Advance Service Reservations with Heterogeneous Customers

1 Introduction

We study a fundamental model of resource allocation in which a finite number of resources must be assigned in an online manner to a heterogeneous stream of customers. The customers arrive randomly over time according to known stochastic processes. Each customer requires a specific amount of capacity and has a specific preference for each of the resources, with some resources being feasible for the customer and some not. The system must find a feasible assignment of each customer to a resource or must reject the customer. The aim is to maximize the total expected capacity utilization of the resources over the time horizon.

This model has application in multiple areas, including services, online advertising, and freight transportation. We now explain a few of the applications.

Service Reservation. In services such as healthcare, the resources can correspond to service sessions. For example, a resource might be a Monday afternoon session from 1 to 5 PM with Dr. Smith. The customers are patients who arrive to book appointments over time. Based on a patient’s urgency, type of visit, arrival time, and preferences, the patient might require a specific length of visit and might be preferably assigned only to a subset of sessions. Upon the arrival of a patient, the system has to reserve a part of a session for the patient. This appointment decision typically takes place immediately. If an appointment cannot be found, the system must reject the patient.

Generalized Adwords. In online advertizing, the resources correspond to advertisers.
The capacity of each resource corresponds to the budget of the corresponding advertiser. Ad impressions arrive randomly over time. Each impression, depending on its characteristics, commands a known non-negative bid from each of the advertisers. When an impression occurs, the ad platform must allocate it to an advertiser for use. The ad platform earns the bid, and the budget of the advertiser is depleted by the same amount. The aim of the ad platform is to maximize the expected revenue earned. Our model is more general than adwords models, as we allow bids to have arbitrary sizes, whereas adwords model tend to assume that bid sizes must be very small relative to the budgets, or that each bid must be truncated by the remaining budget (Mehta 2012).

Our model captures most, if not all, of the features of the above applications. Specifically, we consider a continuous-time planning horizon. There are \( m \) resources with known capacities. There are \( n \) customer types. Each customer type is associated with a known stochastic arrival process. Each customer can be assigned to a known subset of the resources, and consumes a known amount of each resource that it is assigned to. The system aims to assign customers to resources immediately and irrevocably as they arrive in order to maximize the total expected amount of resources used.

A salient feature of our model is that arrivals can be non-stationary. In real applications, demands can be highly non-stationary, changing with the time of day, time of week, seasons and longer-term trends (Huh, Liu and Truong 2012). Kim and Whitt (2014) have shown, for example, that call-center and hospital demands are well-modeled by non-homogeneous Poisson processes. For a problem that essentially aims to match demand with supply over time, capturing this non-stationarity in demand arrivals can lead to significant improvements in performance over stationary models.
In this paper, we will focus on solution methods for the core model. We propose 0.321-competitive online algorithms. Further, we show that an upper bound on the competitive ratio of any algorithm is 1/2. Ours are the first algorithms with performance guarantees for the advance reservation of service with heterogeneous customer needs and preferences. They are also the first algorithms with constant competitive ratios for the adwords problem without any assumption on the bid size and on the stationarity of the arrival process. Despite the conservative performance characterization, we show that our algorithms perform extremely well compared to two common heuristics as demonstrated on a real data set from a large hospital system in New York City.

References


