informs
https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH

Vol. 00, No. 0, Xxxx 0000, pp. 000–000

Submitted to *Operations Research*

ISSN 0030-364X, EISSN 1526-5463

QED Appointments

Junfei Huang
Department of Decisions, Operations and Technology, CUHK Business School, The Chinese University of Hong Kong, Shatin, Hong Kong, junfeih@cuhk.edu.hk

Avishai Mandelbaum
Faculty of Data and Decision Sciences, Technion – Israel Institute of Technology, Haifa, Israel, avim@technion.ac.il

Petar Momčilović
Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX, U.S.A., petar@tamu.edu

Abstract. There is a vast world of service-systems that are driven by appointments (reservations). Many of these systems are large, time-varying, and stochastic, which motivates the present study. Specifically, we optimize appointments to such a service system with soft capacity (e.g., an amusement or a nature park), with penalties for over- and under-capacity utilization ("newsvendor" tradeoffs). As capacity grows, we offer an asymptotically-optimal two-stage solution, first of a macro fluid model and then its diffusion refinement. This solution alternates among three operational regimes, which are temporally determined by the fluid model: under-loaded when the fluid level is below capacity, over-loaded when above capacity, and critically-loaded when fluid equals capacity. In the latter regime, the over- vs. under- tradeoff is ameliorated by exploiting economies-of-scale via square-root safety-appointments. Borrowing the terminology of many-server queueing theory, the operation is then Quality- and Efficiency-Driven (QED).

Funding: AM was supported by ISF (Israeli Science Foundation) 491/22.

Key words: Appointment scheduling, QED regime, Many-server or large-capacity asymptotics, Fluid- and diffusion-models, Infinite-server and time-varying queues, Infinite-dimensional Newsvendor.

1. Introduction

Our daily routines have become dependent on large stochastic service systems, and these have increasingly become driven by appointments – e.g., of patients to doctors and other medical resources; of travelers to park-trails; of last-mile deliveries to customers; and of plaintiffs and lawyers within the justice system. Indeed, the COVID era has exacerbated the appointment trend by often changing pre-COVID unscheduled services into "by-appointment-only"; or, at the very

least, by profoundly affecting delivery-channels – e.g., tele-medicine supplementing, and sometimes replacing face-to-face procedures.

Appointment scheduling processes are typically complex; hence, no single model can come close to capturing their many "dimensions". Models must hence tradeoff complexity vs. applicability by focusing on specific features and compromising on others, which then scopes model relevance.

1.1. Features of the Present Model

Motivated by applications, our focus are *appointment-driven, large-scale*, and *stochastic* systems, for which we offer some theoretical seedlings (Huang et al. 2022). In our model, customers (appointees) arrive for service of random duration, which commences upon arrival and after which they leave. The number of customers in service is the *census process*, for which a *soft* target is given – soft in that it can be deviated from, both over and under, but at a cost. We seek to design a (deterministic) appointment schedule/plan that minimizes expected deviation costs and, in concert with our focus, this gives rise to a model with the following features:

• *Soft occupancy constraints with newsvendor-type costs.* We balance newsvendor costs of *overage* (too many appointments hence excessive customer-waiting) vs. *underage* (too few hence excessive capacity-loss), which are measured relative to a soft constraint on target occupancy. (Appointment decisions at different times are dependent through appointees' positive durations of service.)

• *Transient, time-varying behavior.* Costs and target occupancy are time-varying. This renders our system transient, which leads to an infinite-dimensional newsvendor problem: appointment schedules, optimized over, are functions of time (e.g., during a single business day).

• *Large-scale system, in particular QED.* The target occupancy is large – this is captured by asymptotic analysis as target grows, specifically first-order (fluid) analysis and second-order (diffusion) refinement. In specific circumstances (9), optimal appointment costs are proportional to the square-root of occupancy – we refer to this case as Quality- and Efficiency-driven (QED, Section 4.1), due to analogies with QED staffed many-server queues (Section EC.8).

• *Stochastic behavior.* Punctuality (relative to appointments) and service durations (including no-shows) are random. Their distribution functions are model primitives, on which optimal appointment schedules must depend. This captures systems with high-variability of service durations (e.g., chemotherapy, in contrast to time-slotted models where service durations are deterministic.)

• *Time-horizon is* $\mathbb{R}$*,* to accommodate infinite-support distributions, as they are theoretically-tractable (exponential services with unbounded durations; see Example 1) and practically relevant

(Laplace punctuality, with unbounded earliness; see (Mandelbaum et al. 2020, Fig. 1)). The fact, that appointment systems typically operate over finite time-horizon, is captured by (the weaker assumption of) finite L^1-norm of the target function (10).

- *Off-line optimized total number of appointments.* The total number of appointments is a decision variable. This flexibility is fundamental, as it enables optimality of QED dynamics (Sections 3.3 and 4.1). It is natural in offline settings (e.g., design of appointment books), where targets are apriori given (in contrast to having appointments being added and canceled online).

1.2. Appointment Optimization: Fluid- and Diffusion-Reduction, Asymptotically

We now offer a broad view of our approach, using the minimal formalism necessary. Let $\mathcal{A} = \{\mathcal{A}(t) \geq 0 : t \in \mathbb{R}\}$ be a counting function of appointment times on $\mathbb{R}$, the real line: $\mathcal{A}(t)$ is the number of appointments by time $t \in \mathbb{R}$. We seek to minimize the deviations of the (stochastic) census process $X_{\mathcal{A}}(\cdot)$, induced by a (deterministic) appointment plan $\mathcal{A}(\cdot)$, from a desired goal function $\gamma(\cdot)$:

$$\min_{\mathcal{A}} C(X_{\mathcal{A}} - \gamma), \tag{1}$$

where C is a newsvendor-like cost (4). Denote by G and F the distribution functions of service duration and punctuality, respectively. When the goal γ is "large" (hence, $X_{\mathcal{A}}$ is to be "large"), we show that, at all $t \in \mathbb{R}$, it is asymptotically optimal to substitute $X_{\mathcal{A}}$(t) by its normal approximation, having mean $\tilde{G} * F * \mathcal{A}(t) := F * \mathcal{A}(t) - G * F * \mathcal{A}(t)$ and variance $\Gamma^2_{\mathcal{A}}(t) = (\tilde{G} * F - (\tilde{G} * F)^2) * \mathcal{A}(t)$; here $*$ denotes the convolution operator over $\mathbb{R}$. Moreover, when $X_{\mathcal{A}}$ is "large", its mean $\tilde{G} * F * \mathcal{A}$ and standard-deviation $\Gamma_{\mathcal{A}}$ are on different scales ($\tilde{G} * F * \mathcal{A} \geq \Gamma^2_{\mathcal{A}}$). We refer to these scales as fluid-scale (mean) and diffusion-scale (standard deviation), and their differing magnitudes enable the following decomposition of the intractable problem (1) into two tractable problems:

- *Fluid problem:* First, fluid scale (Section 3) captures predictable (deterministic) variability, for which one minimizes the deviation between the mean of $X_{\mathcal{A}}$ and the target:

$$\bar{\mathcal{A}}_* := \arg\min_{\bar{\mathcal{A}}} C(\tilde{G} * F * \bar{\mathcal{A}} - \gamma),$$

where the optimization is over a class of non-decreasing functions $\bar{\mathcal{A}}$.

- *Diffusion problem:* Second, diffusion scale captures stochastic fluctuations relative to the fluid, for which we minimize (Section 4):

$$\widehat{\mathcal{A}}_* := \arg\min_{\widehat{\mathcal{A}}} C(\tilde{G} * F * (\widehat{\mathcal{A}} + \bar{\mathcal{A}}_*) + \Gamma_{\bar{\mathcal{A}}_*} Z - \gamma),$$

where Z is standard normal, and optimization is over general functions $\widehat{\mathcal{A}}$, yet with additional constraints that ensure "consistency" across fluid- and diffusion-scales (Definition 1).

- *Originating problem:* Finally, a solution to (1) is obtained from the two solutions, $\bar{\mathcal{A}}_*$ and $\widehat{\mathcal{A}}_*$, via $\bar{\mathcal{A}}_* + \widehat{\mathcal{A}}_*$ and an operator (16) that projects this sum (which need not be non-decreasing) onto the space of appointment plans.

The diffusion refinement is relevant when the fluid cost is negligible (Section 3), and the difference $\tilde{G} * F * \bar{\mathcal{A}}_* - \gamma$ is on the smaller diffusion scale – in this case, asymptotic optimality entails square-root safety-appointments, and the system operates in the QED regime (Section 4.1 and EC.8).

1.3. Focused Literature Review

The literature on appointment scheduling is too vast for reviewing here. Indeed, in healthcare by itself (the application area of our preceding Mandelbaum et al. (2020)), the subject has enjoyed multiple comprehensive reviews (Ahmadi-Javid et al. 2017, Berg and Denton 2012, Dantas et al. 2018, Cayirli and Veral 2003, Lamé et al. 2016, Gupta and Denton 2008, Hall 2012, Saville et al. 2019, Zhao et al. 2017). Necessarily then, we focus on three threads that are directly relevant to our research or help assess its contributions: appointment-driven queues, time-varying dynamics, and QED service systems. And even here, one must further narrow down to stochastic service durations (as distinguished from time-slotted (deterministic) durations – the simplicity of which renders tractable a richer scope of features; e.g., as in Gocgun and Puterman (2014) and Liu et al. (2019), via MDPs).

Appointment-driven queues. Operationally viewed as a process, appointment-driven services are queueing systems in which appointments are arrivals and resources are servers; randomness arises, among other things, from service durations, no-shows, punctuality, and variability within appointment-books (Kim et al. 2018). Research on queues controlled by appointments started already in the 1960s (Jansson 1966). The references in Kemper et al. (2014), Luo et al. (2015), Zhou et al. (2021), and Kuiper et al. (2021) jointly offer complementing sources on the relevant body of research. We now single out two themes of that research: multiple servers and asymptotic analysis.

Multi-server systems. Appointment scheduling to multi-servers is analytically less tractable and hence much less understood than the single-server counterparts. Some attempts to address this gap

employ a decomposition approach (e.g., Deng and Shen (2016) or Chatterjee et al. (2025)): customers are first routed to servers, and then appointments are assigned to customers via solving a single-server problem. While such an approach might be justified in certain applications, the decomposition does away with server pooling, which is the main reason for the efficiency of multi-server systems. For example, Kuiper and Lee (2022) study a multi-server appointment system with phase-type service durations. Based on a numerical study of up to four parallel servers, they conclude that significant performance gains can be achieved when servers are pooled. Then, an operation with two to six servers is studied in Soltani et al. (2019), which requires a fit to 3-valued service duration. Finally, Zacharias and Pinedo (2017) and Shnits et al. (2020) study multiple-servers and are simulation-based.

The need to accommodate multiple servers, specifically developing appointments for 10s of infusion beds at a large cancer center (around 1000 patients per day), arose in Mandelbaum et al. (2020); in fact, our present paper provides some theoretical support for the practically successful algorithm in Mandelbaum et al. (2020), which is rooted in ample-server approximation. The heuristic based on this approximation enabled a data-driven creation of high-quality appointment schedules. (Note that this heuristics corresponds to Borst et al. (2004) (optimizing staffing), rather than merely Halfin and Whitt (1981) (QED performance).)

Asymptotic analysis. Here, we are aware of only two relevant papers: Armony et al. (2019) and Lipscomb et al. (2024). The problem in Armony et al. (2019) is offline appointment scheduling, of an overloaded single-server in a large-population asymptotic framework, over a given planning horizon. Overloading is ensured by determining apriori the number of appointees, as well the appropriate cost structure. As in our work, appointment optimization is, asymptotically, first reduced to a fluid optimization problem, and then to its diffusion refinement; the diffusion problem is explicitly solved for the long-horizon limit. Note that Armony et al. (2019) analyze a large appointment system, where "large" is there the fixed processing capacity of a *single* server. Our model, in contrast, is intrinsically transient (capacity and costs are time-varying), and the number of appointees is to be optimally determined. Both distinctions are fundamental; in particular, the latter flexible number of appointees enables QED performance – more on that in the sequel. The second relevant paper is Lipscomb et al. (2024), who analyze a FIFO single-server system with time-dependent punctuality. The authors consider a heavy-traffic fluid approximation to formulate a fluid control problem, the solution of which was used to derive asymptotically optimal appointment schedules for finite-population models. A scheme for numerically solving this fluid control problem was proposed as well.

Transient/time-varying queues. Intraday appointments are mostly scheduled over finite time-horizons. Their models are therefore naturally transient and hence their analysis challenging. Specifically, one must optimize over functions, namely, appointment schedules; this is in contrast to scalars, namely, long-term appointment rates.

Analysis of time-varying queues goes back as early as Palm (1988) in telecommunication and Newell (1968a,b) in transportation. The papers Whitt (2018), Defraeye and Van Nieuwenhuyse (2016), and Schwarz et al. (2016) review the subject, but a definitive text-book treatment has been lacking. This is in stark contrast to classical queueing theory, which perhaps underscores the challenges in analyzing transience. Of the latter, two lines of research bear direct relevance to ours. First is the temporally-alternating operational regime, as in Liu and Whitt (2012a) and Mandelbaum and Massey (1995); specifically, our Figure 1 depicts a QED period $(-\infty, h]$, changing to QD in $(h, T]$, and ending with ED over (T, ∞). Second is the central role played by the time-varying offered-load, which is calculated in Eick et al. (1993), and used as a skeleton of QED server-staffing in Feldman et al. (2008) and Liu and Whitt (2012b); specifically, the skeleton analogue for customer-appointments is our offered-capacity, which is the solution $\bar{\mathcal{A}}$ of equation (8).

QED service/queueing systems. Queueing theory of many-server models (Whitt 2013, van Leeuwaarden et al. 2019) supports server-staffing (given unscheduled arrivals). Its evolution will help assess the status of appointment theory, although the latter focuses on customer scheduled arrivals (given staffing) – see the first paragraph of the next section.

Inspired by large-scale (unscheduled) services, such as call centers and emergency departments, staffing research has been grounded in asymptotic theory as the number of resources increases indefinitely. One of its important outcomes is operational recipes for dimensioning large systems so that they are *both* Quality- and Efficiency-Driven (QED); specifically square-root staffing (see §2.1 in Borst et al. (2004)) that goes back to Erlang (Erlang 1948), who derived it via marginal analysis of the benefit in adding a server. Later, the principle was discussed and analyzed in Grassmann (1988) and then in Kolesar and Green (1998), where both its accuracy and applicability were substantiated. However, a rigorous mathematical justification had to await the seminal asymptotic analysis in Halfin and Whitt (1981), which enabled the practical insights in Whitt (1992). All the research prior to Halfin and Whitt (1981) (except for Erlang's) applied the so-called *infinite-server heuristics*. Following Halfin and Whitt (1981), Borst et al. (2004) offered a framework for optimizing many-server staffing (minimizing staffing and waiting costs or satisfying constraints, again asymptotically) for QED systems, which is robust enough to cover also regimes that are only Quality- or only Efficiency-Driven.

1.4. Contributions

To begin, it is informative to draw a parallel with research on unscheduled arrivals. Early QED research (*"pre-Halfin and Whitt (1981)"*) focused on ample-server approximations of finite-server systems that, in particular, gave rise to staffing heuristics for the latter. This motivated the large body of research on many-server systems, both methodological (as reviewed in van Leeuwaarden et al. (2019)) and empirical (starting in Brown et al. (2005)); in particular, it paved the way to asymptotic staffing-theory that is centered around the offered-load (Whitt 2013). Analogously, our paper is a seedling for future research on appointments to many-server systems; in particular, for asymptotic appointment-theory that will be centered around the *offered-capacity* ($\bar{\mathcal{A}}$ in (8)). As for present contributions, the main ones – practical, theoretical, and technical – can be summarized as follows:

• *Square-root safety appointments*: A *square-root* rule for appointments to *ample servers* is developed (Section 4.1). It guarantees QED performance (Section EC.8) and supports appointment-book (template) design in offline "intraday" scheduling. While infinite-server heuristics can be applied successfully to multiple-server appointment systems (as demonstrated in Mandelbaum et al. (2020)), ample-server models could, in their own right, be the natural models of some appointment-driven large-scale services (e.g., natural- or amusement- parks, as they operate typically under soft occupancy constraints, and visitors' service durations are random; see Leon (2024)).

• *Fluid solution and diffusion (QED) refinement*: An intractable problem of (stochastic) large-scale appointment scheduling is hierarchically solved via two (deterministic) optimization problems. Solutions of the first problem yield optimal appointment plans on a fluid scale (Section 3); these are then optimally refined, by solving the second problem on a diffusion scale (Section 4). In the (pure) QED case (zero fluid-cost), we offer a mechanism (16) that translates the two (deterministic) solutions into an asymptotically-optimal appointment schedule for their originating (stochastic) system (Corollary 2). The subtlety and significance of this translation are well demonstrated by the special case in Section 4.2.

• *S-convergence (as opposed to weak convergence)*: Most early research on the infinite-server model (e.g., Glynn and Whitt (1991)) and subsequent results on its related many-server queue (e.g., Krichagina and Puhalskii (1997), Puhalskii and Reed (2010), and the subsequent line of literature) are within the framework of weak convergence. By contrast, our results are based on convergence of an $\mathbb{R} \to \mathbb{R}$ norm, termed $\mathcal{S}$-convergence in Section 2.2; it is obtained from symmetrizing an L^1 semi-norm which, in turn, is induced by the cost function of the underlying newsvendor problem. To this end, we study convergence of the infinite-server process over the whole real line rather than

on compact intervals (the latter are special cases). Consequently, novel theoretical aspects must be addressed to account for tail behavior at $\pm\infty$ (e.g., see the moment condition and (17) in Lemma 2).

1.5. Notation

Let $D(I)$, $I \subseteq \mathbb{R}$, be the classical space of RCLL functions, namely, functions $x : I \to \mathbb{R}$ that are Right-Continuous and have Left-Limits. In addition, let $D_+(I) := \{x \in D(I) : x(t) \geq 0, \forall t \in I\}$, $D_\wedge(I) := \{x \in D_+(I) : x(t) \leq x(u), \text{ for } t \leq u\}$, and $D_{\mathbb{N}}(I) := \{x \in D_\wedge(I) : x(t) \in \mathbb{N} \cup \{0\}\}$. Clearly, $D_{\mathbb{N}}(I) \subseteq D_\wedge(I) \subseteq D_+(I) \subseteq D(I)$. We set $x(-\infty) = 0$ for $x \in D_\wedge(\mathbb{R})$ and $x \in D_{\mathbb{N}}(\mathbb{R})$. Define $V_\wedge(\mathbb{R}) := \{x : \mathbb{R} \to [0,\infty) : 0 = x(-\infty) \leq x(t) \leq x(u), \text{ for } t \leq u\}$.

Throughout the paper, $(\cdot)^+$ and $(\cdot)^-$ represent the positive and negative parts, respectively; $\vee$ and $\wedge$ denote the maximum and minimum operators, respectively. For a finite set S, let $|S|$ denote the number of its elements. Unless otherwise specified, convergences are with respect to our scaling parameter n increasing without a bound ($n \uparrow \infty$).

1.6. Organization

The paper is organized as follows. In Section 2, we introduce the model and our asymptotic framework. Analysis of appointment plans on the fluid scale is presented in Section 3. A diffusion-level analysis of the zero fluid-cost case (QED) is given in Section 4. Technical proofs of our results are carried out in Section 5. The Electronic Companion (Appendices) contains some auxiliary results.

2. Model, Problem Formulation, and Large-scale Framework

Consider a service system with ample (as-many-as-needed) servers, which is appointment-driven in that customers arrive by appointment-to-service (by any of the servers). Appointments (invitations, reservations, scheduled arrivals), made during the time-interval $\mathbb{R} := (-\infty, \infty)$, are formalized by a *deterministic* sequence of appointment epochs $A = \{A(i) : i = 1, \cdots, \mathcal{A}(\infty)\}$: here $A(i) \in \mathbb{R}$ is the invitation epoch of the ith appointment, thus $A(i) \leq A(i+1)$, for all i, and we assume $\mathcal{A}(\infty) < \infty$. Appointments can be also characterized by an RCLL non-decreasing (in $D_\wedge(\mathbb{R})$) function $\mathcal{A} = \{\mathcal{A}(t) : t \in \mathbb{R}\}$, where

$$\mathcal{A}(t) = \sum_{i=1}^{\mathcal{A}(\infty)} \mathbb{1}\{A(i) \leq t\}$$

counts the cumulative number of scheduled arrivals within $(-\infty, t]$. Note that $\mathcal{A}(-\infty) = 0$ does indeed hold, as there exists a first appointment. Note that *batch* appointments are allowed, that is several appointments at the same epoch; for example, it might be practically convenient to have,

say, three appointments in half-hour intervals, rather than a single appointment every ten minutes. The jump-sizes of $\mathcal{A}$'s are hence natural numbers, and the order of same-epoch appointments, in the corresponding A's, can be arbitrarily determined. With this caveat, an A and its $\mathcal{A}$ equivalently determine an offline *appointment schedule*, or synonymously *appointment plan* and, thus, both will be used as such.

Customers are not necessarily punctual. The *punctuality*, which quantifies deviation from appointment time, has cumulative distribution function (CDF) F that is supported over $\mathbb{R}$: negative values indicate early arrivals (prior to appointment), whereas positive values indicate late arrivals. Denote by ϕ_i the punctuality of the ith invited customer. Then, the actual arrival time of that customer is $A(i) + \phi_i$.

Customers either show up for their appointment or do not. For those who do, their service starts immediately upon arrival (i.e., the ample-servers assumption). The *service duration* of customers (from service-start to departure from the system) has CDF G that is supported over $[0, \infty)$. To simplify notation, no-shows will be modeled by vanishing service durations; hence, for each customer, $p = 1 - G(0)$ is the probability of showing up (and $1 - p$ is the probability of *no-show*). Denote by σ_i (with $\mathbb{E}\sigma_i < \infty$) the service duration of the ith invited customer, and assume that it is independent of the punctuality ϕ_i. Hence, the departure time of the ith invited customer is $A(i) + \phi_i + \sigma_i$, and $\varphi_i(t) := \mathbb{1}\{A(i) + \phi_i \le t < A(i) + \phi_i + \sigma_i\}$ is the indicator of the ith customer being in the system at time t. Customers are assumed independent of each other, and they are statistically identical: that is, we assume that $\{(\phi_i, \sigma_i), i \in \mathbb{N}\}$ is a sequence of i.i.d. random pairs in $\mathbb{R} \times \mathbb{R}^+$ where, furthermore, the two elements of each pair are independent.

Let $X_{\mathcal{A}} = \{X_{\mathcal{A}}(t) : t \in \mathbb{R}\}$ be the *census* process corresponding to $\mathcal{A}(\cdot)$; that is, $X_{\mathcal{A}}(t)$ is the number of customers within the system at time t. Then, its dynamics over $t \in \mathbb{R}$ is

$$\begin{aligned} X_{\mathcal{A}}(t) &= \sum_{i=1}^{\mathcal{A}(\infty)} \varphi_i(t) \\ &= \sum_{i=1}^{\mathcal{A}(\infty)} (\varphi_i(t) - \mathbb{E}[\varphi_i(t)]) \\ &\qquad + \int_{-\infty}^{\infty} (\tilde{G} * F)(t-u)\, \mathrm{d}\mathcal{A}(u) \\ &=: Z_{\mathcal{A}}(t) + (\tilde{G} * F * \mathcal{A})(t), \end{aligned} \tag{2}$$

where $\tilde{G}(t) := \mathbb{1}\{t \geq 0\} - G(t) = \mathbb{1}\{t \geq 0\}\bar{G}(t)$. Since the offline appointment schedule $\mathcal{A}$ is deterministic, $\tilde{G} * F * \mathcal{A} = \{\tilde{G} * F * \mathcal{A}(t) : t \in \mathbb{R}\}$ is also a deterministic function. By contrast, $X_{\mathcal{A}}(\cdot)$ is a stochastic process, which can be thought of as an ample-server G_t/GI/∞ queue, in which $\{A(i) + \phi_i\}$ determines its time-dependent arrivals, and $\{\sigma_i\}$ are i.i.d. service durations.

Let $\gamma = \{\gamma(t) : t \in \mathbb{R}\} \in D_+(\mathbb{R})$ represent a *congestion-goal*. This function constitutes *soft* constraints in the following sense: at all time $t \in \mathbb{R}$, it is desirable to have $\gamma(t)$ appointees within the service facility, yet it is feasible to deviate from γ though at a cost. The goal is to design an offline appointment schedule $\mathcal{A}$, which is optimal in the sense that it minimizes the gap between γ (plan) and $X_{\mathcal{A}}$ (actual). Thus we seek $\mathcal{A}_*$ such that

$$\mathcal{A}_* := \arg\min_{\mathcal{A}} C(X_{\mathcal{A}} - \gamma), \tag{3}$$

for a given cost function C:

$$C(\Delta) := \mathbb{E} \int_{-\infty}^{\infty} \left(c_u(t)\, \Delta^-(t) + c_o(t)\, \Delta^+(t) \right) \mathrm{d}t. \tag{4}$$

Here, $c_u(\cdot)$ and $c_o(\cdot)$ are given $\mathbb{R} \to \mathbb{R}^+$ functions (deterministic), where $c_u(t)$ and $c_o(t)$ quantify, respectively, per-unit underage and overage costs at time $t \in \mathbb{R}$. Correspondingly, $\Delta := X_{\mathcal{A}} - \gamma$ is the deviation process, with $\Delta^-(t)$ and $\Delta^+(t)$ being the actual (random) underage and overage at time $t \in \mathbb{R}$, respectively. We note that, under (4), the mapping $\mathcal{A} \mapsto C(X_{\mathcal{A}} - \gamma)$ is convex.

Problem (3), to the best of our knowledge, is intractable. For insight, either analytical or computational, one must hence resort to approximations, and here we focus on large systems with many appointments – this will be formalized momentarily.

2.1. The Cost Function

Our general cost C in (4) can be represented in terms of the following two building blocks: $C_u(\Delta) := \mathbb{E} \int_{-\infty}^{\infty} c_u(t)\, |\Delta(t)|\, \mathrm{d}t$ and $C_o(\Delta) := \mathbb{E} \int_{-\infty}^{\infty} c_o(t)\, |\Delta(t)|\, \mathrm{d}t$. Specifically,

$$C(\Delta) = C_u(\Delta^-) + C_o(\Delta^+), \tag{5}$$

in which we note that $C_u(\cdot)$ is an L^1-norm with respect to the product measure $\mathbb{E}(\mathrm{d}\omega) \times c_u(t)\, \mathrm{d}t$ on $\Omega \times \mathbb{R}$. More precisely, it is a norm on the vector space L^1_u of functions Δ for which $C_u(\Delta) < \infty$; and standardly, on corresponding equivalence classes. Similarly, $C_o(\cdot)$ is an L^1-norm on L^1_o, namely the functions Δ with $C_o(\Delta) < \infty$. The two L^1 spaces above are partially ordered, by the usual order of functions $\Delta_1(t) \leq \Delta_2(t)$, $t \in \mathbb{R}$. Both spaces are in fact vector lattices (Conradie 2015, Section

3), as $\Delta^+ \in L^1$ if $\Delta \in L^1$. Adding the fact that our L^1 norms are monotone ($C_u(\Delta_1) \geq C_u(\Delta_2)$ if $\Delta_1 \geq \Delta_2 \geq 0$, for example) yields that $C_u(\Delta^-)$ is an asymmetric norm over $\Delta \in L^1_u$ and, similarly, $C_o(\Delta^+)$ is an asymmetric norm over $\Delta \in L^1_o$ (Conradie 2015, Proposition 3.1).

PROPOSITION 1 **(Properties of C).** *The cost function C is an asymmetric norm on $L^1_u \cap L^1_o$, which is the vector lattice of functions Δ for which $C(\Delta) < \infty$.*

Having all our model's primitives, we now formalize the framework proposed in Mandelbaum et al. (2020): seek asymptotically optimal appointment schedules, as γ becomes large and hence gives rise to many appointments. To this end, we introduce next a convergence mode that is tailored to our cost structure, which will be then followed by our asymptotic framework.

2.2. $\mathcal{S}$-convergence in $L_{\mathcal{S}}$

The asymmetric norm C induces a norm $\mathcal{S}$ via symmetrization (Ferrer et al. 1993, Proposition 2.2):

$$\mathcal{S}(\Delta) := C(\Delta) + C(-\Delta). \tag{6}$$

It equals, in fact, to the following proper L^1 norm

$$\mathcal{S}(\Delta) = \mathbb{E} \int_{-\infty}^{\infty} (c_o(t) + c_u(t)) \, |\Delta(t)| \, \mathrm{d}t,$$

since $\Delta^-(t) + \Delta^+(t) = |\Delta(t)|$. Denote by $L_{\mathcal{S}}$ the space of functions $\mathbb{R} \to \mathbb{R}$ with finite $\mathcal{S}$-norm: $L_{\mathcal{S}}$ is thus an L^1-space and, as such, it is complete. *In the sequel, convergence in $L_{\mathcal{S}}$ will play a central role: we shall denote it by $\xrightarrow{s}$, and refer to it as $\mathcal{S}$-convergence.* In particular, $\Delta_n \xrightarrow{s} \Delta$ stands for $\mathcal{S}(\Delta_n - \Delta) \to 0$, as $n \uparrow \infty$. Note that if a sequence $\{\Delta_n\}_n \subseteq L_{\mathcal{S}}$ is $\mathcal{S}$-converging then, by completeness, its $\mathcal{S}$-limit must also be in $\mathcal{S}$. Finally, $\mathcal{S}$-equality in $L_{\mathcal{S}}$ will be denoted by $\overset{s}{=}$; that is, $\Delta_1 \overset{s}{=} \Delta_2$ stands for $\mathcal{S}(\Delta_1 - \Delta_2) = 0$.

REMARK 1 (ON NON-UNIQUENESS OF $\mathcal{S}$). One can symmetrize C via $C(\Delta) \vee C(-\Delta)$ (Conradie 2015, Proposition 3.4) as well. (We have $C(\Delta) \vee C(-\Delta) \leq \mathcal{S}(\Delta) = C(\Delta) + C(-\Delta) \leq 2\, C(\Delta) \vee C(-\Delta)$.) More fundamentally, $\mathcal{S}(\Delta_n) \to 0$ if and only if there is convergences to zero of all the following four sequences, as $n \uparrow \infty$: $C_u(\Delta_n^-)$, $C_u(\Delta_n^+)$, $C_o(\Delta_n^-)$, and $C_o(\Delta_n^+)$. Indeed, $\mathcal{S}(\Delta) = C_u(\Delta^-) + C_u(\Delta^+) + C_o(\Delta^-) + C_o(\Delta^+)$ (hence $L_{\mathcal{S}} \subset L^1_u \cap L^1_o$). It follows that $\mathcal{S}$-convergence $\xrightarrow{s}$ can be characterized by applying, to the above four asymmetric norms, *any* monotone norm on $\mathbb{R}^4$ (as all norms on $\mathbb{R}^4$ are equivalent). We chose to characterize $\mathcal{S}$-convergence via the specific norm $\mathcal{S}$ in (6) for mathematical convenience.

Our choice of the norm $\mathcal{S}$ still has a further nuance when $\gamma \equiv 0$. Then, the values of $c_u(\cdot)$ do not affect the cost $C(X_{\mathcal{A}} - \gamma)$ because $X_{\mathcal{A}} \geq 0$. It follows that $c_u(\cdot)$ can be arbitrary. However, when considering $\mathcal{S}$, the values of $c_u(\cdot)$ can affect the value of the symmetrized cost. In this case, our theory will be best served by setting $c_u \equiv 0$, in order to obtain the smallest $\mathcal{S}$ of the form (6).

2.3. Asymptotic Framework

Large scale (many appointments) is captured by a sequence of systems indexed by $n \in \mathbb{N}$, with corresponding congestion-goal $\gamma^n = \{\gamma^n(t) : t \in \mathbb{R}\}$ that grows large, as $n \uparrow \infty$. Relevant parameters and processes in the nth system will be appended with a superscript n. For example, the appointment schedule and the census process of the nth system will be denoted by $\mathcal{A}^n$ and $X_{\mathcal{A}^n}$, respectively; CDFs G and F, on the other hand, will not vary with n. The cost $C(X_{\mathcal{A}^n} - \gamma^n)$ depends on the distribution of the census process $X_{\mathcal{A}^n}$, which, in turn, is a function of the appointment schedule $\mathcal{A}^n$. In this setup, large-scale asymptotic analysis will be carried out at two levels, with the *microscopic* model above being the starting point.

At the first *macroscopic* level, we carry out fluid analysis in which $\gamma^n(\cdot) \approx n\gamma(\cdot)$, $n \uparrow \infty$, for a given function $\gamma \geq 0$; modeling-wise, n and $\gamma(\cdot)$ capture the *scale* and *shape* of capacity, respectively. Fluid analysis yields a fluid approximation for our model which, among other things, reveals multiple operational regimes: under- or over- or critically-loaded; equivalently quality-driven (QD) or efficiency-driven (ED) or quality- and efficiency-driven (QED) (van Leeuwaarden et al. 2019, Mandelbaum et al. 2020), with QED being the focus here.

At the second *mesoscopic* level (Section 4), we perform a refined diffusion analysis of $\gamma^n \approx n\gamma + \sqrt{n}\widehat{\gamma}$, $n \uparrow \infty$, for some function $\widehat{\gamma}$. This will give rise to QED appointments, centered around the fluid and refined by a square safety, which are both fluid- and diffusion-optimal.

3. Fluid Analysis

We first formulate and optimize our fluid model in Section 3.1. Asymptotic optimality, at the fluid-scale, is then established in Section 3.2. This is followed by two examples: first, Example 1 that demonstrates how fluid evolution determines operational periods; then, in Section 3.3, the case of zero fluid-cost is characterized, paving the way for diffusion analysis in Section 4.

3.1. Fluid (First-Order) Analysis

Introduce the fluid-scaled (deterministic) processes $\bar{\mathcal{A}}^n = \{\bar{\mathcal{A}}^n(t) : t \in \mathbb{R}\}$ and $\bar{\gamma}^n = \{\bar{\gamma}^n(t) : t \in \mathbb{R}\}$, given by $\bar{\mathcal{A}}^n := \frac{1}{n}\mathcal{A}^n$ and $\bar{\gamma}^n := \frac{1}{n}\gamma^n$, respectively. Note that $\bar{X}_{\bar{\mathcal{A}}^n} := \tilde{G} * F * \bar{\mathcal{A}}^n$ is also deterministic, and we refer to it as a *fluid census*. By contrast, recalling (2), $X_{\mathcal{A}^n}$ is random.)

THEOREM 1 **(Fluid continuity)**. *Suppose* $\{\gamma^n\}_n$ *is such that* $\bar{\gamma}^n \xrightarrow{s} \gamma$. *Consider a sequence* $\{\mathcal{A}^n\}_n \subseteq D_{\mathbb{N}}(\mathbb{R})$ *such that* $\sup_n \bar{\mathcal{A}}^n(\infty) < \infty$, *for which* $\tilde{G} * F * \bar{\mathcal{A}}^n \xrightarrow{s} \bar{L}$. *Then,*

$$\frac{1}{n} C(X_{\mathcal{A}^n} - \gamma^n) \to C(\bar{L} - \gamma).$$

Proof. See Section 5.1. □

REMARK 2 (WEAK CONVERGENCE OF APPOINTMENT SCHEDULES). Fluid continuity can be expressed directly via convergence of appointment schedules (rather than by censuses, as in the theorem). To wit, assume that $\{\mathcal{A}^n\}_n \subseteq D_{\mathbb{N}}(\mathbb{R})$ is such that $\bar{\mathcal{A}}^n \Rightarrow \bar{\mathcal{A}}$, for which $\bar{\mathcal{A}}(\infty) < \infty$ (weak convergence of $\{\mathcal{A}^n\}_n$ to $\bar{\mathcal{A}}$, which entails pointwise convergence at all continuity points of $\bar{\mathcal{A}}$). This weak convergence of schedules can replace the $\mathcal{S}$-convergence of censuses, and it suffices to imply $\tilde{G} * F * \bar{\mathcal{A}}^n \to \tilde{G} * F * \bar{\mathcal{A}}$ almost everywhere in $\mathbb{R}$ – see Section EC.1 for the proof. One can then rephrase Theorem 1 via weak convergence of $\{\mathcal{A}^n\}_n$, and adding to it any condition that yields $\mathcal{S}$-convergence (which is L^1-convergence) from almost-everywhere convergence: for example, $\sup_n \tilde{G} * F * \bar{\mathcal{A}}^n(t) \le B(t)$, at all $t \in \mathbb{R}$, where $\mathcal{S}(B) < \infty$; or existence of $\lim_{n\to\infty} \mathcal{S}(\mathbb{1}\{B\}\, \tilde{G} * F * \bar{\mathcal{A}}^n)$, for all measurable sets $B \subseteq \mathbb{R}$.

PROPOSITION 2 **(Fluid Optimality)**. *If* $\mathcal{S}(\gamma) < \infty$ *and*

$$\lim_{M\to\infty} \inf_{\bar{\mathcal{A}} \in V_\wedge(\mathbb{R}) : \bar{\mathcal{A}}(\infty) \ge M} C(\tilde{G} * F * \bar{\mathcal{A}}) = \infty, \tag{7}$$

then there exists $\bar{\mathcal{A}}_* \in D_\wedge(\mathbb{R})$ *that minimizes the fluid cost:*

$$C(\tilde{G} * F * \bar{\mathcal{A}}_* - \gamma) = \min_{\bar{\mathcal{A}} \in V_\wedge(\mathbb{R})} C(\tilde{G} * F * \bar{\mathcal{A}} - \gamma).$$

Proof. See Section 5.2. □

REMARK 3 (ON THE CONDITIONS OF PROPOSITION 2). The conditions of the proposition ensure that there is no (fluid) appointment schedule that leads to bounded (fluid) costs with an unbounded number of (fluid) appointments – one can then focus on fluid schedules with $\bar{\mathcal{A}}(\infty) \le M$, for some finite M, from which existence of a minimizer follows. A sufficient condition for (7) to hold is per-unit overage cost bounded away from zero: if $c_o(\cdot) \ge c > 0$, then (7) holds because $C(\tilde{G} * F * \bar{\mathcal{A}}) \ge c \int_{-\infty}^{\infty} \tilde{G} * F * \bar{\mathcal{A}}(t)\, \mathrm{d}t = c\, \mathbb{E}\sigma\, \bar{\mathcal{A}}(\infty)$. When (7) does not hold, it is possible that no (fluid) appointment schedule achieves the minimum. For example, suppose $c_o(t)$ decays for large values of $|t|$, γ is concentrated around the origin, and punctuality F has a density with infinitely

many local modes, with the modes decaying away from the origin. If modes achieve a better "fit" with the goal γ as they get further away from the origin, then scheduling more appointments further away from the origin can yield better overall costs, as the cost of all but one mode can be negligible due to low $c_o(t)$. It would then follow that the minimal cost can be approached but not attained (see Example EC.1 in Section EC.2).

The condition $\mathcal{S}(\gamma) < \infty$ stands for *both* $C(\gamma) < \infty$ and $C(-\gamma) < \infty$. The former ensures that under (7), $C(\tilde{G} * F * \bar{\mathcal{A}} - \gamma)$ increases without a bound as the number of appointments increases to infinity ($\bar{\mathcal{A}}(\infty) \uparrow \infty$); the latter upper-bounds the minimal cost, being the cost of no appointments $\bar{\mathcal{A}}(\cdot) \equiv 0$.

Note that there need not be a unique optimal solution. In general, uniqueness would depend both on properties of F and G, as well as on the support of the cost functions.

REMARK 4 (RCLL VERSIONS). Appointment plans, by definition, are functions in $D_\wedge$ (RCLL); and so are model primitives and processes: G and F as CDFs, γ by assumption, and census processes because convolution (of well-behaved functions) is RCLL-preserving. In concert, it turns out natural and useful to also have RCLL fluid plans and primitives. However, these arise as $\mathcal{S}$-limits which, being L^1-limits, need not preserve path-properties such as RCLL. One can circumvent the challenge by selecting, when justified, RCLL versions as follows:

- *RCLL fluid* γ can be assumed when $\bar{\gamma}^n \xrightarrow{s} \gamma$, and γ has both right and left limits at all times.
- *RCLL fluid scheduling-plans* can be assumed because every plan has an RCLL version (equals to it almost everywhere) with the same fluid cost.
- *RCLL fluid censuses* can be assumed to be driven by RCLL plans.

The details of justifying the above are provided in Section EC.3.

3.2. Fluid-Scale (Pre-Limit Asymptotic) Optimality

Given the existence of $\bar{\mathcal{A}}_* \in D_\wedge(\mathbb{R})$ that minimizes the fluid costs $C(\tilde{G} * F * \bar{\mathcal{A}} - \gamma)$, there is a natural choice for a sequence of appointment schedules that is asymptotically optimal on the fluid-scale: $\{\lfloor n\bar{\mathcal{A}}_* \rfloor\}_n \subseteq D_{\mathbb{N}}$. Indeed, the following lemma and its corollary provide sufficient conditions that justify this choice. These conditions will be all satisfied in all the examples, subsequent to the corollary.

LEMMA 1. *Suppose* $c_u(t) + c_o(t) \le B$ *for all* $t \in \mathbb{R}$, *and* $\mathbb{E}\sigma < \infty$. *If* $\bar{\mathcal{A}} \in V_\wedge(\mathbb{R})$ *is such that* $\bar{\mathcal{A}}(\infty) < \infty$, *then*

$$\frac{1}{n}\tilde{G} * F * \lfloor n\bar{\mathcal{A}} \rfloor \xrightarrow{s} \tilde{G} * F * \bar{\mathcal{A}}.$$

Proof. See Section 5.3. □

Remark 5 (Representation of fluid limits). If there exists $\{\mathcal{A}^n\}_n \subseteq D_{\mathbb{N}}(\mathbb{R})$ with $\sup_n \bar{\mathcal{A}}^n(\infty) < \infty$, such that $\tilde{G} * F * \bar{\mathcal{A}}^n \xrightarrow{s} \bar{L}$, then there exists $\bar{\mathcal{A}} \in D_\wedge(\mathbb{R})$ with $\bar{\mathcal{A}}(\infty) < \infty$, such that $\tilde{G} * F * \bar{\mathcal{A}} \stackrel{s}{=} \bar{L}$. In general, there could be multiple $\bar{\mathcal{A}}$'s that satisfy $\tilde{G} * F * \bar{\mathcal{A}} \stackrel{s}{=} \bar{L}$, but at least one of those $\bar{\mathcal{A}}$'s is in $D_\wedge(\mathbb{R})$ having the stated properties – see Section EC.4 for an example.

Corollary 1 **(Fluid-scale asymptotic optimality)**. *Suppose $c_u(t) + c_o(t) \leq B$ for all $t \in \mathbb{R}$, $\mathbb{E}\sigma < \infty$, and $\bar{\gamma}^n \xrightarrow{s} \gamma$. If fluid-optimal $\bar{\mathcal{A}}_* = \arg\min_{\bar{\mathcal{A}} \in D_\wedge(\mathbb{R})} C(\tilde{G} * F * \bar{\mathcal{A}} - \gamma)$ is such that $\bar{\mathcal{A}}_*(\infty) < \infty$, then $\{\lfloor n\bar{\mathcal{A}}_* \rfloor\}_n \subseteq D_{\mathbb{N}}(\mathbb{R})$ is asymptotically optimal on the fluid scale:*

$$\begin{aligned}\liminf_{n\to\infty} \frac{1}{n} C\left(X_{\mathcal{A}^n} - \gamma^n\right) &\geq \lim_{n\to\infty} \frac{1}{n} C\left(X_{\lfloor n\bar{\mathcal{A}}_* \rfloor} - \gamma^n\right) \\ &= C(\tilde{G} * F * \bar{\mathcal{A}}_* - \gamma),\end{aligned}$$

for any other sequence of appointment plans $\{\mathcal{A}^n\}_n \subseteq D_{\mathbb{N}}(\mathbb{R})$.

Proof. See Section 5.4. □

Example 1 (Box-goal: Optimal positive fluid-cost). Assume perfect punctuality ($F(t) = \mathbb{1}\{t \geq 0\}$) and exponentially distributed service durations ($\tilde{G}(t) = pe^{-\mu t}\mathbb{1}\{t \geq 0\}$). The goal function is $\gamma^n(t) = n\mathbb{1}\{t \in [0,T)\}$, $t \in \mathbb{R}$ ("box" of length T and height n over the interval $[0,T)$), for a given fixed $T > 0$. Costs are defined by $c_o(t) = c_e\mathbb{1}\{t < 0\} + c_o\mathbb{1}\{0 \leq t < T\} + c_l\mathbb{1}\{t \geq T\}$ and $c_u(t) = c_u\mathbb{1}\{t \in [0,T)\}$, where c_e, c_o, c_l, c_u are non-negative constants that represent per-unit-time costs of earliness, overage, lateness, and underage, respectively. The fluid-cost is hence given by $C(\bar{X}_{\bar{\mathcal{A}}} - \gamma) = \int_{-\infty}^0 c_e\bar{X}_{\bar{\mathcal{A}}}(t)\,\mathrm{d}t + \int_0^T c_o(\bar{X}_{\bar{\mathcal{A}}}(t) - 1)^+\,\mathrm{d}t + \int_0^T c_u(\bar{X}_{\bar{\mathcal{A}}}(t) - 1)^-\,\mathrm{d}t + \int_T^\infty c_l\bar{X}_{\bar{\mathcal{A}}}(t)\,\mathrm{d}t$, where $\bar{X}_{\bar{\mathcal{A}}} := \tilde{G} * \bar{\mathcal{A}}$ is the *fluid census-process*, and $\gamma(\cdot) = \mathbb{1}\{\cdot \in [0,T)\}$.

This fluid cost can be optimized over fluid appointment-plans $\bar{\mathcal{A}} \in D_\wedge(\mathbb{R})$. In particular, the optimal $\bar{\mathcal{A}}_*$ is given by $\bar{\mathcal{A}}_*(t) = \frac{1}{p}\left(\mathbb{1}\{t \geq 0\} + \mu(t^+ \wedge h)\right)\mathbb{1}\{h > 0\}$, where $h = T - \frac{1}{\mu}\ln\left(1 + \frac{c_l}{c_u}\right)$ (see Section EC.5 for details). Thus, it is optimal to have appointments (until time h, $0 < h < T$) only if the width of the "box" T is large relative to c_l/c_u (namely, $e^{\mu T} > 1 + c_l/c_u$); otherwise ($h \leq 0$), it is optimal to have no appointments. The *optimal* fluid census-process $\bar{X}_{\bar{\mathcal{A}}_*} = \tilde{G} * \bar{\mathcal{A}}_*$ is given by $\bar{X}_{\bar{\mathcal{A}}_*}(t) = e^{-\mu(t-h)^+}\mathbb{1}\{t \geq 0, h \geq 0\}$. It is depicted by the blue line in Figure 1 when $h > 0$. Observe that the optimal fluid model is in the critically-loaded regime for $t \leq h$ ($\tilde{G} * \bar{\mathcal{A}}_* = \gamma$: quality- and efficiency-driven, QED); in the under-loaded regime for $t \in (h,T)$, ($\tilde{G} * \bar{\mathcal{A}}_* < \gamma$: quality-driven,

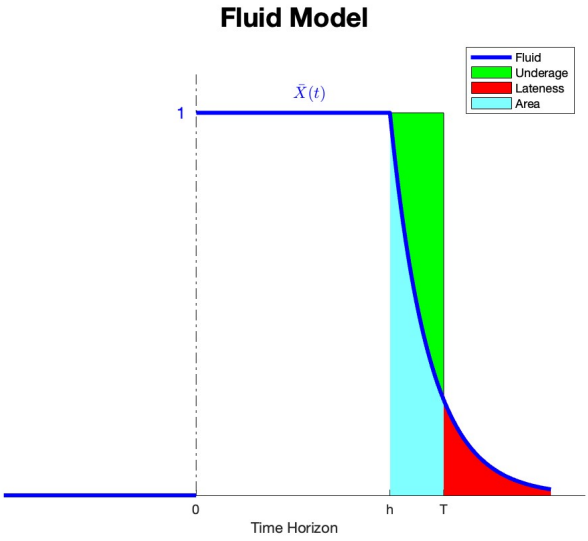

Figure 1 Optimal fluid process $\bar{X}_{\bar{\mathcal{A}}_*}$ in Example 1.

QD); and in the over-loaded regime for $t \geq T$ ($\tilde{G} * \bar{\mathcal{A}}_* > \gamma$: efficiency-driven, ED). Time h is determined by $c_u \int_h^T \bar{X}_{\bar{\mathcal{A}}_*}(t)\,\mathrm{d}t = c_\ell \int_T^\infty \bar{X}_{\bar{\mathcal{A}}_*}(t)\,\mathrm{d}t$ (see the figure), and the optimal fluid-cost is

$$\begin{aligned} C(\bar{X}_{\bar{\mathcal{A}}_*} - \gamma) &= c_u \int_h^T (1 - \bar{X}_{\bar{\mathcal{A}}_*}(t))\,\mathrm{d}t + c_\ell \int_T^\infty \bar{X}_{\bar{\mathcal{A}}_*}\,\mathrm{d}t \\ &= c_u\,(T-h). \end{aligned}$$

Consequently, Corollary 1 yields that $\{\lfloor n\bar{\mathcal{A}}_* \rfloor\}_n \subseteq D_{\mathbb{N}}(\mathbb{R})$ is asymptotically optimal:

$$\begin{aligned} \liminf_{n\to\infty} \frac{1}{n} C(X_{\bar{\mathcal{A}}^n} - \gamma^n) &\geq \lim_{n\to\infty} \frac{1}{n} C(X_{\lfloor n\bar{\mathcal{A}}_* \rfloor} - \gamma^n) \\ &= C(\tilde{G} * \bar{\mathcal{A}}_* - \gamma) = \min_{\bar{\mathcal{A}}} C(\tilde{G} * \bar{\mathcal{A}} - \gamma) \\ &= c_u(T-h), \end{aligned}$$

for any other sequence of appointment plans $\{\mathcal{A}^n\}_n \subseteq D_{\mathbb{N}}(\mathbb{R})$.

3.3. Zero Fluid-Cost

We now focus on (fluid) congestion-goals that are perfectly achievable by some fluid plan; hence, the latter is trivially fluid-optimal as it enjoys *zero fluid-cost*. Formally, such goals γ can be represented

as

$$\gamma \overset{s}{=} \tilde{G} * F * \bar{\mathcal{A}}, \tag{8}$$

for some $\bar{\mathcal{A}} \in D_\wedge(\mathbb{R})$ (equivalently $\bar{\mathcal{A}} \in V_\wedge(\mathbb{R})$). We refer to γ in (8) as a *QED goal* (elaborated in Section 4.1) and its corresponding $\bar{\mathcal{A}}$ as *offered-capacity* (explained momentarily).

Definition (8) depends on the $\mathcal{S}$-norm. For convenience, this dependency will be circumvented by adopting the stronger (pointwise) definition of QED goals:

$$\gamma \equiv \tilde{G} * F * \bar{\mathcal{A}}; \tag{9}$$

such QED goals γ are necessarily non-negative. Importantly, for given G and F, not all goals are QED: for example, the goal in Example 1 is not QED (optimal fluid cost is positive there), but the one in Example 2 below is. An explicit characterization of QED goals (γ's of the form (9)) is an open problem, although some progress has already been made (see Momčilović et al. (2022), where such goals are termed "exhaustive"). For our purpose, the following list of their properties suffices:

- *(Offered Capacity and Renewal Theory.)* Under perfect punctuality, $\gamma := \tilde{G} * \bar{\mathcal{A}}$ is the offered-load of arrivals $\bar{\mathcal{A}}$ (Whitt 2013). It plays a central role in Queueing Theory by being the skeleton for QED (square-root) staffing. Analogously (and writing formally), $\bar{\mathcal{A}} := \tilde{G}^{-1} * \gamma$ is the offered-capacity of goal γ, which will serve (next section) as the skeleton for QED appointments. The analogy of QED staffing and appointments is further substantiated in Section EC.8. The operator $\tilde{G}^{-1}$ is in fact the Renewal Operator associated with G, as (9) is in fact a renewal equation (over $\mathbb{R}$ (Karlïn 1955)): $\bar{\mathcal{A}} = \gamma + \bar{\mathcal{A}} * G$, with $\bar{\mathcal{A}}$ being the unknown. QED goals are hence the functions $\gamma \in D_+$ that are mapped by the Renewal Operator into non-decreasing functions ($\tilde{G}^{-1} * \gamma \in D_\wedge$).
- *(Continuous and Integrable Goals.)* QED goals are continuous if punctuality F is continuous. Also, if $\mathbb{E}\sigma < \infty$, then QED goal $\gamma \in L^1(\mathbb{R})$ if and only if its offered-capacity $\bar{\mathcal{A}}(\infty) < \infty$. This is a consequence of the following representation of what we define as *system capacity*:

$$\int_{-\infty}^{\infty} \gamma(t)\ \mathrm{d}t = \mathbb{E}\sigma\, \bar{\mathcal{A}}(\infty); \tag{10}$$

which, in turn, is a result of integrating (9) (and recalling $\bar{\mathcal{A}}(-\infty) = 0$): $\int_{-\infty}^{\infty} \tilde{G} * F * \bar{\mathcal{A}}(t)\ \mathrm{d}t = \iint_{-\infty}^{\infty} \tilde{G} * F(t-s)\ \mathrm{d}t\ \mathrm{d}\bar{\mathcal{A}}(s) = \iint_{-\infty}^{\infty} \tilde{G} * F(t)\ \mathrm{d}t\ \mathrm{d}\bar{\mathcal{A}}(s) = \mathbb{E}\sigma\, \bar{\mathcal{A}}(\infty)$. Equality (10) implies that for QED goals, results assuming $\bar{\mathcal{A}}(\infty) < \infty$ do prevail if $\gamma \equiv \tilde{G} * F * \bar{\mathcal{A}} \in L^1(\mathbb{R})$ is assumed instead (e.g., in Lemma 1 and Corollary 1).

• *(QED Goals via Transforms.)* The convolutions in (9) correspond to multiplications in the Fourier domain:

$$\tilde{G}^\tau F^\tau \bar{\mathcal{A}}^\tau = \gamma^\tau, \tag{11}$$

where $\tilde{G}^\tau$, F^τ, $\bar{\mathcal{A}}^\tau$, and γ^τ are the Fourier transforms of $\tilde{G}$, F, $\bar{\mathcal{A}}$ and γ, respectively:

$$\tilde{G}^\tau(\omega) = \int_{-\infty}^{\infty} e^{-i\omega t} \tilde{G}(t)\, \mathrm{d}t,\ \gamma^\tau(\omega) = \int_{-\infty}^{\infty} e^{-i\omega t} \gamma(t)\, \mathrm{d}t,$$

and

$$F^\tau(\omega) = \int_{-\infty}^{\infty} e^{-i\omega t}\, \mathrm{d}F(t),\ \bar{\mathcal{A}}^\tau(\omega) = \int_{-\infty}^{\infty} e^{-i\omega t}\, \mathrm{d}\bar{\mathcal{A}}(t).$$

Note that perfect punctuality implies $F^\tau(\omega) = 1$. Equation (11) can be sometimes used to solve for $\bar{\mathcal{A}}^\tau$ from γ^τ. Then, the corresponding $\bar{\mathcal{A}}$ can be calculated via

$$\bar{\mathcal{A}}(t) = \frac{1}{2\pi} \int_{-\infty}^{t} \int_{-\infty}^{\infty} e^{i\omega u} \bar{\mathcal{A}}^\tau(\omega)\, \mathrm{d}\omega\, \mathrm{d}u.$$

Example 2 (Modified box-goal: Zero fluid-cost). Consider Example 1 with a modified goal γ: $\gamma(t) = \mathbb{1}\{t \geq 0\} e^{-\mu(t-T)^+}$. In this case, $\tilde{G}^\tau(\omega) = p/(\mu + i\omega)$ and $\gamma^\tau(\omega) = (1 - e^{-i\omega T})/(i\omega) + e^{-i\omega T}/(\mu + i\omega)$ which, together with (11), imply $\bar{\mathcal{A}}^\tau(\omega) = 1/p + \mu(1 - e^{-i\omega T})/(i\omega p)$. Consequently,

$$\bar{\mathcal{A}}(t) = \frac{1}{p} \mathbb{1}\{t \geq 0\} + \frac{\mu}{p} \min\{t^+, T\}; \tag{12}$$

in words, on the fluid scale: there are $1/p$ appointments at time $t = 0$, and $\mu T/p$ appointments are equi-distributed over $[0, T]$. Note that the amount of "scheduled" work equals the system capacity:

$$\frac{p}{\mu} \bar{\mathcal{A}}(\infty) = \frac{1}{\mu} + T = \int_{-\infty}^{\infty} \gamma(t)\, \mathrm{d}t.$$

On the other hand, let punctuality be a symmetric Laplace distribution, namely with density $e^{-|t|}/2$, $t \in \mathbb{R}$. The corresponding Fourier transform is $F^\tau(\omega) = 1/(1 + \omega^2)$. The $\bar{\mathcal{A}}^\tau$ that solves (11) is such that its $\bar{\mathcal{A}}$ is neither non-negative nor non-decreasing. Hence, with symmetric Laplace punctuality, no fluid plan can achieve zero fluid-cost.

Example 3 (Some QED goals). All non-decreasing γ's are QED goals if punctuality is perfect ($\gamma \in D_\wedge$ implies $\tilde{G}^{-1} * \gamma \in D_\wedge$). Examples of QED goals that are also decreasing arise in Lemma 3 below. Lastly (borrowing from Momčilović et al. (2022)), an additional family of QED goals can be animated in terms of servers that "arrive" to a service system, according to some process s, where each server completes its "shift" after providing exactly k consecutive services.

Formally, consider a system with perfect punctuality and service durations distributed according to G. Suppose $\gamma := s - s * G^{*k}$ for some $s \in D_\wedge(\mathbb{R})$ and $k \geq 1$, where G^{*k} is the k-fold convolution of G with itself. In this case, $\bar{\mathcal{A}} = s * (I + G + \ldots + G^{*(k-1)})$, with $I = \{I(t) = \mathbb{1}\{t \geq 0\} : t \in \mathbb{R}\}$, achieves zero fluid-cost. Indeed, such $\bar{\mathcal{A}}$ renders $\bar{\mathcal{A}} - G * \bar{\mathcal{A}} = \gamma$. The latter can be also obtained from (11), in view of

$$\tilde{G} * (I + G + \cdots + G^{*(k-1)}) = \sum_{i=1}^{k} (-1)^{i-1} \binom{k}{i} \tilde{G}^{*i};$$

then, the desired $\bar{\mathcal{A}}$ follows from

$$\bar{\mathcal{A}}^\tau = \frac{\gamma^\tau}{\tilde{G}^\tau} = \frac{s^\tau}{\tilde{G}^\tau} (I - G^{*k})^\tau = s^\tau \sum_{i=1}^{k} (-1)^{i-1} \binom{k}{i} \left(\tilde{G}^\tau\right)^{i-1}.$$

4. Diffusion Analysis

Our starting point is Corollary 1, with goal $\gamma^n(\cdot) \approx n\gamma(\cdot)$ in which $n \uparrow \infty$ captures the goal's scale. We now develop a diffusion refinement for vanishing optimal fluid-costs, which is equivalent to having order $o(n)$ optimal fluid-costs or, in turn, having a QED fluid goal $\gamma(\cdot)$ (8). This refinement exhibits optimal cost of order-$\sqrt{n}$, which is achievable via square-root safety appointments that leads to operationally-desirable QED performance: optimal fluid census equals the goal.

In general, QED performance could be unachievable at all times, which gives rise to alternating operational regimes: ED when optimal fluid census is above the fluid-goal, and QD when below it (during non-negligible time intervals). This is manifested by positive optimal fluid cost or, equivalently, optimal cost of order-n. Now, if the goal is flexible, one could redesign it as QED, which will reduce cost to order-$\sqrt{n}$.

A case in point is Example 1 (Figure 1): as time evolves, the system is first in the QED regime (until time h), then QD (from h to T), and finally in the ED regime (after T), which incurs order-n cost. Now acknowledging that a vanishing goal is unachievable suggests redesigning it to the one in Example 2 (which could correspond to acknowledging that overtime beyond T is unavoidable and hence better be planned for): the system is then QED at all times thus improving cost to order-$\sqrt{n}$. For a general discussion beyond the examples, see Section EC.6.

4.1. QED Regime

Throughout this section, we fix a sequence of goals $\{\gamma^n\}_n \subseteq D_+(\mathbb{R})$ and a plan $\bar{\mathcal{A}} \in D_\wedge(\mathbb{R})$, for which we assume that $\widehat{\gamma}^n := \frac{1}{\sqrt{n}} (\gamma^n - n\gamma) \xrightarrow{s} \widehat{\gamma}$, for some $\widehat{\gamma} \in D(\mathbb{R})$, and γ is a QED goal (8) with respect to $\bar{\mathcal{A}}$: $\gamma \overset{s}{=} \tilde{G} * F * \bar{\mathcal{A}}$. (Note that, for such a sequence, $\frac{1}{n}\gamma^n \xrightarrow{s} \gamma$ must hold as well.)

We say that a sequence of appointment systems (namely, plans $\{\mathcal{A}^n\}_n \subseteq D_{\mathbb{N}}(\mathbb{R})$ and goals $\{\gamma^n\}_n$ as above) is a *QED sequence* if

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} C(X_{\mathcal{A}^n} - \gamma^n) \in [0, \infty); \tag{13}$$

occasionally, we refer to only the sequence of plans itself as a QED sequence, in case the goals are unambiguous. To analyze (13), we introduce the diffusion-scaled function $\widehat{\mathcal{A}}^n = \{\widehat{\mathcal{A}}^n(t) : t \in \mathbb{R}\}$, where

$$\widehat{\mathcal{A}}^n := \sqrt{n}\left(\bar{\mathcal{A}}^n - \bar{\mathcal{A}}\right) = \frac{1}{\sqrt{n}}\left(\mathcal{A}^n - n\bar{\mathcal{A}}\right); \tag{14}$$

this entails centering of diffusion-scaled plans around $\bar{\mathcal{A}}$ that achieves zero fluid-cost: $0 = \mathcal{S}(\tilde{G} * F * \bar{\mathcal{A}} - \gamma) \geq C(\tilde{G} * F * \bar{\mathcal{A}} - \gamma) \geq 0$. Assuming $\mathcal{A}^n(\infty) < \infty$, the variance function of the diffusion-scaled census (infinite-server process) $X_{\mathcal{A}^n}/\sqrt{n}$ is given by

$$\Gamma_n^2 = \left(\tilde{G} * F - (\tilde{G} * F)^2\right) * \bar{\mathcal{A}}^n;$$

note that $\Gamma_n = \{\Gamma_n(t) : t \in \mathbb{R}\}$ is determined by appointments on the fluid scale. The limiting variance function is defined analogously in terms of $\bar{\mathcal{A}}$:

$$\Gamma^2 = \left(\tilde{G} * F - (\tilde{G} * F)^2\right) * \bar{\mathcal{A}}. \tag{15}$$

THEOREM 2 **(QED continuity)**. *Suppose* $\{\gamma^n\}_n$ *is such that* $\widehat{\gamma}^n \xrightarrow{s} \widehat{\gamma}$. *Consider a sequence* $\{\mathcal{A}^n\}_n \subseteq D_\wedge(\mathbb{R})$ *such that* $\sup_n \bar{\mathcal{A}}^n(\infty) < \infty$, *and* $\tilde{G} * F * \widehat{\mathcal{A}}^n \xrightarrow{s} \widehat{L}$. *If* $\Gamma_n \xrightarrow{s} \Gamma$, *then*

$$\frac{1}{\sqrt{n}} C(X_{\mathcal{A}^n} - \gamma^n) \to C(Z\Gamma + \widehat{L} - \widehat{\gamma}),$$

where Z *is a standard normal r.v.; that is, the sequence of appointment systems* $\{\mathcal{A}^n, \gamma^n\}_n$ *is QED.*

Proof. See Section 5.5. □

REMARK 6. If $\tilde{G} * F * \widehat{\mathcal{A}}^n \xrightarrow{s} \widehat{L}$, then $\mathcal{S}(\tilde{G} * F * \bar{\mathcal{A}}^n - \tilde{G} * F * \bar{\mathcal{A}}) = \frac{1}{\sqrt{n}}\mathcal{S}(\tilde{G} * F * \widehat{\mathcal{A}}^n) \leq \frac{1}{\sqrt{n}}(\mathcal{S}(\tilde{G} * F * \widehat{\mathcal{A}}^n - \widehat{L}) + \mathcal{S}(\widehat{L})) \to 0$; that is, $\tilde{G} * F * \bar{\mathcal{A}}^n \xrightarrow{s} \tilde{G} * F * \bar{\mathcal{A}}$.

QED Optimality – Setup. Next, we create a pre-limit appointment plan $\mathcal{A}^n \in D_{\mathbb{N}}(\mathbb{R})$ from given fluid-scale $\bar{\mathcal{A}}$ and diffusion-scale $\widehat{\mathcal{A}}$ functions. The natural choice would be the rounding of $n\bar{\mathcal{A}} + \sqrt{n}\widehat{\mathcal{A}}$, but it might not be non-decreasing even though $\bar{\mathcal{A}}$ is, e.g., $\widehat{\mathcal{A}}$ might have negative jumps. For a remedy, one could consider its upper or lower envelope, but here, neither ensures $\mathcal{S}$-convergence of the corresponding census processes to the desired $\tilde{G} * F * \widehat{\mathcal{A}}$ on the diffusion scale. Hence, a more subtle mapping $(\bar{\mathcal{A}}, \widehat{\mathcal{A}}) \mapsto \mathcal{A}^n$ is required.

The first step in specifying such a mapping is to characterize points of increase of $\bar{\mathcal{A}}$. We thus say that $t \in \mathbb{R}$ is an *increase point* of $\bar{\mathcal{A}} \in D_\wedge(\mathbb{R})$ if $\bar{\mathcal{A}}(t-\epsilon) < \bar{\mathcal{A}}(t+\epsilon)$, for all $\epsilon > 0$. Denote by $\mathcal{Z}(\bar{\mathcal{A}}) \subseteq \mathbb{R}$ the set of points that are *not* increase points of $\bar{\mathcal{A}}$. The set of increase points is closed; hence $\mathcal{Z}(\bar{\mathcal{A}})$ is open, which implies that it is a countable union of disjoint open intervals: these are the time-intervals with no appointments on the fluid scale which, as it turns out, must be also identified constructively. For that, introduce an inverse of $\bar{\mathcal{A}}$: $\bar{\mathcal{A}}^{\leftarrow}(y) := \sup\{t \in \bar{\mathbb{R}} : \bar{\mathcal{A}}(t) \leq y\}$: it is right-continuous; flat intervals of $\bar{\mathcal{A}}$ correspond to jumps in $\bar{\mathcal{A}}^{\leftarrow}$, and vice versa. Thus, if $\mathcal{J}(\bar{\mathcal{A}}^{\leftarrow})$ is the set of jumps in $\bar{\mathcal{A}}^{\leftarrow}$, then $y \in \mathcal{J}(\bar{\mathcal{A}}^{\leftarrow})$ corresponds to a flat interval $\mathsf{z}_y(\bar{\mathcal{A}}) := \{t \in \mathbb{R} : \bar{\mathcal{A}}(t+\epsilon) = y \text{ for all small enough } \epsilon > 0\}$ of $\bar{\mathcal{A}}$; $\mathsf{z}_y(\bar{\mathcal{A}})$ is $(-\infty, r_y)$, $[l_y, r_y)$, or $[l_y, \infty)$ for $l_y := \inf\{t \in \mathsf{z}_y\} < r_y := \sup\{t \in \mathsf{z}_y\}$. The union of these flat intervals relates to $\mathcal{Z}(\bar{\mathcal{A}})$ via

$$\tilde{\mathcal{Z}}(\bar{\mathcal{A}}) = \bigcup_{y \in \mathcal{J}(\bar{\mathcal{A}}^{\leftarrow})} \mathsf{z}_y(\bar{\mathcal{A}}),$$

where $\tilde{\mathcal{Z}}(\bar{\mathcal{A}})$ is such that $\mathcal{Z}(\bar{\mathcal{A}}) \subseteq \tilde{\mathcal{Z}}(\bar{\mathcal{A}}) := \{t \in \mathbb{R} : (t, t+\epsilon) \subseteq \mathcal{Z}(\bar{\mathcal{A}}) \text{ for some } \epsilon > 0\}$. By contrast, $\mathcal{Z}(\bar{\mathcal{A}})$ is a union of intervals $(-\infty, r_y)$, (l_y, r_y), or (l_y, ∞).

QED Candidate. Our tool for creating pre-limit appointment processes $\mathcal{A}^n \in D_{\mathbb{N}}(\mathbb{R})$, out of fluid-scale $\bar{\mathcal{A}}$ and diffusion-scale $\widehat{\mathcal{A}}$, is a function $\Psi_n(\bar{\mathcal{A}}, \widehat{\mathcal{A}})$, to be now introduced. (The choice of Ψ_n is not unique – see Section EC.7 for an example of an alternative choice.) To this end, fix any $T_{\bar{\mathcal{A}}} > \sup\{u : u \in \mathcal{Z}(\bar{\mathcal{A}})\} \vee 0$ (with the convention that $T_{\bar{\mathcal{A}}} = \infty$ when $\sup\{u : u \in \mathcal{Z}(\bar{\mathcal{A}})\} = \infty$). Then $\bar{\mathcal{A}}$ is strictly increasing on $(T_{\bar{\mathcal{A}}}, \infty)$. Let $\mathcal{G}(\bar{\mathcal{A}}) = \mathcal{Z}(\bar{\mathcal{A}}) \cup (T_{\bar{\mathcal{A}}}, \infty)$, and introduce $\Psi_n(\bar{\mathcal{A}}, \widehat{\mathcal{A}}) = \{\Psi_n(\bar{\mathcal{A}}, \widehat{\mathcal{A}})(t) : t \in \mathbb{R}\}$ by

$$\begin{aligned}\Psi_n(\bar{\mathcal{A}}, \widehat{\mathcal{A}})(t) := &\sup_{u \leq t} \{\sqrt{n}\,\bar{\mathcal{A}}(u) + \widehat{\mathcal{A}}(u)\} \\ &\wedge \inf_{u > t:\, u \in \mathcal{G}(\bar{\mathcal{A}})} \{\sqrt{n}\,\bar{\mathcal{A}}(u) + \widehat{\mathcal{A}}(u)\}. \end{aligned} \tag{16}$$

When $\sqrt{n}\bar{\mathcal{A}} + \widehat{\mathcal{A}}$ is non-decreasing, $\Psi_n(\bar{\mathcal{A}}, \widehat{\mathcal{A}})$ reduces to $\sqrt{n}\bar{\mathcal{A}} + \widehat{\mathcal{A}}$ itself. In general, $\Psi_n(\bar{\mathcal{A}}, \widehat{\mathcal{A}})$ is always non-decreasing for any n (in its definition, both the sup and inf are non-decreasing). Importantly, $\Psi_n(\bar{\mathcal{A}}, \widehat{\mathcal{A}})$ is "close" to $\sqrt{n}\bar{\mathcal{A}} + \widehat{\mathcal{A}}$ asymptotically, in the sense of $\mathcal{S}$-convergence: $\tilde{G} * F * (\Psi_n(\bar{\mathcal{A}}, \widehat{\mathcal{A}}) - \sqrt{n}\bar{\mathcal{A}}) \xrightarrow{s} \tilde{G} * F * \widehat{\mathcal{A}}$. Note that the sup in the definition already suffices to make $\sqrt{n}\,\bar{\mathcal{A}}(u) + \widehat{\mathcal{A}}(u)$ non-decreasing, but the inf correction term is needed to ensure $\mathcal{S}$-convergence:

without it, $\Psi_n(\bar{\mathcal{A}}, \widehat{\mathcal{A}}) - \sqrt{n}\bar{\mathcal{A}}$ and $\widehat{\mathcal{A}}$ can differ asymptotically on time intervals of non-vanishing length where $\bar{\mathcal{A}}$ is non-increasing, i.e., intervals in $\mathcal{Z}(\bar{\mathcal{A}})$.

The following lemma is a diffusion-scale analogue of Lemma 1:

LEMMA 2 **(QED-optimal candidate)**. *Suppose $c_u(t) + c_o(t) \le B$, for all $t \in \mathbb{R}$, and $\mathbb{E}\sigma < \infty$. Let $\mathcal{A}^n = \lfloor \sqrt{n}\Psi_n^+(\bar{\mathcal{A}}, \widehat{\mathcal{A}}) \rfloor \in D_{\mathbb{N}}(\mathbb{R})$, where $\bar{\mathcal{A}} \in D_\wedge(\mathbb{R})$ and $\widehat{\mathcal{A}} \in D(\mathbb{R})$ are such that:*

- *$|\mathcal{J}(\bar{\mathcal{A}}^{\leftarrow})| < \infty$, $\bar{\mathcal{A}}(\infty) < \infty$, and CDF $\bar{\mathcal{A}}(\cdot)/\bar{\mathcal{A}}(\infty)$ has a finite $(1+\delta)$ absolute moment, for some $\delta > 0$;*
- *$\widehat{\mathcal{A}}(-\infty) = 0$, $\widehat{\mathcal{A}}(\infty)$ exists, $\sup_t |\widehat{\mathcal{A}}(t)| < \infty$ and*

$$\lim_{T\to-\infty} \int_{-\infty}^{T} \sup_{u\le t} |\widehat{\mathcal{A}}(u)| \, \mathrm{d}t = \lim_{T\to\infty} \int_{T}^{\infty} \sup_{u\ge t} |\widehat{\mathcal{A}}(u) - \widehat{\mathcal{A}}(\infty)| \, \mathrm{d}t = 0. \tag{17}$$

If $\mathrm{d}\widehat{\mathcal{A}}(t) \ge 0$ for all $t \in \mathcal{Z}(\bar{\mathcal{A}})$, then $\Gamma_n := ((\tilde{G}*F - (\tilde{G}*F)^2) * \frac{1}{n}\mathcal{A}^n)^{1/2} \xrightarrow{s} \Gamma := ((\tilde{G}*F - (\tilde{G}*F)^2) * \bar{\mathcal{A}})^{1/2}$, and

$$\tilde{G} * F * \frac{\mathcal{A}^n - n\bar{\mathcal{A}}}{\sqrt{n}} \xrightarrow{s} \tilde{G} * F * \widehat{\mathcal{A}}.$$

Proof. See Section 5.6. □

REMARK 7 (ON $\mathrm{d}\widehat{\mathcal{A}}(t) \ge 0$). Informally, $\widehat{\mathcal{A}}$ quantifies order-$\sqrt{n}$ refinements to the order-n appointment schedule described by $\bar{\mathcal{A}}$. A decrease of $\widehat{\mathcal{A}}$ in a neighborhood of time t corresponds to a reduction of the number of appointments around time t (deletion of appointments). However, such a reduction is feasible only if $\bar{\mathcal{A}}$ "assigns" appointments in that neighborhood, namely, $\bar{\mathcal{A}}$ is strictly increasing locally – hence, the condition $\mathrm{d}\widehat{\mathcal{A}}(t) \ge 0$, for $t \in \mathcal{Z}(\bar{\mathcal{A}})$. Indeed, for $t \notin \mathcal{Z}(\bar{\mathcal{A}})$, $\bar{\mathcal{A}}$ assigns appointments to an arbitrarily small neighborhood of t, which, due to the difference in scales (n vs. $\sqrt{n}$), is sufficient to compensate for the decrease in $\widehat{\mathcal{A}}$ at t, in the limit as $n \to \infty$. That is, the decrease in $\widehat{\mathcal{A}}(t)$ results in fewer appointments around time t than suggested by $\bar{\mathcal{A}}$. A similar phenomenon has been observed in the literature on skills-based-routing ("instantaneous routing") and stochastic control ("instantaneous displacement").

REMARK 8 (STRUCTURE OF $\widehat{\mathcal{A}}$ UNDER LEMMA 2). When $\bar{\mathcal{A}} \in D_\wedge(\mathbb{R})$ and $\widehat{\mathcal{A}} \in D(\mathbb{R})$, the condition $\mathrm{d}\widehat{\mathcal{A}}(t) \ge 0$, for $t \in \mathcal{Z}(\bar{\mathcal{A}})$, in Lemma 2 is equivalent to $\{\widehat{\mathcal{A}}(t),\, t \in \mathtt{z}_y(\bar{\mathcal{A}})\} \in D_\wedge(\mathtt{z}_y(\bar{\mathcal{A}}))$.

REMARK 9 ($\bar{\mathcal{A}} = 0$). Observe that if $\bar{\mathcal{A}} = 0$, then Lemma 2 reduces to Lemma 1. Indeed, in this case, $\mathcal{Z}(0) = \mathbb{R}$, $\widehat{\mathcal{A}} \in D_\wedge(\mathbb{R})$ with $\widehat{\mathcal{A}}(\infty) < \infty$, and $\Psi_n(0, \widehat{\mathcal{A}}) = \widehat{\mathcal{A}}$. Hence, one obtains $\mathcal{A}^n = \lfloor \sqrt{n}\widehat{\mathcal{A}} \rfloor$ and $n^{-1/2}\,\tilde{G} * F * \lfloor \sqrt{n}\widehat{\mathcal{A}} \rfloor \xrightarrow{s} \tilde{G} * F * \widehat{\mathcal{A}}$.

DEFINITION 1. We say that $\widehat{\mathcal{A}}_* \in D(\mathbb{R})$ is QED-optimal with respect to $\bar{\mathcal{A}} \in D_\wedge(\mathbb{R})$ if

$$\widehat{\mathcal{A}}_* = \operatorname*{arg\,min}_{\widehat{\mathcal{A}} \in D(\mathbb{R}):\, \mathrm{d}\widehat{\mathcal{A}}(t) \geq 0,\, t \in \mathcal{Z}(\bar{\mathcal{A}})} C(Z\Gamma + \tilde{G} * F * \widehat{\mathcal{A}} - \widehat{\gamma}).$$

COROLLARY 2 **(QED optimality)**. *Suppose $c_u(t) + c_o(t) \leq B$ for all $t \in \mathbb{R}$, $\mathbb{E}\sigma < \infty$, and $\widehat{\gamma}^n = \sqrt{n}\,(\bar{\gamma}^n - \gamma) \xrightarrow{s} \widehat{\gamma}$, with $\gamma \stackrel{s}{=} \tilde{G} * F * \bar{\mathcal{A}}$, where $\bar{\mathcal{A}}$ is a unique minimizer of $C(\tilde{G} * F * \cdot - \gamma)$ over $D_\wedge(\mathbb{R})$. Assume that*

$$\liminf_{M \to \infty} \inf_{\mathcal{A} \in V_\wedge(\mathbb{R}):\, \mathcal{A}(\infty) \geq M} C(\tilde{G} * F * \mathcal{A} - \gamma) > 0, \tag{18}$$

and $\widehat{\mathcal{A}}_*$ is QED-optimal (with respect to $\bar{\mathcal{A}}$). If $(\bar{\mathcal{A}}, \widehat{\mathcal{A}}_*)$ satisfies the conditions of Lemma 2, then $\{\lfloor \sqrt{n}\Psi_n^+(\bar{\mathcal{A}}, \widehat{\mathcal{A}}_*) \rfloor\}_n \subseteq D_{\mathbb{N}}(\mathbb{R})$ is a QED sequence that is asymptotically optimal on the diffusion scale:

$$\begin{aligned}
&\liminf_{n \to \infty} \frac{1}{\sqrt{n}} C\left(X_{\mathcal{A}^n} - \gamma^n\right) \\
&\qquad \geq \lim_{n \to \infty} \frac{1}{\sqrt{n}} C\left(X_{\lfloor \sqrt{n}\Psi_n^+(\bar{\mathcal{A}}, \widehat{\mathcal{A}}_*) \rfloor} - \gamma^n\right) \\
&\qquad = C(Z\Gamma + \tilde{G} * F * \widehat{\mathcal{A}}_* - \widehat{\gamma}),
\end{aligned} \tag{19}$$

for any other sequence of appointment plans $\{\mathcal{A}^n\}_n \subseteq D_{\mathbb{N}}(\mathbb{R})$.

Proof. See Section 5.7. □

REMARK 10. Condition (18) eliminates the possibility that multiple sequences of appointment plans with different limits have vanishing fluid costs, and hence, it will guarantee a unique centerings for diffusion scaling. Example EC.2 in Section EC.2 illustrates this point.

4.2. Solution of a Special Case

In this section, we consider the special case of exponential service durations ($\tilde{G}(t) = pe^{-\mu t}\mathbb{1}\{t \geq 0\}$) and perfect punctuality ($F(t) = \mathbb{1}\{t \geq 0\}$). First, in Lemma 3, we identify goals that admit zero fluid-cost. QED-optimal appointment plans are described in Lemma 4. The section concludes with Example 4, where we compute diffusion-optimal appointment plans. For notational convenience,

we define a scaled goal: $\gamma_\mu = \{\gamma_\mu(t) = e^{\mu t}\gamma(t) : t \in \mathbb{R}\}$; note that if $\mathrm{d}\gamma_\mu(t) = 0$ for $t < T$, then $\gamma(t) = 0$ for $t < T$.

LEMMA 3 **(Zero fluid-cost for perfect punctuality and exponential service)**. *If* $\tilde{G}(t) = pe^{-\mu t}\mathbb{1}\{t \geq 0\}$ *and* $\gamma = \tilde{G} * \bar{\mathcal{A}}$, *then* $\bar{\mathcal{A}} \in D_\wedge(\mathbb{R})$ *if and only if* $\gamma_\mu \in D_\wedge(\mathbb{R})$. *Moreover,* $pe^{\mu t}\,\mathrm{d}\bar{\mathcal{A}}(t) = \mathrm{d}\gamma_\mu(t)$ *for all* $t \in \mathbb{R}$.

Proof. The definition of the convolution implies $\mathrm{d}(\tilde{G} * \bar{\mathcal{A}})(t) = p\,\mathrm{d}\bar{\mathcal{A}}(t) - \mu(\tilde{G} * \bar{\mathcal{A}})(t)\,\mathrm{d}t$, or $p\,\mathrm{d}\bar{\mathcal{A}}(t) = \mathrm{d}(\tilde{G} * \bar{\mathcal{A}})(t) + \mu(\tilde{G} * \bar{\mathcal{A}})(t)\,\mathrm{d}t$. Hence, $e^{\mu t}p\,\mathrm{d}\bar{\mathcal{A}}(t) = e^{\mu t}\,\mathrm{d}\gamma(t) + e^{\mu t}\mu\gamma(t)\,\mathrm{d}t = \mathrm{d}\gamma_\mu(t)$, and the result follows. □

The next result characterizes the QED-optimal schedules under perfect punctuality and exponential service durations. Let φ and Φ denote the standard normal density and distribution function, respectively. Define $\beta = \{\beta(t) : t \in \mathbb{R}\}$ by

$$\beta(t) := \Phi^{\leftarrow}\left(\frac{c_o(t)}{c_o(t) + c_u(t)}\right),$$

where $\Phi^{\leftarrow}$ is the inverse of Φ. For notational simplicity, define $c(t,x) := c_o(t)\,x^+ + c_u(t)x^-$ as the instantaneous cost function.

LEMMA 4. *Suppose* c_u, c_o, β, $\widehat{\gamma} \in D(\mathbb{R})$, $F(t) = \mathbb{1}\{t \geq 0\}$, *and* $\tilde{G}(t) = pe^{-\mu t}\mathbb{1}\{t \geq 0\}$. *If* $\mathcal{S}(\widehat{\gamma}) < \infty$, $|\mathcal{J}(\bar{\mathcal{A}}^{\leftarrow})| < \infty$, *and*

$$\lim_{M\to\infty} \inf_{a \in D_\wedge(\mathtt{z}_y(\bar{\mathcal{A}})):\, \sup_{t\in\mathtt{z}_y(\bar{\mathcal{A}})}|a(t)|\geq M} C(\tilde{G} * a) = \infty,$$

for $y \in \mathcal{J}(\bar{\mathcal{A}}^{\leftarrow})$, *then the QED-optimal* $\widehat{\mathcal{A}}_*$ *with respect to* $\bar{\mathcal{A}} \in D_\wedge(\mathbb{R})$, *satisfies*

$$\begin{aligned}\mathrm{d}\widehat{\mathcal{A}}_*(t) = \frac{1}{p}\,&\mathrm{d}\left(\widehat{\gamma}(t) - \Gamma(t)\,\beta(t)\right)\\ &+ \frac{\mu}{p}\left(\widehat{\gamma}(t) - \Gamma(t)\,\beta(t)\right)\mathrm{d}t,\end{aligned} \tag{20}$$

for $t \notin \tilde{\mathcal{Z}}(\bar{\mathcal{A}})$, *and*

$$\begin{aligned}&\{\widehat{\mathcal{A}}_*(t),\, t \in \mathtt{z}_y(\bar{\mathcal{A}})\}\\ &= \operatorname*{arg\,min}_{a\in D_\wedge(\mathtt{z}_y(\bar{\mathcal{A}}))} \int_{\mathtt{z}_y(\bar{\mathcal{A}})} \mathbb{E}c(t, Z\,\Gamma(t) + \tilde{G} * a(t) - \widehat{\gamma}(t))\,\mathrm{d}t\\ &\qquad - \frac{1}{p}\tilde{G} * \widehat{\mathcal{A}}_*(l_y-),\end{aligned}$$

for $y \in \mathcal{J}(\bar{\mathcal{A}}^{\leftarrow})$, where $l_y := \inf\{t \in \mathsf{z}_y(\bar{\mathcal{A}})\}$. The corresponding QED cost is given by

$$\begin{aligned} &C(Z\,\Gamma + \tilde{G} * \widehat{\mathcal{A}}_* - \widehat{\gamma}) \\ &= \int_{t \notin \tilde{\mathcal{Z}}(\bar{\mathcal{A}})} (c_o(t) + c_u(t))\,\Gamma(t)\,\varphi(\beta(t))\,\mathrm{d}t \\ &+ \sum_{y \in \mathcal{J}(\bar{\mathcal{A}}^{\leftarrow})} \inf_{a \in D_\wedge(\mathsf{z}_y)} \int_{\mathsf{z}_y(\bar{\mathcal{A}})} \mathbb{E}c(t, Z\,\Gamma(t) + \tilde{G} * a(t) - \widehat{\gamma}(t))\,\mathrm{d}t. \end{aligned}$$

Proof. See Section 5.8. □

REMARK 11. In Lemma 4, if $\mathcal{J}(\bar{\mathcal{A}}^{\leftarrow}) = \emptyset$, then $C(Z\,\Gamma + \tilde{G} * \widehat{\mathcal{A}}_* - \widehat{\gamma}) = \mathcal{S}(\Gamma \cdot \varphi \circ \beta)$ and $\tilde{G} * \widehat{\mathcal{A}}_* = \widehat{\gamma} - \Gamma\,\beta$. In this special case, a diffusion version of (10) holds: $\mathbb{E}\sigma \int_{-\infty}^{\infty} \mathrm{d}\widehat{\mathcal{A}}_*(t) = \int_{-\infty}^{\infty} (\widehat{\gamma}(t) - \Gamma(t)\beta(t))\,\mathrm{d}t$. The optimal $\widehat{\mathcal{A}}_*$ can be evaluated via a transform: $\tilde{G}^\tau\,\widehat{\mathcal{A}}_*^\tau = (\widehat{\gamma} - \Gamma\beta)^\tau$.

EXAMPLE 4 (MODIFIED BOX: QED). Consider the setup from Example 2. Set $\widehat{\gamma} \equiv 0$ for this example, along with the following costs: $c_o(t) = \mathbb{1}\{t < T\} + 4\,\mathbb{1}\{t \geq T\}$ and $c_u(t) = 2$. In this case, $\beta(t) = \Phi^{\leftarrow}(2/3)(1 - 2\,\mathbb{1}\{t < T\})$. The fluid-optimal schedule (12) and (15) yield $\Gamma^2(t) = e^{-\mu(t-T)^+} - \frac{p}{2}e^{-2\mu(t-T)^+} - \frac{p}{2}e^{-2\mu t}$, for $t \geq 0$; otherwise, $\Gamma^2(t) = 0$. The function $\gamma_\mu^{\leftarrow}$ has two jumps, at $y = 0$ and $y = \bar{\mathcal{A}}(\infty) = 1/p + \mu T/p$, which correspond to two flat intervals $\mathsf{z}_0 = (-\infty, 0)$ and $\mathsf{z}_{\bar{\mathcal{A}}(\infty)} = [T, \infty)$. First,

$$\begin{aligned} &\inf_{a \in D_\wedge(\mathsf{z}_0)} \int_{\mathsf{z}_0} \mathbb{E}c(t, Z\,\Gamma(t) + \tilde{G} * a(t) - \widehat{\gamma}(t))\,\mathrm{d}t \\ &\qquad = \inf_{a \in D_\wedge(\mathsf{z}_0)} \int_{\mathsf{z}_0} \mathbb{E}c(t, \tilde{G} * a(t))\,\mathrm{d}t = 0, \end{aligned}$$

implying that $\widehat{\mathcal{A}}_*(t) = 0$, for $t < 0$.

Second, for $t \in [0, T)$ (or equivalently, $t \notin \mathsf{z}_0 \cup \mathsf{z}_{\bar{\mathcal{A}}(\infty)}$), the optimal $\widehat{\mathcal{A}}_*$ is such that the corresponding asymptotic census process achieves optimality for every t: $\tilde{G} * \widehat{\mathcal{A}}_*(t) = -\Gamma(t)\,\beta(t) = \Gamma(t)\,\Phi^{\leftarrow}(2/3)$, i.e., one solves an instantaneous newsvendor problem for every t. Since G is exponential, the preceding equality implies a differential equation:

$$\mathrm{d}\widehat{\mathcal{A}}_*(t) = \frac{1}{p}\Phi^{\leftarrow}(2/3)\,(\mathrm{d}\Gamma(t) + \mu\Gamma(t)\,\mathrm{d}t)\,, \tag{21}$$

and $\widehat{\mathcal{A}}_*(0) = \Phi^{\leftarrow}(2/3)\sqrt{1-p}/p$.

Third, the parameters of the model are such that the optimal $\widehat{\mathcal{A}}$ on $\mathtt{z}_{\bar{\mathcal{A}}(\infty)} = [T, \infty)$ is not only in $D_\wedge([T,\infty))$ but also non-increasing. Hence, the optimal $\widehat{\mathcal{A}}$ is a constant function over $[T,\infty)$, the value of which is determined at $t = T$: $\widehat{\mathcal{A}}_*(t) = a(T)$, for $t \geq T$. This leads to

$$\begin{aligned}
&\inf_{a \in D_\wedge(\mathtt{z}_{\bar{\mathcal{A}}(\infty)})} \int_{\mathtt{z}_{\bar{\mathcal{A}}(\infty)}} \mathbb{E}c(t, Z\Gamma(t) + \tilde{G} * a(t) - \widehat{\gamma}(t))\, \mathrm{d}t \\
&= \inf_{a \in D_\wedge(\mathtt{z}_{\bar{\mathcal{A}}(\infty)})} \int_{\mathtt{z}_{\bar{\mathcal{A}}(\infty)}} \mathbb{E}c(t, Z\Gamma(t) + \tilde{G} * a(t))\, \mathrm{d}t \\
&= \inf_{a(T) \in \mathbb{R}} \int_T^\infty \mathbb{E}c(t, Z\Gamma(t) + pe^{-\mu(t-T)} a(T))\, \mathrm{d}t \\
&= \inf_{a(T) \in \mathbb{R}} \int_T^\infty \Gamma(t)\, \mathbb{E}c(t, Z + \xi(t, a(T)))\, \mathrm{d}t, \qquad (22)
\end{aligned}$$

where $\xi(t, a(T)) := pe^{-\mu(t-T)} a(T)/\Gamma(t)$, for $t \geq T$. Based on Z being standard normal, the expectation in (22) satisfies $\mathbb{E}c(t, Z + \xi(t, a(T))) = 6\,\varphi(\xi(t, a(T))) + \xi(t, a(T))\,(4\Phi(\xi(t, a(T))) - 2\Phi(-\xi(t, a(T))))$, and $a_*(T)$ that achieves the minimum in (22) solves (a first order condition)

$$\int_T^\infty \left[e^{-\mu(t-T)}\, (6\Phi(\xi(t, a_*(T)) - 2) \right] \mathrm{d}t = 0. \qquad (23)$$

Hence, for $t \geq T$, the optimal diffusion-scale schedule is $\widehat{\mathcal{A}}_*(t) = a_*(T) - p^{-1}\, \tilde{G} * \widehat{\mathcal{A}}_*(T-) = a_*(T) - p^{-1}\Gamma(T-)\, \Phi^{\leftarrow}(2/3)$.

We solve (23) and (21) numerically to obtain a numerical approximation of $\widehat{\mathcal{A}}_*$. Once $\bar{\mathcal{A}}_*$ and $\widehat{\mathcal{A}}_*$ are available, the appointment schedule for the finite system of interest is given by $\mathcal{A}^n = \lfloor \sqrt{n} \Psi_n^+(\bar{\mathcal{A}}_*, \widehat{\mathcal{A}}_*) \rfloor$. Results for $n = 25$, $T = 3$, $p = 0.5$, and $\mu = 1$ are presented in Figure 2. In the figure, we also show $\tilde{\mathcal{A}}^n = n\bar{\mathcal{A}}_* + \sqrt{n}\widehat{\mathcal{A}}_*$ (which is not a feasible appointment plan) and the corresponding $\tilde{G} * F * \tilde{\mathcal{A}}^n$ to highlight the differences with $\mathcal{A}^n$ and $\tilde{G} * F * \mathcal{A}^n$, respectively, i.e., to illustrate the effect of constraining $\mathcal{A}^n$ to a feasible appointment schedule. Finally, we remark that the probability of encountering an overage at time $t \in (0, T)$ is approximately $\frac{c_u(t)}{c_u(t)+c_o(t)} = \frac{2}{3}$ (see Section EC.8 for justification).

5. Proofs

5.1. Proof of Theorem 1

Define $c_{u+o}(\cdot) := c_u(\cdot) + c_o(\cdot)$. Let $\bar{X}_{\mathcal{A}^n} = \{\bar{X}_{\mathcal{A}^n}(t) : t \in \mathbb{R}\}$ and $\bar{Z}_{\mathcal{A}^n} = \{\bar{Z}_{\mathcal{A}^n}(t) : t \in \mathbb{R}\}$, where

$$\bar{X}_{\mathcal{A}^n}(t) := \frac{1}{n} X_{\mathcal{A}^n}(t) \quad \text{and} \quad \bar{Z}_{\mathcal{A}^n}(t) := \frac{1}{n} Z_{\mathcal{A}^n}(t). \qquad (24)$$

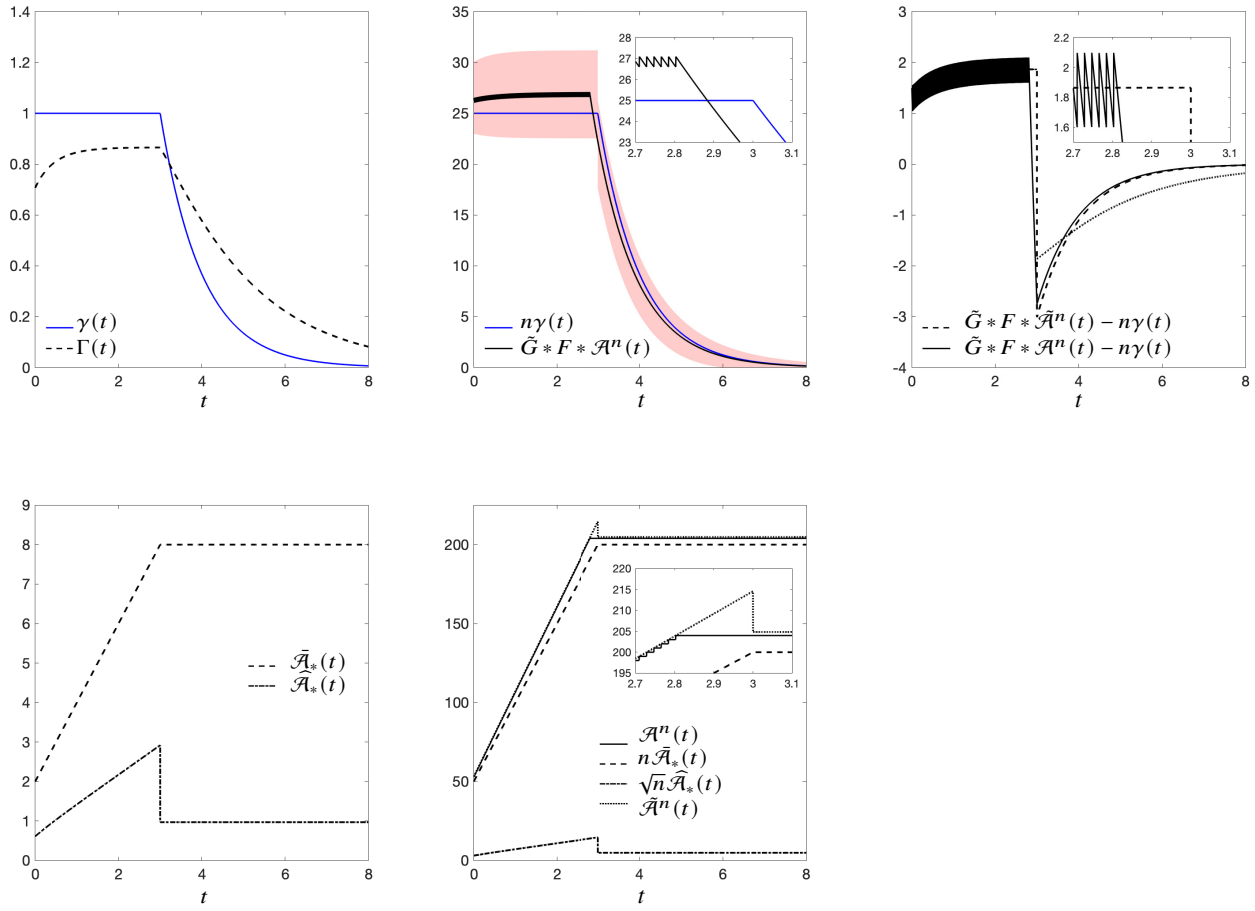

Figure 2 Illustration for Example 4. The two plots on the left correspond to the limiting system; all other plots correspond to a finite system with $n = 25$. Here, $\mathcal{A}^n = \lfloor \sqrt{n}\Psi_n^+(\bar{\mathcal{A}}, \widehat{\mathcal{A}}) \rfloor$, and $\tilde{\mathcal{A}}^n = n\bar{\mathcal{A}}_* + \sqrt{n}\widehat{\mathcal{A}}_*$. In the top-middle plot, the shaded area corresponds to ± one standard deviation of $X_{\mathcal{A}^n}$ around $\mathbb{E}X_{\mathcal{A}^n} = \tilde{G} * F * \mathcal{A}^n$. In the top-right plot, the dotted line corresponds to $\tilde{G} * F * (n\bar{\mathcal{A}}_* + \sqrt{n}\check{\mathcal{A}}) - n\gamma$, where $\check{\mathcal{A}}$ is optimal schedule on the diffusion scale without the constraint that the function is non-decreasing on $[T, \infty)$.

Equality (2) implies $\bar{X}_{\mathcal{A}^n}(t) = \bar{Z}_{\mathcal{A}^n}(t) + \tilde{G} * F * \bar{\mathcal{A}}^n(t)$, which, together with homogeneity of C, yields $\frac{1}{n}C(X_{\mathcal{A}^n} - \gamma^n) = C(\bar{Z}_{\mathcal{A}^n} + \tilde{G} * F * \bar{\mathcal{A}}^n - \bar{\gamma}^n)$. This, combined with $|C(\Delta_1) - C(\Delta_2)| \le \mathcal{S}(\Delta_1 - \Delta_2)$, implies $|\frac{1}{n}C(X_{\mathcal{A}^n} - \gamma^n) - C(\bar{L} - \gamma)| = |C(\bar{Z}_{\mathcal{A}^n} + \tilde{G} * F * \bar{\mathcal{A}}^n - \bar{\gamma}^n) - C(\bar{L} - \gamma)| \le \mathcal{S}(\bar{Z}_{\mathcal{A}^n} + \tilde{G} * F * \bar{\mathcal{A}}^n - \bar{L} + \gamma - \bar{\gamma}^n) \le \mathcal{S}(\bar{Z}_{\mathcal{A}^n}) + \mathcal{S}(\tilde{G} * F * \bar{\mathcal{A}}^n - \bar{L}) + \mathcal{S}(\gamma - \bar{\gamma}^n)$, where the last inequality follows from the triangle inequality. It is sufficient to argue that the first term vanishes, as the last two terms vanish due to the assumptions on $\bar{\mathcal{A}}^n$ and $\bar{\gamma}^n$.

Hence, we focus on proving $\mathcal{S}(\bar{Z}_{\mathcal{A}^n}) \to 0$. First, $\mathbb{E}|\bar{Z}_{\mathcal{A}^n}(t)| \le 2 \cdot \tilde{G} * F * \bar{\mathcal{A}}^n(t)$ and $\tilde{G} * F * \bar{\mathcal{A}}^n \xrightarrow{s} \bar{L}$ yield

$$\begin{aligned}
&\int_{-\infty}^{\infty} \mathbb{E}|\bar{Z}_{\mathcal{A}^n}(t)|\, c_{u+o}(t)\, \mathbb{1}\left\{\bar{L}(t) = 0\right\}\, \mathrm{d}t \\
&\le 2 \int_{-\infty}^{\infty} |\tilde{G} * F * \bar{\mathcal{A}}^n(t)|\, c_{u+o}(t)\, \mathbb{1}\left\{\bar{L}(t) = 0\right\}\, \mathrm{d}t \\
&\le 2\mathcal{S}(\tilde{G} * F * \bar{\mathcal{A}}^n - \bar{L}) \to 0. \qquad (25)
\end{aligned}$$

Second, in $\int_{|\bar{L}(t)|>0} \mathbb{E}|\bar{Z}_{\mathcal{A}^n}(t)|\, c_{u+o}(t)\, \mathrm{d}t = \mathcal{S}(\bar{L}) \int_{|\bar{L}(t)|>0} \frac{\mathbb{E}|\bar{Z}_{\mathcal{A}^n}(t)|}{|\bar{L}(t)|}\, c_{u+o}(t)\, \frac{|\bar{L}(t)|}{\mathcal{S}(\bar{L})}\, \mathrm{d}t$, the term $\mathbb{1}\left\{|\bar{L}(t)| > 0\right\} c_{u+o}(\cdot)\, |\bar{L}(\cdot)|/\mathcal{S}(\bar{L})$ can be interpreted as a probability density, which makes Theorem 1.2 in (Ethier and Kurtz 1986, p. 492) (a version of dominated convergence) applicable, as we now show. First, the conditions of the theorem can be verified as follows:

- $\mathbb{E}|\bar{Z}_{\mathcal{A}^n}(t)| \le 2 \cdot \tilde{G} * F * \bar{\mathcal{A}}^n(t)$.
- $(\mathbb{E}|\bar{Z}_{\mathcal{A}^n}(t)|)^2 \le \mathbb{E}|\bar{Z}_{\mathcal{A}^n}(t)|^2 \le n^{-1}\bar{\mathcal{A}}^n(\infty) \to 0$.
- $\tilde{G} * F * \bar{\mathcal{A}}^n \xrightarrow{s} \bar{L}$ implies the following L^1 convergence:

$$\begin{aligned}
&\int_{|\bar{L}(t)|>0} \left| \frac{\tilde{G} * F * \bar{\mathcal{A}}^n(t)}{|\bar{L}(t)|} - \frac{\bar{L}(t)}{|\bar{L}(t)|} \right| c_{u+o}(t) \frac{|\bar{L}(t)|}{\mathcal{S}(\bar{L})}\, \mathrm{d}t \\
&\qquad \le \frac{\mathcal{S}(\tilde{G} * F * \bar{\mathcal{A}}^n - \bar{L})}{\mathcal{S}(\bar{L})} \to 0,
\end{aligned}$$

which, in turn, implies convergence in probability and weak convergence.

- The preceding item yields

$$\begin{aligned}
&\int_{|\bar{L}(t)|>0} \frac{\tilde{G} * F * \bar{\mathcal{A}}^n(t)}{|\bar{L}(t)|}\, c_{u+o}(t) \frac{|\bar{L}(t)|}{\mathcal{S}(\bar{L})}\, \mathrm{d}t \\
&\to \frac{1}{\mathcal{S}(\bar{L})} \int_{-\infty}^{\infty} \bar{L}(t)\, c_{u+o}(t)\, \mathrm{d}t.
\end{aligned}$$

The theorem then implies

$$\int_{-\infty}^{\infty} \mathbb{E}|\bar{Z}_{\mathcal{A}^n}(t)|\, c_{u+o}(t)\, \mathbb{1}\left\{|\bar{L}(t)| > 0\right\}\, \mathrm{d}t \to 0,$$

which, together with (25), yields the desired $\mathcal{S}(\bar{Z}_{\mathcal{A}^n}) \to 0$. □

5.2. Proof of Proposition 2

The conditions of the proposition and the triangle inequality $C(\tilde{G} * F * \bar{\mathcal{A}} - \gamma) \geq C(\tilde{G} * F * \bar{\mathcal{A}}) - C(\gamma) \geq C(\tilde{G} * F * \bar{\mathcal{A}}) - \mathcal{S}(\gamma)$ yield

$$\lim_{M \to \infty} \inf_{\bar{\mathcal{A}} \in V_\wedge(\mathbb{R}): \bar{\mathcal{A}}(\infty) \geq M} C(\tilde{G} * F * \bar{\mathcal{A}} - \gamma) = \infty.$$

This limit implies that there exists M_γ such that, for all $M \geq M_\gamma$,

$$\begin{aligned} &\inf_{\bar{\mathcal{A}} \in V_\wedge(\mathbb{R}): \bar{\mathcal{A}}(\infty) \geq M} C(\tilde{G} * F * \bar{\mathcal{A}} - \gamma) \\ &\quad > \mathcal{S}(\gamma) \geq C(-\gamma) \\ &\quad \geq \inf_{\bar{\mathcal{A}} \in V_\wedge(\mathbb{R}): \bar{\mathcal{A}}(\infty) \leq M_\gamma} C(\tilde{G} * F * \bar{\mathcal{A}} - \gamma). \end{aligned}$$

As a result, to minimize $C(\tilde{G} * F * \bar{\mathcal{A}} - \gamma)$, it is enough to consider $\bar{\mathcal{A}} \in V_\wedge(\mathbb{R})$ such that $\bar{\mathcal{A}}(\infty) \leq M_\gamma$:

$$\begin{aligned} &\min_{\bar{\mathcal{A}} \in V_\wedge(\mathbb{R}): \bar{\mathcal{A}}(\infty) \leq M_\gamma} C(\tilde{G} * F * \bar{\mathcal{A}} - \gamma) \\ &\qquad = \min_{\bar{\mathcal{A}} \in V_\wedge(\mathbb{R})} C(\tilde{G} * F * \bar{\mathcal{A}} - \gamma). \end{aligned}$$

Consider a sequence $\{\bar{\mathcal{A}}_\alpha\}_\alpha \subseteq V_\wedge(\mathbb{R})$, with $\sup_\alpha \bar{\mathcal{A}}_\alpha(\infty) \leq M_\gamma$, the cost of which converges to $\min_{\bar{\mathcal{A}} \in V_\wedge(\mathbb{R}): \bar{\mathcal{A}}(\infty) \leq M_\gamma} C(\tilde{G} * F * \bar{\mathcal{A}} - \gamma)$. By Helly's Selection Theorem, there is a subsequence $\{\bar{\mathcal{A}}_{\alpha_k}\}_k$ that converges to some $\bar{\mathcal{A}}_*$, for each $x \in \mathbb{R}$. Note that $\bar{\mathcal{A}}_* \in V_\wedge(\mathbb{R})$ but not necessarily $\bar{\mathcal{A}}_* \in D_\wedge(\mathbb{R})$.

For any distribution function H, we have $H * \bar{\mathcal{A}}_{\alpha_k}(t) \to H * \bar{\mathcal{A}}_*(t)$, for all $t \in \mathbb{R}$. This is because, if we denote by v a random variable following the distribution H, then $H * \bar{\mathcal{A}}_{\alpha_k}(t) = \mathbb{E}\bar{\mathcal{A}}_{\alpha_k}(t - v) \to \mathbb{E}\bar{\mathcal{A}}_*(t - v) = H * \bar{\mathcal{A}}_*(t)$, by the dominated convergence theorem (here, for all x, $\bar{\mathcal{A}}_{\alpha_k}(t - x) \to \bar{\mathcal{A}}_*(t - x)$). As F and $G * F$ are both distribution functions, we have $F * \bar{\mathcal{A}}_{\alpha_k} \to F * \bar{\mathcal{A}}_*$ and $G * F * \bar{\mathcal{A}}_{\alpha_k} \to G * F * \bar{\mathcal{A}}_*$, which further implies $\tilde{G} * F * \bar{\mathcal{A}}_{\alpha_k} \to \tilde{G} * F * \bar{\mathcal{A}}_*$ pointwise.

The RCLL regularization of $\bar{\mathcal{A}}_*$, $\tilde{\bar{\mathcal{A}}}_*$, is the minimizer because

$$\begin{aligned}
&\min_{\bar{\mathcal{A}}\in V_\wedge(\mathbb{R}):\,\bar{\mathcal{A}}(\infty)\le M_\gamma} C(\tilde{G}*F*\bar{\mathcal{A}}-\gamma)\\
&\quad=\lim_{k\to\infty} C(\tilde{G}*F*\bar{\mathcal{A}}_{\alpha_k}-\gamma)\\
&\quad\ge C(\tilde{G}*F*\bar{\mathcal{A}}_*-\gamma)\\
&\quad= C(\tilde{G}*F*\tilde{\bar{\mathcal{A}}}_*-\gamma)\\
&\quad\ge \min_{\bar{\mathcal{A}}\in V_\wedge(\mathbb{R}):\,\bar{\mathcal{A}}(\infty)\le M_\gamma} C(\tilde{G}*F*\bar{\mathcal{A}}-\gamma);
\end{aligned}$$

Fatou's Lemma is used in the first inequality, and Lemma EC.2 is used in the second equality. □

5.3. Proof of Lemma 1

The triangle inequality implies

$$\begin{aligned}
&\mathcal{S}(\tilde{G}*F*(n\bar{\mathcal{A}}-\lfloor n\bar{\mathcal{A}}\rfloor))\\
&\quad\le \mathcal{S}(\tilde{G}*F*\Delta^n_{(-\infty,T)})+\mathcal{S}(\tilde{G}*F*\Delta^n_{[-T,T]})\\
&\qquad+\mathcal{S}(\tilde{G}*F*\Delta^n_{(T,\infty)})\\
&\quad\le \mathcal{S}(\tilde{G}*F*\Delta^n_{(-\infty,T)})+\mathcal{S}(F*\Delta^n_{[-T,T]})\\
&\qquad+\mathcal{S}(G*F*\Delta^n_{[-T,T]})+\mathcal{S}(\tilde{G}*F*\Delta^n_{(T,\infty)}),
\end{aligned}\tag{26}$$

where $\Delta^n_I := \{\Delta^n_I(t) = (n\bar{\mathcal{A}}(t) - \lfloor n\bar{\mathcal{A}}(t)\rfloor)\mathbb{1}\{t\in I\} : t\in\mathbb{R}\}$. As $\Delta^n_{[-T,T]}(t)\in[0,1)$ and $\int_{-T}^{T}\Delta^n_{[-T,T]}(t)\,\mathrm{d}t\le 2T$, one has

$$\begin{aligned}
\mathcal{S}(H*\Delta^n_{[-T,T]}) &\le B\int_{-\infty}^{\infty} H*\Delta^n_{[-T,T]}(t)\,\mathrm{d}t\\
&= B\iint_{-\infty}^{\infty}\Delta^n_{[-T,T]}(t-u)\,\mathrm{d}t\,\mathrm{d}H(u)\\
&\le 2TB,
\end{aligned}\tag{27}$$

for any CDF H. On the other hand, both $n\bar{\mathcal{A}}$ and $\lfloor n\bar{\mathcal{A}}\rfloor$ are non-decreasing, and hence,

$$\begin{aligned}
&\mathcal{S}(\tilde{G} * F * \Delta^n_{(-\infty,T)}) \\
&\quad \le \mathcal{S}(\tilde{G} * F * (n\bar{\mathcal{A}}(\cdot)\mathbb{1}\{\cdot \in (-\infty,T)\})) \\
&\qquad + \mathcal{S}(\tilde{G} * F * (\lfloor n\bar{\mathcal{A}}(\cdot)\rfloor\mathbb{1}\{\cdot \in (-\infty,T)\})) \\
&\quad = B\,\mathbb{E}\sigma\,(n\bar{\mathcal{A}}(-T) + \lfloor n\bar{\mathcal{A}}(-T)\rfloor) \\
&\quad \le 2nB\,\mathbb{E}\sigma\,\bar{\mathcal{A}}(-T). \qquad (28)
\end{aligned}$$

Similarly, $\Delta^n_{(T,\infty)}$ is a difference of two non-decreasing functions ($\Delta^n_{(T,\infty)}(t) = (n\bar{\mathcal{A}}(t) - \lfloor n\bar{\mathcal{A}}(T)\rfloor) - (n\bar{\mathcal{A}}(t) - \lfloor n\bar{\mathcal{A}}(T)\rfloor)$, for $t > T$), and therefore,

$$\begin{aligned}
&\mathcal{S}(\tilde{G} * F * \Delta^n_{(T,\infty)}) \\
&\qquad \le 2nB\,\mathbb{E}\sigma\,(\bar{\mathcal{A}}(\infty) - n^{-1}\lfloor n\bar{\mathcal{A}}(T)\rfloor). \qquad (29)
\end{aligned}$$

Due to $\bar{\mathcal{A}}(\infty) < \infty$, for any $\epsilon > 0$, one can choose T such that $\bar{\mathcal{A}}(-T) < \frac{\epsilon}{8B\mathbb{E}\sigma}$ and $\bar{\mathcal{A}}(\infty) - n^{-1}\lfloor n\bar{\mathcal{A}}(T)\rfloor < \frac{\epsilon}{8B\mathbb{E}\sigma}$. Then, for such T, one can choose n_0 such that when $n \ge n_0$, $\frac{2TB}{n} \le \frac{\epsilon}{4}$. Hence for $n \ge n_0$, $\frac{1}{n}\mathcal{S}(\tilde{G} * F * (n\bar{A} - \lfloor n\bar{A}\rfloor)) \le \epsilon$. Combining (26), (27), (28), and (29) yields the statement of the lemma. □

5.4. Proof of Corollary 1

Homogeneity and convexity of C, Jensen's inequality and the triangle inequality yield (see (24)) $\frac{1}{n}C(X_{\mathcal{A}^n} - \gamma^n) \ge C(\mathbb{E}\bar{X}_{\mathcal{A}^n} - \bar{\gamma}^n) \ge C(\tilde{G} * F * \bar{\mathcal{A}}^n - \gamma) - \mathcal{S}(\bar{\gamma}^n - \gamma)$. Hence, the optimality of $\mathcal{A}_*$ and $\bar{\gamma}^n \xrightarrow{s} \gamma$ imply

$$\liminf_{n\to\infty} \frac{1}{n}C(X_{\mathcal{A}^n} - \gamma^n) \ge C(\tilde{G} * F * \bar{\mathcal{A}}_* - \gamma).$$

On the other hand, the conditions of Lemma 1 are satisfied, and we have that $\frac{1}{n}\tilde{G} * F * \lfloor n\bar{\mathcal{A}}_*\rfloor \xrightarrow{s} \tilde{G} * F * \bar{\mathcal{A}}_*$. This, in turn, according to Theorem 1, yields $\frac{1}{n}C(X_{\lfloor n\bar{\mathcal{A}}_*\rfloor} - \gamma^n) \to C(\tilde{G} * F * \bar{\mathcal{A}}_* - \gamma)$. □

5.5. Proof of Theorem 2

Let $\widehat{X}_{\mathcal{A}^n} = \{\widehat{X}_{\mathcal{A}^n}(t) : t \in \mathbb{R}\}$ and $\widehat{Z}_{\mathcal{A}^n} = \{\widehat{Z}_{\mathcal{A}^n}(t) : t \in \mathbb{R}\}$, where $\widehat{X}_{\mathcal{A}^n}(t) := \frac{1}{\sqrt{n}}(X_{\mathcal{A}^n}(t) - n\gamma(t))$ and $\widehat{Z}_{\mathcal{A}^n}(t) := \frac{1}{\sqrt{n}}Z_{\mathcal{A}^n}(t)$. Then, (2) implies $\widehat{X}_{\mathcal{A}^n}(t) = \widehat{Z}_{\mathcal{A}^n}(t) + \tilde{G} * F * \widehat{\mathcal{A}}^n(t) + \sqrt{n}(\tilde{G} * F * \bar{\mathcal{A}}(t) - \gamma(t))$,

which, together with homogeneity of C, yields $\frac{1}{\sqrt{n}}C(X_{\mathcal{A}^n}-\gamma^n) = C(\widehat{Z}_{\mathcal{A}^n}+\tilde{G}*F*\widehat{\mathcal{A}}^n-\widehat{\gamma}^n+\sqrt{n}(\tilde{G}*F*\bar{\mathcal{A}}-\gamma)) = C(\widehat{Z}_{\mathcal{A}^n}+\tilde{G}*F*\widehat{\mathcal{A}}^n-\widehat{\gamma}^n)$. The second equality above is due to $\gamma \overset{s}{=} \tilde{G}*F*\bar{\mathcal{A}}$ and the triangle inequality of C. This, combined with $|C(\Delta_1)-C(\Delta_2)| \le \mathcal{S}(\Delta_1-\Delta_2)$, implies $|\frac{1}{\sqrt{n}}C(X_{\mathcal{A}^n}-\gamma^n) - C(\widehat{Z}_{\mathcal{A}^n}+\widehat{L}-\widehat{\gamma})| = |C(\widehat{Z}_{\mathcal{A}^n}+\tilde{G}*F*\widehat{\mathcal{A}}^n-\widehat{\gamma}^n) - C(\widehat{Z}_{\mathcal{A}^n}+\widehat{L}-\widehat{\gamma})| \le \mathcal{S}(\tilde{G}*F*\widehat{\mathcal{A}}^n-\widehat{L}) + \mathcal{S}(\widehat{\gamma}^n-\widehat{\gamma}) \to 0$, where the limit is due to the assumptions. Hence, it is enough to show $C(\widehat{Z}_{\mathcal{A}^n}+\widehat{L}-\widehat{\gamma}) \to C(Z\Gamma+\widehat{L}-\widehat{\gamma})$. However, this limit follows from Lemma 5 and Lemma 6; here, $C(\widehat{L}-\widehat{\gamma}) \le \mathcal{S}(\widehat{L})+\mathcal{S}(\widehat{\gamma}) < \infty$ due to the $\mathcal{S}$-convergences. □

LEMMA 5. *If* $\Gamma_n \overset{s}{\to} \Gamma$, *then* $C(Z\Gamma_n+\widehat{L}-\widehat{\gamma}) \to C(Z\Gamma+\widehat{L}-\widehat{\gamma})$.

Proof. We have $|C(Z\Gamma_n+\widehat{L}-\widehat{\gamma}) - C(Z\Gamma+\widehat{L}-\widehat{\gamma})| \le \mathcal{S}(Z(\Gamma_n-\Gamma)) = \mathbb{E}|Z|\,\mathcal{S}(\Gamma_n-\Gamma) \to 0$. □

LEMMA 6. *Suppose* $C(\widehat{L}-\widehat{\gamma}) < \infty$. *If* $\sup_n \bar{\mathcal{A}}^n(\infty) < \infty$ *and* $\Gamma_n \overset{s}{\to} \Gamma$, *then*

$$C(\widehat{Z}_{\mathcal{A}^n}+\widehat{L}-\widehat{\gamma}) - C(Z\Gamma_n+\widehat{L}-\widehat{\gamma}) \to 0. \tag{30}$$

Proof. Let $Y^n_{\mathcal{A}}(t) := \widehat{Z}_{\mathcal{A}^n}(t)+\widehat{L}(t)-\widehat{\gamma}(t)$ and $Y^n_{\Gamma}(t) := \Gamma_n(t)Z+\widehat{L}(t)-\widehat{\gamma}(t)$. We first prove that

$$\mathbb{E}(Y^n_{\mathcal{A}}(t))^{\pm} - \mathbb{E}(Y^n_{\Gamma}(t))^{\pm} \to 0, \tag{31}$$

for all $t \in \mathbb{R}$. Note that $\Gamma_n(t) \le \bar{\mathcal{A}}^n(\infty)$. Hence, for any subsequence n_k, there is a further subsequence, still denoted as n_k, such that $\lim_{n_k\to\infty}\Gamma_{n_k}(t) = s \in [0, \sup_n \bar{\mathcal{A}}^n(\infty)]$. Two scenarios are now possible:

1. $\lim_{n_k\to\infty}\Gamma_{n_k}(t) = s = 0$. Since $\{\varphi^{n_k}_i(t) - \mathbb{E}\varphi^{n_k}_i(t)\}_i$ is a sequence of independent r.v.s, we have

$$\begin{aligned}\mathbb{E}|\widehat{Z}_{\mathcal{A}^{n_k}}(t)|^2 &= \frac{1}{n_k}\sum_{i=1}^{\mathcal{A}^{n_k}(\infty)} \mathrm{Var}(\varphi^{n_k}_i(t)) \\ &= \left(\tilde{G}*F - (\tilde{G}*F)^2\right)*\bar{\mathcal{A}}^{n_k}(t) \\ &= \Gamma^2_{n_k}(t) \to 0.\end{aligned}$$

This then implies $\widehat{Z}_{\mathcal{A}^{n_k}}(t) \Rightarrow 0$, and the uniform integrability of the related terms. As a result, $\mathbb{E}(Y^{n_k}_{\mathcal{A}}(t))^{\pm} - \mathbb{E}(Y^{n_k}_{\Gamma}(t))^{\pm} \to \mathbb{E}(\widehat{L}(t)-\widehat{\gamma}(t))^{\pm} - \mathbb{E}(\widehat{L}(t)-\widehat{\gamma}(t))^{\pm} = 0$.

2. $\lim_{n_k\to\infty}\Gamma_{n_k}(t)=s>0$. The definition of $\widehat{Z}_{\mathcal{A}^{n_k}}(t)$ renders

$$\widehat{Z}_{\mathcal{A}^{n_k}}(t)=\Gamma_{n_k}(t)\,\frac{1}{\sqrt{n_k}\Gamma_{n_k}(t)}\sum_{i=1}^{\mathcal{A}^{n_k}(\infty)}\tilde{\varphi}_i^{n_k}(t),$$

where $\tilde{\varphi}_i^{n_k}(t):=\varphi_i^{n_k}(t)-\mathbb{E}[\varphi_i^{n_k}(t)]$. Note that $|\tilde{\varphi}_i^{n_k}(t)|\le 1$. Hence, $\tilde{\varphi}_i^{n_k}(t)$ has a finite second moment for each i, and

$$\sum_{i=1}^{\mathcal{A}^{n_k}(\infty)}\frac{\mathbb{E}[(\tilde{\varphi}_i^{n_k}(t))^2\mathbb{1}\{|\tilde{\varphi}_i^{n_k}(t)|\ge\epsilon\sqrt{n_k}\Gamma_{n_k}(t)\}]}{n_k\Gamma_{n_k}^2(t)}=0,$$

for all large enough n_k. Here, we utilize $\epsilon\sqrt{n_k}\Gamma_{n_k}(t)\to\infty$, for every $\epsilon>0$, which is due to $\lim_{n_k\to\infty}\Gamma_{n_k}(t)=s>0$. Lindeberg's Theorem for the triangular arrays (Billingsley 1995, p. 359, Theorem 27.2) yields $\widehat{Z}_{\mathcal{A}^{n_k}}(t)\Rightarrow sZ$, and the uniform integrability of the related terms follows. As a result, $\mathbb{E}(Y_{\mathcal{A}}^{n_k}(t))^{\pm}-\mathbb{E}(Y_{\Gamma}^{n_k}(t))^{\pm}\to\mathbb{E}(sZ+\widehat{L}(t)-\widehat{\gamma}(t))^{\pm}-\mathbb{E}(sZ+\widehat{L}(t)-\widehat{\gamma}(t))^{\pm}=0$.
Combining the above two scenarios, we have (31).

Now, (31), $|\mathbb{E}(Y_{\mathcal{A}}^{n}(t))^{\pm}-\mathbb{E}(Y_{\Gamma}^{n}(t))^{\pm}|\le\mathbb{E}|\widehat{Z}_{\mathcal{A}^n}(t)|+\Gamma_n(t)\,\mathbb{E}|Z|\le\Gamma_n(t)+\Gamma_n(t)\,\mathbb{E}|Z|\le 2\Gamma_n(t)$, the assumption $\Gamma_n\xrightarrow{s}\Gamma$, and the dominated convergence theorem jointly yield (30). □

5.6. Proof of Lemma 2

The convergence $\Gamma_n\xrightarrow{s}\Gamma$ can be easily proved if we have $\tilde{G}*F*\bar{\mathcal{A}}^n\xrightarrow{s}\tilde{G}*F*\bar{\mathcal{A}}$ (via the dominated convergence theorem and pointwise convergence; because $\bar{\mathcal{A}}^n\Rightarrow\bar{\mathcal{A}}$, from Lemma EC.1, one has $\tilde{G}*F*\bar{\mathcal{A}}^n\to\tilde{G}*F*\bar{\mathcal{A}}$ and $(\tilde{G}*F)^2*\bar{\mathcal{A}}^n\to(\tilde{G}*F)^2*\bar{\mathcal{A}}$ almost everywhere.).

The triangle inequality yields

$$\begin{aligned}
&\mathcal{S}(\tilde{G}*F*\widehat{\mathcal{A}}^n-\tilde{G}*F*\widehat{\mathcal{A}})\\
&=\frac{1}{\sqrt{n}}\mathcal{S}(\tilde{G}*F*(\mathcal{A}^n-n\bar{\mathcal{A}})-\tilde{G}*F*\sqrt{n}\widehat{\mathcal{A}})\\
&\le\frac{1}{\sqrt{n}}\mathcal{S}(\tilde{G}*F*\Delta_1^n)\\
&\qquad+\mathcal{S}(\tilde{G}*F*\Delta_2^n)+\mathcal{S}(\tilde{G}*F*\Delta_3^n),
\end{aligned}\tag{32}$$

where $\Delta_1^n:=\sqrt{n}\Psi_n^+(\bar{\mathcal{A}},\widehat{\mathcal{A}})-\mathcal{A}^n\ge 0$, $\Delta_2^n:=\Psi_n(\bar{\mathcal{A}},\widehat{\mathcal{A}})-\Psi_n^+(\bar{\mathcal{A}},\widehat{\mathcal{A}})\le 0$, and $\Delta_3^n:=\Psi_n(\bar{\mathcal{A}},\widehat{\mathcal{A}})-\sqrt{n}\bar{\mathcal{A}}-\widehat{\mathcal{A}}$. Using the same argument as in the proof of Lemma 1, one can show that the first term in (32) vanishes as $n\to\infty$. To this end, for $\epsilon>0$, let $\Delta_{1,1}^n:=\{\Delta_{1,1}^n(t)=\Delta_1^n(t)\mathbb{1}\{t<-\epsilon\sqrt{n}\}:t\in\mathbb{R}\}$,

$\Delta^n_{1,2} := \{\Delta^n_{1,2}(t) = \Delta^n_1(t)\mathbb{1}\{-\epsilon\sqrt{n} \le t \le \epsilon\sqrt{n}\} : t \in \mathbb{R}\}$, and $\Delta^n_{1,3} := \{\Delta^n_{1,3}(t) = \Delta^n_1(t)\mathbb{1}\{t > \epsilon\sqrt{n}\} : t \in \mathbb{R}\}$. Then, one has

$$\begin{aligned}
&\frac{1}{2B\mathbb{E}\sigma\sqrt{n}}\mathcal{S}(\tilde{G} * F * \Delta^n_{1,1}) \\
&\qquad \le \Psi^+_n(\bar{\mathcal{A}}, \widehat{\mathcal{A}})(-\epsilon\sqrt{n}) \\
&\qquad \le \sqrt{n} \sup_{u \le -\epsilon\sqrt{n}} \left(\bar{\mathcal{A}}(u) + \frac{1}{\sqrt{n}}\widehat{\mathcal{A}}(u)\right)^+ \\
&\qquad \le \sqrt{n}\bar{\mathcal{A}}(-\epsilon\sqrt{n}) + \sup_{u \le -\epsilon\sqrt{n}} \widehat{\mathcal{A}}^+(u) \to 0,
\end{aligned}$$

where the limit is due to the assumptions of the lemma. Similarly, one has

$$\begin{aligned}
&\frac{1}{2B\mathbb{E}\sigma\sqrt{n}}\mathcal{S}(\tilde{G} * F * \Delta^n_{1,3}) \\
&\le \Psi^+_n(\bar{\mathcal{A}}, \widehat{\mathcal{A}})(\infty) - \Psi^+_n(\bar{\mathcal{A}}, \widehat{\mathcal{A}})(\epsilon\sqrt{n}) + \frac{1}{\sqrt{n}} \\
&\quad \le \sqrt{n}(\bar{\mathcal{A}}(\infty) - \bar{\mathcal{A}}(\epsilon\sqrt{n})) \\
&\qquad + \sup_{u > \epsilon\sqrt{n}} \widehat{\mathcal{A}}(u) - \inf_{u > \epsilon\sqrt{n}} \widehat{\mathcal{A}}(u) \to 0,
\end{aligned}$$

where the last inequality is for all n large enough. The term corresponding to $\Delta^n_{1,2}$ can be bounded as follows: $\mathcal{S}(\tilde{G} * F * \Delta^n_{1,2}) \le \mathcal{S}(F * \Delta^n_{1,2}) + \mathcal{S}(G * F * \Delta^n_{1,2}) \le 4B\epsilon\sqrt{n}$. Finally, combining the preceding three estimates yields

$$\frac{1}{\sqrt{n}}\mathcal{S}(\tilde{G} * F * \Delta^n_1) \to 0. \tag{33}$$

As far as the second term in (32) is concerned, the inf in (16) can be analyzed by examining two cases:

- If $\bar{\mathcal{A}}(\infty) = 0$, then $\mathcal{Z}(\bar{\mathcal{A}}) = \mathbb{R}$. Hence $\sqrt{n}\bar{\mathcal{A}}(t) + \widehat{\mathcal{A}}(t) = \widehat{\mathcal{A}}(t) \ge \widehat{\mathcal{A}}(-\infty) = 0$, and the inf-part in the definition of Ψ_n is nonnegative for all $t \in \mathbb{R}$.
- If $\bar{\mathcal{A}}(\infty) \ne 0$, because $|\mathcal{J}(\bar{\mathcal{A}}^{\leftarrow})| < \infty$, there exists a constant $T_l < T_{\bar{\mathcal{A}}}$ (note that $T_{\bar{\mathcal{A}}} > -\infty$) such that $(-\infty, T_l)$ contains at most one flat interval of $\bar{\mathcal{A}}$ and $\bar{\mathcal{A}}(T_l) > 0$. Then, there exists some n_0 such that $\sqrt{n}\bar{\mathcal{A}}(T_l) \ge \sup_{t\in\mathbb{R}} |\widehat{\mathcal{A}}(t)|$ for all $n \ge n_0$. Hence, for $u \ge T_l$, $\sqrt{n}\bar{\mathcal{A}}(u) + \widehat{\mathcal{A}}(u) \ge \sqrt{n}\bar{\mathcal{A}}(T_l) + \widehat{\mathcal{A}}(u) \ge \sup_{t\in\mathbb{R}} |\widehat{\mathcal{A}}(t)| + \widehat{\mathcal{A}}(u) \ge 0$. If $(-\infty, T_l)$ contains one flat interval of $\bar{\mathcal{A}}$, then, as $\bar{\mathcal{A}}(-\infty) =$

$\widehat{\mathcal{A}}(-\infty) = 0$, $\bar{\mathcal{A}}$ and $\widehat{\mathcal{A}}$ are non-decreasing on this flat interval, we have $\sqrt{n}\bar{\mathcal{A}}(t) + \widehat{\mathcal{A}}(t) \geq 0$ for all t in this flat interval. On the other hand, if $(-\infty, T_l)$ contains no flat intervals, then the inf in (16) does not involve arguments in $(-\infty, T_l)$. As a result, the inf-part in (16) is nonnegative for all $t \in \mathbb{R}$ when $n \geq n_0$.

Combining the preceding arguments and the fact that the sup in (16) is nonnegative, it follows that $\Psi_n(t) \geq 0$, or equivalently, $\Delta_2^n(t) = 0$, for all $t \in \mathbb{R}$ and $n \geq n_0$. Therefore,

$$\mathcal{S}(\tilde{G} * F * \Delta_2^n) \to 0. \tag{34}$$

The third term in (32) can be analyzed by considering the difference of the two "appointment" processes (here, let $\tilde{\mathcal{G}}(\bar{\mathcal{A}}) = \tilde{\mathcal{Z}}(\bar{\mathcal{A}}) \cup [T_{\bar{\mathcal{A}}}, \infty)$):

$$\begin{aligned}
&\Delta_3^n(t) \\
&= \sup_{u \leq t}\{\sqrt{n}(\bar{\mathcal{A}}(u) - \bar{\mathcal{A}}(t)) + \widehat{\mathcal{A}}(u) - \widehat{\mathcal{A}}(t)\} \\
&\quad \wedge \inf_{u > t:\, u \in \mathcal{G}(\bar{\mathcal{A}})}\{\sqrt{n}(\bar{\mathcal{A}}(u) - \bar{\mathcal{A}}(t)) + \widehat{\mathcal{A}}(u) - \widehat{\mathcal{A}}(t)\} && (35) \\
&\to \inf_{\delta > 0} \sup_{u \leq t:\, \bar{\mathcal{A}}(u) \geq \bar{\mathcal{A}}(t) - \delta} \{\widehat{\mathcal{A}}(u) - \widehat{\mathcal{A}}(t)\} \\
&\quad \wedge \inf_{u \geq t:\, \bar{\mathcal{A}}(u) = \bar{\mathcal{A}}(t),\, u \in \tilde{\mathcal{G}}(\bar{\mathcal{A}})} \{\widehat{\mathcal{A}}(u) - \widehat{\mathcal{A}}(t)\} && (36) \\
&= \inf_{\delta > 0} \sup_{u \leq t:\, \bar{\mathcal{A}}(u) \geq \bar{\mathcal{A}}(t) - \delta} \{\widehat{\mathcal{A}}(u) - \widehat{\mathcal{A}}(t)\} \\
&\quad \wedge \left(\infty \times \mathbb{1}\left\{t \notin \tilde{\mathcal{G}}(\bar{\mathcal{A}})\right\}\right); && (37)
\end{aligned}$$

the limit in (36) is due to $|\mathcal{J}(\bar{\mathcal{A}}^{\leftarrow})| < \infty$ (the number of flat intervals is finite), $T_{\bar{\mathcal{A}}} > \sup\{u : u \in \mathcal{Z}(\bar{\mathcal{A}})\}$, $\bar{\mathcal{A}} \in D_{\wedge}(\mathbb{R})$ and $\sup_t |\widehat{\mathcal{A}}(t)| < \infty$; the equality in (37) is due to $\widehat{\mathcal{A}} \in D$ and $\mathrm{d}\widehat{\mathcal{A}}(t) \geq 0$ for $t \in \mathcal{Z}(\bar{\mathcal{A}})$; here, we use the convention that $\sup \emptyset = -\infty$ and $\inf \emptyset = \infty$. Based on $\bar{\mathcal{A}}$, any $t \in \mathbb{R}$ can be of two types:

- $(t \in \tilde{\mathcal{G}}(\bar{\mathcal{A}}))$ Note that the inf-sup part of (37) is nonnegative because the sup in the pre-limit of (35) is nonnegative. Hence, $\Delta_3^n(t) \to 0$. This case justifies the presence of the inf in the definition of Ψ_n – without it, the limit could be strictly positive on non-zero length intervals.
- $(t \notin \tilde{\mathcal{G}}(\bar{\mathcal{A}}))$ In this case, the second term is infinite; thus, it can be omitted. Hence, the inf-sup is achieved at either $u = t$ or $u = t-$ or $u = \tau_t-$, where the last two cases occur only when $\bar{\mathcal{A}}$

is continuous at t and τ_t, respectively; here, $\tau_t := \inf\{u \le t : \bar{\mathcal{A}}(u) = \bar{\mathcal{A}}(t)\} \le t$. This is due to $(\tau_t, t) \subseteq \mathcal{Z}(\bar{\mathcal{A}})$ and $\widehat{\mathcal{A}}$ being non-decreasing on that interval. This leads to $\Delta_3^n(t) \to (\widehat{\mathcal{A}}(\tau_t-) - \widehat{\mathcal{A}}(t))^+ \mathbb{1}\{\bar{\mathcal{A}}(\tau_t-) = \bar{\mathcal{A}}(t)\} \vee (\widehat{\mathcal{A}}(t-) - \widehat{\mathcal{A}}(t))^+ \mathbb{1}\{\bar{\mathcal{A}}(t-) = \bar{\mathcal{A}}(t)\}$. The first difference is relevant only when $\tau_t \in \mathcal{J}(\widehat{\mathcal{A}})$; otherwise, $\widehat{\mathcal{A}}(\tau_t-) = \widehat{\mathcal{A}}(\tau_t) \le \widehat{\mathcal{A}}(t-)$, and the second term achieves the maximum. Therefore, for the limit to be non-zero, at least one of the following is necessary: (i) $t \in \mathcal{J}(\widehat{\mathcal{A}})$ and $\bar{\mathcal{A}}(t-) = \bar{\mathcal{A}}(t)$, or (ii) $t > \tau_t \in \mathcal{J}(\widehat{\mathcal{A}})$ and $\bar{\mathcal{A}}(\tau_t-) = \bar{\mathcal{A}}(\tau_t) < \bar{\mathcal{A}}(t+\epsilon)$, for all $\epsilon > 0$. Due to $\widehat{\mathcal{A}} \in D(\mathbb{R})$ and $\bar{\mathcal{A}} \in D_\wedge(\mathbb{R})$, the set of such points is at most countably infinite.
Combining the preceding two cases render that $\Delta_3^n(t)$ vanishes for all but countably many t's; that is, for almost all $t \in \mathbb{R}$,

$$\Delta_3^n(t) \to 0. \tag{38}$$

Next, we consider $\mathcal{S}(\tilde{G} * F * \Delta_3^n)$. If $\sup\{u : u \in \mathcal{Z}(\bar{\mathcal{A}})\} = \infty$, then, as $|\mathcal{J}(\bar{\mathcal{A}}^{\leftarrow})| < \infty$ (the number of flat intervals is finite), there exists a constant $T_0 > 0$ such that $[T_0, \infty) \subseteq \mathcal{G}(\bar{\mathcal{A}})$; if $\sup\{u : u \in \mathcal{Z}(\bar{\mathcal{A}})\} < \infty$, choose any $T_0 \in (|T_{\bar{\mathcal{A}}}|, \infty)$. Hence, there exists $T_0 < \infty$ such that $(T_0, \infty) \subseteq \mathcal{G}(\bar{\mathcal{A}})$. Now, the last term in (32) is estimated using the triangle inequality: $\mathcal{S}(\tilde{G} * F * \Delta_3^n) \le \mathcal{S}(\tilde{G} * F * \Delta^n_{3,(-\infty,-T)}) + \mathcal{S}(\tilde{G} * F * \Delta^n_{3,[-T,T]}) + \mathcal{S}(\tilde{G} * F * \Delta^n_{3,(T,\infty)})$, where $\Delta^n_{3,I} = \{\Delta_3^n(t)\mathbb{1}\{t \in I\} : t \in \mathbb{R}\}$, for $I \in \{(-\infty, -T), [-T, T], (T, \infty)\}$ and $T > T_0 > 0$; here T is chosen such that $(-\infty, -T)$ contains at most one flat interval of $\bar{\mathcal{A}}$ and if $\bar{\mathcal{A}}(\infty) > 0$, $\inf_{\bar{\mathcal{A}}(u)>0:\, u\in\mathcal{G}(\bar{\mathcal{A}})}\{\bar{\mathcal{A}}(u)\} > \bar{\mathcal{A}}(-T)$. Each of the three terms in the preceding inequality can be upper-bounded:

$$\begin{aligned}
&\mathcal{S}(\tilde{G} * F * \Delta^n_{3,I}) \\
&\quad \le B \int_{-\infty}^{\infty} \left(F * |\Delta^n_{3,I}|(t) + G * F * |\Delta^n_{3,I}|(t) \right) \mathrm{d}t \\
&\quad \le 2B \int_I |\Delta_3^n(t)| \, \mathrm{d}t.
\end{aligned} \tag{39}$$

The three cases of interest can be analyzed as follows:

- $I = (-\infty, -T)$. If $\bar{\mathcal{A}}(\infty) = 0$, then the inf-part in (35) is non-negative; if $\bar{\mathcal{A}}(\infty) > 0$, then by noting that $\inf_{\bar{\mathcal{A}}(u)>0:\, u\in\mathcal{G}(\bar{\mathcal{A}})}\{\bar{\mathcal{A}}(u)\} > \bar{\mathcal{A}}(-T)$, there exists an n_0 such that for all $n \ge n_0$, $\sqrt{n}(\inf_{\bar{\mathcal{A}}(u)>0:\, u\in\mathcal{G}(\bar{\mathcal{A}})}\{\bar{\mathcal{A}}(u)\} - \bar{\mathcal{A}}(-T)) \ge 2\sup_u |\widehat{\mathcal{A}}(u)|$, which would then ensure that for all $t \in I$, the inf-part in (35) is non-negative. Consequently, for all $t \in I$ and $n \ge n_0$, one has $|\Delta_3^n(t)| \le \sup_{u\le t}\{\sqrt{n}(\bar{\mathcal{A}}(u) - \bar{\mathcal{A}}(t)) + \widehat{\mathcal{A}}(u) - \widehat{\mathcal{A}}(t)\} \le \sup_{u\le t} |\widehat{\mathcal{A}}(u) - \widehat{\mathcal{A}}(t)| \le 2\sup_{u\le t} |\widehat{\mathcal{A}}(u)|$. This, together with (39), renders $\mathcal{S}(\tilde{G} * F * \Delta^n_{3,(-\infty,-T)}) \le 2B \int_{-\infty}^{-T} |\Delta_3^n(t)| \, \mathrm{d}t \le 4B \int_{-\infty}^{-T} \sup_{u\le t} |\widehat{\mathcal{A}}(u)| \, \mathrm{d}t$, which can be made arbitrarily small due to assumption (17).

- $I = (T, \infty)$. For all $t \geq T > T_0$, one has $|\Delta_3^n(t)| \leq \sqrt{n}\bar{\mathcal{A}}(t) + \widehat{\mathcal{A}}(t) - \inf_{u \geq t}\{\sqrt{n}\bar{\mathcal{A}}(u) + \widehat{\mathcal{A}}(u)\} \leq \sqrt{n}\bar{\mathcal{A}}(t) + \widehat{\mathcal{A}}(t) - (\inf_{u \geq t}\{\sqrt{n}\bar{\mathcal{A}}(t) + \widehat{\mathcal{A}}(u)\}) = \widehat{\mathcal{A}}(t) - \inf_{u \geq t}\{\widehat{\mathcal{A}}(u)\} \leq \sup_{u \geq t} |\widehat{\mathcal{A}}(u) - \widehat{\mathcal{A}}(t)| \leq 2 \sup_{u \geq t} |\widehat{\mathcal{A}}(u) - \widehat{\mathcal{A}}(\infty)|$. This, together with (39), renders $\mathcal{S}(\tilde{G} * F * \Delta^n_{3,(T,\infty)}) \leq 2B \int_T^\infty |\Delta_3^n(t)| \, \mathrm{d}t \leq 4B \int_T^\infty \sup_{u \geq t} |\widehat{\mathcal{A}}(u) - \widehat{\mathcal{A}}(\infty)| \, \mathrm{d}t$, which can be made arbitrarily small due to assumption (17).
- $I = [-T, T]$. Based on (35) and the monotonicity of $\bar{\mathcal{A}}$, one has $|\Delta_3^n(t)| \leq \sup_{u \in \mathbb{R}} |\widehat{\mathcal{A}}(u) - \widehat{\mathcal{A}}(t)| \leq 2 \sup_t |\widehat{\mathcal{A}}(t)| < \infty$, and therefore $\mathcal{S}(\tilde{G} * F * \Delta^n_{3,[-T,T]}) \leq 2B \int_{-T}^T |\Delta_3^n(t)| \, \mathrm{d}t \leq 4BT \sup_t |\Delta_3^n(t)| \leq 8BT \sup_t |\widehat{\mathcal{A}}(t)| < \infty$. This bound, (38), and dominated convergence yield $\mathcal{S}(\tilde{G} * F * \Delta^n_{3,[-T,T]}) \to 0$.

The preceding three cases together imply $\mathcal{S}(\tilde{G} * F * \Delta_3^n) \to 0$.

Finally, combining (32), (33), (34), and the preceding limit yields the desired $\tilde{G} * F * \widehat{\mathcal{A}}^n \xrightarrow{s} \tilde{G} * F * \widehat{\mathcal{A}}$. This completes the proof of Lemma 2. □

5.7. Proof of Corollary 2

First, the equality in (19) is due to Lemma 2. Hence, it is sufficient to prove that, for any sequence of appointment plans $\{\mathcal{A}^n\}_n$, one has

$$\liminf_{n \to \infty} \frac{1}{\sqrt{n}} C\left(X_{\mathcal{A}^n} - \gamma^n\right) \geq C(Z\,\Gamma + \tilde{G} * F * \widehat{\mathcal{A}}_* - \widehat{\gamma}). \tag{40}$$

Second, the convexity of $\mathcal{A} \mapsto C(X_{\mathcal{A}} - \gamma)$, Jensen's inequality and the triangle inequality imply $C(X_{\mathcal{A}^n} - \gamma^n) \geq C(\tilde{G} * F * \mathcal{A}^n - \gamma^n) \geq C(\tilde{G} * F * \mathcal{A}^n - n\gamma) - \sqrt{n}\mathcal{S}(\widehat{\gamma}^n) = n\, C(\tilde{G} * F * \bar{\mathcal{A}}^n - \gamma) - \sqrt{n}\mathcal{S}(\widehat{\gamma}^n)$, where $\mathcal{S}(\widehat{\gamma}^n) \to \mathcal{S}(\widehat{\gamma}) < \infty$ due to $\widehat{\gamma}^n \xrightarrow{s} \widehat{\gamma}$. Hence, it is enough to consider $\{\mathcal{A}^n\}_n$ satisfying $\liminf_{n \to \infty} C(\tilde{G} * F * \bar{\mathcal{A}}^n - \gamma) = 0$ (otherwise, the left-hand side of (40) diverges). Moreover, one can consider only (sub)sequences $\{\mathcal{A}^n\}_n$ such that $C(\tilde{G} * F * \bar{\mathcal{A}}^n - \gamma) \to 0$. Assumption (18) implies that this $\{\mathcal{A}^n\}_n$ satisfies $\sup_n \bar{\mathcal{A}}^n(\infty) < \infty$.

Third, note that $\frac{1}{\sqrt{n}} C\left(X_{\mathcal{A}^n} - \gamma^n\right) \geq C(\widehat{Z}_{\mathcal{A}^n} + \tilde{G} * F * \widehat{\mathcal{A}}^n - \widehat{\gamma}) - C(\widehat{\gamma}^n - \widehat{\gamma}) - \sqrt{n}\mathcal{S}(\tilde{G} * F * \bar{\mathcal{A}} - \gamma) = C(\widehat{Z}_{\mathcal{A}^n} + \tilde{G} * F * \widehat{\mathcal{A}}^n - \widehat{\gamma}) - C(\widehat{\gamma}^n - \widehat{\gamma})$, where the inequality is due to the triangle inequality and

$C(\gamma - \tilde{G} * F * \bar{\mathcal{A}}) \le \mathcal{S}(\tilde{G} * F * \bar{\mathcal{A}} - \gamma)$, and the equality is due to $\gamma \overset{s}{=} \tilde{G} * F * \bar{\mathcal{A}}$. This and the definition of $\widehat{\mathcal{A}}_*$ (from $\mathcal{A}^n \in D_{\mathbb{N}}(\mathbb{R})$, $\bar{\mathcal{A}} \in D_\wedge(\mathbb{R})$, (14), and $\mathrm{d}\widehat{\mathcal{A}}^n(t) \ge 0$ for $t \in \mathcal{Z}(\bar{\mathcal{A}})$) yield

$$\begin{aligned}
&\frac{1}{\sqrt{n}} C\left(X_{\mathcal{A}^n} - \gamma^n\right) \\
&\ge C(Z\Gamma + \tilde{G} * F * \widehat{\mathcal{A}}_* - \widehat{\gamma}) - C(\widehat{\gamma}^n - \widehat{\gamma}) \\
&\quad + C(\widehat{Z}_{\mathcal{A}^n} + \tilde{G} * F * \widehat{\mathcal{A}}^n - \widehat{\gamma}) \\
&\quad - C(Z\Gamma_n + \tilde{G} * F * \widehat{\mathcal{A}}^n - \widehat{\gamma}) \\
&\quad + C(Z\Gamma_n + \tilde{G} * F * \widehat{\mathcal{A}}^n - \widehat{\gamma}) \\
&\quad - C(Z\Gamma + \tilde{G} * F * \widehat{\mathcal{A}}^n - \widehat{\gamma}).
\end{aligned} \tag{41}$$

Finally, (40) follows from (41), assumption $\widehat{\gamma}^n \overset{s}{\to} \widehat{\gamma}$, and Lemmas 7 and 8 below. □

LEMMA 7. *If $C(\tilde{G} * F * \bar{\mathcal{A}}^n - \gamma) \to 0$, $\sup_n \bar{\mathcal{A}}^n(\infty) < \infty$, and there is a unique $\bar{\mathcal{A}} \in D_\wedge(\mathbb{R})$ such that $C(\tilde{G} * F * \bar{\mathcal{A}} - \gamma) = 0$, then $C(Z\Gamma_n + \tilde{G} * F * \widehat{\mathcal{A}}^n - \widehat{\gamma}) - C(Z\Gamma + \tilde{G} * F * \widehat{\mathcal{A}}^n - \widehat{\gamma}) \to 0$.*

Proof. By the triangle inequality in Proposition 1, it is enough to prove $\Gamma_n \overset{s}{\to} \Gamma$. Because $\sup_n \bar{\mathcal{A}}^n(\infty) < \infty$, for any subsequence, by Helly's Selection Theorem, there is a subsequence $\{\bar{\mathcal{A}}_{n_k}\}_k$ that converges to some $\bar{\mathcal{A}}_\diamond$ for each $x \in \mathbb{R}$. Note that $\bar{\mathcal{A}}_\diamond \in V_\wedge(\mathbb{R})$ but not necessarily $\bar{\mathcal{A}}_\diamond \in D_\wedge(\mathbb{R})$. Similar to the proof of Proposition 2, $\tilde{G} * F * \bar{\mathcal{A}}^{n_k} \to \tilde{G} * F * \bar{\mathcal{A}}_\diamond$ pointwise. Then, $\tilde{G} * F * \bar{\mathcal{A}}^{n_k} \overset{s}{\to} \tilde{G} * F * \bar{\mathcal{A}}_\diamond$ can be proved using an argument similar to the proof of Lemma 1. The main difference lies in the proof of $\mathcal{S}(\Delta^{n_k}_{[-T,T]}) \to 0$ for any given $T > 0$; here, $\Delta^{n_k}_I := \{\Delta^{n_k}_I(t) = (n_k \bar{\mathcal{A}}_\diamond(t) - \mathcal{A}^{n_k}(t))\mathbb{1}\{t \in I\} : t \in \mathbb{R}\}$. Instead of obtaining a result similar to (27), $\mathcal{S}(\Delta^{n_k}_{[-T,T]}) \to 0$ for any given $T > 0$ can be established using Scheffe's lemma and the dominated convergence theorem.

Similar to the argument in the proof of Proposition 2, $(\tilde{G} * F)^2 * \bar{\mathcal{A}}^{n_k} \to (\tilde{G} * F)^2 * \bar{\mathcal{A}}_\diamond$ pointwise. Due to $(\tilde{G} * F)^2 * \bar{\mathcal{A}}^{n_k} \le \tilde{G} * F * \bar{\mathcal{A}}^{n_k}$ and $(\tilde{G} * F)^2 * \bar{\mathcal{A}}_\diamond \le \tilde{G} * F * \bar{\mathcal{A}}_\diamond$ the dominated convergence theorem implies $(\tilde{G} * F)^2 * \bar{\mathcal{A}}^{n_k} \overset{s}{\to} (\tilde{G} * F)^2 * \bar{\mathcal{A}}_\diamond$.

From $C(\tilde{G} * F * \bar{\mathcal{A}}^n - \gamma) \to 0$, it follows that $C(\tilde{G} * F * \bar{\mathcal{A}}^{n_k} - \gamma) \to 0$, which, combined with $\tilde{G} * F * \bar{\mathcal{A}}^{n_k} \overset{s}{\to} \tilde{G} * F * \bar{\mathcal{A}}_\diamond$, results in $C(\tilde{G} * F * \bar{\mathcal{A}}_\diamond - \gamma) = 0$ (since $0 \le C(\tilde{G} * F * \bar{\mathcal{A}}_\diamond - \gamma) \le C(\tilde{G} * F * \bar{\mathcal{A}}^{n_k} - \gamma) + \mathcal{S}(\tilde{G} * F * \bar{\mathcal{A}}_\diamond - \tilde{G} * F * \bar{\mathcal{A}}^{n_k}) \to 0$). By the uniqueness of $\bar{\mathcal{A}} \in D_\wedge(\mathbb{R})$ satisfying $C(\tilde{G} * F * \bar{\mathcal{A}} - \gamma) = 0$, $\bar{\mathcal{A}}$ is the RCLL modification of $\bar{\mathcal{A}}_\diamond$. Hence, $\tilde{G} * F * \bar{\mathcal{A}}^{n_k} \overset{s}{\to} \tilde{G} * F * \bar{\mathcal{A}}$ and $(\tilde{G} * F)^2 * \bar{\mathcal{A}}^{n_k} \overset{s}{\to} (\tilde{G} * F)^2 * \bar{\mathcal{A}}$.

As the preceding holds for every subsequence, $\tilde{G} * F * \bar{\mathcal{A}}^n \xrightarrow{s} \tilde{G} * F * \bar{\mathcal{A}}$ and $(\tilde{G} * F)^2 * \bar{\mathcal{A}}^n \xrightarrow{s} (\tilde{G} * F)^2 * \bar{\mathcal{A}}$. Therefore, $\Gamma_n \xrightarrow{s} \Gamma$. □

LEMMA 8. *If the conditions of Lemma 7 hold, then* $C(\widehat{Z}_{\mathcal{A}^n} + \tilde{G} * F * \widehat{\mathcal{A}}^n - \widehat{\gamma}) - C(Z\Gamma_n + \tilde{G} * F * \widehat{\mathcal{A}}^n - \widehat{\gamma}) \to 0$.

Proof. Let $Y^n_{\mathcal{A}}(t) := \widehat{Z}_{\mathcal{A}^n}(t) + \tilde{G} * F * \widehat{\mathcal{A}}^n(t) - \widehat{\gamma}(t)$ and $Y^n_{\Gamma}(t) := Z\Gamma_n(t) + \tilde{G} * F * \widehat{\mathcal{A}}^n(t) - \widehat{\gamma}(t)$. We first prove that, for all $t \in \mathbb{R}$,

$$\mathbb{E}(Y^n_{\mathcal{A}}(t))^{\pm} - \mathbb{E}(Y^n_{\Gamma}(t))^{\pm} \to 0. \tag{42}$$

Consider any subsequence (still denoted by n) such that $\tilde{G} * F * \widehat{\mathcal{A}}^n(t) \to a(t)$ for a constant $a(t) \in [-\infty, \infty]$. Two cases are of interest:

- If $a(t) \in (-\infty, \infty)$ then, as $|x^{\pm} - y^{\pm}| \le |x - y|$, $\mathbb{E}(Y^n_{\mathcal{A}}(t))^{\pm} - \mathbb{E}(\widehat{Z}_{\mathcal{A}^n}(t) + a(t) - \widehat{\gamma}(t))^{\pm} \to 0$ and $\mathbb{E}(Y^n_{\Gamma}(t))^{\pm} - \mathbb{E}(Z\Gamma_n(t) + a(t) - \widehat{\gamma}(t))^{\pm} \to 0$. Similar to (31), $\mathbb{E}(\widehat{Z}_{\mathcal{A}^n}(t) + a(t) - \widehat{\gamma}(t))^{\pm} - \mathbb{E}(Z\Gamma_n(t) + a(t) - \widehat{\gamma}(t))^{\pm} \to 0$. Hence, (42) holds.
- If $a(t) = \infty$ (the case $a(t) = -\infty$ is similar), then because $\mathbb{E}(Y^n_{\mathcal{A}}(t))^{+} - \mathbb{E}(Y^n_{\Gamma}(t))^{+} = (\mathbb{E}[Y^n_{\mathcal{A}}(t)] + \mathbb{E}(Y^n_{\mathcal{A}}(t))^{-}) - (\mathbb{E}[Y^n_{\Gamma}(t)] + \mathbb{E}(Y^n_{\Gamma}(t))^{-})$, it is enough to prove $\mathbb{E}[\widehat{Z}_{\mathcal{A}^n}(t)] - \mathbb{E}[Z\Gamma_n(t)] \to 0$ and $\mathbb{E}(Y^n_{\mathcal{A}}(t))^{-} - \mathbb{E}(Y^n_{\Gamma}(t))^{-} \to 0$. The former can be proved similar to the proof of (31), whereas the latter can be proved by $\tilde{G} * F * \widehat{\mathcal{A}}^n(t) \to \infty$ and the facts that $\sup_n \mathbb{E}[\widehat{Z}^2_{\mathcal{A}^n}(t)] < \infty$ and $\sup_n \Gamma^2_n(t) < \infty$ (because $\sup_n \bar{\mathcal{A}}^n(\infty) < \infty$).

Combining the above two cases, we obtain (42).

Then, from $c_u(t) + c_o(t) \le B$, $\sup_n \bar{\mathcal{A}}^n(\infty) < \infty$, and (10), we have $\sup_n \mathcal{S}(\tilde{G} * F * \bar{\mathcal{A}}^n) < \infty$, and hence, $\sup_n \mathcal{S}(\Gamma_n) < \infty$. Adding this to (42), $|\mathbb{E}(Y^n_{\mathcal{A}}(t))^{\pm} - \mathbb{E}(Y^n_{\Gamma}(t))^{\pm}| \le \mathbb{E}|\widehat{Z}_{\mathcal{A}^n}(t)| + \Gamma_n(t)\,\mathbb{E}|Z| \le \Gamma_n(t) + \Gamma_n(t)\,\mathbb{E}|Z| \le 2\Gamma_n(t)$, and applying the dominated convergence theorem (see the proof of Lemma 7) yield the statement of the lemma. □

5.8. Proof of Lemma 4

The definition of C and the structure of $\bar{\mathcal{A}}$ (see Remark 8) render

$$\begin{aligned} &C(Z\Gamma + \tilde{G} * \widehat{\mathcal{A}} - \widehat{\gamma}) \\ &\quad = \int_{t \notin \tilde{\mathcal{Z}}(\bar{\mathcal{A}})} \mathbb{E}c(t, Z\Gamma(t) + \tilde{G} * \widehat{\mathcal{A}}(t) - \widehat{\gamma}(t))\, \mathrm{d}t \\ &\quad + \sum_{y \in \mathcal{J}(\bar{\mathcal{A}}^{\leftarrow})} \int_{\mathbf{z}_y(\bar{\mathcal{A}})} \mathbb{E}c(t, Z\Gamma(t) + \tilde{G} * \widehat{\mathcal{A}}(t) - \widehat{\gamma}(t))\, \mathrm{d}t. \end{aligned}$$

The first integral can be lower bounded by solving a standard (instantaneous) newsvendor problem:

$$\begin{aligned}
&\int_{t\notin\tilde{\mathcal{Z}}(\bar{\mathcal{A}})} \mathbb{E}c(t, Z\,\Gamma(t) + \tilde{G} * \widehat{\mathcal{A}}(t) - \widehat{\gamma}(t))\ \mathrm{d}t \\
&\geq \int_{t\notin\tilde{\mathcal{Z}}(\bar{\mathcal{A}})} \inf_{x\in\mathbb{R}} \mathbb{E}c(t, Z\,\Gamma(t) + x - \widehat{\gamma}(t))\ \mathrm{d}t \\
&= \int_{t\notin\tilde{\mathcal{Z}}(\bar{\mathcal{A}})} \mathbb{E}c(t, Z\,\Gamma(t) + \tilde{G} * \widehat{\mathcal{A}}_*(t) - \widehat{\gamma}(t))\ \mathrm{d}t \\
&= \int_{t\notin\tilde{\mathcal{Z}}(\bar{\mathcal{A}})} (c_o(t) + c_u(t))\,\Gamma(t)\,\varphi(\beta(t))\ \mathrm{d}t,
\end{aligned}$$

where $(\tilde{G} * \widehat{\mathcal{A}}_*)(t) = \arg\min_{x\in\mathbb{R}} \mathbb{E}c(t, Z\,\Gamma(t) + x - \widehat{\gamma}(t)) = \widehat{\gamma}(t) - \Gamma(t)\,\beta(t)$, and, thus, also (20) due to $\mathrm{d}(\tilde{G} * \widehat{\mathcal{A}}_*)(t) = p\,\mathrm{d}\widehat{\mathcal{A}}_*(t) - (\tilde{G} * \widehat{\mathcal{A}}_*)(t)\,\mu\,\mathrm{d}t$. On the other hand, for each $\mathtt{z}_y(\bar{\mathcal{A}})$, one has

$$\begin{aligned}
&\int_{\mathtt{z}_y(\bar{\mathcal{A}})} \mathbb{E}c(t, Z\,\Gamma(t) + \tilde{G} * \widehat{\mathcal{A}}(t) - \widehat{\gamma}(t))\ \mathrm{d}t \\
&\geq \inf_{a\in D_\wedge(\mathtt{z}_y(\bar{\mathcal{A}}))} \int_{\mathtt{z}_y(\bar{\mathcal{A}})} \mathbb{E}c(t, Z\,\Gamma(t) + \tilde{G} * a(t) - \widehat{\gamma}(t))\,\mathrm{d}t \\
&= \int_{\mathtt{z}_y(\bar{\mathcal{A}})} \mathbb{E}c(t, Z\,\Gamma(t) + \tilde{G} * a_{y*}(t) - \widehat{\gamma}(t))\ \mathrm{d}t,
\end{aligned}$$

where the existence of $a_{y*} = \{a_{y*}(t),\, t \in \mathtt{z}_y(\bar{\mathcal{A}})\} \in D_\wedge(\mathtt{z}_y(\bar{\mathcal{A}}))$ follows from the same argument used in the proof of Proposition 2; even though a_{y*} describes a diffusion-scale appointment plan, Proposition 2 is applicable because only non-decreasing (diffusion) plans are relevant on $\mathtt{z}_y(\bar{\mathcal{A}})$. Moreover, a_{y*} is obtained by considering $\mathtt{z}_y(\bar{\mathcal{A}})$ in isolation. In order to recover $\widehat{\mathcal{A}}_*$ on $\mathbb{R}$, one can decompose $(\tilde{G} * a_{y*})(t)$, $t \in \mathtt{z}_y(\bar{\mathcal{A}})$, as follows: $(\tilde{G} * a_{y*})(t) = \int_{l_y}^{t} \tilde{G}(t-u)\ \mathrm{d}\widehat{\mathcal{A}}_*(u) + \int_{-\infty}^{l_y-} \tilde{G}(t-u)\ \mathrm{d}\widehat{\mathcal{A}}_*(u) = \int_{l_y}^{t} \tilde{G}(t-u)\ \mathrm{d}\widehat{\mathcal{A}}_*(u) + \tilde{G}(t-l_y)\,\frac{1}{p}\tilde{G} * \widehat{\mathcal{A}}_*(l_y-)$, where the last equality is due to $\tilde{G}(t-u) = \tilde{G}(t-l_y)\,\frac{1}{p}\,\tilde{G}(l_y-u)$, $t > l_y$; the two terms correspond to arrival after and before l_y, respectively. This completes the proof. □

References

Ahmadi-Javid A, Jalali Z, Klassen KJ (2017) Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research* 258(1):3–34.

Armony M, Atar R, Honnappa H (2019) Asymptotically optimal appointment schedules. *Mathematics of Operations Research* 44(4):1345–1380.

Ash RB (2000) *Probability and Measure Theorey* (Academic Press), 2nd edition.

Berg B, Denton B (2012) Appointment planning and scheduling in outpatient procedure centers. Hall R, ed., *Handbook of Healthcare System Scheduling*, volume 168 of *International Series in Operations Research & Management Science*, chapter 6, 131–154 (Springer).

Billingsley P (1995) *Probability and Measure* (New York, NY: Wiley), 3rd edition.

Borst S, Mandelbaum A, Reiman MI (2004) Dimensioning large call centers. *Operations Research* 52(1):17–34.

Brown L, Gans N, Mandelbaum A, Sakov A, Shen H, Zeltyn S, Zhao L (2005) Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association* 100(469):36–50.

Cayirli T, Veral E (2003) Outpatient scheduling in health care: A review of literature. *Production and Operations Management* 12(4):519–549.

Chatterjee S, Hebaish Y, Aprahamian H, Ntaimo L (2025) An optimization-based scheduling methodology for appointment systems with heterogeneous customers and nonstationary arrival processes. *INFORMS Journal on Computing* .

Conradie J (2015) Asymmetric norms, cones and partial orders. *Topology and its Applications* 193:100–115.

Dantas LF, Fleck JL, Oliveira FLC, Hamacher S (2018) No-shows in appointment scheduling–a systematic literature review. *Health Policy* 122(4):412–421.

Defraeye M, Van Nieuwenhuyse I (2016) Staffing and scheduling under nonstationary demand for service: A literature review. *Omega* 58:4–25.

Deng Y, Shen S (2016) Decomposition algorithms for optimizing multi-server appointment scheduling with chance constraints. *Mathematical Programming* 157:245–276.

Eick SG, Massey WA, Whitt W (1993) The physics of the $M_t/G/\infty$ queue. *Operations Research* 41(4):731–742.

Erlang A (1948) On the rational determination of the number of circuits. *The Life and Works of AK Erlang* 216–221.

Ethier SN, Kurtz TG (1986) *Markov Processes: Characterization and Convergence*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics (New York: John Wiley & Sons Inc.).

Feldman Z, Mandelbaum A, Massey WA, Whitt W (2008) Staffing of time-varying queues to achieve time-stable performance. *Management Science* 54(2):324–338.

Ferrer J, Gregori V, Alegre C (1993) Quasi-uniform structures in linear lattices. *The Rocky Mountain Journal of Mathematics* 877–884.

Glynn P, Whitt W (1991) A new view of the heavy-traffic limit theorem for the infinite-server queue. *Adv. Appl. Probab.* 23:188–209.

Gocgun Y, Puterman ML (2014) Dynamic scheduling with due dates and time windows: An application to chemotherapy patient appointment booking. *Health Care Management Science* 17(1):60–76.

Grassmann WK (1988) Finding the right number of servers in real-world queuing systems. *Interfaces* 18(2):94–104.

Gupta D, Denton B (2008) Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions* 40(9):800–819.

Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Operations Research* 29(3):567–588.

Hall RW (2012) *Handbook of Healthcare System Scheduling* (Springer).

Huang J, Mandelbaum A, Momčilović P (2022) Appointment-driven service systems with many servers. *Queueing Systems* 1–3.

Jansson B (1966) Choosing a good appointment system—a study of queues of the type (d, m, 1). *Operations Research* 14(2):292–312.

Karlïn S (1955) On the renewal equation. *Pacific J. Math* 5:229–257.

Kemper B, Klaassen CA, Mandjes M (2014) Optimized appointment scheduling. *European Journal of Operational Research* 239(1):243–255.

Kim SH, Whitt W, Cha WC (2018) A data-driven model of an appointment-generated arrival process at an outpatient clinic. *INFORMS Journal on Computing* 30(1):181–199.

Kolesar PJ, Green LV (1998) Insights on service system design from a normal approximation to Erlang's delay formula. *Production and Operations Management* 7(3):282–293.

Krichagina E, Puhalskii A (1997) A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing Syst. Theory Appl.* 25(1-4):235–280.

Kuiper A, de Mast J, Mandjes M (2021) The problem of appointment scheduling in outpatient clinics: A multiple case study of clinical practice. *Omega* 98:102122.

Kuiper A, Lee RH (2022) Appointment scheduling for multiple servers. *Management Science* 68(10):7422–7440.

Lamé G, Jouini O, Stal-Le Cardinal J (2016) Outpatient chemotherapy planning: A literature review with insights from a case study. *IIE Transactions on Healthcare Systems Engineering* 6(3):127–139.

Leon Y (2024) *Optimizing Visitor Load at National Parks: Balancing Accessibility and Overcrowding through Online Reservation Systems*. Master's thesis, Technion – Israel Institute of Technology, Haifa, Israel, URL `https://gality.net.technion.ac.il/files/2024/12/National_parks_Yamit_MSc_thesis.pdf`.

Lipscomb N, Liu X, Kulkarni VG (2024) Asymptotically optimal appointment scheduling in the presence of patient unpunctuality. URL `https://arxiv.org/abs/2412.18215`.

Liu E, Ma X, Sauré A, Weber L, Puterman ML, Tyldesley S (2019) Improving access to chemotherapy through enhanced capacity planning and patient scheduling. *IISE Transactions on Healthcare Systems Engineering* 9(1):1–13.

Liu Y, Whitt W (2012a) The G_t/GI/s_t+GI many-server fluid queue. *Queueing Systems* 71(4):405–444.

Liu Y, Whitt W (2012b) Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Operations Research* 60(6):1551–1564.

Luo J, Kulkarni VG, Ziya S (2015) A tandem queueing model for an appointment-based service system. *Queueing Systems* 79(1):53–85.

Mandelbaum A (2003) QED Q's: Telephone call/contact centers. *Eurandom Workshop on Heavy Traffic Analysis, and Process Limits of Stochastic Networks.* (Available from AM upon request).

Mandelbaum A, Massey W (1995) Strong approximations for time-dependent queues. *Math. Oper. Res.* 20(1):33–64.

Mandelbaum A, Momčilović P, Trichakis N, Kadish S, Leib R, Bunnell CA (2020) Data-driven appointment-scheduling under uncertainty: The case of an infusion unit in a cancer center. *Management Science* 66(1):243–270.

Momčilović P, Mandelbaum A, Carmeli N, Armony M, Yom-Tov G (2022) Resource-driven activity-networks (RANs): A modelling framework for complex operations, preprint.

Newell GF (1968a) Queues with time-dependent arrival rates I: The transition through saturation. *Journal of Applied Probability* 5(2):436–451.

Newell GF (1968b) Queues with time-dependent arrival rates II—The maximum queue and the return to equilibrium. *Journal of Applied Probability* 5(3):579–590.

Palm C (1988) Intensity variations in teletraffic (in german). ericsson technics, 44: 1–189, 1943. *English translation by North Holland, Amsterdam* .

Puhalskii A, Reed J (2010) On many-server queues in heavy traffic. *Ann. Appl. Probab.* 20(1):129–195.

Saville CE, Smith HK, Bijak K (2019) Operational research techniques applied throughout cancer care services: A review. *Health Systems* 8(1):52–73.

Schwarz JA, Selinka G, Stolletz R (2016) Performance analysis of time-dependent queueing systems: Survey and classification. *Omega* 63:170–189.

Shnits B, Bendavid I, Marmor YN (2020) An appointment scheduling policy for healthcare systems with parallel servers and pre-determined quality of service. *Omega* 97:102095.

Soltani M, Samorani M, Kolfal B (2019) Appointment scheduling with multiple providers and stochastic service times. *European Journal of Operational Research* 277(2):667–683.

van Leeuwaarden JSH, Mathijsen BWJ, Zwart B (2019) Economies-of-scale in many-server queueing systems: Tutorial and partial review of the QED Halfin–Whitt heavy-traffic regime. *SIAM Review* 61(3):403–440.

Whitt W (1992) Understanding the efficiency of multi-server service systems. *Management Science* 38(5):708–723.

Whitt W (2013) OM Forum — offered load analysis for staffing. *Manufacturing & Service Operations Management* 15(2):166–169.

Whitt W (2018) Time-varying queues. *Queueing models and service management* 1(2).

Zacharias C, Pinedo M (2017) Managing customer arrivals in service systems with multiple identical servers. *Manufacturing & Service Operations Management* 19(4):639–656.

Zhao P, Yoo I, Lavoie J, Lavoie BJ, Simoes E (2017) Web-based medical appointment systems: A systematic review. *Journal of Medical Internet Research* 19(4):e134.

Zhou S, Ding Y, Huh WT, Wan G (2021) Constant job-allowance policies for appointment scheduling: Performance bounds and numerical analysis. *Production and Operations Management* .

Additional Comments/Results

EC.1. Weak Convergence of Schedules (Remark 2)

Define weak convergence for functions in $D_\wedge$, $\bar{\mathcal{A}}^n \Rightarrow \bar{\mathcal{A}}$, by $\bar{\mathcal{A}}^n(t) \to \bar{\mathcal{A}}(t)$, for all $t \in \mathbb{R}$ at which $\bar{\mathcal{A}}$ is continuous, as well as $\bar{\mathcal{A}}^n(\infty) \to \bar{\mathcal{A}}(\infty)$; see (Ash 2000, p. 125) for finite measure.

REMARK EC.1 (ESCAPE OF MASS TO INFINITY). Convergence at infinity, in the definition, prevents escape of mass to infinity. For example, let $\bar{\mathcal{A}}^n(t) = (t-n)^+ \wedge 1$, which is the CDF of a random variable that is uniformly distributed over $[n, n+1]$. Then, mass escapes to infinity in the sense that $\bar{\mathcal{A}}^n(t) \to \bar{\mathcal{A}}(t) \equiv 0$ for all $t \in \mathbb{R}$.

Denote by $\mathcal{J}(x)$ the set of discontinuity points in $x : \mathbb{R} \to \mathbb{R}$. Recall that $\mathcal{J}(x)$ is at most countable for $x \in D$. For $x_1, x_2 : \mathbb{R} \to \mathbb{R}$, define $\mathcal{J}(x_1) \oplus \mathcal{J}(x_2) := \{u : u = u_1 + u_2, \text{ for some } u_1 \in \mathcal{J}(x_1),\, u_2 \in \mathcal{J}(x_2)\}$.

LEMMA EC.1. *Suppose* $\{\bar{\mathcal{A}}^n\}_n \subseteq D_\wedge$ *is such that* $\sup_n \bar{\mathcal{A}}^n(\infty) < \infty$ *and* $\bar{\mathcal{A}}^n \Rightarrow \bar{\mathcal{A}}$. *If* $H \in D_+(\mathbb{R})$ *is bounded, then* $H * \bar{\mathcal{A}}^n(t) \to H * \bar{\mathcal{A}}(t)$ *for all* $t \notin \mathcal{J}(H) \oplus \mathcal{J}(\bar{\mathcal{A}})$. *In particular, the convergence is almost everywhere in* $\mathbb{R}$, *as the set* $\mathcal{J}(H) \oplus \mathcal{J}(\bar{\mathcal{A}})$ *is at most countable.*

Proof. Let X^n and X be random variables with CDFs $\bar{\mathcal{A}}^n(\cdot)/\bar{\mathcal{A}}^n(\infty)$ and $\bar{\mathcal{A}}(\cdot)/\bar{\mathcal{A}}(\infty)$, respectively. From the assumption, it follows that $X^n \Rightarrow X$. For any $t \notin \mathcal{J}(H) \oplus \mathcal{J}(\bar{\mathcal{A}})$, introduce $h(x) = H(t-x)$ and note that $\mathcal{J}(h) = t - \mathcal{J}(H)$, which implies $\mathcal{J}(h) \cap \mathcal{J}(\bar{\mathcal{A}}) = (t - \mathcal{J}(H)) \cap \mathcal{J}(\bar{\mathcal{A}}) = \emptyset$ because $t \notin \mathcal{J}(H) \oplus \mathcal{J}(\bar{\mathcal{A}})$. Furthermore, the preceding definitions yield

$$
\begin{aligned}
&\mathbb{P}[X \in \mathcal{J}(h)] \\
&\quad = \frac{1}{\bar{\mathcal{A}}(\infty)} \int_{-\infty}^{\infty} \mathbb{1}\{s \in \mathcal{J}(h)\} \, \mathrm{d}\bar{\mathcal{A}}(s) \\
&\quad = \frac{1}{\bar{\mathcal{A}}(\infty)} \int_{-\infty}^{\infty} \mathbb{1}\{s \in t - \mathcal{J}(H)\} \, \mathrm{d}\bar{\mathcal{A}}(s) \\
&\quad = \frac{1}{\bar{\mathcal{A}}(\infty)} \sum_{u \in \mathcal{J}(H)} \int_{-\infty}^{\infty} \mathbb{1}\{s = t - u\} \, \mathrm{d}\bar{\mathcal{A}}(s) \\
&\quad = \frac{1}{\bar{\mathcal{A}}(\infty)} \sum_{u \in \mathcal{J}(H)} \Delta\bar{\mathcal{A}}(t-u),
\end{aligned}
$$

where the notation $\Delta f(\cdot) := f(\cdot) - f(\cdot-)$ stands for jump-sizes of the RCLL function f, and the summation over $\mathcal{J}(H)$ is at most countable. Then, $t - u \notin \mathcal{J}(\bar{\mathcal{A}})$ for each $u \in \mathcal{J}(H)$, as $(t -$

$\mathcal{J}(H)) \cap \mathcal{J}(\bar{\mathcal{A}}) = \emptyset$. It follows that $\Delta\bar{\mathcal{A}}(t-u) = 0$, and consequently $\mathbb{P}[X \in \mathcal{J}(h)] = 0$. From the continuous-mapping theorem, we conclude that $H(t - X^n) = h(X^n) \Rightarrow h(X) = H(t - X)$, for $t \notin \mathcal{J}(H) \oplus \mathcal{J}(\bar{\mathcal{A}})$. Finally, H is bounded by assumption, and hence $\mathbb{E}H(t - X^n) \to \mathbb{E}H(t - X)$, for $t \notin \mathcal{J}(H) \oplus \mathcal{J}(\bar{\mathcal{A}})$; in particular, the latter convergence holds for almost all $t \in \mathbb{R}$. □

REMARK EC.2. Vague convergence of $\bar{\mathcal{A}}^n \xrightarrow{v} \bar{\mathcal{A}}$ is insufficient to claim $F * \bar{\mathcal{A}}^n \xrightarrow{v} F * \bar{\mathcal{A}}$. For example, let $\bar{\mathcal{A}}^n(t) = n^2\mathbb{1}\{t \geq n\}$ and $F(t) = 6\pi^{-2}\sum_{k\leq\lfloor t\rfloor} k^{-2}$, $t \leq 0$. Then, $\bar{\mathcal{A}}^n \xrightarrow{v} \bar{\mathcal{A}} \equiv 0$. However $F * \bar{\mathcal{A}}^n(0) = n^2 \cdot 6\pi^{-2}\sum_{k=n}^{\infty} k^{-2} \geq 6\pi^{-2} > 0$, whereas $F * \bar{\mathcal{A}}(0) = 0$.

COROLLARY EC.1. *Suppose* $\{\bar{\mathcal{A}}^n\}_n \subseteq D_\wedge$ *is such that* $\sup_n \bar{\mathcal{A}}^n(\infty) < \infty$ *and* $\bar{\mathcal{A}}^n \Rightarrow \bar{\mathcal{A}}$. *Then,* $\tilde{G} * F * \bar{\mathcal{A}}^n \to \tilde{G} * F * \bar{\mathcal{A}}$ *almost everywhere.*

EC.2. Examples

EXAMPLE EC.1. Let $c_u(t) = c_o(t) = \mathbb{1}\{t \in [-1, 2]\}$, $\gamma(t) = \mathbb{1}\{t \in [0, 1]\}$, $G(t) = \mathbb{1}\{t \geq 1/2\}$ (the service duration is 1/2), and $F(t) - F(t-) = \frac{1}{2^i}$ for $t = -2i,\ -2i + \frac{1}{2} - \frac{1}{2^i}$ with $i \geq 2$. Then, there is no minimizer of $C(\tilde{G} * F * \cdot - \gamma)$. However, the census processes associated with scheduling plans $\bar{\mathcal{A}}^n(t) = 2^n\mathbb{1}\{t \geq 2n\}$ (that is, 2^n customers are scheduled to arrive at time $2n$) $\mathcal{S}$-converge to γ, i.e., they asymptotically achieve zero fluid-cost: $C(\tilde{G} * F * \bar{\mathcal{A}}^n - \gamma) \to 0$. Indeed, $\tilde{G} * F * \bar{\mathcal{A}}^n(t) = \mathbb{1}\{t \in [0, 1/2]\} + \mathbb{1}\left\{t \in [1/2 - \frac{1}{2^n}, 1 - \frac{1}{2^n}]\right\}$. In this case, $\bar{\mathcal{A}}^n(\infty) < \infty$ for all n, but $\bar{\mathcal{A}}^n(\infty) \to \infty$.

EXAMPLE EC.2. Let $c_u(t) = c_o(t) = \mathbb{1}\{t \in [-1, 2]\}$, $\gamma(t) = \mathbb{1}\{t \in [0, 1]\}$, $G(t) = \mathbb{1}\{t \geq 1/2\}$ (the service duration is 1/2), and

$$F(t) - F(t-) = \begin{cases} \frac{1}{4}, & t = 0,\ -\frac{1}{2}, \\ \frac{1}{2^i}, & t = -2i,\ -2i + \frac{1}{2} - \frac{1}{2^i},\ i \geq 3. \end{cases}$$

Then, $\bar{\mathcal{A}}_*(t) = 4\mathbb{1}\{t \geq 1/2\}$ is the unique minimizer of $C(\tilde{G} * F * \cdot - \gamma)$: $C(\tilde{G} * F * \bar{\mathcal{A}}_* - \gamma) = 0$. However, the census processes associated with scheduling plans $\bar{\mathcal{A}}^n(t) = 2^n\mathbb{1}\{t \geq 2n\}$ (that is, 2^n customers are scheduled to arrive at time $2n$) $\mathcal{S}$-converge to γ, i.e., they asymptotically achieve zero fluid-cost: $C(\tilde{G} * F * \bar{\mathcal{A}}^n - \gamma) \to 0$. Indeed, $\tilde{G} * F * \bar{\mathcal{A}}^n(t) = \mathbb{1}\{t \in [0, 1/2]\} + \mathbb{1}\left\{t \in [1/2 - \frac{1}{2^n}, 1 - \frac{1}{2^n}]\right\}$, and $\bar{\mathcal{A}}^n \to 0$ almost everywhere. In this case, $\bar{\mathcal{A}}^n(\infty) < \infty$ for all n, but $\bar{\mathcal{A}}^n(\infty) \to \infty$. Although $\{\tilde{G} * F * \bar{\mathcal{A}}^n\}_{n=0}^{\infty}$ $\mathcal{S}$-converges to γ, the "limit" of $\{\bar{\mathcal{A}}^n\}_{n=0}^{\infty}$ does not converge to an appointment plan that minimizes the fluid cost.

EC.3. RCLL Versions of Primitives, Plans and Censuses (Remark 4)

The following lemma helps identify when the RCLL property can be assumed or deduced without loss of generality:

LEMMA EC.2. *Fix a function $f:\mathbb{R}\to\mathbb{R}$. If $f(t+):=\lim_{s\to t+}f(s)$ and $f(t-):=\lim_{s\to t-}f(s)$ exist and are finite for all $t\in\mathbb{R}$, then the function $\tilde{f}(t):=f(t+)$ is an* RCLL version *of f, namely, $\tilde{f}$ is RCLL and $\tilde{f}(t)=f(t)$ for almost all $t\in\mathbb{R}$. In the special case that f is non-decreasing (hence, left and right limits exist and are finite), $\tilde{f}$ is also non-decreasing, and $G*f(t)=G*\tilde{f}(t)$ for almost all $t\in\mathbb{R}$. (When G is a continuous CDF, then $G*f\equiv G*\tilde{f}$; that is, equality holds at all $t\in\mathbb{R}$.)*

Proof. The fact that $\tilde{f}$ is an RCLL function and $\tilde{f}(t)=f(t)$ for almost every $t\in\mathbb{R}$ can be seen from (Ethier and Kurtz 1986, Lemmas 2.2.7 and 2.2.8, page 58). The remaining part is to prove that $f*G(t)=\tilde{f}*G(t)$ for almost all $t\in\mathbb{R}$. Note that $f*G(t)-\tilde{f}*G(t)=\mathbb{E}[f(t-\sigma)-\tilde{f}(t-\sigma)]$, where σ follows G. Note that as $f(\cdot)$ can have at most countable jump points, $f(t)-\tilde{f}(t)$ is non-zero for at most countable values t; that is, we can write $f(t)-\tilde{f}(t)=\sum_{t_i\in\mathcal{T}}1_{\{t_i\}}(t)$. Here, $\mathcal{T}$ is the set of points at which $f(t)-\tilde{f}(t)\neq 0$. As a result, $f*G(t)-\tilde{f}*G(t)=\mathbb{E}[\sum_{t_i\in\mathcal{T}}1_{\{t_i\}}(t-\sigma)]=\sum_{t_i\in\mathcal{T}}\mathbb{E}[1_{\{t_i\}}(t-\sigma)]=\sum_{t_i\in\mathcal{T}}\mathbb{P}[\sigma=t-t_i]$. As $\mathbb{P}[\sigma=\delta_j]>0$ for at most countable $\delta_j\in\Delta$ (we use Δ to denote such a set of points), we have $f*G(t)-\tilde{f}*G(t)=\sum_{t_i\in\mathcal{T}}\sum_{\delta_j\in\Delta}\mathbb{P}[\sigma=\delta_j]1_{\{t=t_i+\delta_j\}}$. That is, $f*G(t)=\tilde{f}*G(t)$ for $t\neq t_i+\delta_j$, hence almost every $t\in\mathbb{R}$. If G is continuous, then $\Delta=\emptyset$; hence, $f*G(t)=\tilde{f}*G(t)$ for all $t\in\mathbb{R}$. □

REMARK EC.3 (RCLL VERSION OF γ). Assume that $\bar{\gamma}^n\xrightarrow{s}\gamma$, in which $\gamma(t+):=\lim_{s\to t+}\gamma(s)$ and $\gamma(t-):=\lim_{s\to t-}\gamma(s)$ exist for all t (e.g., when it has finite variation, locally). Then, γ can be replaced by its RCLL version in all cost calculations. The reason is that costs are dt-integrals; hence, changing the integrand in a countable number of times will not change the cost.

REMARK EC.4 (RCLL SCHEDULING PLANS AND CENSUSES). Discrete scheduling plans are RCLL by their definition. Regarding fluid or diffusion plans, any such plan $\mathcal{A}$ is non-decreasing. Denote by $\tilde{\mathcal{A}}$ its RCLL version. Then, $F*\mathcal{A}(t)=F*\tilde{\mathcal{A}}(t)$ and $G*F*\mathcal{A}(t)=G*F*\tilde{\mathcal{A}}(t)$ for almost all $t\in\mathbb{R}$, which implies $\tilde{G}*F*\mathcal{A}(t)=\tilde{G}*F*\tilde{\mathcal{A}}(t)$ for almost all $t\in\mathbb{R}$. Hence, as in the previous remark (re. dt-integrals), costs are equal whether calculated with $\mathcal{A}$ or $\tilde{\mathcal{A}}$. Finally, note that when $\mathcal{A}\in D_\wedge(\mathbb{R})$, then also $F*\mathcal{A}\in D_\wedge(\mathbb{R})$ and $G*F*\mathcal{A}\in D_\wedge(\mathbb{R})$ ($F,G\in D_\wedge(\mathbb{R})$ as CDFs) and, consequently, $\tilde{G}*F*\mathcal{A}\in D_+(\mathbb{R})$.

Remark EC.5 (RCLL censuses can be assumed to be driven by RCLL plans). Suppose that $\tilde{G} * F * \mathcal{A}$ is RCLL, and let $\tilde{\mathcal{A}}$ be the RCLL version of $\mathcal{A}$. We now show that $\tilde{G} * F * \mathcal{A} \equiv \tilde{G} * F * \tilde{\mathcal{A}}$ (equality at all $t \in \mathbb{R}$), implying that $\mathcal{A}$ can be assumed RCLL. To this end, start with $\tilde{G} * F * \tilde{\mathcal{A}}$ being RCLL; continue, by Lemma EC.2, with $\tilde{G} * F * \mathcal{A}(t) = \tilde{G} * F * \tilde{\mathcal{A}}(t)$, for almost all $t \in \mathbb{R}$; and conclude with the fact that two RCLL functions that equal almost everywhere must, in fact, equal everywhere.

EC.4. Representation of Fluid Limits (Remark 5)

By Helly's Selection Theorem, there is a subsequence $\{\bar{\mathcal{A}}^{n_k}\}_k$ that converges to some $\bar{\mathcal{A}}_\diamond$ pointwise; note that $\bar{\mathcal{A}}_\diamond \in V_\wedge(\mathbb{R})$ but not necessarily $\bar{\mathcal{A}}_\diamond \in D_\wedge(\mathbb{R})$. As in the proof of Proposition 2, this yields $\tilde{G} * F * \bar{\mathcal{A}}^{n_k} \to \tilde{G} * F * \bar{\mathcal{A}}_\diamond$ pointwise. First, we show

$$\int_{\bar{L}(t)>0} |\tilde{G} * F * \bar{\mathcal{A}}_\diamond(t) - \bar{L}(t)| c_{u+o}(t)\, \mathrm{d}t = 0, \tag{EC.1}$$

where $c_{u+o}(t) := c_u(t) + c_o(t)$. For this, we interpret $\mathbb{1}\{|\bar{L}(t)| > 0\}\, c_{u+o}(\cdot)\, |\bar{L}(\cdot)|/\mathcal{S}(\bar{L})$ as a probability density. As pointwise convergence implies convergence in probability, we have $\frac{\tilde{G}*F*\bar{\mathcal{A}}^{n_k}(\cdot)}{|\bar{L}(\cdot)|} \to \frac{\tilde{G}*F*\bar{\mathcal{A}}_\diamond(\cdot)}{|\bar{L}(\cdot)|}$ in probability. On the other hand, as in the proof of Theorem 1, one has $\frac{\tilde{G}*F*\bar{\mathcal{A}}^{n_k}(\cdot)}{|\bar{L}(\cdot)|} \to \frac{\bar{L}(\cdot)}{|\bar{L}(\cdot)|}$ in L^1, hence also in probability. Therefore, $\frac{\tilde{G}*F*\bar{\mathcal{A}}_\diamond(\cdot)}{|\bar{L}(\cdot)|} = \frac{\bar{L}(\cdot)}{|\bar{L}(\cdot)|}$ in probability, which then implies (EC.1). Second, as $|\tilde{G} * F * \bar{\mathcal{A}}^{n_k}| \le |\tilde{G} * F * \bar{\mathcal{A}}^{n_k} - \bar{L}(t)| + |\bar{L}(t)|$, by dominated convergence theorem, one has

$$\begin{aligned}&\int_{\bar{L}(t)=0} |\tilde{G} * F * \bar{\mathcal{A}}^{n_k}(t)|\, c_{u+o}(t)\, \mathrm{d}t \\ &\qquad \to \int_{\bar{L}(t)=0} |\tilde{G} * F * \bar{\mathcal{A}}_\diamond(t)|\, c_{u+o}(t)\, \mathrm{d}t.\end{aligned}$$

We also have $\int_{-\infty}^{\infty} |\tilde{G} * F * \bar{\mathcal{A}}^{n_k}(t)|\, c_{u+o}(t)\, \mathbb{1}\{\bar{L}(t) = 0\}\, \mathrm{d}t = \int_{-\infty}^{\infty} |\tilde{G} * F * \bar{\mathcal{A}}^{n_k}(t) - \bar{L}(t)|\, c_{u+o}(t)\, \mathbb{1}\{\bar{L}(t) = 0\}\, \mathrm{d}t \to 0$. Then, $\int_{-\infty}^{\infty} |\tilde{G} * F * \bar{\mathcal{A}}_\diamond(t)|\, c_{u+o}(t)\, \mathbb{1}\{\bar{L}(t) = 0\}\, \mathrm{d}t = 0$. Combining this with (EC.1) yields $\bar{L} \overset{s}{=} \tilde{G} * F * \bar{\mathcal{A}}_\diamond$. Finally, we can choose $\bar{\mathcal{A}}$ as the RCLL version of $\bar{\mathcal{A}}_\diamond$, and we still have $\bar{L} \overset{s}{=} \tilde{G} * F * \bar{\mathcal{A}}$.

Example EC.3. Assume perfect punctuality. If $G(t) = \frac{1}{2}(\mathbb{1}\{t \in [1, \infty)\} + \mathbb{1}\{t \in [2, \infty)\})$, then $\tilde{G} * \bar{\mathcal{A}}^n(t) = \bar{\mathcal{A}}^n(t) - \frac{1}{2}\left(\bar{\mathcal{A}}^n(t-1) + \bar{\mathcal{A}}^n(t-2)\right)$. For $\{\bar{\mathcal{A}}^n\}_n$ such that $\bar{\mathcal{A}}^n(t) \to t^+ =: \bar{\mathcal{A}}(t)$, for all $t \in \mathbb{R}$, we have $\tilde{G} * \bar{\mathcal{A}}^n(t) \to 1.5$ for $t \ge 2$. However, if $\tilde{\mathcal{A}}(t) = 2t - [t]$ for $t \ge 0$ (and 0 otherwise), then we also get $\tilde{G} * \tilde{\mathcal{A}}(t) = (2t - [t]) - \frac{2(t-1)-[t-1]+2(t-2)-[t-2]}{2} = 1.5$ for $t \ge 2$. Hence, if $c_o(t) + c_u(t) = 0$ for $t \le 2$, then $\tilde{G} * \bar{\mathcal{A}}^n \overset{s}{\to} \tilde{G} * \tilde{\mathcal{A}} \overset{s}{=} \tilde{G} * \bar{\mathcal{A}}$, where $\tilde{\mathcal{A}}$ is not a non-decreasing function.

EC.5. Example 1: Details

First, note that $\bar{X}(t) = \tilde{G} * \bar{\mathcal{A}}(t) = \int_{-\infty}^{t} \tilde{G}(t-s)\, \mathrm{d}\bar{\mathcal{A}}(s)$ implies that $\bar{X}(t)$, $t \le T$, does not depend on $\{\bar{\mathcal{A}}(T+s) - \bar{\mathcal{A}}(T), s \ge 0\}$. Consequently, for $t \ge T$, $\bar{X}(t) \ge \int_{-\infty}^{T} \tilde{G}(t-s)\, \mathrm{d}\bar{\mathcal{A}}(s)$, and, when minimizing $C(\bar{X}-\gamma)$, one can consider only $\bar{\mathcal{A}}$'s such that $\bar{\mathcal{A}}(t) = \bar{\mathcal{A}}(T)$ for all $t \ge T$. This results in

$$\bar{X}(t) = \int_{-\infty}^{t\wedge T} \tilde{G}(t-s)\, \mathrm{d}\bar{\mathcal{A}}(s).$$

Second, let

$$h = \left(T - \frac{1}{\mu} \ln\left(1 + \frac{c_l}{c_u}\right)\right)^+ \in [0, T].$$

Based on the definition of the cost function, the following decomposition holds:

$$\begin{aligned} C(\bar{X}-\gamma) = &\int_{-\infty}^{0} c_e \bar{X}(t)\, \mathrm{d}t \\ &+ \int_{0}^{T} (c_o(\bar{X}(t)-1)^+ + c_u(\bar{X}(t)-1)^-)\, \mathrm{d}t \\ &+ \int_{T}^{\infty} c_l \bar{X}(t)\, \mathrm{d}t \\ = &\, C_1 + C_2 + c_u\,(T-h), \end{aligned}$$

where

$$\begin{aligned} C_1 := &\int_{-\infty}^{0} c_e \bar{X}(t)\, \mathrm{d}t \\ &+ \int_{0}^{h} \left(c_o\left(\bar{X}(t)-1\right)^+ + c_u\left(\bar{X}(t)-1\right)^-\right) \mathrm{d}t \\ &+ \int_{h}^{T} (c_o + c_u)\left(\bar{X}(t)-1\right)^+ \mathrm{d}t \ge 0, \end{aligned}$$

and

$$\begin{aligned}
C_2 &:= -\int_h^T c_u \bar{X}(t)\,\mathrm{d}t + \int_T^\infty c_l \bar{X}(t)\,\mathrm{d}t \\
&= -c_u \int_h^T \int_{-\infty}^h \tilde{G}(s-u)\,\mathrm{d}\bar{\mathcal{A}}(u)\,\mathrm{d}s \\
&\quad + c_l \int_T^\infty \int_{-\infty}^h \tilde{G}(s-u)\,\mathrm{d}\bar{\mathcal{A}}(u)\,\mathrm{d}s \\
&\quad - c_u \int_h^T \int_h^s \tilde{G}(s-u)\,\mathrm{d}\bar{\mathcal{A}}(u)\,\mathrm{d}s \\
&\quad + c_l \int_T^\infty \int_h^T \tilde{G}(s-u)\,\mathrm{d}\bar{\mathcal{A}}(u)\,\mathrm{d}s \\
&= \frac{p}{\mu} \int_{-\infty}^h \left((c_u + c_l) e^{-\mu T} - c_u e^{-\mu h} \right) e^{\mu u}\,\mathrm{d}\bar{\mathcal{A}}(u) \\
&\quad + \frac{p}{\mu} \int_h^T \left((c_u + c_l) e^{-\mu(T-u)} - c_u \right) \mathrm{d}\bar{\mathcal{A}}(u);
\end{aligned}$$

the decomposition simplifies for $h = 0$ (sufficiently small T).

Finally, when $\mu T \geq \ln(1 + \frac{c_l}{c_u})$, the choice of h implies a lower bound on C_2:

$$C_2 = \frac{p}{\mu} \int_h^T \left((c_u + c_l) e^{-\mu(T-u)} - c_u \right) \mathrm{d}\bar{\mathcal{A}}(u) \geq 0.$$

Hence, in this case, $\bar{\mathcal{A}}_*(t) = \frac{1}{p}\left(\mathbb{1}\{t \geq 0\} + \mu(t^+ \wedge h)\right)$ is optimal because it achieves lower bounds for both C_1 and C_2: it yields $\bar{X}(t) = \gamma(t)$ for $t \leq h$, $\bar{X}(t) \leq \gamma(t)$ for $t \in (h, H]$, and $\mathrm{d}\bar{\mathcal{A}}_*(t) = 0$ for $t > h$; in fact, when $\mu T = \ln(1 + \frac{c_l}{c_u})$, then any fluid schedule of the form $a\mathbb{1}\{t \geq 0\}$ is optimal for $a \in [0, 1/p]$. On the other hand, when $\mu T < \ln(1 + \frac{c_l}{c_u})$, both lower bounds are achieved by $\bar{\mathcal{A}}_*(t) = 0$ for $t \in \mathbb{R}$. Combining all the cases renders

$$\bar{\mathcal{A}}_*(t) = \frac{1}{p}\left(\mathbb{1}\{t \geq 0\} + \mu(t^+ \wedge h)\right)\mathbb{1}\{h > 0\}.$$

EC.6. QED, ED and QD Regimes

Arguing informally, let $\mathcal{A}^n \approx n\bar{\mathcal{A}} + \sqrt{n}\widehat{\mathcal{A}}$ and $\gamma^n \approx n\gamma + \sqrt{n}\widehat{\gamma}$ for large n. Then, at a specific time t, the system is in the QED or ED or QD regime if $X_{n\bar{\mathcal{A}}}(t) \approx n\gamma(t)$ or $X_{n\bar{\mathcal{A}}^n}(t) > n\gamma(t)$ or $X_{n\bar{\mathcal{A}}}(t) < n\gamma(t)$ (i.e., on the fluid scale), respectively. When the QED regime prevails at all times, the cost satisfies $C(X_{\mathcal{A}^n} - \gamma^n) \approx C(X_{n\bar{\mathcal{A}} + \sqrt{n}\widehat{\mathcal{A}}} - n\gamma - \sqrt{n}\widehat{\gamma}) \approx C(\tilde{G} * F * \mathcal{A}^n + \sqrt{n}Z\Gamma_n - n\gamma - \sqrt{n}\widehat{\gamma}) \approx \sqrt{n}\,C(\tilde{G} * F * \widehat{\mathcal{A}} + Z\Gamma_n - \widehat{\gamma})$, where Γ_n^2 is the variance function of $X_{\mathcal{A}^n}$, and Z is a standard normal random

variable; that is, the total cost is order-$\sqrt{n}$. On the other hand, when the system is in the ED/QD regime at all times, the instantaneous cost function is linear on the $\sqrt{n}$ scale, and, for large n, one has $C(X_{\mathcal{A}^n} - \gamma^n) \approx nC(\tilde{G} * F * \bar{\mathcal{A}} - \gamma) + \sqrt{n}\int_{\mathbb{R}} c(t)\,(\tilde{G} * F * \widehat{\mathcal{A}}(t) - \widehat{\gamma}(t))\,\mathrm{d}t$, where $c(t) = c_o(t)$ (ED) or $c(t) = -c_u(t)$ (QD). That is, the variability of $X_{\mathcal{A}^n}$ around its mean $\tilde{G} * F * \mathcal{A}^n$ averages out. Finally, when the system alternates between the regimes, the larger order-n term is determined by the ED/QD regimes, as the contribution of the QED regime is on the smaller $\sqrt{n}$ scale.

EC.7. Alternative Construction of $\mathcal{A}^n$

Given $(\bar{\mathcal{A}}, \widehat{\mathcal{A}})$, applying (16) is not the only way to create a sequence for appointment plans $\{\mathcal{A}^n\} \subseteq D_{\mathbb{N}}(\mathbb{R})$ such that $\tilde{G} * F * \widehat{\mathcal{A}}^n \xrightarrow{s} \tilde{G} * F * \widehat{\mathcal{A}}$. Instead of $\sup_{u\le t}$ in Ψ_n, one can use $\inf_{u>t}$ as a basis to create a monotonic function. However, the correction term in this case is different. In particular, instead of $\Psi_n(\bar{\mathcal{A}}, \widehat{\mathcal{A}})$, an alternative $\tilde{\Psi}_n(\bar{\mathcal{A}}, \widehat{\mathcal{A}}) = \{\tilde{\Psi}_n(\bar{\mathcal{A}}, \widehat{\mathcal{A}})(t),\, t \in \mathbb{R}\}$ could be used:

$$\begin{aligned}\tilde{\Psi}_n(\bar{\mathcal{A}}, \widehat{\mathcal{A}})(t) := \inf_{u>t}\{\sqrt{n}\,\bar{\mathcal{A}}(u) + \widehat{\mathcal{A}}(u)\}\\ \vee \sup_{u\le t:\, u\in\tilde{\mathcal{G}}^0(\bar{\mathcal{A}})}\{\sqrt{n}\,\bar{\mathcal{A}}(u) + \widehat{\mathcal{A}}(u)\},\end{aligned}$$

where $\tilde{\mathcal{G}}^0(\bar{\mathcal{A}}) = \tilde{\mathcal{Z}}(\bar{\mathcal{A}}) \cup (-\infty, \tilde{T}_{\bar{\mathcal{A}}})$, in which $\tilde{T}_{\bar{\mathcal{A}}}$ is any constant satisfying $\tilde{T}_{\bar{\mathcal{A}}} < \inf\{u : u \in \mathcal{Z}(\bar{\mathcal{A}})\} \wedge 0$ (with the convention that $\tilde{T}_{\bar{\mathcal{A}}} = -\infty$ when $\inf\{u : u \in \mathcal{Z}(\bar{\mathcal{A}})\} = -\infty$). Then, for any $t \notin \mathcal{J}(\bar{\mathcal{A}}) \cup \mathcal{J}(\widehat{\mathcal{A}})$ (note that $\mathcal{J}(\bar{\mathcal{A}}) \cup \mathcal{J}(\widehat{\mathcal{A}})$ is at most countably infinite),

$$\begin{aligned}&\tilde{\Delta}_3^n(t)\\ &:= \inf_{u>t}\{\sqrt{n}(\bar{\mathcal{A}}(u) - \bar{\mathcal{A}}(t)) + \widehat{\mathcal{A}}(u) - \widehat{\mathcal{A}}(t)\}\\ &\quad\vee \sup_{u\le t:\, u\in\tilde{\mathcal{G}}^0(\bar{\mathcal{A}})}\{\sqrt{n}(\bar{\mathcal{A}}(u) - \bar{\mathcal{A}}(t)) + \widehat{\mathcal{A}}(u) - \widehat{\mathcal{A}}(t)\}\\ &\to \sup_{\delta>0}\inf_{u\ge t:\, \bar{\mathcal{A}}(u)\le\bar{\mathcal{A}}(t)+\delta}\{\widehat{\mathcal{A}}(u) - \widehat{\mathcal{A}}(t)\}\\ &\quad\vee \sup_{u\le t:\, \bar{\mathcal{A}}(u)=\bar{\mathcal{A}}(t),\, u\in\tilde{\mathcal{G}}^0(\bar{\mathcal{A}})}\{\widehat{\mathcal{A}}(u) - \widehat{\mathcal{A}}(t)\}\\ &= \sup_{\delta>0}\inf_{u\ge t:\, \bar{\mathcal{A}}(u)\le\bar{\mathcal{A}}(t)+\delta}\{\widehat{\mathcal{A}}(u) - \widehat{\mathcal{A}}(t)\}\\ &\quad\vee \left(-\infty \times \mathbb{1}\left\{t \notin \bar{\mathcal{G}}^0(\bar{\mathcal{A}})\right\}\right) = 0,\end{aligned}$$

where $\bar{\mathcal{G}}^0(\bar{\mathcal{A}})$ is the closure of $\tilde{\mathcal{G}}^0(\bar{\mathcal{A}})$.

EC.8. Operational Characteristics of QED Appointments

Our title "QED Appointments" was chosen to reflect its analogy with "QED Staffing" (originated in Mandelbaum (2003)) which, in turn, had arisen from many-server asymptotics of queues (as reviewed in van Leeuwaarden et al. (2019)). To be more specific, staffing is a response to offered-load from customers, whereas appointments is a response to offered-capacity of servers; and, in both cases, with large enough load and capacity, staffing and appointments could lead to QED performance (high-levels of both service-quality and capacity's efficiency, operationally) by exploiting carefully economies of scale.

In this appendix, we add support to the staffing-appointments QED analogy by first reviewing some essentials of QED staffing, which then sets the stage for a comparison with QED appointments. We restrict attention to special cases, analyzed at the coarsest level that permits a formal convincing comparison.

QED staffing of many-server queues. Consider a sequence, indexed by n, of $M/M/s_n$ queues: in system n, there are s_n servers, each with an individual service rate of μ; and the rate of exogenous arrivals is $\lambda_n \approx n\lambda + \sqrt{n}\hat{\lambda}$, for some constants $\lambda > 0$ and $\hat{\lambda} \in \mathbb{R}$. The following first bullet point prescribes QED staffing, which then leads to QED performances (second and third bullets), as $n \uparrow \infty$:

- Number of servers is $s_n \approx R_n + \beta\sqrt{R_n}$, where $R_n := \lambda_n/\mu$ is the offered-load. This is commonly referred to as *square-root safety staffing*, where the "safety" is relative to the offered-load.
- Probability of delay (all servers busy upon arrival) is strictly between 0 and 1 (asymptotically).
- Queue-length and number of idle-servers are both of order $\sqrt{n}$.

QED appointments to ample-servers. Consider a sequence of appointment systems, indexed by n, in which system n has the following parameters: F and G are cdf's of punctuality and service durations, respectively; and the goal function is $\gamma^n \approx n\gamma + \sqrt{n}\widehat{\gamma}$, in which $\gamma = \tilde{G} * F * \bar{\mathcal{A}}$ is a QED goal (9) associated with the offered-capacity $\bar{\mathcal{A}} \in D_\wedge(\mathbb{R})$, and $\widehat{\gamma} \in D(\mathbb{R})$. The following three bullet points prevail in the setup of our Theorem 2 and Lemma 2. The first one prescribes QED appointments, which then leads to QED performance in the second and third bullets, as $n \uparrow \infty$:

- Appointment plan $\mathcal{A}^n(\cdot)$ satisfies $\mathcal{A}^n(t) = \lfloor \sqrt{n}\Psi_n^+(\bar{\mathcal{A}}, \widehat{\mathcal{A}})(t) \rfloor \approx n\bar{\mathcal{A}}(t) + \sqrt{n}\widehat{\mathcal{A}}(t)$ for almost all $t \in \mathbb{R}$, where $n\bar{\mathcal{A}}$ is offered-capacity (fluid scaled). We have referred to such $\mathcal{A}^n$ as *square-root safety appointments*, where the "safety" is relative to the offered-capacity.
- Probability of encountering overage is strictly between 0 and 1 (asymptotically) for almost all t with $\Gamma(t) > 0$; the same holds for the probability of encountering underage.

- Magnitude of both overage and underage is order $\sqrt{n}$ for almost all $t \in \mathbb{R}$. Consequently, both overage and underage cost-rates are order $\sqrt{n}$ as well.

Justifications. The bullets on QED staffing, and more, were proved in the seminal paper by Halfin and Whitt (1981). We now justify briefly the bullets on QED appointments.

First, (38) implies $\widehat{\mathcal{A}}^n(t) \to \widehat{\mathcal{A}}(t)$ for almost all $t \in \mathbb{R}$, as well as the first bullet point. Next, we justify the second and the third bullet points. Working in the setting of Corollary EC.1, we have $\Gamma_n(t) \to \Gamma(t)$ for $t \notin (\mathcal{J}(F) \bigcup \mathcal{J}(G * F)) \oplus \mathcal{J}(\bar{\mathcal{A}})$. For the following arguments, we fix any $t \notin (\mathcal{J}(F) \bigcup \mathcal{J}(G * F)) \oplus \mathcal{J}(\bar{\mathcal{A}})$. Then, similar to the proof of (31), we have $\widehat{Z}_{\mathcal{A}^n}(t) \Rightarrow \Gamma(t)Z$, and the uniform integrability of the related terms. Since $\sup_u |\widehat{\mathcal{A}}(u)| < \infty$ and $\widehat{\mathcal{A}}^n(\cdot) \to \widehat{\mathcal{A}}(\cdot)$ almost everywhere, we also have $\tilde{G} * F * \widehat{\mathcal{A}}^n(t) \to \widehat{L}(t) = \tilde{G} * F * \widehat{\mathcal{A}}(t)$. As a result, such appointments ensure that the arrival rate and census level are both of order n, where the latter fluctuates around $n\tilde{G} * F * \bar{\mathcal{A}}$. Note that $\widehat{\gamma}^n(t) \to \widehat{\gamma}(t)$ in view of $\gamma^n(t) \approx n\gamma(t) + \sqrt{n}\widehat{\gamma}(t)$. Together with $\tilde{G} * F * \bar{\mathcal{A}} = \gamma$, and similarly to the proof of (31), we argue that $\frac{1}{\sqrt{n}}(X_{\mathcal{A}^n}(t) - \gamma^n(t)) \Rightarrow (\Gamma(t)Z + \widehat{L}(t) - \widehat{\gamma}(t))$, and $\frac{1}{\sqrt{n}}\mathbb{E}(X_{\mathcal{A}^n}(t) - \gamma^n(t))^{\pm} \to \mathbb{E}(\Gamma(t)Z + \widehat{L}(t) - \widehat{\gamma}(t))^{\pm}$. Hence, the probability of encountering an overage is approximately $\bar{\Phi}\left(\frac{\widehat{\gamma}(t) - \widehat{L}(t)}{\Gamma(t)}\right)$, the magnitude of overage is $\sqrt{n}\mathbb{E}(\Gamma(t)Z + \widehat{L}(t) - \widehat{\gamma}(t))^+$, and the overage cost rate at time t is $\sqrt{n}c_o(t)\mathbb{E}(\Gamma(t)Z + \widehat{L}(t) - \widehat{\gamma}(t))^+$; the probability of encountering an underage is approximately $\Phi\left(\frac{\widehat{\gamma}(t) - \widehat{L}(t)}{\Gamma(t)}\right)$, the magnitude of underage is $\sqrt{n}\mathbb{E}(\Gamma(t)Z + \widehat{L}(t) - \widehat{\gamma}(t))^-$, and the underage cost rate is $\sqrt{n}c_u(t)\mathbb{E}(\Gamma(t)Z + \widehat{L}(t) - \widehat{\gamma}(t))^-$. We conclude, with noting that $(\mathcal{J}(F) \bigcup \mathcal{J}(G * F)) \oplus \mathcal{J}(\bar{\mathcal{A}})$ is at most countable, that the above convergences hold for almost all $t \in \mathbb{R}$.

Remark EC.6 (Stabilizing performance). In $M_t/M/s_t$ queues, where arrival rates are time-dependent ($\lambda_n(t)$), one can specify time-varying square-root staffing s_t that stabilizes performance (e.g., delay probability that is constant at all times (Feldman et al. 2008)). This bears direct analogy to the performance of our QED appointments, specifically Example 4, where the probability of encountering overage is constant over $(0, T)$; and similarly for underage.