

The Optimality Gap of Asymptotically-derived Prescriptions with Applications to Queueing Systems

Ramandeep S. Randhawa
Marshall School of Business
University of Southern California

October 22, 2012

Abstract

In complex systems, it is quite common to resort to approximations when optimizing system performance. These approximations typically involve selecting a particular system parameter and then studying the performance of the system as this parameter grows without bound. Indexing this parameter by n , we prove that if the approximation to the objective function is accurate up to $\mathcal{O}(f(n))$ for some function $f(n)$, then under some regularity conditions, the prescriptions that are derived from this approximation are in fact $o(f(n))$ -optimal. A consequence of this result is that the well-known square-root staffing rules for capacity sizing in $M/M/N$ and $M/M/N + M$ queues to minimize the sum of linear expected steady-state customer waiting costs and linear capacity costs are $o(1)$ -optimal, i.e., their optimality gap is asymptotically zero.

1 Introduction and Framework

Introduction. Queueing theory has a rich history of using approximations to derive near-optimal prescriptions to optimization problems. These approximations typically consider a sequence of systems with increasing system scale, which is modeled by increasing arrival rates, and then approximate the performance measures of interest. For instance, the problem of selecting the optimal capacity to minimize the sum total of linear expected steady-state customer waiting costs and linear capacity costs requires an approximation of the expected steady-state queue-length. Suppose one has an approximation of this performance measure, then a capacity prescription is arrived at by optimizing the objective function obtained by replacing the actual performance measure by its approximation. If this approximation is accurate up to $\mathcal{O}(f(n))^1$, then one expects that the corresponding prescription would inherit the $\mathcal{O}(f(n))$ accuracy. It turns out that, under some regularity conditions, this prescription is far more accurate and in fact exhibits an accuracy of $o(f(n))$, i.e., the optimality gap when divided by $f(n)$ converges to zero as n grows without bound. Further,

¹For functions $f, g : \mathbb{N} \rightarrow \mathbb{R}$: $g(n)$ is said to be $\mathcal{O}(f(n))$ if $\limsup_{n \rightarrow \infty} |g(n)/f(n)| < \infty$; and $g(n)$ is said to be $o(f(n))$ if $\lim_{n \rightarrow \infty} |g(n)/f(n)| = 0$.

in many queueing applications the performance measures of interest can be approximated up to $\mathcal{O}(1)$, and in these cases our results imply that the capacity prescriptions are $o(1)$ -optimal, i.e., the optimality gap of the prescriptions is asymptotically zero.

Framework. Consider a sequence of systems indexed with n and let Π_n denote the objective function in system n . The goal of the system manager is to solve:

$$\inf_{x \in X_n} \Pi_n(x), \quad (1)$$

where X_n is a set of feasible decisions. We assume that all optimizers of (1) lie in X_n . Let x_n^* denote an optimizer and Π_n^* the optimal value function.

Suppose the following relation has been established for the system under consideration:

$$\Pi_n(g_n(x)) = a_n + b_n \bar{\pi}(x) + \hat{\pi}(x) + \epsilon_n(x), \quad \text{for } x \in \bar{X}_n, \quad (2)$$

where the following conditions hold:

1. $g_n : \bar{X}_n \rightarrow X_n$ are one-to-one and onto functions, where the sets \bar{X}_n lie in a metric space (M, d) . Further, we have $\bar{X}_n \subseteq \bar{X}_{n+1}$, and defining $\bar{X} = \cup_{n=1}^{\infty} \bar{X}_n$, we have $\bar{\pi}, \hat{\pi} : \bar{X} \rightarrow \mathbb{R}$ and $\epsilon_n : \bar{X}_n \rightarrow \mathbb{R}$;
2. $a_n \in \mathbb{R}$, $b_n \in \mathbb{R}_+$, and $\lim_{n \rightarrow \infty} b_n = \infty$;
3. $\inf_{x \in \bar{X}} \bar{\pi}(x) > -\infty$ and all optimizers of $\inf_{x \in \bar{X}} \bar{\pi}(x)$ lie in \bar{X} ;
4. $\hat{\pi}$ is continuous at every $x \in \arg \min_{x \in \bar{X}} \bar{\pi}(x)$;
5. For any sequence $\{x_n \in \bar{X}\}_n$, we have $\liminf_{n \rightarrow \infty} (\hat{\pi}(x_n) + \epsilon_n(x_n)) > -\infty$; and
6. For any $x \in \arg \min_{x \in \bar{X}} \bar{\pi}(x)$ there exists $\delta_x > 0$ so that $\lim_{n \rightarrow \infty} \sup_{\{y \in \bar{X} : d(x,y) < \delta_x\}} |\epsilon_n(y)| = 0$.

2 Optimality Gap of Approximate Solution

Given the approximation (2), it seems reasonable to solve the optimization problem

$$\min_{x \in \bar{X}} \bar{\pi}(x), \quad (3)$$

and then use one of its optimizers to obtain a prescription for the original system with the knowledge that (2) implies that the performance of this prescription will be near optimal. Formally, we define \bar{x}^* to be an optimizer of (3) such that $\hat{\pi}(\bar{x}^*) = \min_{x \in \bar{X}^*} \hat{\pi}(x)$, where $\bar{X}^* \equiv \arg \min_{x \in \bar{X}} \bar{\pi}(x)$. That is, \bar{x}^* is an optimizer of (3) with the smallest value of $\hat{\pi}$ among all optimizers of (3). Then, the following result proves that the prescription $g_n(\bar{x}^*)$ has an optimality gap of $o(1)$ for the optimization problem (1).

Theorem 1. *If the objective function Π_n satisfies (2) along with conditions 1–6, then $g_n(\bar{x}^*)$ is $o(1)$ -optimal for the optimization problem (1), i.e., we have*

$$\Pi_n(g_n(\bar{x}^*)) - \Pi_n^* \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Proof. We can write the optimality gap of the prescription as

$$\begin{aligned} 0 \leq \left[\Pi_n(g_n(\bar{x}^*)) - \Pi_n^* \right] &= \left[\Pi_n(g_n(\bar{x}^*)) - a_n - b_n \bar{\pi}(\bar{x}^*) \right] - \left[\Pi_n^* - a_n - b_n \bar{\pi}(g_n^{-1}(x_n^*)) \right] \\ &\quad + b_n \left[\bar{\pi}(\bar{x}^*) - \bar{\pi}(g_n^{-1}(x_n^*)) \right], \end{aligned} \quad (4)$$

where n is large enough so that $\bar{x}^* \in \bar{X}_n$.

For the first term on the right-hand-side of (4), using (2) and the fact that $\lim_{n \rightarrow \infty} |\epsilon_n(\bar{x}^*)| = 0$ (which follows from condition 6), we have

$$\lim_{n \rightarrow \infty} \left[\Pi_n(g_n(\bar{x}^*)) - a_n - b_n \bar{\pi}(\bar{x}^*) \right] = \hat{\pi}(\bar{x}^*). \quad (5)$$

We next evaluate the second term on the right-hand-side of (4). Using the optimality of x_n^* , we have

$$\begin{aligned} \limsup_{n \rightarrow \infty} \frac{\Pi_n(x_n^*) - a_n}{b_n} &\leq \limsup_{n \rightarrow \infty} \frac{\Pi_n(g_n(\bar{x}^*)) - a_n}{b_n} \\ &= \bar{\pi}(\bar{x}^*) + \limsup_{n \rightarrow \infty} \left(\frac{\hat{\pi}(\bar{x}^*) + \epsilon_n(\bar{x}^*)}{b_n} \right) \\ &= \bar{\pi}(\bar{x}^*), \end{aligned} \quad (6)$$

where the first equality follows from (2) and the second equality follows from the fact that $\hat{\pi}(\bar{x}^*)$ is bounded, which is a consequence of the continuity of $\hat{\pi}$ at \bar{x}^* , and the fact that $\lim_{n \rightarrow \infty} |\epsilon_n(\bar{x}^*)| = 0$. Considering the sequence $\{g_n^{-1}(x_n^*)\}_n$, we also have

$$\begin{aligned} \liminf_{n \rightarrow \infty} \frac{\Pi_n(x_n^*) - a_n}{b_n} &= \liminf_{n \rightarrow \infty} \left[\bar{\pi}(g_n^{-1}(x_n^*)) + \frac{\hat{\pi}(g_n^{-1}(x_n^*)) + \epsilon_n(g_n^{-1}(x_n^*))}{b_n} \right] \\ &\geq \liminf_{n \rightarrow \infty} \bar{\pi}(g_n^{-1}(x_n^*)), \end{aligned} \quad (7)$$

where the inequality follows by noting that condition 5 implies that $\liminf_{n \rightarrow \infty} [\hat{\pi}(g_n^{-1}(x_n^*)) + \epsilon_n(g_n^{-1}(x_n^*))] > -\infty$. Comparing (6) and (7), we obtain

$$\liminf_{n \rightarrow \infty} \bar{\pi}(g_n^{-1}(x_n^*)) \leq \bar{\pi}(\bar{x}^*)$$

It follows that all cluster points of the sequence $\{g_n^{-1}(x_n^*)\}_n$ must be optimizers of (3). Let us now consider a subsequence indexed by n_k such that $g_{n_k}^{-1}(x_{n_k}^*) \rightarrow \bar{x}$ as $k \rightarrow \infty$ for some $\bar{x} \in \bar{X}^*$. Using

(2), we can then write

$$\lim_{k \rightarrow \infty} \left[\Pi_{n_k}^* - a_{n_k} - b_{n_k} \bar{\pi}(g_{n_k}^{-1}(x_{n_k}^*)) \right] = \lim_{k \rightarrow \infty} \hat{\pi}(g_{n_k}^{-1}(x_{n_k}^*)) = \hat{\pi}(\bar{x}), \quad (8)$$

where the first equality follows because there exists $\delta_{\bar{x}} > 0$ such that $\lim_{n \rightarrow \infty} \sup_{\{y \in \bar{X} : d(\bar{x}, y) \leq \delta_{\bar{x}}\}} |\epsilon_n(y)| = 0$ and the second equality follows from the continuity of $\hat{\pi}$ at \bar{x} .

Finally, consider the third term in the right-hand-side of (4). The optimality of \bar{x}^* and the fact that $g_n^{-1}(x_n^*) \in \bar{X}_n \subseteq \bar{X}$ implies that

$$\left[\bar{\pi}(\bar{x}^*) - \bar{\pi}(g_n^{-1}(x_n^*)) \right] \leq 0. \quad (9)$$

Combining (5), (8) and (9) in (4), we obtain

$$0 \leq \limsup_{k \rightarrow \infty} \left[\Pi_{n_k}(g_{n_k}(\bar{x}^*)) - \Pi_{n_k}^* \right] \leq \hat{\pi}(\bar{x}^*) - \hat{\pi}(\bar{x}).$$

Using the fact that $\hat{\pi}(\bar{x}^*) = \min_{x \in \bar{X}^*} \hat{\pi}(x)$ and $\bar{x} \in \bar{X}^*$, we obtain $\hat{\pi}(\bar{x}^*) \leq \hat{\pi}(\bar{x})$, which implies that $\limsup_{k \rightarrow \infty} \left[\Pi_{n_k}(g_{n_k}(\bar{x}^*)) - \Pi_{n_k}^* \right] = 0$ and this completes the proof. \square

Generalization of Theorem 1. Consider the following general version of (2):

$$\Pi_n(g_n(x)) = a_n + b_n \bar{\pi}(x) + c_n [\hat{\pi}(x) + \epsilon_n(x)], \quad (10)$$

where conditions 1–6 hold as before, and $c_n \in \mathbb{R}_+$ with $c_n = o(b_n)$. Then, proceeding as in the main result, we obtain that \bar{x}^* defined as before is $o(c_n)$ -optimal for the optimization problem (1), that is, we have

$$\frac{\Pi_n(g_n(\bar{x}^*)) - \Pi_n^*}{c_n} \rightarrow 0, \text{ as } n \rightarrow \infty. \quad (11)$$

3 Queueing Applications

3.1 $o(1)$ -optimality of square-root staffing in $M/M/N$ queues

Consider an $M/M/N$ queue in which customers arrive at rate n per unit time and have a mean service time of $1/\mu$ time units. Suppose the system manager faces a capacity planning problem in which she needs to select the number of servers to minimize total system costs, which include delay costs and capacity costs. Such a problem was considered in Borst et al. (2004) and we consider a linear version of this problem in which the customer delay cost is h per customer per unit time and the cost per server is c per unit time. Then, the total system cost of staffing x servers is $\Pi_n(x) = h\mathbb{E}Q_n(x) + cx$, where $Q_n(x)$ is the steady-state queue-length (number of customers waiting in the queue) and we can write its expected value as $\mathbb{E}Q_n(x) = n \left(\frac{1}{x\mu - n} \right) C(x, n)$, where $C(x, n)$ is the blocking probability, i.e., the probability that all servers are busy. Although the function

$C(x, n)$ is defined only for integers, following Janssen et al. (2011), we note that an extension to real numbers is obtained by using $C(x, n) = \left[\frac{n}{\mu} \int_0^\infty t e^{-\frac{n}{\mu}t} (1+t)^{x-1} dt \right]^{-1}$. So, we can consider the optimization problem $\min_{x \in (n/\mu, \infty)} \Pi_n(x)$.

Using Theorem 2 of Janssen et al. (2011) and defining $g_n(x) = \frac{n}{\mu} + \sqrt{\frac{n}{\mu}}x$ for $x \in \bar{X}_n \equiv (0, \infty)$, we can write

$$\mathbb{E}Q_n(g_n(x)) = \sqrt{\frac{n}{\mu}}\bar{q}(x) + \hat{q}(x) + e_n(x), \quad (12)$$

where

$$\bar{q}(x) = \frac{1}{x} \left(1 + \frac{x\Phi(x)}{\phi(x)} \right)^{-1}, \quad (13)$$

$$\hat{q}(x) = x^2 \bar{q}(x)^2 \left(\frac{1}{3} + \frac{x^2}{6} + \frac{\Phi(x)}{\phi(x)} \left(\frac{x}{2} + \frac{x^3}{6} \right) \right), \quad (14)$$

with Φ and ϕ denoting the cumulative distribution and density functions of the standard normal distribution, respectively, and e_n satisfying $\liminf_{n \rightarrow \infty} e_n(x_n) > -\infty$ for any sequence $\{x_n \in \bar{X}\}_n$. We also have that for any $x \in \bar{X}$ there exists $\delta_x > 0$ such that $\lim_{n \rightarrow \infty} \sup_{\{y \in \bar{X}: d(x,y) < \delta_x\}} |e_n(y)| = 0$. We can then write the following approximation for Π_n :

$$\Pi_n(g_n(x)) = c \frac{n}{\mu} + \sqrt{\frac{n}{\mu}}\bar{\pi}(x) + \hat{\pi}(x) + \epsilon_n(x), \quad (15)$$

where $\bar{\pi}(x) = cx + h\bar{q}(x)$, $\hat{\pi}(x) = h\hat{q}(x)$, and $\epsilon_n = he_n(x)$. It can be verified that all the conditions 1–6 are satisfied.

It is easy to check that $\bar{\pi}(x)$ has a unique minimizer \bar{x}^* . The corresponding staffing rule $\frac{n}{\mu} + \sqrt{\frac{n}{\mu}}\bar{x}^*$ is the well-known square-root staffing rule. As the objective function here satisfies (2) and conditions 1–6, Theorem 1 implies that the square-root staffing rule is $o(1)$ -optimal.

Janssen et al. (2011) refines the square-root staffing rule by incorporating the function $\hat{\pi}(x)$ into the optimization problem and proves that the refined staffing prescription thus obtained is optimal up to $\mathcal{O}(1/\sqrt{n})$. It turns out that the square-root staffing rule without any refinement is also optimal up to $\mathcal{O}(1/\sqrt{n})$. To see this, note that Janssen et al. (2011) proves that $\epsilon_n(x) = \mathcal{O}(1/\sqrt{n})$, and thus the proof of Theorem 1 implies that the capacity prescription $\frac{n}{\mu} + \sqrt{\frac{n}{\mu}}\bar{x}^*$ itself would be optimal up to $\mathcal{O}(1/\sqrt{n})$. Thus, our result suggests that the refinement to square-root staffing computed in Janssen et al. (2011) should in fact be $o(1/\sqrt{n})$ -optimal.

3.2 $o(1)$ -optimality in $M/M/N + G$ queues

Consider the same setup that we used for the $M/M/N$ queues in the previous section with the addition of customer abandonments. In particular, customers are endowed with a patience distribution with cumulative distribution function G , a density function g that is strictly positive

and continuously differentiable on $[0, \infty)$, and mean $1/\gamma$, such that the time customers are willing to wait is distributed according to this distribution in an i.i.d. fashion. Consider the same optimization problem considered previously, namely, to select the number of servers (extended to real numbers) to minimize linear customer delay costs and linear capacity costs, i.e., $\min_{x \in [0, \infty)} \Pi_n(x)$, where $\Pi_n(x) = h\mathbb{E}Q_n(x) + cx$.

Using Theorem 5 of Bassamboo and Randhawa (2010) and defining $\bar{w}(x) \equiv \bar{G}^{-1}(\min(1, x\mu))$ with $\bar{G} \equiv 1 - G$, $g_n(x) = nx$ for $x \in \bar{X}_n \equiv [0, \infty)$, we can write

$$\mathbb{E}Q_n(g_n(x)) = n\bar{q}(x) + \hat{q}(x) + e_n(x),$$

where

$$\begin{aligned} \bar{q}(x) &= \int_0^{\bar{w}(x)} \bar{G}(w) dx, \\ \hat{q}(x) &= -\frac{1}{2} \left(\frac{g'(\bar{w}(x))}{\rho g^2(\bar{w}(x))} + 1 \right), \end{aligned}$$

and we have that for any $x < 1/\mu$ there exists $\delta_x > 0$ such that $\lim_{n \rightarrow \infty} \sup_{\{y \in \bar{X}: d(x,y) < \delta_x\}} |e_n(y)| = 0$. (Note that the referred theorem only proves that $\hat{q}(x) = \mathcal{O}(1)$, but the same proofs also yield the additional properties listed here. Further the case $x < 1/\mu$ corresponds to the overloaded regime, or efficiency driven regime as defined in Garnett et al. (2002), in which offered load exceeds unity.) This is the so-called fluid approximation of the queueing system. We then have the following analog of (2):

$$\Pi_n(g_n(x)) = n\bar{\pi}(x) + \hat{\pi}(x) + \epsilon_n(x), \quad (16)$$

where $\bar{\pi}(x) = cx + h\bar{q}(x)$, $\hat{\pi}(x) = h\hat{q}(x)$, and $\epsilon_n(x) = he_n(x)$. It is easy to establish that conditions 1 through 5 are satisfied here, and that $x \in \bar{X}^*$ satisfy $x \leq 1/\mu$ so that condition 6 is satisfied only when $1/\mu \notin \bar{X}^*$. Note that if $1/\mu \notin \bar{X}^*$, then the capacity prescriptions obtained from the approximations lead the system into an overloaded regime. Theorem 1 then implies that if $1/\mu \notin \bar{X}^*$, then $g_n(\bar{x}^*)$ is an $o(1)$ -optimal solution to the cost minimization problem (1) and thus is more accurate than that proved in Bassamboo and Randhawa (2010), which only proved that this fluid-based prescription is $\mathcal{O}(1)$ -optimal.

However, if we have $1/\mu \in \bar{X}^*$, which is equivalent to stating that the optimal capacity prescription leads the system to the critically loaded regime, then we do not obtain the $o(1)$ -optimality of the fluid-based prescription. In fact, in this case square-root staffing is the $o(1)$ -optimal solution, and identifying it requires a finer approach, which is the diffusion approximation. In particular, we set $g_n(x) = \frac{n}{\mu} + \sqrt{\frac{n}{\mu}}x$ with $\bar{X}_n \equiv [-\sqrt{n/\mu}, \infty)$ and expect the following approximation to hold:

$$\mathbb{E}Q_n(g_n(x)) = \sqrt{\frac{n}{\mu}}\bar{q}_1(x) + \hat{q}_1(x) + e_{n,1}(x). \quad (17)$$

We expect the above approximation to hold for generally distributed patience times, especially given the recent work Gurvich et al. (2012) which proves that $\left(\mathbb{E}Q_n(g_n(x)) - \sqrt{\frac{n}{\mu}}\bar{q}_1(x)\right)$ is $\mathcal{O}(1)$. For the case of exponentially distributed patience times, i.e., $M/M/N + M$ queues, Zhang et al. (2012) has recently established (17). In particular, Theorem 5 of Zhang et al. (2012) can be used to obtain

$$\begin{aligned}\bar{q}_1(x) &= \left(\sqrt{\frac{\gamma}{\mu}}H_\gamma(x) - x\right)\frac{\mu}{\gamma}A_*(x), \\ \hat{q}_1(x) &= \mu\bar{q}_1(x) \left[-h_\gamma(x)A_*(x) - \frac{1}{6}x^2H_\gamma(x)\sqrt{\frac{\mu}{\gamma}} + \frac{1}{6}\sqrt{\frac{\gamma}{\mu}}xH_\gamma(x) \left(\sqrt{\frac{\gamma}{\mu}}H_\gamma(x) - x\right)^{-1}\right],\end{aligned}$$

where defining $G(x) = \frac{\Phi(x)}{\phi(x)}$, we have

$$\begin{aligned}H_\gamma(x) &= \frac{\phi(x\sqrt{\mu/\gamma})}{\Phi(-x\sqrt{\mu/\gamma})}, & A_*(x) &= \left(1 + \sqrt{\frac{\gamma}{\mu}}G(x)H_\gamma(x)\right)^{-1}, \\ h_\gamma(x) &= -\frac{1}{6}\sqrt{\frac{\gamma}{\mu}}x^2H_\gamma(x) \left[G(x)H_\gamma(x)\sqrt{\frac{\mu}{\gamma}} - xG(x)\frac{\mu}{\gamma} + 1 + xG(x)\right].\end{aligned}$$

It follows that we can write the approximation to the cost function Π_n in the following form:

$$\Pi_n(g_n(x)) = c\frac{n}{\mu} + \sqrt{\frac{n}{\mu}}\bar{\pi}_1(x) + \hat{\pi}_1(x) + \epsilon_{n,1}(x). \quad (18)$$

It can be established that conditions 1–6 hold so that the optimal square-root staffing $\frac{n}{\mu} + \sqrt{\frac{n}{\mu}}\bar{x}_1^*$ is $o(1)$ -optimal, where \bar{x}_1^* minimizes $\bar{\pi}_1$ and satisfies the relation $\hat{\pi}_1(\bar{x}_1^*) = \min_{x \in \bar{X}_1^*} \hat{\pi}_1(x)$, where $\bar{X}_1^* \equiv \arg \min_{x \in \bar{X}_1} \bar{\pi}_1(x)$.

4 Discussion

We have characterized the optimality gap of prescriptions for optimization problems obtained using asymptotic approximations for the objective function. We have discussed a couple of applications to capacity sizing problems in $M/M/N$ and $M/M/N + G$ systems, and it is easy to see that the results are applicable more generally, for instance, to pricing problems of the form introduced in Maglaras and Zeevi (2003) as long as one can establish an analog of (2). Although the examples we have discussed pertain to unconstrained cost minimization, our results apply to constrained optimization as well by appropriately incorporating the constraints into the feasible set. However, depending on the nature of the constraints, conditions 1–6 may not hold. For instance, in an $M/M/N$ queueing system, consider the problem of minimizing the number of servers to ensure that the probability of waiting (before beginning service) is less than a threshold. In this case, the results of Janssen et al. (2011) imply that square-root staffing is only $\mathcal{O}(1)$ -optimal, and indeed we can verify that condition 1 is violated, in particular $\bar{X}_n \not\subseteq \bar{X}_{n+1}$ (which implies that (9) need not

hold).

In conclusion, I would like to point out that in this paper we have not discussed the issue of optimal control in queueing systems, which is an important topic. In fact, some recent work on this topic has a similar flavor of asymptotic optimality where a control policy derived using an approximation that is accurate up to a certain order leads to an optimality gap of a smaller order. For instance, in a dynamic matching queueing application Gurvich and Ward (2012) proves that a policy derived from solving a static problem, that has an accuracy gap of $\mathcal{O}(\sqrt{n})$, in a dynamic manner (a discrete-review policy) leads to $o(\sqrt{n})$ -optimality. In a revenue management setting, Reiman and Wang (2008) proves that the solution to a static optimization problem, which is expected to have an accuracy gap of $\mathcal{O}(\sqrt{n})$, when resolved once is $o(\sqrt{n})$ -optimal, and recently Jasin and Kumar (2012) proves that re-solving this optimization problem many times can lead to $\mathcal{O}(1)$ -optimality. While these results appear similar at first glance, the dynamic resolving of the static optimization problem allows the approximate control solution to “re-align” with the actual system. This feature is missing in the problems that we have studied here. Nevertheless, the similarity is interesting and calls for a further, potentially unifying, study.

References

- Bassamboo, A. and Randhawa, R. (2010), ‘On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers’, *Operations Research* **58**(5), 1398–1413.
- Borst, S., Mandelbaum, A. and Reiman, M. I. (2004), ‘Dimensioning large call centers’, *Operations Research* **52**, 17–34.
- Garnett, O., Mandelbaum, A. and Reiman, M. (2002), ‘Designing a call center with impatient customers’, *Manufacturing & Service Operations Management* **4**(3), 208–227.
- Gurvich, I., Huang, J. and Mandelbaum, A. (2012), ‘Excursion-based universal approximations for the erlang-a queue in steady-state’, *Working paper*.
- Gurvich, I. and Ward, A. (2012), ‘On the dynamic control of matching queues’, *Working paper*.
- Janssen, A., van Leeuwen, J. and Zwart, B. (2011), ‘Refining square root safety staffing by expanding Erlang C’, *Operations Research* **59**(6), 1512–1522.
- Jasin, S. and Kumar, S. (2012), ‘A re-solving heuristic with bounded revenue loss for network revenue management with customer choice’, *Mathematics of Operations Research* **37**(2), 313–345.
- Maglaras, C. and Zeevi, A. (2003), ‘Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations’, *Management Science* **49**(8), 1018–1038.
- Reiman, M. and Wang, Q. (2008), ‘An asymptotically optimal policy for a quantity-based network revenue management problem’, *Mathematics of Operations Research* **33**(2), 257–282.
- Zhang, B., van Leeuwen, J. and Zwart, B. (2012), ‘Staffing call centers with impatient customers: Refinements to many-server asymptotics’, *Operations Research* **60**(2), 461–474.