Optimal Dynamic Appointment Scheduling
of Base and Surge Capacity

Benjamin Grant*, Itai Gurvich†, R. Kannan Mutharasan MD‡, Jan A. Van Mieghem*  
b-grant@northwestern.edu, gurvich@cornell.edu, kannanm@northwestern.edu, vanmieghem@northwestern.edu

December 19, 2017; revised August 10, 2018

(1) Problem definition: We study dynamic stochastic appointment scheduling when delaying appointments increases the risk of incurring costly failures, such as readmissions in health care or engine failures in preventative maintenance. When near-term base appointment capacity is full, the scheduler faces a trade-off between delaying an appointment at the risk of costly failures versus the additional cost of scheduling the appointment sooner using surge capacity.

(2) Academic/Practical Relevance: Most appointment scheduling literature in operations focuses on the trade-off between waiting times versus utilization. In contrast, we analyze preventative appointment scheduling and its impact on the broader service supply network when the firm is responsible for service and failure costs.

(3) Methodology: We adopt a stochastic dynamic programming (DP) formulation to characterize the optimal scheduling policy and evaluate heuristics.

(4) Results: We present sufficient conditions for the optimality of simple policies. When analytical solutions are intractable, we solve the DP numerically and present optimality gaps for several practical policies in a health care setting.

(5) Managerial Implications: Intuitive appointment policies used in practice are robust under moderate capacity utilization, but their optimality gap can quadruple under high load.

Key words: Transitional care, appointment scheduling, health care, preventive maintenance

1. Introduction

We study dynamic stochastic appointment scheduling when delaying appointments increases the risk of costly failures. Motivated by health care compensation models (CMS 2016) and service contracting (Kim et al. 2007), the appointment provider is responsible for the total cost of service and failure. In particular we investigate early follow-up appointments after major events aimed at identifying and treating preventable causes of future failures. It is common for a service provider to plan base appointment capacity, which has low (or no) incremental cost of usage and have a high cost surge capacity alternative to handle excess demand—such as overtime in health care

* Department of Operations, Kellogg School of Management at Northwestern University, Evanston, IL 60208
† Operations Research and Information Engineering Department, Cornell Tech, New York, NY 10044
‡ Division of Cardiology, Northwestern University Feinberg School of Medicine, Chicago, IL 60611
or external contractors in preventative engine maintenance. When near-term base appointment capacity is full, the scheduler faces a trade-off between delaying a customer’s appointment at the risk of costly failures versus the additional cost of scheduling the appointment sooner using surge capacity. In this paper we develop a framework to analyze dynamic scheduling for preventative appointments and present optimal policies for using base and surge capacity.

Our work was originally inspired by preventative appointments in a health care setting to which we later adapt our model for numerical analysis. Medicare’s implementation of Bundled Payments for Care Improvement (BPCI; see CMS (2016)) is a compensation model where hospitals that provide care for acute health episodes (e.g. Stroke or Heart Failure) are responsible for all costs related to the initial acute illness during hospitalization and for a fixed transition window (e.g. 30, 60, or 90 days) post-discharge. Poor health outcomes—henceforth major health failures, health failures, or simply failures—such as hospital readmissions or mortality are financially costly, and related to poor quality health care. Therefore, the BPCI model aims to increase the value of care via financial incentives for hospitals to reduce health failures (and hence increase quality) by reducing the total cost of care for an acute illness.

Hospitals have taken a variety of steps to adjust to BPCI including preventative transitional care appointments (e.g. follow-up appointments, rehab, etc.) to reduce the risk of major failures post discharge. Early follow-up appointments after a patient is discharged from an acute health episode of care have been shown to reduce readmission rates for patients in the 30-days post-discharge (Hernandez et al. 2010). Further, the risk of readmission increases as the follow-up is delayed (Ryan et al. 2013). We have implemented preventative appointments at Northwestern Memorial Hospital (NMH) where we require that all Congestive Heart Failure (CHF) patients be scheduled for a follow-up appointment within one week of their hospital discharge (NMHC 2016). When planned capacity is full, the central scheduling team pages a physician and adds an overtime appointment to a particular time period. We adapt our framework to the NMH follow-up clinic for CHF patients in our numerical analysis to illustrate how the optimal scheduling policies compare to practical heuristics in terms of total cost and their frequency of surge capacity use.

Our research is more broadly relevant in the context of service and warranty contracts. Railroad companies for example often have service contracts with the locomotive manufacturers where the manufacturer is responsible for servicing and replacing parts throughout the service contract duration. When the manufacturer replaces a locomotive engine, which may be in service for more than 40 years, an early check-up can ensure that the newly installed engine is functioning properly. The follow-up appointment aims to diagnose issues that may cause costly failures for which the service contractor would be responsible. A similar, yet cheaper application involves a local bicycle shop in Evanston, IL, that sells new bikes to customers under warranty. The shop offers a free tune-up
after 10 and 25 hours of riding the cycle, which is an early follow-up appointment to make sure everything is running smoothly. This tune-up serves to increase customer service with a smoother ride but also to prevent potential issues (e.g., a chain breaking) in the year(s) after purchase while the cycle is under warranty. The tune-up has a small overhead cost and reduces the risk of future costly failures and future lost sales due to a poor quality experience.

We formulate the dynamic appointment scheduling problem as a stochastic dynamic program (DP). We capture the decision of when to use base versus surge capacity and the impact that preventative appointments have on total cost. We explicitly model “no-shows” where a customer may not show up for their scheduled appointment. We present sufficient conditions—that can be evaluated in polynomial time—for the optimality of simple policies.

When also run extensive numerical experiments tuned to transitional care appointments for CHF patients. We use our experience at NMH along with medical literature to fit baseline model parameters. We then run the model on a broad test bed of parameters to evaluate how various parameters impact optimal decisions and total cost. We also solve for the steady state probabilities to measure the use of surge capacity under different policies and different demand volumes.

We observe numerically that intuitive appointment policies used in practice are robust under moderate capacity utilization, but their optimality gap can quadruple under high load. We further provide a simple rule of thumb for policy selection based on our optimality conditions, which performs within 12% of optimal for all scenarios in our numerical test bed. Our results give practical insights on how to optimally schedule preventative appointments and use surge capacity when a service provider is responsible for the total costs of appointments and failures.

2. Literature Review

Scheduling is a fundamental and well-studied problem in operations; see Ahmadi-Javid et al. (2017). Within the operations literature there is a growing body of work on scheduling problems in health care. Cayirli and Veral (2003) provide a comprehensive review of the existing literature on outpatient appointment scheduling, at the time of their publication. They overview modeling techniques, problem formulations, and identify gaps in the literature. Gupta and Denton (2008) survey the different outpatient health care and elective-surgery settings where appointment scheduling is used along with the modeling methodologies employed. Ahmadi-Javid et al. (2017) review numerical optimization techniques used in appointment scheduling.

This branch of operations literature focuses on matching demand (appointment requests) and supply (capacity of health care providers) at the time appointments are serviced. The typical trade-off involves patient waiting time vs. physician utilization. In contrast to traditional outpatient appointment scheduling, transitional care appointments focus on reducing the risk of major health
failures. Under Bundled Payments for Care Improvement (BPCI) contracts, the cost and quality of major health failures are first order effects; more so than patient waiting time or resource utilization. Controlling the schedule of transitional care appointments to impact the risk of failure is a crucial decision.

Liu et al. (2010) is the closest work to ours and provides inspiration for some of our modeling choices. Like them, we explicitly model patient no-shows. They model patient appointment cancellations, which can be related to health failures in our model. In their model a cancellation cannot occur after a scheduled appointment. Cancellations create a hole in the schedule and they cost only in so far as the slot may not be filled to match planned capacity. In our model, failures can happen after the scheduled appointment, and regardless of whether the failure happens before or after a scheduled appointment (but within the transition window), the provider will be financially responsible for the cost of the failure.

In our model, we make explicit the decision and cost of adding capacity beyond the planned capacity. The scheduler can choose to surge: to add, at an additional cost, appointments to time periods where planned base capacity is full. In some clinical settings, surge capacity can be thought of as overtime. Scheduling policies differ in the amount of surge capacity they use. When evaluating multiple nearly optimal policies, we will compare them in terms of surge capacity usage—a consideration that is important for practitioners.

Liu et al. (2010)'s model is in some respects more general than ours. It allows for the appointment cancellation and no-show probabilities to depend on the time elapsed since the appointment was booked. This generality makes it difficult to derive analytical results due to the curse of dimensionality. We consider initially a case where failure and no-show rates are constant. This allows us to identify a set of practically appealing policies, prove their optimality in certain parameter regimes, and establish their numerical robustness. At the end of this paper, we numerically relate the general case of a dynamic failure rate to the constant rate case and provide optimality bounds for certain policies.

In addition to the operationally focused literature, there is a branch of medical literature that focuses on the impact of bundled payments on the health care system from both cost and provider perspectives’. Medicare’s BPCI, CMS (2016), program was introduced in 2013 and at this point (2018), publications about its impact are mostly limited to expert opinions, and some initial aggregate statistics. Proper causal studies are still absent. Delisle (2013) outlines the previous Medicare payment structure of fee-for-service and how bundled payments differ. Matchar et al. (2015) discuss bundled payment programs implemented in the United States in the past and consider their implications on acute-stroke care. Mechanic (2014) makes a strong case for increased focus on post-acute care by showing that the cost of this care is substantial in both absolute and relative terms to the
initial admission cost. Froimson et al. (2013) provide a brief history of bundled payments including an overview and commentary on Medicare’s BPCI program. Iorio et al. (2016) use statistical models to evaluate the change in cost related to arthroplasty as a result of BPCI. Morley et al. (2014) evaluate the impact on care facilities in the health care supply chain related to bundled payments and provide descriptive statistics, while Jackson et al. (2016) examine the surgeon’s role in BPCI. Zhang et al. (2016) study how Medicare’s Hospital Readmission Reduction Program effects hospital competition using a game-theoretic model.

Our work builds on current appointment scheduling literature to account for incentives under total-cost payment models, such as BPCI, and control decisions for multiple capacity types. We ground the operational decision model in medical reality by drawing on insights from the medical literature and our experience at NMH. We provide a modeling framework and analytically characterize optimal policies for particular parameter regions. We examine exact numerics for the transitional care setting, and provide actionable guidelines for optimal and near-optimal scheduling policies. We further confirm the numerical robustness of practical scheduling policies being used in practice.

3. Model
We analyze how to dynamically schedule (“book”) a random arrival stream of preventative appointment requests to either future base or surge capacity. Preventative appointments have an indirect delay cost in that earlier appointments reduce the risk of costly failures. We formulate the scheduling problem as an infinite horizon discounted dynamic program (DP). Our objective is to minimize the discounted total cost of capacity and failures.

Timeline and event sequence: At the beginning of period \( t \), the schedule of booked appointments is observed and new appointment requests are received. Scheduling decisions are made based on the current schedule and number of new appointment requests. Each appointment on the schedule represents a customer waiting for her appointment. Customers scheduled for appointments in period \( t \) may either show up for their appointment or “no-show”. However, capacity for period \( t \) must be planned (and fixed) at the beginning of period \( t \) before no-shows are realized. In period \( t \) any of the customers in the appointment book for subsequent periods can experience a failure. A scheduled customer that experiences a failure requires more substantial care or maintenance at substantially higher cost and is removed from the schedule for a preventative appointment. Customers who do not fail before their scheduled appointment may still fail within the contract window: the time the scheduler is responsible for the cost of failures. When customers fail, the scheduler incurs the cost of the failure.

State Space: Let \( S'_t = \{S'_d^t : d = 0, 1, \ldots, D\} \) denote the state at time \( t \), where \( S'_d^t \in \mathbb{Z}_+ \) is the number of appointments booked as of time \( t \) for \( d \) periods in the future, i.e., for period \( t + d \),
and \(d\) is within the scheduling horizon \(D\): appointments cannot be scheduled beyond \(D\) periods in the future. In general \(D\) can be any number of time periods less than or equal to duration of the contract window since a preventative appointment would not be scheduled beyond the contract term. The number of appointments scheduled in any given period is naturally bounded—assuming overbooking in not allowed—by the number of hours in a time period divided by the average appointment time. We refer to the bound on per-period appointments as capacity \(K\). In many cases \(K\) may be bounded further by business decisions such as operating hours. Given \(K\), the state space \(S\) is bounded and contains all possible appointment books \(s \in \{0, 1, \ldots, K\}^D\).

**Appointment Requests:** Let \(A^t \in \mathbb{Z}_+\) be the random number of new appointment requests arriving at time \(t\). We assume that \(A^t, t = 1, 2, \ldots\) are i.i.d. random variables. The uncertainty in \(A^t\) is realized before the scheduling decision is made.

**Actions:** After the state \(S^t\) is observed at time \(t\) and the new appointment requests \(A^t\) are realized, scheduling decisions are made. Define \(U^t_d \in \mathbb{Z}_+\) to be the number of new appointments added to the appointment book at time \(t\) for period \(t + d\). Same-period appointments are not allowed and if, as in our health care application, all requests must be scheduled then we must have \(\sum_{d=1}^{D} U^t_d = A^t\).

Given \(A^t\), a feasible and admissible decision is then a function from the state at the beginning of period \(t\) and the new appointment requests, \(U^t(S^t, A^t)\), that belongs to the set

\[
\mathcal{U} := \left\{ u : \mathbb{Z}_+^{D+1} \times \mathbb{Z}_+ \rightarrow \mathbb{Z}_+^{D+1} : u_0 \equiv 0, \sum_{d=1}^{D} u_d = A^t, u_d + S^t_d \leq K \forall d \in \{1, \ldots, D\} \right\}. \tag{1}
\]

If there exists \(M\) such that \(P(A^t \leq M) = 1\), then the size of the action space is at most \(M^D\).

Since the number of appointments per-period is bounded by \(K\) it is possible that \(\sum_{d=1}^{D} u_d = A^t\) is infeasible when \(u_d + S^t_d \leq K\) holds for all \(1 \leq d \leq D\). In other words there are more appointment requests than available appointment slots on the schedule. In that case \(U = \emptyset\), where \(\emptyset\) denotes the empty set. When this is the case, the decision is to schedule all available appointments such that \(u_d + S^t_d = K\) for all \(1 \leq d \leq D\), and the appointment requests not scheduled, overflow appointments, are sent to another provider and leave the system. We record the number of overflow appointments, \(n_O = \max\{A^t - \sum_{d=1}^{D} u_d, 0\}\), and account for them directly in the cost function below.

**State transitions:** When the decision \(U^t(S^t, A^t)\) is made in period \(t\), the state is changed from \(S^t\) to \(S^t + U^t(S^t, A^t)\) immediately before transitioning to period \(t + 1\). Thus,

\[
\tilde{S}^t := S^t + U^t, \tag{2}
\]

is the state after the scheduling in period \(t\) and before the transition to period \(t + 1\).

Before her appointment, a customer’s time until failure is geometrically distributed with parameter \(p_F\). Alternatively, each of the customers on the appointment book can fail with probability \(p_F\).
each period independently from the past and from other customers. Given the (post-action) state in period \( t \), \( S^t \), the (pre-action) state in period \( t + 1 \) satisfies

\[
S_{d+1}^{t+1} \sim \begin{cases} 
B(S_d^t, p_F), & \text{for } d = 0, \ldots, D-1 \\
0, & \text{for } d = D
\end{cases}
\]

where \( p_F = 1 - p_F \) and \( B(n, p) \) is a binomial random variable with \( n \) trials and probability of success \( p \). The single period transition probability is thus given by

\[
P(S'|S) = \begin{cases} 
\prod_{d=0}^{D-1} S_{d+1}^t (S_d^t) p_F^{S_d^t} (1 - p_F)^{S_{d+1}^t - S_d^t}, & \text{if } S_D^t = 0 \\
0, & \text{otherwise}
\end{cases}
\]

We let \( P_{S,S'} := P(S'|S) \) and write \( P \) for the corresponding matrix with rows and columns indexed by the entire state space \( S \).

No-shows do not appear explicitly in the transition probabilities because at the beginning of time period \( t \) the state is observed, the scheduling decision is made, and capacity for period \( t \) appointments is planned for. All costs for customers with appointments scheduled for time period \( t \) in period \( t \) are explicitly accounted for below, including capacity, no-shows, and subsequent failures after appointment times.

**No-Shows:** Customers that do not experience a failure before their allocated time, might still not show for their appointment. The probability that a customer shows up for her appointment is \( p_S \). The no-show probability is \( p_S = 1 - p_S \). Showing up for the appointment does not rule out a later failure but its likelihood might change. We denote by \( p_{F|S} \) the probability that a customer, who shows-up for her appointment, fails after her appointment but still within the contract window. If a customer no-shows for an appointment, the customer experiences a failure after no-showing after her appointment, but within the contract window, with probability \( p_{F|S} \). The preventative appointments are driven by the assumption that \( p_{F|S} < p_{F|\bar{S}} \), which, in health care, is supported by clinical evidence in Hernandez et al. (2010). We assume for now that \( p_{F|S} \) and \( p_{F|\bar{S}} \) are constants that do not depend on when an appointment is scheduled. This approximation is reasonable when the scheduling window is small relative to the contract window. We relax this assumption in Section 6.2 where we extend our analysis to dynamic \( p_{F|S} \) and \( p_{F|\bar{S}} \).

We define the random variable \( N_{F|S}^t \) to be the number of customers scheduled for period \( t \) who show up for their appointment, and subsequently experience a failure within the contract window. Similarly, we let \( N_{F|\bar{S}}^t \) be the number of customers scheduled for period \( t \) who no-show for their appointment, and subsequently experience a major failure within the contract window.

**Immediate Cost Function:** We denote by \( c_F \) the cost of a single customer’s failure. This could be for example the cost of a readmission in the health care setting, or the cost to replace an engine in a service contract. This cost does not depend on the timing of the failure.
We denote by $h(\cdot)$ the appointment cost function. In a period with $x$ planned appointments, the provider incurs a cost $h(x)$; e.g. the cost of staffing and running a follow-up appointment clinic. No-shows do not reduce this cost. The capacity is planned and held, and hence, must be paid for. We assume that $h(\cdot)$ is a non-decreasing function. Thus, the provider has a cost $h(S_0^t)$ on day $t$ where, we recall, $S_0^t \leq K$ is the number of appointments scheduled for day $t$ (viewed at time $t$).

Consider a service provider that plans for a particular number of appointment slots per period. We denote this base capacity by $K_b$. The scheduler can add a surge appointment beyond $K_b$ at an incremental cost of surge appointments $c_s$; e.g. overtime costs to service appointments beyond base appointments. We denote the number of surge capacity appointments available each period by $K_s$ and note that $K_b + K_s = K$. Of central interest for us is the following instance of the appointment cost function,

$$h(x) = c_s \max\{S_0^t - K_b, 0\}.$$  \hfill (5)

This represents $K_b$ slots of base capacity that is planned for each period and scheduling base capacity incurs no incremental cost. Consider for example the number of appointment slots a cardiologist allocates daily to office visits. Costs starts accruing only when surge appointments are scheduled—that is, when the scheduled appointments exceed the number of base slots.

The expected cost in period $t'$ with discount factor $\beta \in (0, 1)$ is then:

$$\mathbb{E}[g_t(S^t, U')] = h(S_0^t) + \beta c_F \mathbb{E} \left[\sum_{d=0}^{D-1} \left(S_{d+1}^t - S_d^t\right)\right] + c_F \mathbb{E} \left[N_{F|S}^t + N_{F|\bar{S}}^t\right] + c_F n_O$$  \hfill (6)

The first term accounts for the cost of servicing the current period’s appointments. The second term accounts for the cost of major failures occurring between scheduling decisions in period $t$ and $t+1$. Failures that happen after the scheduling decision in $t$ and before $t+1$ are only observed at the beginning of period $t+1$, hence the discount factor $\beta$. The third term accounts for failures of patients scheduled for period $t$ that do not fail before period $t$, but subsequently fail after $t$ and within the contract window. Implicit here is the assumption that $N_{F|S}^t$ and $N_{F|\bar{S}}^t$ are accounted for in period $t$ even though the failure occurs after the appointment at some future time period. The number of no-shows is $n_S \sim Bin(S_0^t, p_S)$-distributed and conditional on $n_S$, $N_{F|S} n_S \sim Bin(n_S, p_F|S)$. Thus, $\mathbb{E}[N_{F|S}] = S_0^t p_S p_{F|S}$. Similarly, $\mathbb{E}[N_{F|\bar{S}}] = S_0^t p_S p_{F|\bar{S}}$. The fourth term is a penalty cost that we set equal to the cost of a failure for each overflow appointment that is forced out of the system.

To simplify notation we introduce the matrix $F$ to count the number of failures that occur in a state transition. Row and columns of $F$ are indexed by the entire state set, and the $(S, S')$ entry is defined as

$$F_{S,S'} = \begin{cases} \sum_{d=0}^{D-1} \left(S_{d+1} - S_d\right), & \text{if } P(S'|S) > 0, \\ 0, & \text{otherwise} \end{cases}$$
We use $P_{(S, \cdot)}$ and $F_{(S, \cdot)}$ to denote, respectively, the row vectors of $P$ and $F$ at the $S$ index; $F_{(S, \cdot)}^T$ is the transpose of $F_{(S, \cdot)}$. The expected cost in period $t$ is then given by

$$
\mathbb{E}[g_t(S^t, U^t)] = h(S^t_0) + \beta c_F \mathbb{E} \left[ \sum_{d=0}^{D-1} \left( \tilde{S}^t_{d+1} - S^t_{d+1} \right) \right] + c_F \mathbb{E} \left[ (N^t_{F|S} + N^t_{F|\bar{S}}) \right] + c_F n_O
$$

$$
= h(S^t_0) + \beta c_F P_{(S, \cdot)} F_{(S, \cdot)}^T + c_F \left( \mathbb{E}[N^t_{F|S}] + \mathbb{E}[N^t_{F|\bar{S}}] \right) + c_F n_O
$$

$$
= h(S^t_0) + \beta c_F P_{(S, \cdot)} F_{(S, \cdot)}^T + S^t_0 c_F \left( p_{SF|S} + p_{SF|\bar{S}} \right) + c_F n_O
$$

(7)

**Bounded Immediate Cost Function:** The number of per-period appointment slots cannot exceed $K$ and we assume the number of new appointment requests is bounded per-period such that $P(A^t \leq M) = 1$ where $M$ is some positive integer. Thus, for a finite scheduling horizon, $D$, the state space is bounded and we can define $\bar{h} = h(K) = \max_{k=1,...,K} h(k) < \infty$. In turn,

$$
\mathbb{E}[g_t(S^t, U^t)] = h(S^t_0) + \beta c_F \mathbb{E} \left[ \sum_{d=0}^{D-1} \left( \tilde{S}^t_{d+1} - S^t_{d+1} \right) \right] + c_F \mathbb{E} \left[ (N^t_{F|S} + N^t_{F|\bar{S}}) \right] + c_F n_O
$$

$$
\leq h(K) + \beta c_F \sum_{d=0}^{D-1} \left( K - 0 \right) + c_F K + c_F M
$$

$$
= \bar{h} + \beta c_F DK + c_F K + c_F M
$$

$$
= \bar{h} + c_F (\beta DK + K + M) < \infty
$$

(8)

so that the expected immediate cost function is bounded.

**Scheduling Policies:** Let $\pi$ be an admissible policy, meaning $\pi$ is a sequence of functions, $\pi = \{U^0, \ldots, U^{T-1}\}$, where $U^t$ maps the state, $S^t$ and the number of appointment requests $A^t$ on period $t$ to action $U^t(S^t, A^t)$. Here admissible means that the control decision $U^t$ must satisfy the constraints in equation (1).

**Objective:** Given the immediate cost function $g_t$ and discount factor $\beta \in (0, 1)$, we wish to find the policy $\pi$ that minimizes, for each initial state $S^0$ and appointment requests $A^0$,

$$
V_\pi(S^0) = \lim_{T \to \infty} \mathbb{E} \left[ \sum_{t=0}^{T-1} \beta^t g_t(S^t, U^t(S^t, A^t)) \right]
$$

(9)

The optimal value function is,

$$
V^*(S^0) := \min_{\pi \in \Pi} J_\pi(S^0),
$$

(10)

where $\Pi$ is the set of admissible policies; those that are non-anticipative and satisfying equation (1). A bounded state space and immediate cost function guarantees that the optimal value function solves the Bellman equation

$$
V^*(S^0) = \min_{u \in U} \left\{ \mathbb{E}[g_0(S^0, U^0)] + \beta \mathbb{E}_u[V(S^1)] \right\},
$$

(11)
and that a stationary policy that minimizes the right-hand side exists and is optimal. Above $\mathbb{E}_s[u(\cdot)]$ is the expectation conditional on $S^0 = s$ and the transition probability induced by $u$. Equation (11) can be solved via value or policy iteration. Later we will use the following linear programming formulation of the DP to quantify the optimality gap of various simple policies.

**LP Formulation:** $V^*(S)$ is the optimal value function of the infinite horizon discounted DP with initial state $S$. The value function solves the linear program

$$\max \sum_S V(S)$$

subject to

$$V(S) \leq \sum_a \mathbb{P}(A = a) g(S, U(a)) + \beta \sum_{S'} \mathbb{P}_{(S + U(S, a), S')} V(S')$$

for each $S, U$.

Including the i.i.d. per-period arrivals as part of the state descriptor gives rise to the same optimal value but allows for easier interpretation. We break $V(S)$ into variables $V(S, A)$.

$$\max \sum_S \sum_a \mathbb{P}(A = a) V(S, a)$$

subject to

$$V(S, a) \leq g(S, U(S, a)) + \beta \sum_{S'} \mathbb{P}_{(S, S')} \sum_{a'} \mathbb{P}(A = a') V(S', a')$$

for each $a, S, U(S, a)$.

The binding constraints in the solution to LP reveal an optimal stationary policy for a given instance of the model parameters.

**Complexity of LP:** Given $K$ appointment slots per period and booking horizon $D$, there are $(K + 1)^D$ possible states at the beginning of any time period. If per-period requests have a finite support $M$ (i.e., $\mathbb{P}\{A^t \leq M\} = 1$), there are $(M + 1) \cdot (K + 1)^D$ decision variables to solve for in LP. Although this problem suffers from the curse of dimensionality, this complexity is naturally constrained in the health care setting we consider below, facilitating our numerical experiments.

## 4. Optimality Results

To derive optimal policies we impose the capacity cost structure of equation (5). In each time period base appointment capacity, $K_b$, is planned and incurs no incremental cost regardless of usage. If the scheduler wishes to schedule more than $K_b$ appointments for a given period, she can do so for a cost $c_s$ per surge appointment over base capacity. There are at most $K_s$ surge appointments available per period for the scheduler to use at her discretion. Thus the total appointment capacity per period is $K = K_b + K_s$.

Consider a customer whose appointment request at time period $t = 0$ is scheduled for $d$ time periods into the future. The probability the customer fails before her scheduled appointment in some period $l$, $0 \leq l < d$, is $p_F p_F^l$. And thus, the probability the customer does not fail before $d$ is
bounded above by

\[ p^d_F. \]

Recall, the customer may experience a failure after they show or no-show for their appointment in period \( d \) with probability \( \left( p_S p_F|S + p_S p_F|S \right) \). The total probability of failure for this customer is then \( \sum_{l=0}^{d-1} p_F p_F^l + p_F^d \left( p_S p_F|S + p_S p_F|S \right) \), and their expected discounted cost of failure is

\[
c_F \left[ \sum_{l=0}^{d-1} \beta^{l+1} p_F p_F^l + \beta^d p_F^d \left( p_S p_F|S + p_S p_F|S \right) \right]. \tag{12}
\]

It is useful to think of the expression inside the brackets of (12) as the discounted probability of a failure during the contract window, at the time scheduling decision is made. This expression is monotone increasing in \( d \) consistent with the expectation that the probability of a failure is increasing the further into the future a preventative appointment is scheduled.

The following proposition justifies referring to (12) as the discounted cost of failure. Roughly, if the cost of a surge appointment is smaller than the discounted failure cost for any appointment delay it is worthwhile to use surge capacity as soon as needed.

**Proposition 1 (First Available (FA)).** The First-Available (FA) policy, which books appointments in the first available time period (filling first base and then surge slots), is optimal if

\[
\frac{c_s}{c_F} \leq \min_{1 \leq d < d' \leq D} \left( p_F^{(d'-d)} \beta^{(d'-d)} - 1 \right) \left( p_S p_F|S + p_S p_F|S \right) + \sum_{l=d}^{d'-1} \beta^{l+1} p_F p_F^l. \tag{13}
\]

The expression on the right-hand side of (13) is not necessarily monotone in \( d' \) in each \( d \), thus it cannot generally be simplified further, say, by setting \( d' = d + 1 \). Nevertheless considering the special case of \( d' = d + 1 \) provides some insight. Setting \( d = 1 \) and \( d' = d + 1 = 2 \) on the right-hand side of (13) gives the intuitive requirement

\[
c_s + c_F \left( 1 - p_F \beta \right) \left( p_S p_F|S + p_S p_F|S \right) \leq c_F p_F \beta, \tag{14}
\]

which stipulates that the cost of giving a customer a surge appointment for the next period—which includes the cost of surge \( (c_s) \) plus the cost of failure if the patient does not fail in the current period and fails after the appointment—is cheaper than postponing the appointment by one more period and using a base slot and absorbing the expected failure cost \( c_F p_F \beta \). While the reduced condition (14) is not sufficient it gives an indication that an FA policy might be optimal, and (14) is easy to evaluate.

**Proof.** This proof is based on an incremental cost analysis, studying the marginal (discounted) cost of scheduling an appointment. The discounted expected cost to service an appointment scheduled for time period \( d \) is at most \( c_s p_F^d \beta^d \). The discounted expected cost for scheduling an appointment for \( d \) time periods into the future, includes this service cost and the failure cost in (12). It is bounded above by

\[
c_s p_F^d \beta^d + c_F \left( p_F^d \beta^d \left( p_S p_F|S + p_S p_F|S \right) + \sum_{l=0}^{d-1} \beta^{l+1} p_F p_F^l \right). \tag{15}
\]
and, dropping the service cost, it is bounded below by

\[ c_F \left( p_F^d \beta^d \left( p_{SPF|S} + p_{SPF|S} \right) + \sum_{i=0}^{d-1} \beta^{i+1} p_F p_F^i \right) \]  \tag{16}

The scheduler must make the decision to schedule an appointment for \( d \) or \( d' \) with \( 1 \leq d < d' \leq D \). Each decision \( d \) and \( d' \) are bounded above by (15) and below by (16). To check if the expected discounted cost of the decision to schedule an appointment for \( d \) is always less than the expected discounted cost of the decision to schedule in \( d' \), there are four cases that must hold. Namely,

\[
0 \geq \left[ c_F^d + c_F \left( p_F^d \beta^d \left( p_{SPF|S} + p_{SPF|S} \right) + \sum_{i=0}^{d-1} \beta^{i+1} p_F p_F^i \right) \right] 
- \left[ c_F^d \beta^d + c_F \left( p_F^d \beta^d \left( p_{SPF|S} + p_{SPF|S} \right) + \sum_{i=0}^{d-1} \beta^{i+1} p_F p_F^i \right) \right],
\tag{17}
\]

\[
0 \geq \left[ c_F^d + c_F \left( p_F^d \beta^d \left( p_{SPF|S} + p_{SPF|S} \right) + \sum_{i=0}^{d-1} \beta^{i+1} p_F p_F^i \right) \right] 
- \left[ c_F^d \beta^d + c_F \left( p_F^d \beta^d \left( p_{SPF|S} + p_{SPF|S} \right) + \sum_{i=0}^{d-1} \beta^{i+1} p_F p_F^i \right) \right],
\tag{18}
\]

\[
0 \geq \left[ c_F \left( p_F^d \beta^d \left( p_{SPF|S} + p_{SPF|S} \right) + \sum_{i=0}^{d-1} \beta^{i+1} p_F p_F^i \right) \right] 
- \left[ c_F^d \beta^d + c_F \left( p_F^d \beta^d \left( p_{SPF|S} + p_{SPF|S} \right) + \sum_{i=0}^{d-1} \beta^{i+1} p_F p_F^i \right) \right],
\tag{19}
\]

\[
0 \geq \left[ c_F \left( p_F^d \beta^d \left( p_{SPF|S} + p_{SPF|S} \right) + \sum_{i=0}^{d-1} \beta^{i+1} p_F p_F^i \right) \right] 
- \left[ c_F \left( p_F^d \beta^d \left( p_{SPF|S} + p_{SPF|S} \right) + \sum_{i=0}^{d-1} \beta^{i+1} p_F p_F^i \right) \right].
\tag{20}
\]

If inequalities (17)-(20) hold for all pairs \( 1 \leq d < d' \leq D \), then it will always be optimal to schedule the first available appointment, base or surge, regardless of the system state. This conclusion does not depend on the distribution of arrivals.

The right hand side of inequality (18) is trivially greater than the right hand side of inequality (17) as the latter has also the term \(-c_F^d \beta^d\) on the second line. Since \( d' > d \) and \( \beta p_F < 1 \), \( p_F^d \beta^d \leq p_F^d \beta^d \), so the right hand side of inequality (17) is greater than that of inequality (20) and the right hand side of (20) is greater than the right hand side of (19). Since (19) and (20) follow immediately from the monotonicity of the (16) it suffices to show (18) to have all inequalities (17)-(20) hold.
The following sequence of simple manipulations shows that (18) is implied by (13) in the statement of the proposition.

\[ 0 \geq c_s p_F^d \beta^d + c_F \left( p_F^d \beta^d (p_{SP} p_{FS} + p_{SP} p_{FS}) + \sum_{i=0}^{d-1} \beta^{i+1} p_F p_F^i \right) - c_F \left( p_F^d \beta^d (p_{SP} p_{FS} + p_{SP} p_{FS}) + \sum_{i=0}^{d-1} \beta^{i+1} p_F p_F^i \right) \]

\[ \iff \]

\[ c_s p_F^d \beta^d \leq c_F \left( p_F^d \beta^d (p_{SP} p_{FS} + p_{SP} p_{FS}) + \sum_{i=0}^{d-1} \beta^{i+1} p_F p_F^i \right) - c_F \left( p_F^d \beta^d (p_{SP} p_{FS} + p_{SP} p_{FS}) + \sum_{i=0}^{d-1} \beta^{i+1} p_F p_F^i \right) \]

\[ \iff \]

\[ c_s p_F^d \beta^d \leq c_F \left( (p_F^d \beta^d - p_F^d \beta^d) (p_{SP} p_{FS} + p_{SP} p_{FS}) + \sum_{i=0}^{d-1} \beta^{i+1} p_F p_F^i \right) \]

\[ \iff \]

\[ c_s p_F^d \beta^d \leq c_F \left( (p_F^d \beta^d - p_F^d \beta^d) (p_{SP} p_{FS} + p_{SP} p_{FS}) + \sum_{i=0}^{d-1} \beta^{i+1} p_F p_F^i \right) \]

\[ \iff \]

\[ \frac{c_s}{c_F} \leq \frac{(p_F^d \beta^d - p_F^d \beta^d) (p_{SP} p_{FS} + p_{SP} p_{FS}) + \sum_{i=0}^{d-1} \beta^{i+1} p_F p_F^i}{p_F^d \beta^d} \]

\[ \iff \]

\[ \frac{c_s}{c_F} \leq \frac{p_F^d \beta^d (p_F^{(d-d)} \beta^{(d-d)} - 1) (p_{SP} p_{FS} + p_{SP} p_{FS}) + \sum_{i=0}^{d-1} \beta^{i+1} p_F p_F^i}{p_F^d \beta^d} \]

\[ \iff \]

\[ \frac{c_s}{c_F} \leq (p_F^{(d-d)} \beta^{(d-d)} - 1) (p_{SP} p_{FS} + p_{SP} p_{FS}) + \frac{\sum_{i=0}^{d-1} \beta^{i+1} p_F p_F^i}{p_F^d \beta^d} \]

which gives the result \( \square \)

Under the condition of Proposition 1, the scheduler uses surge capacity for \( d \) periods into the future even if there is available base capacity in period \( d + 1 \). If surge is sufficiently costly, this is reversed. The scheduler exhausts all base capacity in the scheduling horizon before using any surge capacity.

**Proposition 2 ((First Available Base (FB))).** Scheduling the first available base capacity over the entire booking horizon \( D \) before scheduling any surge capacity is optimal if

\[ \frac{c_s}{c_F} \geq \max_{1 < d < d' < D} \left( p_F^{(d-d)} \beta^{(d-d)} - 1 \right) (p_{SP} p_{FS} + p_{SP} p_{FS}) + \frac{\sum_{i=0}^{d-1} \beta^{i+1} p_F p_F^i}{p_F^d \beta^d}. \]  

(21)

Note that the union of Proposition 1 and 2 does not cover the parameter space. Indeed, in contrast to the FA policy, using surge as infrequently as possible may be optimal:
Proposition 3 (First Available Base Last Available Surge (FBLS)). Scheduling the first available base capacity over the entire booking horizon \( D \) before scheduling the last available surge capacity is optimal if

\[
\frac{c_s}{c_F} \geq \max_{1 \leq d < d' < D} \frac{(p_F^{d'} \beta^{d'} - p_F^d \beta^d)(p_S p_F | S + p_S p_{F|S}) + \sum_{i=d}^{d'-1} \beta^{i+1} p_F p_F^i}{p_F^d \beta^d - p_F^{d'} \beta^{d'}}.
\]

Proofs for Propositions 2 and 3 are in the appendix and follow an incremental cost analysis similar of that of Proposition 1.

In the parameter range satisfying (22), the cost to add a surge appointment is sufficiently high relative to the cost of failure that even if surge must be used (given that all appointment requests must be scheduled when feasible), it should be scheduled as far as possible into the future. It is better in this parameter range to have a customer fail than to expense the surge cost. This is unlikely in the health care setting of Congestive Heart Failure where the average cost of a readmission is approximately $13k and the cost of a surge appointment is approximately $200. However, in a maintenance service contact, it may be the case that preventative maintenance is costly compared to failure costs and Proposition 3 provides sufficient conditions to evaluate if surge is ever worth using.

When the parameters do not satisfy the requirements of either of Proposition 1 or 3, incremental cost arguments do not work. We also consider a hybrid of FA and FBLS that uses base capacity first and then the first available surge capacity when all base capacity slots are full in the booking horizon. We refer to this policy as First Base First Surge (FBFS) which is used in practice by health care schedulers. We shall show in §5.3 that this hybrid nature renders the performance of FBFS robust to the parameters under moderate capacity utilization in the health care setting.

5. Numerical Experiments

As a reference application, we adapt our model to Congestive Heart Failure (CHF) follow-up appointments and draw on our experience at Northwestern Memorial Hospital (NMH); see NMHC (2016). The preventative appointment being scheduled is a follow-up visit with a cardiologist post-discharge from an initial CHF hospital admission. The failure this appointment aims to prevent is a readmission within the first 30 days post-discharge.

NMH has participated in Medicare’s BPCI (CMS 2016) program since 2015. The BPCI contract holds NMH accountable for the total cost of care for CHF patients during their hospitalization and for 30 days post-discharge from their initial admission. Readmissions are one of the highest costs post-discharge. Kilgore et al. (2017) report that Medicare pays an approximate average of $13k per heart failure hospitalization. The NMH cardiology department is pursuing a reduction of 30-day readmissions and operates on the assumption—supported e.g. by Ryan et al. (2013)
and Hernandez et al. (2010)—that an early follow-up appointment post-discharge reduces the readmission probability by approximately 3%. At NMH every CHF patient being discharged from a hospital admission must be scheduled for a follow-up appointment within one week.

5.1. Parameter Test Bed

To evaluate how sensitive policies are to the underlying model parameters, we create a test bed of scenarios starting with NMH’s parameter estimates as the baseline and then vary the parameters relative to their baseline.

NMH staffs the follow-up clinic with cardiologists, nurses, and support staff Monday, Wednesday, and Friday so we break a week into 3 equal time periods, each corresponding to one of these three days. All patients are scheduled within one week after discharge so we limit our booking horizon to $D = 3$. A cardiologist in the clinic has office visits only in the afternoon and the base capacity is $K_b = 4$ appointment slots. These base slots are available in the hospital’s centralized appointment system and are visible to the schedulers. The base capacity is constant and does not incur an incremental cost whether it is used or not. The scheduling policy currently used at NMH is First Available Base, then First Available Surge (FBFS): If a base capacity appointment is not available within one week post-discharge, the cardiologist is paged by the schedulers, and adds a surge capacity appointment to the first day that already has full base capacity. This additional capacity comes with a cost, $c_s$, (overtime for support staff and opportunity costs for the cardiologist’s other responsibilities). We estimate $c_s = $200 based on billing amounts from NMH and consider a range of values in our sensitivity analysis. We use the surge-cost specification from equation (5) with $K_b = 4$. Only one surge appointment can be added in any particular time period (due to limits in overtime, exam room availability, and physician clinic hours) so $K_s = 1$ and therefore $K = 5$.

CHF has a relatively high 30-day readmission rate with a national average of approximately 22%; see CMS (2017). Dharmarajan et al. (2013) study 30 day post-discharge readmissions for CHF patients: for each day 1, 2, ..., 30 they report the fraction of patients that were readmitted after discharge. We use the average of $-0.008$ (0.8%)—as the baseline for the per-day probability of failure. Since there are three periods per week, this translates to a per-period probability of failure, $p_F = \frac{7\text{days}\times 0.008\text{ProbabilityPerDay}}{3\text{PeriodsPerWeek}} = 0.01867$. We evaluate scenarios with $p_F \in \{0.01, 0.01267, 0.01867, 0.02467, 0.03\}$.

Following evidence from Ryan et al. (2013) and Hernandez et al. (2010) we set the baseline impact a follow-up appointment has on 30-day-readmission probability to 3%, that is $p_{F|\bar{S}} - p_{F|S} = 0.03$. We take the perturbation $\{0\%, 1\%, 3\%, 5\%, 7\%\}$ around 3%. Fixing $p_{F|\bar{S}}$ to 0.23, this is the same as varying $p_{F|S} \in \{0.16, 0.18, 0.20, 0.22, 0.23\}$
For the no-show probability baseline, we draw on the estimate of $p_S = 0.93$ in Daggy et al. (2010) for the age group of patients 60 years or older. This is aligned with Medicare patients who are the subject of the BPCI program. We evaluate scenarios with $p_S \in \{0.85, 0.87, 0.93, 0.95, 0.99\}$.

We estimate the cost of a CHF readmission to be $c_F = $13,000 and hold this number constant in all scenarios. This estimate is based on Medicare’s average payments for CHF admissions in Kilgore et al. (2017). We estimate the cost to service a surge appointment is $c_s = $200 and evaluate scenarios with $c_s \in \{50, 100, 200, 500, 1000, 10000, 13000\}$ to capture high and low cost appointments.

To estimate the distribution of appointment request arrivals, $A^t$, we use information on the number of discharged CHF patients per year from NMH. The average yearly discharge of CHF patients is approximately 460. Given 52 weeks per year and 3 periods per week, the average requests per period is $E[A^t] = 3$. Therefore, we fit a uniform distribution on the support $\{0, 1, \ldots, 6\}$, i.e. $P(A^t = a) = \frac{1}{7}$ for $a = 0, 1, \ldots, 6$. This equates 468 average annual appointment requests which is consistent with NMH. Since $K_b = 4$, demand is on average smaller than base capacity, and we consider this system in “low load”. We also investigate how capacity utilization affects optimal policies so we consider a “high load” appointment request arrival process with 50% higher mean: $E[A^t] = 4\frac{3}{2} > K_b$ and distribution: $P(A^t = a) = \frac{1}{41}$ for $a = 0, \ldots, 4$ and $P(A^t = a) = \frac{8}{41}$ for $a = 5, 6$. Both distributions have a common support so that the state and action space remain unchanged and we simply over-weigh the support above base capacity to model high load.

We assume an annual discount rate of 5%, which corresponds to a per period discount factor $\beta = 0.9996795$.

The above construction results in 1750 scenarios (875 per load level) in our sensitivity test bed which is designed to be centered at parameter estimates for NMH and from literature. We vary these values to capture the impact of parameter variation on optimal scheduling policies and test the robustness of different policies. One exception is our sensitivity on surge cost $c_s$, which we consider also at values much higher than the baseline of $200$ so as to capture parameter scenarios where $\frac{c_s}{c_F}$ is high; as in Proposition 3.

5.2. Numerical Solution Method

We compute the optimal stationary policy using a linear program (LP). We show in §3 that this LP has $(M + 1) \cdot (K + 1)^D$ decision variables. This is solvable for the problem sizes of a standard follow-up clinic even at at a large metropolitan hospital such as NMH where: $K = 5$, $D = 3$, and $M = 6$. This model instance has 1512 decision variables and 9618 constraints and takes approximately 5 minutes to solve using Gurobi’s simplex optimization engine and a standard laptop (i7 processor with 16Gb of RAM). Numerical precision is a key factor in computation time due to the relative high cost of failure compared to some extremely small transition probabilities. Increasing the model
instance size along any of the dimensions $M$, $K$, or $D$, begins to push the limit of our current computational practicality.

We compute the value function, $V_\pi(S)$, for four stationary policies $\pi$ (optimal from LP, FA, FBFS, and FBLS) by solving the Bellman equation:

$$V_\pi(S) = \sum_a \mathbb{P}(A = a) \left( g(S, U(a)) + \beta \sum_{S'} \mathbb{P}(S + U(S,a), S') V_\pi(S') \right)$$

for each $S, U$.

The values we report are with the initial state set to the empty state. Because our discount factor $\beta = 0.9996795$ is close to one, the value is relatively insensitive to the initial state. Also, because of irreducibility, a policy that performs the same as the optimal policy for one initial state, has the optimal value in all states. We say FA, FBFS, or FBLS is near-optimal—and henceforth simply optimal—if the value function is within 0.1\% of the optimal value at all initial states. The 0.1\% tolerance buffers for numerical errors from computer rounding, but independently is reasonable evidence for near-optimality.

Separately, we compute the stationary distribution of the Markov chain $S^t$, $t = 0, 1, 2, \ldots$ under each of the policies—optimal from LP, FA, FBFS, and FBLS. Provided that $\mathbb{P}\{A^t > K_b + K_s\} > 0$ all four policies induce an irreducible chain. Since the state space is finite, this guarantees the existence of a stationary distribution. From the stationary distribution we tease out the fraction of periods during which surge capacity is used. This informs an important consideration for management in their choice of a scheduling policy.

### 5.3. Results

In Figure 1 we report the fraction of the 1750 (875 per load) parameter scenarios when each of FA, FBFS, and FBLS is optimal.

![Figure 1](image-url)

**Figure 1** The percent of parameter scenarios where each of FA, FBFS, and FBLS policies is optimal: (LEFT) Low load (RIGHT) High load
**Low load:** The FA policy is optimal in 48.6% of the scenarios. Per Proposition 1, optimality of FA is expected for the scenarios that have low cost surge appointments and high cost failures. In fact, the *sufficient* condition in Proposition 1 is satisfied in all the scenarios where FA is numerically optimal. In other words, that condition is also *necessary* within our numerical test bed and low load capacity utilization.

FBLS and FBFS are optimal in 28.6% and 14.3% of scenarios respectively. In the remaining 8.6% the optimal policy is different from FA, FBFS, and FBLS. Per Proposition 3, optimality of FBLS is expected in those scenarios where the differential impact of the follow-up appointment is small and/or the surge cost is large relative to the failure cost.

**High load:** The FA policy is optimal in 71.4% of the 875 scenarios. As the load increases, surge capacity must be used more often to the point of treating it the same as base capacity. The sufficient condition of Proposition 1 is agnostic to capacity load or arrival rates of new appointment requests. It covers here the same 48.6% scenarios that it covered in the low load case. The remaining 71.4%-48.6% = 22.8% percent of high load scenarios violate the sufficient condition but FA is still optimal.

FBLS and FBFS are optimal in 14.3% and 14.1% of the scenarios respectively. The three heuristics offer almost full coverage in this case, leaving *only* 0.2% of scenarios where the optimal policy is none of the three.

**Robustness:** Figures 2-4 report the sub-optimality of FA, FBFS, and FBLS policies over the entire test bed of parameters. We compute the percentage sub-optimal as the difference in values divided by the optimal value of the infinite horizon discounted expected cost when starting in the empty state (i.e. for a given stationary policy \( \pi \) we report \( \frac{V_\pi(S^0)}{V^*(S^0)} \) for all scenarios where \( \pi \) was not the optimal policy). While we use the empty initial state in our reported output, we found the percentage sub-optimality almost invariant with respect to the initial state under our discount factor.

These figures underscore several important facts:

(i.) Although FA is optimal in 48.6% of the low load scenarios, it performs rather poorly in other scenarios, for example, in those where the optimal policy is FBLS. This is because FA uses more surge than optimal when surge capacity is expensive. In high load scenarios, all policies use surge relatively frequently so the implications of FA’s greediness are not so severe.

(ii.) The FBFS policy—inspired by the practice at NMH—is optimal in only 14.3% of the 875 load load scenarios but is robust *across all low load scenarios*. In its worst parameter scenario, it generates an optimality gap of 3.3% (average optimality gap of 1.5%). Its performance deteriorated in high load scenarios. FBFS introduces an average sub-optimality of 9.5% when the optimal policy is FA and the system is in high load.
(iii.) The FBLS policy performs within 3.4% of optimal when it itself is not optimal in the low load regime. It performs only slightly worse than FBFS in these scenarios. This can be attributed to the infrequent use of surge under low loads. Both policies schedule surge capacity only when all base capacity is full and take the same action when no surge is scheduled. In high load scenarios, FBLS’s “postponement of surge” generates an average optimality gap of 14.5% in the scenarios where FA is optimal.

(iv.) The prescriptive picture that emerges is mixed. In low load scenarios FBFS performs well across the entire test bed. NMH is correct in using this policy as FBFS is robust, unlike FA. A hospital that operates in high load has to be more careful. Here, because surge must be used
Figure 4 Sub-optimality of FBLS: (LEFT) Low load: FBLS stays within 3.4% of optimal in all parameter scenarios and within 0.5% of optimal when FBFS is optimal or the optimal policy is not one of FA, FBFS, or FBLS. (RIGHT) High load: FBLS perform an average of 14.5% worse than optimal in parameter scenarios where FA is optimal. FBLS performs within 0.9% of optimal in all other scenarios.

Figure 5 Sub-optimality of FAFBFS: (LEFT) Low load: FAFBFS stays within 0.4% of optimal in all low load scenarios. (RIGHT) High load: FAFBFS stays within 12.6% of optimal in all high load scenarios.

frequently, FA performs within 15.2% of optimal in all scenarios and is optimal in 71.4% of scenarios. While FBFS has a worst case optimality gap that is only slightly larger, 16.1%, it is sub optimal in 85.9% of scenarios.

A simple rule of thumb that emerges from our numerical experiments, which is robust in both load regimes is as follows: When the condition in Proposition 1 holds use FA, otherwise use FBFS. We call this rule FAFBFS and Figure 5 reports the optimality gap of this rule across all 1750 parameter scenarios. In all of the low load scenarios this simple rule of thumb guarantees an optimality gap of at most 0.4%, and in high load scenarios it guarantees an optimality gap of at most 12.6% in the test bed. Further, FAFBFS picks the optimal policy in 62.7% of all scenarios.

Another benefit of FBFS is its relatively infrequent use of surge capacity. For each of the three heuristics we compute the long-run fraction of periods that surge capacity is utilized (i.e, the stationary probability that the number of appointments scheduled for the current time period strictly exceeds $K_b$). Since in these experiments we have only one surge slot per period, this fraction also equals the average number of surge slots utilized. Figure 6 displays how—with all other
parameters held fixed at their baseline—the fraction of surge periods used by the optimal policy changes with the surge cost $c_s$. As cost of surge increases, the relative value of the appointment $c_F/c_s$ decreases and the optimal policy moves from FA to FBLS.

In low load scenarios, FA uses surge capacity approximately 28% of the time in steady state, while FBFS and FBLS surge in only 0.2% of the periods. Since, even in scenarios where FA is optimal, FBFS performs extremely well, (non-monetary) concern for frequent surging further supports the use of FBFS. Even in high load scenarios, FA surges significantly more than FBFS and FBLS. In the health care setting reducing the amount of surge use may be an important managerial concern.

6. Extension to Dynamic and Appointment-Dependent Failure Rates

So far we have assumed a constant failure rate, $p_F$, in each period a customer awaits their scheduled appointment and constant probabilities of failure after a scheduled appointment, $p_{F|S}$ and $p_{F|\bar{S}}$, independent of when the appointment is scheduled. In reality it is possible that $p_F$ changes depending on how long a customer waits for their appointment. If a new engine is installed improperly, complications can begin compiling and the risk of failure increase over time. In contrast, a patient released from the hospital after a major health episode may be in a fragile state immediately following discharge, but as time passes her body heals and the risk of failure decreases. The probability a customer fails during the contract window after their scheduled appointment, $p_{F|S}$ and $p_{F|\bar{S}}$, is more clearly a function of when an appointment is scheduled, since this determines the amount of time in the contract window remaining after the appointment. In this section we aim to address these modeling limitations by extending our numerical analysis to dynamic failure rates in each period and appointment-dependent risk of failure in the contract window following appointments.

A full dynamic programming formulation that captures dynamic failure rates requires keeping track of the “age” of each customer —the number of time periods that have elapsed from
the patient’s discharge and hence from the appointment request—rendering the state-space too large for the computation of optimal policies and the model too challenging for the derivation of structural properties. Instead, we study a dynamic $p_F$ through bounds, and separately analyze appointment-dependent $p_{F|S}$ and $p_{F|\bar{S}}$ by augmenting the state space of our model with partial “age” information. In the numerical experiments of this section we continue to use the example of CHF at NMH and therefore use customer and patient interchangeably.

6.1. Dynamic $p_F$

Dharmarajan et al. (2013) argue that a constant failure rate, $p_F$, is a reasonable approximation for Heart Failure patients. It is plausible, however, that the failure/hazard rate be in some cases increasing, decreasing or non-monotone in the time-from-discharge. Patients might be the most fragile in the first days after discharge. In later days after discharge, non-adherence to prescriptions may have an effect that is increasing in time. Similarly in preventative maintenance, it may be the case that failure is unlikely in the near-term but increases over time with normal machine usage.

Let $V_\pi[p_F](S^0)$ be the infinite horizon discounted expected cost of a stationary policy $\pi$ under the constant failure rate $p_F$ and initial state $S^0$. It is not generally true that $V_\pi[p]$ is decreasing in $p$. If the load is sufficiently high that surge must be used occasionally but this surge is very expensive relatively to the cost of failure ($c_s \gg c_F$), it may be better to let patients fail than utilize surge. Realistically the cost of a major failure exceeds the cost of a surge capacity slot for a preventative appointment, i.e., $c_F \geq c_s$. In this case, monotonicity of optimal values and under FA and FBFS follows from standard sample path arguments that we omit.

**Lemma 1.** Suppose that $c_F \geq c_s$. Then, for every initial state $S^0$ and any $\overline{p_F} \geq p_F$

$$V_\pi[p_F](S^0) \leq V_\pi[\overline{p_F}](S^0),$$

and for $\pi \in \{FA, FBFS\}$

$$V_\pi[p_F](S^0) \leq V_\pi[\overline{p_F}](S^0).$$

**Dynamic Failure Rate:** The failure probability is given by a function $\tilde{p}_F(w) : \mathbb{Z}_+ \to \mathbb{R}_+$ where $w$ is the “age” of a patient—the number of time periods that have elapsed from the patient’s discharge (and hence from the appointment request). A patient with an appointment scheduled for day $d$, conditional on “surviving” up to (including) day $w-1$, fails on day $w < d$ with probability $\tilde{p}_F(w)$. We let

$$\underline{p_F} = \min_w \tilde{p}_F(w), \text{ and } \overline{p_F} = \max_w \tilde{p}_F(w).$$

With dynamic failure rates a more elaborate state-descriptor is needed. It must include the “age” of each of the patients in the appointment book; see Liu et al. (2010). We do not need to fully spell
out this state descriptor for our purposes. In what follows “stationary policy” refers to stationarity relative to the larger state descriptor. Evidently, policies that are stationary relative to our original state descriptor, are also stationary relative to the more granular one.

While the state space is (significantly) larger it is still finite and standard arguments guarantee the existence of an optimal stationary policy, $\tilde{\pi}^*$ for the dynamic failure rate $\tilde{p}_F$. Argued as in Lemma 1, it follows that

$$V_{\tilde{\pi}}[\tilde{p}_F] \leq V_{\pi^*}[\tilde{p}_F],$$

and for $\pi \in \{FA, FBFS\}$

$$V_{\pi}[\tilde{p}_F] \leq V_{\pi^*}[\tilde{p}_F],$$

so that for these policies

$$V_{\pi}[\tilde{p}_F] - V_{\pi^*}[\tilde{p}_F] \leq V_{\pi}[\tilde{p}_F] - V_{\pi^*}[\tilde{p}_F] \leq V_{\pi^*}[\tilde{p}_F] - V_{\pi^*}[p_F],$$

where $\tilde{\pi}^*$, $\pi^*$ are, respectively, the optimal policies for $\tilde{p}_F$ and $p_F$. That is, the sub-optimality of a policy for dynamic readmission rates can be at most the gap between the performance of this policy for the upper and lower bounds on $\tilde{p}_F(w)$.

In turn, for any stationary policy $\pi$,

$$\frac{V_{\pi}[\tilde{p}_F] - V_{\pi^*}[\tilde{p}_F]}{V_{\pi^*}[\tilde{p}_F]} \leq \frac{V_{\pi}[\tilde{p}_F] - V_{\pi^*}[p_F]}{V_{\pi^*}[p_F]}.$$  \hspace{1cm} (23)

To illustrate the bounds on the optimality gap given in (23) we return to the health care setting of CHF at NMH. We use the baseline estimates for the probability a patient no-shows, $1 - p_S = 0.07$, and the conditional probabilities of failure, $p_{F|S} = 0.23$ and $p_{F|\bar{S}} = 0.20$, which is a follow-up appointment impact of 3% on readmissions.

We fix $p_F = 0.01$ and vary $\overline{p}_F$ from 0.01 to 0.03. Dharmarajan et al. (2013) report that the readmission probability for each of the 30 days post-discharge from heart failure ranges from approximately 0.0057 to 0.01. Given 3 time periods per week this is equivalent to approximately 0.01 to 0.02 failure probability per period in our model (i.e. $0.01 \leq \overline{p}_F \leq 0.02$). We report results for the optimal policy and leading heuristic for each load regime in Figures 7 and 8. In each of the graphs we vary the (upper) cost of failure $\overline{p}_F$ and report the right-hand side of (23).

We make several observations:

1. Naturally, as one allows the support of $\tilde{p}_F(w)$ to be larger, the less accurate the bound. Throughout this analysis we keep $p_F = 0.01$. The extreme right of each graph corresponds to allowing the dynamic failure rate to oscillate between 0.01 and 0.03 which is a large range. Yet, even in that case, the sub-optimality introduced by using the optimal policy for $\overline{p}_F = 0.03$ is at most 10%.
2. With the support of $\tilde{p}_F$ restricted to $[0.1, 0.2]$ (consistent with the medical literature), the optimal policy for $\overline{p}_F = 0.02$ leads to an error of at most 5% in either the low-load or high-load scenarios.

3. Robustness of FBFS in low load: In the low load scenarios one can be agnostic to the dynamics in failure probability. FBFS, being nearly optimal in this load regime, mimics very closely the optimal for each $\overline{p}_F$. As long as the upper bound on the per period failure probability is less than 2.0% the optimality gap will be, at most, 5%.

4. Robustness of FA in high load: In high load scenarios, FA, recall, is always within 15% of optimal. Ignoring the structure of the failure probability and applying FA would cost at most 10% in the high load regime with the exception of the (non-realistic) cases where surge capacity costs are as large as $10k$ per slot.

6.2. Appointment-Dependent $p_{F|\bar{S}}$ and $p_{F|S}$

In the context of early follow-up appointments the scheduling window is significantly shorter than the contract window, making a constant $p_{F|\bar{S}}$ and $p_{F|S}$ a reasonable approximation. As the scheduling window grows with respect to the contract window, this assumption degrades as we illustrate.
with the following hypothetical: First consider a scheduling window equal to the contract window, i.e. the scheduler can schedule a preventative appointment for anytime in the contract window. Next, let Patient A be scheduled for a follow-up appointment immediately after discharge. The time remaining in the contract window after Patient A’s appointment is the entire contract window less one time period. Further, let Patient B be scheduled for his appointment in the next-to-last period in the contract window. There is only a single period in which the scheduler is responsible for the cost of a failure for Patient B but potentially many periods where the scheduler is responsible for failure costs of Patient A. In this hypothetical it is not reasonable to assume that the probability of failure after the appointment, \( p_{F|S} \) and \( p_{F|\bar{S}} \), is the same for Patient A and Patient B.

To capture variability in failure probabilities after appointments we augment the state descriptor \( S^t \) to include a parameter \( R^t = \{ R^t_d : d = 0, 1, \ldots, D \} \), which indicates the risk type for the group of patients scheduled for each period on the schedule. The risk parameter \( R^t \) can be thought of as capturing partial age or risk information for each group of patients scheduled for each period in the schedule. Specifically the risk type \( R^t_d \) defines the risk of failure after appointments and within the contract window, \( p_{F|S} \) and \( p_{F|\bar{S}} \), for the patients \( S^t_d \) (i.e. all patients scheduled at time \( t \) for \( d \) periods into the future). While a fully descriptive model would keep track of each patient on the schedule, we assign a single risk type to all patients scheduled for a particular time period, therefore only retaining partial information of patient age (or risk of failure after an appointment). Although this is a limitation, it allows for computational tractability of optimal stationary policies. We define the augmented state at period \( t \) as \((S^t, R^t)\).

We further limit the risk descriptor such that \( R^t_d \in \{1, 2, 3\} \) due to the instance size of our problem and for computational considerations. Let \( R^t_d = 1 \) be low risk, \( R^t_d = 2 \) be medium risk, and \( R^t_d = 3 \) be high risk of failure after the scheduled appointment.

States transition from \((S^t, R^t)\) to \((S^{t+1}, R^{t+1})\). We assume that the first component of the state \( S^t \) transitions to \( S^{t+1} \) as described in (4). If there are no failures for a given time period then we assume that the risk type \( R^t \) remains the same. If there is a failure, the risk type either remains the same or moves to a lower risk type. We define the risk parameter transition probabilities as,

\[
R^{t+1}_d = \begin{cases} 
R^t_d & \text{w.p. 1, if } S^{t+1}_d = S^t_d, \\
3 \text{ w.p. } \frac{1}{3} & \text{if } S^{t+1}_d < S^t_d, \text{ and } R^{t+1}_d = 3 \\
2 \text{ w.p. } \frac{1}{3} & \text{if } 0 < S^{t+1}_d < S^t_d, \text{ and } R^{t+1}_d = 3 \\
1 \text{ w.p. } \frac{1}{3} & \text{if } 0 < S^{t+1}_d < S^t_d, \text{ and } R^{t+1}_d = 2 \\
2 \text{ w.p. } \frac{1}{2} & \text{if } S^{t+1}_d < S^t_d, \text{ and } R^{t+1}_d = 2 \\
1 \text{ w.p. } \frac{1}{2} & \text{if } S^{t+1}_d < S^t_d, \text{ and } R^{t+1}_d = 1 \\
1 \text{ w.p. 1} & \text{if } S^{t+1}_d = 0 \\
\end{cases}
\] (24)
This says that if a patient fails then the remaining patients scheduled for that same period either have the same risk type as before or a reduced risk type. The argument for a reduced risk of failure is that it is those patients with the most risk that would tend to fail, reducing the average risk of remaining patients.

We assume that the risk type of the patients scheduled for \( d \) periods into the future can impact and be impacted by the actions of the scheduler. In particular the action \( U^t \) may depend on \( S^t \) through (1) but may also depend on the risk type \( R^t \).

A patient scheduled far into the scheduling horizon will have a lower probability of failure after their appointment than one scheduled soon after discharge since there is more time remaining in the contract window for the former. With the risk parameter \( R^t \) at the beginning of period \( t \), let \( \tilde{R}^t \) be the risk parameter at the end of period \( t \) after decision \( U^t \). Recall that \( \tilde{S}^t \) is the state after the action \( U^t \) is taken defined by (2). We define the impact of the scheduling decision on the risk parameter as,

\[
\tilde{R}^t_0 = R^t_0 \\
\tilde{R}^t_1 = \begin{cases} 
R^t_1, & \text{if } \tilde{S}^t_1 = S^t_1 \\
3, & \text{if } \tilde{S}^t_1 > S^t_1
\end{cases} \\
\tilde{R}^t_2 = \begin{cases} 
R^t_2, & \text{if } \tilde{S}^t_2 = S^t_2 \\
2, & \text{if } \tilde{S}^t_2 > S^t_2
\end{cases} \\
\tilde{R}^t_3 = 1
\]

Therefore at the end of period \( t \) the state is \((\tilde{S}^t, \tilde{R}^t)\).

The objective is to solve (10) with the addition that \( U^t(S^t, A^t, R^t) \) is now also a function of \( R^t \). The probability of moving from state \((S^t, R^t)\) to \((S^{t+1}, R^{t+1})\) is defined by (4) and (24). Note that any stationary policy that takes the same action for a given \( S^t \) and \( A^t \) regardless of \( R^t \) can be reduced to a policy using only the original model’s state definition. In addition any stationary policy from the original model can be implemented in the new expanded state definition by taking a single action for each \( S^t \) and \( A^t \) for all \( R^t \).

The number of decision variables in the corresponding LP grows from 1512 to 7392 and the time it takes to solve grows from 5 minutes to approximately 20 hours using the same machine specifications as Section 5.2. Even a small increase in the model structure complexity requires approximately 300 times the amount of time to run the numerical analysis. For the numerical experiments that follow we use the baseline estimates from Section 5.1 for \( c_s, c_F, p_S, \) and \( p_F \). We set \( p_F|\bar{S} - p_F|S = 0.03 \) so that the appointment impact remains constant at the baseline estimate of a 3% reduction in failure risk. We then test how the optimal policy changes with different possible risk types, \( R \). If we first define the risk types as \( R = 1 \Rightarrow p_{F|\bar{S}} = 0.23, R = 2 \Rightarrow p_{F|\bar{S}} = 0.23, \) and
Table 1  Policy Performance with appointment-dependent $p_{F|S}$ and $p_{F|S}$. Two output metrics are computed for each of the policies FA, FBFS, and FBLS on each of the 4 cases analyzed. OPT is the percentage of 7392 states where the policy takes the optimal action in stationarity. $\frac{V_*(S^0) - V^*(S^0)}{V^*(S^0)}$ is the optimality gap for a given policy and risk type. We can see in the original risk type, FA is optimal. Further as the magnitude of appointment-dependent $p_{F|S}$ and $p_{F|S}$ changes, FBFS and FBLS perform better with respect to the optimality gap and FA performs worse. In all cases we analyze the optimality gap of any of the policies is less than 1%.

<table>
<thead>
<tr>
<th></th>
<th>$R_{Original}$</th>
<th>$R_{Low}$</th>
<th>$R_{Mid}$</th>
<th>$R_{High}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPT</td>
<td>$\frac{V_<em>(S^0) - V^</em>(S^0)}{V^*(S^0)}$</td>
<td>OPT</td>
<td>OPT</td>
<td>OPT</td>
</tr>
<tr>
<td>FA</td>
<td>100.0 %</td>
<td>90.5 %</td>
<td>79.7 %</td>
<td>63.0 %</td>
</tr>
<tr>
<td></td>
<td>0.00 %</td>
<td>0.00 %</td>
<td>0.02 %</td>
<td>0.84 %</td>
</tr>
<tr>
<td>FBFS</td>
<td>45.0 %</td>
<td>49.1 %</td>
<td>50.2 %</td>
<td>50.0 %</td>
</tr>
<tr>
<td></td>
<td>0.70 %</td>
<td>0.70 %</td>
<td>0.68 %</td>
<td>0.15 %</td>
</tr>
<tr>
<td>FBLS</td>
<td>42.9 %</td>
<td>47.0 %</td>
<td>48.1 %</td>
<td>47.8 %</td>
</tr>
<tr>
<td></td>
<td>0.70 %</td>
<td>0.70 %</td>
<td>0.68 %</td>
<td>0.15 %</td>
</tr>
</tbody>
</table>

$R = 3 \Rightarrow p_{F|S} = 0.23$, we get the point estimates used in the Section 5 and the problem reduces to the original model. We find the same optimal policy of FA and the policy can be reduced to a policy on the original state descriptor. We test three additional definitions of $R$ where the impact of the scheduling decision on the probability of failure after an appointment increase from Low to Mid to High. We define the three risk profiles as,

$$R_{Original} = \begin{cases} 
1 \Rightarrow p_{F|S} = 0.23 \Rightarrow p_{F|S} = 0.20 \\
2 \Rightarrow p_{F|S} = 0.23 \Rightarrow p_{F|S} = 0.20 \\
3 \Rightarrow p_{F|S} = 0.23 \Rightarrow p_{F|S} = 0.20 
\end{cases}$$

$$R_{Low} = \begin{cases} 
1 \Rightarrow p_{F|S} = 0.2113 \Rightarrow p_{F|S} = 0.1813 \\
2 \Rightarrow p_{F|S} = 0.2207 \Rightarrow p_{F|S} = 0.1907 \\
3 \Rightarrow p_{F|S} = 0.23 \Rightarrow p_{F|S} = 0.20 
\end{cases}$$

$$R_{Mid} = \begin{cases} 
2 \Rightarrow p_{F|S} = 0.2160 \Rightarrow p_{F|S} = 0.1860 \\
3 \Rightarrow p_{F|S} = 0.23 \Rightarrow p_{F|S} = 0.20 
\end{cases}$$

$$R_{High} = \begin{cases} 
2 \Rightarrow p_{F|S} = 0.2113 \Rightarrow p_{F|S} = 0.1813 \\
3 \Rightarrow p_{F|S} = 0.23 \Rightarrow p_{F|S} = 0.20 
\end{cases}$$

Table 1 summarizes the results of the numerical analysis. Recall that for the baseline parameters, the FA policy was optimal, however, the FBFS policy that mimics practice was close in value to optimal, which is illustrated in the first column of Table 1. In the above three cases, as we move from $R_{Low}$ to $R_{Mid}$ to $R_{High}$ the percentage of states that use the FA policy decreases, which is as expected since delaying an appointment is, in a sense, less costly than the original model. Further the optimality gap of FA increases as we move from $R_{Low}$ to $R_{Mid}$ to $R_{High}$, while it decreases for both FBFS and FBLS. These results provide reinforcing evidence that the policy implemented at NMH is nearly optimal in total cost and has the managerial benefit of using less surge capacity. We further note that the optimality gap for all policies and scenarios analyzed in Table 1 are less than 1%, which gives evidence that $p_{F|S}$ and $p_{F|S}$ have a small impact in the setting of CHF early follow-up appointments.
7. Conclusion

We analyzed preventative appointment scheduling and its impact on the broader service supply network when the firm is responsible for service and failure costs. We presented sufficient conditions for the optimality of simple policies that can be evaluated without solving a dynamic program. Adapting our model to preventative appointments in the health care setting at Northwestern Memorial Hospital, we show that the FBFS policy implemented for CHF follow-ups performs within 3.3% of optimal under low load parameter scenarios. We further advocate policy selection based on our optimality conditions and show that this FAFBFS performs within 12.6% of optimal for all load and parameter scenarios in our test bed. Our results give implementable insights on how to optimally schedule preventative appointments and use surge capacity when a firm is responsible for the total costs of appointments and failures.

Appendix

Proof: Proposition 2 This proof is based on incremental cost analysis, studying the marginal (discounted) cost of scheduling an appointment. The discounted expected cost to service an appointment scheduled for time period $d$ is at most $c_s p_F^d \beta^d$. The total cost for scheduling an appointment for $d$ time periods into the future, includes this service cost and the failure cost in (12). It is bounded above by

$$c_s p_F^d \beta^d + c_F \left( p_F^d \beta^d (p_S p_{I|S} + p_S p_{I|S}) + \sum_{l=0}^{d-1} \beta^{l+1} p_F p_F^l \right)$$

and, dropping the service cost, it is bounded below by

$$c_F \left( p_F^d \beta^d (p_S p_{I|S} + p_S p_{I|S}) + \sum_{l=0}^{d-1} \beta^{l+1} p_F p_F^l \right)$$

For $1 < d < d' < D$ that satisfy the following inequality, then it is cheaper in expectation to schedule for a base capacity slot in $d'$ than a surge capacity slot for $d$ regardless of other appointments on the schedule and the evolution of the system. If (27) holds for all pairs $1 \leq d < d' \leq D$, then it will always be optimal to schedule all available base capacity before scheduling any surge capacity. Further since (26) is increasing in $d$ it will always be optimal to schedule the first available base capacity appointment. This conclusion does not depend on the distribution of arrivals.

$$0 \leq \left[ c_s p_F^d \beta^d + c_F \left( p_F^d \beta^d (p_S p_{I|S} + p_S p_{I|S}) + \sum_{l=0}^{d-1} \beta^{l+1} p_F p_F^l \right) \right]$$

$$- \left[ c_F \left( p_F^{d'} \beta^{d'} (p_S p_{I|S} + p_S p_{I|S}) + \sum_{l=0}^{d'-1} \beta^{l+1} p_F p_F^l \right) \right].$$

(27)

The following sequence of simple manipulations shows that (27) is implied by (21) in the statement of the corollary.
\[0 \leq c_s p_F^d \beta^d + c_F \left( p_F^d \beta^d (p_{SF} + p_{SF}) + \sum_{l=0}^{d-1} \beta^{l+1} p_F^l \right)
\]

\[- \left[ c_F \left( p_F^d \beta^d (p_{SF} + p_{SF}) + \sum_{l=0}^{d-1} \beta^{l+1} p_F^l \right) \right]\]

\[c_s p_F^d \beta^d \geq c_F \left( p_F^d \beta^d (p_{SF} + p_{SF}) + \sum_{l=0}^{d-1} \beta^{l+1} p_F^l \right) - \left[ c_F \left( p_F^d \beta^d (p_{SF} + p_{SF}) + \sum_{l=0}^{d-1} \beta^{l+1} p_F^l \right) \right]\]

\[c_s p_F^d \beta^d \geq c_F \left( p_F^d \beta^d - p_F^d \beta^d \right) \left( p_{SF} + p_{SF} \right) + \sum_{l=0}^{d-1} \beta^{l+1} p_F^l\]

\[\frac{c_s}{c_F} \geq \left( p_F^d \right) \left( d^d - d \right) - 1 \left( p_{SF} + p_{SF} \right) + \sum_{l=0}^{d-1} \beta^{l+1} p_F^l\]

which gives the result. \(\square\)

**Proof: Proposition 3** This proof is based on incremental cost analysis, studying the marginal (discounted) cost of scheduling an appointment. The discounted expected cost to service an appointment scheduled for time period \(d\) is at most \(c_s p_F^d \beta^d\). The total cost for scheduling an appointment for \(d\) time periods into the future, includes this service cost and the failure cost in (12). It is bounded above by

\[c_s p_F^d \beta^d + c_F \left( p_F^d \beta^d \left( p_{SF} + p_{SF} \right) + \sum_{l=0}^{d-1} \beta^{l+1} p_F^l \right)\] (28)

and, dropping the service cost, it is bounded below by

\[c_F \left( p_F^d \beta^d \left( p_{SF} + p_{SF} \right) + \sum_{l=0}^{d-1} \beta^{l+1} p_F^l \right)\] (29)

For \(1 < d < d' < D\) that satisfy the following inequality, it is cheaper in expectation to schedule for a surge capacity slot in \(d'\) than for a surge capacity slot for \(d\) regardless of other appointments on the schedule and the evolution of the system. This conclusion does not depend on the distribution of arrivals.

\[0 \leq c_s p_F^d \beta^d + c_F \left( p_F^d \beta^d \left( p_{SF} + p_{SF} \right) + \sum_{l=0}^{d-1} \beta^{l+1} p_F^l \right)
\]

\[- \left[ c_s p_F^d \beta^d + c_F \left( p_F^d \beta^d \left( p_{SF} + p_{SF} \right) + \sum_{l=0}^{d-1} \beta^{l+1} p_F^l \right) \right]\]

\[c_s p_F^d \beta^d + c_F \left( p_F^d \beta^d \left( p_{SF} + p_{SF} \right) + \sum_{l=0}^{d-1} \beta^{l+1} p_F^l \right)\]

\[\left( p_F^d \beta^d - p_F^d \beta^d \right) \left( p_{SF} + p_{SF} \right) + \sum_{l=0}^{d-1} \beta^{l+1} p_F^l\]

(30)

Note that (30) \(\Rightarrow (21)\) and therefore if (30) holds for all pairs \(1 < d < d' < D\) an optimal policy will be schedule the first available base capacity until all base capacity is full, then schedule the last available surge capacity when no base capacity is available.

The following sequence of simple manipulations shows that (30) is implied by (22) in the statement of the corollary.

\[0 \leq c_s p_F^d \beta^d + c_F \left( p_F^d \beta^d \left( p_{SF} + p_{SF} \right) + \sum_{l=0}^{d-1} \beta^{l+1} p_F^l \right)
\]

\[- \left[ c_s p_F^d \beta^d + c_F \left( p_F^d \beta^d \left( p_{SF} + p_{SF} \right) + \sum_{l=0}^{d-1} \beta^{l+1} p_F^l \right) \right]\]

\[c_s p_F^d \beta^d + c_F \left( p_F^d \beta^d \left( p_{SF} + p_{SF} \right) + \sum_{l=0}^{d-1} \beta^{l+1} p_F^l \right)\]
\[ c_s \beta^d \geq c_F \left[ (p_F^d - p_F^d) (p_S p_F|S + p_S p_F|\bar{S}) + \sum_{i=d}^{d-1} \beta^{i+1} p_F p_F^i \right] \]

\[ c_s \geq \frac{(p_F^d - p_F^d) (p_S p_F|S + p_S p_F|\bar{S}) + \sum_{i=d}^{d-1} \beta^{i+1} p_F p_F^i}{p_F^d - p_F^d} \]

which gives the result. \(\Box\)

References


