Control of Non-Conventional Queueing Networks: A Parametric and Approximate Dynamic Programming Approach

Xiaoting Chen  Emmanuel Fernandez
Department of Electrical Engineering and Computing Systems
University of Cincinnati
Cincinnati, OH 45221, USA

Abstract

In this paper, we investigate the control of non-conventional queueing networks, where multiple concurrent state transitions following non-exponential and general sojourn distributions are allowed. Two approximation schemes are discussed that produce an approximated Markovian model. We further propose to model the problem as an uncertain Markov decision processes (MDP) by considering the induced approximation error. A new simulation-based approach is investigated here, to give an overall optimal policy beyond the classic approach to such problems, e.g., a robust formulation. In particular, we approach the problem as one of finding a best overall control policy, via exploration and exploitation within a heuristic policy set. An approximate dynamic programming algorithm is then used, in connection with a parametric cost function, for efficiently learning and finding policies in this set. We show that the problem of finding the best control policy within this new policy set can be understood, equivalently, as one of finding the best set of parameters for one-stage cost function of the problem. Later, an integrated framework, denoted as extended actor-critic, is proposed to give a comprehensive treatment for those types of problems. Results of a case study are also presented and discussed.

Keywords

1. Introduction

Queueing networks have been used as a major mathematical tool for modeling certain stochastic processes. Classic models of queueing networks usually assume Poisson arrival and exponential service time. These models are well understood with vast literature available on the subject. Departing from these models, problems with more general assumptions about the arrival and service time are of much interest for researchers and practitioners [1].

With the increasing complexity of queueing networks, the related control problem, i.e., queueing discipline, becomes also an important factor in their analysis. For example, multi-class queueing networks usually require delicate treatments for sequencing among different job classes. For networks with classic assumptions, e.g., Poisson arrival and exponential service time, their control problems can be modeled and investigated relatively easily within the framework of Markov decision processes (MDP) [2]. However, this does not hold for networks with general assumptions. In order to study more general problems, one often turns to investigate an easier one through approximation, e.g., to study a similar fluid or Brownian network. For example, a fluid network can be used to provide a deterministic first-order approximation. Its optimal control, and the connection to the actual stochastic network, are well understood given many available theoretical studies and contributions in literature, cf. [3,4] and references therein.

The fluid and Brownian approximations, with their various assumptions, are mainly used to approximate the scaled or normalized queue lengths within a network. A different approach we investigated in this paper is to approximate the dynamics of state transitions in a network, which no longer can be completely modeled by discrete Markov chain, due to general assumptions of the network. Such an approximation framework has been studied and shown to be useful in [5], as well as in [6]. However, similar to other approximation methods, connections between the approximated
problem and the actual network needs to be further studied. This approximation framework here is based on the
dynamics of state transitions in a network, and essentially, through considering matching problems to the arrival
and service time distributions. This utilizes parametric representations of continuous phase-type (PH) distributions
[7]. It is known that PH distributions can have variable structures depending on several factors, therefore our first
focus is to examine the potential parameter biases with this framework for approximating transition dynamics in
queueing networks. To start, we briefly discuss in the following about the modeling of the queueing networks as
well as the reasons behind such approximation. To further distinguish networks considered here from their classic
counterparts, we refer queueing networks with general assumptions about arrival and service time distributions as
“non-conventional” type in the rest of this paper.

To have an adequate model for non-conventional queueing networks, we consider the so-called generalized semi-
Markov processes (GSMP) formulation here [8]. GSMP has been proposed as a complement for classic Markovian or
semi-Markovian models, such that, it gives a comprehensive framework for capturing essential dynamical structure of
more general types in a system. As a result, the state transitions of GSMP model depend continuously on the history
of time elapses. For queueing networks considered here, this corresponds to time elapsed for all service stations since
their last service events. When decision making is incorporated into the formulation, we then have the generalized
semi-Markov decision processes (GSMDP). To find a good control policy for those problems, indirect approaches
with model approximations are usually needed, as discussed early on. Although the discussed approximation frame-
work above can provide a fairly good way of seeking policies for non-conventional queueing networks modeled with
GSMDP, it is not well understood whether the derived policies serve as an overall optimal one for the actual process.
This is because the optimal policy for the approximated problem can be very sensitive to parameter biases and uncer-
tainties induced with the approximation procedure. In this paper, we first investigate impacts of parameter biases in
the approximation framework used here. We discuss two approximation schemes under this framework, and examine
their parameter biases respectively. In order to further alleviate negative impacts from approximations, a parameter
uncertainty formulation is considered later, along with a new simulation-based approach, aiming to give an overall
optimal policy for control of non-conventional queueing networks.

The rest of the paper is organized as follows. In Section 2, we first review the generalized semi-Markov decision
processes, then we discuss and compare two approximation schemes under what we call here a general competing
transition scenario. In Section 3, we give a parameter uncertainty formulation of the optimization problem, as to
further alleviate negative impacts from approximations. In addition, we discuss a heuristic approach for finding overall
optimal policies to the formulated optimization problem, through using a parametric and simulation-based approach.
In Section 4, we present a case study using the proposed approach for optimizing sequencing control problem in a
small non-conventional queueing network. Conclusions are then given in Section 5.

2. Generalized Semi-Markov Decision Processes and Its Approximation

2.1 Generalized Semi-Markov Decision Processes

Consider models of generalized semi-Markovian decision processes (GSMDP), given by the tuple \( < T, S, A, P, Q, c, E > \).
By conventions, \( T = \{ t_0, t_1, \ldots, t_k, \ldots \} \) is the discrete time index of the state transitions, \( S \) is state space, \( A \) is action space,
\( P \) is transition matrix, \( Q \) is the matrix of joint cumulative distribution of sojourn time and state transition, and \( c \) is the
cost rate function [9]. Similar to MDP models, we restrict the control actions to be only applied when state transitions
occur. The model here is extended with the inclusion of the so-called event set \( E \), which is associated with each state
\( s \in S \) to indicate active events. For each event \( e \in E_s \), a clock \( c_e \) is used to indicate the time duration since the event be-
comes active, e.g., time since work station becomes active. To properly model state transitions, the following notation
is used for the joint cumulative distribution of sojourn time and state transition, conditioning on the clock readings,
i.e., \( Q_{s,e}(\tau, a|c_e) \triangleq \Pr \{ t_{k+1} - t_k \leq \tau, s_{k+1} = s'|s_k = s, a_k = a, c_e \} \). For a continuous time process, a value function can be
declared for the problem, for example, with a discounted cost (DC) criterion as follows:

\[
J^\pi(s) \triangleq \lim_{N \to \infty} \mathbb{E}[\pi] \left\{ \int_0^{\tau_N} e^{-\beta t} c(s(t), a(t)) \, dt | s_0 = s \right\}.
\] (1)

The control goal here is to find a Markovian and deterministic control policy \( \pi: s \to a \) [2], to minimize the value
function, e.g., \( \min_{\pi \in \Pi} J^\pi(s), \forall s \in S \), where \( \Pi \) is the set of all feasible control policies of this type. To solve this
minimization problem, one may follow the principle of optimality [2], and write the value function into two parts as:

\[ J^\pi(s) = c_0 + \lim_{N \to \infty} E^\pi \left\{ \int_{t_1}^{t_N} e^{-\beta t} c(s(t), a(t)) \, dt | s_{t_1} = s', c_e \right\}, \]

\[ = c_0 + E^\pi \left\{ e^{-\beta t} J^\pi(s') | c_e \right\}. \]

Where \( c_0 \) is the expected cost before the first state transition and the right hand side term is the well-known expected cost-to-go. When the state transitions of the model are synchronous type, e.g., every work station event triggers at the same time, or if there are no concurrent events with non-exponential and general sojourn distributions, then the value function and the cost-to-go term do not depend conditionally on the clock reading \( c_e \). In the former case, the clock readings \( c_e \) become \( 0 \) at epochs of state transitions, and in the latter case, the so-called memoryless property applies. Therefore, a closed-form can be derived for the expected cost-to-go term independent to the clock readings, and essentially the minimization problem can be solved efficiently with many existing algorithms [10]. However, this is not the case when there are asynchronous events that follow non-exponential and general sojourn distributions. The formulation can remain Markovian if all clock readings are included, e.g., by expanding the state space to be \( (s, c_e) \), however the poor analytical tractability of this formulation would not allow optimal policies to be found efficiently. Although pioneer works have been reported for MDP problems with state space expanded by continuous time [11, 12], it is in general considered impractical to find optimal policies due to the immense search space.

The above formulation is understood as a direct approach, that is, to find a discrete equivalent model for the continuous time problem, and then solve it in the discrete time domain. Although the GSMDP model provides a great extension, it is in general considered impractical to find optimal policies due to the immense search space.

The direct approach mentioned above fails due to the inclusion of continuous time, either explicitly included in an expanded state space, or implicitly being conditioned on as in Equation (3). As an alternative, an indirect approach can be used by utilizing a model approximation framework, as discussed early on. In this study, we consider the approximation framework based on continuous phase-type (PH) distributions. Essentially, one may approximate state transitions with non-exponential sojourn delays with PH distributions, so that an approximated discrete time Markovian model can be recovered and later solved with well-established techniques. PH distributions introduce the use of multiple memoryless and transient phases, where sojourn delays in between follow exponential distributions. Due to the fact that the set of PH distributions is dense with all positive distributions, in theory, it can be used to approximate any positive-valued distribution. Existing algorithms thus allow matching first two or three moments of an arbitrary distribution, while maintaining minimum number of phases [13, 14].

Although the approach provides a fairly good framework for obtaining discrete time Markovian models for non-conventional queueing networks, unlike other established approximation methods, several key issues are not fully addressed and understood. For instance, it is to be seen: 1) whether severe parameter biases will be introduced under certain conditions, and 2) the quality of the derived optimal policy across different approximation schemes. Here, our focus is to try to answer the first question. In particular, we examine whether additional bias beyond the normal approximation residue will be introduced when such approximation scheme is used, under what we call a general competing transition scenario. The competition between different state transitions here is understood in a same way as in classic Markov models, except the distributions of sojourn delay are allowed to be of general types.

To understand why additional parameter biases will be introduced into the process, we consider the following continuous time model with only three states, i.e., \( S = \{ s_0, s_1, s_2 \} \). For simplicity, we consider the case where state transitions are with synchronous type, while the subsequent results discussed here can be easily generalized to the asynchronous case. The process starts at \( s_0 \) and tries to get to one of the states \( \{ s_1, s_2 \} \) according to some distributions \( G_1 \) and \( G_2 \). And when the state is in either \( s_1 \) or \( s_2 \), it goes back to \( s_0 \) by following exponential distributions \( E_1 \) and \( E_2 \) respectively. Let \( \{ s_{iG_1}, s_{jG_2} \} | i, j = 0, 1, \ldots \) be phases used in approximating \( G_1 \) and \( G_2 \). In particular, let \( s_0^{G_1} = s_0^{G_2} = s_0 \) be a replacement of the original \( s_0 \), since \( s_0 \) will later be understood in a broader sense. Approximation with PH distributions thus allows the sojourn delays, e.g., \( s_0 \rightarrow s_{1G_1} \rightarrow s_{1G_1}^{j+1} \rightarrow \cdots \rightarrow s_1 \), and \( s_0 \rightarrow s_{2G_2} \rightarrow s_{2G_2}^{j+1} \rightarrow \cdots \rightarrow s_2 \), to match distribution \( G_1 \) and \( G_2 \) respectively, up to the first two or three moments.

One of the problem with such approximation scheme is that the transition to \( s_1 \) or \( s_2 \) becomes deterministic after the first jump from \( s_0 \) to \( s_{1G_1} \), or \( s_{2G_2} \). This is counter-intuitive, since all phases introduced represent different stages of original state \( s_0 \), where its transition to \( s_1 \) and \( s_2 \) should be stochastic in nature. To properly correct the deterministic
nature of this approximation, additional transit paths between phases need to be added, e.g., \( s^j_{G_1} \rightarrow s^j_{G_2} \). However, due to the complexity of the matching problem, which essentially requires a multivariate extension of PH distribution with multiple absorbing states, there is no analytical tool available for this task. Furthermore, under the current framework, parameters for the first jump from \( s_0 \) to \( s^1_{G_1} \) or \( s^1_{G_2} \) are functions of their target distributions only, say \( f^1_1(G_1) \) and \( f^1_2(G_2) \). Consequently, when these parameters are not properly set to offset the deterministic nature of the followed phase transitions, additional bias are introduced.

To see how those biases would affect the problem, a comparative example is given below with this simple model between two approximation schemes. Namely, Scheme A: approximating with PH distribution for matching up to first two moments; and Scheme B: approximating with PH distribution for matching only up to first moment only, namely, by using an exponential distribution. The detailed distributions and parameters are given in Table 1; see also Equation (1) for detailed formulation.

### Table 1: Distributions and Parameters for Comparative Example

<table>
<thead>
<tr>
<th>( c(s_0) )</th>
<th>( c(s_1) )</th>
<th>( c(s_2) )</th>
<th>( \beta )</th>
<th>( G_1 )</th>
<th>( G_2 )</th>
<th>( E_1 )</th>
<th>( E_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0.5</td>
<td>0.02</td>
<td>UNIF(0,24)</td>
<td>UNIF(3,21)</td>
<td>EXPO(5)</td>
<td>EXPO(3)</td>
</tr>
</tbody>
</table>

In addition to above two approximation schemes, empirical evaluated values are also provided as a benchmark. And three types of criterion are considered here, namely, the value functions with DC criterion: \( J(s_0), J(s_1), \) and \( J(s_2) \), the transition probability: \( p_{s_0,s_1}, \) and \( p_{s_0,s_2} \), as well as the steady state distribution: \( \pi_0, \pi_1, \) and \( \pi_2 \). It needs to be pointed out that since two uniform distributions \( G_1 \) and \( G_2 \) are to be approximated here, which have coefficient of variation less than 0.5, generalized Erlang structures [13] are used for their approximations.

### Table 2: Results for Comparative Example with Approximation Scheme A & B

<table>
<thead>
<tr>
<th></th>
<th>( J(s_0) )</th>
<th>( J(s_1) )</th>
<th>( J(s_2) )</th>
<th>( \pi_0 )</th>
<th>( \pi_1 )</th>
<th>( \pi_2 )</th>
<th>( p_{s_0,s_1} )</th>
<th>( p_{s_0,s_2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empirical Value &amp; 56.51 &amp; 60.36 &amp; 54.94 &amp; 0.6770 &amp; 0.2045 &amp; 0.1185 &amp; 0.5003 &amp; 0.4997</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appr. Scheme A &amp; 52.18 &amp; 56.53 &amp; 50.64 &amp; 0.7433 &amp; 0.1180 &amp; 0.1387 &amp; 0.3397 &amp; 0.6620</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appr. Scheme B &amp; 58.07 &amp; 61.88 &amp; 56.20 &amp; 0.6 &amp; 0.25 &amp; 0.15 &amp; 0.5 &amp; 0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The results in Table 2 shows that severe parameter biases are introduced in Scheme A, such that, all its approximated values with respect to the listed criterion are worse than Scheme B. The results are reasonable, since each PH distribution utilizes a unique structure to approximate its target distribution, and when there is no coordination between multiple approximations, e.g., additional transit paths or corrections towards parameters for first transits, those different structures can make the overall approximation unreliable.

### 3. Parameter Uncertainty Formulation and A Parametric Approach

#### 3.1 Approximated Nominal Problem and Its Parameter Uncertainty

To avoid additional parameter biases, approximation scheme B is used in the rest of this paper. Following such scheme, an approximated discrete time Markovian model given by a tuple \( \{S,A,\hat{P},g\} \) can be obtained with a uniformization procedure [10], and we denote it as a nominal problem for the original process. In here, \( \hat{P} \) is the approximated transition matrix, and \( g \) is the derived one-stage cost function with uniformization. A value function \( V^\pi_{\hat{P}}(s) \) is then associated with the nominal problem, and for simplicity we write it in vector form as follows:

\[
V^\pi_{\hat{P}} = (I - \alpha \hat{P})^{-1} g, \quad (4)
\]

where \( \alpha \in (0,1) \) is the derived discount rate for the nominal problem, \( I \) is identity matrix, and \( g \) is vector of cost functions. Under a DC criterion, it is known that an optimal policy exists, e.g., \( \pi^*_\hat{P} \), and would simultaneously minimize all components of this value function, i.e., \( V^\pi_{\hat{P}} \leq c_w V^\pi_{\hat{P}}, \forall \pi \in \Pi \). For the indirect approach, the control goal is usually to find this policy and hope it serves as a minimizer for the continuous time problem as well.

However, this may or may not be true since the optimal policy for the nominal problem does not necessarily infer an optimal, or even a good, policy for the original process. In addition to this approximation, we consider a parameter uncertainty formulation of the problem here. The reason to propose such formulation is because it is known that the approximated problem, and MDP problems in general, can be very sensitive to small parameter changes [15].
Although certain parameter biases can be avoided by using approximation scheme B, approximation errors are not completely eliminated due to the unmatched higher-order moments. As a result, a better approximation may exist in the sense that it leads to more appealing policies for the original process.

Departing from the nominal problem above, we consider the following uncertainty set for the approximated transition matrix. Given an approximated transition probability matrix $\hat{P}$, e.g., with an exponential approximation, let $p_{ij}$ be component of its $i$th row and $j$th column. We then define the following set of all possible realizations of approximated transition matrices within an interval, as $\mathcal{P} = \{P | p_{ij} \leq p_{ij} \leq \bar{p}_{ij}; \sum_j p_{ij} = 1\}$, where $\{\underline{p}_{ij}, \bar{p}_{ij}\}$ are some given lower and upper bounds of approximated transition matrices. The interval matrix here then models the potential parameter deviations from an exponential approximation due to the unmatched higher-order moments. It is worth to point out that although our formulation draws similarity with uncertain MDP problems studied in the past, e.g., in [16], the problem considered here is different in several ways. First, most of the parameter uncertainty problems considered in the literature are approached via parameter estimation problems, while the uncertainty here is approached via model approximation. And secondly, under our formulation, changes in the approximations, e.g., picking a different $\hat{P} \in \mathcal{P}$, will lead to different derived discount rate $\alpha$ and cost function $g$, which are usually assumed to be fix in existing studies for uncertain MDP problems. Therefore our formulation serves as a generalization to such other approaches.

To formally address the optimization problem to be solved, we first introduce a new performance index for the GSMDP problem. The reason is that unlike its approximated discrete time MDP problem, a Markovian and determinisitic policy that can simultaneously minimize the value function for all initial states in Equation (1) may not exist for GSMDP formulation (2). Therefore, a new performance index is used here to evaluate different policies. Let $(s, a) | s \in S, a \in A, t \geq 0\} be a stochastic process of GSMDP type under some control policy $\pi$, the performance index is defined as: $W^\pi \triangleq E^\pi \{G(s_i, a, t)\}$, where $G : S \times A \times T \rightarrow \mathbb{R}^+$ servers as a general functional of cost. Let $\hat{P}$ be an approximation drawn from the uncertainty set $\mathcal{P}$ above. We then associate each approximation $\hat{P}$ with a policy $\pi_{\hat{P}}$, denoted as nominal-optimal policy, which serves as a minimizer for Equation (4). Furthermore, the union of these nominal-optimal policies are defined as: $\Pi_{\hat{P}} \triangleq \bigcup_{\hat{P} \in \mathcal{P}} \{\pi_{\hat{P}}\}$, and we denote it as transition-induced nominal-optimal policy set. To optimize the original GSMDP problem, we then consider the following minimization problem:

$$\min_{\pi \in \Pi_{\hat{P}}} W^\pi. \quad (5)$$

Essentially Equation (5) allows one to examine different approximations within the uncertainty set as well as their associated nominal-optimal policies for the original GSMDP problem.

### 3.2 A Heuristic and Parametric Approach

To find the policy that minimizes Equation (5), one needs to go through all possible transition matrices within the uncertainty set $\mathcal{P}$. This is impractical due to the complexity of the search space, and usually a robust solution is used as a trade-off. However, since our formulation here differs from existing parameter uncertainty formulation and serves as a generalized problem, the robust approach cannot be used, e.g., due to the uniformized model parameters changes for different pick of $\hat{P}$. Instead, a new parametric and simulation-based approach is investigated in the sequel.

It needs to be pointed out that $\Pi_{\hat{P}} \subset \Pi$, and in addition, each transition-induced nominal-optimal policy is defined over the products of the approximated transition matrix and the one-stage cost function, e.g., $(1 - \alpha \hat{P})^{-1} g$. Thus, preference of a policy would be impacted by changes in the approximated transition matrix $\hat{P}$, as well as changes in the one-stage cost function $g$. Therefore, we consider a heuristic version of the problem as follows:

$$\min_{\pi \in \Pi_{\hat{P}}} W^\pi, \quad (6)$$

where $\Pi_{\hat{P}} \subset \Pi$ is said to be the cost-induced nominal-optimal policy set. Before formally defining $\Pi_{\hat{P}}$, we first introduce a parametric form of cost function. It is assumed here that the cost function $g(s)$ for a nominal problem, e.g., with approximated transition matrix $\hat{P}$, can be always decomposed into a set of $n$ basis functions $\{g^{(1)}(s), g^{(2)}(s), \ldots, g^{(n)}(s)\}$. Our assumption here is reasonable, since complex process tends to exhibit a local utility preference over different parts of the process [17]. And for queuing networks, natural preference decomposition towards different sub-queues within the process can be utilized. We thus introduce a parametric form of the cost function as $g_0(s) = \sum_{i=1}^{n} \theta_i g^{(i)}(s)$, where $g^{(i)}(s)$ corresponding to cost for a sub-queue within the network. The parameter $\theta_i$, here, denoted as cost coefficients, satisfies $\theta_i \geq 0, \forall i = 1, 2, \ldots, n$, and $\sum \theta_i = 1$. When the approximated transition matrix $\hat{P}$ is held fix, then a nominal-optimal policy $\pi_{\hat{P}, 0}$, or simply $\pi_0$, is uniquely determined by its cost coefficients $\theta_0 \triangleq [\theta_0, \theta_1, \ldots, \theta_{n}]^T \in \Theta$; c.f. Equation (4). Let the union of those nominal-optimal policies be $\Pi_{\hat{P}} \triangleq \bigcup_{m=1}^{\infty} \{\pi_0^m\}$, which we denote as cost-induced nominal-optimal policy set. The minimization problem in Equation (6), then can be equivalently understood as one
offering the best parameter for the underlying cost function, as follows:

$$\min_{\theta \in \Theta} W^{\pi^*}. \tag{7}$$

The optimization problem is well defined, with the heuristic assumption that the two policy sets above have big overlap over each other. We do not provide a detailed discussion of conditions for this here, since certain analytical arguments are needed, and they are surely beyond the presentation of this paper. In the sequel, we discuss efficient algorithms used in searching both cost-induced nominal-optimal policy set and parameter space of cost functions, along with an integrated framework for solving the minimization problem in Equation (7).

### 3.3 Extended Actor-Critic Framework

In order to find a set of optimal cost coefficients for the process, a new simulation-based framework denoted as extended actor-critic is introduced in this study, as shown in Figure 1. This framework contains two layers, namely, the outer layer and the inner layer. The formal one is further denoted as the policy space layer, since it is used to guide the parameter updating for the cost coefficients and the search of nominal-optimal policies; while the latter one is denoted as the value space layer, and serves to find each of those policies. In the following, we discuss algorithms used in both layers.

![Figure 1: Extended Actor-Critic Framework](image)

We start with the inner layer, the problem here is to find nominal-optimal policies that minimize the value function, given the approximated transition matrix \( \tilde{P} \) and the corresponding cost coefficient \( \theta \) obtained from the policy space (outer) layer. Following our early discussions, it is known that this corresponds to solving a classic discrete MDP problem, and therefore many algorithms can be used, e.g., value iteration, policy iteration, and linear programming. In here, we are interested in not only finding these policies, but also want to evaluate it against the actual process via simulations. This motivates the use of algorithms that can be used with simulation models and have on-line updating capability. Therefore, for the inner layer, the well-known approximate policy iteration scheme is adopted, along with temporal difference (TD) algorithm for its value function learning \([10, 18]\). This introduces three components of the inner layer: 1) a simulation model for the approximated MDP problem; 2) a critic part for learning the value function; and 3) an actor part for running policy iterations. For more detailed discussion of the algorithmic use of the approach, we refer to our earlier work \([19, 20]\).

Finding the nominal-optimal policies only solves half of the puzzle, essentially one wants to compare different policies within our heuristic set \( \Pi_C \) in order to find the overall optimal one for the actual process. The outer layer here is used for this purpose. Due to the continuous search space for \( \theta \), even if each nominal-optimal policy can be determined relatively quickly, it is still impractical to have a complete and exhaustive search over the entire parameter space. Therefore an adaptive approach is needed. To pick a good search algorithm here, we first list two facts that help our selection of the search algorithms:

- The relationship between \( W^{\pi^*} \) and \( \theta \) is complicated, and the direct gradient information \( \nabla W^{\pi^*} \) is not available.
- The measurement of \( W^{\pi^*} \) is conducted with simulations, therefore certain noise level is expected.

For search problems of this kind, stochastic approximation (SA) approaches have been extensively studied and well tested. In here, we adopt an adaptive version of the SA approach, which is known as Simultaneous Perturbation Stochastic Approximation (SPSA) for our parameter search task \([21]\). It is known that SPSA algorithm does not require exact gradient information and requires less amount of performance evaluations compared with other approaches \([22]\).
while global convergence property for SPSA can be guaranteed for a large spectrum of objective functions. The SPSA algorithm is summarized below:

\[
\hat{\theta}_{k+1} = \hat{\theta}_k - \alpha \hat{h}_k(\hat{\theta}_k),
\]

where \(\hat{h}_k(\hat{\theta}_k)\) is an approximation to the gradient:

\[
\hat{h}_k(\hat{\theta}_k) = (2c_k\Delta_k)^{-1}(y_k^+ - y_k^-).
\]

c_k here denotes the perturbation level at each step, while \(\Delta_k\) is vector of perturbed direction, that follows standard assumptions, e.g., vector of mutually independent zero mean Bernoulli trials. In here, the inverse of vector \((2c_k\Delta_k)^{-1}\) corresponds to the vector of inversed element, and \(y_k^+\) and \(y_k^-\) are defined as noisy performance measure of the underlying process, probably through simulated experiments:

\[
y_k^+ = W\pi^*(\hat{\theta}_k + c_k\Delta_k) + \varepsilon^+_k,\quad (10)
\]
\[
y_k^- = W\pi^*(\hat{\theta}_k - c_k\Delta_k) - \varepsilon^-_k.\quad (11)
\]

where \(\varepsilon^+_k\) and \(\varepsilon^-_k\) are noise associated with each evaluation trial. The extended actor-critic framework introduced here is able to provide a complete treatment for the control problem associated in non-conventional queueing networks. Next, we present a case study utilizing the proposed approach.

4. Case Study
4.1 Model and Assumptions
We consider here a special type of multiclass queueing network that is usually known as a reentrant queue. Reentrant queueing networks are often used to model semiconductor fabrication operations. In this case study, we consider the control problem of a relatively small network, as shown in Figure 2. There are two independent stations associated with the process, each with two shared processing steps. It is assumed here there is only a single type of product in this network and will need to go through all four processing steps sequentially. Departing from models with classic assumptions studied before, the service time here at each station for different processing steps are allowed to be non-exponential with general distributions, with details summarized in Table 3. The external arrival to the process is assumed to be Poisson type and with a rate of 0.13 per hour. We then consider the state of the process to be the tuple of queue lengths at each buffer along with the occupancy of the service stations from each processing step respectively. And with a realistic consideration, we limit each buffer size to be 20. The network is said to be open type, and the external arrivals are released to the process once they arrived if there are additional slots in buffer 1, otherwise they will be held temporarily and to be released later. The control goal here is to find an non-preemptive Markovian and deterministic control policy for station 1 and 2 with their shared competing buffers, e.g., buffer 1 & 3, and buffer 2 & 4, in order to minimizes the production cycle time (CT).

![Figure 2: A Reentrant Manufacturing Process](image)

It is worth to point out that if infinite buffer size and exponential service time for each processing step are assumed, then the state transitions here constitute a discrete Markovian chain. However, this can be a bad model for such a manufacturing process. This is because with the strict decreasing density function of exponential distribution, it implies that the highest probability of service time would be near zero. In this study, we model the service time as a composition of two parts, starting with a deterministic period of time, and followed by a stochastic period of time, of exponential type; the particular choice here was made to allow certain variability of the process to be considered, and other arbitrary choices could be make as well. Furthermore, we restrict the control policy here to be non-idling for each stations, e.g., the control policy would always select buffers with non-zero queue length to be served next.

In order to minimize the production cycle time, a common approach is to minimize functionals of work-in-process (WIP), based on the well-known Little’s Law and its extensions. In this study, we use quadratic functions of
Table 3: Machine Processing Rate

<table>
<thead>
<tr>
<th>Proc. Rate (Hour)</th>
<th>Step 1</th>
<th>Step 2</th>
<th>Step 3</th>
<th>Step 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.37+EXPO(3.71)</td>
<td>0.3+EXPO(3.03)</td>
<td>0.18+EXPO(1.82)</td>
<td>0.37+EXPO(3.71)</td>
</tr>
</tbody>
</table>

each queue length at those buffers. Our selection here has two reasons, they are: 1) quadratic functions gives better differentiations for networks with limited capacities; and 2) it is more sensitive to the changes of parameter $\theta$ as indicated in our previous work [19].

4.2 Simulation Results

In this part, we first discuss parameters and conditions used in this study, later we present our simulation results. It needs to be pointed out that simulation models for both the approximated model and actual process are implemented in the simulation tool Arena [28]. Next, we start with the parameters used for both the TD($\lambda$) algorithm and SPSA.

Some key parameters and settings for temporal difference algorithm are summarized below:

- The forgetting factor $\lambda$ is selected with 0.95.
- Linear function approximation is used for the cost-to-go, with basis functions selected to be quadratic functions of the individual queue length at each buffer.
- The control actions are updated following both greedy and optimistic rules [10].

Key parameters for SPSA are given below:

- Perturbations to the parameter $\theta$ starts at 0.12, with a decaying factor of 0.3.
- $\theta_0$ is initialized to $[0.25 \ 0.25 \ 0.25 \ 0.25]^T$ to indicate a non-preferential setting towards any buffer.

Simulation conducted with this study are given sufficient length in order to have a reliable estimate of the key performance index as well as policy learning, in particular, all cycle time are measured with 150 replications of a 400 day simulation length. It further needs to be pointed out that with each selection of $\theta$, all underlying parameters of the inner layer are reset to their initial value, so the inner layer does not keep memories of past learned policies.

In next, we present results of the simulation study. First, the learning curve of parameter $\theta$ for cost function is given in Figure 3, and in Figure 4, its stepwise absolute increment is given.

It can be observed the parameter $\theta$ converges after about the first 300 updates. Since the updates of the $\theta$ is based on the approximated gradient of the cycle time, it is worth to point out that this indicates slightly more weights for buffer 3 & 4 will generally leads to a lower cycle time. In Table 4 the measured cycle time and its corresponding 95% confidence interval (CI) are listed for policies learned before and after the parameter updates for $\theta$.

Table 4: Measured Cycle Time

<table>
<thead>
<tr>
<th>No. of $\theta$ Updates</th>
<th>CT (hour) with 0.95 CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>80.56 ± 0.49</td>
</tr>
<tr>
<td>800</td>
<td>75.34 ± 0.44</td>
</tr>
</tbody>
</table>
The results here thus justifies the parameter changes of $\theta$. After 800 updates, the newly derived nominal-optimal policy gives around 6.5% improvements over the original one. For the sequencing problem here, it is always more interesting to visualize the derived nominal-optimal policies. In here, due to the dimensionality of the problem, we give marginalized difference plots with respect to nominal-optimal policies derived with $\theta_0$ and $\theta_{800}$. The difference of the control policies here for station 1 and 2 are then marginalized over the rest of the state space. The results are shown in Figure 5 and 6, where the percentage difference scale are attached. It can be seen that, for station 2, the major difference for prioritizing between buffer 2 and 4 lie in a relative small region in upper left corner of the state space, while this difference region becomes much larger for station 1. This suggests that the prioritizing discipline between buffer 1 and 3 is a key factor in the control policy for reducing cycle time for this process.

\begin{figure}[h]
\centering
\includegraphics[width=0.4\textwidth]{fig5.png}
\caption{Policy Difference for Station 1}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=0.4\textwidth]{fig6.png}
\caption{Policy Difference for Station 2}
\end{figure}

5. CONCLUSIONS

In this study, we investigated the control problem for queueing network with general assumptions of the arrival and service distributions. Due to the induced concurrent events with general sojourn distributions, equivalent discrete Markovian model for those type of networks do not exist. Different from popular fluid and Brownian approximation methods, an indirect approach by approximating network’s state transition dynamics with continuous phase-type distribution is investigated. In particular, we examined parameter bias of such approximation under what we call a general competing transition scenario, and compared two approximation schemes. Our investigation suggests severe approximation bias can be introduced with existing algorithms when seeking higher-order approximations. Later in this study, a parameter uncertainty formulation of the optimization problem was proposed, aiming to alleviate the potential side-effect from model approximation, and also seeking an overall optimal policy for the problem. We gave a heuristic solution to the problem by using a parametric and simulation-based ADP approach. A case study was presented by using the proposed approach for sequencing control of a small manufacturing process with general service time. The results here validate the proposed approach investigated in this paper.

References


Chen and Fernandez


