Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

SAMPLE-PATH ALGORITHM FOR GLOBAL OPTIMAL SOLUTION OF RESOURCE ALLOCATION IN QUEUEING SYSTEMS WITH PERFORMANCE CONSTRAINTS

Mengyi Zhang

Mechanical Engineering
Politecnico di Milano
Via La Masa 1
Milan, 20156, ITALY

ABSTRACT

Resource allocation problems with performance constraints (RAP–PC) are a category of optimization problems on queueing system design. They can be often found in operations management of manufacturing and service systems. RAP–PC aims at finding the system with the minimum cost while guaranteeing a target performance, which usually must be obtained by simulation due to complexity of practical systems. It proposes an algorithm providing a sample–path exact solution within finite time. Specifically, the algorithm works on the mathematical programming model of RAP–PC and uses logic–based exact and gradient–based approximate feasibility cuts to define and reduce the feasible region. Results show that the proposed approach can solve at optimality problems on lines with up to 9–stage within two hours and feasible good quality solutions can be found faster than the state–of–the–art algorithm.

1. PROBLEM DEFINITION

The resource allocation problem (RAP) represents a category of optimization problems in queueing system design, which deals with the decision about the amount of resources, such as servers and queue capacity, to allocate to each system stage. The main aspect of RAP, from an operational perspective, is the trade–off between the system performance and the overall cost. Roughly speaking, increasing the quantity of the allocated resources allows to reach a higher performance, but, on the other hand, it will also increase the cost. RAPs can be categorized depending on whether the performance and/or the cost is managed through objective function or constraints. This work considers, specifically, the RAP with performance constraints (RAP–PC) in queueing systems. In the studied problem, once a set of performance indicators have been identified, a target is set to each of them as constraint, and the objective is to achieve the minimum total cost. RAP–PC can be mathematically formulated as follows:

$$\min_{\mathbf{x}\in\mathbb{X}}\{\mathbf{c}^T\mathbf{x}\}$$
$$s.t.\ \ h_l(\mathbf{x}) \le p_l^*, \forall\, l = 1, \dots, L$$

The vector $\mathbf{x}$ is the compact representation of the set of all the decision variables, i.e., the capacity of each stage. The initial search area $\mathbb{X}$ is considered as a box–shape subset of integer–ordered lattice. The set $\mathbb{X}$ should be well defined to avoid unstable systems. Vector $\mathbf{c}$ collects in a compact form the cost of each single resource. The left hand side of the constraints represents the performance of the system with resource capacity $\mathbf{x}$, and the right hand side is the target performance. Multiple performance constraints (indexed by l) can be handled. The variable vector $\mathbf{x}$ is not a generic integer–ordered variable vector; instead, it is a resource–type variable vector. A variable is defined as resource–type if the system performance is monotonic on it.

2. METHODOLOGY

As queueing systems representing practical settings are usually subject to blocking, dispatching policies and non-Markov property, discrete event simulation (DES) is one of the most used tools for performance evaluation, and, hence, simulation–optimization algorithms have to be used for RAP–PC. This work proposes a sample average approximation (SAA) algorithm, which finds a global optimal solution for RAP–PC of given sample paths within finite time. Once applying SAA, a stochastically constrained problem becomes a deterministic problem, but still with unknown feasible region.

The algorithm is based on a MIP whose feasible region is defined by feasibility cuts, which are iteratively generated and added to the MIP. The main focus of this paper is how to generate and manage the feasibility cuts so as to find good solution at early time of the solution process and guaranteeing the global optimality at its termination. Gradient–based approximate cuts are first proposed. The gradient estimation on the integer variable, i.e., the resource capacity, is enabled by exploring the structure of the DES through its mathematical programming representations (MPR) (Chan and Schruben 2008). This gradient estimation approach uses the simulation experiment related to a single design point, which provides the algorithm high efficiency in finding good solution within finite time. Using gradient to generate approximate cut represents the original contribution of this work. Then, logic–based exact cuts are applied for finding the global optimum.

3. NUMERICAL RESULTS

The proposed approach is applied to the server allocation problem of serial–parallel queueing system with inter-stage buffers of finite capacity. The objective is to find the server number in each stage that allows to achieve the minimum cost while guaranteeing that the average system time over all the jobs does not exceed a target value. Jobs arrive at the first stage of the system, which has a buffer with infinite capacity, with a general arrival process. After being processed by the last server stage, each job immediately leaves the system.

Numerical analysis has shown that the proposed approach can solve up to 9–stage problems within two hour time limit. A feasible solution with good quality, which is proved to be the global optimum at the end, can be found at very early time. For the 9–stage system with truncated normal distributed inter-arrival time and service time, the algorithm terminates in 5025.5 seconds with an optimal solution equal to 63, but a feasible solution equal to 63 has been visited at time 1.0 second, after having visited only 9 solutions.

Comparing with state-of-the-art approach, which is a combination of the adaptive hyperbox algorithm (Xu, Nelson, and Hong 2013) and the penalty–type framework proposed in Park and Kim (2015), the proposed approach can provided better solution within shorter time. The superiority is significant especially in higher dimension problems. For a 20-stage case, the proposed approach finds a feasible solution with good quality with computational time equal to 395.9 seconds on average, while the state-of-the-art approach cannot find the solution with the same quality within 2-hour time limit.

REFERENCES

Chan, W. K., and L. Schruben. 2008. “Optimization models of discrete-event system dynamics”. *Operations Research* 56(5):1218–1237.

Park, C., and S.-H. Kim. 2015. “Penalty function with memory for discrete optimization via simulation with stochastic constraints”. Operations Research 63(5):1195–1212.

Xu, J., B. L. Nelson, and L. J. Hong. 2013. “An adaptive hyperbox algorithm for high-dimensional discrete optimization via simulation problems”. INFORMS Journal on Computing 25(1):133–146.