Proceedings of the 2003 Winter Simulation Conference
S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, eds.

USING SIMULATION TO APPROXIMATE SUBGRADIENTS OF CONVEX PERFORMANCE MEASURES IN SERVICE SYSTEMS

Júlíus Atlason
Marina A. Epelman

Department of Industrial
and Operations Engineering
University of Michigan
Ann Arbor, MI 48109-2117, U.S.A.

Shane G. Henderson

School of Operations Research
and Industrial Engineering
Cornell University
Ithaca, NY 14853, U.S.A.

ABSTRACT

We study the problem of approximating a subgradient of a convex (or concave) discrete function that is evaluated via simulation. This problem arises, for instance, in optimization problems such as finding the minimal cost staff schedule in a call center subject to a service level constraint. There, subgradient information can be used to significantly reduce the search space. The problem of approximating subgradients is closely related to the one of approximating gradients and we suggest and compare how three existing methods for computing gradients via simulation, i.e., finite differences, the likelihood ratio method and infinitesimal perturbation analysis, can be applied to approximate subgradients when the variables are discrete. We provide a computational study to highlight the properties of each approach.

1 INTRODUCTION

Simulation is becoming an increasingly popular tool for optimization of complex systems. There are several different approaches to simulation optimization available (Fu 2002). Some of these are developed from classical gradient-based optimization techniques. For problems with non-differentiable functions one cannot use gradients. An alternative for convex non-differentiable functions is to compute subgradients and use them in place of gradients.

A subgradient of a convex function at a particular point is a normal vector to a hyperplane that passes through the corresponding point on the graph on the function and lies below the function graph everywhere else. The set of all subgradients at a point is called the subdifferential at the point. For convex functions the subdifferential is non-empty at all points. If the function is differentiable at a point then the gradient is the unique subgradient at that point. Although convexity and subgradients are only defined for functions with a continuous domain, we can still use the subgradients in the above sense for discrete functions.

Discrete functions arise naturally in many optimization problems. One such problem, and the one that motivates our discussion, is the problem of scheduling agents (employees) in an inbound call center over a specific planning horizon (e.g., a week) at the lowest cost possible while at the same time maintaining a minimum level of service. We consider a somewhat simplified version of the real world problem for the purpose of developing the theory. We assume, for instance, that there is an unlimited number of trunk lines, one customer class, one type of agent and no abandonments. The call load typically varies throughout the planning horizon and therefore it is desirable to vary the number of agents accordingly. To do that we divide the planning horizon into smaller periods (e.g., 30 minutes in duration) and fix the number of agents in each period.

There is an extensive literature on staffing problems in call centers (Mandelbaum 2003). In many cases it is reasonable to model the call center using a queuing model for which an analytic expression for the service level function exists. For call centers that cannot be adequately analyzed in this way, simulation is a useful and popular tool. If the staffing level (number of agents) in one period does not affect the service level in other periods then it is relatively easy (given a mathematical model of the call center) to determine (or estimate) the number of agents required in each period to maintain a minimum level of service. It is possible, however, that the "best" staffing level based on this simplification of the call center is either suboptimal, or results in an unsatisfactory service level. This can occur, in particular, if the service times are relatively long with respect to the length of a period, or if there is great variability in the call load (Green, Kolesar, and Soares 2001). This makes an analytic approach using queuing models even less adequate, whereas a simulation model can encompass such interdependency between periods.

The difficulty with using simulation to solve this problem is that the number of alternatives can be very large. A cutting plane optimization algorithm (Atlason, Epelman, and Henderson 2003) counters this problem in an attempt to explore only a few alternatives en route to finding the optimal number of agents in each period. We start by formulating a linear integer program (IP) that excludes the service level constraints. The decision variables in the IP are the numbers of agents on each shift. The shift patterns are predetermined including constraints such as meal breaks, and a linear function is given that maps the number of agents assigned to each shift to the number of agents in each period. We solve the IP to obtain the lowest cost staffing level. We use a simulation model to evaluate the service level in all periods. If the service level is satisfactory we conclude that the current staffing level is optimal. Otherwise, we compute a (sub)gradient of the service level function and use it to create a new constraint (cut) in the IP. We re-solve the IP and repeat the process until we find an optimal solution or conclude that no solution is feasible. We use common random numbers for each simulation, so in fact the service level function is a sample average approximation (SAA) (see Kleywegt, Shapiro, and Homem-de-Mello (2001) for more detail on the SAA method).

The algorithm is guaranteed to converge to an optimal solution if one exists and if the service level constraints form a convex set. A common measure for level of service, and the one we use, is the percentage of calls that do not have to wait longer than a certain amount of time. It is reasonable to assume, at least for a range of staffing levels, that the service level increases with diminishing marginal returns as the number of agents increases, which suggests that the service level function is concave. Then the minimum service level constraint forms a convex set in the number of agents in each period. See Atlason, Epelman, and Henderson (2003) for additional discussion and references.

In this paper we discuss what is perhaps the most challenging part of the algorithm described above: Computation of the (sub)gradients of the SAA of the service level function. The service level function is a discrete function of the number of agents, so a gradient does not exist. In addition, we do not have a closed form expression of the function.

The problem of estimating gradients via simulation has received considerable attention over the past fifteen years. Among the most prominent approaches are the method of finite differences (FD), the likelihood ratio method (LR) (also called the score function method) and infinitesimal perturbation analysis (IPA). We study each of the FD, LR and IPA in an attempt to obtain a subgradient of the SAA service level function. We explore the advantages and disadvantages of each method for this particular problem.

The remainder of the paper is organized as follows. We formulate the subgradient approximation problem in Section 2, the FD method is discussed in Section 3, the LR method in Section 4 and IPA in Section 5. We include a computational study to further enhance the comparison in Section 6 and provide concluding remarks in Section 7.

2 SUBGRADIENT APPROXIMATION PROBLEM

In this section we study the service level function in more detail. Let p be the number of periods in the planning horizon and let $y \in \mathbb{Z}^p$ be the vector whose ith component denotes the number of agents in period i. Let ξ, a random vector, denote all the random quantities in the problem, i.e., the interarrival and service times in one planning horizon, and let $\xi^1, \ldots, \xi^n$ denote independent realizations of ξ. The service level in each period i is the fraction of calls received in that period answered within a certain amount of time τ. Therefore, there is one service level function for each period and the problem is to approximate a subgradient for each service level function. Fortunately, the value of the service level functions for particular staffing levels can be estimated for all periods simultaneously from a single simulation.

For an arbitrary period let $N(\xi)$ be the number of calls received in that period and let $S(y, \xi)$ be the number of those calls answered on time. The fraction of customers receiving adequate service in this period in the long run is then

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

If $E[N(\xi)] < \infty$ 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 $E[S(y, \xi)]/E[N(\xi)]$. Thus,

$$\frac{E[S(y, \xi)]}{E[N(\xi)]} \tag{1}$$

is a natural representation of the service level function (excluding the pathological case $E[N(\xi)] = 0$).

Note that the denominator in (1) does not depend on y, so the service level function has the same properties as $E[S(y, \xi)]$, the expected number of calls received in the period that are answered within time τ. The SAA of $s(y) \equiv E[S(y, \xi)]$ with sample size n is $\bar{s}(y; n) \equiv n^{-1} \sum_{d=1}^{n} S(y, \xi^d)$. Given the realizations $\xi^1, \ldots, \xi^n$, $\bar{s}(y; n)$ is a deterministic function of y. In the cutting plane algorithm we compute $\bar{s}(y; n)$ via simulation and also need a subgradient of $\bar{s}(y; n)$.

Under the assumption that $s(y)$ is a concave function of y it is also reasonable to assume that $\bar{s}(y; n)$ is concave in y, at least when the sample size, n, is large. Thus, at every point y^* there exists a subgradient $q(y^*)$ (some may prefer the term "supergradient" for a concave function) such that

$$\bar{s}(y; n) \le \bar{s}(y^*; n) + q(y^*)^T (y - y^*)$$

holds for any y and y^* in the range that $\bar{s}(y; n)$ is concave. The problem is therefore to compute such a vector $q(y^*)$ for a given y^*.

The finite difference method works directly with the discrete function $\bar{s}(y; n)$. For IPA and LR to work we must first approximate $\bar{s}(y; n)$ with a function of a continuous variable rather than the discrete variable y. For this problem, and queuing problems in general, a natural candidate is a function of the service rates in each period. When we use the service rates as the variables instead of the staffing levels we get a new function. The good news is that there are numerous studies that compare performance measures of queuing systems having the same total service rate but different number of servers. The most relevant results can be found in Chao and Scott (2000).

Let the service rate in period i be equal to μ_i and let $\mu = (\mu_1, \ldots, \mu_p)$. In the original problem $\mu_1 = \ldots = \mu_p$. Let $r(\mu; y)$ be the service level function as a function of the rates μ at staffing level y and let $s(y; \mu)$ be the original service level function with the service rates μ.

The functions r and s are indeed the same function. We choose to represent it as two functions to make clear the distinction that in the original problem we are interested in the service level as a function of the number of agents y for a fixed service rate vector μ, but to estimate a gradient we work with the service level as a function of the service rate μ for a fixed staffing level y. When we estimate the gradient at different staffing levels we are in fact working with a new function $r(\mu; y)$ parameterized by the staffing level y. By definition $r(\mu; y) = s(y; \mu)$ but, as an example, $r(2\mu; y) \ne s(2y; \mu)$ in general, so even if the functions agree for identical service rates and staffing levels then the effect of increasing the total service rate $(\mu_i y_i)$ in each period by changing the service rates is not the same as when the same change is accomplished by changing the staffing levels.

Suppose that we add one agent in period i. The effect of that is to increase the total service rate in period i by μ_i. If on the other hand we increase the service rate in period i by 1 then the total service rate in that period increases by y_i. This suggests that

$$s(y + e_i; \mu) - s(y; \mu) \approx \frac{\mu_i}{y_i} \frac{\partial r(\mu; y)}{\partial \mu_i}, \qquad (2)$$

where e_i is the ith unit vector in $\mathbb{R}^p$. We can use this approximation and simulation gradient estimation techniques to approximate the subgradient of $\bar{s}(y; n)$.

In Section 4 we study how to estimate $\partial r(\mu; y)/\partial \mu_i$ with the LR method and in Section 5 we use IPA. In the next section we compute

$$\bar{s}(y + 1; n, \mu) - \bar{s}(y + e_i; n, \mu) \quad \text{for } i = 1, \ldots, p \qquad (3)$$

and discuss the properties of (3) as a potential subgradient for $\bar{s}(y; n, \mu)$ at y.

3 FINITE DIFFERENCES

The simplest and perhaps the most intuitive method for estimating a gradient (or subgradient) when an expression for the function is unknown is by the method of finite differences. The FD method can easily be extended to discrete functions. There is a price to pay, however, for this ease of implementation. The number of simulations to get one gradient estimate is rather large and this method can fail to produce a subgradient even under rather stringent conditions on the service level function.

To estimate the partial derivative with respect to a continuous variable the function is evaluated at two different points. Then an estimate of the derivative at a value at or between these two values can be estimated by linear interpolation. When the variable is integer, as in the staffing problem, then the smallest difference between the two points is one.

Let $\bar{q}(y^*)$ be an estimate of a subgradient $q(y^*)$ of $\bar{s}(y; n)$ at y^*. The finite forward difference estimator of q at y^* is given by

$$\bar{q}_i(y^*) = \bar{s}(y + e_i; n) - \bar{s}(y; n)$$

for $i = 1, \ldots, p$. As we can see this estimator is easy to implement. To estimate the subgradient at the staffing level y, given $\bar{s}(y; n)$, simply run p simulations with the number of agents in period i increased by one in the ith simulation. This of course requires $p + 1$ simulations to get a subgradient estimate at a single point, but we also get the subgradients of the service level functions in the other periods from these $p + 1$ simulations.

When the FD method is used to estimate a gradient of a strictly convex or concave function of continuous variables at a single point, and if the function is indeed differentiable at that point, then for any given $\epsilon > 0$ there is a $\delta > 0$ such that the resulting estimator is a subgradient with respect to all points outside of a ball with radius ϵ centered at the point, so long as the difference is less than δ for each coordinate direction. One might think that this would also work at a point where the function is not differentiable.

That is not true in general. We demonstrate this with a simple example. Consider the function

$$f(y_1, y_2) = \begin{cases} y_1, & 0 \le y_1 \le y_2, \\ y_2, & 0 \le y_2 \le y_1. \end{cases}$$

The function f is concave. If we try to estimate a subgradient at any point y^* on the diagonal by the FD method we get $(0,0)^T$ as our estimate. Clearly this is not a valid subgradient since the hyperplane $h(y) = f(y^*) < f(y_1^* + a, y_1^* + b)$ for any positive numbers a and b.

What does this mean in the context of the staffing problem? If we look for the reason why the FD method failed for the function f above we see that the function f increases slower if we only change one variable at a time than if we change both variables. This would happen in the call center if there would be greater marginal benefit of adding one agent to two periods than the combined marginal benefit of adding one agent to the two periods separately. It seems reasonable to assume that this would not occur in a typical call center. Then, the marginal return on the service level of adding an agent in any period decreases every time we add an agent *regardless of* what period the previous agent was added to, i.e., assume f is submodular (Topkis 1998).

This is a stronger assumption than the concavity assumption we previously worked with, so do we get a subgradient by the FD method if the service level function is submodular? Again, the short answer to that question is no. We also demonstrate this via a simple example. Consider the points and their corresponding function values (numbers by the dots) depicted in Figure 1. That function has a subgradient at every point and is submodular. Nevertheless, the FD estimator fails to produce a subgradient at the point (0,1).

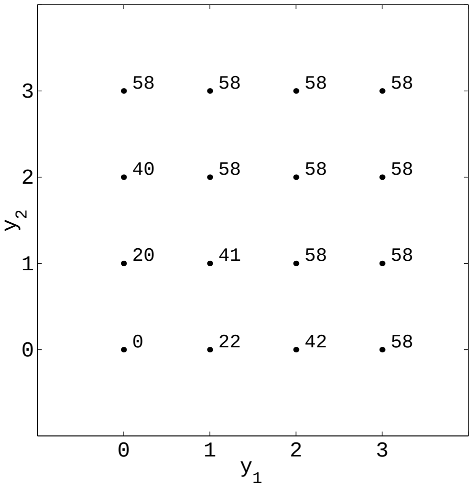

Figure 1: Submodular Function

We have shown by two examples that the FD method does not necessarily produce a subgradient of the service level function. There are still many other examples where the FD method will produce a subgradient. Consider, for instance, any point other than (0,1) of the function in Figure 1. For that point the FD method will indeed produce a subgradient. Therefore, and because of how easy the FD method is to implement, the FD method for obtaining a subgradient is a plausible approach when the number of periods p, i.e., input variables, is not too large.

The implication of having an invalid subgradient is adding an invalid cut which may “cut off” an optimal solution. Thus, we might terminate the algorithm with a suboptimal solution (or none at all). That is obviously a concern when the goal of the cutting plane algorithm is to find the best staffing levels. In other cases the underlying problem might be so complicated that obtaining a “good” solution is an appropriate goal that can be reached even though the “subgradients” are invalid.

4 LIKELIHOOD RATIO METHOD

Previously we saw that the FD method for approximating a subgradient of a convex or concave function can fail to produce a subgradient. Furthermore, if the number of input variables, p, is large then the computational requirements to obtain a single subgradient estimate are rather formidable. That motivates us to explore other options such as the likelihood ratio method. The LR method is an entirely different method. It is intended to estimate a gradient from a *single* simulation run of an expected value function by differentiating the elements of the expected value, i.e., the densities and the integrand.

The input variables in our setting are the number of agents in each period. We cannot differentiate with respect to these variables since they are discrete. We can, however, approximate a subgradient of the service level function by estimating a gradient of the service level as a function of the service rates as in Section 2. There we defined $r(\mu; y)$ as the expected service level function as a function of the service rates μ given the staffing levels y. If we let P_μ be the distribution of ξ as a function of the service rates μ then the expected service level can be written as

$$r(\mu; y) = \int \rho(\mu, \xi; y) dP_\mu(\xi) \qquad (4)$$

where $\rho(\mu, \xi; y)$ is the service level in a particular period as a function of the random elements in ξ and the service rates μ given the staffing level y. There are a few different ways to represent the service rates in the model. We can either let the sample path $\rho(\mu, \xi; y)$, or the distribution P_μ, or both depend on μ. In the next section, when we discuss IPA, we let only the sample path depend on μ. Here we let only the distribution P_μ depend on μ. The central idea of the likelihood ratio method is to rewrite (4) such that the distribution that is being integrated over does not depend on μ. To do that there must exist a distribution Q such that P_μ is absolutely continuous with respect to Q. In that case we can write

$$r(\mu; y) = \int \rho(\xi; y)\frac{dP_\mu(\xi)}{dQ(\xi)}dQ(\xi) \tag{5}$$

and then under certain conditions on P_μ, Q and ρ

$$\nabla_\mu r(\mu; y) = \int \rho(\xi; y)\nabla_\mu\left(\frac{dP_\mu(\xi)}{dQ(\xi)}\right)dQ(\xi). \tag{6}$$

To estimate the gradient at a single value of μ, say μ^*, Q can often be taken as P_{μ^*} (L'Ecuyer 1990, L'Ecuyer 1995).

Recall that ξ contains all the interarrival and service times of all calls in the planning horizon. Suppose that there are C calls in the planning horizon and let $a = (a_1, \ldots, a_C)$ be the arrival times of calls $1, \ldots, C$. Also let $x = (x_1, \ldots, x_C)$ denote the service times of the C calls. In this model we assume that the service time of each call is determined by the service rate in the period in which the call begins service, i.e., once a call enters service it is served at the same rate until it is completed. Thus, the period in which a call enters service depends on its arrival time, and arrival times and service times of all the previous calls.

This dependence potentially makes the distributions P_μ and Q difficult to compute. On the other hand, once $a_1, \ldots, a_j$ and $x_1, \ldots, x_{j-1}$ are known (as they would be in a simulation study) then it is relatively easy to determine x_j (and a_{j+1}, of course).

Thus, we first generate the interarrival times and then generate the service times, depending on which period they occur in. Let B_j be the time at which call j enters service, b_j be a realization of B_j and let $\pi(b_j)$ be the period containing b_j. Let $G(x; \mu_{\pi(b_j)})$ denote the distribution of the service time of call j. Verifying that the differentiation can be taken inside the integral in (6) is difficult in general. If, however, the service rate μ is a scale parameter of a family of distributions, i.e., there exist a random variable $\hat{X}$ such that $\hat{X}/\mu_{\pi(B_j)}$ has the same distribution as the service time of the jth call, then the problem simplifies somewhat. Thus, if we let X_j denote the random variable for the service time of the jth call then $X_j = \hat{X}_j/\mu_{\pi(B_j)}$ where $\hat{X}, \hat{X}_1, \hat{X}_2, \ldots$ are i.i.d. random variables and are independent of the period in which the corresponding call enters service. The gamma, Weibull and exponential distributions all have a scale parameter. If in addition we assume that G has a density g then the gradient of $r(\mu; y)$ at μ^* is

$$\nabla_\mu r(\mu; y)\Big|_{\mu=\mu^*} = \int \rho(a, x; y)\frac{\nabla_\mu \prod_{j=1}^{C} g(x_j; \mu_{\pi(b_j)})\Big|_{\mu=\mu^*}}{\prod_{j=1}^{C} g(x_j; \mu^*_{\pi(b_j)})}dP_{\mu^*}(a, x) \tag{7}$$

under appropriate conditions on the arrival process and on g. The left-hand side of (7) is fairly easy to estimate from a single simulation when the derivative of the density g can be computed.

The final step is to translate (7) into an approximation of the subgradient of the service level function $s(y; \mu)$. We use (2) to get

$$q_i(y^*) \approx \left(\frac{\mu_i}{y_i^*}\frac{\partial r(\mu; y^*)}{\partial \mu_i}\right)_{\mu=\mu^*}, \tag{8}$$

where the partial derivative is estimated by (7) and μ^* are the actual service rates.

Equation (8) is an approximation of the subgradient of the *expected* service level function. Our objective was to compute a subgradient that can be used to create a valid cut for the SAA of the call center problem. For the cut to be valid we require a subgradient of the *sample average* of the service level function.

For a fixed y^* and at μ^* the sample average of $r(\mu; y)$ as defined by (5) agrees with the sample average of the service level function $s(y; \mu)$ since the likelihood ratio $dP_\mu(\xi)/dQ(\xi)$ at $\mu = \mu^*$ equals 1 when $Q = P_{\mu^*}$. In (7) we compute a gradient of the sample average of r. Now, if we change either the staffing levels to y' or the service rates in to μ' (i.e., let $Q = P_{\mu'}$) then we get a new sample average function for r. Therefore we cannot guarantee that a gradient estimated by (7) is a subgradient of the sample average of r.

5 INFINITESIMAL PERTURBATION ANALYSIS

In the previous two sections we studied very different methods to approximate a subgradient for the service level function. The FD method suffers from being computationally expensive and can fail to produce a subgradient. The LR method is computationally efficient but the resulting subgradient approximation may not have the desired properties as a subgradient of the sample path function. In this section we study a third method, infinitesimal perturbation analysis (see e.g. Glasserman 1991).

IPA is related to the LR method and can be thought of as a special case of the LR method where only the sample path and not the distribution depends on the variable of interest (L'Ecuyer 1990, L'Ecuyer 1995). An IPA estimator of a gradient is a gradient of the sample path function with respect to the variable of interest. In this case the variables of interest are the number of servers in each period. These variables are integer so we approximate the service level function with a function of the service rate in each period as described in Section 2.

The service level in any period as a function of the service rates on each sample path is a finite sum of indicator functions where the jth indicator function equals 1 if the jth call arrives in the period and begins service on time. The derivative of this function with respect to any of the service rates equals zero where it exists and is undefined at the points where there is a jump in any of the indicator functions. It is reasonable to assume, however, that the *expected* service level function is differentiable everywhere and that the gradient is nonzero in general. Applying IPA directly to the service level function in that case does not yield useful gradient information. Instead, before applying IPA we *smooth* the sample path of the service level function (Glasserman 1991, Fu and Hu 1997).

Smoothing is done by conditioning on some of the random elements in the problem in order to obtain a function that is continuous on every (or almost every) sample path. At the same time, the conditioning argument ensures that this new function has the same expected value as the original function. In this case, for each call we condition on the value of the interarrival and service times of the previous calls so the information that we condition on increases with time as in filtered Monte Carlo (Glasserman 1993). If we let C be the number of calls in the planning horizon, A_j be the arrival time of call j and $L_j(\mu; y)$ equal 1 if call j arrives in the period and begins service on time when the rates and number of servers are as in μ and y then

$$r(\mu; y) = E\left[\sum_{j=1}^{C} L_j(\mu; y)\right] = E\left[\sum_{j=1}^{C+1} J_j(\mu; y)\right]. \quad (9)$$

Here,

$$\begin{aligned} J_j(\mu; y) &= \mathbf{1}\{\beta_j(\mu; y) \le V_e + \tau\}(F(V_e - A_{j-1}; A_{j-1}) \\ &\quad - F(\max\{V_s, \beta_j(\mu; y) - \tau\} - A_{j-1}; A_{j-1})) \end{aligned}$$

where V_s and V_e are the start and end points of the period, β_j is the (random) time when a server becomes available to serve the jth call, A_j is the time of the jth call and $F(\cdot; a)$ is the distribution of the interarrival time given that the last call arrived at time a. In words, J_j is the probability, given the arrival time of the previous call, that call j will arrive in the period and no sooner than τ time units before a server will become available to serve call j. The function J_j is continuous and piecewise differentiable whenever F and β_j are. If, for instance, the arrival process can be modeled as a nonhomogeneous Poisson process with an absolutely continuous cumulative rate function then this assumption is satisfied for F.

The next step is to show that $\beta_j(\mu; y)$ is continuous and piecewise differentiable in μ. For a general y this is not the case. Suppose that there is a reduction in the number of servers between periods and that $\beta_j(\mu; y)$ occurs just before the end of the period. Then there can be a jump in $\beta_j(\mu; y)$ if the service rate is slightly reduced so that the server that previously became available for that call is now working into the next period and is off once that server finishes service. If, however, the number of servers in all periods is constant and if the service rates changes instantaneously then it can be shown for bounded and positive service rates (similarly to Corollary 3.7 in Glasserman (1991)) that β_j is indeed continuous and piecewise differentiable. (By an instantaneous change in service rates we mean that when a new period starts then all calls in service are served at the service rate of the new period rather than at the service rate of the period when the call entered service. Thus a service time of a call will depend on what periods it is being serviced in, as opposed to only the service rate that is in effect when it starts service as in the model in the previous section.)

When all these assumptions are satisfied then

$$\nabla_\mu r(\mu; y) = E\left[\sum_{j=1}^{C+1} \nabla_\mu J_j(\mu; y)\right].$$

This estimator is generally not as easy to compute as the LR estimator (7) since we must keep track of how changes in the service rate *propagate* on the sample path. To approximate the subgradient as in (2) we must be careful about how to choose the fixed number of agents. One choice is $\hat{y} \equiv \max_{i\in\{1,\ldots,p\}} y_i^*$ and then $\mu_i^* = \tilde{\mu}\frac{y_i^*}{\hat{y}}$ where $\tilde{\mu}$ is the true service rate in all periods. Then

$$q_i(y^*) \approx \frac{\tilde{\mu}}{\hat{y}}\left(\frac{\partial r(\mu; \hat{y}e)}{\partial \mu_i}\right)_{\mu=\mu^*} \quad (10)$$

where e is the vector of all ones in $\mathbb{R}^p$.

Here we have estimated the subgradient of $s(y; \mu)$ at y^*. If we change the staffing levels then we get a new sample average function for (9) just as when we computed the subgradients for the LR method (see the end of Section 4). Alternatively, fix the staffing levels at an upper bound, $\hat{y} \in \mathbb{R}$ for the staffing level in any period, and let the service rate be $\mu_i = \mu_0 y_i/\hat{y}$ where μ_0 is the true service rate (same in all periods) and y_i is the staffing level in period i. Then, if the service times are exponential, we can use the bounds on the queue length process as in Theorem 1 of Chao and Scott (2000) to conclude that the sample average of $r(\mu; \hat{y}e)$ is an upper bound on the sample average of the service level function $s(y; \mu_0 e)$. Therefore a subgradient obtained by

(10) can be used to generate a valid (but possibly loose) cut.

6 NUMERICAL RESULTS

In the previous sections we developed three different methods for approximating subgradients of a discrete service level function via simulation. We mentioned some of the advantages and disadvantages of each method. In this section we present a small numerical example in order to shed light on the practical performance of each method. In particular we discuss the computational effort of each method, the variance of the subgradient estimators and the validity of the estimators as a subgradient of the SAA of the service level function.

We consider an $M(t)/M/s(t)$ queue with $p = 2$ periods of equal length of 30 minutes. The service rate is $\mu = 4$ calls/hour. The arrival process is a nonhomogeneous Poisson process with the arrival rate a function of the time t in minutes equal to $\lambda(t) = \lambda(1 - |t/60 - .65|)$. We set $\lambda = 120$ calls/hour, which makes the average arrival rate over the 2 periods equal 87.3 calls/hour. We say that a call begins service on time if it enters service less than 90 seconds after it arrives.

Our reasoning for presenting such a simple example, rather than a more realistic model of a call center, is that this example captures every complexity of the problem, it is easy to verify the properties of the subgradient approximation and a complete visualization of the service level function is possible.

We computed the average number of calls in each period answered in less than 90 seconds after they arrive. We did this for servers in period 1 ranging from 10 to 30 and the number of servers in period 2 ranging from 22 to 40. Our sample size was 999. We also computed at each point an approximation of a subgradient from each of the FD, LR and IPA methods. The staffing level in period 2 does not have a great effect on the service level in period 1 so in the remainder of the discussion we focus on the service level in period 2 as a function of the staffing levels in periods 1 and 2.

Figure 2 shows the number of calls received in period 2 that are answered on time as a function of the staffing levels in periods 1 and 2. Figures 3-5 show a contour plot of the same function (curved lines going across), the subgradient approximation at each point (arrows) and 95% confidence regions for selected points (ellipses). The subgradient arrows originate at their corresponding point and show both the magnitude and direction of the gradient. The ellipses are centered at the endpoint of the corresponding arrow. The confidence regions for the FD and IPA methods are so small that the ellipses are barely visible in the plots.

From Figure 3 we see that the finite difference method gives a good approximation of the subgradients over the whole domain, so if the computational work of running $p + 1$ simulations to get a subgradient at a single point is not overwhelming then the FD method would certainly be a good candidate.

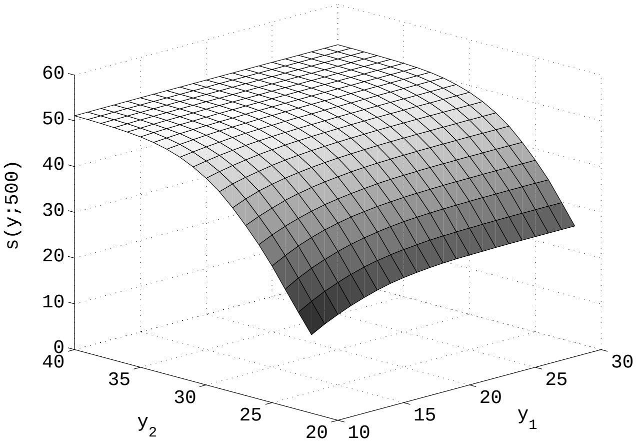

Figure 2: Sample Average Approximation (Sample Size 999) of the Number of Calls that are Answered on Time

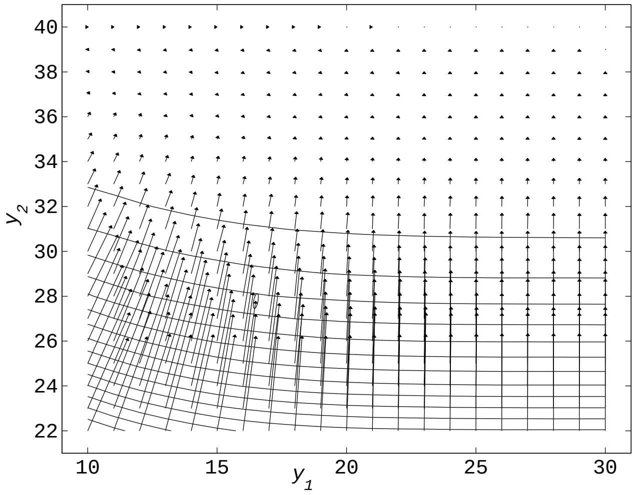

Figure 3: Subgradient Estimates and Confidence Regions via the Finite Difference Method

For exponential service times the likelihood ratio estimator (7) for the partial derivative of the number of period answered on time in period i w.r.t. the staffing level in period k simplifies to

$$\left.\frac{\partial r^i(\mu; y)}{\partial \mu_k}\right|_{\mu=\mu^*} \tag{11}$$
$$= \quad E\left[S^i(y;\mu^*) \sum_{j=1}^{C} \mathbf{1}\{\pi(B_j) = k\} \left(\frac{1}{\mu_k^*} - X_j \right) \right],$$

where X_j is the service time of the jth call. For this example $\mu_1^* = \mu_2^* = 15$ minutes. We can see from (11) that the LR

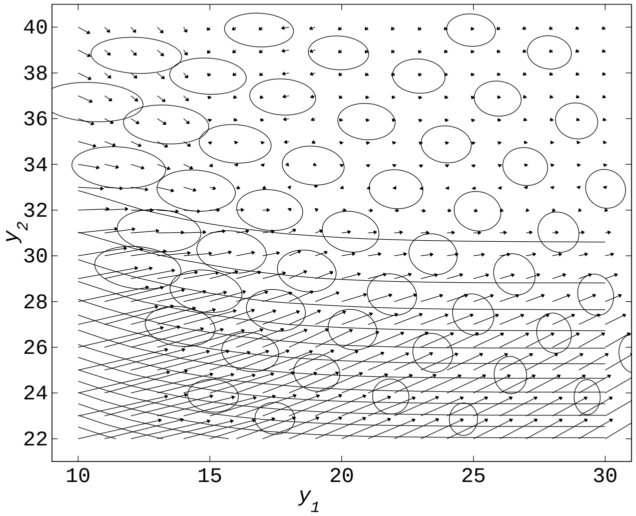

Figure 4: Subgradient Estimates and Selected Confidence Regions via the Likelihood Ratio Method

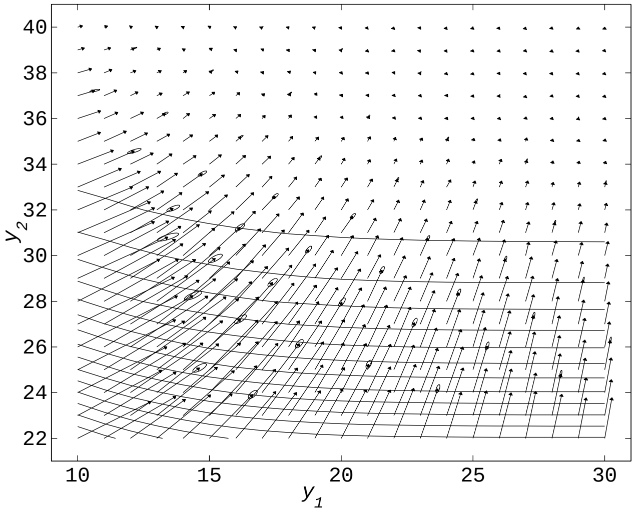

Figure 5: Subgradient Estimates and Confidence Regions via Infinitesimal Perturbation Analysis

method requires only the additional work of summing up the terms in (11) and multiplying by the respective service level function. On the other hand, we see that the confidence regions are much larger for the LR method than the other two methods. High variance of the LR method has also been observed by several others in the literature (see e.g. Fu (2002)).

A second observation from Figure 4 is that in the lower right corner the LR gradients suggest that the service level in period 2 will improve significantly if servers are added in period 1, even when there are plenty of servers in period 1. This is because the LR estimator interprets the increase as an increase in the service rate. Therefore, even if only a fraction of the servers are busy at the end of period 1, increasing the service rate will reduce their residual service times (recall that with the LR estimator the service time distribution depends only on when the call begins service). As a result, the state of the servers in period 2 is quite strongly impacted by an increase in service rate in period 1. A possible remedy would be to modify the LR method to take into account how much time a call actually spends in each period when computing service times.

Our last comment on the LR method is that some of the subgradient estimates have a negative component which contradicts a nondecreasing service level function. This is due to the large variance of the LR method and we can see that many of the confidence regions cover the origin of the corresponding arrow which means that the subgradient is not statistically different from zero.

The variance of the IPA estimator (Figure 5) is much lower than the variance of the LR estimator. The computational effort is only slightly greater for IPA than LR. It can be seen, however, that the estimates differ from the FD estimates, especially for low staffing levels in period 1. This occurs partly because, in computing the IPA estimator, we fix the number of servers at the maximum of the number of servers in periods 1 and 2 and adjust the service rates accordingly. This approach overestimates the performance, as can be seen by a coupling argument.

7 CONCLUSIONS

We have proposed three different approaches to a difficult problem. The FD approach seems to be the most consistent while at the same time being the most computationally expensive. Our future agenda related to this problem is to study in more detail the properties of using a change in service rates as a proxy for a change in staffing levels. Also, we would like to improve the quality of the IPA and LR estimators as subgradients, for instance by considering different approaches to account for the appropriate service rates as mentioned in the previous section.

ACKNOWLEDGMENTS

We thank Michael Beyer, Maria Ng and Shang Yen See, students at Cornell University, for their contributions to the computational study. This work was supported by National Science Foundation grant DMI-9984717.

REFERENCES

Atlason, J., M. A. Epelman, and S. G. Henderson. 2003. Call center staffing with simulation and cutting plane methods. *Annals of Operations Research.* To appear.

Chao, X., and C. Scott. 2000. Several results on the design of queueing systems. *Operations Research* 48 (6): 965–970.

Fu, M. C. 2002. Optimization for simulation: Theory vs. practice. *INFORMS Journal on Computing* 14 (3): 192–215.

Fu, M. C., and J. Q. Hu. 1997. *Conditional Monte Carlo: Gradient Estimation and Optimization Applications*. The Netherlands: Kluwer.

Glasserman, P. 1991. *Gradient Estimation via Perturbation Analysis*. The Netherlands: Kluwer.

Glasserman, P. 1993. Filtered Monte Carlo. *Mathematics of Operations Research* 18:610–634.

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.

Kleywegt, A. J., A. Shapiro, and T. Homem-de-Mello. 2001. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization* 12 (2): 479–502.

L'Ecuyer, P. 1990. A unified view of the IPA, SF, and LR gradient estimation techniques. *Management Science* 36 (11): 1364–1383.

L'Ecuyer, P. 1995. Note: On the interchange of derivative and expectation for likelihood ratio derivative estimators. *Management Science* 41 (4): 738–748.

Mandelbaum, A. 2003. Call centers (centres): Research bibliography with abstracts. Version 4. Accessible online via `<http://ie.technion.ac.il/serveng/References/ccbib.pdf>` [accessed June 17, 2003].

Topkis, D. M. 1998. *Supermodularity and Complementarity*. Princeton, New Jersey: Princeton University Press.

AUTHOR BIOGRAPHIES

JÚLÍUS ATLASON is a Ph.D. Candidate at the University of Michigan. His e-mail address is `<jatlason@umich.edu>`, and his web page is `<www-personal.umich.edu/~jatlason>`.

MARINA A. EPELMAN is an Assistant Professor at the University of Michigan. Her e-mail address is `<mepelman@umich.edu>`, and her web page is `<www.personal.engin.umich.edu/~mepelman>`.

SHANE G. HENDERSON is an Assistant Professor at Cornell University. His e-mail address is `<sgh9@cornell.edu>`, and his web page is `<www.orie.cornell.edu/~shane>`.