arbitrary, it follows that a and b can be found so that the infimum is achieved in $[a,b]$ for all large t.

Next, by the explicit representation of Q_t^β in terms of Γ_1, it follows that $|Q_t^\beta-Q_t^{\beta'}|\le 2|\beta-\beta'|t$. Hence, $\mathbb{E}[Q_t^\beta]$ is continuous in β in the compact interval $[a,b]$. Moreover, as already mentioned in the first paragraph of the proof, $\mathbb{E}[Q_t^\beta]$ is monotone in t. Because we have that $F(t,\beta)=c_w\mathbb{E}[Q_t^\beta]-\tilde{c}_o\beta$ converge pointwise as $t\to\infty$ to a continuous function $F(\beta)$, we can use Dini's theorem and obtain that the convergence is uniform in β. We conclude that, as $t\to\infty$, $\inf_{\beta\in[a,b]}F(t,\beta)\to\inf_{\beta\in[a,b]}F(\beta)$. This proves part 1 of the lemma.

It remains to solve the optimization problem (76) or, equivalently, the RHS of (81). As already stated, for $\beta<0$, the first moment of $m_{\mathrm{RBM}(\beta)}$ is given by $\sigma^2/(2|\beta|)$. We are, therefore, interested in minimizing

$$c_w\frac{\sigma^2}{2|\beta|}+\tilde{c}_o|\beta|$$

over $\beta \in (-\infty, 0)$. By a direct calculation, the minimum is attained at $\beta^* = -\sigma(c_w/2\tilde{c}_o)^{1/2}$ and is given by $\sigma(2c_w\tilde{c}_o)^{1/2}$. Q.E.D.

Proof of Lemma 7. Fix $\varepsilon > 0$. Given any H and any $U \in \mathcal{U}_H$, we have for $Q = Q^U$ the relation

$$Q_t = U_t + X_t + \sup_{s \le t}(-U_s - X_s)^+, \qquad t \in [0, H].$$

Define $q = q^U$ as

$$q_t = U_t + \sup_{s \le t}(-U_s)^+, \qquad t \in [0, H].$$

Then

$$\|Q - q\|_H \le 2\|X\|_H, \tag{82}$$

by the Lipschitz continuity of the Skorokhod reflection map. Consider U for which $\mathbb{E}[Q_H] \ge \varepsilon H$. Then $q_H \ge \varepsilon H - 2\mathbb{E}[\|X\|_H]$. Fix H_0 so large that, for every $H > H_0$, $2\mathbb{E}[\|X\|_H] < \varepsilon H/4$ (H_0 exists due to Doob's maximal inequality and Jensen's inequality), and consider in what follows only $H > H_0$. Then $q_H > 3\varepsilon H/4$. Let $\tilde{U}$ be defined as $\tilde{U} = U$ on $[0, H)$ and $\tilde{U}_H = U_H - q_H$ (here, $q = q^U$). Denote $\tilde{Q} = Q^{\tilde{U}}$ and $\tilde{q} = q^{\tilde{U}}$. Clearly, on $[0, H)$, we have $\tilde{q} = q$ and $\tilde{Q} = Q$. As for the time H, we have

$$\tilde{q}_H = U_H - q_H + \sup_{s \le H}(-U_s + q_H 1_{\{s=H\}})^+ = 0.$$

As a result, $\tilde{Q}_H = \tilde{q}_H + \tilde{Q}_H - \tilde{q}_H = \tilde{Q}_H - \tilde{q}_H \le 2\|X\|_H$, by the Lipschitz continuity of the Skorokhod reflection map. This shows $\mathbb{E}[\tilde{Q}_H] \le 2\mathbb{E}[\|X\|_H] < \varepsilon H/4$. Moreover, by (75),

$$\hat{J}_H(\tilde{U}) \le \hat{J}_H(U) + \frac{\tilde{c}_o}{H} 2\mathbb{E}[\|X\|_H] \le \hat{J}_H(U) + c_4 H^{-1/2},$$

for a suitable constant c_4 (again, by Doob's and Jensen's inequalities). This proves the lemma. Q.E.D.

Proof of Lemma 8. Denote $Q = Q^U$ and $Q^\# = Q^{U^\#}$. First, we provide lower estimates on $Q - Q^\#$ on each of the three intervals separately.

The interval $[0, z)$. Here we use the trivial lower bound $Q_t \ge 0$. As for $Q^\#$,

$$Q^\#_t = k + X_t + \sup_{s \le t}(-k - X_s)^+ \le 2k + 2\|X\|_z.$$

The interval $[z, H - z)$. We have

$$\begin{aligned} Q_t &= U_t + X_t + \sup_{s \le t}(-U_s - X_s)^+ \\ &\ge U_t + X_t + \sup_{s \in [z,t]}(-U_s - X_s)^+ \\ &\ge U_t + X_t + \max\left[r, \sup_{s \in [z,t]}(-U_s - X_s)^+\right] - r, \end{aligned}$$

where we have used the fact that $a \ge a \vee b - b$ provided $a \ge 0$ and $b \ge 0$. Recalling that for $t \in [z, H - z)$ U and $U^\#$ agree and that $U^\# = k$ on $[0, z)$, the expression is equal to

$$U^\#_t + X_t + \sup_{s \in [0,t]}(-U^\#_s - X_s)^+ - r.$$

It follows that $Q_t \ge Q^\#_t - r$.

The interval $[H - z, H]$. In fact, we only need a lower estimate on $Q_H - Q^\#_H$. We have

$$Q_H = U_H + X_H + \sup_{s \le H}(-U_s - X_s)^+,$$

and

$$Q_H^{\#} = U_H + X_H + \max\Big[\sup_{s<H-z}(-U_s^{\#} - X_s)^+, \sup_{s\in[H-z,H]}(-U_H - X_s)^+\Big].$$

Hence,

$$\begin{aligned}
Q_H &\geq U_H + X_H + \sup_{s\in[z,H]}(-U_s - X_s)^+ \\
&\geq U_H + X_H + \max\Big[r, \sup_{s\in[z,H]}(-U_s - X_s)^+\Big] - r \\
&\geq U_H + X_H + \max\Big[r, \sup_{s\in[z,H-z)}(-U_s - X_s)^+, (-U_H - X_H)^+\Big] - r \\
&\geq U_H + X_H + \max\Big[r, \sup_{s\in[z,H-z)}(-U_s - X_s)^+, \sup_{s\in[H-z,H]}(-U_H - X_s)^+\Big] - r - \hat{r},
\end{aligned}$$

where $\hat{r} = \sup_{s\in[H-z,H]}|X_s - X_H|$. This shows $Q_H \geq Q_H^{\#} - r - \hat{r}$.

Next,

$$\begin{aligned}
\hat{J}_H(U) &= \frac{c_w}{H}\mathbb{E}\int_0^H Q_s ds + \frac{\tilde{c}_o}{H}\mathbb{E}[Q_H - U_H] \\
&\geq \frac{c_w}{H}\mathbb{E}\int_z^{H-z} Q_s ds + \frac{\tilde{c}_o}{H}\mathbb{E}[Q_H - U_H] \\
&\geq \frac{c_w}{H}\Big[\mathbb{E}\int_z^{H-z} Q_s^{\#} ds - (H-2z)\mathbb{E}[r]\Big] + \frac{\tilde{c}_o}{H}\{\mathbb{E}[Q_H^{\#} - U_H^{\#}] - \mathbb{E}[r+\hat{r}]\}.
\end{aligned}$$

Also,

$$\begin{aligned}
\hat{J}_H(U^{\#}) &= \frac{c_w}{H}\mathbb{E}\int_0^H Q_s^{\#} ds + \frac{\tilde{c}_o}{H}\mathbb{E}[Q_H^{\#} - U_H^{\#}] \\
&\leq \frac{c_w}{H}\Big[2k + 2\mathbb{E}[\|X\|_z] + \mathbb{E}\int_z^{H-z} Q_s^{\#} ds + z\mathbb{E}[Q_H^{\#} + \hat{r}]\Big] + \frac{\tilde{c}_o}{H}\mathbb{E}[Q_H^{\#} - U_H^{\#}],
\end{aligned}$$

where we used that for $t \in [H-z, H]$ one has $Q_t^{\#} = X_t + U_t^{\#} + L_t^{\#} \leq X_t + U_H^{\#} + L_H^{\#} = X_t - X_H + Q_H^{\#}$; hence, $Q_t^{\#} \leq Q_H^{\#} + \hat{r}$. Combine these two bounds to obtain

$$\begin{aligned}
\hat{J}_H(U^{\#}) - \hat{J}_H(U) &\leq \frac{c_w}{H}\Big[(H-2z)\mathbb{E}[r] + 2k + 2\mathbb{E}[\|X\|_z] + z\mathbb{E}[Q_H^{\#} + \hat{r}]\Big] + \frac{\tilde{c}_o}{H}\mathbb{E}[r+\hat{r}] \\
&\leq \frac{c_w}{H}\Big[(H-2z)\mathbb{E}[r] + 2k + 2\mathbb{E}[\|X\|_z] + z\mathbb{E}[Q_H + r + 2\hat{r}]\Big] + \frac{\tilde{c}_o}{H}\mathbb{E}[r+\hat{r}] \\
&\leq \frac{c_w}{H}\Big[(H-2z)\mathbb{E}[r] + 2k + 2\mathbb{E}[\|X\|_z] + z\varepsilon H + z\mathbb{E}[r + 2\hat{r}]\Big] + \frac{\tilde{c}_o}{H}\mathbb{E}[r+\hat{r}].
\end{aligned}$$

Denote by $e_2(z,k,H)$ the expression on the last line. Note that although $\hat{r}$ depends on H, its expectation does not (and is finite). Thus, e_2 satisfies (80). Q.E.D.

Proof of Lemma 9. Fix z, k, H and a control $U \in \mathcal{U}^{\#}(z,k,H)$. Then $Q_t = Q_t^U = X_t + U_t + \sup_{s\leq t}[-X_s - U_s]^+ \geq \sup_{s\leq t}[X_t - X_s + U_t - U_s]$. Hence, for $z + t \leq H$,

$$\begin{aligned}
\int_0^{z+t} Q_s ds &\geq \int_z^{z+t} Q_s ds \\
&\geq \int_z^{z+t} \sup_{\theta\in[0,s]}[X_s - X_{s-\theta} + U_s - U_{s-\theta}] ds \\
&\geq \int_z^{z+t} \sup_{\theta\in[0,z]}[X_s - X_{s-\theta} + U_s - U_{s-\theta}] ds.
\end{aligned}$$

For each $s \geq z$, the stochastic process $\{X_s - X_{s-\theta}\}_{\theta\in[0,z]}$ is equal in law to $\{X_z - X_{z-\theta}\}_{\theta\in[0,z]}$. As a result,

$$\mathbb{E}\int_0^{z+t} Q_s ds \geq \mathbb{E}\int_z^{z+t} \sup_{\theta\in[0,z]} [X_z - X_{z-\theta} + U_s - U_{s-\theta}]ds.$$

We now use the inequality

$$\int_a^b \sup_\theta f(s,\theta)ds \geq \sup_\theta \int_a^b f(s,\theta)ds.$$

This gives

$$\begin{aligned}\frac{1}{t}\mathbb{E}\int_0^{z+t} Q_s ds &\geq \frac{1}{t}\mathbb{E}\sup_{\theta\in[0,z]}\int_z^{z+t}[X_z - X_{z-\theta} + U_s - U_{s-\theta}]ds \\ &= \mathbb{E}\sup_{\theta\in[0,z]}\Big\{[X_z - X_{z-\theta}] + \frac{1}{t}\int_z^{z+t}[U_s - U_{s-\theta}]ds\Big\}.\end{aligned}$$

We have

$$\int_z^{z+t}[U_s - U_{s-\theta}]ds = \int_z^{z+t} U_s ds - \int_{z-\theta}^{z-\theta+t} U_s ds = \int_{z-\theta+t}^{z+t} U_s ds - \int_{z-\theta}^{z} U_s ds.$$

Moreover, by the assumption on U, we have $U = k$ on $[0,z)$, $U = U_H$ on $[H-z,H]$. Thus, with $t + z = H$,

$$\frac{1}{H-z}\mathbb{E}\int_0^H Q_s ds \geq \mathbb{E}\sup_{\theta\in[0,z]}\Big\{[X_z - X_{z-\theta}] + \theta\frac{1}{H-z}(U_H - k)\Big\}.$$

Thus,

$$\begin{aligned}\hat{J}_H(U) &= \frac{c_w}{H}\mathbb{E}\int_0^H Q_t dt + \frac{\tilde{c}_o}{H}\mathbb{E}[Q_H - U_H] \\ &\geq c_w\mathbb{E}\sup_{\theta\in[0,z]}\Big\{\frac{H-z}{H}[X_z - X_{z-\theta}] + \theta\Big(\frac{U_H}{H} - \frac{k}{H}\Big)\Big\} - \tilde{c}_o\frac{U_H}{H} \\ &\geq \inf_{\beta\in\mathbb{R}}\Big\{c_w\mathbb{E}\sup_{\theta\in[0,z]}\Big\{\frac{H-z}{H}[X_z - X_{z-\theta}] + \theta\Big(\beta - \frac{k}{H}\Big)\Big\} - \tilde{c}_o\beta\Big\}.\end{aligned}$$

Denoting $\delta = z/H$,

$$\begin{aligned}\hat{J}_H(U) &\geq \inf_{\beta\in\mathbb{R}}\Big\{c_w\mathbb{E}\sup_{\theta\in[0,z]}\Big\{(1-\delta)[X_z - X_{z-\theta}] + \theta\beta\Big\} - \tilde{c}_o\beta\Big\} - k\delta \\ &\geq \inf_{\beta\in\mathbb{R}}\Big\{c_w\mathbb{E}\sup_{\theta\in[0,z]}\Big\{[X_z - X_{z-\theta}] + \theta\beta\Big\} - \tilde{c}_o\beta\Big\} - k\delta - 2\delta\mathbb{E}\|X\|_z.\end{aligned}$$

Send $H \to \infty$ (hence, $\delta \to 0$) to obtain

$$\liminf_{H\to\infty}\inf_{U\in\mathcal{U}^\#(z,k,H)}\hat{J}_H(U) \geq \inf_{\beta\in\mathbb{R}}\Lambda(z,\beta),$$

where

$$\Lambda(z,\beta) = \Big\{c_w\mathbb{E}\sup_{\theta\in[0,z]}\Big\{[X_z - X_{z-\theta}] + \theta\beta\Big\} - \tilde{c}_o\beta\Big\}.$$

Now, if we let $Q_z = \sup_{\theta\in[0,z]}\{[X_z - X_{z-\theta}] + \theta\beta\}$, then Q_z is a (β,σ) RBM starting at origin. Hence, by Lemma 6(1), $\lim_{z\to\infty}\inf_{\beta\in\mathbb{R}}\Lambda(z,\beta) = V^*$. This proves the lemma. Q.E.D.

4.3. Lower Bound on the Diffusion-Scale Cost

Recall from (75) the definitions of $\hat{J}_H$ and $\hat{V}_H$, the cost and value of the BOP. Also recall the diffusion-scale cost $\hat{J}_{n,H}(\{T_i\})$ and value $\hat{V}_{n,H} = n^{1/2}[V_{n,H} - \bar{V}_H]$, where $V_{n,H}$ is defined in (7) and $\bar{V}_H$ is the FOP value defined in (36). This subsection is devoted to proving the following result.

Theorem 3. *Let Assumption 1 hold. Fix H. Then*

$$H^{-1} \liminf_{n\to\infty} \hat{V}_{n,H} \geq \hat{V}_H. \tag{83}$$

Assumption 1 is in force throughout this subsection. (It is used in the proof of Lemma 11.)

We say that a sequence $\{T_i^n\} \in \mathcal{T}_n$, $n \in \mathbb{N}$ *achieves the limit inferior in* (83) if

$$\lim_{n\to\infty} \hat{J}_{n,H}(\{T_i^n\}) = \liminf_{n\to\infty} \hat{V}_{n,H}. \tag{84}$$

If the expression on the left-hand side (LHS) of (83) is infinite, then there is nothing to prove. Hence, we may and will assume that, for any such $\{T_i^n\}$, the sequence $\hat{J}_{n,H}(\{T_i^n\})$ of (84) is bounded. The following lemma provides various convergence results based on, to a large extent, the boundedness of the sequence of costs.

Lemma 10. *There exists a sequence $\{T_i^n\} \subset \mathcal{T}_n$, $n \in \mathbb{N}$ that achieves the limit inferior in* (83) *and for which assertions* (i)–(iv) *hold.*

i. $\sup_n \mathbb{E}[(\hat{\tau}_n^-)^2] \vee \mathbb{E}[\hat{\tau}_n^+] < \infty$ *(in particular, $\hat{\tau}_n$ are tight, and $\tau_n \Rightarrow \bar{\tau}$ as $n \to \infty$).*

ii. $\sup_{t\in[0,\tau_n]} |B_n(t) - t| \Rightarrow 0$ *as* $n \to \infty$.

iii. $\sup_n \int_0^H \Gamma_1[p\hat{E}_n](t)dt < \infty$ *and* $\sup_n\{\Gamma_1[p\hat{E}_n](H-) - p\hat{E}_n(H-)\} < \infty$.

iv. $\sup_{t\in[0,H)} |\bar{E}_n(t) - p^{-1}\mu t| \to 0$ *and* $|\bar{E}_n^H - \lambda^H| \to 0$ *as* $n \to \infty$.

Fix a sequence $\{T_i^n\}$ as in Lemma 10. Let all the processes and RVs, such as W_n, Q_n, τ_n, etc., denote those associated with the schedule $\{T_i^n\}$ for each n. Recall the diffusion-scale expression $\hat{W}_n$ from (64). Also recall $\hat{W}_n = \hat{W}_n(1) + \hat{W}_n(2)$. Then, by (23),

$$\hat{J}_{n,H}(\{T_i^n\}) = \mathbb{E}[c_w\hat{W}_n(1) + c_w\hat{W}_n(2) + c_o\hat{\tau}_n]. \tag{85}$$

Recall that $(\hat{\Xi}_n, \hat{S}_n) \Rightarrow (X^{(1)}, X^{(2)})$; that these two BMs are mutually independent; and that, by its definition, X is equal in law to the sum $X^{(1)}(\tilde{\mu}\cdot) + X^{(2)}(\cdot)$.

In the derivation of the BOP in Section 4.1, we took formal limits in the equations that describe the scaled processes, such as (58) for $\hat{Q}_n$, (63) for $\hat{\tau}_n$, etc. *We are unable to turn this into a rigorous argument* in a straightforward manner because there is no apparent precompactness for the sequence of functions $\{\hat{E}_n\}$. For example, in (58), there is no justification to replace the limit of the term $p\hat{E}_n$ by some control U. We, therefore, take a different route, in which we are able to provide a lower bound in which the error term converges to zero as $n \to \infty$ only because of the convergence of the stochastic processes and RVs involved (such as τ_n, $\hat{\Xi}_n$), not relying on any convergence associated with $\hat{E}_n$.

An outline of the argument is as follows. We appeal to Skorokhod's representation theorem and derive (in Lemma 11) a bound of the form

$$\mathbb{E}[c_w\hat{W}_n(1) + c_w\hat{W}_n(2) + c_o\hat{\tau}_n] \geq H\hat{J}_H(p\hat{E}_n) - \mathbb{E}[\varepsilon_n], \tag{86}$$

where ε_n is a sequence of RVs satisfying $\mathbb{E}[\varepsilon_n] \to 0$ as $n \to \infty$. The argument leading to this estimate is based on the closeness of $(\hat{\Xi}_n, \hat{S}_n)$ to $(X^{(1)}, X^{(2)})$. However, it does not require $\hat{E}_n$ to be close to any candidate limit and, thus, allows us to avoid the aforementioned issue regarding precompactness of this sequence. Now, for each n, $p\hat{E}_n$ is a member of $\mathcal{U}_H$. Thus, using (85), it follows from the definition of $\hat{V}_H$ that

$$\hat{J}_{n,H}(\{T_i^n\}) \geq H\hat{V}_H - \mathbb{E}[\varepsilon_n]. \tag{87}$$

In view of the fact that $\mathbb{E}[\varepsilon_n] \to 0$, the result follows.

Toward stating Lemma 11, note that the convergence $\hat{\Xi}_n \Rightarrow X^{(1)}$ and the one stated in Lemma 10(iv) imply that $\hat{\Xi}_n \circ \bar{E}_n \Rightarrow X^{(1)}(\tilde{\mu}\cdot)$ (recall that $\tilde{\mu} = p^{-1}\mu$). By Lemma 10(i), $\tau_n \geq H$ with probability tending to one; hence, by Lemma 10(ii), $\sup_{t\in[0,H]} |B_n(t) - t| \Rightarrow 0$. Consequently, $\hat{S}_n \circ B_n \Rightarrow X^{(2)}$, and so $X_n \Rightarrow X$ in the uniform topology on $[0, H)$. Moreover, for $t \in [H, \tau_n)$ (on the event $\tau_n > H$), $B_n(t) = B_n(H) + (t - H)$ by the nonidling property. It

follows that $(\hat{S}_n \circ B_n)(\cdot \wedge \tau_n) \Rightarrow X^{(2)}(\cdot \wedge \bar{\tau})$. We now appeal to Skorokhod's representation theorem, by which we may assume without loss of generality that, a.s.,

$$\tau_n \to \bar{\tau}, \quad \text{and } (\hat{\Xi}_n, \hat{\Xi}_n \circ \bar{E}_n, (\hat{S}_n \circ B_n)(\cdot \wedge \tau_n)) \to (X^{(1)}, X^{(1)}(\tilde{\mu}\cdot), X^{(2)}(\cdot \wedge \bar{\tau})), \tag{88}$$

uniformly on compacts. If we let $X = X^{(1)}(\tilde{\mu}\cdot) - X^{(2)}$, then we also have $X_n(\cdot \wedge \tau_n) \to X(\cdot \wedge \bar{\tau})$ a.s.

Lemma 11. *The estimate* (86) *holds with a sequence of RVs* ε_n *for which* $\mathbb{E}[\varepsilon_n] \to 0$ *as* $n \to \infty$.

Proof of Theorem 3. The estimate (86) holds by Lemma 11. Hence, (87) is valid. Taking the limit inferior and using the convergence $\mathbb{E}[\varepsilon_n] \to 0$ stated in Lemma 11 establishes the result. Q.E.D

We turn to the proofs of the lemmas. We use uniform second-moment bounds as follows. For every t,

$$\sup_n \mathbb{E}[\|\hat{\Xi}_n\|_t^2] < \infty, \qquad \sup_n \mathbb{E}[\|\hat{S}_n\|_t^2] < \infty, \tag{89}$$

where the first assertion follows by Doob's L^2 maximum inequality (Durrett [12, section 4.4]), and the second is shown in Krichagina and Taksar [21, theorem 4].

Recall that c denotes a generic positive constant (nonrandom, independent of n), whose value may change from line to line. Moreover, $\{\Theta_n\}$ denotes a generic sequence of nonnegative RVs that are uniformly square integrable; that is, $\sup_n \mathbb{E}[\Theta_n^2] < \infty$. The value of the sequence $\{\Theta_n\}$ may also change from line to line.

Note that, in view of (89), for every fixed t, one has $\|\hat{S}_n\|_t \le \Theta_n$ and $\|\hat{\Xi}_n\|_t \le \Theta_n$. Moreover, $\|\hat{S}_n \circ B_n\|_t \le \Theta_n$ because $0 \le B_n(s) \le s$ for all s.

Proof of Lemma 10. We argue that assertions (i)–(iii) hold for any sequence $\{T_i^n\}$ that achieves the limit inferior in (83). Hence, we fix such a sequence and denote it by $\{T_i^n\}$ (thus, (84) is valid for this sequence). On the other hand, the proof of part (iv) requires a certain construction; it is achieved by modifying this fixed sequence in a suitable way.

It follows from (85) and the fact that the sequence of costs is bounded that

$$c_w \mathbb{E}[\hat{W}_n(1)] + c_w \mathbb{E}[\hat{W}_n(2)] + c_o \mathbb{E}[\hat{\tau}_n] \le c. \tag{90}$$

We would like to deduce from (90) that each of the terms on the LHS is bounded above by a constant. Before we may do so, we must provide a lower bound on each of these terms. The first term is nonnegative by its definition.

Next we show that $\mathbb{E}[(\hat{\tau}_n^-)^2] \le c$, equivalently $\hat{\tau}_n^- \le \Theta_n$. By (58), using the boundedness of $\bar{E}_n(H-)$ and $B_n(H-)$ and the nonnegativity of I_n, we have $\hat{Q}_n(H-) \ge -\Theta_n + p\hat{E}_n(H-)$. By (57), $\hat{A}_n^H \ge -\Theta_n - p\hat{E}_n(H-)$. Thus,

$$\hat{Q}_n(H) \ge -\Theta_n. \tag{91}$$

Consider the event $\hat{\tau}_n \le 0$ and use (63). On this event, the expression $B_n(\tau_n)$ is bounded above by $\bar{\tau}$. Hence, the term $\hat{S}_n(B_n(\tau_n))$ is bounded in absolute value by $\|\hat{S}_n\|_{\bar{\tau}}$. Using this and the lower bound (91) in (63) gives $\hat{\tau}_n 1_{\{\hat{\tau}_n \le 0\}} \ge -\Theta_n$. This gives

$$\mathbb{E}[(\hat{\tau}_n^-)^2] \le c, \qquad n \in \mathbb{N}. \tag{92}$$

A lower bound on $\mathbb{E}[\hat{W}_n(2)]$ is achieved by considering the three expressions $l_n = \mathbb{E}[\{\hat{W}_n(2)1_{\{\tau_n < H\}}]$, $l_n' = \mathbb{E}[\{\hat{W}_n(2)1_{\{\tau_n \in [H,\bar{\tau}]\}}]$, and $l_n'' = \mathbb{E}[\{\hat{W}_n(2)1_{\{\tau_n > \bar{\tau}\}}]$. For l_n, we use the lower bound $-cn^{1/2}$ on $\hat{W}_n(2)$ and (92), by which $\mathbb{P}(\tau_n < H) = \mathbb{P}((\tau_n - \bar{\tau})^- > (\bar{\tau} - H)) = \mathbb{P}(\hat{\tau}_n^- > n^{1/2}(\bar{\tau} - H)) \le cn^{-1}$. This shows $l_n \ge -cn^{-1/2}$.

For l_n', on the event $\tau_n \in [H, \bar{\tau}]$, we use (68). By this equation, we have

$$\hat{W}_n(2)1_{\{\tau_n \in [H,\bar{\tau}]\}} \ge -\Theta_n - \frac{\mu}{2} n^{-1/2} (\hat{\tau}_n^-)^2 + p\alpha\hat{\tau}_n.$$

By (92), $\hat{\tau}_n \ge -\Theta_n$. This shows $l_n' \ge -c$.

As for l_n'', it follows from (66) by a calculation similar to that leading to (68) that, on $\{\tau_n > \bar{\tau}\}$,

$$\hat{W}_n(2) = \int_H^{\bar{\tau}} [q_n - \hat{S}_n(B_n(H) + t - H) + S_n(B_n(H))]dt + n^{1/2} \int_{\bar{\tau}}^{\tau_n} Q_n(t)dt.$$

The last term is nonnegative on the indicated event; hence,

$$\hat{W}_n(2)1_{\{\tau_n > \bar{\tau}\}} \ge -\Theta_n.$$

Thus, $l_n'' \geq -c$. We conclude that $\mathbb{E}[\hat{W}_n(2)] \geq -c$.

In view of these lower bounds, (90) now implies

$$(a)\ \mathbb{E}[\hat{W}_n(1)] \leq c, \qquad (b)\ \mathbb{E}[\hat{W}_n(2)] \leq c, \qquad (c)\ \mathbb{E}[\hat{\tau}_n^+] \leq c. \tag{93}$$

The bounds (92) and (93)(c) prove part (i) of the lemma.

Next, the tightness of $\hat{\tau}_n$ used in (63) implies the tightness of q_n. By (60), this gives the tightness of $n^{1/2}I_n(H-)$. Because $I_n(H-) = I_n(H) = H - B_n(H)$, we obtain $B_n(H) \Rightarrow H$. By the property $|B_n(t) - B_n(s)| \leq |t-s|$, this implies $\sup_{t\in[0,H]} |B_n(t) - t| \Rightarrow 0$, and because for $t \in [H, \tau_n]$ we have $B_n(t) - B_n(H) = t - H$, the result stated in part (ii) of the lemma follows.

The bound (93)(a) clearly implies the tightness of $\hat{W}_n(1)$. By the expression (65) for $\hat{W}_n(1)$ and the convergence $\tau_n \Rightarrow \bar{\tau}$, this gives the tightness of $\int_0^H \hat{Q}_n(t)dt$. Using (59) and the Lipschitz property of Γ_1 (with constant two),

$$\int_0^H \Gamma_1[p\hat{E}_n](t)dt \leq \int_0^H \Gamma_1[p\hat{E}_n + X_n](t)dt + 2H\|X_n\|_H = \int_0^H \hat{Q}_n(t)dt + 2H\|X_n\|_H,$$

Hence, the tightness of the RHS implies that of the LHS. However, the expression on the LHS is deterministic; thus, it is, simply, bounded. This gives the first assertion in part (iii).

The aforementioned tightness of the RVs $n^{1/2}I_n(H-)$, along with the equality between the two expressions (60), implies the tightness of the RVs $\Gamma_1[p\hat{E}_n(H-)] - p\hat{E}_n(H-)$ (in view of the tightness of the terms involving $\hat{\Xi}_n$ and $\hat{S}_n$ in these expressions). Arguing by the Lipschitz continuity of Γ_1 establishes the second assertion in part (iii).

Finally, we prove part (iv). Given $\varepsilon > 0$, we first show that $\sup_{t\in[0,H-\varepsilon]}(\bar{E}_n(t) - p^{-1}\mu t) \leq \varepsilon$ provided n is sufficiently large. Recall that $\bar{E}_n(t)$ is nondecreasing. Thus, if the inequality $\sup_{t\in[0,H-\varepsilon]}(\bar{E}_n(t) - p^{-1}\mu t) > \varepsilon$ is valid for some n, then there exists $t = t_n \in [0, H-\varepsilon]$ such that, for all $s \in [t, t+\varepsilon_0 \wedge \varepsilon]$, where $\varepsilon_0 = \varepsilon p\mu^{-1}/2$, $\bar{E}_n(s) - p^{-1}\mu s > \varepsilon - p^{-1}\mu(s-t) > \frac{\varepsilon}{2}$. Hence, by the definition of $\hat{E}_n$, $\hat{E}^n(s) \geq \sqrt{n}\frac{\varepsilon}{2}$ for the same set of times s. By the definition of Γ, we have $\Gamma[p\hat{E}^n](s) \geq p\hat{E}_n(s)$, for each s in the interval alluded to. Hence, $\int_0^H \Gamma(p\hat{E}^n)ds \geq c_\varepsilon\sqrt{n}$, where $c_\varepsilon > 0$ depends on ε but not on n. By part (iii), this can occur for only finitely many n. Hence, the claim.

Next, given $\varepsilon > 0$, assume that the inequality $\bar{E}_n(t) - p^{-1}\mu t > 3\varepsilon p^{-1}\mu$ is valid for some n and $t = t_n \in [H-\varepsilon, H)$. Then $E_n(t) > p^{-1}\mu(t+3\varepsilon)n \geq p^{-1}\mu(H+2\varepsilon)n$. Construct from E_n another schedule $E_n^{(1)}(s) = E_n(s) \wedge p^{-1}\mu(H+2\varepsilon)n$ for $s \in [0, H)$. Then the schedules agree on $s \in [0, t)$. Moreover, the new schedule satisfies the constraint $\bar{E}_n(t) - p^{-1}\mu t < c\varepsilon$ for all times t. The resulting queue length can only be decreased by this modification. Moreover, the effect of the modification on the overage time is negligible at the scaling limit as follows by the following argument that shows under both schedules no idle time is accumulated during $[t, H)$ with high probability. This is because, for $s \in [t, H)$,

$$\bar{E}_n^{(1)}(s) = p^{-1}\mu(H+2\varepsilon).$$

Hence, according to (34), with the error term ε_n as in that equation,

$$\begin{aligned}
\bar{Q}_n(s) &= p\bar{E}_n(s) - \mu s + \varepsilon_n(s) + \sup_{u\leq s}(-p\bar{E}_n(u) + \mu u - \varepsilon_n(u))^+ \\
&= \mu(H+2\varepsilon) - \mu s + \varepsilon_n(s) + \sup_{u\leq s}(-p\bar{E}_n(u) + \mu u - \varepsilon_n(u))^+ \\
&\geq \mu(H+2\varepsilon) - \mu s + \varepsilon_n(s) \\
&\geq 2\mu\varepsilon - \|\varepsilon_n\|_H.
\end{aligned}$$

This shows that for any $\varepsilon > 0$ one can construct E_n for which $\sup_{t\in[0,H)}(\bar{E}_n(t) - p^{-1}\mu t) < \varepsilon$, and (84) and assertions (i)–(iii) of the lemma hold. A diagonal argument may now be used to take $\varepsilon = \varepsilon(n) \downarrow 0$.

Next, a similar use of the formula (34) now for the term μI_n shows that if

$$\inf_{t\in[0,H]}(\bar{E}_n(t) - p^{-1}\mu t) \leq -\varepsilon,$$

then one has $I_n(H) \geq c\varepsilon$ with high probability. In this case, the cost associated with the scaled overage time $\hat{\tau}_n$ grows without bound as $n \to \infty$.

Finally, the second assertion of part (iv) follows from the first one by using the constraint $\bar{E}_n(H) = N_n$. Q.E.D.

Proof of Lemma 11. We first estimate $\hat{W}_n(1)$ from below. By (58), for $t \in [0,H)$, $\hat{Q}_n(t)$ is given by $p\hat{E}_n(t) + X_n(t) + n^{1/2}\mu I_n(t)$, where we recall that $X_n(t) = \hat{\Xi}_n(\bar{E}_n(t)) - \hat{S}_n(B_n(t))$. We use (59) and the fact that Γ_1 is Lipschitz with constant 2 in the supremum norm to write

$$\begin{aligned}\hat{W}_n(1) &= \int_0^{H\wedge\tau_n} \hat{Q}_n(t)dt \\ &= \int_0^{H\wedge\tau_n} \{\Gamma_1[p\hat{E}_n + X_n](t) - \Gamma_1[p\hat{E}_n + X](t)\}dt + \int_0^{H\wedge\tau_n} \Gamma_1[p\hat{E}_n + X](t)dt \\ &\geq \int_0^{H} \Gamma_1[p\hat{E}_n + X](t)dt - \varepsilon_n^0,\end{aligned}$$

where

$$\varepsilon_n^0 = 2(H\wedge\tau_n)\|X_n - X\|_{H\wedge\tau_n} + \int_{H\wedge\tau_n}^{H} \Gamma_1[p\hat{E}_n + X](t)dt.$$

The term q_n (see (60)) appears in the expression for both $\hat{W}_n(2)$ and $\hat{\tau}_n$. We have

$$\begin{aligned}q_n &= \Gamma_1[p\hat{E}_n + X_n](H-) + \hat{\Xi}_n(\alpha) - \hat{\Xi}_n(\bar{E}_n(H-)) - p\hat{E}_n(H-) \\ &\geq \Lambda_n - \varepsilon_n^1,\end{aligned} \tag{94}$$

where

$$\begin{aligned}\Lambda_n &= \Gamma_1[p\hat{E}_n + X](H-) - p\hat{E}_n(H-) + X^{(1)}(\alpha) - X^{(1)}(\tilde{\mu}H), \\ \varepsilon_n^1 &= 2\|X_n - X\|_H + |\hat{\Xi}_n(\alpha) - X^{(1)}(\alpha)| + |\hat{\Xi}_n(\bar{E}_n(H-)) - X^{(1)}(\tilde{\mu}H)|.\end{aligned}$$

Recall that, on the event $\tau_n < H$, $\hat{W}_n(2) = -n^{1/2}\bar{W}$. On the event $\tau_n \geq H$, we have the expression (68). We obtain (in both cases)

$$\hat{W}_n(2) \geq (\tau_n - H)^+(\hat{Q}_n(H-) + \hat{A}_n^H) - \int_H^{H\vee\tau_n} (X^{(2)}(t\wedge\bar{\tau}) - X^{(2)}(H))dt + p\alpha\hat{\tau}_n - \varepsilon_n^2 - \varepsilon_n^3 - \varepsilon_n^4,$$

where

$$\varepsilon_n^2 = 2\|\hat{S}_n \circ B_n(\cdot\wedge\tau_n) - X^{(2)}(\cdot\wedge\bar{\tau})\|_{\tau_n}, \quad \varepsilon_n^3 = \frac{\mu}{2}n^{1/2}(\tau_n - \bar{\tau})^2, \quad \varepsilon_n^4 = n^{1/2}\bar{W}1_{\{\tau_n<H\}}.$$

Hence,

$$\hat{W}_n(2) \geq (\bar{\tau} - H)\Lambda_n - \int_H^{\bar{\tau}} (X^{(2)}(t\wedge\bar{\tau}) - X^{(2)}(H))dt + p\alpha\hat{\tau}_n - \varepsilon_n^2 - \varepsilon_n^3 - \varepsilon_n^4 - \varepsilon_n^5,$$

where

$$\varepsilon_n^5 = |\tau_n - \bar{\tau}|(\hat{Q}_n(H-) + \hat{A}_n^H) + \bar{\tau}\varepsilon_n^1 + 2|\tau_n - \bar{\tau}|\,\|X^{(2)}\|_{\bar{\tau}}.$$

As for $\hat{\tau}_n$, using (63),

$$\hat{\tau}_n \geq \mu^{-1}\Big[\Lambda_n - X^{(2)}(\bar{\tau}) + X^{(2)}(H) - \varepsilon_n^1 - \varepsilon_n^2\Big].$$

We now combine the lower bounds obtained on $\hat{W}_n(1)$, $\hat{W}_n(2)$, and $\hat{\tau}_n$ and compare with expression (71) with $p\hat{E}_n$ substituted for U. The term $L(H-)$ appearing in (71) is related to Λ_n via

$$\mathbb{E}[L(H-)] = \mathbb{E}[\Gamma_1[p\hat{E}_n + X](H-) - p\hat{E}_n(H-)] = \mathbb{E}[\Lambda_n].$$

We have, thus, shown that (86) holds with $\varepsilon_n = c_5(\varepsilon_n^0 + \cdots + \varepsilon_n^5)$, where c_5 is a constant.

It now remains to show that each of the terms $\mathbb{E}[\varepsilon_n^i]$, $i = 0, 1, \ldots, 5$ converges to zero.

The first term in ε_n^0 converges to zero a.s. by (88). It follows from (89) that this term is uniformly integrable. Hence, its expectation also converges to zero. As for the second term in ε_n^0, using the Lipschitz property of Γ_1, this term is bounded by

$$1_{\{\tau_n<H\}}\left\{\int_0^H \Gamma_1[p\hat{E}_n](t)dt + 2H\|X\|_H\right\} \le 1_{\{\tau_n<H\}}\left\{c + 2H\|X\|_H\right\},$$

where we used Lemma 10(iii), and c is a suitable constant. The indicator function converges to zero a.s. by (88), and the expectation of $\|X\|_H$ is finite. Hence, by dominated convergence, $\mathbb{E}[\varepsilon_n^0] \to 0$.

The a.s. convergence of ε_n^1 and ε_n^2 to zero follows from (88), whereas their uniform integrability follows from (89). Thus, $\mathbb{E}[\varepsilon_n^1] \to 0$ and $\mathbb{E}[\varepsilon_n^2] \to 0$.

Next, by (62), $n^{1/2}(\tau_n - \bar{\tau})^2 = n^{-1/2}\hat{\tau}_n^2$. Thus, to show that $\mathbb{E}[\varepsilon_n^3] \to 0$ it suffices to improve the estimate from Lemma 10(i) to show that $\mathbb{E}[\hat{\tau}_n^2]$ is bounded.

To this end, recall expression (63). Because we already established the boundedness of the second moment of q_n, it suffices to show that also the second and third terms in (63) have bounded second moments. Because $B_n(H) \le H$, the second moment of the last term in (63) is bounded by $\mathbb{E}[\|\hat{S}_n\|_H^2]$; hence, (89) gives a uniform bound. It remains to show that

$$\sup_n \mathbb{E}[\hat{S}_n(B_n(\tau_n))^2] < \infty. \tag{95}$$

Here we use the $3+\varepsilon$ moment assumption. First, note that $B_n(\tau_n)$ is the total time the server works on jobs, which thus is equal to the total arriving work. A bound on this is given by $\sum_{i=1}^{N_n} \nu_{i,n} = n^{-1}\sum_{i=1}^{N_n}\nu_i$. Denoting $w_n = \sum_{i=1}^{N_n}\nu_i$, we have

$$\begin{aligned}
\mathbb{E}[\hat{S}_n(B_n(\tau_n))^2] &= \mathbb{E}[1_{\{n^{-1}w_n<1\}}\hat{S}_n(B_n(\tau_n))^2] + \sum_{k=0}^{\infty}\mathbb{E}[1_{\{n^{-1}w_n\in[2^k,2^{k+1})\}}\hat{S}_n(B_n(\tau_n))^2] \\
&\le \mathbb{E}[1_{\{n^{-1}w_n<1\}}\|\hat{S}_n\|_{n^{-1}w_n}^2] + \sum_{k=0}^{\infty}\mathbb{E}[1_{\{n^{-1}w_n\in[2^k,2^{k+1})\}}\|\hat{S}_n\|_{n^{-1}w_n}^2] \\
&\le \mathbb{E}[\|\hat{S}_n\|_1^2] + \sum_{k=0}^{\infty}\mathbb{E}[1_{\{n^{-1}w_n\in[2^k,2^{k+1})\}}\|\hat{S}_n\|_{2^{k+1}}^2] \\
&\le c + \sum_{k=0}^{\infty}\mathbb{P}(w_n \ge n2^k)^{1/p}\,\mathbb{E}[\|\hat{S}_n\|_{2^{k+1}}^{2q}]^{1/q},
\end{aligned}$$

where (89) is used for the first term, and for the sum, Hölder's inequality is used, where $p^{-1} + q^{-1} = 1$. Fix $\beta \in (3, 3+\varepsilon)$. Then by Minkowski's inequality we have $\mathbb{E}[w_n^\beta] \le cn^\beta\mathbb{E}[\nu_1^\beta]$. Hence, $\mathbb{P}(w_n \ge n2^k) \le c2^{-\beta k}$, where c is finite and does not depend on n or k. Next we appeal again to Krichagina and Taksar [21, theorem 4], which states that $\mathbb{E}[\|\hat{S}_n\|_t^\beta]^{1/\beta} \le c(t^{1/2}+1)$ (under the hypothesis that the βth moment of ν_1 is finite and $\beta \ge 2$), where c does not depend on t or n. We use this estimate with $t = 2^{k+1}$, and $q = \beta/2$ (accordingly, p is determined). This gives

$$\mathbb{E}[\hat{S}_n(B_n(\tau_n))^2] \le c + c\sum_{k=0}^{\infty} 2^{-\beta k/p}(2^{k+1}+1).$$

Now, $p = (1-2/\beta)^{-1}$, and because $\beta > 3$, we have $p < 3$. In particular, $\beta > p$. Therefore, the sum is finite. This proves (95) and, hence, follows the estimate on $\mathbb{E}[\varepsilon_n^3]$.

To show that $\mathbb{E}[\varepsilon_n^4] \to 0$ amounts to showing that $n^{1/2}\mathbb{P}(\tau_n < H) \to 0$. Now, $\tau_n \to \bar{\tau} > H$, and we have just shown that $\sup_n \mathbb{E}[\hat{\tau}_n^2] < \infty$. Hence, for $c = \bar{\tau} - H$,

$$n^{1/2}\mathbb{P}(\tau_n < H) \le n^{1/2}\mathbb{P}(n^{-1/2}|\hat{\tau}_n| > c) \le \tilde{c}n^{-1/2},$$

for some constant $\tilde{c}$. This shows $\mathbb{E}[\varepsilon_n^4] \to 0$.

Finally, for a bound on ε_n^5, we use Cauchy–Schwartz to write

$$\mathbb{E}[\varepsilon_n^5] \le \mathbb{E}[(\tau_n - \bar{\tau})^2]^{1/2}\mathbb{E}[q_n^2]^{1/2} + \bar{\tau}\mathbb{E}[\varepsilon_n^1] + c\mathbb{E}[(\tau_n - \bar{\tau})^2]^{1/2}.$$

We have already shown that q_n are uniformly square integrable. Hence, the convergence of the first and last terms to zero follows from the boundedness of $\mathbb{E}[\hat{\tau}_n^2]$. The second term has already been argued to converge to zero.

This concludes the proof that $\mathbb{E}[\varepsilon_n^i] \to 0$ for $i = 0, 1, \ldots, 5$ and completes the proof of the lemma. Q.E.D.

4.4. Upper Bound on the Diffusion-Scale Cost

In this section, we propose a sequence of schedules (indexed by n) whose diffusion-scaled cost converges to the cost we obtained as a solution to the BOP. This establishes that the BOP cost is an asymptotic upper bound for diffusion-scaled cost, which, together with the lower bound, establishes the asymptotic optimality of our proposed schedule.

Let β^* be the optimal drift associated with the problem defined in Proposition 3; namely β^* is such that the control $U(t) = \beta^* t$ is optimal for the problem $\lim_{H\to\infty} \inf_{U\in\mathcal{U}_H^{\mathrm{lin}}} \hat{J}_H(U)$. Note that $\beta^* < 0$. Let $n > (\beta^*/\mu)^2$, and for $t \in [0, H]$, define

$$E_n^d(t) = \begin{cases} \left\lfloor \dfrac{nt}{p}\left(\mu + \dfrac{\beta^*}{\sqrt{n}}\right)\right\rfloor & t < H, \\ N_n & t = H. \end{cases} \tag{96}$$

To see that $E_n^d(t)$ is an admissible schedule, we need to verify that $E_n^d(t) \geq 0$ and that it is nondecreasing in t. Both of these requirements follow from the condition that $n > (\beta^*/\mu)^2$ and from our assumption that $p^{-1}\mu H < \alpha$ in (18). The latter is used to verify that the jump of E_n^d at $t = H$ is nonnegative.

Next, consider the diffusion-scaled schedule $\hat{E}_n^d := \sqrt{n}\big(n^{-1}E_n^d - \lambda^*\big)$; recall that λ^* is the fluid-optimal schedule defined in (37). A straightforward computation shows that, as $n \to \infty$,

$$\hat{E}_n^d(t) \to \hat{u}(t) := \begin{cases} \beta^* t, & t \in [0, H) \\ 0, & t \geq H. \end{cases} \tag{97}$$

Furthermore, it can also be easily shown that the convergence is uniformly on compact sets of the time index. It follows that as $H \to \infty$, $\hat{u}$ converges to the drift of the stationary optimal RBM associated with the problem $\lim_{H\to\infty} \inf_{U\in\mathcal{U}_H^{\mathrm{lin}}} \hat{J}_H(U)$. The main result of this section proves that this sequence of schedules asymptotically achieves the large time horizon value of the BOP determined in Proposition 3.

Theorem 4. *The sequence of schedules* $\{T_{i,n}^d, n \geq 1\}$ *corresponding to the scheduling functions* $\{E_n^d, n \geq 1\}$ *of* (96) *satisfies*

$$\lim_{H\to\infty} \limsup_{n\to\infty} H^{-1}\hat{J}_{n,H}(\{T_{i,n}^d\}) = \lim_{H\to\infty} \hat{V}_H.$$

Proof. Recall the diffusion-scaled cost in (85):

$$\hat{J}_{n,H}(\{T_{i,n}^d\}) = \mathbb{E}[c_w \hat{W}_n(1) + c_w \hat{W}_n(2) + c_o \hat{\tau}_n],$$

where $\hat{W}_n(1), \hat{W}_n(2)$, and $\hat{\tau}_n$ are defined in (65), (66), and (71) (respectively). Applying Skorokhod's representation theorem as in (88), we have that a.s.

$$\tau_n \to \bar{\tau}, \quad \text{and } (\hat{\Xi}_n, \hat{\Xi}_n \circ \bar{E}_n^d, (\hat{S}_n \circ B_n)(\cdot \wedge \tau_n)) \to (X^{(1)}, X^{(1)}(\tilde{\mu}\cdot), X^{(2)}(\cdot \wedge \bar{\tau})) \text{ as } n \to \infty.$$

As before, let $X := X^{(1)}(\tilde{\mu}\cdot) - X^{(2)}$ so that $X_n(\cdot \wedge \tau_n) \to X(\cdot \wedge \bar{\tau})$ a.s.

Consider $\hat{W}_n(1)$ first. From (18), (88), and (97) and the fact that $B_n(t) \in o(n^{1/2})$ a.s. for all $t \in [0, H)$, it follows that $\hat{W}_n(1) = \int_0^{H\wedge\tau_n} \hat{Q}_n(t)dt - n^{-1/2}B_n(\tau_n)$ satisfies

$$\hat{W}_n(1) \to \int_0^H (p\hat{u} + X)(t)dt \text{ a.s. as } n \to \infty. \tag{98}$$

Next, recall from (68) that on the event $\{\tau_n \geq H\}$

$$\hat{W}_n(2) = \int_H^{\tau_n} \Big[q_n - \hat{S}_n(B_n(t)) + \hat{S}_n(B_n(H))\Big]dt + p\alpha\hat{\tau}_n - \frac{\mu}{2}\sqrt{n}(\tau_n - \bar{\tau})^2,$$

where $q_n = \hat{Q}_n(H-) + \hat{A}_n^H$ and $\hat{A}_n^H = \hat{\Xi}_n(\alpha) - \hat{\Xi}_n(\bar{E}_n^d(H-)) - p\hat{E}_n^d(H-)$. Again, using (88) and (97), it can be easily seen that

$$q_n \to \Gamma_1(p\hat{u} + X)(H-) + X^{(1)}(\alpha) - X^{(1)}(\tilde{\mu}H-) - p\hat{u}(H-) = \hat{q} \text{ a.s. as } n \to \infty. \tag{99}$$

It follows from (63), (88), and (99) that

$$\hat{\tau}_n \to \frac{1}{\mu}(\hat{q} - X^{(2)}(\bar{\tau}) + X^{(2)}(H)) = \hat{\tau} \text{ a.s. as } n \to \infty. \tag{100}$$

Now, because $\bar{\tau} > H$, it follows from (68), (99), and (100) that

$$\hat{W}_n(2) \to \int_H^{\bar{\tau}} \left[\hat{q} - X^{(2)}(t) + X^{(2)}(H)\right] dt + p\alpha\hat{\tau} \text{ a.s. as } n \to \infty, \tag{101}$$

where we have used the fact that $\sqrt{n}(\tau_n - \bar{\tau})^2 = (\tau_n - \bar{\tau})\hat{\tau}_n \to 0$ a.s. as $n \to \infty$.

At this point, we have shown that, as $n \to \infty$, a.s.

$$c_w\hat{W}_n(1) + c_w\hat{W}_n(2) + c_o\hat{\tau}_n \to c_w \int_0^H \Gamma_1(p\hat{u} + X)(t)dt + c_w \int_H^{\bar{\tau}} \left[\hat{q} - X^{(2)}(t) + X^{(2)}(H)\right] dt + (c_o + p\alpha)\hat{\tau}.$$

To prove convergence in L^1, following the analysis in Section 4.3, we prove that the second moments of $\hat{W}_n(1)$, $\hat{W}_n(2)$, and $\hat{\tau}_n$ are bounded. This, however, follows directly from (89) and the analysis in the proof of Lemma 11 and will not be repeated here. Therefore, we have that

$$\hat{J}_{n,H}(\{T_{i,n}^d\}) = c_w\hat{W}_n(1) + c_w\hat{W}_n(2) + c_o\hat{\tau}_n$$

is uniformly integrable, implying that $\lim_{n\to\infty} \hat{J}_{n,H}(\{T_{i,n}^d\}) = J_H(\hat{u})$, where

$$\begin{aligned} J_H(\hat{u}) = {} & \frac{c_w}{H}\mathbb{E}\left[\int_0^H \Gamma_1(p\hat{u} + X)(t)dt\right] \\ & + \frac{c_w}{H}\mathbb{E}\left[\int_H^{\bar{\tau}} \left(\hat{q} - X^{(2)}(t) + X^{(2)}(H)\right)dt\right] + \frac{c_o + p\alpha}{H}\mathbb{E}[\hat{\tau}]. \end{aligned} \tag{102}$$

Using the fact that $X^{(1)}$ and $X^{(2)}$ are Brownian motion processes and the definition of $\hat{\tau}$ in (100), $\hat{J}_H(\hat{u})$ simplifies to

$$\hat{J}_H(\hat{u}) = \frac{c_w}{H}\mathbb{E}\left[\int_0^H \Gamma_1(p\hat{u} + X)(t)dt\right] + \frac{\tilde{c}_o}{H}\mathbb{E}[\Gamma_1(p\hat{u} + X)(H-) - p\hat{u}(H-)],$$

where we recall that $\tilde{c}_o = c_o/\mu + c_w(\bar{\tau} - H)$. Now, using the fact that $\hat{u}$ converges to the drift function $\beta^* e$ as $H \to \infty$, it follows that $\lim_{H\to\infty} J_H(\hat{u}) = \lim_{H\to\infty} \hat{V}_H$, thus completing the proof. Q.E.D.

4.5. The Stochasticity Gap at the Diffusion Scale

Continuing our investigation of quantifying the effect of the stochasticity on the scheduling problem, we now consider the asymptotic SG in the diffusion scale. We study the limit of $\hat{\gamma}_{n,H} = \sqrt{n}(V_{n,H} - V_{n,H}^{CI}) \geq 0$ as $n \to \infty$ and then $H \to \infty$. We first show that the value of the CI problem in the diffusion scale is asymptotically null.

Lemma 12. *Fix $H > 0$. We have $\hat{\gamma}_{n,H}^{CI} = \sqrt{n}(V_{n,H}^{CI} - \bar{V}_H) \to 0$ as $n \to \infty$.*

We delay the proof of the lemma to after the main result of this section.

Proposition 4. *Let Assumption 1 hold. Then the large horizon SG is positive. More precisely,* $\lim_{H\to\infty} \liminf_{n\to\infty} \hat{\gamma}_{n,H} = \lim_{H\to\infty} \hat{V}_H^* =: \hat{V}^*$.

Proof. First note that $\hat{\gamma}_{n,H} = \sqrt{n}(V_{n,H} - \bar{V}_H) - \sqrt{n}(V_{n,H}^{CI} - \bar{V}_H)$, where $\bar{V}_H$ is the FOP value from (36). Also recall the definition $\hat{V}_{n,H} := \sqrt{n}(V_{n,H} - \bar{V}_H)$. Then, Theorem 3 and Lemma 12 imply that

$$\lim_{H\to\infty} \liminf_{n\to\infty} \hat{\gamma}_{n,H} \geq \hat{V}^*. \tag{103}$$

On the other hand, Theorem 4 and Lemma 12 together imply that

$$\lim_{H\to\infty} \limsup_{n\to\infty} \hat{\gamma}_{n,H} \leq \hat{V}^*, \tag{104}$$

completing the proof. Q.E.D.

Proof of Lemma 12. Straightforward algebraic manipulation of (30) shows that

$$\hat{\gamma}^{CI}_{n,H} = \mathbb{E}\left[\frac{c_w}{H}\sqrt{n}\left(\frac{1}{n}\int_H^{H\vee\tau_n^*}(\Xi(N_n)-S_n(t)-1)\,dt - \int_H^{\bar\tau}(\alpha p-\mu t)dt\right) + \frac{c_o}{H}\sqrt{n}\big((\tau_n^*-H)^+-(\bar\tau-H)\big)\right],$$

where we have used the fact that $\bar V_H = H^{-1}c_w\int_H^{\bar\tau}(\alpha p-\mu t)dt + H^{-1}c_o(\bar\tau-H)$.

In the proof of Lemma 4, it was shown that $\tau_n^*\to\bar\tau$ a.s. as $n\to\infty$. Let $x\in\mathbb{R}$ and consider the event $\{\sqrt{n}(\tau_n^*-\bar\tau)>x\}$ or, equivalently, $\{\tau_n^*>n^{-1/2}x+\bar\tau\}$. Recall that $\tau_n^* := \inf\{t>0 : S_n(t)\ge\Xi(N_n)\}$ so that

$$\{\tau_n^*>n^{-1/2}x+\bar\tau\} = \big\{S_n\left(n^{-1/2}x+\bar\tau\right)<\Xi(N_n)\big\} = \{\hat\Xi_n(\alpha)-\hat S_n(n^{-1/2}x+\bar\tau)>x\mu\}, \tag{105}$$

where the last equality follows by simple algebraic manipulations and recognizing that $\bar\tau=\mu^{-1}\alpha p$.

Now, consider

$$\hat\Xi_n(\alpha)-\hat S_n(n^{-1/2}x+\bar\tau) = \big(\hat\Xi_n(\alpha)-\hat S_n(\bar\tau)\big)+\big(\hat S_n(\bar\tau)-\hat S_n(n^{-1/2}x+\bar\tau)\big).$$

Recall that $(\hat\Xi_n,\hat S_n)\Rightarrow(X^{(1)},X^{(2)})$ as $n\to\infty$. Therefore, $\hat\Xi_n(\alpha)-\hat S_n(\bar\tau)\Rightarrow X^{(1)}(\alpha)-X^{(2)}(\bar\tau)$ as $n\to\infty$. On the other hand, because $n^{-1/2}x+\bar\tau\to\bar\tau$ as $n\to\infty$, we have $\hat S_n(\bar\tau)-\hat S_n(n^{-1/2}x+\bar\tau)\Rightarrow 0$ as $n\to\infty$. Therefore,

$$\hat\Xi_n(\alpha)-\hat S_n(n^{-1/2}x+\bar\tau)\Rightarrow X^{(1)}(\alpha)-X^{(2)}(\bar\tau). \tag{106}$$

Displays (105) and (106) together imply, as $n\to\infty$,

$$\hat\tau_n^*\Rightarrow\frac{1}{\mu}\big(X^{(1)}(\alpha)-X^{(2)}(\bar\tau)\big). \tag{107}$$

Now, consider $\underline{\hat\tau}_n^* = \sqrt{n}((\tau_n^*-H)^+-(\bar\tau-H))$. For each $n\ge1$, we write $\underline{\hat\tau}_n^* = \underline{\hat\tau}_n^*1_{\{\tau_n^*>H\}}+\underline{\hat\tau}_n^*1_{\{\tau_n^*\le H\}}$. The first term on the right-hand side is simply $\hat\tau_n^*1_{\{\tau_n^*>H\}}$. Because $\tau_n^*\Rightarrow\bar\tau>H$ as $n\to\infty$, it follows for large enough n that $\tau_n^*>H$, implying that $\underline{\hat\tau}_n^*\Rightarrow\frac{1}{\mu}\big(X^{(1)}(\alpha)-X^{(2)}(\bar\tau)\big)$. Following (89), it is straightforward to deduce that $\underline{\hat\tau}_n^*$ is uniformly integrable, implying that

$$\mathbb{E}[\underline{\hat\tau}_n^*]\to\mathbb{E}\big[\mu^{-1}\big(X^{(1)}(\alpha)-X^{(2)}(\bar\tau)\big)\big]=0 \text{ as } n\to\infty. \tag{108}$$

Next, the first term in the definition of $\hat\gamma^{CI}_{n,H}$ can be written as

$$\begin{aligned}
&\sqrt{n}\left(\frac{1}{n}\int_H^{H\vee\tau_n^*}\Big(\Xi(N_n)-S_n(t)-1\Big)dt-\int_H^{\bar\tau}(\alpha p-\mu t)dt\right)\\
&\quad=\sqrt{n}\left(\frac{1}{n}\int_H^{\tau_n^*}(\Xi(N_n)-S_n(t))dt-\frac{1}{n}(\tau_n^*-H)-\int_H^{\bar\tau}(\alpha p-\mu t)dt\right)1_{\{\tau_n^*>H\}}\\
&\qquad+\sqrt{n}\left(-\int_H^{\bar\tau}(\alpha p-\mu t)dt\right)1_{\{\tau_n^*\le H\}}.
\end{aligned}$$

Consider the first term on the right-hand side under the event $\{\tau_n^*>H\}$,

$$\sqrt{n}\left(\frac{1}{n}\int_H^{\tau_n^*}(\Xi(N_n)-S_n(t))dt-\int_H^{\bar\tau}(\alpha p-\mu t)dt\right)-\frac{1}{\sqrt{n}}(\tau_n^*-H),$$

and focus on the term under the integral first:

$$\sqrt{n}\left(\frac{1}{n}\int_H^{\bar\tau}(\Xi(N_n)-S_n(t))dt-\int_H^{\bar\tau}(\alpha p-\mu t)dt\right)+\frac{1}{\sqrt{n}}\int_{\bar\tau}^{\tau_n^*}(\Xi(N_n)-S_n(t))dt. \tag{109}$$

Note that breaking up the integral is justified because $S_n(t)$ is well defined for all $t\ge0$. Consider the latter integral:

$$\frac{1}{\sqrt{n}}\int_{\bar\tau}^{\tau_n^*}(\Xi(N_n)-S_n(t))dt=\int_{\bar\tau}^{\tau_n^*}(\hat\Xi_n(\alpha)-\hat S_n(t))dt+\sqrt{n}\int_{\bar\tau}^{\tau_n^*}(\alpha p-\mu t)dt.$$

It follows that the right-hand side equals

$$\int_{\bar{\tau}}^{\tau_n^*} (\hat{\Xi}_n(\alpha) - \hat{S}_n(t))dt + \alpha p \hat{\tau}_n^* - \frac{\mu}{2}\hat{\tau}_n^*(\tau_n^* + \bar{\tau}),$$

and as $n \to \infty$

$$\frac{1}{\sqrt{n}}\int_{\bar{\tau}}^{\tau_n^*} (\Xi(N_n) - S_n(t))dt \Rightarrow \alpha p \hat{\tau}^* - \mu \bar{\tau}\hat{\tau}^*.$$

Now, using (89), it can be shown that $\frac{1}{\sqrt{n}}\int_{\bar{\tau}}^{\tau_n^*}(\Xi(N_n) - S_n(t))dt$ is uniformly integrable so that

$$\mathbb{E}\left[\frac{1}{\sqrt{n}}\int_{\bar{\tau}}^{\tau_n^*} (\Xi(N_n) - S_n(t))dt\right] \to \mathbb{E}[\alpha p \hat{\tau}^* - \mu\bar{\tau}\hat{\tau}^*] = 0 \text{ as } n \to \infty. \tag{110}$$

Returning to (109), consider the term

$$\sqrt{n}\left(\frac{1}{n}\int_H^{\bar{\tau}} (\Xi(N_n) - S_n(t))dt - \int_H^{\bar{\tau}} (\alpha p - \mu t)dt\right) = \sqrt{n}\left(\int_H^{\bar{\tau}} (\hat{\Xi}_n(\alpha) - \hat{S}_n(t))dt\right).$$

Figure 1. Expected queue length comparisons at different population sizes and at different overload conditions when traffic is scheduled using the fluid- and diffusion-scale AO schedules. (a) Small overload with $n = 10$. (b) Small overload with $n = 100$. (c) Large overload with $n = 10$. (d) Large overload with $n = 100$.

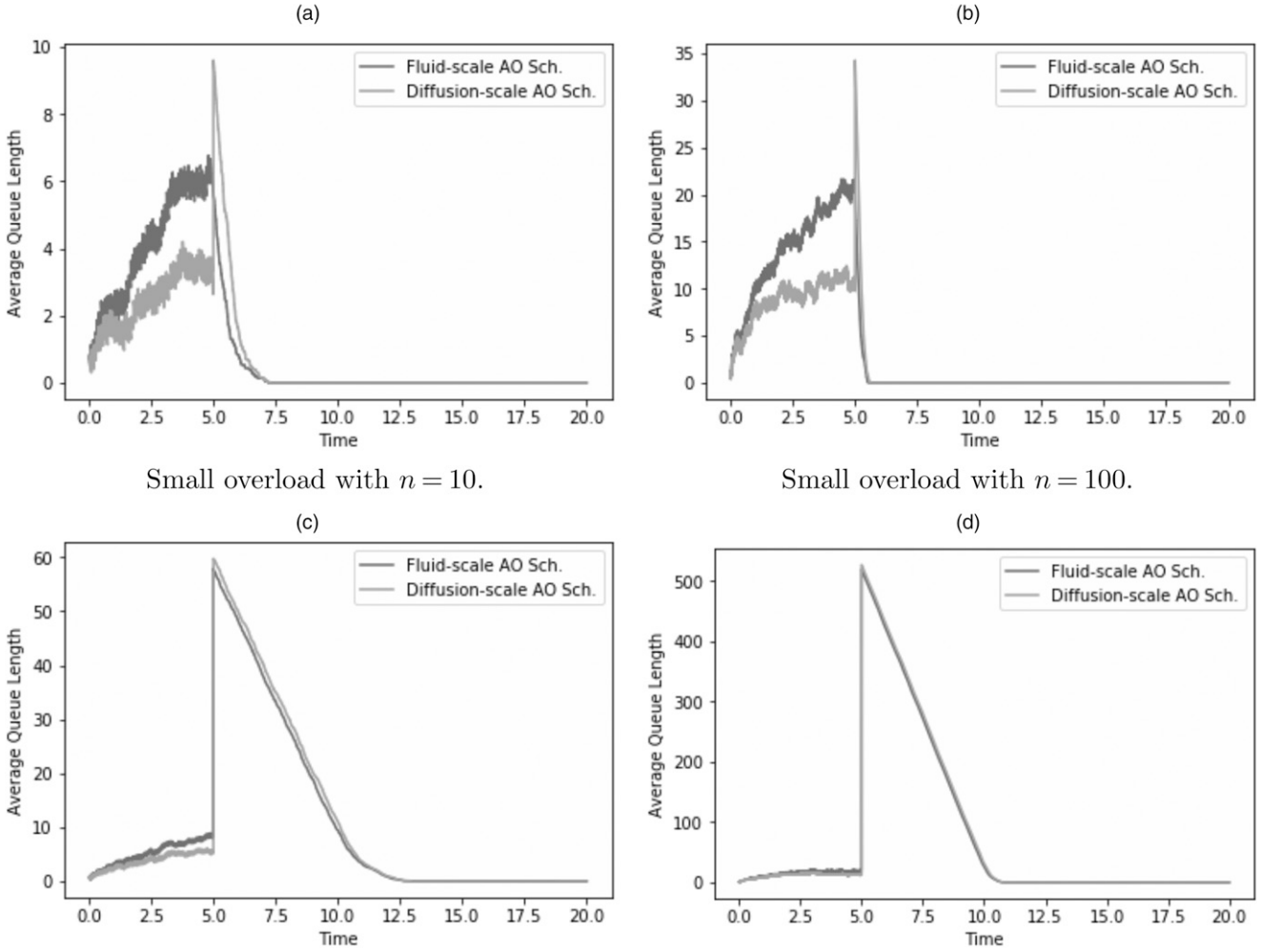


Table 1. Estimated expected costs of the fluid- and diffusion-scale AO schedules.

H		$n = 10$	$n = 100$
1.0	Fluid scale	1.8144	0.9597
	Diffusion scale	1.7368	0.9895
5.0	Fluid scale	18.8153	9.818
	Diffusion scale	15.7282	8.7902
10.0	Fluid Scale	64.7717	27.0111
	Diffusion scale	47.0690	23.067
25.0	Fluid Scale	210.1238	116.3872
	Diffusion scale	144.7316	99.5102
50.0	Fluid Scale	499.5357	339.8137
	Diffusion scale	343.1662	304.6898

It is straightforward to see that this sequence of RVs converges weakly to $\int_H^{\bar{\tau}}(X^{(1)}(\alpha) - X^{(2)}(t))dt$. Furthermore, it is again true that the prelimit integral is uniformly integrable, implying that as $n \to \infty$

$$\mathbb{E}\left[\sqrt{n}\left(\int_H^{\bar{\tau}}(\hat{\Xi}_n(\alpha) - \hat{S}_n(t))dt\right)\right] \to \mathbb{E}\left[\int_H^{\bar{\tau}}(X^{(1)}(\alpha) - X^{(2)}(t))dt\right] = 0. \tag{111}$$

From (108) and (111), it follows that $\hat{\gamma}^{CI}_{n,H} \to 0$ as $n \to \infty$ for all $H > 0$. Q.E.D.

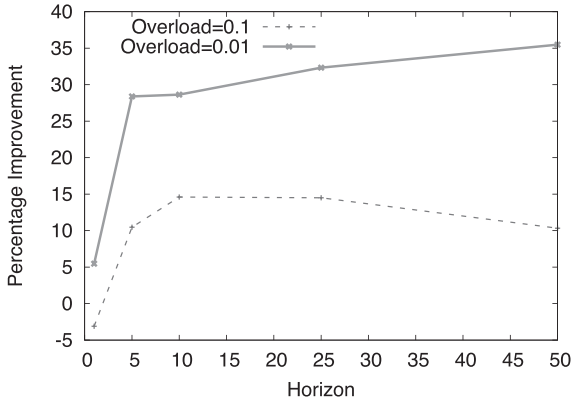
5. Numerics

We illustrate the analytical results obtained in the prior sections with a few simulation results. Recall that we focus on optimization problems in which the overload condition $p\alpha > \mu H$ holds. In the simulations, we consider a "small" overload condition in which $p\alpha$ is barely larger than μH and a "large" overload condition in which $p\alpha$ is appreciably larger than μH.

Figure 1 reports the expected queue length (computed from 30 Monte Carlo repetitions) when traffic is scheduled using the fluid and diffusion AO schedules (respectively) in Equations (21) and (26) (respectively). We set the service rate $\mu = 1.0$ with exponentially distributed service times, the horizon H to 5.0, the no-show probability to $1 - p = 0.2$, and we vary α to be $5.01/p$ in the small-overload case and $10.0/p$ in the large-overload case (observe that $\mu H = 5.0$). First, compare the figures longitudinally: the first row depicts the small-overload experiment and the second the large-overload one. We make the following *qualitative* observations.

One can immediately note that, in the small-overload case, using the diffusion-scale AO schedule in (26), the average queue length is closer to zero in the interval $[0, H]$ with a large increase in queue length at $H = 5.0$, just as predicted by the (fluid) optimal schedule. On the other hand, as expected, the fluid-scale AO schedule tends to schedule more jobs in $[0, H]$, thereby increasing wait times, but it also has fewer overage jobs. Note that the overage time is decreased when $n = 100$ because the service rate has been accelerated by n. In the large-overload case, we observe that there is minimal qualitative difference between the fluid- and diffusion-scale AO schedules with the latter scheduling marginally fewer jobs in $[0, H)$.

Figure 2. Percentage improvement using diffusion-scale AO schedule.



The results in Table 1 and Figure 2 show that the diffusion-scale AO schedule affords substantial improvements in terms of expected cost over the fluid AO schedule. The table displays the estimated expected scaled cost under the overload condition $p\alpha = 1.1H > \mu H$ (because $\mu = 1.0$) for $n = 10$ and $n = 100$. Figure 2 shows the percentage improvement in the expected cost of using the diffusion-scale AO schedule over that of the fluid-scale AO schedule.

Two important points stand out from these displays: First, Table 1 demonstrates that the diffusion-scale AO schedule shows a marked improvement over using the fluid-scale AO schedule when the horizon is large for a fixed overload condition. Second, the "degree" of overload is critical; as Figure 2 demonstrates, the diffusion-scale AO schedule has an improvement of between 10% and 15% over the fluid-scale AO schedule when the overload condition satisfies $p\alpha - \mu H = 0.1$ (as in Table 1). On the other hand, when the overload condition is smaller at 0.01, we observe that the diffusion-scale AO schedule is *always* better (over the chosen range of horizons), and furthermore, the improvement ranges between 30% and 35% for larger horizon lengths. In general, one would expect that a system operator would prefer a small overload, implying that the system is not excessively undercapacitated, and in this case, the diffusion-scale AO schedule is appropriate.

6. Conclusions

Exact solutions for the optimal scheduling problem studied in this paper are intractable in general. The analytical results provide the first rigorously justified approximate solutions to this problem in the large population limit. We have taken the approach of first formulating an optimization problem that is expected to govern the asymptotics in the respective scales (based on formal limits), then solving it, and finally proving that the value of these limit problems indeed gives the limit of the values for the rescaled scheduling problems. A by-product of the last step, which is important in its own right, is to derive asymptotically optimal schemes for the prelimit scheduling problems.

It is customary to distinguish a *control* problem, in which online information on the state of the system is available to the decision maker, from an *optimization* problem, in which decisions are made at the initial time. The three-step approach alluded to is well established in work on control in asymptotic regimes, specifically in the heavy-traffic literature, but it is less studied in optimization problems. As far as the authors know, an optimization problem involving diffusion of the type of the BOP we have formulated has not been considered before in relation to heavy-traffic applications or in the context of solving stochastic programs without recourse. It seems that versions of this problem might be relevant in applications far beyond the present model.

Although the FOP is easy to solve, we have not been able to find an explicit solution to the BOP over a time interval of finite horizon. As we have mentioned, the BOP is a convex optimization problem, and thus, it is plausible that one could treat it via numerical schemes; this is left for future work. However, one of our main findings is that, when set on an infinite time horizon, this BOP is solvable explicitly. Its solution, in the form of a reflected BM with constant drift, is rather simple. The form of the optimal drift captures the trade-off between the two parts of the cost.

Another main ingredient of our work is the notion of an SG, which we have introduced as a means of quantifying the performance loss resulting from the inherent stochasticity in the model as compared with the complete information problem. As one may expect, we have shown that the gap converges to zero in the fluid limit but remains positive in the diffusion limit. It is natural to associate this to (but it certainly does not automatically follow from) the fact that the FOP is a deterministic problem, whereas the BOP is stochastic. Moreover, our calculation of the gap in the diffusion limit shows that it is proportional to the diffusion coefficient σ. Thus, the loss in performance resulting from stochasticity is proportional to the standard deviation of the underlying noise. It is also interesting to note that the CI problem can be viewed as a single-stage stochastic program *with* recourse (Shapiro et al. [25]) because the optimization is conducted after the stochastic values are revealed to the decision maker. The SG, thus, provides a useful measure of the impact of recourse on such problems.

A possible source of uncertainty not accounted for in this work but that is important in practice is that of nonpunctual arrivals. This aspect may be addressed in future work. Moreover, the analysis has focused exclusively on a single-server queue, and consequently, the limit-optimization problems are one-dimensional. Appointment scheduling in multiqueue networks is natural to consider next as the limit problems are concerned with multidimensional diffusion processes that are constrained to lie within a quadrant or, more generally, a cone.

We also note that admission control is often used to manage customer scheduling, especially with walk-ins. Walk-ins are not considered in the current model, and incorporating them requires a reworking of our current model. In particular, admission control is more of a real-time/stochastic optimal control problem. It is

interesting to note that walk-ins also provide a means of "recourse" in this setting as they give the service system more flexibility in how many customers to schedule at each time instant. We leave this to future study.

Acknowledgments

The authors thank Assaf Zeevi for the idea behind the proof of Lemma 9. They also thank an associate editor and two referees for numerous valuable suggestions and comments that have significantly improved this paper.

References

[1] Araman VF, Glynn PW (2012) Fractional Brownian motion with $H < 1/2$ as a limit of scheduled traffic. *J. Appl. Prob.* 49(3):710–718.
[2] Avriel M, Williams A (1970) The value of information and stochastic programming. *Oper. Res.* 18(5):947–954.
[3] Benjaafar S, Jouini O (2011) Queueing systems with appointment-driven arrivals, non-punctual customers, and no-shows. Technical report, University of Minnesota, Minneapolis.
[4] Billingsley P (1968) *Convergence of Probability Measures* (John Wiley & Sons, New York).
[5] Birge JR (1982) The value of the stochastic solution in stochastic linear programs with fixed recourse. *Math. Programming* 24(1):314–325.
[6] Brown DB, Haugh MB (2017) Information relaxation bounds for infinite horizon Markov decision processes. *Oper. Res.* 65(5):1355–1379.
[7] Brown DB, Smith JE, Sun P (2010) Information relaxations and duality in stochastic dynamic programs. *Oper. Res.* 58(4-part-1):785–801.
[8] Cayirli T, Veral E (2003) Outpatient scheduling in health care: A review of literature. *Production Oper. Management* 12(4):519–549.
[9] Cayirli T, Veral E, Rosen H (2006) Designing appointment scheduling systems for ambulatory care services. *Healthcare Management Sci.* 9(1): 47–58.
[10] Chen H, Yao DD (2001) *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization* (Springer-Verlag, New York).
[11] Dempster MAH (1981) The expected value of perfect information in the optimal evolution of stochastic systems. Arató M, Vermes D, Balakrishnan AV, eds. *Stochastic Differential Systems*, Lecture Notes in Control and Information Sciences, vol. 36 (Springer, Berlin, Heidelberg), 25–40.
[12] Durrett R (2010) *Probability: Theory and Examples* (Cambridge University Press, Cambridge, UK).
[13] Ethier SN, Kurtz TG (2009) *Markov Processes: Characterization and Convergence*, vol. 282 (John Wiley & Sons, New York).
[14] Gupta D, Denton B (2008) Appointment scheduling in health care: Challenges and opportunities. *IIE Trans.* 40(9):800–819.
[15] Hall RW (2012) *Handbook of Healthcare System Scheduling* (Springer, New York).
[16] Harrison JM (1985) *Brownian Motion and Stochastic Flow Systems* (John Wiley & Sons, New York).
[17] Hassin R, Mendel S (2008) Scheduling arrivals to queues: A single-server model with no-shows. *Management Sci.* 54(3):565–572.
[18] Honnappa H, Jain R, Ward AR (2015) A queueing model with independent arrivals, and its fluid and diffusion limits. *Queueing Systems* 80(1–2):71–103.
[19] Honnappa H, Jain R, Ward AR (2016) On transitory queueing. Preprint arXiv:1412.2321, submitted December 7, https://arxiv.org/abs/1412.2321.
[20] Kim SH, Whitt W, Cha WC (2018) A data-driven model of an appointment-generated arrival process at an outpatient clinic. *INFORMS J. Comput.* 30(1):181–199.
[21] Krichagina EV, Taksar MI (1992) Diffusion approximation for $GI/G/1$ controlled queues. *Queueing Systems* 12(3):333–367.
[22] Kuiper A, Mandjes M, de Mast J (2017) Optimal stationary appointment schedules. *Oper. Res. Lett.* 45(6):549–555.
[23] LaGanga LR, Lawrence SR (2012) Appointment overbooking in health care clinics to improve patient service and clinic performance. *Production Oper. Management* 21(5):874–888.
[24] Mitter SK (2008) Convex optimization in infinite dimensional spaces. Blondel VD, Boyd SP, Kimura H, eds. *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, vol. 371 (Springer, London), 161–179.
[25] Shapiro A, Dentcheva D, Ruszczyński A (2009) *Lectures on Stochastic Programming: Modeling and Theory* (SIAM, Philadelphia).
[26] Zacharias C, Pinedo M (2014) Appointment scheduling with no-shows and overbooking. *Production Oper. Management* 23(5):788–801.