Annals of Operations Research 127, 333–358, 2004

Call Center Staffing with Simulation and Cutting Plane Methods

JÚLÍUS ATLASON* and MARINA A. EPELMAN {jatlason, mepelman}@umich.edu
Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI 48109-2117, USA

SHANE G. HENDERSON shane@orie.cornell.edu
School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853, USA

Abstract. We present an iterative cutting plane method for minimizing staffing costs in a service system subject to satisfying acceptable service level requirements over multiple time periods. We assume that the service level cannot be easily computed, and instead is evaluated using simulation. The simulation uses the method of common random numbers, so that the same sequence of random phenomena is observed when evaluating different staffing plans. In other words, we solve a sample average approximation problem. We establish convergence of the cutting plane method on a given sample average approximation. We also establish both convergence, and the rate of convergence, of the solutions to the sample average approximation to solutions of the original problem as the sample size increases. The cutting plane method relies on the service level functions being concave in the number of servers. We show how to verify this requirement as our algorithm proceeds. A numerical example showcases the properties of our method, and sheds light on when the concavity requirement can be expected to hold.

Keywords: simulation optimization, call centers, cutting planes, sample average approximation

1. Introduction

In this paper we present the theoretical properties of a cutting plane method for minimizing staffing costs in a service system subject to satisfying acceptable service level requirements over multiple time periods. This method was proposed by Henderson and Mason (1998) and combines simulation and integer programming in an iterative cutting plane algorithm. Simulation is a powerful method for analyzing complex systems, but optimization with simulation can be difficult. Linear integer programming problems, along with many other mathematical programming models, are well studied and many algorithms have been developed for solving problems in this form, but a simplification of the system is often required for modelling. Our iterative cutting-plane algorithm combines simulation and linear (integer) programming to solve resource allocation problems where the objective function, or constraints, or both, are evaluated via simulation. The algorithm relies on the concavity of the problem constraints, but in our algorithm we have a built-in robustness, so that nonconcavity can be detected.

* Corresponding author.

The method of combining simulation and optimization in this way has potential applications in various service systems, such as call center staffing (which will be the focus of this paper) and emergency vehicle dispatching (which we are currently investigating). In fact, the method could potentially, with appropriate modifications, be utilized in many other areas where simulation is an appropriate modelling tool.

The problem of determining optimal staffing levels in a call center (see, e.g., Thompson (1997)) is a motivating example for our work. The decision maker faces the task of creating a collection of tours (work schedules) for the call center of low cost that together ensure a satisfactory service level. A tour is comprised of several shifts and has to observe several restrictions related to labor contracts, management policies, etc. We divide the planning horizon (typically a day or a week) into small periods (15–60 minutes) and focus on the service level in each period. We define the service level in a given period as the fraction of calls received in that period answered within a specified time limit. In this paper we focus on the problem of minimizing cost while satisfying the service level and scheduling constraints.

The traditional method for solving this problem involves two steps. First, the required staffing level in each period is estimated, independently period by period, often using i.e. queueing theory. Second, an integer program is solved to determine how many workers should be assigned to each of the tours in order to "cover" the previously assigned staffing levels. Our method combines these two steps and allows for dependence between periods. There are examples in the literature (Green, Kolesar, and Soares, 2001; Ingolfsson, Haque, and Umnikov, 2002; Jennings et al., 1996) that show that significant cost savings can be obtained by doing so, or that a staffing level obtained by assuming independence does not meet performance criteria when there is, in fact, dependence between periods. Indeed, we present an example at the end of the paper showing that the staffing level in one period can have a considerable effect on the service level in another period. Green, Kolesar, and Soares (2001) and Jennings et al. (1996) suggest a relatively simple method for determining the required staffing levels that accounts for such dependence. Their method is based on infinite server queuing models, but requires that the call center can be accurately modelled as a $G(t)/G(t)/s(t)$ queuing system.

The cost function is usually relatively straightforward to calculate. We can calculate the cost of each tour (salary costs, appeal to employees, etc.), and multiply by the number of employees working each tour to get the overall cost. The service level, on the other hand, can be very difficult to obtain. Queuing models can be used for simple problems, but simulation must be used to accurately model complex systems. The difficulty with using simulation is the large number of possible solutions since it is impractical to evaluate all of them. By using integer programming, we hope that we only need to simulate a small portion of the solution set.

Simulation has been widely used to analyze the impact of different staffing levels on service levels and commercial simulation packages, specially designed for call centers, are available. Integer programming has also been used, in which case the staff requirements in each period are usually needed as an input in the model (see Mehrotra, Murphy, and Trick (2000)).

We present a cutting plane method based on the one developed by Kelley, Jr. (1960). The method solves a linear (integer) program to obtain the staffing levels, and the solution is used as an input for a simulation to calculate the service level. If the service level is unsatisfactory, we add a constraint to the linear program and go to the next iteration.

Kelley's cutting plane method applies to minimization problems where both the objective function and feasible region (of the continuous relaxation of the integer problem) need to be convex. The costs in the call center problem are linear and we will assume that the service level function is concave, so that (see equation (1)), the feasible region, relaxing the integer restriction, is convex. Since the service level function is unknown beforehand, we need to incorporate a mechanism into the method to verify that the concavity assumption holds. In section 5 we present a numerical method for checking concavity of a function, when the function values and possibly gradients are only known at a finite number of points.

Morito et al. (1999) use simulation in a cutting-plane algorithm to solve a logistic system design problem at the Japanese Postal Service. Their problem is to decide where to sort mail provided that some post offices have automatic sorting machines but an increase in transportation cost and handling is expected when the sorting is more centralized. The algorithm proved to be effective for this particular problem and found an optimal solution in only 3 iterations where the number of possible patterns (where to sort mail for each office) was 2^{30}. Their discussion of the algorithm is ad hoc, and they do not discuss its convergence properties.

Ingolfsson, Haque, and Umnikov (2002) present an algorithm for solving a call center staffing problem that uses a genetic algorithm for the optimization component and numerical solution of differential equations for evaluating the service level. Ingolfsson and Cabral (2002) have developed a cutting plane algorithm for this problem using queuing models instead of simulation to calculate the service levels. The cuts are generated using a heuristic, based on approximating the service level in each period as a function of the staffing level in that period, and may not be valid, although examples suggest good performance.

Cutting plane methods have been successfully used to solve two stage stochastic linear programs. In many applications the sample space becomes so large that one must revert to sampling to get a solution (Birge and Louveaux, 1997; Infanger, 1994). The general cutting plane algorithm for two stage stochastic programming is known as the L-shaped method (van Slyke and Wets, 1969) and is based on Benders decomposition (Benders, 1962). Stochastic decomposition (Higle and Sen, 1991) for solving the two stage stochastic linear program starts with a small sample size, which is increased as the algorithm progresses and gets closer to a good solution. Stochastic decomposition could also be applied in our setting, but that is not within the scope of this paper.

The random nature of our problem and the absence of an algebraic form for the service level function makes the optimization challenging. We use sampling to get an estimate of the service level function, and optimize the sample average approximation.

An important question is whether the solution to the sample average approximation converges to a solution to the original problem, and if so, how fast.

We apply the strong law of large numbers to prove conditions for almost sure convergence and apply a result due to Dai, Chen, and Birge (2000) to prove an exponential rate of convergence of the optimal solutions as the sample size increases. Vogel (1994) proved almost sure convergence in a similar setting, but we include proofs for reasons listed in section 4.1. Shapiro and Homem-de-Mello (2000) established conditions for an exponential rate of convergence of the probability that the solution to the sample average approximation is exactly the solution to the original problem in the case of a discrete distribution and Vogel (1988) proved a polynomial rate of convergence in a similar setting, but under weaker conditions than we require. The optimization of sample average approximations has also been studied in the simulation context (Chen and Schmeiser, 2001; Healy and Schruben, 1991; Robinson, 1996; Rubenstein and Shapiro, 1993).

The main contribution of this paper is to demonstrate the potential of bringing simulation and traditional optimization methods together. We establish the properties of a new method for solving call center staffing problems. The method is carefully developed because we believe that the same idea can be applied to resource allocation problems other than staffing problems, as previously mentioned. In addition, we present a numerical method for checking the concavity of a function when the function value and possibly gradient is only known at a finite number of points.

The computing requirements of the algorithm presented here, as applied to realistically-sized problems, are rather large. Indeed, it is often the case that the covering integer programs alluded to earlier (with predetermined staffing levels in each period) are difficult to solve, so that iterating such a step with simulation appears to be a rather formidable computational task. We view this work as a first step in the process of integrating the steps of determining work requirements and covering the work requirements with tours. Subsequent work will focus on exploring methods for making the approach computationally feasible. We have many ideas for how this could be achieved; see section 7 for more comments on this issue.

The paper is organized as follows. We formulate the call center staffing problem in section 2. We present the cutting plane algorithm and its convergence properties in section 3. The convergence and the rate of convergence of the solutions of the sample average approximation to solutions of the original problem are proved in section 4. The numerical method for checking concavity is described in section 5 and an implementation of the overall method is described in section 6. Conclusions and considerations for future research are given in section 7.

2. Call center staffing

In this section we formulate and discuss in more detail the call center staffing problem of minimizing cost subject to service level constraints.

2.1. Formulation and notation

The management of a call center needs some criteria to follow when they decide on a set of staffing levels. It is not unusual in practice to determine the staffing levels from a service level perspective. In an emergency call center, for example, it might be required that 90% of received calls should be answered within 10 seconds.

We introduce terminology and notation before we formulate the problem. The set of permissible tours (predefined work schedules over the planning horizon) can be conveniently set up in a matrix (see Dantzig (1954)). More specifically we have

$$A_{ij} = \begin{cases} 1, & \text{if period } i \text{ is included in tour } j, \\ 0, & \text{otherwise.} \end{cases}$$

From the above we see that a column in A represents a feasible tour and a row in A represents a specific period. We let p be the total number of periods and m be the number of feasible tours. If we let $x \in \mathbb{R}^m$ be a vector where the jth component represents the number of employees that work tour j, then $Ax = y \in \mathbb{R}^p$ is a vector where the ith component of y corresponds to the number of employees that are working in period i. We let c be the cost vector, where c^j is the cost per employee working tour j.

Next we define the service level constraints. We let $l \in \mathbb{R}^p$ be the vector whose ith component is the minimum acceptable service level in period i, for example, 90%. Since, for example, the arrival and service times of customers are not known but are random, the service level in each period will be a random variable. Let ξ, a random vector, denote all the random quantities in the problem and let $\xi^1, \ldots, \xi^n$ denote independent realizations of ξ. Let $N^i(\xi)$ be the number of calls received in period i and let $S^i(y, \xi)$ be the number of those calls answered within a pre-specified time limit, for example, 10 seconds, based on the staffing level y. The fraction of customers receiving adequate service in period i in the long run is then

$$\lim_{n\to\infty} \frac{\sum_{d=1}^{n} S^i(y, \xi^d)}{\sum_{d=1}^{n} N^i(\xi^d)} = \frac{\lim_{n\to\infty} n^{-1}\sum_{d=1}^{n} S^i(y, \xi^d)}{\lim_{n\to\infty} n^{-1}\sum_{d=1}^{n} N^i(\xi^d)}.$$

If $E[N^i(\xi)] < \infty$ then the strong law of large numbers can be applied separately to both the numerator and denominator of this expression, and then the desired long-run ratio is $E[S^i(y, \xi)]/E[N^i(\xi)]$. Thus, $E[S^i(y, \xi)]/E[N^i(\xi)] \geqslant l^i$ is a natural representation of the service level constraint (excluding the pathological case $E[N^i(\xi)] = 0$) in period i. If we define $G^i(y, \xi) := S^i(y, \xi) - l^i N^i(\xi)$ then we can conveniently write the service level constraint as $E[G^i(y, \xi)] \geqslant 0$. Define $g^i(y) := E[G^i(y, \xi)]$ as the expected service level in period i as a function of the server allocation vector y and let $g : \mathbb{R}^p \to \mathbb{R}^p$ be a function whose ith component is g^i.

We are now ready to formulate the problem of minimizing staffing costs subject to satisfying a minimum service level in each period. It is

$$\begin{array}{ll} \min\ c^T x & \\ \text{subject to} & Ax \geqslant y, \\ & g(y) \geqslant 0, \\ & x \in X, \\ & x, y \geqslant 0 \text{ and integer}, \end{array} \tag{1}$$

where X is a compact set. The compactness of X can be easily justified in practice. It is, for example, impossible to hire an infinite number of employees, and there are usually budget constraints which impose an upper bound on x since c is generally positive. We also define, for future reference,

$$Y := \{y \geqslant 0 \text{ and integer: } \exists\, 0 \leqslant x \in X \text{ and integer with } Ax \geqslant y\}.$$

Note that Y is a finite set since X is compact and the entries in A are either 0 or 1.

The functions $g^i(y)$ are expected values, and the underlying model might be so complex that an algebraic expression for $g(y)$ can not be easily obtained. Therefore, simulation could be the only viable method for estimating $g(y)$. In the next subsection we formulate (1) as an approximate problem, where the expected values are replaced by sample averages.

2.2. *Sample average approximation of the call center problem*

In this paper we assume that the algebraic form of the service level function $g(y)$ is not available, and that its value is evaluated using simulation. Suppose we run a simulation with sample size n, where we independently generate the realizations $\{\xi^d\}_{d=1}^n$ from the distribution of ξ, to get an estimate of the expected values $g(y)$. Let $\bar{g}_n(y) = (1/n)\sum_{d=1}^n G(y, \xi^d)$ be the resulting estimates and let $\bar{g}_n^i(y)$ denote the ith component of $\bar{g}_n(y)$. We use this notation to formulate the sample average approximation

$$\begin{array}{ll} \min\ c^T x & \\ \text{subject to} & Ax \geqslant y, \\ & \bar{g}_n(y) \geqslant 0, \\ & x \in X, \\ & x, y \geqslant 0 \text{ and integer}. \end{array} \tag{2}$$

The problem above is linear except for the service level function $\bar{g}_n(y)$. We assume that each of the component functions $\bar{g}_n^i(y)$ are concave so that we can approximate them with piecewise linear concave functions and solve the sample average approximation by using cutting plane methods. In the next subsection we discuss the concavity assumption in more detail.

2.3. *Concave service levels*

Intuitively, we would expect that the service level increases if we increase the number of employees in any given period. We also conjecture that the marginal increase in service level decreases as we add more employees. If these speculations are true then $g^i(y)$ is increasing and concave in each component of y for all i. We will make the stronger assumption that $g^i(y)$ and $\bar{g}^i_n(y)$ are increasing componentwise and jointly concave in y, for all i. Our initial computational results suggest that this is a reasonable assumption, at least within a region containing practical values of y (see section 6). Others have also studied the convexity of performance measures of queuing systems. Akşin and Harker (2001) show that the throughput of a call center is stochastically increasing and directional concave in the sample path sense as a function of the allocation vector y in a similar setting. Analysis of the steady state waiting time of customers in an $M/M/s$ queue shows that its expected value is a convex and decreasing function of the number of servers s (Dyer and Proll, 1977), its expected value is convex and increasing as a function of the arrival rate (Chen and Henderson, 2001) and its distribution function evaluated at any fixed value, is concave and decreasing as a function of the arrival rate (Chen and Henderson, 2001). See other references in Chen and Henderson (2001) for further studies in this direction. Koole and van der Sluis (2003) developed a local search algorithm for a call center staffing problem with a global service level constraint. When the service level constraint satisfies a property called multimodularity their algorithm is guaranteed to terminate with a global optimal solution. There are, however, examples where the service level constraint, as defined in this paper, is not multimodular even for nondecreasing and concave service level functions.

If the concavity assumption holds, then we can approximate the service level function with piecewise linear concave functions, which can be generated as described below. The following definition is useful.

Definition 1 (Rockafellar, 1970, p. 308). Let $y^k \in \mathbb{R}^p$ be given. If $h:\mathbb{R}^p \to \mathbb{R}$ is a concave function and $q(y^k) \in \mathbb{R}^p$ is such that

$$h(y) \leqslant h\big(y^k\big) + q(y^k)^T\big(y - y^k\big) \quad \forall y \in \mathbb{R}^p \tag{3}$$

then $q(y^k)$ is a subgradient of h at y^k.

The term "supergradient" might be more appropriate since the hyperplane $\{h(y^k) + q(y^k)^T(y - y^k)\}$ lies above the function h, but we use "subgradient" to conform with the literature. A concave function has at least one subgradient at every point (see theorem 3.5.2 in Bazaraa, Sherali, and Shetty (1993)). The notion of concavity and subgradients is defined for functions of continuous variables, but we are dealing with functions of integer variables. We say that such a function h is concave if no points of the form $(x, h(x)) \in \mathbb{R}^{p+1}$ (with $x \in \mathbb{Z}^p$) lie in the interior of the convex hull of the set $\{(y, h(y))\colon\ y \in \mathbb{Z}^p\} \subseteq \mathbb{R}^{p+1}$. We replace $\mathbb{R}^p$ with $\mathbb{Z}^p$ in definition 1 to define the subgradient of a function with integer domain.

Let $q^i(y^k)$ and $\bar{q}_n^i(y^k)$ be subgradients at y^k of g^i and $\bar{g}_n^i$, respectively. There are many potential methods one might consider to obtain the subgradients. Finite differences using differences of length 1 appear reasonable since we are working with integer variables. There are, however, examples where that fails to produce a subgradient, even for a concave nondecreasing function. Still, we used finite differences in our numerical study and converged to an optimal solution of the sample average approximation. Gradients might also be obtained using infinitesimal perturbation analysis (IPA) (see, e.g., Glasserman (1991)). Before using IPA we would have to extend the service level function to a differentiable function defined over a continuous domain, since IPA is applied in settings where the underlying function is differentiable.

The subgradients are used to approximate the service level constraints. Let y^k be a given server allocation vector, and suppose that $\bar{g}_n^i(y^k)$ and $\bar{q}_n^i(y^k)$ are obtained via simulation. If our assumptions about concavity hold then by definition 1 we must have $\bar{g}_n^i(y) \leqslant \bar{g}_n^i(y^k) + \bar{q}_n^i(y^k)^T(y - y^k)$ for all allocation vectors y, and all i. We want y to satisfy $\bar{g}_n(y) \geqslant 0$ and therefore it is necessary that

$$0 \leqslant \bar{g}_n^i(y^k) + \bar{q}_n^i(y^k)^T(y - y^k), \tag{4}$$

for all i.

In the next section we show how to use the subgradients in a cutting plane algorithm to solve the sample average approximation (2).

3. The cutting plane method

In this section we present a cutting plane algorithm for solving the sample average approximation (2). We select a fixed sample size at the beginning of the algorithm and use the same sample (common random numbers) in each iteration. This minimizes the effect of sampling in that we only work with one function $\bar{g}_n$ instead of getting a new $\bar{g}_n$ function in each iteration, which could, for example, invalidate the concavity assumption.

The typical cutting plane algorithm for (2) works as follows. We relax the nonlinear service level constraints to convert the call center staffing problem into a linear integer problem. We solve the linear integer problem and run a simulation with the staffing levels obtained from the solution. If the service levels meet the service level constraints as approximated by the sample average then we stop with an optimal solution to (2). If a service level constraint is violated then we add a linear constraint to the relaxed problem that eliminates the current solution but does not eliminate any feasible solutions to the sample average approximation.

Our algorithm fits the framework of Kelley's cutting plane method (Kelley, Jr., 1960). It differs from the traditional description of the algorithm only in that we use a stimulation to generate the cuts and evaluate the function values instead of having an algebraic form for the function and using analytically determined gradients to generate the cuts. Nevertheless, we include a proof of convergence of our cutting plane method, since its statement is specific to our algorithm and it makes the results clearer.

The relaxed problem for (2) that we solve in each iteration is

$$\begin{aligned} &\min\ c^T x \\ &\text{subject to} \quad Ax \geqslant y, \\ &\phantom{\text{subject to}} \quad D_k y \geqslant d_k, \\ &\phantom{\text{subject to}} \quad x \in X, \\ &\phantom{\text{subject to}} \quad x, y \geqslant 0 \text{ and integer.} \end{aligned} \tag{5}$$

We replaced the constraints $\bar{g}_n(y) \geqslant 0$ with linear constraints $D_k y \geqslant d_k$. The subscript k indicates the iteration number in the cutting plane algorithm. The constraint set $D_k y \geqslant d_k$ is initially empty but we add more constraints to it as the algorithm evolves.

At iteration k we solve an instance of (5) to obtain the solution pair (x^k, y^k). For the server allocation vector y^k we run a simulation to calculate $\bar{g}_n(y^k)$. If we find that the service level is unacceptable, i.e., if $\bar{g}_n^i(y^k) < 0$ for some i, then we add the constraint (4) to the set $D_k y \geqslant d_k$, i.e., we add the component $-\bar{g}_n^i(y^k) + \bar{q}_n^i(y^k)^T y^k$ to d_k and the row vector $\bar{q}_n^i(y^k)^T$ to D_k. We add a constraint for all periods i where the service level is unacceptable. Otherwise, if the service level is acceptable in all periods then we terminate the algorithm with an optimal solution to the sample average approximation (2).

Algorithm 1.

Initialization. Generate n independent realizations from the distribution of ξ. Let $k \leftarrow 1$, D_1 and d_1 be empty.

Step 1. Solve (5) and let (x^k, y^k) be an optimal solution.

Step 1a. Stop with an error if (5) was infeasible.

Step 2. Run a simulation to obtain $\bar{g}_n(y^k)$.

Step 2a. If $\bar{g}_n(y^k) \geqslant 0$ then stop. Return (x^k, y^k) as an optimal solution.

Step 3. Compute, by simulation, $\bar{q}_n^i(y^k)$ for all i for which $\bar{g}_n^i(y^k) < 0$, and add the cuts (4) to D_k and d_k.

Step 4. Let $d_{k+1} \leftarrow d_k$ and $D_{k+1} \leftarrow D_k$. Let $k \leftarrow k+1$. Go to step 1.

It is usually not necessary to store the n independent realizations referred to in the initialization phase. Instead, we only need to store a few numbers, called seeds, and reset the random number generators (streams) in the simulation with the seeds at the beginning of each iteration. See Law and Kelton (2000) for more details on this approach to using common random numbers. To speed up the algorithm it is possible to start with D_1 and d_1 nonempty. Ingolfsson and Cabral (2002) developed, for example, lower bounds on y. They point out that if there is an infinite number of servers in all periods except period i and if $\tilde{y}_i$ is the minimum number of employees required in period i in this setting so that the service level in period i is acceptable, then $y_i \geqslant \tilde{y}_i$ for all y satisfying $g(y) \geqslant 0$. We could select D_1 and d_1 to reflect such lower bounds.

If the algorithm terminates in step 1a then the sample average approximation is infeasible. That could be due to either a sampling error, i.e., the sample average approximation does not have any feasible points even though the original problem is feasible, or that the original problem is infeasible. As a remedy, either the sample size should be increased, or the original problem should be reformulated, e.g., the acceptable service level should be lowered, or more employees should be allocated (expand X).

In the algorithm above we solve an integer linear program and add constraints to it in each iteration until we terminate. The integer linear problem always has a larger feasible region than the sample average approximation (2), so $c^T x^k \leqslant c^T x^{k+1} \leqslant c^T x_n^*$, where (x_n^*, y_n^*) is an optimal solution for (2). An important question is whether $\lim_{k\to\infty} c^T x^k = c^T x_n^*$. The following theorem answers this question in the positive.

Theorem 1.

1. The algorithm terminates in a finite number of iterations.
2. Suppose that each component of $\bar{g}_n$ is concave in y. Then the algorithm terminates with an optimal solution to (2) if and only if (2) has a feasible solution.

Proof. 1. Y is a finite set and it is therefore sufficient to show that no point in Y is visited more than once. Suppose that the algorithm did not terminate after visiting point y^t. That means that $\bar{g}_n(y^t) \ngeqslant 0$ and we added one or more cuts of the form

$$0 \leqslant \bar{g}_n^i(y^t) + \bar{q}_n^i(y^t)^T(y - y^t), \quad i \in \{1, \ldots, p\}$$

to (5). Suppose that $y^k = y^t$, for some $k > t$. Since y^k is the solution for (5) at step k it must satisfy the cuts added at iteration t, i.e., $0 \leqslant \bar{g}_n^i(y^t) + \bar{q}_n^i(y^t)^T(y^k - y^t) = \bar{g}_n^i(y^t)$, which is a contradiction because this constraint was added since $\bar{g}_n^i(y^t) < 0$. Hence, we visit a new point in the set Y in each iteration and thus the algorithm terminates in a finite number of iterations.

2. Suppose first that (2) does not have a feasible solution. Then no $y \in Y$ satisfies $\bar{g}_n(y) \geqslant 0$. The algorithm only visits points in Y, so the optimality condition in step 2a is never satisfied. Since the algorithm terminates in a finite number of iterations it must terminate with the relaxed problem being infeasible. Suppose now that (2) is feasible. The problem (5) solved in step 1 is a relaxed version of (2) since $\bar{g}_n$ is concave, so (5) is feasible in every iteration. Therefore, the algorithm terminates in step 2a with (x^k, y^k) as the solution. But $\bar{g}_n(y^k) \geqslant 0$ by the termination criteria, so it is an optimal solution to (2). □

The theorem above states that we terminate with an optimal solution to the sample average approximation so long as one exists. In the next section, we discuss the convergence of that solution to an optimal solution to the original problem (1) as the sample size n increases.

4. Convergence of solutions of the sample average approximation to solutions of the original problem

We have established that the cutting plane algorithm will identify an optimal solution of the problem (2). The problem (2) was formed by approximating the expected service level constraints of problem (1), and we will next investigate if solutions of the sample average approximation converge to a solution of the original problem. We show, by using the strong law of large numbers (SLLN), that the set of optimal solutions of the sample average approximation is a subset of the set of optimal solutions for the original problem w.p. 1 as the sample size gets large. Furthermore, we show that the probability of this event approaches 1 exponentially fast when we increase the sample size. These results require the existence of at least one optimal solution for the original problem to satisfy the expected service level constraints with strict inequality, but this regularity condition can be easily justified for practical purposes as will be discussed later.

4.1. Almost sure convergence of optimal solutions of the sample average approximation

The results in this section may be established by specializing the results in Vogel (1994). We choose to provide direct proofs in this section for 3 main reasons:

1. The additional structure in our setting allows a clearer statement and proof of the results.
2. The proofs add important insight into why solving the sample average approximation is a sensible approach.
3. The proofs serve as an excellent foundation to develop an understanding of the "rate of convergence" results that follow in section 4.2.

The effect of the sampling on the optimization problem is to change the shape of the feasible region. It directly affects the service level constraint, so we will rewrite the problems (1) and (2) to make the effect more transparent and to make the proofs easier to read. First define

$$f(y) := \min_{\{x \geqslant 0 \text{ and integer: } x \in X,\ Ax \geqslant y\}} c^T x,$$

where $f(y) = +\infty$ if the set $\{x \geqslant 0 \text{ and integer: } x \in X,\ Ax \geqslant y\}$ is empty. Now we can rewrite problem (1) as

$$\begin{array}{ll} \min & f(y) \\ \text{subject to} & y \in Y, \\ & g(y) \geqslant 0 \end{array} \tag{6}$$

and its sample average approximation, which is equivalent to (2), as

$$\begin{array}{ll} \min & f(y) \\ \text{subject to} & y \in Y, \\ & \bar{g}_n(y) \geqslant 0. \end{array} \tag{7}$$

We are interested in the properties of the optimal solutions of (7) as the sample size n gets large. It turns out, by an application of the SLLN, that any optimal solution of (6) that satisfies $g(y) > 0$, i.e., $g^i(y) > 0$ for all i, is an optimal solution of (7) with probability 1 (w.p. 1) as n goes to infinity. We make a few more definitions before we prove this. Let

$$\bar{g}_\infty(y) := \lim_{n\to\infty} \bar{g}_n(y),$$
$$F^* := \text{the optimal value of (6)}$$

and define the sets

$$Y^* := \text{the set of optimal solutions to (6)},$$
$$Y_0^* := \{y \in Y^*\colon\ g(y) > 0\},$$
$$Y_1 := \{y \in Y\colon\ f(y) \leqslant F^*,\ g(y) \ngeqslant 0\},$$
$$Y_n^* := \text{the set of optimal solutions to (7)}.$$

Note that Y_1 is the set of solutions to (6) that have the same or lower cost than an optimal solution, and satisfy all constraints except the service level constraints. We are concerned with solutions in this set since they could be feasible (optimal) to the sample average approximation (7) if the difference between the sample average, $\bar{g}_n$, and g is sufficiently large. We show that when Y_0^* is not empty, $Y_0^* \subseteq Y_n^* \subseteq Y^*$ for all n large enough w.p. 1. We say that property $E(n)$ holds for all n large enough w.p. 1 if and only if $P[\exists N < \infty\colon\ E(n) \text{ holds } \forall n \geqslant N] = 1$. (Here N should be viewed as a random variable.) Sometimes such statements are communicated by saying that $E(n)$ holds *eventually*.

We start with two lemmas. The first one establishes properties of $\bar{g}_\infty(y)$ by repeatedly applying the SLLN. The second shows that solutions to (6) satisfying $g(y) > 0$, and infeasible solutions, will be feasible and infeasible, respectively, w.p. 1 for problem (7) when n gets large. The only condition $g(y)$ has to satisfy is that it has to be finite for all $y \in Y$. That assumption is easily justified by noting that the absolute value of each component of $g(y)$ is bounded by the expected number of arrivals in that period, which would invariably be finite in practice.

Even though we restrict attention to optimal solutions, the overall approach would not change if we wanted to prove that all "interior" feasible solutions for (6) are eventually feasible for (7) w.p. 1 and that all infeasible solutions for (6) are eventually infeasible for (7). This may lend some intuition, since it will then almost invariably be the case that the feasible region of the sample average approximation converges to the feasible region of the original problem and therefore the set of optimal solutions converges. Define

$$\|g\| = \max_{y\in Y} \big\|g(y)\big\|_\infty = \max_{y\in Y} \max_{i=1,\dots,p} \big|g^i(y)\big|.$$

Lemma 2.

1. Suppose that $\|g(y)\|_\infty < \infty$ for some fixed $y \in \mathbb{Z}^p$, $y \geqslant 0$. Then $\bar{g}_\infty(y) = g(y)$ w.p. 1.
2. Suppose that $\|g\| < \infty$ and $\Gamma \subseteq Y$. Then $\bar{g}_\infty(y) = g(y)\ \forall y \in \Gamma$ w.p. 1.

Proof. 1. The SLLN (see theorem 6.1 in Billingsley (1995)) gives $\bar{g}^i_\infty(y) = g^i(y)$ w.p. 1. So

$$P\big[\bar{g}_\infty(y) = g(y)\big] \geqslant 1 - \sum_{i=1}^{p} P\big[\bar{g}^i_\infty(y) \neq g^i(y)\big] = 1.$$

2. Note that

$$P\big[\bar{g}_\infty(y) = g(y)\ \forall y \in \Gamma\big] \geqslant 1 - \sum_{y\in\Gamma} P\big[\bar{g}_\infty(y) \neq g(y)\big] = 1$$

since Γ is finite. □

Lemma 3. Suppose that $\|g\| < \infty$. Then

1. $\bar{g}_n(y) \geqslant 0\ \forall y \in Y_0^*$ for all n large enough w.p. 1.
2. All $y \in Y_1$ are infeasible for the sample average approximation (7) for all n large enough w.p. 1.

Proof. 1. The result is trivial if Y_0^* is empty, so suppose it is not. Let

$$\epsilon = \min_{y\in Y_0^*} \min_{i\in\{1,\dots,p\}} \{g^i(y)\}.$$

Then $\epsilon > 0$ by the definition of Y_0^*. Let

$$N_0 = \inf\Big\{n_0\colon \max_{y\in Y_0^*} \|\bar{g}_n(y) - g(y)\|_\infty < \epsilon \quad \forall n \geqslant n_0\Big\},$$

with the infimum defined as $+\infty$ if the set is empty, and then $\bar{g}_n \geqslant 0\ \forall y \in Y_0^*\ \forall n \geqslant N_0$. Now, the set Y_0^* is a subset of Y, so $\lim_{n\to\infty} \bar{g}_n(y) = g(y)\ \forall y \in Y_0^*$ w.p. 1 by part 2 of lemma 2. Therefore, $N_0 < \infty$ w.p. 1.

2. The result is trivial if Y_1 is empty, so suppose it is not. Let

$$\epsilon = \min_{y\in Y_1} \max_{i\in\{1,\dots,p\}} \big\{-g^i(y)\big\}.$$

Then $\epsilon > 0$, since $g^i(y) < 0$, for at least one $i \in \{1,\dots,p\}\ \forall y \in Y_1$. Let

$$N_1 = \inf\big\{n_1\colon \ \big\|g(y) - \bar{g}_n(y)\big\|_\infty < \epsilon \quad \forall n \geqslant n_1\big\}$$

and then all $y \in Y_1$ are infeasible for (7) for all $n \geqslant N_1$. Now, the set Y_1 is a subset of Y, so $\lim_{n\to\infty} \bar{g}_n(y) = g(y)\ \forall y \in Y_1$ w.p. 1 by part 2 of lemma 2. Therefore, $N_1 < \infty$ w.p. 1. □

Lemma 3 shows that all the "interior" optimal solutions for the original problem are eventually feasible for the sample average approximation and remain so as the sample size increases. Furthermore, all solutions that satisfy the constraints that are common for both problems, but not the service level constraints, and have at most the same cost as an optimal solution, eventually become infeasible for the sample average approximation. Hence, we have the important result that for a large enough sample size an optimal solution for the sample average approximation is indeed optimal for the original problem.

Theorem 4. Suppose that $\|g\| < \infty$. Then $Y_0^* \subseteq Y_n^*$ for all n large enough w.p. 1. Furthermore, if Y_0^* is nonempty then $Y_0^* \subseteq Y_n^* \subseteq Y^*$ for all n large enough w.p. 1.

Proof. The first inclusion holds trivially if Y_0^* is empty, so assume that Y_0^* is not empty. On each sample path let $N = \sup\{N_0, N_1\}$, where N_0 and N_1 are the same as in lemma 3. When $n \geqslant N$ we know that all $y \in Y_0^*$ are feasible for (7) and that all $y \in Y_1$ are infeasible for (7). Hence, all $y \in Y_0^*$ are optimal for (7) and no $y \notin Y^*$ is optimal for (7) whenever $n \geqslant N$. Thus, $Y_0^* \subseteq Y_n^* \subseteq Y^*$ for all $n \geqslant N$. Finally, $P[N < \infty] = P[N_0 < \infty,\ N_1 < \infty] \geqslant P[N_0 < \infty] + P[N_1 < \infty] - 1 = 1$. □

Corollary 5. Suppose that $\|g\| < \infty$ and that (1) has a unique optimal solution, y^*, such that $g(y^*) > 0$. Then y^* is the unique optimal solution for (2) for all n large enough w.p. 1.

Proof. In this case $Y_0^* = Y^* = \{y^*\}$ and the result follows from the previous theorem. □

The conclusion of theorem 4 relies on existence of an "interior" optimal solution for the original problem. A simple example illustrates how the conclusion can fail if this requirement is not satisfied. Let ξ be a uniform random variable on $[-0.5, 0.5]$ and consider the following problem:

$$\begin{array}{ll} \min\ y & \\ \text{subject to} & y \geqslant \left|E[\xi]\right|, \\ & y \geqslant 0 \text{ and integer.} \end{array}$$

Then $y^* = 0$ for this problem since $E[\xi] = 0$. We form the sample average approximation by replacing $E[\xi]$ with $\bar{\xi}_n$, the sample average of n independent realizations of ξ. Then $0.5 > |\bar{\xi}_n| > 0$ w.p. 1 for all $n > 0$ and thus we get that $y_n^* = 1$ w.p. 1.

We mentioned earlier that the existence of an "interior" optimal solution is merely a regularity condition. In reality it is basically impossible to satisfy the service level constraints in any period exactly, since the feasible region is discrete. Even if this occurred, we could subtract an arbitrarily small positive number, say ε, from the right-hand side of each service level constraint and solve the resulting ε-perturbed problem. Then all solutions with $g^i(y) = 0$ for some i satisfy $g^i(y) > -\varepsilon$ and it is sufficient for the problem to have an optimal solution (not necessarily satisfying $g(y) > 0$) for theorem 4 to hold. This rationale also applies to the next subsection where we prove an exponential rate of convergence as the sample size increases.

4.2. *Exponential rate of convergence of optimal solutions of the sampled problems*

In the previous subsection, we showed that we can expect to get an optimal solution for the original problem (1) by solving the sample average approximation (2) if we choose a large sample size. In this section we show that the probability of getting an optimal solution this way approaches 1 exponentially fast as we increase the sample size. We use large deviations theory and a result due to Dai, Chen, and Birge (2000) to prove our statement. Vogel (1988) shows, under weaker conditions, that the feasible region of a sample average approximation for a chance constraint problem approaches the true feasible region at a polynomial rate and conjectures, without giving a proof, that an exponential rate of convergence is attainable under similar conditions to those we impose.

The following theorem is an intermediate result from theorem 3.1 in Dai, Chen, and Birge (2000).

Theorem 6. Let $H : \mathbb{R}^p \times \mathbb{Z} \to \mathbb{R}$ and assume that there exist $\gamma > 0$, $\theta_0 > 0$ and $\eta : \mathbb{Z} \to \mathbb{R}$ such that

$$\left|H(y, \xi)\right| \leqslant \gamma\eta(\xi), \qquad E\left[e^{\theta\eta(\xi)}\right] < \infty,$$

for all $y \in \mathbb{R}^p$ and for all $0 \leqslant \theta \leqslant \theta_0$, where ξ is an integer valued random variable. Then for any $\delta > 0$, there are $a > 0$, $b > 0$, such that for any $y \in \mathbb{R}^p$

$$P\left[\left|h(y) - \bar{h}_n(y)\right| \geqslant \delta\right] \leqslant a e^{-bn},$$

for all $n > 0$, where $h(y) = E[H(y, \xi)]$, and $\bar{h}_n(y)$ is a sample mean of n independent and identically distributed realizations of $H(y, \xi)$.

In our setting take $H(y, \xi) = G^i(y, \xi)$ and note that $|G^i(y, \xi)| \leqslant N^i(\xi)$, where N^i is the number of calls received in period i. If the arrival process is, for example, a (nonhomogeneous or homogeneous) Poisson process, which is commonly used to model incoming calls at a call center, then N^i satisfies the condition of theorem 6 since it is a Poisson random variable, which has a finite moment generating function.

Before we prove the exponential rate we prove a lemma that shows that for any n, $Y_0^* \subseteq Y_n^* \subseteq Y^*$, precisely when all the solutions in Y_0^* are feasible for the sample average approximation, and all infeasible solutions for (6) that are equally good or better, i.e., are in the set Y_1, are also infeasible for (7).

Lemma 7. Let $n > 0$ be an arbitrary integer. The properties

1. $\bar{g}_n(y) \geqslant 0 \; \forall y \in Y_0^*$, and
2. $\bar{g}_n(y) \ngeqslant 0 \; \forall y \in Y_1$

hold if and only if $Y_0^* \subseteq Y_n^* \subseteq Y^*$.

Proof. Suppose properties 1 and 2 hold. Then by property 1 all $y \in Y_0^*$ are feasible for (7) and the optimal value of (7) is at most F^*. By property 2 there are no solutions with a lower objective that are feasible for (7), so $Y_0^* \subseteq Y_n^*$. By property 2, no solutions outside Y^* with objective value equal to F^* are feasible for (7). Hence, $Y_0^* \subseteq Y_n^* \subseteq Y^*$.

Suppose $Y_0^* \subseteq Y_n^* \subseteq Y^*$. Then F^* is the optimal value for (7). Now, since all $y \in Y_0^*$ are optimal for (7) they are also feasible for (7) and property 1 holds. All $y \in Y_1$ are infeasible for (7) since $Y_n^* \subseteq Y^*$ and therefore property 2 holds. □

Theorem 8. Suppose $G^i(y,\xi)$ satisfies the assumptions of theorem 6 for all $i \in \{1,\dots,p\}$ and that Y_0^* is nonempty. Then there exist $\alpha > 0$, $\beta > 0$ such that

$$P\big[Y_0^* \subseteq Y_n^* \subseteq Y^*\big] \geqslant 1 - \alpha \mathrm{e}^{-\beta n}.$$

Proof. Define

$$\begin{aligned}
\delta_1 &:= \min_{y\in Y_0^*} \min_{i\in\{1,\dots,p\}} \big\{g^i(y)\big\},\\
i(y) &:= \arg \max_{i\in\{1,\dots,p\}} \big\{-g^i(y)\big\},\\
\delta_2 &:= \min_{y\in Y_1} \big\{-g^{i(y)}(y)\big\}, \quad \text{and}\\
\delta &:= \min\{\delta_1, \delta_2\}.
\end{aligned}$$

Here $\delta_1 > 0$ is the minimal amount of slack in the constraints "$g(y) \geqslant 0$" for any solution $y \in Y_0^*$. Similarly $\delta_2 > 0$ is the minimal violation in the constraints "$g(y) \geqslant 0$" induced by any solution $y \in Y_1$. Thus,

$$\begin{aligned}
&P\big[Y_0^* \subseteq Y_n^* \subseteq Y^*\big]\\
&\quad= P\big[\bar{g}_n(y) \geqslant 0\ \forall y \in Y_0^*,\ \bar{g}_n(y) \ngeqslant 0\ \forall y \in Y_1\big] && (8)\\
&\quad= 1 - P\big[\bar{g}_n(y) \ngeqslant 0 \text{ for some } y \in Y_0^* \text{ or } \bar{g}_n(y) \geqslant 0 \text{ for some } y \in Y_1\big]\\
&\quad\geqslant 1 - \sum_{y\in Y_0^*}\sum_{i=1}^{p} P\big[\bar{g}_n^i(y) < 0\big] - \sum_{y\in Y_1} P\big[\bar{g}_n(y) \geqslant 0\big] && (9)\\
&\quad\geqslant 1 - \sum_{y\in Y_0^*}\sum_{i=1}^{p} P\big[\big|\bar{g}_n^i(y) - g^i(y)\big| \geqslant \delta\big]\\
&\qquad - \sum_{y\in Y_1} P\big[\big|\bar{g}_n^{i(y)}(y) - g^{i(y)}(y)\big| \geqslant \delta\big] && (10)\\
&\quad\geqslant 1 - \sum_{y\in Y_0^*}\sum_{i=1}^{p} a_i \mathrm{e}^{-b_i n} - \sum_{y\in Y_1} a_{i(y)} \mathrm{e}^{-b_{i(y)} n} && (11)\\
&\quad\geqslant 1 - \alpha \mathrm{e}^{-\beta n}.
\end{aligned}$$

Here

$$\alpha = |Y_0^*| \sum_{i=1}^{p} a_i + \sum_{y\in Y_1} a_{i(y)} \quad \text{and} \quad \beta = \min_{i\in\{1,\dots,p\}} b_i,$$

where $|Y_0^*|$ is the cardinality of the set Y_0^*. Equation (8) follows by lemma 7. Equation (9) is Boole's inequality. Equation (10) follows since $P[\bar{g}_n(y) \geqslant 0] \leqslant P[\bar{g}_n^{i(y)}(y) \geqslant 0]$ and $g^i(y) \geqslant \delta_1 \geqslant \delta$ for $y \in Y_0^*$ and $g^{i(y)}(y) \geqslant \delta_2 \geqslant \delta$ for $y \in Y_1$. Equation (11) follows from theorem 6. □

The case where Y_0^* is empty but Y^* is not would almost certainly never arise in practice. But in such a case one can solve an ε-perturbation of (2) as described at the end of section 4.1, and the results of theorem 8 hold for $0 < \varepsilon < \delta$.

5. Numerically checking the concavity of a function

The success of the cutting plane algorithm relies on concavity of each component of the service level function $\bar{g}_n$. If a component of $\bar{g}_n$ is not concave, then the algorithm may "cut off" a portion of the feasible set and terminate with a nonoptimal solution. In each iteration of the algorithm we obtain new information about $\bar{g}_n$. To improve the robustness of the algorithm, we would like to ensure that the information we receive is consistent with the notion that each component of $\bar{g}_n$ is concave.

There are 2 cases to consider. The first is where the vectors $\bar{q}_n^i(y)$ as returned by the simulation are guaranteed to be subgradients of $\bar{g}_n^i$ *if* $\bar{g}_n^i$ is concave. For example, this would occur if the vectors were exact gradients of the function $\bar{g}_n^i$ at y (assuming that it had a differentiable extension to $\mathbb{R}^p$ from $\mathbb{Z}^p$). In this case there is an easy test for nonconcavity, as we will see. The second case, that appears more likely to occur in practice, is where the vectors $\bar{q}_n^i(y)$ are obtained using some heuristic, and are therefore not guaranteed to be subgradients, even if $\bar{g}_n^i$ is indeed concave. In this case, we may decide to disregard some of the potentially-unreliable "subgradient" information and focus only on the function values themselves. (This setting may also be useful if one does not have "subgradient" information at all points, as arises using the finite-differencing heuristic mentioned earlier. When evaluating the "subgradient" at y, we also compute the function value, *but not gradient information*, at points of the form $y + e_i$ where e_i is the usual ith basis vector.) If the function values themselves are consistent with the notion that the function is concave, then we may view our heuristically-derived "subgradients" with some suspicion, and even drop some of them from the optimization. An alternative would be to attempt to restrict the feasible region to a region where the functions are concave. We view the analysis of the cutting plane algorithm under these conditions as beyond the scope of this paper, partly because it is then possible that we then need to deal with the usual difficulties of nonconvex optimization. If the function values alone suggest nonconcavity, then the algorithm results should be viewed with some caution. Indeed, values reported as optimal by the algorithm could, in this case, be nonoptimal. The ability to detect when the key assumption of the cutting plane algorithm may not apply is, we believe, a strength of our approach.

Of course, one may either implement a check for nonconcavity either inline on each iteration of the cutting plane algorithm, or after the algorithm halts, or not at all. The choice depends on how conservative one wishes to be, and is therefore application dependent, and so we do not enter into a discussion of which approach to take here.

To simplify the presentation, let us consider the concavity of a real-valued function $f : \mathbb{R}^p \to \mathbb{R}$ instead of $\bar{g}_n^i$. Hopefully no confusion will arise since the previously-defined function f plays no role in this section. We assume that we are given a set of points $y^1, y^2, \ldots, y^k \in \mathbb{R}^p$ and their corresponding function values

$f(y^1), f(y^2), \ldots, f(y^k)$. The tests below allow one to conclude that either f is nonconcave, or that there exists a concave function that matches the given function values. Of course, the tests cannot conclude that f is concave unless they examine all points in its domain, so that the conclusions that these tests reach are the best possible in that sense.

5.1. Concavity check with function values and "subgradients"

Suppose that we know the vectors $q(y^1), q(y^2), \ldots, q(y^k)$ in addition to the set of points and their function values. Here $q(y^v)$ should have the property that *if* f is concave, then $q(y^v)$ is a subgradient at y^v $(v = 1, \ldots, k)$. If they are in fact subgradients then they need to satisfy (3), i.e., all k points must lie below the k hyperplanes defined by the $q(y^v)$'s and the corresponding function values. This means that for each point y^v, $v \in \{1, \ldots, k\}$, we must check that

$$f(y^w) \leqslant f(y^v) + q(y^v)^T (y^w - y^v) \quad \forall w \in \{1, \ldots, k\}. \tag{12}$$

If this inequality is violated by some v and w, then we conclude that f is not concave in y. Otherwise, the known values of f do not contradict the concavity assumption and

$$h(y) := \inf_{v \in \{1,\ldots,k\}} f(y) + q(y^v)^T (y - y^v)$$

is a concave function (see theorem 5.5 in Rockafellar (1970)), such that $h(y^w) = f(y^w)$ $\forall w \in \{1, \ldots, k\}$. In other words if (12) holds $\forall v \in \{1, \ldots, k\}$ then a concave function exists that agrees with the observed function values and "subgradients" $q(y^v)$, $v = 1, \ldots, k$.

When this test is implemented in the framework of algorithm 1, where in each iteration k we obtain y^{k+1}, $\bar{g}_n^i(y^{k+1})$ and $\bar{q}_n^i(y^{k+1})$, we need only check that (for each period i) the new point lies below all the previously defined hyperplanes and that all previous points lie below the hyperplane defined by the new "subgradient."

5.2. Concavity check with function values only

Now consider the case when only f is known at a finite number of points.

We want to know whether or not there is a concave function, say h, which passes through f at all the given points. If such a function does not exist then we conclude that f is not concave. (This problem appeared in Murty (1988, p. 539).)

We present a method where we solve a linear program (LP) and draw our conclusions based on the results of the LP. The idea behind this method is that if a one-dimensional function is concave then it is possible to set a ruler above each point and rotate it until the function lies completely below the ruler. This can also be done when dealing with functions of higher dimensions, and then the ruler takes the form of a plane $(p = 2)$ or a hyperplane $(p > 2)$.

The LP changes the given function values so that a supporting hyperplane for the convex hull of the points can be fitted through each point. The objective of this LP is to minimize the change in the function values that needs to be made to accomplish this goal.

If the changes are measured in the L_1- or L_∞-norm then the objective function is linear. The LP also gives an idea of how far, in some sense, the function is from being concave if a concave function cannot be fitted through the given points. If a concave function can be fitted then the LP will return such a function, namely the pointwise minimum of the hyperplanes computed by the LP.

It is most straightforward to use the L_1-norm to measure the changes in the function values. Then the LP can be formulated as follows:

$$\begin{aligned}
&\min \sum_{v=1}^{k} |b_v| \\
&\text{subject to} \\
&\quad a_{0v} + \left(a^v\right)^T y^v = f\left(y^v\right) + b_v \qquad \forall v \in \{1, \ldots, k\}, \\
&\quad a_{0v} + \left(a^v\right)^T y^w \geqslant f\left(y^w\right) + b_w \qquad \forall v \in \{1, \ldots, k\}\ \forall w \in \{1, \ldots, k\},\ w \neq v.
\end{aligned} \tag{13}$$

To linearize the objective function we adopt the standard trick of writing $b_v = b_v^+ - b_v^-$ and replace $|b_v|$ with $b_v^+ + b_v^-$, where b_v^+ and b_v^- are nonnegative. The decision variables are:

$a_{0v} \in \mathbb{R}$, $v \in \{1, \ldots, k\}$:	intercepts of the hyperplanes,
$a^v \in \mathbb{R}^p$, $v \in \{1, \ldots, k\}$:	slopes of the hyperplanes and
$b_v^+, b_v^- \in \mathbb{R}$, $v \in \{1, \ldots, k\}$:	change in the function values.

The number of variables in this LP is $k(p + 1) + 2k = k(p + 3)$ and the number of constraints is $k + k(k - 1) = k^2$. We could split the LP up into k separate linear programs if that would speed up the computations, as might occur if we could run them on multiple processors in parallel, or if the LP solver was unable to detect the separable structure in this problem and exploit it. Here, the vth separate linear program tries to fit a hyperplane through the point $(y^v, f(y^v))$ that lies above all other points.

The LP is always feasible, since a feasible solution is given by $a^v = 0$, $a_{0v} = 0$ and $b_v = -f(y^v)$ for all $v \in \{1, \ldots, k\}$. It is also bounded below by 0, since the objective function is a sum of absolute values. Therefore, this problem has a finite minimum. If the minimum value is 0, then the function defined by

$$h(y) := \inf_{v=1,\ldots,k} a_{0v} + \left(a^v\right)^T y$$

is concave and $f(y^v) = h(y^v)$ for all $v \in \{1, \ldots, k\}$. On the other hand, if f is indeed concave, then there exists a subgradient at every point of f (see theorem 3.2.5 in Bazaraa, Sherali, and Shetty (1993)) and hence the constraints of the LP can be satisfied with $b_v = 0$ for all $v \in \{1, \ldots, k\}$. We have proved the following result.

Theorem 9. Consider the LP (13).

1. If the optimal objective value of the LP is 0 then there exists a concave function $h(y)$ such that $h(y^v) = f(y^v)$ for all $v \in \{1, \ldots, k\}$.
2. If f is concave then the optimal objective value of the LP is 0.

So we see that a necessary condition for f to be concave is that the optimal objective value of the LP (13) is zero. Thus we have the following corollary.

Corollary 10. If the optimal objective value of the LP (13) is positive, then f is not concave.

Note that the hyperplanes obtained from the LP are generally not subgradients of f, so we cannot use them in algorithm 1 as such. Hence, we have to solve this LP after step 2 in each iteration, or as a check after the algorithm terminates. Given the computational demands of the cutting plane algorithm, repeatedly solving this LP in each iteration does not represent a significant increase in computational effort.

6. Computational study

In this section we present a small numerical example that showcases the properties of our method. The example is far from being a realistic representation of a call center, but captures many issues in setting call center staffing levels. We will study 3 aspects of the problem in the context of the example:

1. Convergence of the cutting plane algorithm and the quality of the resulting solution.
2. Dependence of the service level in one period on staffing levels in other periods. This is of particular practical interest since traditional methods assume independence between periods.
3. Concavity of $\bar{g}_n(y)$.

Our implementation creates the integer programs (5) in AMPL and uses the CPLEX solver to solve them in step 1 of the algorithm, and a simulation model built in ProModel to perform steps 2 and 3. We used Microsoft Excel to pass data between the simulation and optimization components and to run the iterations of the algorithm. The implementation was exactly as described in algorithm 1 except for the initialization, where we started with y^1 at the lower bounds described in section 3 instead of starting with D_1 and d_1 empty.

6.1. Example

We consider an $M(t)/M/s(t)$ queue with $p = 5$ periods of equal length of 30 minutes. We let the service rate be $\mu = 4$ customers/hour. The arrival process is a nonhomogeneous Poisson process with the arrival rate a function of the time t in minutes equal to $\lambda(t) = \lambda(1 - |t/150 - 0.65|)$, i.e., the arrival rate is relatively low at the beginning of the first period, then increases linearly at rate λ until it peaks partway through the fourth period and decreases at rate λ after that. We set $\lambda = 120$ customers/hour, which makes the average arrival rate over the 5 periods equal 87.3 customers/hour.

Table 1
The iterates of the algorithm and the resulting service level function values and their 95% confidence intervals (CI). $f(y^k)$ is the objective value at y^k.

k	y^k					$\bar{g}_{100}(y^k) \pm$ 95% CI half width (% of calls received that are answered in less than 90 sec.)					$f(y^k)$
1	11	19	27	30	29	0.4 ± 1.0 (81.5%)	-1.1 ± 2.2 (77.2%)	-1.8 ± 3.1 (76.6%)	-3.5 ± 3.2 (73.8%)	-2.1 ± 2.3 (75.5%)	125.0
2	11	21	27	33	29	0.5 ± 0.9 (81.9%)	3.4 ± 1.5 (88.5%)	0.2 ± 2.6 (80.5%)	4.1 ± 2.2 (87.4%)	-0.3 ± 2.7 (79.4%)	127.5
3	11	21	27	34	29	0.5 ± 0.9 (81.9%)	3.4 ± 1.5 (88.5%)	0.4 ± 2.6 (80.7%)	5.8 ± 1.8 (90.4%)	0.0 ± 2.6 (80.0%)	128.0

The goal is to answer 80% of received calls in each period in less than 90 seconds. The customers form a single queue and are served on a first come first serve basis. If a server is still in service at the end of a period it finishes that service before becoming unavailable. For example, if there are 8 busy servers at the end of period 3 and period 4 only has 6 servers then the 8 servers will continue to serve the customers already in service, but the next customer in the queue will not get service until 3 customers have finished service.

There are 6 permissible tours, including 4 tours that cover 2 adjacent periods, i.e., periods 1 and 2, 2 and 3, 3 and 4, and finally 4 and 5. The remaining 2 tours cover only one period, namely the first and the last. The cost of the tours covering 2 periods is \$2 and the single period tours cost \$1.50.

6.2. *Results*

We selected a sample size of $n = 100$ for running the algorithm. The lower bounds on y are depicted in the row $k = 1$ in table 1. Note that the staffing levels at the lower bounds result in an unacceptable level of service and thus a method which would treat the periods independently, would give an infeasible solution, since the service level is as low as 73.8% in period 4. The algorithm terminates after only 3 iterations with an optimal solution to the sample average approximation. To verify that this is indeed an optimal solution we ran a simulation for all staffing levels that have lower costs than the optimal solution and satisfy the initial lower bounds. None of these staffing levels satisfied $\bar{g}_{100}(y) \geqslant 0$, so the solution returned by the algorithm is the optimal solution for the sample average approximation. By including the 95% confidence interval we get information about the quality of the solution as a solution of the original problem. In the example, the confidence intervals in periods 1, 3 and 5 cover zero, which is a concern since we cannot say with conviction that our service level is acceptable in those periods. To get a better idea of whether the current solution is feasible for the original problem we calculated $\bar{g}_{999}(y^3) = (0.5 \pm 0.3, 3.0 \pm 0.5, 2.3 \pm 0.7, 5.1 \pm 0.7, 0.0 \pm 0.8)^T$, so we are more confident that the service levels in periods 1 and 3 are acceptable. The service level in period 5 is close to being on the boundary, hence our difficulty in determining

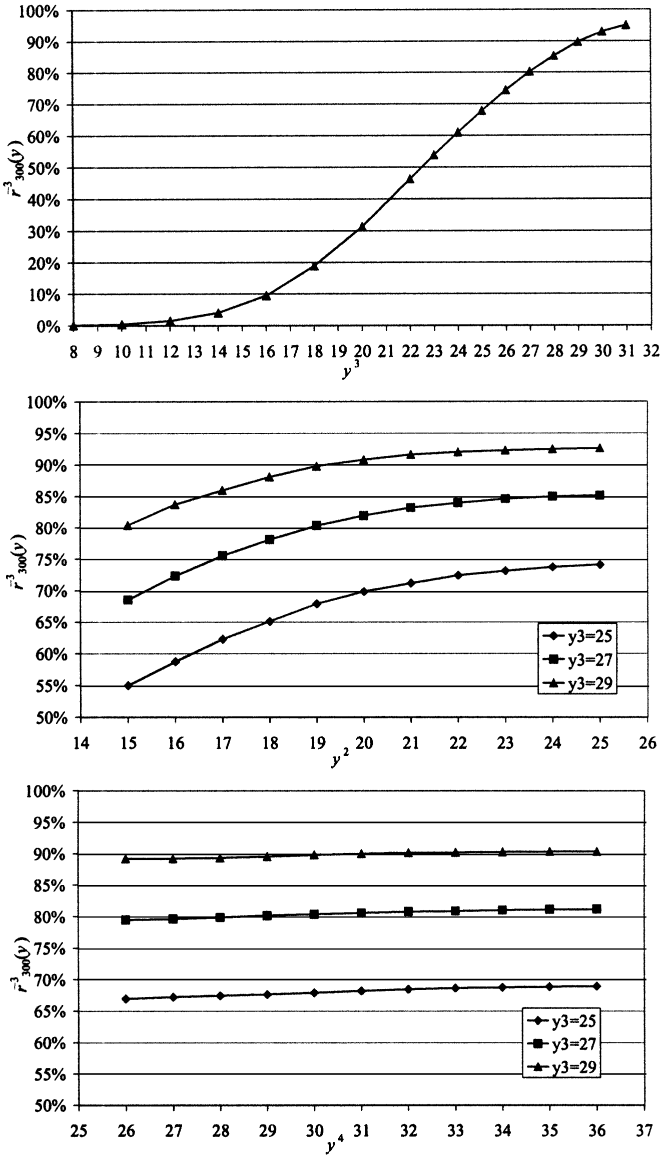

Figure 1. Dependence of staffing levels on the service level in period 3 of the example in section 6.1.

whether the solution is feasible or not. From a practical standpoint, if we are infeasible, then we are so close to being feasible that it probably is of little consequence.

We already noted that there is dependence between periods. To investigate the dependence further we calculated $\bar{r}^3_n(y)$, the percentage of calls received in period 3 answered in less than 90 seconds, i.e., $\bar{r}^3_n(y) := \sum_{d=1}^n S^3(y, \xi^d) / \sum_{d=1}^n N^3(y, \xi^d)$. We chose period 3 to demonstrate how the service level depends on staffing level in both the period before and after. Figure 1 illustrates this point. The graphs show the service level

Table 2
Concavity study. Low, medium and high staffing levels in each period and the optimal values of (13).

Period	1	2	3	4	5
Low	10	18	26	29	29
Medium	12	20	28	31	31
High	14	22	30	33	33
Optimal value	0.0	0.0	$4.0 \cdot 10^{-3}$	$1.5 \cdot 10^{-2}$	$5.7 \cdot 10^{-4}$

in period 3 as a function of the staffing level in period 3 (1.a), period 2 (1.b) and period 4 (1.c) when the staffing levels in other periods are fixed. The service level depends more on the staffing level in the period before than the period after as could be expected. That is because a low staffing level in an earlier period results in a queue buildup, which increases waiting in the next period. The reason why the staffing level in a later period affects the service level in an earlier period is that customers that called in the earlier period may still be waiting at the beginning of the next period and thus receive service earlier if there are more servers in that period. We noted dependence between periods as far apart as from the first period to the last. Figure 1 also supports the concavity assumption of the service level function when y is within a region of reasonable values, i.e., at least satisfies some lower bounds. It is, however, clear that the service level function looks like an s-shaped function over a broader range of y's as pointed out by Ingolfsson and Cabral (2002). That would not be problematic if one were to include the aforementioned lower bounds on y and if the concavity assumption holds for all y above the lower bounds.

We also performed a separate concavity check based on the method in section 5.2. In an effort to demonstrate these ideas as clearly as possible we performed the concavity check *outside* the scope of the cutting plane algorithm itself, using a selection of points that appear reasonable from a performance standpoint. We used a sample size of 300 and calculated $\bar{g}_{300}(y)$ at 3 different staffing levels (labelled low, medium and high in table 2) for each period, i.e., at $3^5 = 243$ points. We solved the linear program (13) for each $\bar{g}^i_{300}(y)$, $i \in \{1, \ldots, 5\}$, and obtained the results in table 2. We see that the service level functions in periods 1 and do not violate the concavity assumption at the observed points. The other functions violate the concavity condition. The values of the b_v's, i.e., the changes needed to satisfy the concavity assumption are all small, as can be seen by the objective value. We examined the points at which nonconcavity was detected, and noted that they occurred when a change in staffing level in a different period was made. (It is a strength of the LP-based concavity check that we were able to discover a region where the nonconcavity was exhibited.) The service level in period 3 increased, for example, more when the staffing level in period 1 was increased from 12 to 14 than when it was increased from 10 to 12 at staffing levels 22 and 30 in periods 2 and 3, respectively. The reason for this violation of the concavity assumption is not obvious.

One possible explanation is that our measure of service quality is binary for each customer, so that "rounding" may contribute to the nonconcavity. To elaborate, in the above example it is possible that unusually many customers exceed the waiting time limit of 90 seconds by very little when there are 12 servers in period 1, so that the effect of adding servers at this staffing level is more than when servers are added at a lower level. We would expect such a "rounding" effect to be averaged out in a longer simulation. In fact, we increased the sample size to 999 (the maximum number of replications in ProModel 4.2) and calculated the service level at the problematic points. We discovered that the nonconcavity vanished. Therefore, we make the following conjecture.

Conjecture 11. For $M(t)/M/s(t)$ queues of the form considered here there exists a finite $y_0 \geqslant 0$ such that the service level function g is nondecreasing and concave in y in the region $y \geqslant y_0$. Furthermore, $\bar{g}_n$ is nondecreasing and concave in y in the region $y_0 \leqslant y \leqslant y_1$ for all n large enough w.p. 1, for any fixed $y_1 \geqslant y_0$.

7. Conclusions

In this paper we have shown that combining simulation and cutting plane methods is a promising approach for solving optimization problems in which some of the constraints can only be assessed through simulation. As a motivating example we studied the problem of minimizing staffing costs in call centers when traditional methods fail, either because of the characteristics of the problem, or if a detailed model of the call center dynamics are desired. We performed a computational study, which supports the use of the cutting plane method and demonstrates how it can be implemented.

There are several interesting directions for future research. We established the theoretical foundation of the method, but an obvious drawback of the method is the large computational effort required to solve realistic problems. More research (in progress) is needed to make this a practical method. In relation to the integer programs one should investigate integer programming algorithms that can utilize the special structure of the relaxed problems solved in each iteration and consider allowing approximate solutions of the IPs, especially in early stages of the algorithm. One might consider Lagrangian relaxation techniques for solving these problems, still in the context of simulation and optimization, since we are approximating the constraints.

Other areas of interest include coming up with methods for obtaining subgradients or improving the current heuristic of using finite differences. It would add to the robustness of the method to study the properties of the solutions when some of the conditions set forth in this paper, e.g., the concavity of the service levels, are violated.

We only tested our algorithm on one simple example. It would be informative to run the algorithm on more complicated problems and include in the simulation model factors such as absenteeism and skill-based routing, not to mention implementing the algorithm in other types of service systems than call center staffing. We are currently pursuing many of these issues.

Acknowledgments

We would like to thank Ryan Anthony, Ernest Fung and Paul Rosenbaum, undergraduate students at Cornell University, for their contributions to the computational study. We would also like to thank the anonymous referees for their comments. This work was supported by National Science Foundation Grant DMI-9984717.

References

Akşin, O.Z. and P.T. Harker. (2001). "Modeling a Phone Center: Analysis of a Multichannel, Multiresource Processor Shared Loss System." *Management Science* 47(2), 324–336.

Bazaraa, M.S., H.D. Sherali, and C.M. Shetty. (1993). *Nonlinear Programming: Theory and Algorithms.* New York: Wiley.

Benders, J.F. (1962). "Partitioning Procedures for Solving Mixed-Variables Programming Problems." *Numerische Mathematik* 4, 238–252.

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

Birge, J.R. and F. Louveaux. (1997). *Introduction to Stochastic Programming.* Springer Series in Operations Research. New York: Springer.

Chen, B.P.K. and S.G. Henderson. (2001). "Two Issues in Setting Call Centre Staffing Levels." *Annals of Operations Research* 108(1), 175–192.

Chen, H. and B.W. Schmeiser. (2001). "Stochastic Root Finding via Retrospective Approximation." *IIE Transactions* 33(3), 259–275.

Dai, L., C.H. Chen, and J.R. Birge. (2000). "Convergence Properties of Two-Stage Stochastic Programming." *Journal of Optimization Theory and Applications* 106(3), 489–509.

Dantzig, G.B. (1954). "A Comment on Edie's "Traffic Delays at Toll Booths"." *Operations Research* 2(3), 339–341.

Dyer, M.E. and L.G. Proll. (1977). "On the Validity of Marginal Analysis for Allocating Servers in $M/M/c$ Queues." *Management Science* 23(9), 1019–1022.

Glasserman, P. (1991). *Gradient Estimation Via Perturbation Analysis.* Norwell, MA: Kluwer.

Green, L., P.J. Kolesar, and J. Soares. (2001). "Improving the SIPP Approach for Staffing Service Systems that Have Cyclic Demands." *Operations Research* 49(4), 549–564.

Healy, K. and L.W. Schruben. (1991). "Retrospective Simulation Response Optimization." In B.L. Nelson, W.D. Kelton, and G.M. Clark (eds.), *Proceedings of the 1991 Winter Simulation Conference,* pp. 901–906. Piscataway, NJ: IEEE.

Henderson, S.G. and A.J. Mason. (1998). "Rostering by Iterating Integer Programming and Simulation." In D.J. Medeiros, E.F. Watson, J.S. Carson, and M.S. Manivannan (eds.), *Proceedings of the 1998 Winter Simulation Conference,* pp. 677–683. Piscataway, NJ: IEEE.

Higle, J.L. and S. Sen. (1991). "Stochastic Decomposition: An Algorithm for Two-Stage Stochastic Linear Programs with Recourse." *Mathematics of Operations Research* 16, 650–669.

Infanger, G. (1994). *Planning under Uncertainty: Solving Large-Scale Stochastic Linear Programs.* Danvers, MA: Boyd and Fraser.

Ingolfsson, A. and E. Cabral. (2002). "Combining Integer Programming and the Randomization Method to Schedule Employees." Research Report No. 02-1, University of Alberta.

Ingolfsson, A., M.A. Haque, and A. Umnikov. (2002). "Accounting for Time-Varying Queueing Effects in Workforce Scheduling." *European Journal of Operational Research* 139, 585–597.

Jennings, O.B., A. Mandelbaum, W.A. Massey, and W. Whitt. (1996). "Server Staffing to Meet Time-Varying Demand." *Management Science* 42(10), 1383–1394.

Kelley, J.E., Jr. (1960). "The Cutting-Plane Method for Solving Convex Programs." *Journal of the Society for Industrial and Applied Mathematics* 8(4), 703–712.

Koole, G. and E. van der Sluis. (2003). “Optimal Shift Scheduling with a Global Service Level Constraint.” *IIE Transactions* 35(11), 1049–1055.

Law, A.M. and W.D. Kelton. (2000). *Simulation Modeling and Analysis*, 3rd ed. Boston, MA: McGraw-Hill.

Mehrotra, A., K.E. Murphy, and M.A. Trick. (2000). “Optimal Shift Scheduling: A Branch-and-Price Approach.” *Naval Research Logistics* 47(3), 185–200.

Morito, S., J. Koida, T. Iwama, M. Sato, and Y. Tamura. (1999). “Simulation-Based Constraint Generation with Applications to Optimization of Logistic System Design.” In P.A. Farrington, H.B. Nembhard, D.T. Sturrock, and G.W. Ewans (eds.), *Proceedings of the 1999 Winter Simulation Conference,* pp. 531–536. Piscataway, NJ: IEEE.

Murty, K.G. (1988). *Linear Complementarity, Linear and Nonlinear Programming.* Berlin: Heldermann.

Robinson, S.M. (1996). “Analysis of Sample-Path Optimization.” *Mathematics of Operations Research* 21(3), 513–528.

Rockafellar, R.T. (1970). *Convex Analysis.* Princeton, NJ: Princeton University Press.

Rubenstein, R.Y. and A. Shapiro. (1993). *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method.* Chichester: Wiley.

Shapiro, A. and T. Homem-de-Mello. (2000). “On the Rate of Convergence of Optimal Solutions of Monte Carlo Approximations of Stochastic Programs.” *SIAM Journal on Optimization* 11(1), 70–86.

Thompson, G.M. (1997). “Labor Staffing and Scheduling Models for Controlling Service Levels.” *Naval Research Logistics* 44(8), 719–740.

van Slyke, R.M. and R. Wets. (1969). “L-Shaped Linear Programs with Applications to Optimal Control and Stochastic Programming.” *SIAM Journal on Applied Mathematics* 17(4), 638–663.

Vogel, S. (1988). “Stability Results for Stochastic Programming Problems.” *Optimization* 19(2), 269–288.

Vogel, S. (1994). “A Stochastic Approach to Stability in Stochastic Programming.” *Journal of Computational and Applied Mathematics* 56, 65–96.