Parametric Forecasting and Stochastic Programming Models for Call-Center Workforce Scheduling

Noah Gans
OPIM Department
The Wharton School
U. Pennsylvania

Haipeng Shen
Department of Statistics and OR
UNC Chapel Hill
IIM, U. Hong Kong

Yong-Pin Zhou
ISOM Department
The Foster School
U. Washington

Nikolay Korolev Alan McCord Herbert Ristock
Genesys Telecommunications Laboratories, Inc.

April 1, 2015

Abstract

We develop and test an integrated forecasting and stochastic programming approach to workforce management in call centers. We first demonstrate that parametric forecasts, discretized using Gaussian quadrature, can be used to drive stochastic programs whose results are stable with relatively small numbers of scenarios. We then extend our approach to include forecast updates and two-stage stochastic programs with recourse and provide a general modeling framework for which recent, related models are special cases. In our formulations, the inclusion of multiple arrival-rate scenarios allows call centers to meet long-run average quality-of-service targets, while the use of recourse actions help them to lower long-run average costs. Experiments with two large sets of call-center data highlight the complementary nature of these elements.

1 Introduction

Inbound telephone call centers handle service requests that originate from customers calling in, and they use a hierarchical staffing and scheduling system (Gans et al. 2003, Akşin et al. 2007). The process begins with forecasts of the arrival rate of calls over a planning horizon for scheduling agents, which may range from a day to several weeks. The forecasts then drive queueing models that determine how staffing levels affect system congestion over short, 15-minute to 1-hour, time intervals within the horizon. The queueing formulae determine staffing levels over the short time intervals, and, in turn, constraints to be met as the call center develops staff schedules. A rostering process then matches employees with required schedules. Through this sequence, the forecasted arrival process of calls to the center drives employee schedules.

Traditionally, call centers assume that arrival-rate forecasts are correct. They use point forecasts of arrival rates to determine staffing levels and, in turn, deterministic staffing-level requirements to drive scheduling decisions. But arrival-rate forecasts are often not perfect, and when realized arrival rates do not match those

forecasted, system performance naturally deviates from managers' expectations. Higher-than-expected arrival rates lead to understaffing, which drives up waiting times and abandonment rates, while unexpectedly low arrival rates result in overstaffing and the "overservice" of customers.

Work within the statistics and operations management literatures has begun to address the problem of how call centers – and other high volume service businesses – can better manage the capacity-demand mismatch that results from arrival-rate uncertainty. Earlier papers have explored the impact of arrival-rate uncertainty (Grassman 1988, Chen and Henderson 200, Jongbloed and Koole 2001, Ross 2001), and more recent work has explicitly modeled arrival-rate uncertainty and its effects (Robbins et al. 2006, Steckley et al. 2009). Statistical models in Whitt (1999), Avramidis et al. (2004), Brown et al. (2005), Weinberg et al. (2007), Shen and Huang (2008), Aldor-Noiman et al. (2009), Taylor (2012), Ibrahim et al. (2012), Ibrahim and L'Ecuyer (2013), Kim and Whitt (2014), Oreshkin et al. (2014), and others have sought to better characterize the distribution of arrival rates, by time of day, as they evolve.

Operations management papers account for uncertainty when making staffing and scheduling decisions. Maman (2009) extends many-server heavy-traffic limits to explicitly account for an arrival-rate distribution. Papers by Harrison and Zeevi (2005), Bassamboo et al. (2005), Bassamboo et al. (2006), Bassamboo and Zeevi (2009), Bertsimas and Doan (2010), Gurvich et al. (2010), Zan et al. (2013), and Ding and Koole (2014) have used stochastic programming (Birge and Louveaux 1997) and related approaches to account for arrival-rate uncertainty when making short-run staffing and call-routing decisions. Jouini et al. (2010) propose an alternative on-line formulation that does not use arrival rate information at all. More recent papers, such as Robbins et al. (2010), Robbins and Harrison (2010), and Liao et al. (2012) extend the stochastic programming framework to employee scheduling, and Mehrotra et al. (2010) uses mid-day recourse actions to adjust pre-scheduled staffing levels in reaction to realized deviations from arrival-rate forecasts.

While each of these streams of research has made progress in addressing elements of the problems caused by arrival-rate uncertainty, none addresses the whole problem. Statistical papers dedicated to forecasting have used traditional measures of fit for realized arrival counts to assess forecast quality. They have not, however, considered the downstream cost and quality of service (QoS) implications of arrival-rate forecast errors. Furthermore, the high dimensionality of within day arrival-rate profiles makes many of the more complex methods difficult to use in the context of stochastic scheduling algorithms. While operations management papers have looked carefully at the cost and QoS implications of stochastic scheduling methods, they have not considered how to integrate sophisticated statistical forecasting methods to better capture the nature of arrival-rate uncertainty. In turn, their measures of cost and QoS improvements may not accurately reflect the gains that can be made when better forecasting and scheduling methods are used in concert.

In this paper we integrate these statistical and operations-management approaches, marrying the use of

relatively sophisticated forecasting methods with stochastic programming formulations of call-center staffing and scheduling problems. Our work is data driven, and we use two large sets of call-center data to evaluate the elements of our approach. A first set of empirical results shows that parametric forecasting methods can be used to more efficiently solve stochastic scheduling problems that have traditionally been solved using sampling-based scenarios. A second set of empirical tests highlights the complementary roles that the use of scenario-based stochastic programming and the use of recourse actions can play in addressing arrival-rate uncertainty. More specifically, we make the following contributions.

In §2 we develop low-dimensional, parametric arrival-rate forecasts and use Gaussian quadrature to transform their continuous distributions into discrete scenarios (Miller and Rice 1983). We apply this scenario generation scheme to demonstrate that only a small number of scenarios is needed to capture the bulk of arrival-rate uncertainty in simple, one-stage stochastic workforce-scheduling problems.

Because scenarios are based on arrival rates, while updates are based on the realization of arrival counts rather than rates, the development of forecasts suitable for two-stage stochastic programs with recourse is not trivial. Nevertheless, in §3 we are able to use our parametric approach as the basis of such a procedure.

We first develop a Bayesian procedure that uses realized arrival counts in the early stage of the planning horizon to derive an *ex post* update of the forecast distribution for arrival rates during the later stage. In turn, the forecast update becomes an input to a second stochastic program that revises the initial staffing plan in a manner that lowers costs while maintaining QoS performance.

While the *ex post* updating scheme above allows managers to react to forecast errors, it does not affect initial schedules that are constructed without regard for the potential for later schedule revisions. To account for this possibility, we use our problem's Gaussian structure to develop two-stage forecasts that include *a priori* anticipation of potential forecast updates. These forecasts, in turn, drive two-stage stochastic programs whose initial schedules are able to further lower costs by exploiting the relative expense of anticipated recourse actions.

In Section 4 we test these joint forecasting-scheduling schemes using two sets of call center data. In both sets of tests we find that the use of multiple scenarios helps to stabilize system performance and leads to average abandonment rates that match *a priori* targets, while the use of recourse actions helps to systematically lower average costs.

More broadly, our work provides a general scheme for the formulation and solution of joint forecasting-scheduling models that support workforce management, a framework within which previous, related work represents special cases. Our results show that effective solutions to these problems can take the form of relatively simple, Gaussian arrival-rate forecasts married with computationally tractable stochastic programs with recourse. This approach explicitly accounts for arrival-rate uncertainty, as well as the ability to change scheduling decisions in response to forecast updates.

2 Parametric Forecasts for Stochastic Programming

In this section, we formulate and solve simple, single-stage arrival-rate forecasting and workforce scheduling models for call centers. While the scheduling model is nearly identical to that in Robbins and Harrison (2010), our approach to scenario generation uses Gaussian quadrature to discretize continuous forecast distributions and differs from the sampling schemes that are common to that paper, Bertsimas and Doan (2010), and others. This section's empirical tests show that, in fact, our Gaussian quadrature approach can greatly reduce the numbers of scenarios needed to stably solve these stochastic programs.

We also show that, for measures of QoS that are convex in the number of agents staffed during a given interval, we can further reduce the nonlinear constraints that model system performance across multiple scenarios into a single set of piecewise linear constraints. The transformation does not rely on the method by which scenarios have been generated – through discretization or via sampling – and allows us to efficiently perform tests that compare the two methods on problems with large numbers of scenarios. The ability to work with few scenarios, as well as to collapse large numbers of scenarios, when warranted, becomes particularly important when solving the two-stage stochastic programs with recourse that we analyze in §3.

2.1 Parametric Forecast

Our historical data comprise a $D \times I$ matrix of arrival counts, $\gamma = (\gamma_{di})$, where $d \in \mathcal{D} = \{1, \ldots, D\}$ indexes days and $i \in \mathcal{I} = \{1, \ldots, I\}$ indexes (30-minute) intervals within each day. We refer to the dth row of γ, denoted as $\gamma_d = (\gamma_{d1}, \ldots, \gamma_{dI})$, as the *intraday call volume profile* of the dth day.

We model the distribution of each count, γ_{di}, as a Poisson random variable (RV), Υ_{di}, with an uncertain arrival-rate Λ_{di}. Denote $\mathbf{\Lambda}_d = (\Lambda_{d1}, \ldots, \Lambda_{dI})$ as the *dth intraday arrival-rate profile*. We are interested in forecasting $\mathbf{\Lambda}_{D+h}$, the intraday arrival rate profile for a future day D+h, where h is a positive integer. Because the underlying rate profiles are uncertain and unobservable, however, our forecasting model uses the count profiles $\{\gamma_1, \ldots, \gamma_D\}$ to form an I-dimensional time series that we use for forecasting.

We begin by using a square-root transformation to stabilize the variance of the count data and approximately normalize the observations. Together, these features improve forecast accuracy and make the transformed counts amenable for standard statistical modeling. The proof of the following proposition can be found in Brown et al. (2010).

Proposition 1 (Brown et al. 2010) *Suppose an RV Υ has a Poisson distribution with deterministic arrival rate λ. As $\lambda \to \infty$, $Y \equiv \sqrt{\Upsilon + 1/4}$ has a Gaussian distribution with mean $\sqrt{\lambda}$ and variance $1/4$.*

Thus, instead of directly modeling the call volumes γ_{di}, we build our forecasting model using the square

root of the call volumes. Such a square-root transformation has been used in the call center forecasting literature (Brown et al. 2005, Weinberg et al. 2007, Shen and Huang 2008). We then view the square-root-transformed count distributions, $Y_{di} \equiv \sqrt{\Upsilon_{di} + 1/4}$ ($d \in \mathcal{D}$, $i \in \mathcal{I}$), as driven by a hidden arrival-rate process:

$$\begin{aligned} \omega_d - \alpha_{l_d} &= \beta(\omega_{d-1} - \alpha_{l_{d-1}}) + \eta_d, \quad \eta_d \sim \mathrm{N}(0, \phi^2), \\ \vartheta_{l_d,i} &\geq 0, \quad \sum_{i=1}^{I} \vartheta_{l_d,i} = 1, \\ \sqrt{\Lambda_{di}} &= \omega_d \vartheta_{l_d,i}, \\ Y_{di} &= \sqrt{\Lambda_{di}} + \epsilon_{di}, \quad \epsilon_{di} \sim \mathrm{N}(0, \sigma^2). \end{aligned} \tag{1}$$

Here ω_d is the daily total arrival rate (on the square-root scale), l_d is day-of-the-week of day d, α_{l_d} is a daily (zero-centering) adjustment for the day of the week l_d, $\vartheta_{l_d,i}$ is the fraction of the daily arrival rate that falls over time interval i on day of the week l_d, and Λ_{di} is the arrival rate for interval i on day d, represented on the natural scale. From Proposition 1, we know that the square-root counts Y_{di} have (approximately) Gaussian distributions with mean $\sqrt{\Lambda_{di}}$ and variance $\sigma^2 \approx 1/4$.

Our forecasting model (1) can be understood as follows. First, on the square-root-transformed scale, the daily total rate (ω_d) follows an AR(1) (order-one autoregressive) time series model, adjusting for the day of the week (α_{l_d}). We assume that the AR(1) daily rate process $\{\omega_d\}$ is Gaussian, which implies that, for a given day, d, the within-day rates, $\{\omega_d \vartheta_{l_d,i} \,|\, i = 1, \ldots, I\}$, are Gaussian as well. Second, each day of the week has its own intraday arrival proportion profile, $(\vartheta_{l_d,1}, \ldots, \vartheta_{l_d,I})$, and the square-root arrival rates $\sqrt{\Lambda_{di}}$ are assumed to follow a multiplicative model. Third, arrival rates are hidden, and we observe only square-root scaled arrival counts, rather than rates. Remark 1 in Online Supplement A describes how the use of a daily total arrival rate, together with an intraday arrival-rate profile, reduces forecast dimensionality.

Thus, we model the arrival-rate process as a hidden, Gaussian, AR(1) process, with observed counts that reflect Gaussian measurement errors. This representation is commonly called a *Gaussian state space* model and can be estimated using maximum likelihood methods (Douc et al. 2011).

We use observed square-rooted call volumes, $\{y_{di}\}$, to generate simple parameter estimates that are unbiased and have variance that is only slightly greater than that of computationally more intensive maximum likelihood estimates (Brown et al. 2005),

$$\widehat{\omega}_d = \sum_i y_{di}, \quad \widehat{\alpha}_{l_d} = \frac{\sum_{\{d':l_{d'}=l_d\}} \sum_i y_{d'i}}{\#\{d' : l_{d'} = l_d\}}, \quad \widehat{\vartheta}_{l_d,i} = \frac{\sum_{\{d':l_{d'}=l_d\}} y_{d'i}}{\sum_{\{d':l_{d'}=l_d\}} \sum_i y_{d'i}}, \quad d = 1, \ldots, D; i = 1, \ldots, I. \tag{2}$$

We then estimate the autoregressive coefficient $\hat{\beta}$ via linear regression. Estimates for the two variance parameters $\hat{\sigma}^2$ and $\hat{\phi}^2$ can be obtained from the residual sum of squares.

Once Model (1) is estimated, we use it to obtain the h-day-ahead forecast distribution, from day D for day

D+h, which we denote as $\omega_{D,D+h}$ and characterized below in Proposition 2.

Proposition 2 *The distribution of $\omega_{D,D+h}$, $h \geq 1$, is normal with mean $\zeta_{D,D+h}$ and variance $\psi^2_{D,D+h}$:*

$$\zeta_{D,D+h} \equiv \widehat{\alpha}_{l_{D+h}} + \hat{\beta}^h(\widehat{\omega}_D - \widehat{\alpha}_{l_D}) \qquad \text{and} \qquad \psi^2_{D,D+h} \equiv \hat{\phi}^2 \sum_{d=0}^{h-1} \hat{\beta}^{2d}. \tag{3}$$

Thus, the forecast mean for day D+h is the same as the estimate for day D's normalized mean, $\widehat{\omega}_D - \widehat{\alpha}_{l_D}$, with an adjustment for h days of AR(1) innovations, $\hat{\beta}^h$, and the day of the week on day D+h, $\widehat{\alpha}_{l_{D+h}}$. The forecast variance reflects the estimated variance associated with h days of AR(1) innovations, $\hat{\phi}^2 \sum_{d=0}^{h-1} \hat{\beta}^{2d}$.

We define $\Lambda_{D,D+h,i}$ as the day-D forecast for $\Lambda_{D+h,i}$. Observing from (1) that $\Lambda_{D,D+h,i} = \left(\omega_{D,D+h}\vartheta_{l_{D+h},i}\right)^2$, the forecast distributions for the arrival rates in intervals $i \in \{1, \ldots, I\}$ of day D+h follow easily from $\omega_{D,D+h}$.

2.2 Gaussian Quadrature for Scenario Generation

To simplify notation, we now drop the day subscript in ω_d, $\vartheta_{l_d,i}$ and $\Lambda_{d,i}$, and consider an arbitrary day. The uncertain arrival rate during the ith time period satisfies

$$\Lambda_i = (\omega\vartheta_i)^2 \tag{4}$$

where ω has a (forecast) distribution that is Gaussian with mean ζ and variance ψ^2. Note that, if we are at day D and are forecasting h days ahead, the mean and variance should be replaced by $\zeta_{D,D+h}$ and $\psi^2_{D,D+h}$ in (3).

To account for the uncertainty of Λ_i, recent papers have used stochastic programs with scenarios generated via random sampling from the forecast distribution (Bertsimas and Doan 2010, Robbins et al. 2010, Gurvich et al. 2010, Robbins and Harrison 2010). It is well known, however, that a large number of scenarios may be needed for the sampling approach to be effective (Shapiro and Philpott 2007, §2.2).

In this paper, we use Gaussian quadrature to derive a discrete approximation, ω^*, for ω, where $\omega^* = \omega_k$ with probability p_k, $k \in \mathcal{K} = \{1, \ldots, K\}$, and ω^* and ω have identical first $2K-1$ moments. Given these ω_ks and p_ks, the relation (4) naturally leads to the discrete approximation of Λ_i as

$$\Lambda^*_i \;=\; \lambda_{ik} \;\equiv\; \omega_k^2\vartheta_i^2 \qquad \text{with probability } p_k,$$

for $k \in \mathcal{K}$. Details of the discretization procedure can be found in Miller and Rice (1983). The specific expressions for our discretized Gaussian forecast distribution are provided in Online Supplement B.

Observe that (1) and (4) imply that, for all time periods within a given day, arrival-rate uncertainty is essentially only driven by the one-dimensional random scaling factor ω. Thus, we are able to use Gaussian quadrature efficiently for a one-dimensional distribution and avoid the usual "curse of dimensionality" problem.

2.3 Stochastic Programming Formulation

In the retail banking setting from which we have collected data, there is no explicit customer waiting time or abandonment cost. Rather, these types of call centers often minimize staffing costs, subject to explicit QoS constraints. In this paper we impose a 3% limit on the expected fraction of incoming calls that abandon before service, a QoS limit that can be regularly attained in larger, well-run call centers.

Let T be the length of the planning horizon, which may range from one day to several weeks. In the context of the mathematical programs we solve in this section, we fix T to be one day (i.e. $T = I$), but in practice it can exceed one day. As before, $\mathcal{I} = \{1, \ldots, I\}$ is a set of equally-divided subintervals within a day.

Let $\mathcal{J} = \{1, \ldots, J\}$ be the set of all feasible work schedules, each of which dictates which intervals within the planning horizon an agent answers calls. For interval $i \in \mathcal{I}$ and schedule $j \in \mathcal{J}$,

$$a_{ij} = \begin{cases} 1, & \text{if schedule } j \in \mathcal{J} \text{ has an agent answer calls during interval } i \in \mathcal{I}, \text{ and} \\ 0, & \text{otherwise.} \end{cases}$$

We let c_j be the cost of assigning an agent to schedule j. Costs include hourly wages and overtime pay, if the schedule requires it. Depending on the setting, costs may or may not include a prorated share of benefit payments. The principal decision variables are the numbers of agents to assign to the various schedules: $\{x_j \mid j \in \mathcal{J}\}$.

Recall that λ_{ik} is the arrival-rate forecast during interval $i \in \mathcal{I}$ under scenario $k \in \mathcal{K}$, which occurs with probability p_k. Observe that, when $K = 1$, then $p_k = 1$, and the stochastic program with one scenario collapses to become a traditional, deterministic workforce-scheduling integer program (IP). For *i.i.d.* scenarios based on sampling, each scenario, k, occurs with equal probability $p_k = 1/K$ and is determined by sampling ω and then using the relation (4) to determine the λ_{ik}s.

In any given interval, i, and scenario, k, our stochastic program determines the quality of service experienced by arriving customers by using a stationary measure of performance from standard queueing models. In particular, in this paper we track customer abandonment as the measure of QoS and use results from Mandelbaum and Zeltyn (2007) that characterize the stationary behavior of the M/M/n+M (Erlang-A) model. Given Poisson arrivals of constant rate λ, *i.i.d.* exponentially distributed service times with mean $1/\mu$, *i.i.d.* exponentially distributed times until customer abandonment (sometimes called patience) with mean $1/\theta$, and n servers, the paper provides explicit expressions for the calculation of the fraction of arriving customers that abandon before being served, $f(\lambda, \mu, \theta, n)$. We include these expressions in Online Supplement C.

In our context, the arrival rate during period i under scenario k is λ_{ik}, and the number of agents on hand during interval i is $n_i = \sum_{j \in \mathcal{J}} a_{ij} x_j$. Together with μ and θ they determine the expected number of customers abandoning during period i under scenario k, $\lambda_{ik} f(\lambda_{ik}, \mu, \theta, n_i)$. Green et al. (2001) call this the stationary

independent period-by period (SIPP) approach, and Remark 2 in Online Supplement A discusses its implicit assumptions.

Let α^* be an upper bound on the expected abandonment rate over the planning horizon. Then we wish to solve the following nonlinear stochastic integer program, which minimizes total staffing cost, subject to a constraint on expected abandonments.

$$\begin{array}{lllll} & \min \sum_{j\in\mathcal{J}} c_j x_j & & & \\ s.t. & \sum_{k\in\mathcal{K}} p_k \lambda_{ik} f(\lambda_{ik}, \mu, \theta, \sum_{j\in\mathcal{J}} a_{ij} x_j) & = & \alpha_i & i \in \mathcal{I} \\ & \sum_{i\in\mathcal{I}} \alpha_i & \leq & \alpha^* \bar{\lambda} & \\ & x_j & \in & \mathbb{Z}^+ & j \in \mathcal{J}, \end{array} \tag{5}$$

where $\bar{\lambda} = \sum_{k\in\mathcal{K}} p_k \sum_{i\in\mathcal{I}} \lambda_{ik}$ is the expected number of arrivals over the planning horizon, and $\mathbb{Z}^+$ is the set of non-negative integers. The first set of nonlinear constraints defines the expected number of abandoning calls for each interval, i. The second constraint defines the upper bound on the expected global abandonment rate.

We note that the upper bound, $\alpha^*\bar{\lambda}$, holds only in expectation, across the entire arrival-rate distribution. If we consider every potential abandonment to have the same (unknown) implicit cost, a Lagrangian relaxation of (5) would minimize expected total cost of staffing and abandonment. Because the cost of abandonment is unknown, however, call centers instead place direct constraints on expected QoS. Remarks 3 and 4 in Online Supplement A discuss the long-run average interpretation of the constraint, as well as a Lagrangian analogue.

The fact that $f(\lambda_{ik}, \mu, \theta, n_i)$ may be nonlinear in n_i makes the stochastic program (5) potentially difficult to solve. Nevertheless, Armony et al. (2009) show that, given $\mu \geq \theta$, $f(\lambda_{ik}, \mu, \theta, n_i)$ is nonincreasing in n_i, with decreasing differences (discretely convex). This is typically the case. For example, see Zohar et al. (2002) and Brown et al. (2005). In Section 2.4 we will see that the desired relationship also holds in our data.

Given $\mu \geq \theta$, we can use a common transformation to replace the nonlinear constraints with a larger set of linear constraints that provides a lower bound on $\sum_{k\in\mathcal{K}} p_k \lambda_{ik} f(\lambda_{ik}, \mu, \theta, n_i)$. For each $i \in \mathcal{I}$ and $n > 0$ we define slopes, m_{in}, and intercepts, b_{in},

$$\begin{array}{rcl} m_{in} & = & \sum_{k\in\mathcal{K}} p_k \left[\lambda_{ik}(f(\lambda_{ik}, \mu, \theta, n) - f(\lambda_{ik}, \mu, \theta, n-1))\right] \\ b_{in} & = & \sum_{k\in\mathcal{K}} p_k \lambda_{ik} f(\lambda_{ik}, \mu, \theta, n) - n \cdot m_{in}, \end{array} \tag{6}$$

where $m_{i0} = -\mu$ and $b_{i0} = \sum_{k\in\mathcal{K}} p_k \lambda_{ik}$. Then we replace each of the I constraints that define the α_is in (5) with a set of $\mathcal{N}_i = \{0, \ldots, N_i\}$ linear constraints,

$$(\textstyle\sum_{j\in\mathcal{J}} a_{ij} x_j) m_{in} + b_{in} \leq \alpha_i, \qquad n \in \mathcal{N}_i,$$

where N_i is large enough that the expected abandonment count is nearly zero: $\sum_{k\in\mathcal{K}} p_k \lambda_{ik} f(\lambda_{ik}, \mu, \theta, N_i) \approx 0$.

The revised linear, integer stochastic program becomes

$$\begin{array}{lllll} & \min \sum_{j\in\mathcal{J}} c_j x_j & & & \\ s.t. & (\sum_{j\in\mathcal{J}} a_{ij}x_j)m_{in} + b_{in} & \leq & \alpha_i & i\in\mathcal{I},\ n\in\mathcal{N}_i \\ & \sum_{i\in\mathcal{I}} \alpha_i & \leq & \alpha^*\bar{\lambda} & \\ & x_j & \in & \mathbb{Z}^+ & j\in\mathcal{J}. \end{array} \tag{7}$$

Essentially, we are replacing the abandonment rate function in (5) by the maximum of the linear functions defined by all the m_{in}s and the b_{in}s, which is piece-wise linear and convex, a standard practice in mathematical programming. In our case, this substitution is exact because the variables only take on integer values.

Thus, rather than solving the stochastic program (5), which is nonlinear and may have a large number of scenarios, we use the definitions (6) to develop a deterministic, piecewise linear, "certainty equivalent" program (7). Equation (6) suggests that the computational effort needed for the transformation is linear in the number of scenarios. Computational results reported in Online Supplement G.1 are consistent with this observation.

We note that, while it is straightforward to define analogues of (5) for other QoS measures, such as delay in queue, or to place limits on tail probabilities or require that QoS targets be met over sub-intervals within the planning horizon or within individual scenarios, schemes that use these alternatives are likely to be more computationally intensive. For example, in systems with abandonment, alternative QoS measures, such as $\mathrm{P}\{\text{Delay} > t\}$, often have a concave-convex structure. (It is often the case that the concave region of these functions can be roughly modeled as a simple threshold, for example a fixed staffing level below which $\mathrm{P}\{\text{Delay} > t\} = 1$. See Figures 4-8 in Garnett et al. (2002).) While concave-convex functions can be modelled to be piecewise-linear, we must introduce an additional 0-1 integer variable for each distinct concave segment within a scenario, and we cannot collapse scenarios into a single "certainty equivalent" as we did above. Nevertheless, computational results reported in Online Supplement G.2 suggest that only small numbers of scenarios are needed in the mathematical programs, and it may therefore be computationally practical to solve problems for concave-convex measures of QoS in much the same fashion. To the extent that large numbers of scenarios are needed, however, the L-shaped decomposition method (Birge and Louveaux 1997, §5.1) used in Robbins and Harrison (2010) could be of value.

2.4 Setup for Empirical Tests of Quadrature and Sampling-Based Scenarios

With the machinery developed in §2.1–2.3 at our disposal, we are now in the position to perform large-scale tests of the efficacy of our quadrature-driven scenarios. We run these tests using a dataset from a European retail bank's call center operations.

Our dataset consists of historical arrival counts, abandonment counts, and service-time averages for 30-

minute intervals on each of 176 weekdays in 2007. Online Supplement D provides summary plots of arrival counts over this horizon. The call center is open 13 hours each weekday, from 8 a.m. to 9 p.m. Hence, for each day we have 26 intervals of data, and we set the planning horizon to be $T = 26$ intervals, or one day.

We also have the rules and parameters that the bank's workforce management system uses to schedule agents. Agents work either 7 or 9-hour days, without overtime, and with specific rules for meal breaks: a lunch break lasts half an hour and must occur between 11 a.m. and 2 p.m.; a late lunch break also lasts half an hour and must occur between 4:30 p.m. and 6 p.m. An agent qualifies for either of the two breaks if her/his shift contains a half-hour period within that time window. If her/his shift contains a half-hour period within each time window, then s/he qualifies for both breaks. Enumeration shows that there are $J = 262$ feasible schedules.

The bank did not share payroll information with us. The fact that it used no overtime in constructing schedules motivates us to use a normalized cost of 1 for each half hour of work, and we let $c_j = \sum_{i \in \mathcal{I}} a_{ij}$.

We apply the approach described in §2.1 to forecast future arrival rates. Each forecast uses the previous 100 days of arrival counts to forecast the next day's rates. Therefore we have 76 days (days 101 to 176) of out-of-sample forecasts that we use to develop scenarios and run stochastic programs.

We use these parametric forecasts as the basis for quadrature and sampling-based scenario-generation schemes. For the former, we follow the procedure detailed in Online Supplement B. For the latter, we sample ω from its normal distribution, once for each scenario, and apply (4).

Other data used in the stochastic program include the following. The service times in our dataset average 121 seconds, so we set the service rate to be $\mu = 1800/121 \approx 14.6$ services per agent per 30-minute interval. To estimate the abandonment rate, we divide the total number of abandoned calls in our dataset by the total waiting time in queue for all served calls. This provides an upper bound on the abandonment rate of $\theta = 3.93$ calls per 30-minute interval, or equivalently, a lower bound on average caller patience of $1800/3.93 \approx 458$ seconds. (For an explanation why this ratio is an upper bound, see Remark 5 in Online Supplement A.) This implies that, on average, customers are willing to wait (at least) about 3.7 service times before abandoning the queue. The data clearly satisfy the requirement that $\mu \geq \theta$. All of the mathematical programs use an expected daily abandonment rate target of $\alpha^* = 3\%$. In most cases, the QoS constraint in the optimal solution is tight or nearly tight.

2.5 Empirical Comparison of Quadrature and Sampling-Based Scenario Generation

We compare the results of stochastic programs that use scenarios generated using Gaussian quadrature to those that use scenarios based on sampling of the forecast distribution. Our results suggest that the quadrature-based approach is more efficient and stable, in that very few scenarios are needed to obtain performance that is reliable

with respect to our two performance measures: cost and abandonment rate. Our comparison has three parts.

2.5.1 Distributions of Total Cost for One Day

We begin with a detailed look at the two methods' performance for a single out-of-sample forecast day. Using data from days 1 through 100 we generate a normally-distributed forecast for ω for day 101 and then use the forecast to create 909 stochastic programs that we solve. We generate scenarios for 9 of the stochastic programs – with 1, 4, 9, 16, 25, 49, 100, 225, and 400 scenarios – using Gaussian quadrature. For each of these 9 stochastic programs, we also create 100 analogous sets of scenarios for day 101, each set using an appropriate number of *i.i.d.* samples from the arrival-rate forecast. Thus we generate, run, and evaluate 101 instances – 100 realizations of the sampling approach, plus one quadrature-based instance – of 9 stochastic programs for a total of 909. For each of these instances, we record two statistics: the stochastic program's objective function value, which equals the total cost of staffing day 101, and the computation time required to set up and solve the IP. We report cost results below and computation times in Online Supplement G.1.

Figure 1 summarizes the total costs (the objective function value) of the 909 solutions. Results are grouped by number of scenarios along the horizonal axis, from 1 up to 400. For each number of scenarios, the vertical axis reports total cost. Each box of the box-and-whisker plot shows the 25th, 50th, and 75th percentiles of the cost of the 100 sampling-based instances of the problem for that number of scenarios. The whiskers are 1.5 times the interquartile range (75th percentile - 25th percentile) and are used as thresholds for determining outliers. The circles above and below the whiskers are the outliers. The dashed lines running across the whiskers display the 2.5% and 97.5% percentiles of the sampling-based results, and the solid line running across the plot's boxes shows the cost of the analogous quadrature-based program.

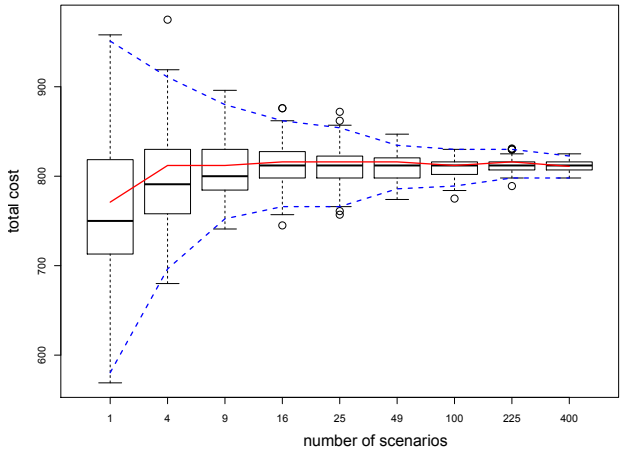

Figure 1: Day 101 Costs, by Number of Scenarios

Several features of Figure 1 are of note. As expected, as the number of scenarios increases, the distribution of results for sampling-based instances becomes less dispersed. Similarly, it is not surprising that, as the number of scenarios increases, the average cost of sampling-based programs and the cost of quadrature-based programs, are (generally) non-decreasing. Formulations with fewer scenarios display a well-known, systematic downward bias that results from solving a convex minimization problem with stochastic data (Shapiro 2000, §5.2). Of more interest to us is the fact that the cost of quadrature-based solutions remains nearly constant from 4 to 400 scenarios, and for instances with 16 scenarios or more, the cost is nearly identical to the median costs of the sampling-based programs.

2.5.2 First Differences in Total Cost for Each of 76 Days

Our second set of tests compares the results of all 76 out-of-sample days. In these tests, we formulate and solve 1,368 mathematical programs: 9 stochastic programs – with 1, 4, 9, 16, 25, 49, 100, 225, and 400 scenarios – for each of the two scenario-generation schemes on each of the 76 days. For each of 18 of the optimal solutions found on a given day, we record the total cost.

To develop a consistent measure of cost performance across all 76 days, we then use first differences. Given the relative stability of results for 400 scenarios, we use the 400-scenario results as benchmarks against which we compare those of formulations with fewer scenarios. For each type of scenario-generation scheme, on each day, we record the first difference between the performance of schemes with 1 through 225 scenarios and that with 400 scenarios.

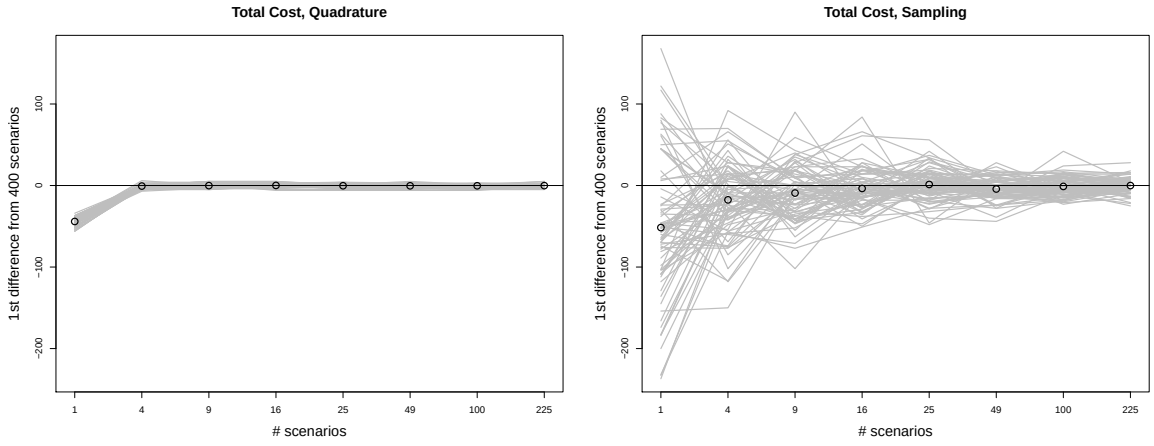

Figure 2: First Differences in Cost, by Number of Scenarios

Figure 2 plots the first differences of the total costs, by number of scenarios. The left panel plots the differences for mathematical programs with quadrature-based scenarios, and the right panel plots the differences for those using sampling-based scenarios. Each of the 76 lines in a panel plots the first differences for one day, by number of scenarios. The black circles show the means of the first differences across all 76 days.

Figure 2's results again suggest that, for quadrature-based formulations with more than one scenario, the means of the first differences are all quite close to zero; there is no apparent bias introduced by using fewer scenarios, as long as there are more than one. Formulations using one scenario, again, display the expected bias toward understaffing that was noted in Figure 1. Visual inspection also suggests that the results for sampling-based formulations are systematically noisier than those using quadrature-generated scenarios.

2.5.3 Abandonment Rates Across 76 Days

Our use of the objective function value as a measure of solution stability is crude: two optimal solutions, $x \neq y$, can have $\sum_{j \in \mathcal{J}} c_j x_j = \sum_{j \in \mathcal{J}} c_j y_j$. Nevertheless, Figures 1 and 2 both show that, even with this rough measure of stability, sampling-based solutions appear to be relatively unstable for smaller numbers of scenarios.

We would similarly like to rule out the possibility that the apparent stability of quadrature-based solutions is overstated by the use of objective function values. While one alternative would be to measure the difference between pairs of optimal solutions using an L^2 (or some other) norm, the possibility that an IP such as (7) has multiple optimal solutions motivates us to look instead at differences in the abandonment rates across solutions.

For our third test, we therefore consider abandonment rates for the 1,368 problem instances tested in §2.5.2, each of 18 problem instances on all 76 out-of-sample days. For each day we use each stochastic program's optimal solution to determine each half hour's staffing level. We then generate a single sample path of arrival times, virtual service times, and virtual times to abandonment that we use to drive a common set of discrete event simulations for that day. Results of the discrete event simulations allow us to calculate a sample realization of the abandonment rate for that day. More specifically, the numbers of agents on hand each half hour is determined by the optimal solution to the IP (7). For each of the 76 out-of-sample days we simulate one sample path of a 13-hour day. The simulated arrival process is driven by our dataset's 30-minute arrival counts. Because we do not have the arrival time of each call in the dataset, we turn each 30-minute count into the realization of a Poisson arrival process by distributing the counted number of arrivals as *i.i.d.* samples, each of which has an arrival time that is uniformly distributed over the half-hour. (See §2.3 in Ross (1996).) For each arrival, we also sample a virtual service time and virtual patience time, each of which is exponentially distributed with mean $1/\mu$ and $1/\theta$, respectively.

Within each simulation run, the number of agents on hand may decrease from one half hour interval to the next. In such cases, we remove agents with shortest remaining service time first. (Idle agents have zero remaining service times.) At the end of the simulated day, we also check to make sure that the number of unserved calls left in queue is not large enough to bias abandonment-rate statistics, which are calculated as fractions of arriving calls. Across 1,368 test simulations, the average number of calls left in queue at the end of

the day was 1.21, out of a total of 8,435.79 average daily arrivals. The maximum across all simulations was 15, on a day with 8,631 arrivals.

Figure 3 reports the abandonment-rate performance of the two discretization approaches: the left panel for quadrature and the right panel for sampling. In both panels, the horizonal axis shows the number of scenarios for a given test, and the vertical axis shows the realized abandonment rate. For each number of scenarios, each panel plots the 95% confidence interval (CI) associated with the estimate of the long-run average abandonment rate obtained over 76 days' performance. The CIs are constructed assuming each day's abandonments are independent of the others' and weighting each day's abandonment performance by that day's number of arrivals.

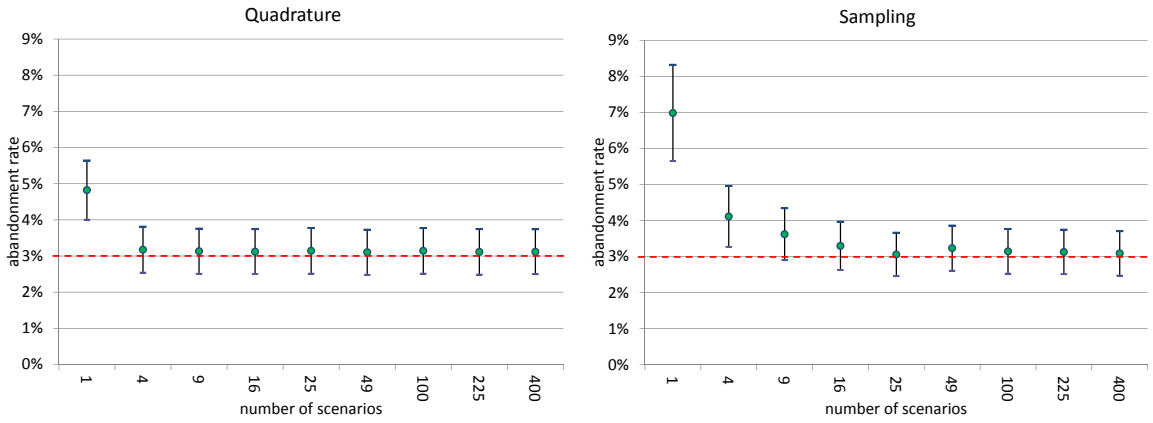

Figure 3: 95% CIs for Abandonment Rates, by Number of Scenarios

Figure 3's results are analogous to those shown in Figure 2. The left panel shows that, for quadrature, the entire CI for a single scenario falls above the 3% target, while those for 4 or more scenarios are quite stable and centered on 3%. The right panel shows that, with sampling, it takes 25 or more scenarios for the CIs to fall squarely over the 3% target. Again, the quadrature scheme appears to provide stable performance in the intended range using very few scenarios.

3 Stochastic Programs with Recourse

In the previous section, we demonstrated the effectiveness of using Gaussian AR(1) forecasts and quadrature to generate scenarios for simple, one-stage stochastic programs. In this section, we build on the machinery developed in §2 to develop schemes that, part-way through the planning horizon for agent schedules, use observed arrival-count data to revise the arrival-rate forecast and adjust schedules accordingly. These types of forecast updates and recourse procedures are commonly used to adjust staffing levels part-way through each day (Mehrotra et al. 2010).

For example, if a forecast update suggests higher-than-anticipated arrival rates, then additional agent capac-

ity may be requested for the second part of the horizon. Often this additional capacity comes in the form of extended hours for existing agents, the addition of short-term call-in agents, or the use of outsourcing capacity. Conversely, if a forecast revision suggests lower-then-anticipated arrival rates, then capacity may be reduced during the second part of the horizon, typically by encouraging agents to take unpaid time off. The use of forecast updates, combined with these so-called "recourse" actions, adds complexity to both the forecasting and scheduling approaches described in §2.

We consider two forms of updates. In §3.1, we describe a less sophisticated scheme that determines initial staffing levels using the one-stage approach of §2. Part-way through the horizon, this method then utilizes realized arrival counts to derive an *ex post* update of the arrival-rate forecast and solves a related one-stage stochastic program to determine recourse actions that adjust staffing levels for the remaining time intervals. In §3.2 we detail a more complex scheduling method. This second scheme follows the same overall approach as the simpler one, but it uses a more sophisticated, two-stage recourse program in the initial planning phase. In setting initial staffing levels, the two-stage program uses an *a priori* forecast of potential *ex post* forecast updates to explicitly account for the later use of recourse actions.

3.1 Simple Forecast Updates and Recourse Actions

The simple update scheme begins by forecasting and scheduling using the approach developed in §2, without considering the fact that recourse actions subsequently can be taken part-way through the planning horizon. On day D, it generates an arrival-rate forecast for day D+h, discretizes the forecast distribution, and solves (7) to determine an initial set of agent schedules, x.

Then after some update interval $i^* \in \{1, \ldots, I-1\}$ of day D+h, the scheme uses arrival-count data obtained from the start of day D+1 through i^* of day D+h to develop a revised arrival-rate forecast for the remaining intervals, $\{i^*+1, \ldots, I\}$, of day D+h. To account for the partial day, we denote the *early stage* intervals of day D+h as $\mathcal{I}_e \equiv \{1, \ldots, i^*\}$ and the *late-stage* intervals as $\mathcal{I}_l \equiv \{i^*+1, \ldots, I\}$.

With the revised forecast for $i \in \mathcal{I}_l$ in hand, the scheme runs a new, one-stage stochastic program whose decision variables represent recourse actions that add or remove agent capacity from the initial schedule, x. Additional constraints limit recourse actions according to work rules and the numbers of available agents. The solution of the second mathematical program determines how employee schedules will change in response to the updated arrival-rate forecast.

Note that this simpler scheme is a generalization of that proposed by Mehrotra et al. (2010), in which the initial mathematical program and the one-stage recourse program are simple IPs that use only one scenario. The formulation below generalizes that scheme to use multiple scenarios.

3.1.1 Forecast Update

Our forecast update uses available count data to generate a normally distributed revision of the original, h-day-ahead forecast, $\omega_{D,D+h} \sim \mathrm{N}(\zeta_{D,D+h}, \psi^2_{D,D+h})$, that is generated according to Proposition 2. For days $D+1$ through $D+h-1$, we have full days of (square-root scaled) count data available, $\{\mathbf{y}_{D+1}, \ldots, \mathbf{y}_{D+h-1}\}$, where $\mathbf{y}_d$ denotes the vector of count data for day d. For day $D+h$, however, we only have count data over intervals $i \in \mathcal{I}_e$, and, in a slight abuse of notation, we let $D+h_e$ denote the "index" of early-stage data on day $D+h$ and let $\mathbf{y}_{D+h_e}$ denote the associated vector of count data.

To use the count data $\{\mathbf{y}_{D+1}, \ldots, \mathbf{y}_{D+h-1}, \mathbf{y}_{D+h_e}\}$ to update the arrival-rate forecast for the late intervals of day $D+h$, we perform a sequence of h one-day forecast updates. We begin with an initial one-day-ahead forecast from day D to $D+1$, $\omega_{D,D+1}$, that we derive using Proposition 2 with forecast horizon $h = 1$. We then use day $D+1$'s count data, y_{D+1}, to derive a posterior distribution of ω for day $D+1$, which we denote $\omega_{D+1,D+1}$. The following proposition provides a general statement of the posterior derived from such a one-day-ahead update. Its proof can be found in Online Supplement E.

Proposition 3 *Suppose the prior distribution for day d is $\omega_{d-1,d} \sim N(\zeta_{d-1,d}, \psi^2_{d-1,d})$, and we observe counts $\{y_{di} \mid i = 1, \ldots, \hat{\imath}\}$ on day d. Letting*

$$a_{d,\hat{\imath}} \equiv \sum_{i=1}^{\hat{\imath}} \vartheta_{l_d,i}\, y_{d,i} \qquad \textit{and} \qquad \nu_{d,\hat{\imath}} \equiv \sum_{i=1}^{\hat{\imath}} \vartheta^2_{l_d,i}, \tag{8}$$

the posterior distribution $\omega_{d_{\hat{\imath}},d}$ is normal with mean $\zeta_{d_{\hat{\imath}},d}$ and variance $\psi^2_{d_{\hat{\imath}},d}$:

$$\zeta_{d_{\hat{\imath}},d} = \frac{\psi^2_{d-1,d} a_{d,\hat{\imath}} + \sigma^2 \zeta_{d-1,d}}{\psi^2_{d-1,d} \nu_{d,\hat{\imath}} + \sigma^2}, \tag{9}$$

$$\psi^2_{d_{\hat{\imath}},d} = \frac{\sigma^2 \psi^2_{d-1,d}}{\psi^2_{d-1,d} \nu_{d,\hat{\imath}} + \sigma^2}. \tag{10}$$

When we calculate the posterior for the entire day d, we let $\hat{\imath} = I$, and we call the posterior $\omega_{d,d} \sim \mathrm{N}(\zeta_{d,d}, \psi^2_{d,d})$. When the calculation is based on data from the early stage of day d's planning horizon, we let $\hat{\imath} = i^*$ and call the posterior $\omega_{d_e,d} \sim \mathrm{N}(\zeta_{d_e,d}, \psi^2_{d_e,d})$.

The proposition is consistent with the well known behavior of Bayesian updates that use normally distributed conjugate pairs (DeGroot 1970). That is, given a normally distributed prior and normally distributed data, the Bayesian posterior is also normally distributed, with a mean that is a weighted average of the prior mean and the observed data and a variance that is independent of the values of the observed data but grows systematically smaller with the quantity of data observed.

Once we have the posterior $\omega_{D+1,D+1}$ we can use our Gaussian AR(1) forecasting model (1) with $h = 1$ to provide the next one-day-ahead forecast, $\omega_{D+1,D+2}$. The proposition below provides a general statement of a

one-day-ahead update for a generic day, d.

Proposition 4 *Given a posterior for day d,* $\omega_{d,d} \sim N(\zeta_{d,d}, \psi^2_{d,d})$*, the prior for day* $d+1$*,* $\omega_{d,d+1}$*, is*

$$\zeta_{d,d+1} = \alpha_{l_{d+1}} + \beta(\zeta_{d,d} - \alpha_{l_d}), \tag{11}$$

$$\psi^2_{d,d+1} = \beta^2\, \psi^2_{d,d} + \phi^2. \tag{12}$$

The proposition's expressions for $\zeta_{d,d+1}$ and $\psi^2_{d,d+1}$ follow immediately from (1).

Thus, starting with $\omega_{D,D+1}$ we can apply Propositions 3 and 4 sequentially (h-1) times, with $\hat{\imath} = I$, to generate $\omega_{D+h-1,D+h}$. At this point, we apply Proposition 3 one final time, with $\hat{\imath} = i^*$ to calculate $\omega_{D+h_e,D+h} \sim \mathrm{N}(\zeta_{D+h_e,D+h}, \psi^2_{D+h_e,D+h})$. The associated arrival rates, $\{\Lambda_{D+h_e,D+h,i} \,|\, i \in \mathcal{I}_l\}$, naturally follow in light of (4).

3.1.2 Scenario Generation for Simple Forecast Updates

As in §2.2, we drop the day subscripts in the updated arrival-rate forecasts, $\{\Lambda_{D+h_e,D+h,i} \mid i \in \mathcal{I}_l\}$ and call the updated arrival-rate forecast $\{\Lambda'_i \mid i \in \mathcal{I}_l\}$, and we use Gaussian quadrature to generate a new set of scenarios. The number of updated scenarios may (in theory) differ from that previously used, and we differentiate updated scenarios by labeling them $k \in \mathcal{K}' = \{1, \ldots, K'\}$, with probabilities $\{p'_k \mid k \in \mathcal{K}'\}$ and scenario-dependent rates $\{\lambda'_{ik} \mid i \in \mathcal{I}_l, k \in \mathcal{K}'\}$.

3.1.3 Recourse Program for Simple Updates

With the initial set of schedules and scenarios for the revised forecasts in hand, we solve a schedule-update problem. Its solution determines a set of recourse actions to be used to adjust staffing levels over late-stage intervals of the scheduling horizon.

For each schedule assignment, $j \in \mathcal{J}$, we define a set of feasible recourse actions, $\mathcal{G}_j = \{1, \ldots, G_j\}$. If $a_{ij} = 0$ for some $i \in \mathcal{I}_l$ then we may be able to extend schedule j to have an agent assigned to that schedule to work during interval i. Similarly, if $a_{ij} = 1$ for some $i \in \mathcal{I}_l$ then we may be able to reduce schedule j to have an agent assigned to that schedule idle during interval i. We therefore let

$$r_{ijg} = \begin{cases} +1, & \text{if recourse action } g \text{ extends schedule } j \text{ by having an agent work during interval } i; \\ -1, & \text{if recourse action } g \text{ reduces schedule } j \text{ by having an agent idle during interval } i; \\ 0, & \text{otherwise,} \end{cases}$$

for $i \in \mathcal{I}_l$, $j \in \mathcal{J}$ and $g \in \mathcal{G}_j$. If we define a dummy schedule J to have $a_{iJ} = 0$ for all i and $c_J = 0$, then we can represent the ability to use call-in agents or outsourcing capacity in a similar fashion. As with the a_{ij}s, feasible columns – $(r_{i^*+1,j,g}, \ldots, r_{T,j,g})^{\mathsf{T}}$ – are determined by company policy and employee work rules.

Each of the decision variables $\{z_{jg} \mid j \in \mathcal{J}, g \in \mathcal{G}_j\}$ denotes the number of agents who were initially on schedule j and are assigned recourse action g after the schedule update. For each z_{jg}, a coefficient d_{jg} defines the unit cost of taking the recourse action. Positive costs are associated with the addition of work hours, either through schedule extensions or the use of call-in capacity. Negative costs (savings) may results from the ability to reduce agents' working hours.

The mathematical program, below, is an analogue of that for the first-stage problem, (7), with piecewise-linear constraints providing lower bounds to expected abandonment rates across the K' scenarios. As in (7), each $\mathcal{N}_i'$ is the set of linear constraints used to bound the expected number of abandoning customers in interval $i \in \mathcal{I}_l$, with $\{(m_{in}, b_{in}) \mid n \in \mathcal{N}_i'\}$ defining the slopes and intercepts. Here the x_j's are numbers that were previously determined via (7).

Given a fixed initial schedule, x, the following mathematical program then finds a set of recourse actions, z, that minimize recourse costs, subjected to a revised QoS constraint.

$$
\begin{array}{lllll}
 & \min \sum_{j\in\mathcal{J}} \sum_{g\in\mathcal{G}_j} d_{jg} z_{jg} & & & \\
s.t. & (\sum_{j\in\mathcal{J}} a_{ij} x_j + \sum_{j\in\mathcal{J}} \sum_{g\in\mathcal{G}_j} r_{ijg} z_{jg}) m_{in} + b_{in} & \leq & \alpha_i & i \in \mathcal{I}_l;\ n \in \mathcal{N}_i' \\
 & \sum_{i\in\mathcal{I}_l} \alpha_i & \leq & \alpha' \bar{\lambda}' & \\
 & \sum_{g\in\mathcal{G}_j} z_{jg} & \leq & x_j & j \in \mathcal{J} \\
 & z_{jg} & \in & \mathbb{Z}^+ & j \in \mathcal{J},\ g \in \mathcal{G}_j.
\end{array}
\tag{13}
$$

Here, the first two constraints of (13) define the piecewise-linear lower bounds on expected abandonment rates and enforce a revised QoS limit, $\alpha'\bar{\lambda}'$, only over $\mathcal{I}_l$. Specifically, we let $\bar{\lambda}' = \sum_{i\in\mathcal{I}_l} \sum_{k\in\mathcal{K}'} p_k' \lambda_{ik}'$ be the revised expected arrival rate over the second part of the planning horizon of day D+h and let $\alpha' = \left[\sum_{k\in\mathcal{K}} p_k' \sum_{i\in\mathcal{I}_l} \lambda_{ik}' f(\lambda_{ik}', \mu, \theta, \sum_{j\in\mathcal{J}} a_{ij} x_j)\right] / \bar{\lambda}'$ denote the expectation of the late-stage abandonment rate that would have occurred had the original staffing plan, x, been maintained.

Our definition of α' ensures that, on a sample-path basis, the bound on the expected number of abandonments over the late part of the planning horizon remains the same with and without recourse actions. In turn, over the entire planning horizon, our simple recourse scheme will achieve the same expected QoS as the one-stage stochastic schedule, developed in §2, at a (weakly) lower cost. Thus, the recourse scheme offers a Pareto improvement of scheduling without recourse. Remark 6 in Online Supplement A discusses other possible definitions of α'.

To summarize, we operationalize the simple scheme in two stages. Before the start of the planning horizon, we forecast arrival rates as in §2 and solve (7) to determine an initial set of schedules, $\{x_j \mid j \in \mathcal{J}\}$, and the numbers of agents on hand in intervals $i \in \mathcal{I}_e$ of day $(D + h)$. At the end of interval i^* on day $(D + h)$ we then use the available arrival-count over days $\{D + 1, \ldots, D + h_e\}$ to update the arrival-rate forecast for

intervals $i \in \mathcal{I}_l$ in the later part of target day D+h. We feed the initial schedule, $\{x_j \mid j \in \mathcal{J}\}$, along with the updated forecast, into the recourse program (13), whose solution determines optimal schedule adjustments, $\{z_{jg} \mid j \in \mathcal{J}, g \in \mathcal{G}_j\}$, and in turn the numbers of agents on hand in intervals $i \in \mathcal{I}_l$.

3.2 Forecasting and Scheduling Using Two-Stage Recourse Programs

A more complex approach to forecast and schedule updating explicitly takes the ability to use recourse actions into account when determining the initial staffing plan. For example, if adding agent capacity after the update is more expensive than initially overstaffing and then sending agents home, then we may wish to set initial staffing levels to be higher than would have the simpler early-stage program, which does not account for the relative costs of later capacity increases and decreases. To support this approach, both the initial forecast (3) and the scheduling program (7) become more elaborate.

3.2.1 A Priori Distribution of Potential Forecasts Updates

The update procedure described in §3.1.1 makes an iterative set of one-day forecast updates – for day D+1, then day D+2, up through day D+h_e – ultimately generating a forecast update for the late-stage intervals of day D+h, $\omega_{D+h_e,D+h}$. In this section, we extend that approach so that our original, day-D forecast includes an entire distribution of potential posterior distributions for intervals $i \in \mathcal{I}_l$ on day D+h. We can think of each sample path of counts $\{\mathbf{y}_{D+1}, \ldots, \mathbf{y}_{D+h-1}, \mathbf{y}_{D+h_e}\}$ as being one sample point of a distribution of potential sample paths that we forecast as of day D, $\{\mathbf{Y}_{D,D+1}, \ldots, \mathbf{Y}_{D,D+h_e}\}$. This distribution of potential sample paths generates an associated distribution of posteriors, and we can use an analogous day-by-day updating scheme to iteratively derive the entire *a priori* distribution of potential posterior forecast distributions.

Two forms of insensitivity highlighted in Proposition 3 make a distributional generalization of our forecast-update procedure relatively straightforward to derive and implement. First, given a normally distributed prior distribution, ω, and normally distributed sample data, any given sample, $\mathbf{y}$, will generate a normally distributed posterior. Second, all posterior forecast distributions have the same variance, defined according to the repeated application of (12) and (10), and differ only in their means. Thus, the *a priori* distribution of posterior forecast distributions is one-dimensional and can be characterized as the distribution of the mean of the posterior.

In a distributional analogue to (8), we let $Y_{D,D+d,i}$ denote the day D forecast for the arrival count that is to occur on interval i of day D+d. Let $\hat{\imath}$ denote the interval for a *potential update* on day D+d. For the focal day, day D+h, it is the interval i^*; for the days before D+h, it is the end of day. Thus,

$$A_{D+d,\hat{\imath}} = \sum_{i=1}^{\hat{\imath}} \vartheta_{l_{D+d},i}\, Y_{D,D+d,i}, \tag{14}$$

where $\hat{\imath} = I$ for $d \in \{1 \ldots, h-1\}$ and $\hat{\imath} = i^*$ for $d = h$. After the dth day's potential update, we denote the posterior mean as $\zeta'_{D,D+d}$. As a parallel to (9), we then have

$$\zeta'_{D,D+d} = \frac{\psi^2_{D+d-1,D+d} A_{D+d,\hat{\imath}} \; + \; \sigma^2 \zeta_{D+d-1,D+d}}{\psi^2_{D+d-1,D+d} \nu_{D+d,\hat{\imath}} \; + \; \sigma^2}, \tag{15}$$

where $\nu_{D+d,\hat{\imath}}$ can be obtained from (8) and $\psi^2_{D+d-1,D+d}$ from (10).

The following proposition uses the above expressions to characterize the *a priori* distribution of the posterior mean. Its proof can be found in Online Supplement E.

Proposition 5 *The a priori distribution of the posterior mean, $\zeta'_{D,D+d}$, is normally distributed with mean $\Theta_{D,D+d}$ and variance $\Psi^2_{D,D+d}$,*

$$\Theta_{D,D+d} = \zeta_{D,D+d} \qquad \textit{and} \qquad \Psi^2_{D,D+d} \; = \; \frac{\psi^4_{D+d-1,D+d}(\nu^2_{D+d,\hat{\imath}} \psi^2_{D,D+d} + \sigma^2 \nu_{D+d,\hat{\imath}})}{(\psi^2_{D+d-1,D+d} \nu_{D+d,\hat{\imath}} + \sigma^2)^2}, \tag{16}$$

where $\zeta_{D,D+d}$ and $\psi^2_{D,D+d}$ are defined as in (3)*, and $\psi^2_{D+d-1,D+d}$ is calculated via the d-fold application of* (12) *and* (10)*.*

Three facts about the proposition are worth noting. First, the mean of $\zeta'_{D,D+d}$, $\Theta_{D,D+d}$, always equals the simple d-day-ahead forecast as of day D, $\zeta_{D,D+d}$. Given an unbiased day-D forecast, this equivalence is to be expected. Second, as is noted in Online Supplement E, the expression for $\Psi^2_{D,D+d}$ can be interpreted as a scaled version of the variance of $A_{D+d,\hat{\imath}}$, where the scale corresponds to $A_{D+d,\hat{\imath}}$'s weight in the calculation of the posterior mean (15). Third and most important is the fact that the use of d days of AR(1) data does not require the use of an d-stage recourse tree to generate the *a priori* variance of the posterior mean on day $D+d$.

Thus, the arrival rates' Gaussian structure allows us to generate a 1-dimensional posterior by using (3) at day D to derive $\psi^2_{D,D+1}$, repeatedly applying (10) and (12) to calculate $\psi^2_{D,D+h_e}$, and then applying Proposition 5 to find the mean and variance of the posterior mean. It significantly simplifies our construction of an h-day-ahead *a priori* forecast of the posterior mean, even for very large h.

3.2.2 Scenario Generation for Two-Stage Forecasts

As before, we divide the planning horizon on day D+h into two stages, an early stage, $\mathcal{I}_e$, and a late stage, $\mathcal{I}_l$. The early-stage intervals occur before the forecast update, and their staffing levels cannot be adjusted by using late-stage recourse actions. In contrast, staffing levels in intervals $i \in \mathcal{I}_l$ can be adjusted using recourse.

Because recourse actions are not available in early-stage intervals, the forecasting and scheduling problem for $i \in \mathcal{I}_e$ is the same as that in §2.2. We use the original forecast distribution for $\omega_{D,D+h} \sim \mathrm{N}(\zeta_{D,D+h}, \psi^2_{D,D+h})$, and we use quadrature to generate a set of E early-stage rate scenarios $e \in \mathcal{E} = \{1, ..., E\}$ with arrival rates $\lambda_{ie} \; \equiv \; \omega_e^2 \vartheta_i^2$, $i = 1, \ldots, i^*$ and associated probabilities $\{p_e \geq 0 \mid \sum_{e \in \mathcal{E}} p_e = 1\}$.

After interval i^* we can use recourse actions to modify staffing levels, and we construct a set of K scenarios $k \in \mathcal{K}$, each associated with a distinct set of recourse actions to be considered. These *recourse* scenarios discretize the *a priori* distribution of the posterior. Recall that the distribution of the posterior mean is normal with mean $\Theta_{D,D+h_e} = \zeta_{D,D+h}$ and variance $\Psi^2_{D,D+h_e}$, as calculated according to Proposition 5. We denote the K discretized values of the mean as $\{\zeta'_k \mid k \in \mathcal{K}\}$, and they occur with probabilities $\{p_k \geq 0 \mid \sum_{k\in\mathcal{K}} p_k = 1\}$. Thus, quadrature yields K posterior distributions, each normally distributed with a different mean, ζ'_k, but all with identical variance, $\psi^2_{D+h_e,D+h}$, derived according to the h-fold application of Propositions 3 and 4.

Finally, for each recourse scenario $k \in \mathcal{K}$, we then associate a distinct set of late-stage rate scenarios, $\mathcal{L}_k = \{1, \ldots, L_k\}$. More specifically, given scenario k's ζ'_k, the associated posterior distribution is $\omega_k \sim \mathrm{N}(\zeta'_k, \psi^2_{D+h_e,D+h})$. We again use quadrature to discretize this posterior distribution into L_k scenarios, $\{\omega_{kl} \mid l \in \mathcal{L}_k\}$, with probabilities $\{p_{kl} \geq 0 \mid \sum_{l\in\mathcal{L}_k} p_{kl} = 1\}$. For each k, we transform the resulting ω_{kl}s according to (4) to yield the late-stage arrival rates, $\{\lambda_{ikl} = (\omega_{kl}\vartheta_i)^2 \mid i \in \mathcal{I}_l,\ k \in \mathcal{K},\ l \in \mathcal{L}_k\}$.

3.2.3 Two-Stage Recourse Program

As in the simpler scheme, the two-stage recourse program includes decision variables, $\{x_j \mid j \in \mathcal{J}\}$, that represent initial scheduling decisions, implemented before the start of the planning horizon. It also determines possible late-horizon recourse decisions, which can vary by recourse scenario. Decision variables $\{z_{jgk} \mid j \in \mathcal{J}, g \in \mathcal{G}_j, k \in \mathcal{K}\}$ represent the full set of these recourse decisions. If recourse scenario k is realized, then recourse decisions $\{z_{jgk} \mid j \in \mathcal{J}, g \in \mathcal{G}_j\}$ are taken.

We now formulate the piecewise-linear version of the two-stage recourse program. As before, each $\mathcal{N}_i$ is a set of linear constraints that bounds the expected number of abandoning customers in intervals $i \in \mathcal{I}_e$, with $\{(m_{in}, b_{in}) \mid n \in \mathcal{N}_i\}$ defining slopes and intercepts. We similarly let $\mathcal{N}_{ik} = \{0, \ldots, N_{ik}\}$ define the analogous sets of constraints for each interval $i \in \mathcal{I}_l$ in the later part of the planning horizon and under each recourse scenario, $k \in \mathcal{K}$, with slopes and intercepts $\{(m_{ikn}, b_{ikn}) \mid n \in \mathcal{N}_{ik}\}$. In both cases, we eliminate the explicit representation of *rate* scenarios – those indexed e in early intervals $i \in \mathcal{I}_e$ and those indexed l in late intervals $i \in \mathcal{I}_l$ – by exploiting the fact that, for a fixed arrival rate, the expected number of abandonments in a given interval, i, is decreasing and convex in the number of agents. For the two-stage recourse formulation we do not eliminate the recourse scenarios $k \in \mathcal{K}$, however, precisely because we wish to track how each k's recourse actions affect the conditional expectation of the number of abandonments in each interval $i \in \mathcal{I}_l$.

We let $\{\alpha_i \mid i \in \mathcal{I}_e\}$ be the expected number of abandoning calls in each interval of the early part of the planning horizon. For each recourse scenario, $k \in \mathcal{K}$, we let $\{\alpha_{ik} \mid i \in \mathcal{I}_l\}$ be the analogous quantities in the second part of the planning horizon, conditioned on falling into recourse scenario k. We then let α^* be an upper

bound on the expected abandonment rate over the entire horizon and across all recourse scenarios.

The solution to the stochastic integer program, below, minimizes the expected cost of initial scheduling and recourse decisions, subject to an upper bound on the expected number of abandonments over the planning horizon, $\alpha^*\bar{\lambda}$, where

$$\bar{\lambda} \equiv \sum_{e\in\mathcal{E}}\sum_{i\in\mathcal{I}_e} p_e\lambda_{ie} + \sum_{k\in\mathcal{K}}\sum_{l\in\mathcal{L}_k}\sum_{i\in\mathcal{I}_l} p_k p_{kl}\lambda_{ikl}$$

is the expected number of arrivals over the planning horizon. The first set of constraints, below, provides a lower bound on the expected numbers of abandonments during each interval $i \in \mathcal{I}_e$, and the second set provides a similar set of bounds for each recourse scenario $k \in \mathcal{K}$ during $i \in \mathcal{I}_l$. The third constraint enforces an upper bound on the expected number of abandonments over the planning horizon, and the fourth set of constraints ensures that at most one recourse action may be taken for each employee.

$$\begin{array}{lllll}
 & \min \sum_{j\in\mathcal{J}} c_j x_j + \sum_{k\in\mathcal{K}} p_k \sum_{j\in\mathcal{J}} \sum_{g\in\mathcal{G}_j} d_{jg} z_{jgk} & & & \\
s.t. & (\sum_{j\in\mathcal{J}} a_{ij} x_j) m_{in} + b_{in} & \leq & \alpha_i & i \in \mathcal{I}_e, n \in \mathcal{N}_i \\
 & (\sum_{j\in\mathcal{J}} a_{ij} x_j + \sum_{j\in\mathcal{J}} \sum_{g\in\mathcal{G}_j} r_{ijg} z_{jgk}) m_{ikn} + b_{ikn} & \leq & \alpha_{ik} & i \in \mathcal{I}_l;\ k \in \mathcal{K};\ n \in \mathcal{N}_{ik} \\
 & \sum_{i\in\mathcal{I}_e} \alpha_i + \sum_{k\in\mathcal{K}} p_k \sum_{i\in\mathcal{I}_l} \alpha_{ik} & \leq & \alpha^*\bar{\lambda} & \\
 & \sum_{g\in\mathcal{G}_j} z_{jgk} & \leq & x_j, & j \in \mathcal{J},\ k \in \mathcal{K} \\
 & x_j & \in & \mathbb{Z}^+ & j \in \mathcal{J} \\
 & z_{jgk} & \in & \mathbb{Z}^+ & j \in \mathcal{J},\ g \in \mathcal{G}_j,\ k \in \mathcal{K}
\end{array} \tag{17}$$

Remark 7 in Online Supplement A discusses several extensions to this scheme.

3.2.4 Determination of Initial Schedules and Recourse Actions

We operationalize the two-stage scheme as follows. On day D we use the procedure defined in §3.2.1–§3.2.3 to find an optimal set of initial schedules and recourse actions, and for $i \in \mathcal{I}_e$ of day D+h, we use the optimal x to determine agent schedules in the early part of the planning horizon. Then after interval i^* of day D+h, we use the observed count data, $\{\mathbf{y}_{D+1}, \ldots, \mathbf{y}_{D+h_e}\}$, together with the *ex post* update procedure defined in §3.1.1, to generate $\zeta_{D+h_e,D+h}$, the mean of the realized posterior distribution of ω for day D+h. If $\zeta_{D+h_e,D+h}$ equals the mean of the posterior distribution for recourse scenario, $k \in \mathcal{K}$, then we update agents' schedules according to scenario k's z_{jgk}s. If $\zeta_{D+h_e,D+h}$ does not match one of the scenario's means exactly, then we implement the recourse actions associated with a scenario whose mean is either just above or just below $\zeta_{D+h_e,D+h}$.

As the number of recourse scenarios increases without bound, $K \to \infty$, the actual posterior mean always matches that of one of the recourse scenarios. For finite K, however, it will not match, and we must decide whether to choose the recourse scenario whose mean is closest to the posterior mean or to round the actual posterior up or down. We have found through experimentation that rounding up consistently provides expected

abandonment-rate performance, and in §4 we therefore round up to the next scenario. In the case that the realized posterior mean is above that of the highest scenario, we round down.

Remark 8 in Online Supplement A describes a more common alternative that solves (13) – using the *ex post* forecast update together with the optimal x and abandonment rates from (17) – to directly (re)identify a set of optimal recourse actions .

4 Numerical Tests of Nine Scheduling Schemes

We now have the machinery necessary to precisely define each of nine scheduling schemes we will evaluate. Schemes labeled SPm ($m = 1, 4, 100$) solve the stochastic program (7) with m scenarios to find an initial set of schedules x and perform no updating. Schemes denoted UPm ($m = 1, 4, 100$) solve (7) with m scenarios at the start of the planning horizon, update the initial arrival-rate forecast after interval i^* for the focal day, and then solve (13) with m scenarios to determine a set of recourse actions z to take in $\mathcal{I}_l$. At the start of the planning horizon, schemes labeled RPm ($m = 1, 4, 100$) solve (17) with $E = m$ early-interval rate scenarios, $K = \sqrt{m}$ recourse scenarios and, for each of the recourse scenarios, $k \in \mathcal{K}$, $L_k = \sqrt{m}$ late-interval posterior arrival-rate scenarios. Its solution generates an initial set of schedules, x, as well as K potential sets of recourse actions, $\{z_k \mid k \in \mathcal{K}\}$. After interval i^* of the focal day, the RPm schemes calculate the realized posterior mean for the arrival rate and then round to the nearest recourse scenario, k (as described in §3.2.4), to find the set of recourse actions, z_k, to implement in $i \in \mathcal{I}_l$. Thus, the RPm schemes use $E = m$ scenarios for intervals $i \in \mathcal{I}_e$, as well as $K \times L = m$ total scenarios for intervals $i \in \mathcal{I}_l$, numbers that are comparable to those in SPm and UPm.

We note that four of these schemes correspond to approaches found elsewhere: SP1 is a traditional IP, driven by a point forecast; SP4 and SP100 are quadrature-based versions of the approach used in Robbins and Harrison (2010) evaluated in §2; and UP1 is the simple forecast-update approach evaluated in Mehrotra et al. (2010).

To be consistent with practice, which generates initial agent schedules for days or weeks at a time, we construct forecasts and agent schedules one week at a time. For example, our first set of forecasts use data from days 1 to 100 to generate a set of 1-day-ahead forecasts for day 101, a set of 2-day-ahead forecasts for day 102, and so on through day 105. We use these forecasts in the mathematical programs (7) and (17) to determine five sets of initial agent schedules, one set for each day of the week. Then for each day $d \in \{101, \ldots, 105\}$ we use count data obtained from the start of day 101 through interval i^* of day d to derive d's posterior rate distribution. We run (13) to determine the recourse actions to be used in the UPm schemes, and we use the arrival rate's posterior mean to select the recourse scenario k that will be used for each of the RPm schemes. Similarly we use data from days 6 through 105 to derive initial schedules for days 106 through 110 and then use count data from days 106 though 110 to determine the appropriate recourse actions for each day of that week.

As in §2, for each day and each scheme we use optimal initial schedules and recourse actions to determine the scheme's staffing counts for each half-hour of the day. Then we generate a single, common sample path of arrival times, virtual service times, and virtual times to abandonment that we use in discrete event simulations to generate the numbers of realized abandonments for the schemes. From each day's optimal schedules, costs, arrival counts, and abandonment counts we calculate realized abandonment rates and costs per handled call.

In the following sections we test these schemes on two sets of data. The first set of tests uses the European retail bank data described in §2.4. A second set of tests uses arrival-count data from a network of call centers operated by a North American (NA) retail bank.

4.1 Empirical Results for the European Retail Bank

In addition to the data described in §2.4, our empirical tests for the European bank use the definition of recourse actions and costs that are used in (13) and (17). However, because the European bank's labor practices are highly restricted (compared with those in the US), it does not currently use recourse actions. Therefore, we use recourse actions and costs that are analogous to those used by Mehrotra et al. (2010) .

Specifically, we consider three sets of recourse actions. For any worker assigned to an initial schedule, $j \in \mathcal{J}$, we consider two generic actions: 1) the ability to extend the worker's shift beyond the time it would normally end; and 2) the ability to send the worker home early, before his or her schedule would normally end. We require schedule adjustments to be made over contiguous sets of intervals. For example, an agent who is originally scheduled to work until 5:00 p.m. and is asked to work from 6:30 p.m. to 7:00 p.m. must work from 5:00 p.m. to 6:30 p.m. as well. For shift extensions we assume a traditional overtime cost of 1.5 per agent per half hour, a 50% premium over the base rate of 1 per interval, and when sending someone home we define the cost to be -0.75. A third set of recourse actions is the ability to (outsource or) call in workers who are not scheduled to work on a given day. We assume that the cost of this action is 2 per agent per half hour; these agents receive double-time pay.

We choose a fixed forecast update interval after $i^* = 6$ at 11 a.m., just after what is typically the first peak in the day's arrival rate, and we use the observed arrival counts to revise the initial forecast and determine recourse actions. Given an i^* of 6 and other problem parameters – 26 half-hour intervals per day and 262 feasible initial schedules – the number of feasible recourse actions totals 4,973.

As in §2.5 we use 176 days of weekday arrival data from the European bank. For each of the last 76 days, we use the week-by-week forecasting, scheduling, and simulation procedures described above to generate daily abandonment rates and average costs per handled call. Online Supplement G.2 reports and discusses average computation times.

Figure 4 summarizes the results of all 76 days. The left panel plots 95% CIs for the 76 realized abandonment rates, and the right panel analogous intervals for the average of the 76 average costs per handled call. The intervals' point estimates are calculated as weighted averages of the 76 days' results, with the number of calls handled on a given day acting as weights. Similarly, the 1/2-widths of the CIs are calculated using standard deviations whose points are weighted by numbers of calls.

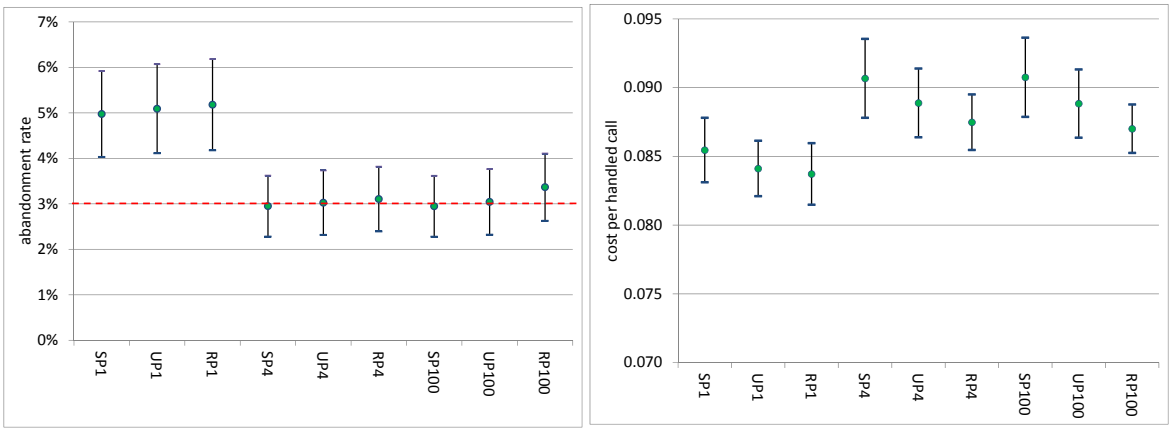

Figure 4: 95% CIs for European Bank Abandonment Rates (left) and Average Costs per Handled Call (right)

The left panel of Figure 4 shows that the six schemes that use multiple scenarios have average abandonment rates that are close to 3%, while one-scenario schemes that use point forecasts have significantly higher average abandonment rates. Conversely, the six schemes that use multiple scenarios have average costs that are significantly higher than those that use point forecasts. These results echo those shown in Figures 1 and 2.

The figure's right panel shows that there is a small but perceptible decline in average cost per handled call as one moves from SPm to UPm to RPm scheme for a given m. For example, the average cost for UP4 and RP4 are, respectively, 2.0% and 3.5% lower than that for SP4. Similarly, average costs for UP100 and RP100 are, respectively, 2.1% and 4.1% lower than that for SP100.

While the CIs' broad overlaps suggest that differences in average costs are not significant, we can take advantage of the fact that each day's forecasts, schedules, and simulation results generate 9 matched data points to compare results day by day across schemes. Specifically, we perform more sensitive (one-tailed) Wilcoxon Signed-Rank tests, non-parametric analogues of paired t-tests, to compare the nine schemes' performance. The results, detailed in Online Supplement F, show that in fact there are systematic differences. Average costs for UP4 are significantly lower than those for SP4, and average costs for RP4 are significantly lower than those for both SP4 and UP4. The same relationship holds for SP100, UP100, and RP100.

Conversely, Wilcoxon tests indicate that abandonment rates for UP4 and RP4 are not significantly higher than those for SP4, an indication that the cost improvements enabled by the schemes' recourse actions are not associated with declines in service quality. While the same holds true for UP100 versus SP100, the Wilcoxon

tests indicate that abandonments rates for RP100 are significantly greater those for SP100 and UP100, a fact that can also be seen in the left panel of Figure 4. As we will see in §4.2, the decline in RP100's QoS does not occur in our second set of tests.

If we compare each 4-scenario scheme's average costs and abandonment rates with those of its 100-scenario counterpart, we also find that the performance does not improve as we move from 4 to 100 scenarios. More specifically, the Wilcoxon tests show that abandonment rates and average costs are not significantly higher for SP4 than for SP100 nor for UP4 when compared to UP100. Abandonment rates are significantly higher for RP100 than for RP4, a reflection of the fact, seen in the right panel of Figure 4, that RP100 appears to be slightly understaffing.

The results have at least three important implications for the European bank's call center. First, schemes that are based on point estimates of arrival rates do not appear to provide adequate staffing to meet long-run QoS targets, while those that are based on distributional forecasts appear to meet QoS goals. Second, the cost advantage of using recourse actions plays out largely as predicted: the UPm schemes' *ex post* updates lower average costs a bit, while the RPm schemes' two-stage, *a priori* approach lowers costs a bit more. In general, the significance and magnitude of cost improvements will vary with details of a call center's cost and scheduling data, however. Third, the computationally less demanding 4-scenario schemes performed as well as the 100-scenario schemes – in the case of the RPm schemes one might say even better.

4.2 Empirical Results for a North American Retail Bank

As a check on the robustness of Section 4.1's results, we construct a second set of tests that use arrival-count data from another organization, a network call centers operated by a North American retail bank. This North American bank's call centers operate at a larger scale, with call volumes that are more than six and one half times those of the European call center studied in §4.1. Summary plots of the North American bank's arrival counts can be found in Online Supplement D.

We control these new tests so that their results are comparable to those from the European bank. In the new tests we continue to use the scheduling, service rate, individual abandonment rate, and cost parameters used in the original experiments. The only differences between the two sets of numerical tests are the historical half-hour arrival counts and the number of days we test.

We employ the same sampling and scheduling schemes as before, using the previous 100 days' arrival-count data to forecast arrival-rate profiles for upcoming sets of 5 days. We have 210 days of arrival-count data for the North American bank and therefore have 110 out-of-sample points that we test. As before, we determine staffing numbers and costs according to the nine scheduling schemes and then use a common sample path for

discrete event simulations that generate numbers of realized abandonments. We report computation times in Online Supplement G.2.

Figure 5 summarizes the results of the 110 out-of-sample days. The left panel plots 95% CIs for the realized abandonment rates, and the right panel analogous intervals for costs per handled call. As before, the intervals' point estimates and 1/2-widths are determined using weighted calculations, with weights that are numbers of calls handled on each day. To make the plots visually comparable to Figure 4's, we use the same vertical scales.

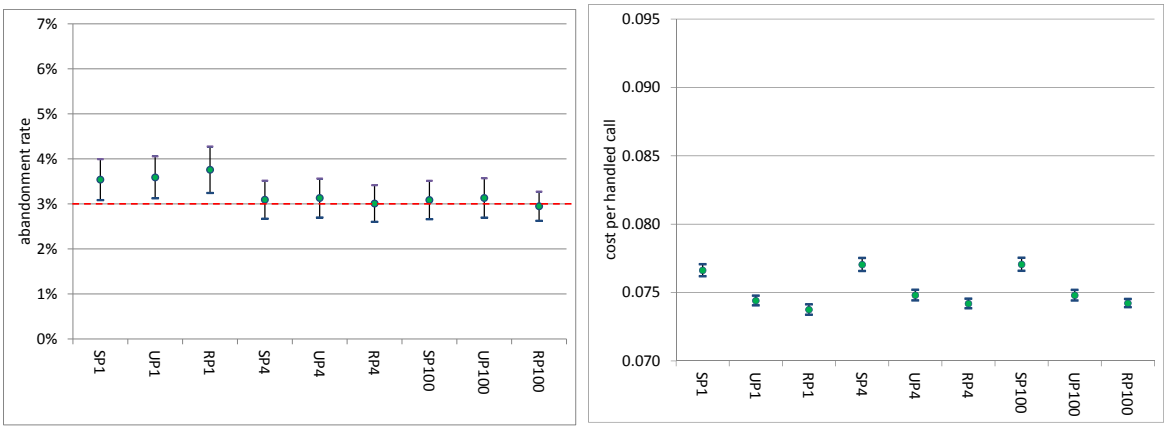

Figure 5: 95% CIs for NA Bank Abandonment Rates (left) and Average Costs per Handled Call (right)

The results are analogous to those of the European bank. As before, the schemes that use distributional forecasts do a good job of reaching a long-run abandonment target of 3%, while those that use point forecasts appear to have abandonment rates that are biased upward. The use of recourse actions also provides similar magnitude of cost savings: average costs for UP4 and UP100 are, respectively, 2.9% below those of SP4 and SP100, and RP4's and RP100's costs are both 3.7% below their SP4 and SP100 benchmarks.

In all cases, CIs for performance are much narrower, a reflection of systematically lower forecast variances for the North American data. The Wilcoxon tests detailed in Online Supplement F show that the schemes' results are quite well behaved. While UPm and RPm schemes again have significantly better cost performance than SPm schemes, they do not have significantly higher abandonment rates. Furthermore, within each set of schemes – SPm, UPm, and RPm – those with $m = 4$ perform as well as those with $m = 100$.

In addition to narrower CIs, we also see that the schemes' cost per handled call is, in most cases, about 15% lower in the North American data than in the European data. An interesting question, then, is the extent to which the reduction in forecast uncertainty and costs enjoyed by the North American bank is the result of an increase in scale. For example, an arrival process that is an aggregate of separate, independent arrival processes – such as that obtained by pooling across independent geographic areas – should enjoy a reduced CV of the overall arrival rate. This phenomenon would represent an as-yet unaccounted for source of economies of scale that warrants further investigation.

5 Conclusion

Our analysis has provided a number of insights into the value of stochastic programming and recourse for call-center workforce management. We used a parametric forecasting scheme to generate stochastic programs that needed only small numbers of scenarios, and our use of a convex measure of QoS allowed us to collapse scenarios in linear time to create efficient, deterministic, piecewise linear, certainty-equivalent versions of these stochastic programs. Together, the forecasting scheme and certainty-equivalent formulation allowed us to simply generate and solve large numbers of two-stage recourse programs.

Numerical tests of the forecasting and scheduling schemes showed how our use of multiple scenarios and of recourse actions provided complementary benefits. Multiple scenarios were needed to achieve long-term QoS goals, and recourse actions improved system costs. A comparison of the European bank's and North American banks's results also suggested that, with respect to forecast, the same variance-reducing pooling effect that is widely recognized in inventory systems might provide an as-yet unaccounted for source of economies of scale in call center operations. This mechanism is different from the traditional pooling of queueing systems.

The results so far are promising, and they suggest a number of directions for follow-on work that will strengthen both the theoretical underpinnings and the practical value of our framework. As a start, we are currently working to show formally that the scheme used in this paper is long-run average optimal (Gans et al. 2014). Other directions include the integration of more complex forecasting models and the use of other, non-convex measures of QoS.

Also of interest are the computational demands associated with longer planning horizons. Specifically, in the numerical tests in §4 we generated 5-day sets of one-day schedules. In practice agent schedules can span many days or weeks at a time, however, and in these settings the number of feasible schedules can explode. For example, given the flexibility to combine any of 262 feasible daily schedules across 5 days, SP would have 262^5 or more than 1.2 trillion possible weekly agent schedules. While there are other practical factors that may restrict the combinations of daily schedules, and while call centers could consider using the rostering process to knit together multi-day schedules from consecutive one-day schedules (that are identified as we have in §4), significant work will be required to develop computationally tractable scheduling procedures that identify optimal multi-day or multi-week agent schedules.

Acknowledgments

The authors gratefully acknowledge the helpful comments of the Associate Editor and Referees. This material is based upon work supported by the National Science Foundation under Grant Numbers CMMI-0645075, CMMI-0800575, CMMI-0800645, DMS-1106912, and DMS-1407655, as well as by the University of Hong Kong Stanley Ho Alumni Challenge Fund.

References

Z. Akşin, M. Armony, and V. Mehrotra. 2007. The Modern Call-Center: A Multi-Disciplinary Perspective on Operations Management Research. *Production and Operations Management*, 16(6):655-688.

S. Aldor-Noiman, P. D. Feigin, and A. Mandelbaum. 2009. Workload Forcasting for a Call Center: Methodology and a Case Study. *Annals of Applied Statistics*, 3(4):1403-1447.

M. Armony, E. Plambeck, and S. Seshadri. 2009. Sensitivity of Optimal Capacity to Customer Impatience in an Unobservable M/M/S Queue (Why You Shouldn't Shout at the DMV). *Manufacturing & Service Operations Management*, 11(1):19-32.

A. N. Avramidis, A. Deslauriers, and P. L'Ecuyer. 2004. Modeling Daily Arrivals to a Telephone Call Center. *Management Science*, 50:896-908.

A. Bassamboo, J. M. Harrison, and A. Zeevi. 2005. Dynamic Routing and Admission Control in High Volume Service Systems: Asymptotic Analysis via Multi-Scale Fluid Limits. *Queueing Systems Theory and Applications,* 51:249–285.

A. Bassamboo, J. M. Harrison and A. Zeevi. 2006. Design and Control of a Large Call Center: Asymptotic Analysis of an LP-Based Method. *Operations Research,* 54:419–435.

A. Bassamboo and A. Zeevi. 2009. On a Data-Driven Method for Staffing Large Call Centers. *Operations Research*, 57(3):714-726.

G. P. Bhattacharjee. 1970. Algorithm AS 32: The Incomplete Gamma Integral. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 19(3):285-287.

D. Bertsimas and X. Doan. 2010. Robust and Data-Driven Approaches to Call Centers. *European Journal of Operational Research*, 207(2):1072-1085.

J. R. Birge and F. Louveaux. 1997. *Introduction to Stochastic Programming*. New York: Springer.

L. D. Brown, T. Cai, R. Zhang, L. Zhao, and H. Zhou. 2010. The Root-Unroot Algorithm for Density Estimation as Implemented via Wavelet Block Thresholding. *Probability Theory and Related Fields*, 146:401-433.

L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. 2005. Statistical Analysis of a Telephone Call Center: a Queueing Science Perspective. *J. of the Am. Statistical Association*, 100:36-50.

B. Chen and S. G. Henderson. 2001. Two Issues in Setting Call Centre Staffing Levels. *Annals of Operations Research*, 108:175-192.

P. Coolen-Schrijner and E. A. Van Doorn. 2001. On the Convergence to Stationarity of Birth-Death Processes. *Journal of Applied Probability*, 38(3):696-706

M. H. DeGroot. 1970. *Optimal Statistical Decisions*. McGraw-Hill, New York.

S. Ding and G. Koole. 2014. Optimal Call Center Forecasting and Staffing under Arrival Rate Uncertainty. Working Paper, VU University.

R. Douc, E. Moulines, J. Olsson, and R. van Handel. 2011. Consistency of the Maximum Likelihood Estmator for General Hidden Markov Models. *Annals of Statistics*, 39(1):474-513.

Z. Feldman, A. Mandelbaum, W. A. Massey, and W. Whitt. 2007. Staffng of Time-Varying Queues to Achieve Time-Stable Performance. *Management Science*, 54:324-338.

N. Gans, G. Koole, and A. Mandelbaum. 2003. Telephone Call Centers: Tutorial, Review, and Research Prospects. *Manufacturing & Service Operations Management*, 5:79-141.

N. Gans, H. Shen, H. Ye, and Y.-P. Zhou. 2014. Long-run Average Optimality of AR(1)-Driven Workforce Scheduling Models. In preparation.

O. Garnett, A. Mandelbaum, and M. Reiman. 2002. Designing a Call Center with Impatient Customers. *Manufacturing & Service Operations Management*, 4:208-227.

W. K. Grassman. 1988. Finding the Right Number of Servers in Real-World Queuing Systems. *Interfaces*, 18(2):94-104.

L. V. Green, P. J. Kolesar, and J. Soares. 2001. Improving the SIPP Approach for Staffing Service Systems that Have Syclic Demands. , 49(4):549-564.

L. V. Green, P. J. Kolesar, and W. Whitt. 2007. Coping with Time-Varying Demand When Setting Staffng Requirements for a Service System. *Production and Operations Management*, 16:13-39.

I. Gurvich, J. Luedtke, and T. Tezcan. 2010. Staffing Call Centers with Uncertain Demand Forecasts: A Chance-Constrained Approach. *Management Science*, 56(7):1093-1115.

J. M. Harrison and A Zeevi. 2005. A Method for Staffing Large Call Centers Based on Stochastic Fluid Models. *Manufacturing & Service Operations Management*, 7:20-36.

R. Ibrahim and P. L'Ecuyer. 2013. Forecasting Call Center Arrivals: Fixed-Effects, Mixed-Effects, and Bivariate Models. *Manufacturing & Service Operations Management*, 15(1):72-85.

R. Ibrahim, P. L'Ecuyer, N. Régnard, and H. Shen. 2012. On the Modeling and Forecasting of Call Center Arrivals. *Proceedings of the 2012 Winter Simulation Conference*, IEEE Press, 256-267.

G. Jongbloed and G. Koole. 2001. Managing Uncertainty in Call Centres Using Poisson Mixtures. *Applied Stochastic Models in Business and Industry*, 17:307-318.

O. Jouini, A. Pot, G. Koole, and Y. Dallery. 2010. Online Scheduling Policies for Multiclass Call Centers with Impatient Customers. *European Journal of Operational Research* 207:258-268.

S.-H. Kim and W. Whitt. 2014. Are Call Center and Hospital Arrivals Well Modeled by Nonhomogeneous Poisson Processes? *Manufacturing & Service Operations Management*, 16(3):464-480.

S. Liao, G. Koole, C. van Delft, and O. Jouini. 2012. Staffing a Call Center with Uncertain Non-Stationary Arrival Rate and Flexibility. *OR Spectum*, 34(3):691-721.

S. Maman. 2009. *Uncertainty in the Demand for Service: the Case of Call Centers and Emergency Departments.* Master's Thesis, Technion, Israel Institute of Technology.

A. Mandelbaum and S. Zeltyn. 2007. The M/M/n+G Queue: Summary of Performance Measures. Technical Note, Technion, Israel Institute of Technology.

V. Mehrotra, O. Ozluk, and R. Saltzman. 2010. Intelligent Procedures for Intra-Day Updating of Call Center Agent Schedules. *Production and Operations Management*, 19(3):353-367.

A. C. Miller and T. R. Rice. 1983. Discrete Approximations of Probability Distributions. *Management Science*, 29:352-362.

B. N. Oreshkin, N. Regnard, and P. L'Ecuyer. 2014. Rate-Based Daily Arrival Process Models with Application to Call Centers. Working Paper, Université de Montréal.

T. R. Robbins and T. P. Harrison. 2010. A Stochastic Programming Model for Scheduling Call Centers with Global Service Level Agreements. *European Journal of Operational Research*, 207:1608-1617.

T. R. Robbins, D. J. Medeiros, and P. Dum. 2006. Evaluating Arrival Rate Uncertainty in Call Centers. In L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto (eds.), *Proceedings of the 2006 Winter Simulation Conference*. Piscataway NJ: IEEE, 2180-2187.

T. R. Robbins, D. J. Medeiros, and T. J. Harrison 2010. Cross Training in Call Centers with Uncertain Arrivals and Global Service Level Agreements. *International Journal of Operations and Quantitative Management*, 16(3):307-329.

A. M. Ross. 2001. Queueing Systems with Daily Cycles and Stochastic Demand with Uncertain Parameters. Ph.D. Dissertation, University of California, Berkeley.

S. M. Ross. 1996. *Stochastic Processes*, 2nd ed. New York: Wiley.

A. Shapiro. 2000. Stochastic Programming by Monte Carlo Simulation Methods. Stochastic Programming e-Print Series, 2000(3), `http://www.speps.org`.

A. Shapiro and A Philpott. 2007. A Tutorial on Stochastic Programming. Technical Note, Georgia Institute of Technology.

H. Shen and J. Z. Huang. 2008. Interday Forecasting and Intraday Updating of Call Center Arrivals. *Manufacturing & Service Operations Management*, 10:391-410.

S. G. Steckley, S. G. Henderson, and V. Mehrotra. 2009. Forecast Errors in Service Systems. *Probability in the Engineering and Informational Sciences*, 23(2):305-332.

J. W. Taylor. 2012. Density Forecasting of Intraday Call Center Arrivals Using Models Based on Exponential Smoothing. *Management Science*, 58:534-549.

W. Wang, S. Ahmed. 2008. Sample Average Approximation of Expected Value Constrained Stochastic Programs. *Operations Research Letters*, 36:515-519.

J. Weinberg, L. D. Brown, and J. R. Stroud. 2007. Bayesian Forecasting of an Inhomogeneous Poisson Process with Applications to Call Center Data. *Journal of the American Statistical Association*, 102:1185-1199.

W. Whitt. 1999. Dynamic Staffing in a Telephone Call Center Aiming to Immediately Answer All Calls. 1999. *Operations Research Letters*, 24:205-212.

W. Whitt. 2006. Fluid Models for Many-Server Queues with Abandonments. *Operations Research*, 54:37-54.

J. Zan, J. Hasenbein, and D. Morton. 2013. Staffing Large Service Systems Under Arrival-rate Uncertainty. Working Paper, UT Austin.

E. Zohar, A. Mandelbaum, and N. Shimkin. 2002. Adaptive Behavior of Impatient Customers in Tele-Queues: Theory and Empirical Support. *Management Science*, 48(4):566-583.

Online Supplement

A Remarks

Remark 1 The dimensionality of the vector time series is typically high, for example, $I = 26$ half-hour periods in a 13-hour working day. Our multiplicative model easily captures the two-way (intraday and interday) time dependence that is common to call centers and other large-scale service systems and reduces potentially high-dimensional forecasts to a single dimension, the daily rate, which can be discretized using Gaussian quadrature. Our model is a parametric analogue of the data-driven approach of Shen and Huang (2008) and is an extension of the approach of Whitt (1999), which proposes a similar approach for un-transformed arrival rates and assumes that all days share a common intraday arrival-rate profile.

In terms of forecast accuracy, our model's performance is roughly comparable to that of the models proposed in Weinberg et al. (2007), Shen and Huang (2008), and Aldor-Noiman et al. (2009). While the arrival count data used in our numerical tests do not exhibit monthly or annual seasonal effects, these could be easily incorporated into (1) as well. There are also more complex models of intraday performance, such as those found in Ibrahim and L'Ecuyer (2013) and Oreshkin et al. (2014), that could further reduce within-day forecast variance, but at the cost of higher forecast dimensionality. While the numerical results reported in §2.5 and §4 suggest that our one-dimensional model is sufficiently accurate to provide desired levels of operational performance, future work could consider methods of dimension reduction of these more complex within-day forecasting approaches. □

Remark 2 Our use of the above results implicitly makes two common assumptions. The first is that, in each of the scenarios, the arrival rate is constant over interval i. While common, this assumption is not necessarily innocuous. Nevertheless, effective measures can be taken to account for time-inhomogeneity within intervals. For a characterization of time-inhomogeneity, see Brown et al. (2005), and for effective responses see Feldman et al. (2007) and Green et al. (2007), as well as the recent article by Kim and Whitt (2014). The second is that, even if the arrival rate were constant, the use of stationary performance measures assumes that the event rate during interval i under scenario k is large enough that transient effects, due to initial conditions at the start of the interval, are not significant. Because our systems include abandonment, relaxation times are similar to those for infinite-server Markovian systems and occur on the order of an expected service time or time to abandonment, whichever is longer (Coolen-Schrijner and Van Doorn 2001). □

Remark 3 While for a specific arrival-rate realization the QoS constraint may be violated, given a correctly forecasted arrival-rate distribution and many *i.i.d.* days, the long-run average abandonment rate should be less than or equal to α^* (Wang and Ahmed 2008). Of course, in our AR(1) setting, arrival-rate distributions need

not be *i.i.d.*, and forecast distributions need not be correct. Nevertheless we conjecture that, in our case, the long-run average abandonment rate will still fall at or below α^* and are working to formally prove it is so (Gans et al. 2014). □

Remark 4 Robbins and Harrison (2010) use a variant of (7) in which the QoS constraint becomes $\sum_{i\in\mathcal{I}} \alpha_i \leq \alpha^*\bar{\lambda} + \delta$, and the objective function is augmented to include a penalty for abandonments above the nominal target of α^*: $\min \sum_{j\in\mathcal{J}} c_j x_j + p\delta$. When abandonments carry explicit, rather than implicit costs, for example in certain contractual arrangements used for outsourcing, this alternate form is appropriate. □

Remark 5 The Kaplan-Meier estimator for exponentially-distributed patience divides the total number of abandonments by the sum of the delays of all calls, including those that are served and those that abandon the queue before being served. See Zohar et al. (2002). Our dataset includes records of average delay in queue only for served calls, however, and we include the total delay of served calls in the denominator of our calculation. If we were to include the waiting time of abandoning calls, it would therefore lower the estimate of θ.

Remark 6 Our definition of α' has the virtue of being straightforward to calculate and analyze. Nevertheless, there are other definitions of α' that we can consider. For example, we can use the optimal solution to (7) to define the expected late-period abandonment rate given only the initial forecast (1): $\alpha' = (\sum_{i\in\mathcal{I}_l} \alpha_i)/(\sum_{i\in\mathcal{I}_l} \lambda_i)$. More generally, one may look for mappings $\{\mathbf{y}_{D+1}, \ldots, \mathbf{y}_{D+h_e}\} \mapsto \alpha'$ that satisfy other objectives, such as the stabilization of late-interval abandonment. In §3.2 we present one such scheme. □

Remark 7 There are several extensions to this scheme that we could consider. Rather than setting a fixed update interval, i^*, Mehrotra et al. (2010) use a sequential procedure that looks for the first period, i^*, for which they can reject the null hypothesis that the arrival-rate pattern comes from the initial forecast distribution. An alternative that follows along the lines of (17) would be to allow for forecast updates and recourse actions after every interval of the scheduling horizon, $i \in \{1, \ldots, I-1\}$. The computation time required to solve such an *a priori* program would grow exponentially with the number of updates, however. A computationally more practical approach would use (13) for repeated *ex post* updates, since computing time would grow only linearly with the number of updates. Even less complex, and perhaps of greatest practical interest, would be an empirical search across days for an optimal static update interval, i^*. Such a test would be straightforward, though time consuming, and we do not pursue in this paper. □

Remark 8 An alternative scheme uses the *ex post* forecast update after i^*, along with the optimal x from RP (17), within UP (13) to recommend the set of recourse actions that are ultimately taken. As in §3.2.4, we identify an appropriate scenario k by rounding up or down. Rather than directly using the associated z_{jgk}s

from (17), however, we substitute $\sum_{i \in \mathcal{I}_l} \alpha_{ik}$, the sum of the optimal α_{ik}s from recourse scenario k, for the right-hand side of (13)'s QoS constraint, $\alpha' \bar{\lambda}'$, and solve (13) to find a low-cost set of recourse actions whose abandonment performance nearly matches that of the original set of z_{jgk}s. □

B Discretization of the Arrival-Rate Forecast Using Gaussian Quadrature

Let ω be a normal random variable with mean μ and variance ϕ, and let ω^* be a discrete random variable such that $\omega^* = \omega_k$ with probability p_k, $1 \leq k \leq K$. In this section, we detail the Gaussian quadrature procedure that we use to compute ω_k and p_k, $1 \leq k \leq K$ so that ω^* matches ω for the first $2K-1$ moments.

Without loss of generality, we first consider a standard normal random variable $Z \sim \mathrm{N}(0,1)$. Denote its first $2K-1$ moments by $\mu_k \equiv \mathrm{E}(Z^k), k = 1, \ldots, 2K-1$. Let Z^* be a discrete random variable such that $Z^* = z_k$ with probability p_k, $1 \leq k \leq K$.

Following the work of Miller and Rice (1983), we proceed to compute $\{(z_k, p_k) \mid 1 \leq k \leq K\}$, for Z^* so that its first $2K-1$ moments match $\mu_k, k = 1, \ldots, 2K-1$:

$$\begin{aligned} p_1 + p_2 + \cdots + p_K &= 1, \\ p_1 z_1 + p_2 z_2 + \cdots + p_K z_K &= \mu_1, \\ p_1 z_1^2 + p_2 z_2^2 + \cdots + p_K z_K^2 &= \mu_2, \\ &\vdots \\ p_1 z_1^{2K-1} + p_2 z_2^{2K-1} + \cdots + p_K z_K^{2K-1} &= \mu_{2K-1}. \end{aligned} \tag{18}$$

The above equations can be solved using a standard method. The derivation below slightly differs from the description of Miller and Rice (1983). A polynomial is first defined as

$$q(z) = (z - z_1)(z - z_2) \ldots (z - z_K) \equiv \sum_{k=0}^{K} q_k z^k. \tag{19}$$

It then follows that $q_K = 1$ and $q(z_k) = 0$ for $1 \leq k \leq K$.

Then, consider the first $K+1$ equations in (18), multiply the first one by q_0, the second one by q_1, etc., and sum them up to obtain:

$$\sum_{k=0}^{K} q_k \mu_k = 0.$$

Similarly, take the second through $(K+2)$th equations in (18), and repeat the above multiplication and summation procedure to obtain:

$$\sum_{k=0}^{K} q_k \mu_{k+1} = 0.$$

The above process can be repeated K times to yield the following set of equations:

$$\begin{aligned} q_0 + \mu_1 q_1 + \mu_2 q_2 \cdots + \mu_{K-1} q_{K-1} &= -\mu_K, \\ \mu_1 q_0 + \mu_2 q_1 + \mu_3 q_2 \cdots + \mu_K q_{K-1} &= -\mu_{K+1}, \\ \mu_2 q_0 + \mu_3 q_1 + \mu_4 q_2 \cdots + \mu_{K+1} q_{K-1} &= -\mu_{K+2}, \\ &\vdots \\ \mu_{K-1} q_0 + \mu_K q_1 + \mu_{K+1} q_2 \cdots + \mu_{2K-2} q_{K-1} &= -\mu_{2K-1}. \end{aligned} \tag{20}$$

The coefficients of polynomial (19), q_k, can be obtained by solving the equations in (20), and the roots of the polynomial determine the discrete atoms $\{z_k \mid 1 \leq k \leq K\}$. The corresponding probabilities, $\{p_k \mid 1 \leq k \leq K\}$, can then be found from the original equations in (18) with the z_k substituted. The whole procedure is implemented in the R function *gauss.quad.prob*, available within the R package *statmod*. The computation of the Gaussian quadrature approximation is very fast.

Since normal distributions form a location-scale family, our variable of interest ω is related with Z through $\omega = \zeta + \phi Z$. Once the z_k and p_k for the standard normal distribution are obtained, we can easily obtain the discrete approximation ω^* for ω as follows,

$$\omega^* = \omega_k \text{ with probability } r_k, \quad \text{where } \omega_k = \zeta + \phi z_k \text{ and } r_k = p_k, \ \forall 1 \leq k \leq K. \tag{21}$$

It can be easily shown that ω and ω^* share the same first $2K-1$ moments.

As a side note, for the degenerate one-scenario case (i.e. $K = 1$), special care is needed to make sure that Λ_i has the correct mean. In this case, we set $\omega_1 = \sqrt{\zeta^2 + \phi^2}$ instead of the default value ζ. The reason is that Gaussian quadrature only matches the first moment of ω when $K = 1$, which won't guarantee the mean matching for Λ_i since $\Lambda_i = (\omega\theta_i)^2$.

C Abandonment Calculations for the Erlang-A Model

The primitives for the Erlang A model, also called the M/M/n+M model, are Markovian interarrival times with mean $1/\lambda$, Markovian service times with mean $1/\mu$, n servers, and Markovian patience times with mean $1/\theta$. Given these data, we use the exact expressions presented in Mandelbaum and Zeltyn (2007) to calculate the stationary probability that an arriving customer abandons queue before being served, $f(\lambda, \mu, \theta, n)$.

Recall that the incomplete gamma function is defined as $\gamma(x, y) \triangleq \int_0^y t^x e^{-t} dt$ and, for $x > 0, y \geq 0$ and can be calculated recursively using standard procedures. We have used the algorithm described in Bhattacharjee (1970) and is available in MATLAB. In turn, we use system primitives and the incomplete gamma function to define two useful quantities,

$$J \triangleq \left(\frac{e^{\lambda/\theta}}{\theta}\right) \cdot \left(\frac{\theta}{\lambda}\right)^{\frac{n\mu}{\theta}} \cdot \gamma\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right) \qquad \text{and} \qquad \mathcal{E} \triangleq \frac{\sum_{j=0}^{n-1} (1/j!) \cdot (\lambda/\mu)^j}{1/(n-1)! \cdot (\lambda/\mu)^{n-1}}.$$

The two quantities then allow us to calculate the desired probability, $f(\lambda, \mu, \theta, n) = \frac{1+(\lambda - n\mu)J}{\mathcal{E}+\lambda J}$.

D Arrival Count Data Used in Numerical Tests

Figure 6 displays summary plots of the arrival data used in the European bank tests reported in Sections 2.5 and 4.1. The left plot shows daily counts for each of 176 days, and the right plot shows average within-day count profiles by day of the week.

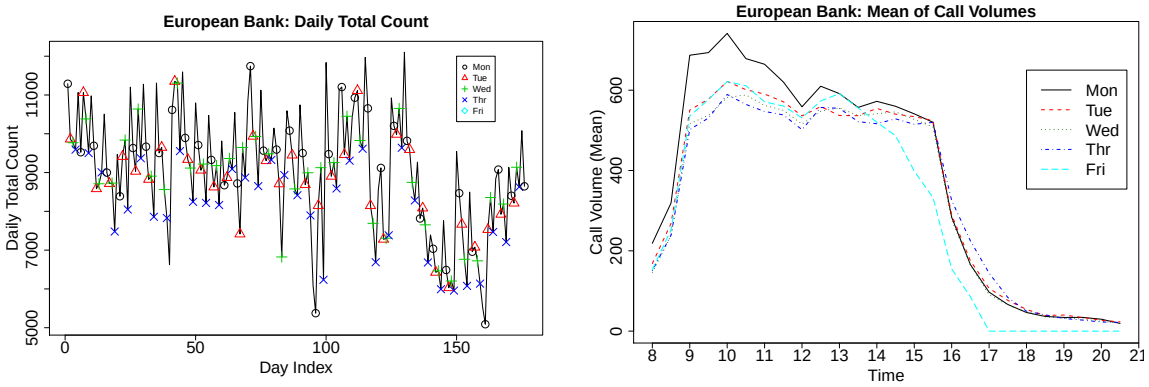

Figure 6: Daily Arrival Counts (left) and Within-Day Average Count Profiles (right) for European Bank

Figure 7 displays summary plots of the arrival data used in the North American bank tests reported in Section 4.2. The left plot shows daily counts for each of 210 days, and the right plot shows average within-day count profiles by day of the week.

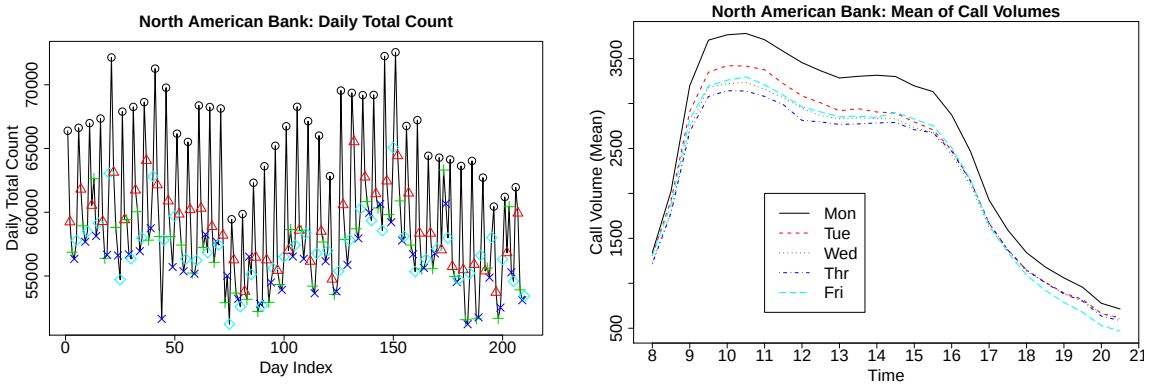

Figure 7: Daily Arrival Counts (left) and Within-Day Average Count Profiles (right) for NA Bank

E Proofs of the Propositions

E.1 Proof of Proposition 3

For notational simplicity, we first drop the subscript d and obtain some generic results.

Suppose we observe y_i, $i = 1, ..., \hat{\imath}$, and denote the corresponding observed vector as $\mathbf{y}_{\hat{\imath}} = (y_1, \ldots, y_{\hat{\imath}})^{\mathsf{T}}$. According to the forecasting model (1), we then have

$$y_i \text{ is a sample of } Y_i = \sqrt{\Lambda_i} + \epsilon_i = \omega\vartheta_i + \epsilon_i, \quad i = 1, ..., \hat{\imath}, \tag{22}$$

where $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ and $\omega \sim \mathrm{N}(\zeta, \psi^2)$.

Now consider the posterior density $f(\omega|\mathbf{y}_{\hat{\imath}})$ as follows

$$\begin{aligned} f(\omega|\mathbf{y}_{\hat{\imath}}) \;=\; \frac{f(\omega, \mathbf{y}_{\hat{\imath}})}{f(\mathbf{y}_{\hat{\imath}})} = \frac{\prod_{i=1}^{\hat{\imath}} f(y_i|\omega) \cdot f(\omega)}{f(\mathbf{y}_{\hat{\imath}})} &= c \cdot e^{-\sum_{i=1}^{\hat{\imath}} \frac{\vartheta_i^2\omega^2 - 2y_i\vartheta_i\omega}{2\sigma^2} - \frac{\omega^2 - 2\zeta\omega}{2\psi^2}} \\ &= c \cdot e^{-\frac{\sum_{i=1}^{\hat{\imath}} \vartheta_i^2\omega^2 - 2\sum_{i=1}^{\hat{\imath}} y_i\vartheta_i\omega}{2\sigma^2} - \frac{\omega^2 - 2\zeta\omega}{2\psi^2}}, \\ &= c \cdot e^{-\frac{\psi^2\sum_{i=1}^{\hat{\imath}} \vartheta_i^2 + \sigma^2}{2\sigma^2\psi^2}\omega^2 + \frac{\psi^2\sum_{i=1}^{\hat{\imath}} y_i\vartheta_i + \sigma^2\zeta}{\sigma^2\psi^2}\omega}, \end{aligned}$$

where c is a leading constant. Recalling that $a_{\hat{\imath}} = \sum_{i=1}^{\hat{\imath}} \vartheta_i \; y_i$ and $\nu_{\hat{\imath}} = \sum_{i=1}^{\hat{\imath}} \vartheta_i^2$, we can see from the above density function that the posterior of ω, given $\mathbf{y}_{\hat{\imath}}$, is the normal distribution with mean and variance

$$\frac{\psi^2 a_{\hat{\imath}} \;+\; \sigma^2\zeta}{\psi^2\nu_{\hat{\imath}} \;+\; \sigma^2}, \qquad \text{and} \qquad \frac{\sigma^2\psi^2}{\psi^2\nu_{\hat{\imath}} \;+\; \sigma^2}.$$

Now suppose that the prior distribution for day d is $\omega_{d-1,d} \sim \mathrm{N}(\zeta_{d-1,d}, \psi^2_{d-1,d})$, and we observe counts $\{y_{di} \mid i = 1, \ldots, \hat{\imath}\}$ on day d. Let $\nu_{d,\hat{\imath}} = \sum_{i=1}^{\hat{\imath}} \vartheta^2_{l_d,i}$ and $a_{d,\hat{\imath}} \equiv \sum_{i=1}^{\hat{\imath}} \vartheta_{l_d,i} \; y_{d,i}$. It then follows from the above generic results that the posterior distribution $\omega_{d_{\hat{\imath}},d}$ is normal with mean and variance

$$\zeta_{d_{\hat{\imath}},d} = \frac{\psi^2_{d-1,d} a_{d,\hat{\imath}} \;+\; \sigma^2\zeta_{d-1,d}}{\psi^2_{d-1,d}\nu_{d,\hat{\imath}} \;+\; \sigma^2}, \qquad \text{and} \qquad \psi^2_{d_{\hat{\imath}},d} \;= \frac{\sigma^2\psi^2_{d-1,d}}{\psi^2_{d-1,d}\nu_{d,\hat{\imath}} \;+\; \sigma^2}.$$

E.2 Lemma 1 and Its Proof

To prove Proposition 5, we first state and prove the following Lemma 1.

Lemma 1 *Suppose $Y_i = \omega\vartheta_i + \epsilon_i, \quad i = 1, \ldots, \hat{\imath}$, where $\omega \sim N(\zeta, \psi^2)$, $\epsilon_i \sim N(0, \sigma^2)$, and ω and ϵ_i are independent. Define $A = \sum_{i=1}^{\hat{\imath}} \vartheta_i Y_i$. Then A is normally distributed with mean $\nu_{\hat{\imath}}\zeta$ and variance $\nu^2_{\hat{\imath}}\psi^2 + \nu_{\hat{\imath}}\sigma^2$, where $\nu_{\hat{\imath}} = \sum_{i=1}^{\hat{\imath}} \vartheta_i^2$.*

Proof We note that $Y_i|\omega \sim \mathrm{N}(\omega\vartheta_i, \sigma^2)$, and the Y_i's are independent conditional on ω. It then follows that

$$A|\omega \sim \mathrm{N}(\omega\nu_{\hat{\imath}}, \sigma^2\nu_{\hat{\imath}}).$$

Given that $\omega \sim \mathrm{N}(\zeta, \psi^2)$, the density function of A at a can be written as

$$f(a) = \int f(a|\omega) f(\omega) d\omega = \int \frac{1}{\sqrt{2\pi\sigma^2\nu_{\hat{\imath}}}} e^{-\frac{(a-\omega\nu_{\hat{\imath}})^2}{2\sigma^2\nu_{\hat{\imath}}}} \cdot \frac{1}{\sqrt{2\pi\psi^2}} e^{-\frac{(\omega-\zeta)^2}{2\psi^2}} d\omega,$$

which can be reorganized as

$$f(a) = \frac{1}{\sqrt{(2\pi)^2\nu_{\hat{\imath}}\sigma^2\psi^2}} e^{-\frac{a^2}{2\sigma^2\nu_{\hat{\imath}}} - \frac{\zeta^2}{2\psi^2}} \int e^{-\frac{\nu_{\hat{\imath}}^2\omega^2 - 2\nu_{\hat{\imath}}a\omega}{2\sigma^2\nu_{\hat{\imath}}} - \frac{\omega^2 - 2\zeta\omega}{2\psi^2}} d\omega.$$

Note that the exponent of the integrand can be re-expressed as

$$-\frac{\nu_{\hat{\imath}}\psi^2 + \sigma^2}{2\sigma^2\psi^2}(\omega - \frac{a\psi^2 + \zeta\sigma^2}{\nu_{\hat{\imath}}\psi^2 + \sigma^2})^2 + \frac{(a\psi^2 + \zeta\sigma^2)^2}{2\sigma^2\psi^2(\nu_{\hat{\imath}}\psi^2 + \sigma^2)}.$$

Then, the density function $f(a)$ equals to

$$f(a) = \frac{1}{\sqrt{2\pi\nu_{\hat{\imath}}(\nu_{\hat{\imath}}\psi^2 + \sigma^2)}} e^{-\frac{(a-\nu_{\hat{\imath}}\zeta)^2}{2\nu_{\hat{\imath}}(\nu_{\hat{\imath}}\psi^2 + \sigma^2)}},$$

which implies that $A \sim \mathrm{N}(\nu_{\hat{\imath}}\zeta, \nu_{\hat{\imath}}^2\psi^2 + \nu_{\hat{\imath}}\sigma^2)$. □

E.3 Proof of Proposition 5

Consider the following model:

$$Y_{D+d,i} = \omega_{D+d}\nu_{l_{D+d},i} + \epsilon_{D+d,i}, \qquad i = 1, \ldots, I, \tag{23}$$

where $\omega_{D+d} \sim \mathrm{N}(\zeta_{D,D+d}, \psi^2_{D,D+d})$, $\epsilon_{D+d,i} \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$, and ω_{D+d} and $\epsilon_{D+d,i}$ are independent.

Given (14), it follows from Lemma 1 that

$$A_{D+d,\hat{\imath}} \sim \mathrm{N}(\nu_{D+d,\hat{\imath}}\zeta_{D,D+d}, \ \ \nu^2_{D+d,\hat{\imath}}\psi^2_{D,D+d} + \nu_{D+d,\hat{\imath}}\sigma^2). \tag{24}$$

The definition of $\zeta'_{D,D+d}$ in (15) implies that $\zeta'_{D,D+d}$ is a linear shifted and scaled transformation of $A_{D+d,\hat{\imath}}$; hence $\zeta'_{D,D+d}$ is also normally distributed, and its variance is a scaled version of the variance of $A_{D+d,\hat{\imath}}$, whose expression is given in the right equality of (16).

Below we use induction to prove that $E(\zeta'_{D,D+d}) = \zeta_{D,D+d}$. To begin with, consider $d = 1$. Note that

$$\zeta'_{D,D+1} = \frac{\psi^2_{D,D+1}A_{D+1,\hat{\imath}} \ + \ \sigma^2\zeta_{D,D+1}}{\psi^2_{D,D+1}\nu_{D+1,\hat{\imath}} \ + \ \sigma^2},$$

which implies that

$$E(\zeta'_{D,D+1}) = \frac{\psi^2_{D,D+1}\nu_{D+1,\hat{\imath}}\zeta_{D,D+1} \ + \ \sigma^2\zeta_{D,D+1}}{\psi^2_{D,D+1}\nu_{D+1,\hat{\imath}} \ + \ \sigma^2} = \zeta_{D,D+1}.$$

Now suppose that, for $d \geq 2$, we have proven that $E(\zeta'_{D,D+d-1}) = \zeta_{D,D+d-1}$, for any $\hat{\imath} \in \{1, \ldots, I\}$. We then have

$$\zeta_{D,D+d} - E(\zeta'_{D,D+d}) = \frac{\sigma^2}{\psi^2_{D+d-1,D+d}\nu_{D+d,\hat{\imath}} \ + \ \sigma^2} \ (\zeta_{D,D+d} - \zeta_{D+d-1,D+d}). \tag{25}$$

Hence, it suffices to show that $\zeta_{D,D+d} - \zeta_{D+d-1,D+d} = 0$. For that, we note

$$\begin{aligned}
\zeta_{D,D+d} - \zeta_{D+d-1,D+d} &= \zeta_{D,D+d} - E(E(\omega_{D,D+d}|\mathbf{Y}_{D+1},\ldots,\mathbf{Y}_{D+d-1})) \\
&= \zeta_{D,D+d} - \{\alpha_{l_{D+d}} + \beta[E(E(\omega_{D,D+d-1}|\mathbf{Y}_{D+1},\ldots,\mathbf{Y}_{D+d-1})) - \alpha_{l_{D+d-1}}]\} \\
&= \zeta_{D,D+d} - \alpha_{l_{D+d}} - \beta[E(\zeta'_{D,D+d-1}) - \alpha_{l_{D+d-1}}] \\
&= \zeta_{D,D+d} - \alpha_{l_{D+d}} - \beta(\zeta_{D,D+d-1} - \alpha_{l_{D+d-1}}) \\
&= \beta^d(\omega_D - \alpha_{l_D}) - \beta^d(\omega_D - \alpha_{l_D}) = 0.
\end{aligned}$$

Thus, we have shown that $E(\zeta'_{D,D+d}) = \zeta_{D,D+d}$ for $d \geq 1$, which concludes the induction proof.

F Results for Wilcoxon Signed Rank Tests

The figures below show results for Wilcoxon signed rank tests that check to see if the median of the difference between two matched sets of values is significantly different from zero. The tests were implemented using the *signrank* function in MATLAB R2013a.

Results of 1-Sided Wilcoxon Signed Rank Tests Using European Bank Abandonment Rates
Left-Tail Probabilities for [Colum Values - Row Values]

	SP1	UP1	RP1	SP4	UP4	RP4	SP100	UP100	RP100
SP1	1.000	0.750	0.024	1.000	1.000	1.000	1.000	1.000	1.000
UP1	0.251	1.000	0.032	1.000	1.000	1.000	1.000	1.000	1.000
RP1	0.976	0.968	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SP4	0.000	0.000	0.000	1.000	0.992	0.413	0.868	0.845	0.002
UP4	0.000	0.000	0.000	0.008	1.000	0.323	0.042	0.310	0.000
RP4	0.000	0.000	0.000	0.589	0.678	1.000	0.553	0.639	0.006
SP100	0.000	0.000	0.000	0.134	0.959	0.450	1.000	0.966	0.002
UP100	0.000	0.000	0.000	0.156	0.693	0.363	0.035	1.000	0.001
RP100	0.000	0.000	0.000	0.998	1.000	0.994	0.998	0.999	1.000

Figure 8: Results of One-Tailed Wilcoxon Signed Rank Tests for European Bank Abandonment Data

Results of 1-Sided Wilcoxon Signed Rank Tests Using European Bank Cost per Handled Call
Right-Tail Probabilities for [Column Values - Row Values]

	SP1	UP1	RP1	SP4	UP4	RP4	SP100	UP100	RP100
SP1	1.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000	0.999
UP1	1.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
RP1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SP4	0.000	0.000	0.000	1.000	0.000	0.000	0.985	0.000	0.000
UP4	0.000	0.000	0.000	1.000	1.000	0.006	1.000	0.371	0.000
RP4	0.000	0.000	0.000	1.000	0.994	1.000	1.000	0.993	0.028
SP100	0.000	0.000	0.000	0.015	0.000	0.000	1.000	0.000	0.000
UP100	0.000	0.000	0.000	1.000	0.631	0.007	1.000	1.000	0.000
RP100	0.001	0.000	0.000	1.000	1.000	0.973	1.000	1.000	1.000

Figure 9: Results of One-Tailed Wilcoxon Signed Rank Tests for European Bank Average Cost Data

Figure 8 reports the results of one-tailed tests that compare daily abandonment rates across the nine schemes. The result shown in each cell of the figure's table is based on a set of 76 values, where each value is the

abandonment rate associated with the column label, less the abandonment rate associated with the row label, for a single day. The one-sided, left-tailed p-value shown in the cell is the probability that the median value is less than zero, and we interpret the p-value to be a measure of the probability that the abandonment rates associated with the column label's scheme are significantly less than those associated with the row label's scheme. Below, we denote a particular cell by its (row,column)-heading coordinates.

For example, cell (SP1,UP1) shows that the p-value associated with the probability that the median of UP1-SP1 differences is less than zero is 0.75. In this case, we cannot reject the null hypothesis that the median value of the differences is less than zero, and we conclude that UP1's abandonment rates do not appear to be significantly higher than SP1's.

In Figure 8, clusters of adjacent (blue) cells with solid borders compare different schemes that use the same number of scenarios. For example, from (SP1,UP1) we see that UP1's daily abandonment rates do not appear to be significantly higher than SP1's. In contrast, from the p-values reported in (SP1,RP1) and (UP1,RP1) we see that RP1's daily abandonment rates appear to be significantly higher than those of the other two 1-scenario schemes. In contrast, results for (SP4,UP4), (SP4,RP4), and (UP4,RP4) show that UP4's daily abandonment rates do not appear to be significantly higher than SP4's, nor do RP4's as compared to SP4's and UP4's. While UP100's rates do not appear to be significantly higher than SP100's, RP100's abandonment rates do appear to be systematically higher than those of the other two schemes. This last fact is visible to the eye in the left panel of Figure 4.

The set of three diagonal (green) cells with dashed borders in Figure 8 reports tests of differences between the results of 4 and 100-scenario versions that use the same scheduling approach. From (SP100,SP4), (UP100,UP4), and (RP100,RP4) we see that the 4-scenario version of these schemes have abandonment-rates that do not appear to be significantly higher than that of their 100-scenario counterparts. This is evidence that only a small number of scenarios is needed to ensure that the schemes properly compensate for arrival-rate variability.

Figure 9 reports analogous results for daily cost per handled call. Here, we report right-tail probabilities and interpret small p values to suggest that average costs associated with the focal column label's scheme are systematically larger than those associated with the focal row label's scheme. In this figure we see that, for any number of scenarios, the UP scheme appears to have lower average costs than its analogous SP scheme, and the RP scheme appears to have lower average costs than either its SP or UP counterpart. When we compare SP4 to SP100 and UP4 to UP100, we see that the schemes with higher numbers of scenarios do not have significantly lower costs. In contrast RP100 has average costs that appear to be significantly lower than those of RP4, but as Figure 8 indicates, the cost savings also come with significantly higher abandonment rates.

In sum, for the European Bank call center data, the use of UP schemes provides significantly lower costs

than SP schemes, without appearing to degrade abandonment rate performance. Furthermore, the 4-scenario versions of the SP and UP schemes appear to perform as well as the 100-scenario versions. While the same observations hold for RP4, they do not hold for RP100.

Results of 1-Sided Wilcoxon Signed Rank Tests Using North American Bank Abandonment Rates
Left-Tail Probabilities for [Colum Values - Row Values]

	SP1	UP1	RP1	SP4	UP4	RP4	SP100	UP100	RP100
SP1	1.000	0.091	0.000	1.000	1.000	1.000	1.000	1.000	1.000
UP1	0.909	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
RP1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SP4	0.000	0.000	0.000	1.000	0.207	0.998	0.621	0.231	0.924
UP4	0.000	0.000	0.000	0.794	1.000	0.997	0.914	0.441	0.978
RP4	0.000	0.000	0.000	0.002	0.003	1.000	0.003	0.004	0.214
SP100	0.000	0.000	0.000	0.380	0.086	0.997	1.000	0.110	0.864
UP100	0.000	0.000	0.000	0.770	0.560	0.996	0.891	1.000	0.965
RP100	0.000	0.000	0.000	0.077	0.022	0.787	0.137	0.036	1.000

Figure 10: Results of One-Tailed Wilcoxon Signed Rank Tests for NA Bank Abandonment Data

Results of 1-Sided Wilcoxon Signed Rank Tests Using North American Bank Cost per Handled Call
Right-Tail Probabilities for [Column Values - Row Values]

	SP1	UP1	RP1	SP4	UP4	RP4	SP100	UP100	RP100
SP1	1.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000
UP1	1.000	1.000	0.000	1.000	1.000	0.000	1.000	1.000	0.000
RP1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SP4	0.000	0.000	0.000	1.000	0.000	0.000	0.766	0.000	0.000
UP4	1.000	0.000	0.000	1.000	1.000	0.000	1.000	0.440	0.000
RP4	1.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000	0.125
SP100	0.000	0.000	0.000	0.235	0.000	0.000	1.000	0.000	0.000
UP100	1.000	0.000	0.000	1.000	0.562	0.000	1.000	1.000	0.000
RP100	1.000	1.000	0.000	1.000	1.000	0.875	1.000	1.000	1.000

Figure 11: Results of One-Tailed Wilcoxon Signed Rank Tests for NA Bank Average Cost Data

Figures 10 and 11 present the results of analogous Wilcoxon tests for the 110 sets of abandonment rates and costs per handled call generated with the North American data. From the figures' tables we see that the RP schemes are better behaved in these tests. As we would expect, the RP schemes appear to have systematically lower costs per call than the SP or UP schemes, without having systematically higher abandonment rates. Similarly the 100-scenario versions of SP, UP, and RP all have abandonment rates and costs per call that do not appear to be significantly better than those for their 4-scenario counterparts.

G Computation Times

All the numerical tests were run using OPL-CPLEX Optimization Studio 12.6 on a dedicated PC with a 2.4 GHz Intel Core 2 Quad CPU and 8 GB of RAM.

G.1 Computation Times for Quadrature and Sampling-Based Scenario Generation

In this part, we report computation times for the tests performed in Section 2.5.1.

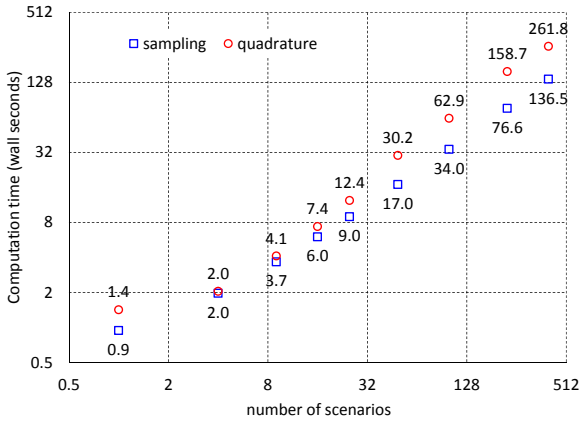

Figure 12: Day 101 Computation Times, by Number of Scenarios

Figure 12 reports so-called "wall" computation times for each method and number of scenarios. That is, the times are the elapsed time recorded by the computer's clock and account for the sometimes significant preprocessing time needed to set up the optimization problems. For each number of scenarios, results for quadrature reflect the solution of a single IP, while those for sampling are the average of solution times of the 100 IP's in each associated set.

The panel's plot shows that, for both methods, computation times appear to grow roughly linearly with the number of scenarios and that the times for quadrature are about double the average computation time (across 100 samples) for larger numbers of scenarios. The longest times, at 400 scenarios, are on average 136 seconds for sampling-based scenarios, and about 260 seconds when using quadrature. (These computation times seem reasonable for practical implementation.) Of course, the longer computation times required by quadrature-based scenarios are compensated by the fact that IP solutions for quadrature appear to be stable with many fewer scenarios, as little as 4.

G.2 Computation Times for Tests of SPm, UPm, and RPm Schemes

In this part, we report computation times for the tests performed in Section 4. Figure 13 shows the nine schemes' average wall computation times for the European retail bank call center dataset, in the left panel, and for the North American retail bank call center dataset, in the right panel, respectively

Average computation times for the European retail bank call center dataset increase with the number of scenarios and with the complexity of the scheme. All 1 and 4-scenario schemes took less than 20 seconds of wall clock time to set up and solve a given day's mathematical program, on average. RP100 was the most

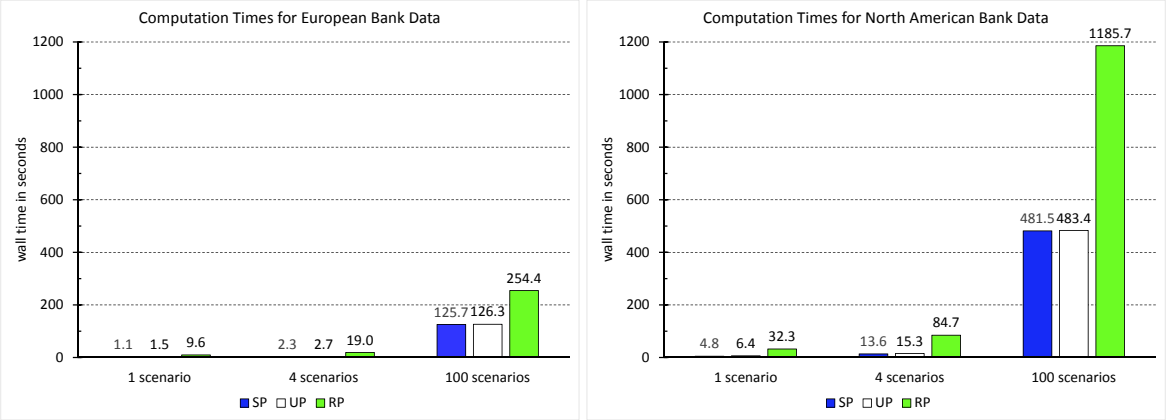

Figure 13: Nine Schemes' Average Computation Times for European (left) and NA (right) Data

computationally intensive scheme, requiring an average of 254 seconds, about twice the 126-second average required to set up and solve UP100.

Average computation times for the North American retail bank call center dataset were systematically longer than those for the European bank, but with the same relative patterns. For example, UP4 took less than 15 seconds of wall clock time, on average, to set up and solve, while RP4 took a bit less than 85 seconds on average. RP100 was again the most computationally intensive scheme, requiring an average of 1186 seconds – almost 20 minutes – to set up and solve, more than twice the 484-second average required to set up and solve UP100. As noted above, for practical purposes it is sufficient to solve the 4-scenario versions of the recourse programs.