ABSTRACT

Title of Dissertation: ESSAYS IN STOCHASTIC MODELING AND OPTIMIZATION

Jiaqi Zhou
Doctor of Philosophy, 2021

Dissertation Directed by: Professor Ilya Ryzhov
Department of Decision, Operations, and Information Technologies

Stochastic modeling plays an important role in estimating potential outcomes where randomness or uncertainty is present. This type of modeling forecasts the probability distributions of potential outcomes by allowing for random variation in one or more inputs over time under different conditions. One of the classic topics of stochastic modeling is queueing theory.

Hence, the first part of the dissertation is about a stylized queueing model motivated by paid express lanes on highways. There are two parallel, observable queues with finitely many servers: one queue has a faster service rate, but charges a fee to join, and the other is free but slow. Upon arrival, customers see the state of each queue and choose between them by comparing the respective disutility of time spent waiting, subject to random shocks. This framework encompasses both the multinomial logit and exponomial customer choice models. Using a fluid limit analysis, we give a detailed characterization of the equilibrium in this system. We show that social welfare is optimized when the express queue is exactly at (but not

over) full capacity; however, in some cases, revenue is maximized by artificially creating congestion in the free queue. The latter behaviour is caused by changes in the price elasticity of demand as the service capacity of the free queue fills up.

The second part of the dissertation is about a new optimal experimental design for linear regression models with continuous covariates, where the expected response is interpreted as the value of the covariate vector, and an "error" occurs if a lower-valued vector is falsely identified as being better than a higher-valued one. Our design optimizes the rate at which the probability of error converges to zero using a large deviations theoretic characterization. This is the first large deviations-based optimal design for continuous decision spaces, and it turns out to be considerably simpler and easier to implement than designs that use discretization. We give a practicable sequential implementation and illustrate its empirical potential.

ESSAYS IN STOCHASTIC MODELING AND OPTIMIZATION

by

Jiaqi Zhou

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:
Professor Ilya O. Ryzhov, Chair/Advisor
Professor Michael Fu
Professor Zhi-Long Chen
Professor Kunpeng Zhang
Professor Paul Smith

© Copyright by
Jiaqi Zhou
2021

Dedication

I dedicate this dissertation to my boyfriend, Tianyu Zhang, and my parents, Ke Zhou and Dong Xu.

Acknowledgments

First and foremost I'd like to thank my advisor, Professor Ilya O. Ryzhov for giving me an invaluable opportunity to work on several projects that I was really interested in over the past three years. When I first met with him three years ago, I almost had no idea what research is. However, with his patience and constantly been encouraging me to brainstorm, I finally made this dissertation to be possible. He has always made himself available for helping me whenever I had questions. Even during the COVID-19 period, he always replied to my e-mail within a couple of hours. It has been a pleasure to work with Professor Ryzhov and learn from such an extraordinary individual.

I would also like to thank Professor Michael Fu, Professor Zhi-Long Chen, Professor Kunpeng Zhang, and Professor Paul Smith for serving on my committee. I appreciate their invaluable advice on my dissertation.

I would like to thank all my friends and relatives. Especially Shihangyin Zhang, Yousheng Shi, Jiahao Su who encouraged me during my Ph.D. life.

I would like to acknowledge financial support from the Department of Mathematics, for all the projects discussed herein.

I owe my deepest thanks to my parents Ke Zhou and Dong Xu for their unconditional support and love. I also thank my grandparents for encouraging me to pursue a Ph.D. degree. Finally, I want to thank my boyfriend, Tianyu Zhang, and my kitty, Mineuchi for the happiness they have been bringing to me.

It is impossible to remember all, and I apologize to those I've inadvertently left out.

Lastly, thank you all!

Table of Contents

List of Figures

Chapter 1: Introduction

1.1 Equilibrium analysis of observable express service with customer choice

In this discussion, we mainly focus on work that involves observable queues and heterogeneous customers. There are many other interesting problems that do not deal with those particular issues; for instance, [1] and [2] both consider pricing in queueing systems, but assume that system states are unobservable to customers and/or do not model individual customer decisions. [3] gives a survey of the broad rational queueing literature that encompasses these types of problems, and so we will not delve more deeply into them here.

Our paper has some commonality with the stream of literature on priority queues, where customers are given the option to receive faster service by paying a fee. This is usually accomplished by moving paying customers in front of non-paying customers [4], so that both types of customers are handled by the same set of servers (perhaps a single server). In some cases, the customers do not observe the queue state [5,6] or make any choice at all [7]. In other cases, customers observe the queue state but have no choice of service type: for example, in [8] customers do not choose their priority class. Many of these papers focus on single-server models, thus streamlining the issue of capacity. Recent work by [9] and [10] considered multi-server settings, but made the queue unobservable to the customers; [11] studied a multi-server priority queue with two customer classes, but did not include any form of customer choice.

The notion of customer heterogeneity has many possible meanings: customers may have different valuations of the service, different patience levels, or access to different levels of information. Many papers, for instance [12] or [13], introduce distinct customer classes, but assume homogeneity within any given class. In [14] or [15], rather than purchasing faster service, customers can pay to make the queue observable, though their utilities are homogeneous. Common approaches to representing customer heterogeneity include modeling purely exogenous, i.i.d. valuations of the service [16,17] or abandonment times [18], or using a linear disutility of waiting with a randomly generated slope [6,19,20]. [21] used the multinomial logit (MNL) choice model to represent heterogeneous customer decisions in an unobservable queue. Our paper uses a general random utility model within an observable system; the MNL model falls under our framework, but so does, for example, the recent exponomial choice model of [22].

Some authors have considered forms of express service that are closer to the one in our paper. For example, [23] allows customers to move to a separate "fast lane" by paying a fee; however, the fast lane is not explicitly represented in the model, so these customers essentially disappear from the system altogether. By contrast, a major distinguishing feature of our work is the inclusion of the state and service capacity of the express queue into the customer's decision. Two closely-related studies by [24] and [25] explicitly model express queues, also motivated by the application of paid lanes on highways. Both studies consider customer heterogeneity, but their focus is on time-dependent pricing rather than equilibrium analysis, making it more difficult to obtain tractable results. Thus, the analysis in [24] assumes linear disutility (with random slope); on the other hand, [25] uses the MNL choice model, but primarily relies on numerical simulations for insight. In comparison, we simplify the service provider's decision by considering the equilibrium performance of a fixed price, with the upside that we can obtain a much more detailed characterization of revenue and social welfare under much more general utility and choice models. Our analytical approach builds on a recent series of papers by [26–28], which to our knowledge were the first to study the equilibrium behaviour of paired queueing systems under MNL choice. However, the focus of these papers is on delayed information, and they do not include any dimension of pricing, optimization, or even the notion of choosing between two different service types (they assume that both queues have identical service rates).

1.2 A new rate-optimal design for linear regression

In this work, we derive a new optimal experimental design for linear regression models with continuous covariates, where the expected response is interpreted as the value of the covariate vector, and an "error" occurs if a lower-valued vector is falsely identified as being better than a higher-valued one. Our design optimizes the rate at which the probability of error converges to zero using a large deviations theoretic characterization. This is the first large deviations-based optimal design for continuous decision spaces, and it turns out to be considerably simpler and easier to implement than designs that use discretization. We give a practicable sequential implementation and illustrate its empirical potential.

Consider the linear regression model

$$y = \beta^\top x + \epsilon, \tag{1.1}$$

where $\beta \in \mathbb{R}^d$ is a fixed, but unknown vector of regression coefficients, $x \in \mathbb{R}^d$ is a vector of data, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a residual noise. The expectation $\mathbb{E}(y|x) = \beta^\top x$ is interpreted as the "value" of x. For example, the elements of x could represent various attributes of a combination treatment for cancer, with the response y being the health outcome [29]. We assume that x is "better" if $\mathbb{E}(y|x)$ is larger. The set of possible x need not be discrete.

Suppose that we have the ability to design the data vector: given a sample size of n, we may choose $x_1, ..., x_n$ anywhere in some compact subset of $\mathbb{R}^d$ called the "design space." This choice may be made either all at once, before any observations are collected, or sequentially, where each x_i may depend on $x_1, y_1, ..., x_{i-1}, y_{i-1}$, perhaps through a vector b_i of least-squares regression coefficients estimated using these previously collected data. The first, static setting has been extensively studied in the literature on *experimental design* [30]. Because static decisions are made without any information on the response, one builds the design to optimize some summary statistic of the covariance matrix of the least-squares estimator b_n. There are many possible criteria, known by such "alphabet-optimal" names as A-optimality [31], D-optimality [32, 33], G-optimality [34], etc. All of these evaluate designs in a purely statistical sense, with no other notion of the value of x.

The second, sequential setting has been considered by the community working on simulation-based optimization. This literature grew out of the *ranking and selection* problem, in which the goal is to identify the highest-valued alternative (unlike experimental design, ranking and selection always has some notion of value to maximize) from some finite set using independent samples of the value. An early effort to apply algorithmic concepts from ranking and selection to the linear regression setting was by [35], which also assumed that each x_i could take values only in a finite set; similar settings were considered by [36] and [37]. [38] provided approximation algorithms for combinatorial design spaces, while [39] and [40] handled low-dimensional, continuous design spaces with special structure (e.g., the value being a quadratic function of a scalar control). In the computer science literature, [41], [42] and others studied related "linear bandit" problems where one maximizes the total value of the sampled vectors.

However, the static setting can also be used to examine the problem of finding the best x, and this approach has yielded deep insights into the development of sequential algorithms. In the simulation community, [43] used large deviations theory to derive a new type of optimal design where the optimality criterion was connected to the value through the probability of incorrect selection (i.e., the event that a suboptimal alternative is erroneously estimated to have a higher value than the optimal one). Similar ideas motivate the literature on optimal computing budget allocation [44–46], which uses various approximations of this error probability. Later work by [47–49] generalized this notion to a broader class of simulation-based optimization problems. In all of these papers, however, the optimal design depends on the underlying unknown problem parameters (in regression, this is the vector β) which determine the value of each alternative. Thus, although an optimal static design exists, it cannot be computed statically, but rather must be learned as data are acquired. In a sense, the purpose of a sequential algorithm is to do this efficiently; see [50] and [51] for examples of such algorithms in the context of ranking and selection. The computer science literature has also developed similar insights, with [52] and [53] proposing sequential variants of G-optimal design.

1.3 Outline of Dissertation

In Chapter 2, we describe the stylized queueing system that we use for modeling express service with customer choice and presents the fluid limit approximation used to study the equilibrium behaviour of this system and investigates its properties. Then we found the dependence of the equilibrium solution on c will determine the shape of any relevant revenue function that we might define; for this reason, we start by examining this dependence in Section 2.3.1. Then, in Section 2.3.2, we propose and study two objective functions related to the revenue of the service provider and the social welfare of the customers. Finally, we present an additional analysis and numerical illustrations for the setting where customer choice follows the MNL model and the exponomial choice model.

In Chapter 3, we derive a new, large deviations theoretic optimality criterion for linear regression, and propose a new design that optimizes this criterion. We first derive the large deviations law for b_n. Then we focus on studying error events for countable collections $\{v_k\}_{k=1}^{\infty}$ that are dense in some uncountable set of interests, where v is the vector we are studying. In Section 3.2, we define and solve the optimization problem to make the probability of error events converge to zero at the fastest possible rate. Finally, we state a very simple algorithm (which we call "LD-optimal") for implementing the optimal design in practice and conduct a numerical experiment comparing this algorithm with some other algorithms.

Chapter 4 provides the conclusion to the thesis.

Chapter 2: Equilibrium analysis of observable express service with customer choice

2.1 Introduction

This work was motivated by an increasingly common sight in urban beltways and surrounding arterial highways – the availability of paid express lanes with higher speed limits. To reduce congestion in the transportation network, and to generate revenue for the state authority, drivers are given the option to pay a fee and gain access to a special set of lanes running parallel to the general-purpose lanes on the same highway. The magnitude of the entry fee has an impact on how many drivers are willing to make the switch, which also affects the quality of service in both free and paid lanes because the capacity of both types of lanes is finite. Thus, the entry fee can be used to manipulate the amount of congestion in the system, either to improve the overall quality of service or to maximize revenue.

This behaviour is not limited to transportation networks; there are other types of service systems where faster service can be obtained at a price, such as express lines in theme parks, or expedited service in document processing. In this paper, we develop a stylized queueing model that is somewhat abstracted from the specifics of any one application in particular, but provides insight into the broader problem of pricing in service systems with a paid express option. In our framework, there are two $M/M/\bar{q}$ queues operating in parallel. The "express" queue has a faster service rate, but charges a fixed fee to join, whereas the "regular" queue is slower but free. The value of the service itself is the same in either queue, but customers prefer to wait less and, upon arrival, will choose between the two queues based on their perception of the waiting times. The following key dimensions are present in the model:

1. Both queues are *observable*: a newly arriving customer will see the exact state of both queues at the moment of arrival, and determine whether the reduced (conditional expected) waiting time in the express queue is worth the entry fee.

2. Customers are *heterogeneous*: their valuations of waiting time are subject to random variation, reflecting their differing perceptions of the waiting times or of the inconvenience of waiting.

3. Both queues have limited service *capacity*: all else being equal, a newly arriving customer will be less likely to choose the express queue if all of its servers are busy and other customers are waiting in line.

In short, customer choice follows a random utility model and is based on the observed queue lengths at the moment of arrival. Thus, customers all have different willingness to pay and the magnitude of the entry fee affects the proportion of customers that prefer express service to regular. However, these proportions also depend on the queue lengths at any given moment and thus change dynamically over time even though the fee is kept fixed.

2.1.1 Contributions and Insights

We use a fluid limit equilibrium analysis; for other applications of this technique in service operations, see, e.g., [54] or [55]. We characterize the long-run average queue lengths and choice probabilities for both express and regular service, and then study the dependence of these quantities on the entry fee, which drives the behaviour of various objectives related to revenue and social welfare. Below, we summarize our key findings and insights.

Classification of equilibrium. The finite capacity of the system plays a vital role in the structure of the equilibrium. Given a fixed entry fee and a fixed set of other problem inputs, the equilibrium can belong to one of four "regimes" depending on whether the express and regular queues are above or below capacity. The distinctions between these regimes essentially determine the way in which the entry fee impacts revenue and social welfare.

Transitions of equilibrium as a function of price. If we vary the entry fee while keeping the other problem inputs fixed, the equilibrium changes: as one might expect, the express queue length decreases in the price, while the regular queue length increases. When one of the queues approaches capacity, the equilibrium transitions from one regime to another, completely changing the structure of revenue and social welfare.

We provide a full characterization of all possible sets of transitions. Any given set of problem inputs will yield one, and only one, of six possible cases. For example, in one of these cases, low prices will lead to congestion in the express queue and unused capacity in the regular queue; mid-range prices will cause enough customers to move to the regular queue so as to eliminate congestion entirely; and high prices will create congestion in the regular queue while leaving unused capacity in the express queue.

Social welfare. A natural way to measure social welfare in this problem is in terms of the expected disutility of waiting per arrival; in other words, a customer is better off when he or she spends less time in the system, regardless of whether it is in the express or regular queue. We find that, under virtually any utility function and choice model, social welfare is optimized by choosing a price that is high enough to avoid creating congestion in the express queue, but otherwise low enough to minimize congestion in the regular queue. Customers are not always better off if the express queue is free to join, because congestion in the express queue also reduces service quality.

Revenue maximization. We find that the shape of the revenue function is problem-specific and (depending on which of the six cases applies) there may be multiple locally optimal prices. This behaviour arises because the finite capacity of the regular queue effects a change in the price elasticity of demand. If both queues are under capacity, a new customer obtains a constant improvement in waiting time by choosing express over regular; however, as the price increases and the regular queue fills up, the benefit of switching to express starts to grow, partially offsetting additional price increases. One possible consequence of this phenomenon is that the revenue-maximizing price can artificially create congestion in the regular queue, while deliberately maintaining unused capacity in the regular queue, even though a different price may have eliminated all congestion entirely.

We note that these findings are obtained in a very general setting that encompasses many possible disutility functions and random choice models. If one makes additional assumptions, it is possible to obtain even more detailed characterizations – for example, under the MNL model, we derive the equilibrium queue lengths in closed form. However, the general setting also applies to, e.g., the exponomial choice model, and all of our general results continue to hold in that context.

2.2 Model and analytical framework

Section 2.2.1 describes the stylized queueing system that we use for modeling express service with customer choice. Section 2.2.2 presents the fluid limit approximation used to study the equilibrium behaviour of this system and investigates its properties.

2.2.1 Setup with general choice probabilities

Consider the following queueing system. Customers arrive according to a Poisson process with rate λ. Upon arrival, a customer can choose to enter one of two queues: a "regular" queue (free highway) with exponential service rate μ_r, or an "express" queue (paid express highway) with rate $\mu_e > \mu_r$. Each queue has $\bar{q}$ servers. If neither queue is desirable, in a certain sense to be defined, the customer may also choose an "outside option" (e.g., taking a back road) and leave the system entirely. Once the choice has been made, it cannot be revisited; if the customer chooses one of the queues, he or she remains in that queue until service is completed, and subsequently leaves the system. We assume that passing through the system (arriving at home) has the same positive value regardless of how it was achieved, so the choice between the three options (regular, express, outside) is made by comparing their respective disutility of time spent waiting. We focus on disutility (as does, e.g., 11) because, in the highway application, every commuter needs the service.

Let $Q_r(t)$ and $Q_e(t)$ denote the lengths of the two queues at time t. We will formally define the dynamics of the queue lengths at the end of this discussion; for now, let us focus on how they affect customer choice. Both queues are observable: the choice made by a customer arriving at time t will depend on $Q_r(t)$ and $Q_e(t)$. A customer expecting to wait s time units in a queue will evaluate the disutility of waiting as $u(s)$, where $u : \mathbb{R}_+ \to \mathbb{R}_+$ satisfies the following properties:

U1) The disutility of not waiting at all is zero, i.e., $u(0) = 0$.

U2) The disutility of waiting infinitely long is infinite, i.e., $\lim_{s\to\infty} u(s) = \infty$.

U3) Disutility strictly increases with the waiting time, i.e., $u' > 0$.

Thus, the "ideal" disutility of waiting in queue $i \in \{r, e\}$, evaluated by a customer arriving at time t, is given by

$$u_i(Q_i(t)) = \begin{cases} u(\frac{1}{\mu_i}) & Q_i(t) < \overline{q}, \\ u(\frac{Q_i(t)}{\mu_i \overline{q}}) & Q_i(t) \geq \overline{q}. \end{cases}$$

The disutility of the outside option is assumed to be a fixed positive number $\overline{u} > \max\{u(\frac{1}{\mu_e}), u(\frac{1}{\mu_r})\}$, meaning that it is not preferable to either queue as long as the latter is under capacity.

The *total* disutility of joining either queue, as evaluated by a customer arriving at time t, is given by

$$U_e(t) = u_e(Q_e(t)) + c + \tau_e, \tag{2.1}$$

$$U_r(t) = u_r(Q_r(t)) + \tau_r, \tag{2.2}$$

where the quantity $c \geq 0$ in (2.1) is the fixed dollar cost of joining the express queue; this term is absent from (2.2) since the regular queue is free. The terms τ_e,τ_r are random shocks (independent of each other, as well as the arrival process and queue lengths) used to model heterogeneity between customers, e.g., differences in their individual valuations, or differences in their individual perceptions of the queue lengths.

Thus, the disutility function u can be seen as converting waiting time into a dollar equivalent so that it might be directly traded off against the actual dollar cost of entering the express queue. This determines the endogenous arrival rates for the two queues. For example, a customer arriving at time t will choose the express queue if $U_e(t) \leq \min\{U_r(t), \bar{U}\}$ where $\bar{U} = \bar{u} + \tau_o$ is the disutility of the outside option subject to the random shock τ_o.

With additional assumptions on the distributions of τ_e, τ_r, τ_o, customer choice can be made to follow a standard choice model such as multinomial logit (if the shocks are Gumbel distributed) or exponomial (if exponentially distributed). We consider both of these examples in Section 2.4; for the time being, however, we work with the general form of the model. Denote by

$$p_e\Big(u_e(Q_e(t)) + c, u_r(Q_r(t)), \overline{u}\Big) = P\Big(U_e(t) \leq \min\{U_r(t), \overline{U}\} | Q_r(t), Q_e(t)\Big) \tag{2.3}$$

the probability (conditional on the observed queue lengths) that a customer arriving at time t chooses the express queue. Similarly, we can define conditional probabilities that the customer arriving at time t will choose the regular queue or the outside option. We will use the notation p_e, p_r, p_o to refer to these probabilities, sometimes without explicitly writing out their dependence on the various components of the disutility calculations in order to make the notation less cumbersome.

The choice probabilities are assumed to add up to 1 and satisfy the following conditions:

P1) All three probability functions (for example, the function $p_e(u_e, u_r, \bar{u})$ in (2.3)) are differentiable and have uniformly bounded first derivatives with respect to each argument.

P2) The derivative of each choice probability with respect to the disutility of that particular choice is strictly negative (for example, $\frac{\partial p_e}{\partial u_e} < 0$), whereas the derivative with respect to the disutility of a different choice is strictly positive. In words, if the disutility of joining a particular queue goes up, the probability of joining that same queue should go down, and the probability of joining a different queue should go up.

P3) For any δ, $p_i(u_e + \delta, u_r + \delta, \bar{u} + \delta) = p_i(u_e, u_r, \bar{u})$ for all $i \in \{r, e, o\}$. In words, the choice probabilities are unaffected if all the disutilities are changed by the same amount.

As will be discussed in Section 2.4, these assumptions can be verified for both the MNL and exponomial choice models.

We can now formally define the dynamics of the queue length processes. For $i \in \{r, e\}$, let Π_i^{arr}, Π_i^{dep} be independent Poisson processes with rate 1. Then,

$$Q_i(t) = Q_i(0) + \Pi_i^{arr}\left(\int_0^t \lambda p_i(u_e(Q_e(s)+c), u_r(Q_r(s)), \bar{u})ds\right) - \Pi_i^{dep}\left(\int_0^t \mu_i \min\{Q_i(s), \bar{q}\}ds\right), \tag{2.4}$$

with $Q_i(0) = 0$ by convention. Thus, the arrival rate of each queue depends explicitly on the choice probabilities, while the departure rate depends only on the queue lengths.

2.2.2 Equilibrium analysis using fluid limit

We analyze the long-run behaviour of (2.4) using a fluid limit approximation. Essentially, we construct a deterministic dynamical system that strongly approximates the scaled queue length processes

$$Q_i^n(t) = \frac{1}{n}\Pi_i^{arr}\left(n\int_0^t \lambda p_i(u_e(Q_e^n(s)+c), u_r(Q_r^n(s)), \bar{u})ds\right) - \frac{1}{n}\Pi_i^{dep}\left(n\int_0^t \mu_i \min\{Q_i^n(s), \bar{q}\}ds\right), \tag{2.5}$$

in the limit as $n \to \infty$. Essentially, we are scaling up the arrival and departure rates by a factor of n, resulting in a large number of customers passing through the system (as might happen during peak traffic), but we correspondingly scale the resulting numbers of arrivals and departures back down to the magnitude of the original process. This has the effect of averaging out the stochasticity in the choice probabilities, leading to a purely deterministic limit, which is rigorously justified in the following result.

Theorem 1. *The sequence of stochastic processes $Q^n(t) = (Q_e^n(t), Q_r^n(t))$ converges a.s. and uniformly on compact sets of time to the dynamical system $q(t) = (q_e(t), q_r(t))$ described by*

$$q_e'(t) = \lambda p_e(u_e(q_e(t)) + c, u_r(q_r(t)), \overline{u}) - \mu_e \min\{q_e(t), \overline{q}\}, \tag{2.6}$$

$$q_r'(t) = \lambda p_r(u_e(q_e(t)) + c, u_r(q_r(t)), \overline{u}) - \mu_r \min\{q_r(t), \overline{q}\}. \tag{2.7}$$

Proof. For convenience, we assume $Q_i(0) = Q_i^n(0) = q_i(0) = 0$. We follow [28] in using the following result from [56]:

Lemma 1. *A standard Poisson process $\{\Pi(t)\}_{t>0}$ can be realized on the same proba-*

bility space as a standard Brownian motion $\{W(t)\}_{t>0}$ *in such a way that the almost surely finite random variable*

$$Z \equiv \sup_{t\geq 0} \frac{|\Pi(t) - t - W(t)|}{\log(2 \vee t)}$$

has finite moment generating function in the neighborhood of the origin and in particular finite mean.

Lemma 1 allows us to rewrite Q_i^n in terms of two standard Brownian motions B_i^{arr} , B_i^{dep} via

$$\begin{aligned}
\frac{1}{n}\Pi_i^{arr}\left(n\int_0^t \lambda p_i(Q_e^n(s), Q_r^n(s))ds\right) &= \int_0^t \lambda p_i(Q_e^n(s), Q_r^n(s))ds \\
&\quad + \frac{1}{n}B_i^{arr}\left(n\int_0^t \lambda p_i(Q_e^n(s), Q_r^n(s))ds\right) + O\left(\frac{\log n}{n}\right) \\
&= \int_0^t \lambda p_i(Q_e^n(s), Q_r^n(s))ds \\
&\quad + \frac{1}{\sqrt{n}}B_i^{arr}\left(\int_0^t \lambda p_i(Q_e^n(s), Q_r^n(s))ds\right) + O\left(\frac{\log n}{n}\right)
\end{aligned} \tag{2.8}$$

and

$$\frac{1}{n}\Pi_i^{dep}\left(n\int_0^t \mu_i \min\{Q_i^n(s),\overline{q}\}ds\right) = \int_0^t \mu_i \min\{Q_i^n(s),\overline{q}\}ds$$
$$+\frac{1}{n}B_i^{dep}\left(n\int_0^t \mu_i \min\{Q_i^n(s),\overline{q}\}ds\right) + O\left(\frac{\log n}{n}\right)$$
$$= \int_0^t \mu_i \min\{Q_i^n(s),\overline{q}\}ds$$
$$+\frac{1}{\sqrt{n}}B_i^{dep}\left(\int_0^t \mu_i \min\{Q_i^n(s),\overline{q}\}ds\right) + O\left(\frac{\log n}{n}\right) \tag{2.9}$$

Now, we calculate the difference between the scaled length process for queue $i \in \{e, r\}$ and its fluid limit, given by

$$Q_i^n(t) - q_i(t) = \frac{1}{n}\Pi_i^{arr}\left(n\int_0^t \lambda p_i(Q_e^n(s), Q_r^n(s))ds\right) - \int_0^t \lambda p_i(q_e(s), q_r(s))ds$$
$$-\frac{1}{n}\Pi_i^{dep}\left(n\int_0^t \mu_i \min\{Q_i^n(s),\overline{q}\}ds\right) + \int_0^t \mu_i \min\{q_i(s),\overline{q}\}ds$$

whence

$$\mid Q_i^n(t) - q_i(t) \mid \leq \left|\frac{1}{n}\Pi_i^{arr}\left(n\int_0^t \lambda p_i(Q_e^n(s), Q_r^n(s))ds\right) - \int_0^t \lambda p_i(q_e(s), q_r(s))ds\right|$$
$$+\left|\frac{1}{n}\Pi_i^{dep}\left(n\int_0^t \mu_i \min\{Q_i^n(s),\overline{q}\}ds\right) - \int_0^t \mu_i \min\{q_i(s),\overline{q}\}ds\right| \tag{2.10}$$

Substituting (2.8) into the first term of (2.10), we obtain

$$\begin{aligned}
&\left|\frac{1}{n}\Pi_i^{arr}\left(n\int_0^t \lambda p_i(Q_e^n(s),Q_r^n(s))ds\right) - \int_0^t \lambda p_i(q_e(s),q_r(s))ds\right| \\
\leq &\left|\int_0^t \lambda p_i(Q_e^n(s),Q_r^n(s))ds - \int_0^t \lambda p_i(q_e(s),q_r(s))ds\right| \\
&+\left|\frac{1}{\sqrt{n}}B_i^{arr}\left(\int_0^t \lambda p_i(Q_e^n(s),Q_r^n(s))ds\right)\right| + O\left(\frac{\log n}{n}\right),
\end{aligned}$$

with the Brownian term satisfying

$$\begin{aligned}
\lim_{n\to\infty}\sup_{t'\leq t}\left|\frac{1}{\sqrt{n}}B_i^{arr}\left(\int_0^t \lambda p_i(Q_e^n(s),Q_r^n(s))ds\right)\right| &\leq \lim_{n\to\infty}\left|\frac{1}{\sqrt{n}}B_i^{arr}(\lambda t)\right| \\
&= 0.
\end{aligned}$$

Substituting (2.9) into the second term of (2.10) yields

$$\begin{aligned}
&\left|\frac{1}{n}\Pi_i^{dep}\left(n\int_0^t \mu_i \min\{Q_i^n(s),\overline{q}\}ds\right) - \int_0^t \mu_i \min\{q_i(s),\overline{q}\}ds\right| \\
\leq &\left|\int_0^t \mu_i \min\{Q_i^n(s),\overline{q}\}ds - \int_0^t \mu_i \min\{q_i(s),\overline{q}\}ds\right| \\
&+\left|\frac{1}{\sqrt{n}}B_i^{dep}\left(\int_0^t \mu_i \min\{Q_i^n(s),\overline{q}\}ds\right)\right| + O\left(\frac{\log n}{n}\right),
\end{aligned}$$

with the Brownian term satisfying

$$\begin{aligned}
\lim_{n\to\infty}\sup_{t'\leq t}\left|\frac{1}{\sqrt{n}}B_i^{dep}\left(\int_0^t \mu_i \min\{Q_i^n(s),\overline{q}\}ds\right)\right| &\leq \lim_{n\to\infty}\left|\frac{1}{\sqrt{n}}B_i^{dep}(\mu_i\overline{q}t)\right| \\
&= 0.
\end{aligned}$$

Thus, (2.10) has become

$$
\begin{aligned}
| \ Q_i^n(t) - q_i(t) \ | \leq & \left| \int_0^t \lambda p_i(Q_e^n(s), Q_r^n(s)) ds - \int_0^t \lambda p_i(q_e(s), q_r(s)) ds \right| \\
& + \left| \int_0^t \mu_i \min\{Q_i^n(s), \overline{q}\} ds - \int_0^t \mu_i \ \min\{q_i(s), \overline{q}\} ds \right| + o(n).
\end{aligned}
$$

The choice probability p_i and the departure function both have uniformly bounded derivatives by assumption P1, so there exist constants C and ϵ such that, for large enough n, we have

$$
| \ Q_i^n(t) - q_i(t) \ | \leq C \int_0^t \sup_{0 \leq s' \leq s} | \ Q_i^n(s') - q_i(s') \ | \ ds + \epsilon.
$$

Applying Gronwall's lemma [57], we obtain

$$
\sup_{0 \leq s \leq t} | \ Q_i^n(s) - q_i(s) \ | \leq \epsilon \cdot e^{Ct}.
$$

Letting $\epsilon \to \infty$ completes the proof. □

Theorem 1 provides us with the system (2.6)-(2.7), which can be studied to obtain insight into the long-run behaviour of the original queue. The validity of the fluid limit is argued analogously to [28];

The equilibrium of the system (2.6)-(2.7) consists of two values q_e, q_r satisfying the equations

$$\lambda p_e(u_e(q_e) + c, u_r(q_r), \overline{u}) = \mu_e \min\{q_e, \overline{q}\}, \tag{2.11}$$

$$\lambda p_r(u_e(q_e) + c, u_r(q_r), \overline{u}) = \mu_r \min\{q_r, \overline{q}\}, \tag{2.12}$$

which are obtained by setting the time derivatives in (2.6)-(2.7) equal to zero. The solution can also be related to the outside option through the equation

$$\lambda = \lambda p_o(u_e(q_e) + c, u_r(q_r), \bar{u}) + \mu_e \min\{q_e, \bar{q}\} + \mu_r \min\{q_r, \bar{q}\}. \tag{2.13}$$

Since we focus on the equilibrium from this point on, we abuse notation slightly by using q_e, q_r to denote the fixed solution to (2.11)-(2.13), rather than the time-dependent quantities in (2.6)-(2.7).

Remark 1 As will be shown further down in Theorem 2, the system (2.11)-(2.13) has a unique solution. However, there are four possible interpretations of this solution depending on which arguments attain the minima in (2.11)-(2.13). We call these the four possible "regimes" of the equilibrium:

R1) Both queues are over capacity ($q_e, q_r \geq \bar{q}$);

R2) Both queues are under capacity ($q_e, q_r < \bar{q}$);

R3) Only the express queue is over capacity ($q_e \geq \bar{q} > q_r$);

R4) Only the regular queue is over capacity ($q_r \geq \bar{q} > q_e$).

Different problem instances lead to different regimes: for example, if λ is very small, the equilibrium will likely be in regime R2, whereas if the entry cost c is very large, we may see regime R4. The distinctions between R1-R4 are quite important for pricing because, if we vary c while keeping the other problem inputs fixed, the equilibrium may "jump" from one regime to another, affecting the revenues earned from the express queue. Section 2.3 will explore this issue in much more detail.

The following results state some general properties of the equilibrium;

Theorem 2. *The equilibrium of the system (2.11)-(2.13) exists and is unique.*

Proof. Existence of the equilibrium. We first show the existence of the equilibrium using Brouwer's fixed point theorem, which states that, if f is a continuous function mapping a compact convex set to itself, there exists a point x_0 satisfying $f(x_0) = x_0$.

We rewrite the equilibrium conditions (2.11)-(2.12) as

$$\lambda p_e - \mu_e \min\{q_e, \bar{q}\} + q_e = q_e, \tag{2.14}$$

$$\lambda p_r - \mu_r \min\{q_r, \bar{q}\} + q_r = q_r. \tag{2.15}$$

We can then express the system (2.14)-(2.15) as $f(q) = q$, where $q = (q_e, q_r)$. Because we have assumed continuity of p_e, p_r (assumption P1), it straightforwardly follows that f is continuous.

To show that $f = (f_e, f_r)$ maps a compact convex set to itself, let us consider the first component f_e and suppose that $q_e < \bar{q}$. In this case, we have the bound $f_e(q_e) \leq \lambda + \bar{q}$.

When $q_e \geq \bar{q}$, we have $f_e(q_e) = \lambda p_e\Big(u\Big(\frac{q_e}{\mu_e \overline{q}}\Big), u_r, \overline{u}\Big) - \mu_e \overline{q} + q_e$. Note that, if $\bar{q} \geq \frac{\lambda}{\mu_e}$, then $f_e(q_e) < q_e$ and the codomain of f_e is automatically contained in the domain.

If $\bar{q} < \frac{\lambda}{\mu_e}$, let $\tilde{q}_e$ be a value satisfying

$$\lambda p_e\Big(u\Big(\frac{\tilde{q}_e}{\mu_e \overline{q}}\Big), u(\infty), \overline{u}\Big) = \mu_e \overline{q}.$$

Then, for $q_e \geq \tilde{q}_e$, we have

$$\begin{aligned} \lambda p_e\Big(u\Big(\frac{q_e}{\mu_e \overline{q}}\Big), u_r, \overline{u}\Big) &\leq \lambda p_e\Big(u\Big(\frac{\tilde{q}_e}{\mu_e \overline{q}}\Big), u_r, \overline{u}\Big) \\ &\leq \lambda p_e\Big(u\Big(\frac{\tilde{q}_e}{\mu_e \overline{q}}\Big), u(\infty), \overline{u}\Big) \\ &= \mu_e \overline{q}, \end{aligned}$$

implying $f(q_e) \leq q_e$. Finally, for $\bar{q} < q_e < \tilde{q}_e$, we have

$$\lambda p_e\Big(u\Big(\frac{q_e}{\mu_e \overline{q}}\Big), u_r, \overline{u}\Big) - \mu_e \overline{q} + q_e \leq \lambda p_e\Big(u\Big(\frac{1}{\mu_e}\Big), u_r(\infty), \overline{u}\Big) - \mu_e \overline{q} + \tilde{q}_e. \tag{2.16}$$

Denote by $\hat{q}_e$ the right-hand side of (2.16). Then, for any $0 \leq q_e \leq \max\{\overline{q}, \tilde{q}_e, \hat{q}_e, \frac{\lambda}{\mu_e}\}$, we have $f(q_e)$ in the same interval, regardless of q_e. Thus, the conditions for Brouwer's fixed point theorem hold and the equilibrium exists.

Uniqueness of the equilibrium. Let λ, μ_e, μ_r, and the disutility function u be given. Suppose that there are two non-identical equilibrium solutions $\Big(q_e^{(1)}, q_r^{(1)}\Big)$ and $\Big(q_e^{(2)}, q_r^{(2)}\Big)$. Let us focus on the case where $q_e^{(1)} < q_e^{(2)}$ (the other case where we start with $q_r^{(1)} < q_r^{(2)}$ is handled symmetrically).

We first show that, if $q_e^{(1)} < q_e^{(2)}$, then $q_r^{(1)} < q_r^{(2)}$ as well. To see this, let us assume the contrary, i.e., that $q_r^{(1)} \geq q_r^{(2)}$. We derive

$$\begin{aligned}
0 &= \lambda p_e\Big(u_e\Big(q_e^{(1)}\Big) + c, u_r\Big(q_r^{(1)}\Big), \bar{u}\Big) - \mu_e \min\{q_e^{(1)}, \bar{q}\} \\
&\geq \lambda p_e\Big(u_e\Big(q_e^{(2)}\Big) + c, u_r\Big(q_r^{(1)}\Big), \bar{u}\Big) - \mu_e \min\{q_e^{(1)}, \bar{q}\} \qquad (2.17) \\
&\geq \lambda p_e\Big(u_e\Big(q_e^{(2)}\Big) + c, u_r\Big(q_r^{(2)}\Big), \bar{u}\Big) - \mu_e \min\{q_e^{(1)}, \bar{q}\} \qquad (2.18) \\
&\geq \lambda p_e\Big(u_e\Big(q_e^{(2)}\Big) + c, u_r\Big(q_r^{(2)}\Big), \bar{u}\Big) - \mu_e \min\{q_e^{(2)}, \bar{q}\} \qquad (2.19) \\
&= 0,
\end{aligned}$$

where (2.17) is obtained from $q_e^{(1)} < q_e^{(2)}$ and the fact that $u' > 0$ (assumption U3) while p_e is deceasing in u_e; equation (2.18) follows from the assumption that $q_r^{(1)} \geq q_r^{(2)}$ and the fact that $u' > 0$ while p_e is increasing in u_r; and (2.19) follows from $q_e^{(1)} < q_e^{(2)}$. However, since the first and last line both equal zero due to the equilibrium conditions, (2.17)-(2.19) must all hold with strict equality. Consequently, (2.18)-(2.19) imply that

$$\min\{q_e^{(1)}, \bar{q}\} = \min\{q_e^{(2)}, \bar{q}\}, \qquad (2.20)$$

whence we conclude $\bar{q} \leq q_e^{(1)} < q_e^{(2)}$. From that, however, (2.17) yields

$$p_e\left(u\left(\frac{q_e^{(1)}}{\mu_e \bar{q}}\right) + c, u_r\Big(q_r^{(1)}\Big), \bar{u}\right) = p_e\left(u\left(\frac{q_e^{(2)}}{\mu_e \bar{q}}\right) + c, u_r\Big(q_r^{(1)}\Big), \bar{u}\right),$$

and this is impossible since $u' > 0$ with strict inequality. Therefore, $q_e^{(1)} < q_e^{(2)}$ implies $q_r^{(1)} < q_r^{(2)}$.

Next, we claim that $q_r^{(2)} > \bar{q}$. To see this, let us assume the opposite, i.e. $q_r^{(2)} \leq \bar{q}$, whence $u_r\left(q_r^{(1)}\right) = u_r\left(q_r^{(2)}\right) = u\left(\frac{1}{\mu_r}\right)$. We then have

$$\begin{aligned} \mu_e \min\{q_e^{(2)}, \bar{q}\} &= \lambda p_e\left(u_e\left(q_e^{(2)}\right) + c, u\left(\frac{1}{\mu_r}\right), \bar{u}\right) \\ &\leq \lambda p_e\left(u_e\left(q_e^{(1)}\right) + c, u\left(\frac{1}{\mu_r}\right), \bar{u}\right) \\ &= \mu_e \min\{q_e^{(1)}, \bar{q}\}, \end{aligned} \tag{2.21}$$

and $q_e^{(1)} < q_e^{(2)}$ implies that (2.21) holds with strict equality. This again implies (2.20) and the same reasoning as before can be repeated to obtain a contradiction. Therefore, $q_r^{(2)} > \bar{q}$. A symmetric argument can be used to show $q_e^{(2)} > \bar{q}$.

Combining the previous facts, (2.13) yields

$$p_o\left(u_e\left(q_e^{(1)}\right), u_r\left(q_r^{(1)}\right), \bar{u}\right) = p_o\left(u\left(\frac{q_e^{(2)}}{\mu_e \bar{q}}\right), u\left(\frac{q_r^{(2)}}{\mu_r \bar{q}}\right), \bar{u}\right).$$

However, from $q_e^{(1)} < q_e^{(2)}$ and $q_r^{(1)} < q_r^{(2)}$ we obtain

$$p_o\left(u_e\left(q_e^{(1)}\right), u_r\left(q_r^{(1)}\right), \bar{u}\right) < p_o\left(u\left(\frac{q_e^{(2)}}{\mu_e \bar{q}}\right), u\left(\frac{q_r^{(2)}}{\mu_r \bar{q}}\right), \bar{u}\right),$$

regardless of whether $q_e^{(1)}$ and $q_r^{(1)}$ are under or over capacity, because p_o satisfies assumption P2. We conclude that it is impossible to have $q_e^{(1)} < q_e^{(2)}$ and still satisfy the equilibrium conditions for both solutions. □

Theorem 3. *The equilibrium of the system (2.11)-(2.13) is locally stable.*

Proof. We examine each of regimes R1-R4 separately. In each regime, we write

(2.6)-(2.7) as

$$(q_e', q_r') = (f_e(q_e, q_r), f_r(q_e, q_r)),$$

obtain all of the first-order partial derivatives $\frac{\partial f_i}{\partial q_i}$ for $i \in \{e, r\}$, put them into matrix form (the Jacobian) and evaluate this matrix at the equilibrium $(q_e^\star, q_r^\star)$, which we know exists and is unique from the preceding. The equilibrium is locally stable if both eigenvalues of the Jacobian are negative [58].

Regime R1. We have $q_e^\star, q_r^\star \geq \bar{q}$ and the Jacobian is given by

$$J^{R1} = \lambda \begin{bmatrix} \frac{\partial p_e}{\partial u_e}\frac{\partial u_e}{\partial q_e} & \frac{\partial p_e}{\partial u_r}\frac{\partial u_r}{\partial q_r} \\ \frac{\partial p_r}{\partial u_e}\frac{\partial u_e}{\partial q_e} & \frac{\partial p_r}{\partial u_r}\frac{\partial u_r}{\partial q_r} \end{bmatrix}.$$

Letting e_1, e_2 be the eigenvalues, the characteristic equation is given by

$$\left(\lambda\frac{\partial p_e}{\partial u_e}\frac{\partial u_e}{\partial q_e} - e_1\right) \cdot \left(\lambda\frac{\partial p_r}{\partial u_r}\frac{\partial u_r}{\partial q_r} - e_2\right) - \lambda^2\frac{\partial p_e}{\partial u_r}\frac{\partial u_r}{\partial q_r} \cdot \frac{\partial p_r}{\partial u_e}\frac{\partial u_e}{\partial q_e} = 0,$$

which can be rewritten as

$$\lambda^2\frac{\partial u_e}{\partial q_e}\frac{\partial u_r}{\partial q_r}\left(\frac{\partial p_e}{\partial u_e}\frac{\partial p_r}{\partial u_r} - \frac{\partial p_e}{\partial u_r}\frac{\partial p_r}{\partial u_e}\right) + e_1e_2 = \lambda e_1\frac{\partial p_r}{\partial u_r}\frac{\partial u_r}{\partial q_r} + \lambda e_2\frac{\partial p_e}{\partial u_e}\frac{\partial u_e}{\partial q_e}. \tag{2.22}$$

We argue that

$$\det(J^{R1}) = \lambda^2\frac{\partial u_e}{\partial q_e}\frac{\partial u_r}{\partial q_r}\left(\frac{\partial p_e}{\partial u_e}\frac{\partial p_r}{\partial u_r} - \frac{\partial p_e}{\partial u_r}\frac{\partial p_r}{\partial u_e}\right)$$

is positive, which would imply that the product e_1e_2 in (2.22) is also positive. We first observe that the product $\frac{\partial u_e}{\partial q_e}\frac{\partial u_r}{\partial q_r}$ is positive since, for example,

$$\frac{\partial u_e}{\partial q_e} = u'\left(\frac{q_e^\star}{\mu_e \bar{q}}\right)\frac{1}{\mu_e \bar{q}} > 0$$

by assumption U3. The same is true of $\frac{\partial u_r}{\partial q_r}$.

Assumption P3 implies

$$\frac{\partial p_e}{\partial u_e} + \frac{\partial p_e}{\partial u_r} + \frac{\partial p_e}{\partial \bar{u}} = 0$$

since changing all the disutilities by the same amount does not change the probability of any choice. Since $\frac{\partial p_e}{\partial \bar{u}} > 0$ by assumption P2, it follows that $\frac{\partial p_e}{\partial u_e} + \frac{\partial p_e}{\partial u_r} < 0$, whence

$$\frac{\partial p_e}{\partial u_e} < -\frac{\partial p_e}{\partial u_r} \tag{2.23}$$

and, symmetrically,

$$\frac{\partial p_r}{\partial u_r} < -\frac{\partial p_r}{\partial u_e}. \tag{2.24}$$

From this we obtain

$$\frac{\partial p_e}{\partial u_e}\frac{\partial p_r}{\partial u_r} > -\frac{\partial p_e}{\partial u_r}\frac{\partial p_r}{\partial u_r} > \frac{\partial p_e}{\partial u_r}\frac{\partial p_r}{\partial u_e},$$

where the first inequality is obtained from (2.23) and the fact that $\frac{\partial p_e}{\partial u_e} < 0$, while the second inequality is obtained from (2.24) and the fact that $\frac{\partial p_e}{\partial u_r} > 0$. Thus, we conclude that $\det(J^{R1}) > 0$ and so both e_1, e_2 have the same sign.

From the preceding, it follows that the left-hand side of (2.22) is positive. On the right-hand side of (2.22), suppose that e_1, e_2 are both positive; then we have

$$\lambda e_1 \frac{\partial p_r}{\partial u_r} \frac{\partial u_r}{\partial q_r} < 0, \qquad \lambda e_e \frac{\partial p_e}{\partial u_e} \frac{\partial u_e}{\partial q_e} < 0$$

since $\frac{\partial p_r}{\partial u_r}, \frac{\partial p_e}{\partial u_e} < 0$ while $\frac{\partial u_r}{\partial q_r}, \frac{\partial u_e}{\partial q_e} > 0$. Therefore, both e_1, e_2 must be negative, as required.

Regime R2. We have $q_e^\star, q_r^\star < \bar{q}$ and the Jacobian is given by

$$J^{R2} = \lambda \begin{bmatrix} -\mu_e & 0 \\ 0 & -\mu_r \end{bmatrix},$$

from which the conclusion directly follows.

Regime R3. We have $q_r^\star < \bar{q} \leq q_e^\star$ and the Jacobian is given by

$$J^{R3} = \lambda \begin{bmatrix} \lambda \frac{\partial p_e}{\partial u_e} \frac{\partial u_e}{\partial q_e} & 0 \\ \lambda \frac{\partial p_r}{\partial u_e} \frac{\partial u_e}{\partial q_e} & -\mu_r \end{bmatrix},$$

which is a lower triangular matrix, meaning that its eigenvalues are on the diagonal. It is easy to see that both are negative.

Regime R4. The proof is very similar to the previous case and is omitted.

□

The next result illustrates the distinctions between regimes. Suppose that customer disutility becomes "steeper," i.e., customers are more dissatisfied with the same waiting time. If the cost remains unchanged, one might expect that the load on the express queue should increase, as express service is perceived as more beneficial. However, this is not guaranteed to happen in every regime.

Theorem 4. *Let u and v be disutility functions satisfying assumptions U1-U3, and suppose that $v' > u'$, that is, v grows more steeply than u; suppose also that the disutility of the outside option similarly changes to $\bar{v} > \bar{u}$ satisfying $v^{-1}(\bar{v}) = u^{-1}(\bar{u})$. Let (q_e^u, q_r^u) and (q_e^v, q_r^v) be the equilibria under u and v. Then, if (q_e^u, q_r^u) belongs to regime R2 or R4, we have $q_e^v > q_e^u$.*

Proof. The assumptions on v imply that, for any $s_1 < s_2$, we have

$$v(s_2) - v(s_1) > u(s_2) - u(s_1). \tag{2.25}$$

This fact will be used to show the desired result in each of the relevant regimes.

Regime R2. Since both queues are under capacity, we have $u_e(q_e^u) = u\left(\frac{1}{\mu_e}\right)$ and $u_r(q_r^u) = u\left(\frac{1}{\mu_r}\right)$. From (2.25), we obtain

$$v\left(\frac{1}{\mu_e}\right) - u\left(\frac{1}{\mu_e}\right) < v\left(\frac{1}{\mu_r}\right) - u\left(\frac{1}{\mu_r}\right) < \bar{v} - \bar{u}. \tag{2.26}$$

We will now show that $q_e^u < q_e^v$ by contradiction. Suppose that $q_e^u \geq q_e^v$. It follows that the express queue continues to be under capacity when we switch to v. We then derive

$$\lambda p_e(v_e(q_e^v) + c, v_r(q_r^u), \bar{v}) = \lambda p_e(v_e(q_e^u) + c, v_r(q_r^u), \bar{v}) \tag{2.27}$$

$$= \lambda p_e\left(v\left(\frac{1}{\mu_e}\right) + c, v\left(\frac{1}{\mu_r}\right), \bar{v}\right) \tag{2.28}$$

where (2.27)-(2.28) follow because both $q_e^u, q_e^v < \bar{q}$. Next, we let $\delta = v\left(\frac{1}{\mu_r}\right) - u\left(\frac{1}{\mu_r}\right)$, noting that $\delta > 0$, and observe that

$$\lambda p_e\left(v\left(\frac{1}{\mu_e}\right) + c, v\left(\frac{1}{\mu_r}\right), \bar{v}\right) = \lambda p_e\left(v\left(\frac{1}{\mu_e}\right) + c - \delta, v\left(\frac{1}{\mu_r}\right) - \delta, \bar{v} - \delta\right) \tag{2.29}$$

$$= \lambda p_e\left(v\left(\frac{1}{\mu_e}\right) + c - \delta, u\left(\frac{1}{\mu_r}\right), \bar{v} - \delta\right)$$

$$> \lambda p_e\left(u\left(\frac{1}{\mu_e}\right) + c, u\left(\frac{1}{\mu_r}\right), \bar{u}\right) \tag{2.30}$$

$$= \mu_e q_e^u \tag{2.31}$$

$$\geq \mu_e q_e^v. \tag{2.32}$$

Above, (2.29) is due to assumption P3, (2.30) follows from assumption P2 combined with (2.26), and (2.31) follows by (2.11).

To obtain the desired contradiction, we consider two cases, one where $q_r^v < \bar{q}$ and one where $q_r^v \geq \bar{q}$. Suppose that $q_r^v < \bar{q}$. Then, both queues are under capacity with v as the disutility function, so (2.11) implies

$$\lambda p_e\left(v\left(\frac{1}{\mu_e}\right) + c, v\left(\frac{1}{\mu_r}\right), \bar{v}\right) = \mu_e q_e^v.$$

On the other hand, if $q_r^v \geq \bar{q}$, we have $q_r^v > q_r^u$ and

$$\begin{aligned}\lambda p_e(v_e(q_e^v) + c, v_r(q_r^u), \bar{v}) &< \lambda p_e(v_e(q_e^v) + c, v_r(q_r^v), \bar{v}) \\ &= \mu_e q_e^v,\end{aligned}$$

by assumption P2. Either case, when combined with (2.32), yields $q_e^v < q_e^v$, which is impossible; therefore, we must have $q_e^u < q_e^v$.

Regime R4. In this regime, only the express queue is under capacity. There are two possible permutations

$$u\Big(\frac{1}{\mu_e}\Big) < \bar{u} \leq u\Big(\frac{q_r^u}{\mu_r \bar{q}}\Big), \qquad u\Big(\frac{1}{\mu_e}\Big) < u\Big(\frac{q_r^u}{\mu_r \bar{q}}\Big) < \bar{u}.$$

Applying (2.25) to both of these yields

$$v\Big(\frac{1}{\mu_e}\Big) - u\Big(\frac{1}{\mu_e}\Big) < \bar{v} - \bar{u} \leq v\Big(\frac{q_r^u}{\mu_r \bar{q}}\Big) - u\Big(\frac{q_r^u}{\mu_r \bar{q}}\Big), \tag{2.33}$$

$$v\Big(\frac{1}{\mu_e}\Big) - u\Big(\frac{1}{\mu_e}\Big) < v\Big(\frac{q_r^u}{\mu_r \bar{q}}\Big) - u\Big(\frac{q_r^u}{\mu_r \bar{q}}\Big) < \bar{v} - \bar{u}. \tag{2.34}$$

First, let us suppose that permutation (2.33) is correct. Again, we proceed by contradiction and assume that $q_e^u \geq q_e^v$. Since the express queue is under capacity with either disutility function, we have

$$\lambda p_e(v_e(q_e^v) + c, v_r(q_r^u), \bar{v}) = \lambda p_e(v_e(q_e^u) + c, v_r(q_r^u), \bar{v})$$
$$= \lambda p_e\left(v\left(\frac{1}{\mu_e}\right) + c, v\left(\frac{q_r^u}{\mu_r \bar{q}}\right), \bar{v}\right).$$

Letting $\delta = \bar{v} - \bar{u}$, we further derive

$$\lambda p_e\left(v\left(\frac{1}{\mu_e}\right) + c, v\left(\frac{q_r^u}{\mu_r \bar{q}}\right), \bar{v}\right) = \lambda p_e\left(v\left(\frac{1}{\mu_e}\right) + c - \delta, v\left(\frac{q_r^u}{\mu_r \bar{q}}\right) - \delta, \bar{v} - \delta\right) \tag{2.35}$$

$$= \lambda p_e\left(v\left(\frac{1}{\mu_e}\right) + c - \delta, v\left(\frac{q_r^u}{\mu_r \bar{q}}\right) - \delta, \bar{u}\right)$$
$$> \lambda p_e\left(u\left(\frac{1}{\mu_e}\right) + c, u\left(\frac{q_r^u}{\mu_r \bar{q}}\right), \bar{u}\right) \tag{2.36}$$
$$= \mu_e q_e^u$$
$$\geq \mu_e q_e^v, \tag{2.37}$$

where, as before, (2.35) is due to assumption P3, while (2.36) follows from (2.33) combined with assumption P2. From (2.37) and assumption P2, we conclude that $q_r^u > q_r^v$, otherwise there will be no way to satisfy (2.11).

Now, (2.13) yields

$$
\begin{aligned}
0 &= \lambda p_o\left(u\left(\frac{1}{\mu_e}\right)+c, u\left(\frac{q_r^u}{\mu_r\bar{q}}\right), \bar{u}\right) - \lambda + \mu_e q_e^u + \mu_r \min\{q_r^u, \bar{q}\} \\
&\geq \lambda p_o\left(u\left(\frac{1}{\mu_e}\right)+c, u\left(\frac{q_r^u}{\mu_r\bar{q}}\right), \bar{u}\right) - \lambda + \mu_e q_e^v + \mu_r \min\{q_r^v, \bar{q}\} \\
&= \lambda p_o\left(u\left(\frac{1}{\mu_e}\right)+c, u\left(\frac{q_r^u}{\mu_r\bar{q}}\right), \bar{u}\right) - \lambda p_o\left(v\left(\frac{1}{\mu_e}\right)+c, v_r(q_r^v), \bar{v}\right), \qquad (2.38)
\end{aligned}
$$

whence

$$
p_o\left(u\left(\frac{1}{\mu_e}\right)+c, u\left(\frac{q_r^u}{\mu_r\bar{q}}\right), \bar{u}\right) \leq p_o\left(v\left(\frac{1}{\mu_e}\right)+c, v_r(q_r^v), \bar{v}\right). \qquad (2.39)
$$

At the same time, letting $\delta = v\left(\frac{q_r^u}{\mu_r\bar{q}}\right) - u\left(\frac{q_r^u}{\mu_r\bar{q}}\right)$, we obtain

$$
\begin{aligned}
p_o\left(u\left(\frac{1}{\mu_e}\right)+c, u\left(\frac{q_r^u}{\mu_r\bar{q}}\right), \bar{u}\right) &= p_o\left(u\left(\frac{1}{\mu_e}\right)+c+\delta, u\left(\frac{q_r^u}{\mu_r\bar{q}}\right)+\delta, \bar{u}+\delta\right) \\
&= p_o\left(u\left(\frac{1}{\mu_e}\right)+c+\delta, v\left(\frac{q_r^u}{\mu_r\bar{q}}\right), \bar{u}+\delta\right) \\
&> p_o\left(v\left(\frac{1}{\mu_e}\right)+c, v\left(\frac{q_r^u}{\mu_r\bar{q}}\right), \bar{v}\right) \qquad (2.40) \\
&> p_o\left(v\left(\frac{1}{\mu_e}\right)+c, v_r(q_r^v), \bar{v}\right), \qquad (2.41)
\end{aligned}
$$

where the first equality is due to assumption P3, while (2.40) follows from (2.33) and assumption P2, while (2.41) follows from assumption P2 and $q_r^u > q_r^v$. Clearly (2.39) and (2.41) contradict each other, whence we conclude that $q_e^u < q_e^v$.

Finally, we suppose that permutation (2.34) is correct. In this case, however, the proof is nearly identical. The only difference is that, in order to obtain (2.37), we use $\delta = v\left(\frac{q_r^u}{\mu_r \bar{q}}\right) - u\left(\frac{q_r^u}{\mu_r \bar{q}}\right)$, while (2.40) is obtained by using $\delta = \bar{v} - \bar{u}$. The same contradiction then follows.

$\square$

For regimes R1 and R3, it is possible to design numerical examples where the result of Theorem 4 does not hold. In other words, if the express queue is over capacity to begin with, increased customer impatience may lead to increased loads on the outside option or the regular queue. Here we see an example of how service capacity affects the behaviour of the system.

2.3 Pricing observable express service

We now suppose that $c \in \mathbb{R}_+$ is a decision variable, with all of the other problem inputs (such as λ, μ_e, μ_r, the disutility function u etc.) remaining fixed. Let $(q_e(c), q_r(c))$ denote the solution to (2.11)-(2.12) for fixed, but arbitrary c. The dependence of the equilibrium solution on c will determine the shape of any relevant revenue function that we might define; for this reason, we start by examining this dependence in Section 2.3.1. Then, in Section 2.3.2, we propose and study two objective functions related to the revenue of the service provider and the social welfare of the customers.

2.3.1 Dependence of the equilibrium on the entry cost

First, we present a key result on the monotonicity of the equilibrium solution with respect to c. Because this result is important for what follows, the proof is placed in the text.

Theorem 5. *Consider a fixed cost c_0 and let $(q_e(c_0), q_r(c_0))$ be the corresponding equilibrium solution. Then:*

1. If $q_e(c_0) \geq \bar{q}$, then $\frac{\partial q_e}{\partial c}\Big|_{c=c_0} < 0$ and $\frac{\partial q_r}{\partial c}\Big|_{c=c_0} = 0$.

2. If $q_e(c_0) < \bar{q}$, then $\frac{\partial q_e}{\partial c}\Big|_{c=c_0} < 0$ and $\frac{\partial q_r}{\partial c}\Big|_{c=c_0} > 0$.

Proof. We consider each of the four possible regimes separately. In each regime, we differentiate both sides of (2.11)-(2.12) with respect to c and manipulate the resulting expressions. A slight abuse of notation should be clarified: when we write, e.g., $\frac{\partial p_e}{\partial u_e}$, we are referring to the generic first argument u_e of the function $p_e(u_e, u_r, \bar{u})$, not to the actual disutility $u_e(q_e(c))$ of the equilibrium.

Regime R1. Differentiating both sides of (2.11)-(2.12), we obtain

$$\lambda \frac{\partial p_e}{\partial u_e}\left(\frac{\partial u_e}{\partial c} + 1\right) + \lambda \frac{\partial p_e}{\partial u_r}\left(\frac{\partial u_r}{\partial c}\right) = 0,$$

$$\lambda \frac{\partial p_r}{\partial u_e}\left(\frac{\partial u_e}{\partial c} + 1\right) + \lambda \frac{\partial p_r}{\partial u_r}\left(\frac{\partial u_r}{\partial c}\right) = 0,$$

which can be expanded as

$$\frac{\partial p_e}{\partial u_e}\left(u'\left(\frac{q_e}{\mu_e\overline{q}}\right)\cdot\frac{1}{\mu_e\overline{q}}\frac{\partial q_e}{\partial c}+1\right)+\frac{\partial p_e}{\partial u_r}\left(u'\left(\frac{q_r}{\mu_r\overline{q}}\right)\cdot\frac{1}{\mu_r\overline{q}}\frac{\partial q_r}{\partial c}\right)=0, \tag{2.42}$$

$$\frac{\partial p_r}{\partial u_e}\left(u'\left(\frac{q_e}{\mu_e\overline{q}}\right)\cdot\frac{1}{\mu_e\overline{q}}\frac{\partial q_e}{\partial c}+1\right)+\frac{\partial p_r}{\partial u_r}\left(u'\left(\frac{q_r}{\mu_r\overline{q}}\right)\cdot\frac{1}{\mu_r\overline{q}}\frac{\partial q_r}{\partial c}\right)=0. \tag{2.43}$$

Equations (2.42)-(2.43) can be written in matrix form as

$$\begin{bmatrix}\frac{\partial p_e}{\partial u_e} & \frac{\partial p_e}{\partial u_r}\\ \frac{\partial p_r}{\partial u_e} & \frac{\partial p_r}{\partial u_r}\end{bmatrix}\cdot\begin{bmatrix}u'(\frac{q_e}{\mu_e\overline{q}})\cdot\frac{1}{\mu_e\overline{q}}\frac{\partial q_e}{\partial c}\\ u'(\frac{q_r}{\mu_r\overline{q}})\cdot\frac{1}{\mu_r\overline{q}}\frac{\partial q_r}{\partial c}\end{bmatrix}=-\begin{bmatrix}\frac{\partial p_e}{\partial u_e}\\ \frac{\partial p_r}{\partial u_e}\end{bmatrix}.$$

The matrix $A=\begin{bmatrix}\frac{\partial p_e}{\partial u_e} & \frac{\partial p_e}{\partial u_r}\\ \frac{\partial p_r}{\partial u_e} & \frac{\partial p_r}{\partial u_r}\end{bmatrix}$ is invertible, as in the proof of Theorem 2 it is shown that $\det(A)>0$. Consequently,

$$\begin{bmatrix}u'(\frac{q_e}{\mu_e\overline{q}})\cdot\frac{1}{\mu_e\overline{q}}\frac{\partial q_e}{\partial c}\\ u'(\frac{q_r}{\mu_r\overline{q}})\cdot\frac{1}{\mu_r\overline{q}}\frac{\partial q_r}{\partial c}\end{bmatrix}=-A^{-1}\cdot\begin{bmatrix}\frac{\partial p_e}{\partial u_e}\\ \frac{\partial p_r}{\partial u_e}\end{bmatrix}.$$

We then calculate

$$A^{-1}\cdot\begin{bmatrix}\frac{\partial p_e}{\partial u_e}\\ \frac{\partial p_r}{\partial u_e}\end{bmatrix}=\frac{1}{\frac{\partial p_e}{\partial u_e}\frac{\partial p_r}{\partial u_r}-\frac{\partial p_e}{\partial u_r}\frac{\partial p_r}{\partial u_e}}\begin{bmatrix}\frac{\partial p_r}{\partial u_r} & -\frac{\partial p_e}{\partial u_r}\\ -\frac{\partial p_r}{\partial u_e} & \frac{\partial p_e}{\partial u_e}\end{bmatrix}\cdot\begin{bmatrix}\frac{\partial p_e}{\partial u_e}\\ \frac{\partial p_r}{\partial u_e}\end{bmatrix}=\begin{bmatrix}1\\ 0\end{bmatrix},$$

whence

$$\begin{bmatrix} \frac{\partial q_e}{\partial c} \\ \\ \frac{\partial q_r}{\partial c} \end{bmatrix} = - \begin{bmatrix} u'(\frac{q_e}{\mu_e \overline{q}}) \cdot \frac{1}{\mu_e \overline{q}} & 0 \\ \\ 0 & u'(\frac{q_r}{\mu_r \overline{q}}) \cdot \frac{1}{\mu_r \overline{q}} \end{bmatrix}^{-1} \cdot \begin{bmatrix} 1 \\ \\ 0 \end{bmatrix} = - \begin{bmatrix} \left(u'(\frac{q_e}{u_e \overline{q}}) \cdot \frac{1}{\mu_e \overline{q}} \right)^{-1} \\ \\ 0 \end{bmatrix},$$

which proves the claim.

Regime R2. Differentiating both sides of (2.11)-(2.12), we obtain

$$\frac{\partial q_e}{\partial c} = \frac{\lambda}{\mu_e} \cdot \frac{\partial p_e}{\partial u_e},$$

$$\frac{\partial q_r}{\partial c} = \frac{\lambda}{\mu_r} \cdot \frac{\partial p_r}{\partial u_e},$$

and the claim follows straightforwardly from assumption P2.

Regime R3. Differentiating both sides of (2.11)-(2.12), we obtain

$$\lambda \frac{\partial p_e}{\partial u_e} \cdot \left(\frac{\partial u_e}{\partial c} + 1 \right) = 0, \tag{2.44}$$

$$\lambda \frac{\partial p_r}{\partial u_e} \cdot \left(\frac{\partial u_e}{\partial c} + 1 \right) = \mu_r \frac{\partial q_r}{\partial c}, \tag{2.45}$$

and (2.44) becomes

$$\frac{\partial p_e}{\partial u_e} \cdot \left(u' \left(\frac{q_e}{\mu_e \overline{q}} \right) \cdot \frac{1}{\mu_e \overline{q}} \frac{\partial q_e}{\partial c} + 1 \right) = 0 \tag{2.46}$$

analogously to (2.42). From (2.46), it follows that

$$\frac{\partial q_e}{\partial c} = -\left(u'\left(\frac{q_e}{\mu_e\overline{q}}\right)\cdot\frac{1}{\mu_e\overline{q}}\right)^{-1},$$

which is strictly negative, as claimed. From (2.45), we obtain

$$\frac{\partial q_r}{\partial c} = \frac{\lambda}{\mu_r}\frac{\partial p_r}{\partial u_e}\cdot\left(\frac{\partial u_e}{\partial c}+1\right) = 0,$$

where the last equality is due to the fact that $\frac{\partial u_e}{\partial c}+1=0$, which follows from (2.44).

Regime R4. Differentiating both sides of (2.11)-(2.12), we obtain

$$\lambda\frac{\partial p_e}{\partial u_e} + \lambda\frac{\partial p_e}{\partial u_r}\cdot\frac{\partial u_r}{\partial c} = \mu_e\frac{\partial q_e}{\partial c}, \tag{2.47}$$

$$\lambda\frac{\partial p_r}{\partial u_e} + \lambda\frac{\partial p_r}{\partial u_r}\cdot\frac{\partial u_r}{\partial c} = 0. \tag{2.48}$$

From (2.48), we find

$$\frac{\partial u_r}{\partial c} = -\frac{\frac{\partial p_r}{\partial u_e}}{\frac{\partial p_r}{\partial u_r}}, \tag{2.49}$$

which is strictly positive due to assumption P2. Because of this, we also have

$$\frac{\partial q_r}{\partial c} = \left(u'\left(\frac{q_r}{\mu_r\overline{q}}\right)\cdot\frac{1}{\mu_r\overline{q}}\right)^{-1}\cdot\frac{\partial u_r}{\partial c} > 0.$$

From (2.47), we find

$$\begin{aligned}\frac{\partial q_e}{\partial c} &= \frac{\lambda}{\mu_e}\cdot\frac{\partial p_e}{\partial u_e}+\frac{\lambda}{\mu_e}\cdot\frac{\partial p_e}{\partial u_r}\cdot\frac{\partial u_r}{\partial c} \\ &= \frac{\lambda}{\mu_e}\cdot\frac{\frac{\partial p_r}{\partial u_r}\frac{\partial p_e}{\partial u_e}-\frac{\partial p_e}{\partial u_r}\frac{\partial p_r}{\partial u_e}}{\frac{\partial p_r}{\partial u_r}},\end{aligned} \tag{2.50}$$

where (2.50) is obtained via (2.49). In the proof of Theorem 2, it was shown that $\frac{\partial p_r}{\partial u_r}\frac{\partial p_e}{\partial u_e}-\frac{\partial p_e}{\partial u_r}\frac{\partial p_r}{\partial u_e}>0$, which completes the proof of the claim.

□

Earlier, we observed that, for any fixed value of c, the equilibrium $(q_e(c), q_r(c))$ can belong to one of four possible regimes, based on whether the regular and express queues are under or over capacity. If we then vary c, it is possible for the equilibrium to transition from one regime to another. Theorem 5 provides us with a way to categorize all possible transitions, which are summarized in Figure 2.1. The nodes represent possible regimes of the equilibrium, labeled R1- R4 as defined in Remark 1. In any given instance of this problem (that is, for a given disutility function u, given parameters λ, μ_e, μ_r, etc.), as c increases from zero to infinity, the equilibrium must make transitions between regimes according to one, and only one, of the six cases labeled C1-C6 in Figure 2.1, with the first node in each case representing the regime at $c = 0$.

To understand how this categorization is made, let us first consider an extreme situation where $c \to \infty$ (we call this the "terminal condition"). In this situation, the express queue is never preferable to any other option regardless of

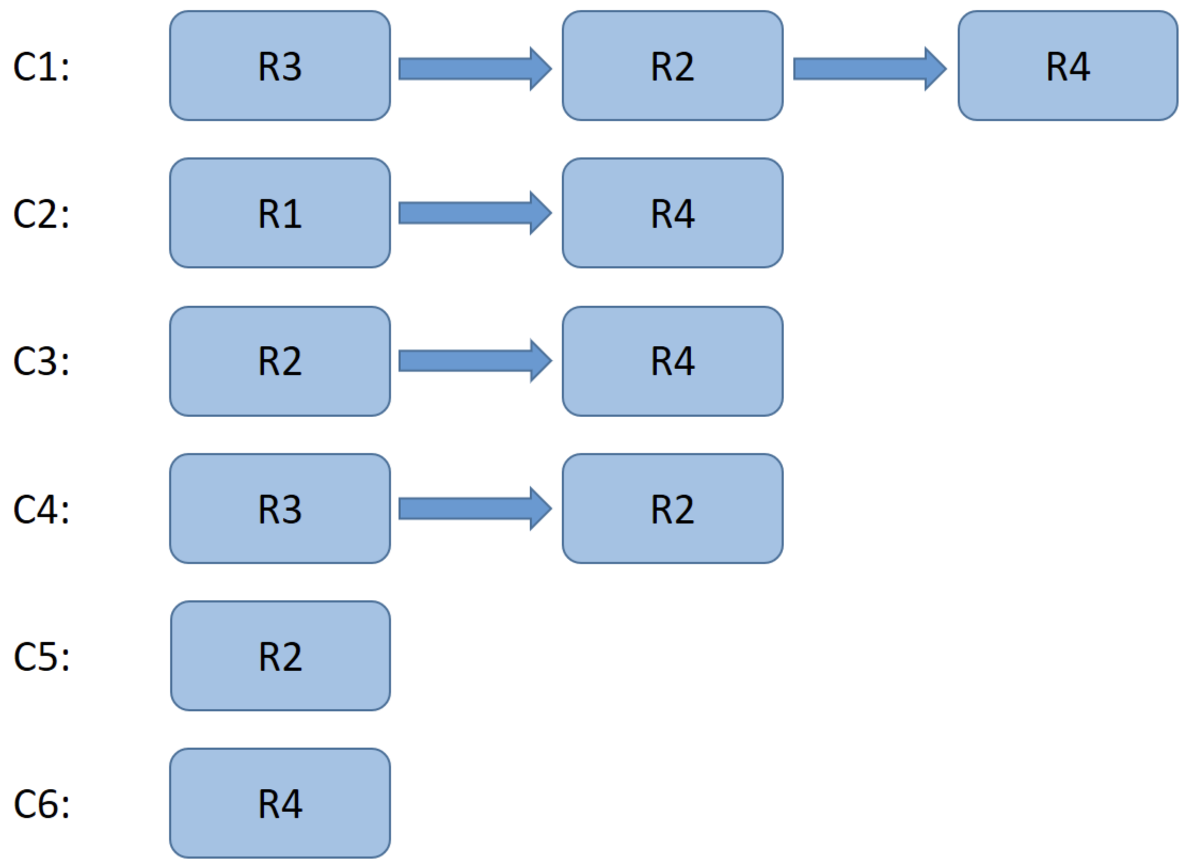


Figure 2.1: Flowchart describing possible transitions of the equilibrium as c increases.

how many customers are in the system. Therefore, the express queue cannot be over capacity in equilibrium; one can think of q_r in this case as measuring the congestion that would occur in the regular queue if the express queue did not exist at all. Under the terminal condition, only two regimes are possible, namely R2 and R4: either both queues are under capacity (e.g., if λ is small), or the regular queue is over capacity. The following result gives a precise way to determine the terminal regime using only the problem inputs and the distribution of the random shocks.

Proposition 1. *The terminal regime is R2 if and only if*

$$P\left(u\left(\frac{1}{\mu_r}\right) + \tau_r < \bar{u} + \tau_o\right) < \frac{\mu_r}{\lambda}\bar{q}.$$

Proof. From (2.12), we have

$$\begin{aligned} \frac{\mu_r}{\lambda}\min\{q_r(\infty), \bar{q}\} &= \lim_{c\to\infty} p_r(u_e(q_e(c)) + c, u_r(q_r(c)), \bar{u}) \\ &= P(u_r(q_r(\infty)) + \tau_r < \bar{u} + \tau_o) \qquad (2.51) \\ &\leq P\Big(u\Big(\frac{1}{\mu_r}\Big) + \tau_r < \bar{u} + \tau_o\Big), \end{aligned}$$

where (2.51) follows because the disutility of the express queue becomes infinitely large as $c \to \infty$. Now, if R4 is the terminal regime, we obtain $\frac{\mu_r}{\lambda}\bar{q} \leq P\Big(u\Big(\frac{1}{\mu_r}\Big) + \tau_r < \bar{u} + \tau_o\Big)$, as required. On the other hand, if R2 is the terminal regime, (2.51) becomes

$$\frac{\mu_r}{\lambda}\bar{q} > P\Big(u\Big(\frac{1}{\mu_r}\Big) + \tau_r < \bar{u} + \tau_o\Big),$$

completing the proof.

□

Now, if we can identify the regime of the equilibrium for $c = 0$ (the "initial condition"), Theorem 5 will then fill in the transitions in between. For example, if the initial regime is R1 (both queues over capacity, case C2 of Figure 2.1), we know that, as c increases, there will be only one transition to regime R4 (express queue under capacity), because $q_e(c)$ is decreasing in c while $q_r(c)$ is non-decreasing. If the initial regime is R2 (both queues under capacity, cases C3 and C5), at most one transition can occur to regime R4 (regular queue over capacity), as more of the load is shifted from the express queue to the regular queue. In fact, there is only one case (C1) where more than one transition is possible, arising only when R3 is the initial regime.

We may think of case C5 as representing situations where there is no real congestion in the system to begin with (the regular queue is under capacity even under the terminal condition). Cases C1, C3 and C4 represent situations where the congestion in the regular queue can be relieved by the presence of an express queue, if the entry fee is suitably chosen. Case C2 represents a situation where the congestion is so heavy that the express queue will never be able to relieve it entirely (though the service provider will still generate revenue from it). Finally, case C6 represents a surprising situation where, even if it is free to join the express queue, we will continue to see congestion in the regular queue even though some unused capacity remains in the express queue. As will be illustrated later, this can occur in instances where λ is moderately large (if λ is too large, we will be in case C2 instead) and μ_e is significantly larger than μ_r. The high service rate in the express queue reduces the queue length; although a large proportion of customers may be choosing this queue, they are processed so quickly that the queue does not become congested. However, the presence of random shocks will still direct some small proportion of customers to the regular queue, which can lead to congestion when combined with a much slower service rate.

2.3.2 Optimization of expected revenue and social welfare

We now propose two objective functions. Recalling from (2.11) that $p_e = \frac{\mu_e}{\lambda} \min\{q_e(c), \bar{q}\}$, the function

$$R(c) = p_e \cdot c = \frac{\mu_e c}{\lambda} \min\{q_e(c), \bar{q}\} \tag{2.52}$$

represents the expected revenue per arrival to the system, a natural objective to maximize for the service provider. The function

$$\begin{aligned} D(c) &= p_e \cdot (u_e(q_e(c)) + c) + p_r \cdot u_r(q_r(c)) + p_o \cdot \overline{u} \\ &= \overline{u} + R(c) + \frac{\mu_e}{\lambda} \min\{q_e(c), \overline{q}\}(u_e(q_e(c)) - \overline{u}) + \frac{\mu_r}{\lambda} \min\{q_r(c), \overline{q}\}(u_r(q_r(c)) - \overline{u}) \end{aligned}$$

represents the expected total disutility incurred by each customer, consisting of the expected cost paid as well as the expected disutility of waiting. Since D is a measure of negative value, the function $c \mapsto R(c) - D(c)$ can be viewed as a measure of the overall social welfare. Optimizing the social welfare is equivalent to finding the value of c that minimizes

$$W(c) = \overline{u} + \frac{\mu_e}{\lambda} \min\{q_e(c), \overline{q}\}(u_e(q_e(c)) - \overline{u}) + \frac{\mu_r}{\lambda} \min\{q_r(c), \overline{q}\}(u_r(q_r(c)) - \overline{u}),$$

the expected disutility of waiting.

The analysis of the six cases in Section 2.3.1 helps us understand the shape of these functions. For example, it is obvious that, in regimes R1 and R3 where the express queue is over capacity, the revenue (2.52) grows linearly in the cost. On the other hand, in these same two regimes, the total disutility D is *unaffected* by cost, as shown below.

Proposition 2. *Consider a fixed cost c_0 and let $q_e(c_0), q_r(c_0)$ be the corresponding equilibrium solution. If $q_e(c_0) \geq \bar{q}$, then $\frac{\partial D}{\partial c}\big|_{c=c_0} = 0$.*

Proof. The relevant regimes to consider are R1 and R3. We first consider regime R1. Define

$$D^{R1}(c) = \bar{u} + \frac{\mu_e \bar{q}}{\lambda}\left(u\left(\frac{q_e(c)}{\mu_e \bar{q}}\right) + c - \bar{u}\right) + \frac{\mu_r \bar{q}}{\lambda}\left(u\left(\frac{q_r(c)}{\mu_r}\right) - \bar{u}\right). \tag{2.53}$$

Taking the derivative with respect to c, we find

$$\frac{\partial D^{R1}}{\partial c} = \frac{\mu_e \bar{q}}{\lambda}\left(u'\left(\frac{q_e}{\mu_e \bar{q}}\right)\frac{1}{\mu_e \bar{q}}\frac{\partial q_e}{\partial c} + 1\right),$$

noting that, due to Theorem 5, the last term on the right-hand side of (2.53) vanishes when differentiated with respect to c. When we are in regime R1, differentiating both sides of (2.11) with respect to c yields

$$\frac{\partial p_e}{\partial u_e}\left(u'\left(\frac{q_e}{\mu_e \bar{q}}\right)\frac{1}{\mu_e \bar{q}}\frac{\partial q_e}{\partial c}+1\right)=0,$$

and since $\frac{\partial p_e}{\partial u_e}<0$ by assumption, it follows that $\frac{\partial D^{R1}}{\partial c}=0$.

In regime R3, we consider the function

$$D^{R3}(c)=\bar{u}+\frac{\mu_e \bar{q}}{\lambda}\left(u\left(\frac{q_e(c)}{\mu_e \bar{q}}\right)+c-\bar{u}\right)+\frac{\mu_r}{\lambda}\left(u\left(\frac{1}{\mu_r}\right)-\bar{u}\right)q_r(c).$$

Taking the derivative with respect to c, we find that $\frac{\partial D^{R3}}{\partial c}=\frac{\partial D^{R1}}{\partial c}$ due to Theorem 5, and since (2.11) has the same form in both R1 and R3, the result follows from the previous analysis.

□

Thus, regimes R2 and R4 are crucial to the understanding of both the revenue and the social welfare. In fact, we can obtain a complete characterization of the social welfare optimization problem under general choice probabilities.

Theorem 6. *The social welfare R-D is maximized (and W is minimized) as follows:*

1. In cases C1 and C4, W is minimized by setting c equal to the threshold between regimes R3 and R2.

2. In case C2, W is minimized by setting c equal to the threshold between regimes

R1 and R4.

3. In cases C3, C5 and C6, W is minimized by setting $c = 0$.

Proof. We examine the six cases in reverse order, because results obtained for the simpler cases can be reused for the more complicated ones.

Case C6. In this case, the equilibrium is always in regime R4 and the social welfare function is identical to

$$W^{R4}(c) = \overline{u} + \frac{\mu_e}{\lambda}\left(u\left(\frac{1}{\mu_e}\right) - \overline{u}\right)q_e(c) + \frac{\mu_r}{\lambda}\overline{q}\left(u\left(\frac{q_r(c)}{\mu_r\overline{q}}\right) - \overline{u}\right).$$

Therefore,

$$\frac{\partial W^{R4}}{\partial c} = \frac{\mu_e}{\lambda}\left(u\left(\frac{1}{\mu_e}\right) - \overline{u}\right)\frac{\partial q_e}{\partial c} + \frac{1}{\lambda}u'\left(\frac{q_r}{\mu_r\overline{q}}\right)\frac{\partial q_r}{\partial c}. \tag{2.54}$$

Both terms on the right-hand side of (2.54) are positive. The first term is a product of two negative quantities since $\bar{u} \geq u\left(\frac{1}{\mu_e}\right)$ and $\frac{\partial q_e}{\partial c} \leq 0$ by Theorem 5. The second term is positive since $\frac{\partial q_r}{\partial c} \geq 0$ by Theorem 5, and the disutility function u is assumed to be increasing. Thus, $\frac{\partial W^{R4}}{\partial c} \geq 0$ at any c for which the equilibrium solution belongs to regime R2. Consequently, in case C5, $W4$ is minimized by setting $c = 0$.

Case C5. In this case, the equilibrium is always in regime R2. From (2.13), we know that

$$\frac{\partial p_o}{\partial c} + \frac{\mu_e}{\lambda}\frac{\partial q_e}{\partial c} + \frac{\mu_r}{\lambda}\frac{\partial q_r}{\partial c} = 0. \tag{2.55}$$

From the construction of the choice probabilities, we know that $\frac{\partial p_o}{\partial c} \geq 0$, because, for any $c_1 \le c_2$, the event that

$$\bar{u} + \tau_o \le \min\left\{u\left(\frac{1}{\mu_e}\right) + c_1 + \tau_e, u\left(\frac{1}{\mu_r}\right) + \tau_r\right\}$$

implies

$$\bar{u} + \tau_o \le \min\left\{u\left(\frac{1}{\mu_e}\right) + c_2 + \tau_e, u\left(\frac{1}{\mu_r}\right) + \tau_r\right\}.$$

Then (2.55) implies

$$\frac{\partial q_r}{\partial c} \le -\frac{\mu_e}{\mu_r}\frac{\partial q_e}{\partial c}. \tag{2.56}$$

In regime R2, the social welfare function is identical to

$$W^{R2}(c) = \overline{u} + \frac{\mu_e}{\lambda}\left(u\left(\frac{1}{\mu_e}\right) - \overline{u}\right)q_e(c) + \frac{\mu_r}{\lambda}\left(u\left(\frac{1}{\mu_r}\right) - \overline{u}\right)q_r(c).$$

Therefore,

$$\begin{aligned}
\frac{\partial W^{R2}}{\partial c} &= \frac{\mu_e}{\lambda}\left(u\left(\frac{1}{\mu_e}\right) - \overline{u}\right)\frac{\partial q_e}{\partial c} + \frac{\mu_r}{\lambda}\left(u\left(\frac{1}{\mu_r}\right) - \bar{u}\right)\frac{\partial q_r}{\partial c} \\
&\ge \frac{\mu_e}{\lambda}\left(u\left(\frac{1}{\mu_e}\right) - \overline{u}\right)\frac{\partial q_e}{\partial c} - \frac{\mu_e}{\lambda}\left(u\left(\frac{1}{\mu_r}\right) - \bar{u}\right)\frac{\partial q_e}{\partial c} \qquad (2.57) \\
&= \frac{\mu_e}{\lambda}\left(u\left(\frac{1}{\mu_e}\right) - u\left(\frac{1}{\mu_r}\right)\right)\frac{\partial q_e}{\partial c} \qquad (2.58) \\
&\ge 0,
\end{aligned}$$

where (2.57) is obtained by combining (2.56) with the fact that $\bar{u} \ge u(\frac{1}{\mu_r})$ and the last line follows because (2.58) is the product of two negative quantities, since $\frac{\partial q_e}{\partial c} \le 0$ by Theorem 5 and $\mu_e > \mu_r$ with u increasing. Thus, $\frac{\partial W^{R2}}{\partial c} \ge 0$ at any c for which the equilibrium solution belongs to regime R2. Consequently, in case C5, W is minimized by setting $c = 0$.

Case C4. Let c_0 be the threshold value such that $(q_e(c), q_r(c))$ belongs to regime R3 for $c \in [0, c_0]$, and to regime R2 for $c > c_0$. From the preceding analysis, $W(c_0) \leq W(c)$ for all $c > c_0$. However, from Proposition 2 we know that D is constant on the interval $[0, c_0]$, while R increases linearly on the same interval. Consequently,

$$\arg\min_{0 \leq c \leq c_0} W(c) = \arg\max_{0 \leq c \leq c_0} R(c) - D(c) = c_0.$$

Case C3. Let c_0 be the threshold value such that $(q_e(c), q_r(c))$ belongs to regime R2 for $c \in [0, c_0]$, and to regime R4 for $c > c_0$. Then, $W(c) = W^{R2}(c)$ for $c \in [0, c_0]$ and $W(c) = W^{R4}(c)$ for $c > c_0$. It follows that $\frac{\partial W}{\partial c} \geq 0$ at all $c \geq 0$, and thus is minimized by setting $c = 0$.

Cases C1-C2. The analysis follows straightforwardly from the above.

□

Theorem 6 shows that social welfare is maximized when the express queue is running exactly at full capacity (or as close to it as possible), but without going over. Customers do not always benefit from being allowed to access the express queue for free, because this would lead to congestion and reduced service quality. Rather, the price should be low enough to alleviate the congestion in the regular queue where possible, but high enough to avoid congestion in the express queue.

The shape of the revenue function R is more difficult to characterize. It is possible to show, in a fairly general setting, that R has a unique maximum in regime R2.

Proposition 3. *Suppose that each random shock τ_e, τ_r, τ_o has a log-concave density on $\mathbb{R}_+$. Then, the mapping*

$$c \mapsto c \cdot p_e\left(u\left(\frac{1}{\mu_e}\right) + c, u\left(\frac{1}{\mu_r}\right), \bar{u}\right) \tag{2.59}$$

is log-concave in c.

Proof. Let $\tau = (\tau_e, \tau_r, \tau_o)$ and $u = (u_e, u_r, \bar{u})$. From p. 107 of [59], we know that the mapping $u \mapsto P(\tau + u \in C)$ is log-concave for any given convex set C. The set

$$C = \{(c_e, c_r, c_o) : c_e \leq c_r, c_e \leq c_o\}$$

is convex, being described by linear inequalities. Consequently, the mapping

$$c \mapsto p_e\left(u\left(\frac{1}{\mu_e}\right) + c, u\left(\frac{1}{\mu_r}\right), \bar{u}\right)$$

is log-concave, being a composition of a log-concave function and a linear function. The log-concavity of (2.59) easily follows.

□

Unfortunately, there is no guarantee that R2 will always be associated with higher revenue. In other words, it is possible to design some instances where the revenue is maximized in regime R2, and others where it is maximized in R4. In the latter case, a revenue-maximizing service provider will prefer to artificially drive up congestion in the regular queue, while deliberately leaving unused capacity in the express queue, simply because it is more profitable to serve a small proportion of customers with high willingness to pay.

2.4 Specific choice models

The multinomial logit and exponomial choice models represent two standard and well-known sets of assumptions for the distributions of the random shocks. Both models can be used together with general disutility functions, which makes them the two most natural contexts in which to study our problem. In this section, we discuss both models, both to illustrate the generality of our framework, and to show that the presence of multiple peaks in the revenue function is not confined to one particular choice model.

Section 2.4.1 presents additional analysis and numerical illustrations for the setting where customer choice follows the MNL model. Section 2.4.2 considers the exponomial choice model.

2.4.1 Multinomial logit (MNL) choice model

Under the MNL model, we assume that all random shocks in the problem are i.i.d. Gumbel distributed. Using standard parameter choices, we obtain the following explicit forms for the choice probabilities:

$$p_e(q_e, q_r, \bar{u}) = \frac{e^{-u_e(q_e)-c}}{e^{-u_e(q_e)-c} + e^{-u_r(q_r)} + e^{-\bar{u}}},$$

$$p_r(q_e, q_r, \bar{u}) = \frac{e^{-u(q_r)}}{e^{-u_e(q_e)-c} + e^{-u_r(q_r)} + e^{-\bar{u}}},$$

$$p_o(q_e, q_r, \bar{u}) = \frac{e^{-\bar{u}}}{e^{-u_e(q_e)-c} + e^{-u_r(q_r)} + e^{-\bar{u}}},$$

We can verify that these probabilities satisfy the assumptions listed in Section 2.2.1. Thus, all the results of Section 2.2.2 and Section 2.3 apply.

We now write (2.11)-(2.13) as

$$\lambda \frac{e^{-u_e(q_e)-c}}{e^{-u_e(q_e)-c} + e^{-u_r(q_r)} + e^{-\bar{u}}} = \mu_e \min\{q_e, \bar{q}\}, \tag{2.60}$$

$$\lambda \frac{e^{-u_r(q_r)}}{e^{-u_e(q_e)-c} + e^{-u_r(q_r)} + e^{-\bar{u}}} = \mu_r \min\{q_r, \bar{q}\}, \tag{2.61}$$

$$\lambda \frac{e^{-\bar{u}}}{e^{-u_e(q_e)-c} + e^{-u_r(q_r)} + e^{-\bar{u}}} = \lambda - \mu_e \min\{q_e, \bar{q}\} - \mu_r \min\{q_r, \bar{q}\}. \tag{2.62}$$

These equations lead to closed-form expressions for the equilibrium solution in every possible regime. It then becomes possible to identify which of the four regimes holds in a specific problem instance.

Regime R1. $q_e, q_r \geq \bar{q}$. Equations (2.60)-(2.62) become

$$\lambda \frac{e^{-u(\frac{q_e}{\mu_e \bar{q}})-c}}{e^{-u(\frac{q_e}{\mu_e \bar{q}})-c} + e^{-u(\frac{q_r}{\mu_r \bar{q}})} + e^{-\bar{u}}} = \mu_e \bar{q}, \tag{2.63}$$

$$\lambda \frac{e^{-u(\frac{q_r}{\mu_r \overline{q}})}}{e^{-u(\frac{q_e}{\mu_e \overline{q}})-c} + e^{-u(\frac{q_r}{\mu_r \overline{q}})} + e^{-\overline{u}}} = \mu_r \overline{q}, \tag{2.64}$$

$$\lambda \frac{e^{-\overline{u}}}{e^{-u(\frac{q_e}{\mu_e \overline{q}})-c} + e^{-u(\frac{q_r}{\mu_r \overline{q}})} + e^{-\overline{u}}} = \lambda - \mu_e \overline{q} - \mu_r \overline{q}. \tag{2.65}$$

Dividing (2.63) and (2.64), respectively, by (2.65) produces

$$q_e^{R1} = \mu_e \overline{q} \cdot u^{-1} \Bigg(- \log \left(\frac{\mu_e \overline{q}}{\lambda - \mu_e \overline{q} - \mu_r \overline{q}} \right) + \overline{u} - c \Bigg),$$

$$q_r^{R1} = \mu_r \overline{q} \cdot u^{-1} \Bigg(- \log \left(\frac{\mu_r \overline{q}}{\lambda - \mu_e \overline{q} - \mu_r \overline{q}} \right) + \overline{u} \Bigg).$$

Regime R2. $q_e, q_r \leq \bar{q}$. Equations (2.60)-(2.62) directly lead to

$$q_e^{R2} = \frac{\lambda}{\mu_e} \cdot \frac{e^{-u(\frac{1}{\mu_e})-c}}{e^{-u(\frac{1}{\mu_e})-c} + e^{-u(\frac{1}{\mu_r})} + e^{-\overline{u}}},$$

$$q_r^{R2} = \frac{\lambda}{\mu_r} \cdot \frac{e^{-u(\frac{1}{\mu_r})}}{e^{-u(\frac{1}{\mu_e})-c} + e^{-u(\frac{1}{\mu_r})} + e^{-\overline{u}}}.$$

Regime R3. $q_e \geq \bar{q} > q_r$. Equations (2.60)-(2.62) become

$$\lambda \frac{e^{-u(\frac{q_e}{\mu_e \overline{q}})-c}}{e^{-u(\frac{q_e}{\mu_e \overline{q}})-c} + e^{-u(\frac{1}{\mu_r})} + e^{-\overline{u}}} = \mu_e \overline{q}, \tag{2.66}$$

$$\lambda \frac{e^{-u(\frac{1}{\mu_r})}}{e^{-u(\frac{q_e}{\mu_e \overline{q}})-c} + e^{-u(\frac{1}{\mu_r})} + e^{-\overline{u}}} = \mu_r q_r, \tag{2.67}$$

$$\lambda \frac{e^{-\overline{u}}}{e^{-u(\frac{q_e}{\mu_e \overline{q}})-c} + e^{-u(\frac{1}{\mu_r})} + e^{-\overline{u}}} = \lambda - \mu_e \overline{q} - \mu_r q_r. \tag{2.68}$$

Dividing (2.66) and (2.67), respectively, by (2.68) produces

$$-\,u\Big(\frac{q_e}{\mu_e\overline{q}}\Big) - c + \overline{u} = \log\left(\frac{\mu_e\overline{q}}{\lambda - \mu_e\overline{q} - \mu_r q_r}\right), \tag{2.69}$$

$$-\,u\Big(\frac{1}{\mu_r}\Big) + \overline{u} = \log\left(\frac{\mu_r q_r}{\lambda - \mu_e\overline{q} - \mu_r q_r}\right). \tag{2.70}$$

The system (2.69)-(2.70) is solved by

$$q_e^{R3} = \mu_e\overline{q}\cdot u^{-1}\left(-\log\left(\frac{\mu_e\overline{q}}{\lambda - \mu_e\overline{q}}\cdot\frac{e^{-u(\frac{1}{\mu_r})} + e^{-\overline{u}}}{e^{-\overline{u}}}\right) + \overline{u} - c\right),$$

$$q_r^{R3} = \frac{\lambda - \mu_e\overline{q}}{\mu_r}\cdot\frac{e^{-u(\frac{1}{\mu_r})}}{e^{-u(\frac{1}{\mu_r})} + e^{-\overline{u}}}.$$

Regime R4. $q_r \geq \bar{q} > q_e$. We proceed similarly to the derivation for regime R3 and obtain

$$q_e^{R4} = \frac{\lambda - \mu_r\overline{q}}{\mu_e}\cdot\frac{e^{-u(\frac{1}{\mu_e})-c}}{e^{-u(\frac{1}{\mu_e})-c} + e^{-\overline{u}}},$$

$$q_r^{R4} = \mu_r\overline{q}\cdot u^{-1}\left(-\log\left(\frac{\mu_r\overline{q}}{\lambda - \mu_r\overline{q}}\cdot\frac{e^{-u(\frac{1}{\mu_e})-c} + e^{-\overline{u}}}{e^{-\overline{u}}}\right) + \overline{u}\right).$$

Given a specific problem instance (some specific disutility function u, parameters λ, μ_e, μ_r, etc.), we can identify which of the four regimes holds by calculating these four solutions and checking which of them actually falls into its correct range. Thus, for example, if we calculate q_e^{R1}, q_r^{R1}, but find that at least one of these quantities is strictly less than $\bar{q}$ (contrary to the definition of regime R1), it necessarily follows that R1 is not the correct regime for the equilibrium of this problem instance. In fact, for any given instance, only one of the four solutions will be in the correct range, corresponding to the regime of the equilibrium.

By applying this analysis for $c = 0$ and $c \to \infty$, we can further identify the case, among C1-C6 in Figure 2.1, to which the given problem instance belongs. Note that each of the six cases is described by a unique combination of initial and terminal regime. Thus, identifying the initial and terminal regimes is enough to tell us how many transitions, and between which regimes, will occur as c increases from zero to infinity.

We can use this approach to obtain further insight into how the problem inputs determine which case among C1-C6 is realized. Let us first focus on the initial regime (fixing $c = 0$). The threshold between regimes R2 and R3 occurs when $q_e^{R2} = \bar{q}$. This is equivalent to the condition

$$\lambda = \mu_e \overline{q} + \frac{e^{-\overline{u}} + e^{-u(\frac{1}{\mu_r})}}{e^{-u(\frac{1}{\mu_e})}} \mu_e \overline{q}, \tag{2.71}$$

which (for a given disutility function u) defines a curve on the space of all possible (μ_r, μ_e, λ), on which small changes in these inputs will cause the initial regime to change from R2 to R3. In a similar fashion, the threshold between R3 and R1 is found by setting $q_r^{R3} = \bar{q}$, yielding the curve

$$\lambda = \mu_e \overline{q} + \frac{e^{-\overline{u}} + e^{-u(\frac{1}{\mu_r})}}{e^{-u(\frac{1}{\mu_r})}} \mu_r \overline{q}. \tag{2.72}$$

The threshold between R4 and R1 can be found by setting $q_e^{R4} = \bar{q}$, yielding

$$\lambda = \mu_r \overline{q} + \frac{e^{-\overline{u}} + e^{-u(\frac{1}{\mu_e})}}{e^{-u(\frac{1}{\mu_e})}} \mu_e \overline{q}, \tag{2.73}$$

and the threshold between R2 and R4 is found by setting $q_r^{R2} = \bar{q}$, yielding

$$\lambda = \mu_r \overline{q} + \frac{e^{-\overline{u}} + e^{-u(\frac{1}{\mu_e})}}{e^{-u(\frac{1}{\mu_r})}} \mu_r \overline{q}. \tag{2.74}$$

Under the terminal condition (now taking $c \to \infty$), we observed before that R2 and R4 are the only possible regimes. The threshold between them is found by setting $q_r^{R2} = \bar{q}$, yielding

$$\lambda = \frac{e^{-\overline{u}} + e^{-u(\frac{1}{\mu_r})}}{e^{-u(\frac{1}{\mu_r})}} \mu_r \overline{q}. \tag{2.75}$$

In Figure 2.2, we take a linear utility function, standardize $\mu_r = 1$, and plot all five curves (2.71)-(2.75) on the (μ_e, λ)-plane. We explain how this diagram can be used to match any (μ_e, λ) pair to one of the six cases. First, the terminal threshold (2.75) is represented in Figure 2.2 by the dashed horizontal black line. Any (μ_e, λ) pair below this threshold must have R2 as the terminal regime; from Figure 2.1, we know that this is only possible in cases C4 and C5. Thus, it follows that (as one might expect) case C5 occurs only when λ is sufficiently small. Conversely, any (μ_e, λ) pair above the terminal threshold must have R4 as the terminal regime.

The initial regime is found as follows. For any fixed $\mu_e \geq 1$, the initial regime will be R2 if λ is very low, and R1 if λ is very high. For "moderate" values, either R3 or R4 can occur, but these two are mutually exclusive under a fixed μ_e value. In

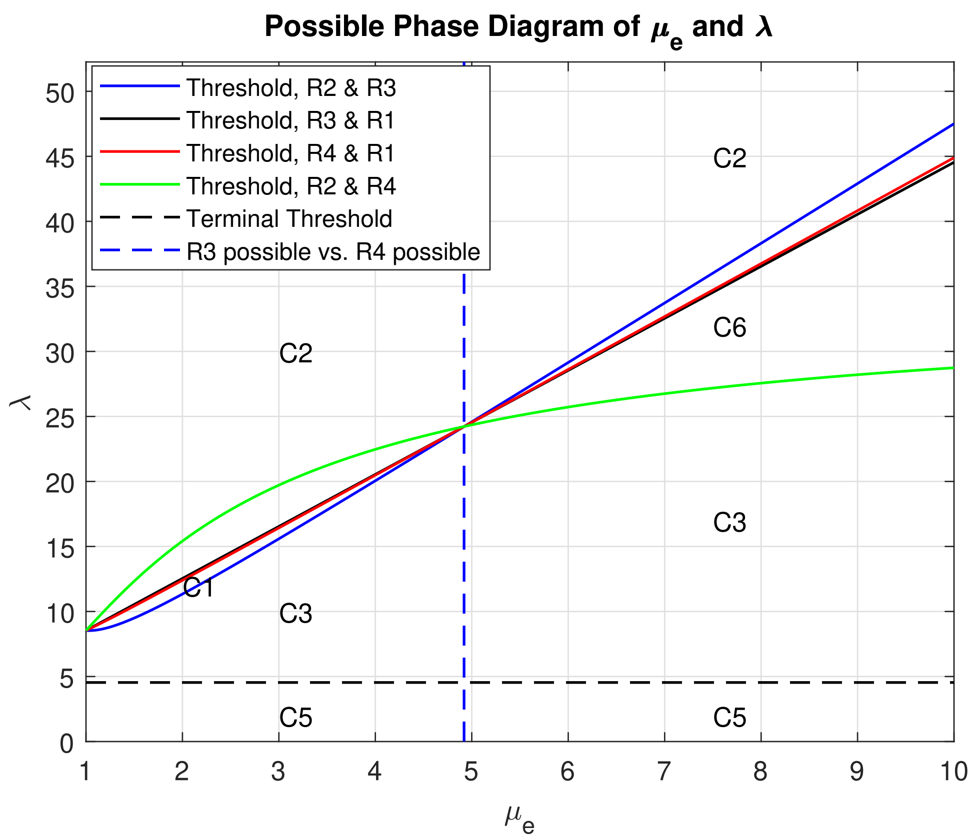


Figure 2.2: Phase diagram illustrating the impact of μ_e, λ on the equilibrium regimes.

other words, if μ_e is fixed to a "low" value, then R3 is the initial regime for moderate λ, so increasing λ from zero to infinity under this fixed μ_e will take us from R2 to R3 to R1, with the relevant thresholds being (2.71) and (2.72). However, if μ_e is "high," then R4 is the initial regime for moderate λ, so increasing λ under such a μ_e will take us from R2 to R4 to R1, with the relevant thresholds being (2.74) and (2.73). The precise threshold of μ_e separating "low" and "high" values is shown in Figure 2.2 by the dashed vertical blue line.

Thus, the complete reading of Figure 2.2 is as follows:

1. Any point below the dashed horizontal black line belongs to case C5.
2. Any point to the left of the dashed vertical blue line belongs to:
   (a) Case C3 if it is above the dashed horizontal black line, but below the blue curve;
   (b) Case C1 if it is above the blue curve, but below the black curve;
   (c) Case C2 if it is above the black curve.
3. Any point to the right of the dashed vertical blue line belongs to:
   (a) Case C3 if it is above the dashed horizontal black line, but below the green curve;
   (b) Case C6 if it is above the green curve, but below the red curve;
   (c) Case C2 if it is above the red curve.

Note that case C4 is not present in Figure 2.2. We found that this case is somewhat rare, occurring only if the blue curve dips below the dashed horizontal black line on the left side of the graph. It is possible to design instances where this happens, if $\bar{u}$ is very close to $u(1)$ and the disutility function is extremely steep around 1, but even then the region in which C4 occurs will be very small.

Once the correct case has been identified, we can obtain the revenue function (2.52) by plugging in the appropriate expressions for $q_e(c)$ attained in each of the relevant regimes. In regimes R1 and R3, we have $\bar{q} \leq q_e(c)$ and so the revenue grows linearly in c as long as $(q_e(c), q_r(c))$ belong to one of these regimes. In regime R2, the revenue function is log-concave by Proposition 3, and in regime R4, this can also be shown by direct computation. Consequently, there is a unique revenue-maximizing price in cases C2, C4, C5 and C6, but two local optima in cases C1 and C3, caused by the transition from R2 to R4. Figure 2.3 gives numerical illustrations of R and $R - D$ in all six cases.

The second peak only occurs when the transition from R2 to R4 is present, and is caused by a change in the behavior of the price elasticity of demand between these regimes. Unfortunately, it is not possible in general to guarantee that one of the two peaks will always be better; one can design instances of either case C1 or C3 in which either R2 or R4 generates more revenue.

Note that in Figure 2.3, R has double peaks in cases C1 and C3, while $R - D$ has nonzero maxima in cases C1, C2 and C4. The figures were obtained under different parameter choices, which are omitted here as the purpose is illustrative.

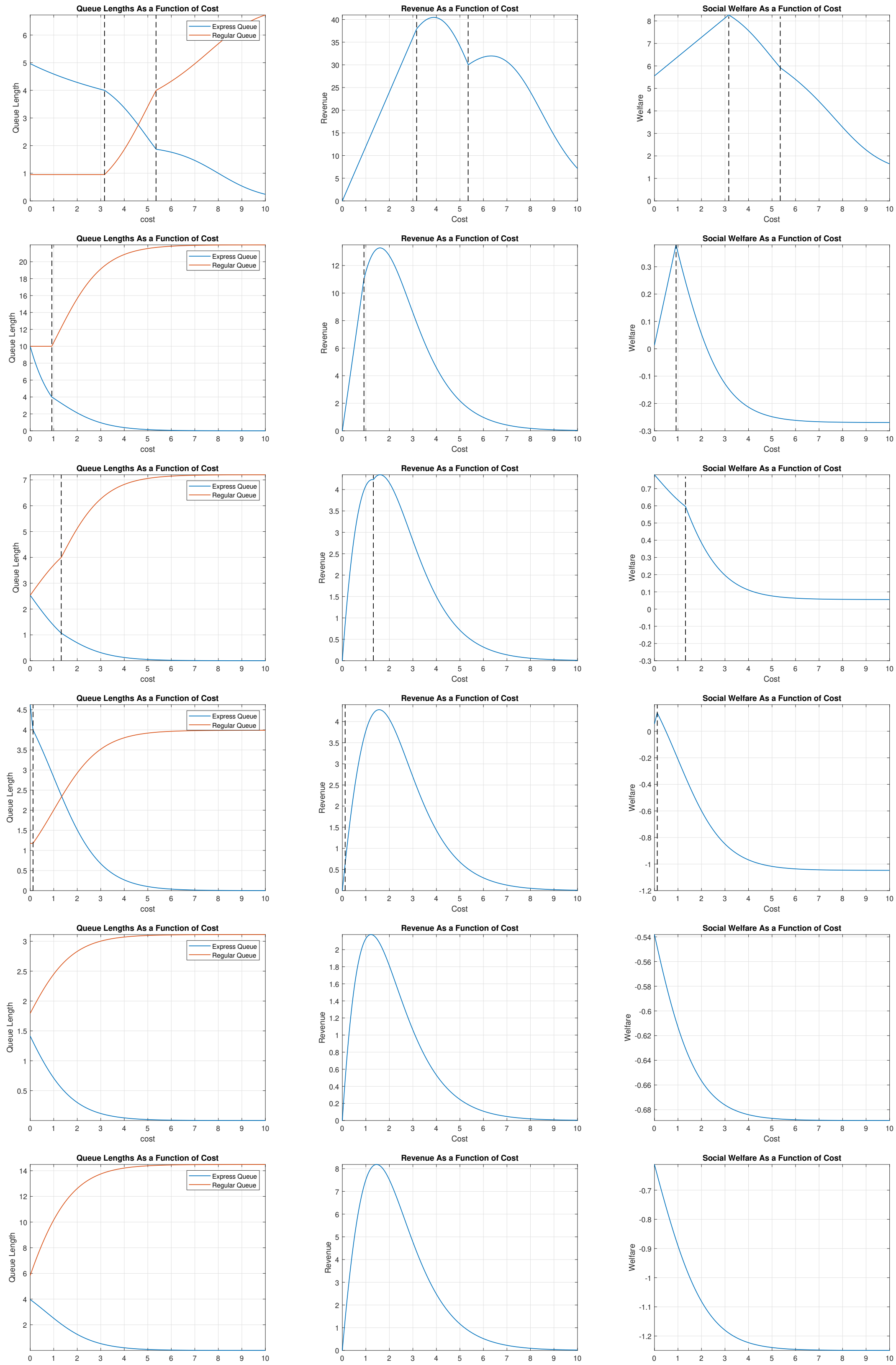


Figure 2.3: Illustrations of equilibrium queue lengths, revenue, and social welfare in cases C1- C6.

2.4.2 Exponomial choice model

Under the exponomial model [22], we assume that all random shocks in the problem are i.i.d. exponentially distributed. In this model, the expression for the choice probabilities depends on the utility order; for example, if $u_e \leq u_r \leq \bar{u}$, we have

$$p_e(u_e, u_r, \bar{u}) = 1 - \frac{1}{2}e^{-\ell(u_r - u_e)} - \frac{1}{6}e^{-\ell((\bar{u} - u_r) + (\bar{u} - u_e))},$$

$$p_r(u_e, u_r, \bar{u}) = \frac{1}{2}e^{-\ell(u_r - u_e)} - \frac{1}{6}e^{-\ell((\bar{u} - u_r) + (\bar{u} - u_e))},$$

$$p_o(u_e, u_r, \bar{u}) = \frac{1}{3}e^{-\ell((\bar{u} - u_r) + (\bar{u} - u_e))},$$

where ℓ is the fixed rate parameter of the exponential distribution. One can, however, examine all of the possible permutations and directly verify that our assumptions in Section 2.2.1 hold. For example, in the case shown above, $\frac{\partial p_e}{\partial u_e} < 0$ and $\frac{\partial p_e}{\partial u_r} > 0$, and the derivatives are uniformly bounded since each exponential term must take values between 0 and 1. It follows that all of the general results from Section 2.2.2 and Section 2.3 apply.

Unfortunately, we do not have closed-form expressions for the equilibrium queue lengths, so we cannot explicitly solve for the thresholds between the four regimes. However, for $\mu_r = 1$ and a given disutility function u, we can still construct a phase diagram numerically, as shown in Figure 2.4. The interpretation of this diagram is the same as in the case of MNL; in particular, we see that the same cases are present. (Case C4 is again rare, but possible for some choices of u.)

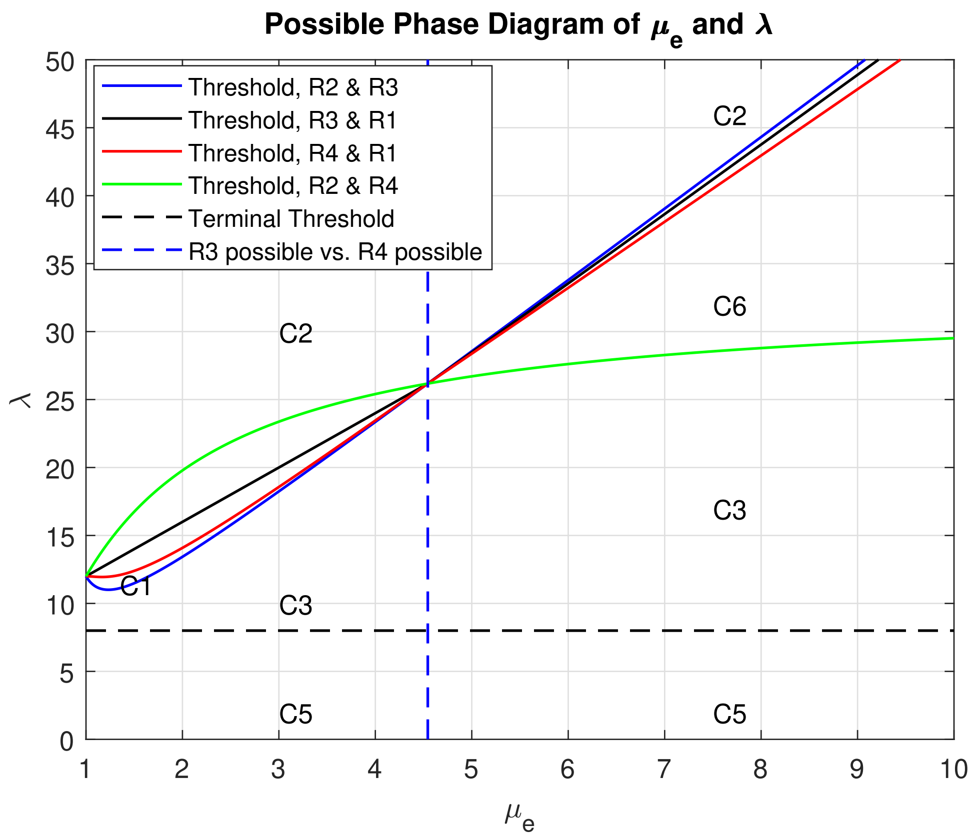


Figure 2.4: Phase diagram illustrating the impact of μ_e, λ on the equilibrium regimes.

We can also numerically evaluate the revenue function. Figure 2.5 shows that, just as in Figure 2.3, the revenue function may still have multiple peaks, and that this behaviour cannot be eliminated by simply using a different choice model.

2.5 Conclusion

We have studied a service system where paying customers join a separate queue with a faster service rate. The system is observable, and newly arriving customers

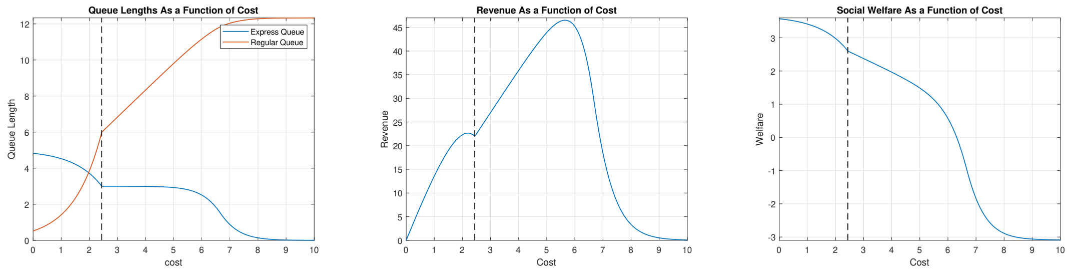


Figure 2.5: Illustration of double peaks under exponomial choice probabilities (case C3).

make the decision to purchase express service based not only on the cost, but also on the lengths of both regular and express queues at that moment. Customer heterogeneity is represented with a general probabilistic choice model, and our analysis can accommodate both multinomial logit and exponomial choice probabilities, together with very general disutility functions.

We find that the limited service capacity in both queues (modeled using $\bar{q}$ servers per queue, in contrast with the $M/M/1$ models studied in much of the related literature) plays a key role in how customers react to the entry fee. Depending on how we change the fee, the equilibrium may transition between different regimes; for example, very low prices may cause crowding in the express queue, very high prices may cause crowding in the free queue, and mid-range prices may eliminate congestion entirely. As a result, the revenue function may exhibit multiple local optima – a revenue-maximizing service provider may opt to artificially drive up congestion in the regular queue, while leaving unused capacity in the express queue, because the benefit of switching from regular to express starts to grow once the regular queue becomes congested. By contrast, if the goal is to optimize social welfare, the price should be low enough to eliminate congestion from the regular queue (or to reduce it by as much as possible) without creating congestion in the express queue.

The main limitation of our model is the assumption of a fixed arrival rate λ and a fixed cost c, though this is consistent with most of the related literature. Several studies have examined the complementary setting of nonstationary arrivals and time-dependent prices, but these elements make the problem much less tractable. The upside of our assumptions is that they allow us to work with very general random utility models, capturing many different forms of customer valuation and heterogeneity. If one is willing to make additional assumptions on the choice model, it can even be possible to solve for the optimal prices in closed form. However, the fundamental structure of the revenue function is quite robust with respect to the particular choice model being used.

Chapter 3: A new rate-optimal design for linear regression

3.1 Introduction

In this paper, we derive a new, large deviations theoretic optimality criterion for linear regression, and propose a new design that optimizes this criterion. Unlike all of the existing work on large deviations-based designs, we do *not* discretize the design space, but rather allow any x on the L^2 sphere $\{x : ||x|| = 1\}$. This requires substantial new technical developments over past work (which is limited to finite sets), and leads to a completely different interpretation of the design. In [43] and related papers, each alternative is assigned a certain nonzero proportion of the sample, which is no longer possible when x is a continuous variable. However, due to the structure of the linear model, we can instead characterize the design as an allocation of the budget to an *orthonormal basis* for the design space, with β itself being one of the basis vectors. We then obtain exceptionally simple closed-form calculations for the optimal proportions to assign to each basis vector. In fact, these optimal proportions are *almost* uniform: one samples β with a certain small probability (computable in closed form) that does not depend on β itself, and otherwise chooses one of the other basis vectors uniformly at random.

Due to this structure, our design is much easier to learn sequentially than any existing design of this type. In discrete problems, such designs require enumeration of all possible alternatives, and make a special distinction between the allocation to the best alternative vs. all the others. As a result, any sequential implementation first has to guess which alternative is the best, and if this guess is incorrect, the estimated proportions will be very inaccurate. In our case, however, by changing the focus to an orthonormal basis for the design space, we do not require any information about which x is optimal; we simply estimate β and extend the estimate to a suitable basis. For this reason, our approach has considerable practical utility (also illustrated in a numerical example) and can serve as a natural benchmark for continuous optimal design in linear regression.

3.2 Large deviations in least squares regression

Return to the model (1.1) and assume, without loss of generality, that $||\beta|| = 1$. Suppose that $\{x_n\}_{n=1}^{\infty}$ is a deterministic sequence satisfying

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top = A \tag{3.1}$$

where A is a symmetric, positive definite matrix. Let $y_i = \beta^\top x_i + \epsilon_i$ with the residuals $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ being independent. The ordinary least-squares estimator b_n of β, given the data (x_i, y_i) for $i = 1, ..., n$, is defined as $b_n = \arg\min_b \sum_{i=1}^{n} (y_i - b^\top x_i)^2$.

3.2.1 Large deviations laws

We derive the following large deviations law for b_n.

Theorem 7. *For any $E \subseteq \mathbb{R}^d$ such that $\beta \notin E$,*

$$\lim_{n\to\infty} \frac{1}{n} \log P(b_n \in E) = -\inf_{u\in E} I(u) \tag{3.2}$$

where $I(u) = \frac{1}{2\sigma^2}(u-\beta)^\top A(u-\beta)$.

Proof. We first describe the major steps in the proof and then complete the computations. First, for any n, we let $\Psi_n(\gamma) = \log \mathbb{E}\left(e^{\gamma^\top b_n}\right)$ be the log-mgf of b_n. Assuming that the scaled limit $\Psi(\gamma) = \lim_{n\to\infty} \frac{1}{n}\Psi_n(n\gamma)$ exists, we let

$$I(u) = \sup_{\gamma} \gamma^\top u - \Psi(\gamma) \tag{3.3}$$

be the Fenchel-Legendre transform of Ψ. The large deviations law (3.2) then follows from the Gartner-Ellis theorem [60]. It remains to explicitly compute Ψ and I.

For any n, b_n can be written [61] as

$$b_n = \beta + \left(\sum_{i=1}^{n} x_i x_i^\top\right)^{-1} \sum_{j=1}^{n} x_j \epsilon_j.$$

Using this representation, we calculate

$$\begin{aligned}
\Psi_n(\gamma) &= \gamma^\top \beta + \log \mathbb{E}\left(e^{\gamma^\top (\sum_{i=1}^n x_i x_i^\top)^{-1} \sum_{j=1}^n x_j \epsilon_j} \right) \\
&= \gamma^\top \beta + \log \mathbb{E}\left(e^{\sum_{j=1}^n [\gamma^\top (\sum_{i=1}^n x_i x_i^\top)^{-1} x_j] \epsilon_j} \right) \\
&= \gamma^\top \beta + \sum_{j=1}^n \frac{1}{2}\sigma^2 \left[\gamma^\top \left(\sum_{i=1}^n x_i x_i^\top \right)^{-1} x_j \right]^2 .
\end{aligned}$$

Consequently, the scaled limit Ψ is found to be

$$\begin{aligned}
\Psi(\gamma) &= \gamma^\top \beta + \lim_{n\to\infty} \sum_{j=1}^n \frac{1}{2}\sigma^2 n \left[\gamma^\top \left(\sum_{i=1}^n x_i x_i^\top \right)^{-1} x_j \right]^2 \\
&= \gamma^\top \beta + \lim_{n\to\infty} \sum_{j=1}^n \frac{1}{2}\sigma^2 n \gamma^\top \left(\sum_{i=1}^n x_i x_i^\top \right)^{-1} x_j x_j^\top \left(\sum_{i=1}^n x_i x_i^\top \right)^{-1} \gamma \\
&= \gamma^\top \beta + \lim_{n\to\infty} \frac{1}{2}\sigma^2 n \gamma^\top \left(\sum_{i=1}^n x_i x_i^\top \right)^{-1} \left(\sum_{j=1}^n x_j x_j^\top \right) \left(\sum_{i=1}^n x_i x_i^\top \right)^{-1} \gamma \\
&= \gamma^\top \beta + \lim_{n\to\infty} \frac{1}{2}\sigma^2 \gamma^\top \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right)^{-1} \gamma \\
&= \gamma^\top \beta + \frac{1}{2}\sigma^2 \gamma^\top A^{-1} \gamma.
\end{aligned}$$

Then, (3.3) becomes

$$I(u) = \sup_\gamma \gamma^\top (u - \beta) - \frac{1}{2}\sigma^2 \gamma^\top A^{-1} \gamma.$$

The supremum is achieved at $\gamma^\star$ satisfying

$$\sigma^2 A^{-1}\gamma^\star = u - \beta \implies \gamma^\star = \frac{1}{\sigma^2}A(u-\beta).$$

Substituting $\gamma^\star$ into (3.3) yields $I(u) = \frac{1}{2\sigma^2}(u-\beta)^\top A(u-\beta)$, as required.

□

In words, since the true coefficients β satisfy $\beta \notin E$, the event $\{b_n \in E\}$ represents an "error" of some sort. As $n \to \infty$, the probability of error decays exponentially, but the exponent can be controlled by changing the matrix A. Although we have treated the data sequence $\{x_n\}$ as deterministic in this discussion, intuitively one can think of (3.1) as a kind of "law of large numbers" for the data-generating process. For example, if we were given some desired A, we could generate x_n i.i.d. from some distribution, independent of $\{\epsilon_n\}_{n=1}^\infty$ and satisfying $\mathbb{E}(x_n x_n^\top) = A$, and still achieve the large deviations law.

In the remainder of this paper, we will primarily focus on error events of the form

$$E_v = \left\{u \in \mathbb{R}^d : u^\top v \leq 0\right\} \tag{3.4}$$

for various fixed vectors $v \in \mathbb{R}^d$ that satisfy $\beta^\top v > 0$. The rate exponent for any such event can be computed in closed form, as shown in the following result.

Proposition 4. *Suppose that $\beta^\top v > 0$. Then,*

$$\lim_{n\to\infty} \frac{1}{n} \log P(b_n^\top v \leq 0) = -\frac{1}{2\sigma^2} R(v)$$

where $R(v) = \frac{(\beta^\top v)^2}{v^\top A^{-1} v}$.

Proof. From Theorem 7, it follows that $R(v)$ is the optimal value of the convex program

$$\begin{aligned} &\min_{u \in \mathbb{R}^d} (u - \beta)^\top A (u - \beta) \\ &\text{s.t.} \qquad v^\top u \leq 0. \end{aligned} \tag{3.5}$$

Letting λ be the Lagrange multiplier of the single linear constraint, the optimality conditions of (3.5) are given by

$$A(u - \beta) + \lambda v = 0, \tag{3.6}$$

$$v^\top u = 0, \tag{3.7}$$

where (3.7) follows because the linear constraint should be binding at optimality. Now, (3.6) yields

$$u = \beta - \lambda A^{-1} v, \tag{3.8}$$

and plugging (3.8) into (3.7) leads to

$$v^\top \beta - \lambda v^\top A^{-1} v = 0 \implies \lambda = \frac{v^\top \beta}{v^\top A^{-1} v}.$$

Plugging this back into (3.8), we obtain

$$u^\star = \beta - \frac{v^\top \beta}{v^\top A^{-1} v} A^{-1} v,$$

whence

$$\begin{aligned} I(u^\star) &= (u^\star - \beta)^\top A (u^\star - \beta) \\ &= \left(\frac{v^\top \beta}{v^\top A^{-1} v} \right)^2 v^\top A^{-1} A A^{-1} v \\ &= \frac{(v^\top \beta)^2}{v^\top A^{-1} v}, \end{aligned}$$

as required.

□

Thus, the convergence rate of $P(b_n \in E_v)$ is governed by the exponent $R(v)$, which depends on the specific vector v we are studying; note that $R(v)$ is invariant with respect to $||v||$, so we can assume $||v|| = 1$ whenever it is convenient to do so. We can now study error events of the form $\bigcup_k E_{v_k}$ for countable collections $\{v_k\}_{k=1}^\infty$ that are dense in some uncountable set of interests. A straightforward consequence of Theorem 7 and Proposition 4 is that

$$\lim_{n \to \infty} \frac{1}{n} \log P \left(b_n \in \bigcup_k E_{v_k} \right) = - \inf_k R(v_k), \tag{3.9}$$

provided that $\beta \notin \bigcup_k E_{v_k}$ (or, equivalently, $\inf_k \beta^\top v_k > 0$). Intuitively, the probability that at least one error event in the collection occurs is determined by the slowest convergence rates among the individual error events.

3.2.2 Optimal designs

Let $x^\star \in \mathbb{R}^d$ be some fixed "reference solution," possibly obtained from some optimization problem that will not be explicitly modeled here. The value of this solution is $\beta^\top x^\star$. We assume that larger values are better, so

$$\mathcal{X}(x^\star) = \left\{ x \in \mathbb{R}^d : \beta^\top (x^\star - x) > 0 \right\}$$

is interpreted as the set of all inferior solutions. If there is any $x \in \mathcal{X}(x^\star)$ for which $b_n^\top (x^\star - x) \leq 0$, this means that the estimated coefficients b_n have led us to erroneously identify x as being superior to $x^\star$. This is clearly an example of (3.4) with $v = x^\star - x$. Note that the convergence rate of $P(b_n \in E_v)$ only depends on $x^\star$ and x through the "optimality gap" v.

Potentially, any $x \in \mathcal{X}(x^\star)$ can generate an error. Consider a countable collection $\{x_k\}_{k=1}^\infty \subseteq \mathcal{X}(x^\star)$. Each x_k corresponds to an error vector $v_k = x^\star - x_k$, motivating an optimization problem of the form

$$\sup_{A \in \mathbb{S}_{++}^d} \inf_k \frac{(v_k^\top \beta)^2}{v_k^\top A^{-1} v_k}, \tag{3.10}$$

where $\mathbb{S}_{++}^d$ is the set of all $d \times d$ symmetric positive definite matrices. Through (3.9), this problem chooses the matrix A to make $P(b_n \in \bigcup_k E_{v_k})$ converge to zero at the fastest possible rate. Of course, to ensure that (3.10) is not unbounded, we would also need to impose a simple constraint on the magnitude of A, such as an upper bound on the trace. Such an upper bound serves as a scale factor on $R(v_k)$ for all k, but otherwise does not change the geometry of the optimal A.

However, we require $\beta \notin \bigcup_k E_{v_k}$ in order to use Theorem 7, which means that we cannot make $\{x_k\}$ dense in the entire set $\mathcal{X}(x^\star)$. Instead, we will focus on $\{v_k\} \subseteq V_\delta$ where

$$V_\delta = \left\{ v : ||v|| = 1, \beta^\top v \geq \delta \right\}$$

and $\delta > 0$ is a small constant. This ensures that $\inf_k R(v_k)$ is strictly positive (for any fixed positive definite A) and allows (3.9) to be applied. Essentially, we are now willing to accept $x \in \mathcal{X}(x^\star)$ whose value is sufficiently close to that of $x^\star$, and we focus on eliminating errors generated by solutions that are outside this tolerance level. Note that our *design space* need not be restricted to V_δ. The parameter δ only imposes restrictions on the error events that we are trying to eliminate.

With this modification, one can rewrite (3.10) as

$$\sup_{A \in \mathbb{S}^d_{++}} \min_{v \in V_\delta} \frac{(v^\top \beta)^2}{v^\top A^{-1} v}. \tag{3.11}$$

Since A is symmetric and positive definite, we can write $A = \sum_{i=1}^d p_i \zeta_i \zeta_i^\top$ where $p_i > 0$ and $(\zeta_1, ..., \zeta_d)$ is an orthonormal basis for $\mathbb{R}^d$. We may assume that $\sum_i p_i = 1$ without loss of generality; as discussed earlier, this condition scales the optimal A without changing its geometry. Recalling the interpretation of A as an expected value, p_i can be seen as the probability of sampling ζ_i.

So far, (3.11) requires us to jointly choose both eigenvalues and eigenvectors. We will simplify this problem by setting $\zeta_1 = \beta$, that is, β itself will be an eigenvector. With this, the orthonormal basis can be straightforwardly completed, and the only remaining decision variable is the vector p of eigenvalues. We first give some intuition for this choice. For any fixed positive definite B, the ratio $\frac{(\beta^\top v)^2}{v^\top Bv}$ can in general be made arbitrarily small. However, if we allow the positive *semidefinite* matrix $B = \beta\beta^\top$, the ratio evaluates to 1 for any v with $\beta^\top v \neq 0$. This suggests that, when we choose a positive definite B, its principal eigenvector should also be aligned with β.

Before providing more rigorous support for this idea, we first manipulate the problem setup as follows. Let $B = \sum_i r_i \zeta_i \zeta_i^\top$, where $(\zeta_1, ..., \zeta_d)$ is an orthonormal basis for $\mathbb{R}^d$, and $r_1 > r_2 \geq ... \geq r_d > 0$ are the eigenvalues. It can easily be seen that $\min_{v \in V_\delta} \frac{(\beta^\top v)^2}{v^\top Bv}$ is attained on the boundary $\partial V_\delta = \{v : ||v|| = 1, \beta^\top v = \delta\}$. Then, the problem $\max_B \min_{v \in \partial V_\delta} \frac{(\beta^\top v)^2}{v^\top Bv}$ has the same optimal solution as the problem $\min_B \max_{v \in \partial V_\delta} v^\top Bv$. We then show that, when δ is small, the optimal value of the inner maximization can be bounded below by the second-largest eigenvalue of B, regardless of the choice of orthonormal basis.

Proposition 5. *For sufficiently small δ, we have $\max_{v \in \partial V_\delta} v^\top Bv \geq r_2$.*

Proof. We consider two cases: one where $\beta = \zeta_i$ for some i, and one where $\beta \neq \zeta_i$ for any i. In the first case, an optimal solution can be found by taking $v = \delta\zeta_1 +$

$\sqrt{1-\delta^2} \cdot \zeta_2$ if $\beta = \zeta_1$, or $v = \delta\zeta_i + \sqrt{1-\delta^2} \cdot \zeta_1$ if $\beta = \zeta_i$ for $i \neq 1$. Either way, the optimal value is bounded below by r_2 for sufficiently small δ.

Now consider the case where $\beta \neq \zeta_i$ for any i. Define $v = \delta\beta + Pw$, where $P = I - \beta\beta^\top$ is the projection onto the orthogonal complement of β. Then, the objective $\frac{v^\top \beta v}{v^\top v}$, which coincides with $v^\top \beta v$ when $v^\top v = 1$, can be rewritten in terms of w as

$$f(w) = \frac{w^\top PBPw + 2\delta w^\top PB\beta + \delta^2 \beta^\top B\beta}{w^\top Pw + \delta^2}.$$

Observe that

$$\frac{\partial f}{\partial w} = \frac{1}{w^\top Pw + \delta^2}(2PBPw + 2\delta PB\beta - 2f(w)Pw).$$

Setting the derivative equal to zero yields

$$PBPw + \delta PB\beta = f \cdot Pw. \tag{3.12}$$

Given any solution (f, w) of (3.12), we can obtain a feasible $v = \delta\beta + Pw$ whose objective value is f. Observe, however, that such a solution may be found for almost any f value: we may rewrite (3.12) as $(fI - PB)Pw = \delta PB\beta$, where the matrix $fI - PB$ is invertible as long as f is not equal to any of the eigenvalues $s_1 \geq \ldots \geq s_d$ of PB. Consequently, given any f satisfying $f \neq s_i$ for all i, we can obtain $Pw = \delta(fI - PB)^{-1}PB\beta$ such that $v = \delta\beta + Pw$ satisfies $\frac{v^\top \beta v}{v^\top v} = f$.

However, we also require v to satisfy the normalization condition $v^\top v = 1$.

Equivalently, we must have $w^\top P^2 w = 1 - \delta^2$, which becomes

$$\frac{1-\delta^2}{\delta^2} = b^\top BP(fI - PB)^{-2}PB\beta. \tag{3.13}$$

Thus, the optimal value of $\max_{v \in \partial V_\delta} v^\top Bv$ is the largest f for which (3.13) holds. Since the right-hand side of (3.13) has a cusp at $f = s_1$ and decreases monotonically on (s_1, ∞), the largest root satisfies $f > s_1$. By the Courant-Fischer theorem [62], Thm. 4.2.6, we have $r_1 \geq s_1 \geq r_2$, whence $f \geq r_2$.

□

Proposition 5 shows that, no matter how we choose the orthonormal basis, the inner maximum $\max_{v \in \partial V_\delta} v^\top Bv$ cannot be reduced below r_2. Thus, we may simply set $\zeta_1 = \beta$, in which case $\max_{v \in \partial V_\delta} v^\top Bv = \delta^2 r_1 + (1 - \delta^2) r_2$, a quantity that can be made arbitrarily close to the lower bound for sufficiently small δ. Returning to our original problem (3.11), since we are primarily interested in the small-δ regime, we will impose the structure

$$A = p_1 \beta\beta^\top + \sum_{i>1} p_i \zeta_i \zeta_i^\top, \tag{3.14}$$

where the other vectors $\zeta_2, ..., \zeta_d$ in the orthonormal basis are unique (up to multiplication by -1). The remainder of this paper will derive the optimal eigenvalues p_i. In fact, we will see that $p_1 = \min_i p_i$ in the optimal solution, confirming the intuition that β should be the principal eigenvector of A^{-1}.

3.3 Solving for the optimal design

Suppose that the sequence $\{v_k\}$ is dense in V_δ. Since $R(v)$ is invariant with respect to $||v||$, we can focus on unit vectors without loss of generality. For fixed K, we consider the problem

$$\max_p \min_{k \leq K} \frac{(v_k^\top \beta)^2}{\frac{1}{p_1}(v_k^\top \beta)^2 + \sum_{i>1} \frac{1}{p_i}(v_k^\top \zeta_i)^2} \tag{3.15}$$

subject to the constraints $p \geq 0$, $\sum_i p_i = 1$. Equation (3.15) is a version of (3.10) with (3.14) plugged into the denominator. As $K \to \infty$, the inner minimum in (3.15) will behave like a minimum over all $v \in V_\delta$. Since we are mainly interested in this asymptotic regime, we can choose the elements of $\{v_k\}$ in any way we want, as long as the sequence remains dense in V_δ.

The objective function in (3.15) is concave in p and can be rewritten as $\max_{p,z} z$ subject to

$$z \leq \frac{(v_k^\top \beta)^2}{\frac{1}{p_1}(v_k^\top \beta)^2 + \sum_{i>1} \frac{1}{p_i}(v_k^\top \zeta_i)^2}, \qquad k = 1, ..., K. \tag{3.16}$$

in addition to the original constraints on p. The Lagrangian of this optimization problem is given by

$$L(z, p, \mu, \nu) = -z + \sum_{k=1}^{K} \mu_k \left(z - \frac{(v_k^\top \beta)^2}{\frac{1}{p_1}(v_k^\top \beta)^2 + \sum_{i>1} \frac{1}{p_i}(v_k^\top \zeta_i)^2} \right) + \nu \left(\sum_i^d p_i - 1 \right),$$

with the terms corresponding to the nonnegativity constraints on p_i omitted, in order to ensure that A is positive definite. The optimality conditions are as follows:

1. First-order conditions:

$$\sum_{k=1}^{K} \mu_k \frac{(v_k^\top \beta)^4}{\left[\frac{1}{p_1}(v_k^\top \beta)^2 + \sum_{i>1} \frac{1}{p_i}(v_k^\top \zeta_i)^2\right]^2} = p_1^2 \nu, \tag{3.17}$$

$$\sum_{k=1}^{K} \mu_k \frac{(v_k^\top \beta)^2 (v_k^\top \zeta_i)^2}{\left[\frac{1}{p_1}(v_k^\top \beta)^2 + \sum_{i>1} \frac{1}{p_i}(v_k^\top \zeta_i)^2\right]^2} = p_i^2 \nu, \qquad i = 2, ..., d \tag{3.18}$$

$$\sum_{k=1}^{K} \mu_k = 1. \tag{3.19}$$

2. Primal feasibility: (3.16) and $\sum_i p_i = 1$, $p_i > 0$ for all i.

3. Dual feasibility: $\mu_k \geq 0$.

4. Complementary slackness:

$$\mu_k \left(z - \frac{(v_k^\top \beta)^2}{\frac{1}{p_1}(v_k^\top \beta)^2 + \sum_{i>1} \frac{1}{p_i}(v_k^\top \zeta_i)^2} \right) = 0, \qquad k = 1, ..., K. \tag{3.20}$$

The first-order conditions (3.17)-(3.18) can be viewed as a system of d linear equations in K variables $\mu_1, ..., \mu_K$. For large K, this system may have many solutions. In particular, we can construct a basic solution by taking d linearly independent vectors $v_{k_1}, ..., v_{k_d}$ from $\{v_k\}_{k=1}^K$ and setting $\mu_k = 0$ if $k \notin \{k_1, ..., k_d\}$. Since $\{v_k\}$ is dense in a set of dimension d, we can choose individual v_k to take certain values in that set without affecting the asymptotic result. For our analysis, it is convenient to take $w_1 = \beta$ and let w_j be a linear combination of β and ζ_j, for $j = 2, ..., d$, with $w_j^\top \zeta_i = 0$ for any $i \neq j$. We may assume that, for any j, there exists $k_j \leq K$ such that $w_j = v_{k_j}$.

With this choice of w_j, we can rewrite (3.17)-(3.18) as

$$p_1^2\mu_{k_1} + \sum_{j>1} \mu_{k_j} \frac{(w_j^\top \beta)^4}{\left[\frac{1}{p_1}(w_j^\top \beta)^2 + \frac{1}{p_j}(w_j^\top \zeta_j)^2\right]^2} = p_1^2 \nu \tag{3.21}$$

$$\mu_{k_j} \frac{(w_j^\top \beta)^2 (w_j^\top \zeta_j)^2}{\left[\frac{1}{p_1}(w_j^\top \beta)^2 + \frac{1}{p_j}(w_j^\top \zeta_j)^2\right]^2} = p_j^2 \nu, \qquad j = 2, ..., d. \tag{3.22}$$

Substituting (3.22) into (3.21) yields

$$p_1^2\mu_{k_1} + \nu \sum_{j>1} p_j^2 \frac{(w_j^\top \beta)^2}{(w_j^\top \zeta_j)^2} = p_1^2 \nu. \tag{3.23}$$

If we set $\mu_{k_1} = 0$, the dual variable ν cancels out of (3.23), yielding

$$p_1^2 = \sum_{j>1} p_j^2 \frac{(w_j^\top \beta)^2}{(w_j^\top \zeta_j)^2}. \tag{3.24}$$

Note that, for any p, it is easy to find $\mu_{k_j} > 0$ and ν to satisfy (3.22). Condition (3.19) can also be easily satisfied by rescaling these values. The complementary slackness condition (3.20) is satisfied for any $k \notin \{k_2, ..., k_d\}$ since the corresponding dual variables μ_k are set to zero. To satisfy the condition for the remaining values of k, it is sufficient to ensure that $R(w_i) = R(w_j)$, that is,

$$\frac{(w_i^\top \beta)^2}{\frac{1}{p_1}(w_i^\top \beta)^2 + \frac{1}{p_i}(w_i^\top \zeta_i)^2} = \frac{(w_j^\top \beta)^2}{\frac{1}{p_1}(w_j^\top \beta)^2 + \frac{1}{p_j}(w_j^\top \zeta_j)^2}, \qquad i, j \neq 1. \tag{3.25}$$

Thus, as long as p is chosen to satisfy (3.24)-(3.25), we can find feasible μ, ν to satisfy (3.17)-(3.20). Essentially, most of the optimality conditions for the problem (3.15) have reduced to the conditions (3.24)-(3.25) on p, which generalize those derived in Example 1 of [43] for large deviations of pairwise comparisons between scalar normal distributions.

In fact, there is only one optimality condition for (3.15) that has not yet been treated, namely (3.16). Our choice of p must also imply $R(w_i) \leq R(v_k)$ for all $i = 2, ..., d$ and $k = 1, ..., K$. Recalling that we have the freedom to pick w_j , we further suppose that $(w_j^\top \beta)^2 = \delta^2$ for $j = 2, ..., d$. Since each w_j is a unit vector, it follows that $(w_j^\top \zeta_j)^2 = 1 - \delta^2$. Consequently, (3.25) now implies that $p_i = p_j = c$ for $i, j \neq 1$ and some constant c. Then, for any $v \in V_\delta$, the rate exponent $R(v)$ simplifies to

$$R(v) = \frac{(v^\top \beta)^2}{\frac{1}{p_1}(v^\top \beta)^2 + \frac{1}{c}\sum_{i>1}(v^\top \zeta_i)^2}.$$

Note that, since $(v^\top \beta)^2 \geq \delta^2$ for any $v \in V_\delta$, we must also have $\sum_{i>1}(v^\top \zeta_i)^2 \leq 1 - \delta^2$ because v is a unit vector. Consequently,

$$R(v) \geq \frac{\delta^2}{\frac{1}{p_1}\delta^2 + \frac{1}{c}(1-\delta^2)} = R(w_j)$$

for any $j = 2, ..., d$. Thus, our choice of w has caused (3.16) to be satisfied for any $v \in V_\delta$. Therefore, the solution $p^\star$ of (3.24)-(3.25), for this choice of w, is optimal for *any* arbitrarily large K, and therefore

$$p^\star = \arg \max_{p:\sum_i p_i = 1} \min_{v \in V_\delta} R(v)$$

also optimizes the convergence rate of the probability that an error arises from any $v \in V_\delta$.

It remains to calculate $p^\star$. Letting $\Delta = \frac{\delta^2}{1-\delta^2}$, we find that (3.24) reduces to

$$p_1^2 = (d-1)\Delta c^2.$$

At the same time, $p_1 = 1-(d-1)c$, whence

$$1-(d-1)c = c\sqrt{(d-1)\Delta},$$

leading to the closed-form solution

$$p_1^\star = \frac{\sqrt{(d-1)\Delta}}{(d-1)+\sqrt{(d-1)\Delta}}, \tag{3.26}$$

$$p_i^\star = \frac{1}{(d-1)+\sqrt{(d-1)\Delta}}, \qquad i = 2, ..., d. \tag{3.27}$$

Recalling our earlier interpretation of A as an expected value, the representation (3.14) allows us to view the design as a discrete probability distribution where each p_i represents the probability of collecting a data point using ζ_i as the covariate vector. The solution (3.26)-(3.27) indicates that the optimal distribution is *almost* uniform: any basis vector that is orthogonal to β can be sampled with the same probability. However, the probability assigned to the first eigenvector β is different from the others; as δ becomes smaller, this probability is reduced, which means that

A^{-1} will correspondingly place more weight on $\beta\beta^\top$, as expected.

One especially striking aspect of this solution is that the probabilities $p_i^\star$ are completely deterministic. Thus, the only unknown quantity in (3.14) is β itself, as suitable ζ_i can be straightforwardly computed if β is known. However, one does not need to know $x^\star$ in order to apply the optimal design. Another way to interpret our results is that, for *any* $x^\star$, the probability that $b_n^\top(x^\star - x) > 0$ for *all* x satisfying $\beta^\top(x^\star - x) \geq \delta$ converges to 1 at the fastest possible rate.

3.4 Algorithm and numerical example

Figure 3.1 states a very simple algorithm (which we call "LD-optimal") for implementing the optimal design in practice. Essentially, we use the least-squares estimator b_n in place of β. The estimator itself can be updated recursively, but in every iteration we have to extend it to an orthonormal basis. A simple way to do this is to take d arbitrary linearly independent vectors $(\zeta_1, ..., \zeta_d)$ and apply the Gram-Schmidt process to $(b_n, \zeta_1, ..., \zeta_d)$. Note that the algorithm does not need to know or estimate $x^\star$, unlike virtually every known large deviations-based optimal design.

To evaluate this procedure, we consider the following test setting in five dimensions. First, for $i = 1, ..., 100$, we generate vectors $x_i^\star \in \mathbb{R}^5$, where each component of $x_i^\star$ is drawn from a uniform distribution on [-1,1]. We also generate a vector β

Step 0: Let $n = 1$, initialize $b_1 \in \mathbb{R}^d$ and $A_1 \in \mathbb{S}^d_{++}$.

Step 1: Calculate vectors $\zeta_{n,i}$ such that $\left(\frac{b_n}{\|b_n\|}, \zeta_{n,2}, ..., \zeta_{n,d}\right)$ is an orthonormal basis for $\mathbb{R}^d$.

Step 2: Set

$$x_{n+1} = \begin{cases} \frac{b_n}{\|b_n\|} & \text{w.p. } p_1^* \\ \zeta_{n,i} & \text{w.p. } p_i^*. \end{cases}$$

Step 3: Observe $y_{n+1} = \beta^\top x_{n+1} + \varepsilon_{n+1}$ and update

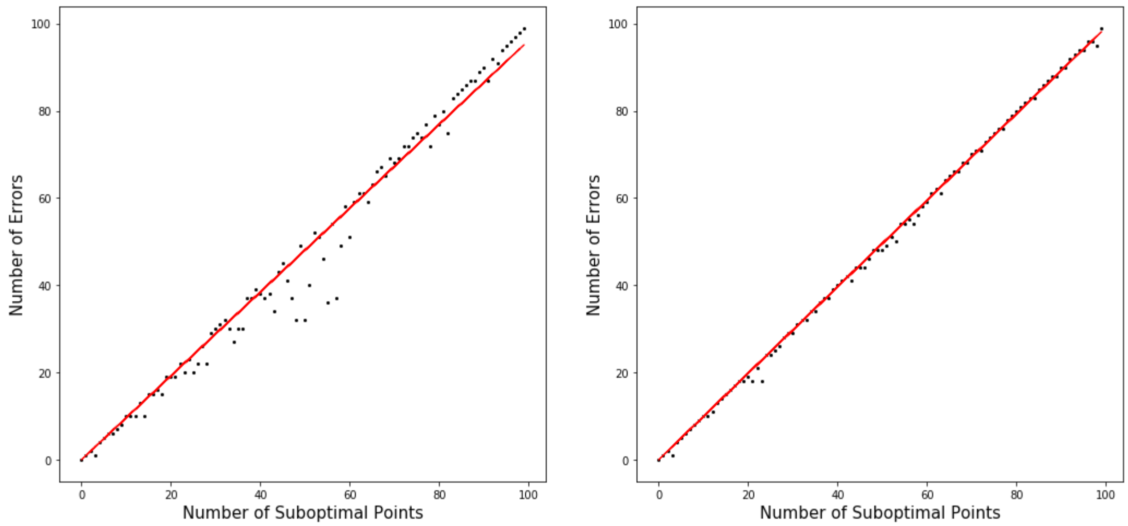
$$\begin{aligned} b_{n+1} &= b_n + \frac{y_{n+1} - b_n^\top x_{n+1}}{1 + x_{n+1}^\top A_n x_{n+1}} A_n x_{n+1}, \\ A_{n+1} &= A_n - \frac{A_n x_{n+1} x_{n+1}^\top A_n}{1 + x_{n+1}^\top A_n x_{n+1}^\top}. \end{aligned}$$

Increment n by 1 and return to step 1.

Figure 3.1: LD-optimal algorithm for sequential implementation of the optimal design.

in the same way and normalize it. Suppose that the vectors $x_i^\star$ are sorted in order of decreasing $\beta^\top x_i^\star$. Thus, for fixed $1 \leq i \leq 100$, there are exactly $i-1$ vectors that are suboptimal relative to $x_i^\star$. We can then collect n observations using the algorithm in Figure 3.1 and calculate, for each i, how many of these $i-1$ suboptimal choices are mistakenly identified as being superior to $x_i^\star$. Figure 3.2 gives an illustration of this calculation for two values of n: the horizontal axis represents the index i, while the vertical axis gives the number of errors for that i value. Note that the number of errors can never be greater than i itself. The red line in Figure 3.2 is the zero-intercept regression line drawn through the points, which can help to visualize how well we are doing (if there are no errors, the slope of this line will be 1).

We compare our approach against two benchmarks: the Randomized Adaptive



(a) $n = 100$ (b) $n = 1000$

Figure 3.2: Illustration of error counts.

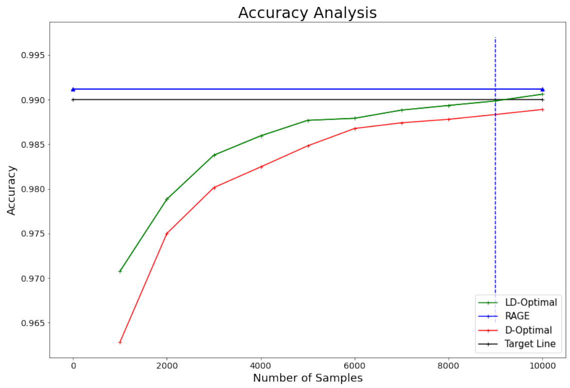
Gap Elimination (RAGE) procedure of [53], and a classical design of experiments procedure known as D-optimal [32]. The RAGE method assumes that the sampling decision is restricted to a pre-specified finite set of vectors, which does not need to be the same as the set of $x_i^\star$ vectors whose values we are learning. With such a discretization of the design space, the D-optimal method can be formulated as a convex optimization problem [63] that can be solved efficiently. Thus, for $i = 1, ..., 100$, we generate vectors z_i uniformly on the unit sphere in $\mathbb{R}^5$, and use these as the input to both benchmarks. Our proposed LD-optimal algorithm does not use these values since it can sample anywhere on the unit sphere.

Remark2 In fact, like our method, RAGE does not require any knowledge of $x^\star$, which is why we see it as the most natural benchmark. We do not compare against, e.g., the knowledge gradient method of [38], or the Thompson sampling method of [64], because these focus on identifying a particular $x^\star$ value.

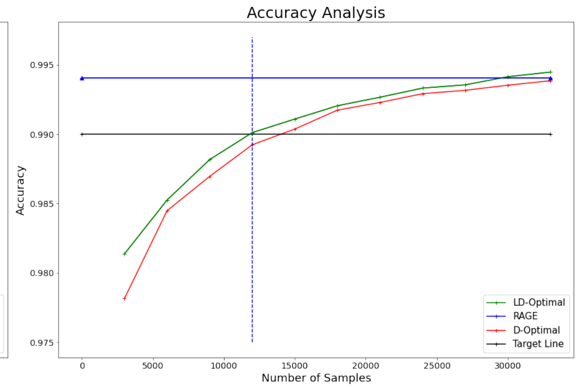
Although both benchmarks are designed for learning in linear regression, they make different assumptions about how the problem proceeds. D-optimal generates vector independently from the set $\{z_i\}$ according to a probability mass function that maximizes $\log\det \mathbb{E}(x_i x_i^\top)$. Recall that we assumed the long-run behaviour for $x_i^\top x_i$ equals A, so we may think D-optimal is equivalent to maximize $\log\det(A) = \log \prod p_i = \sum \log p_i$. The problem can be maximized by setting p_i equal to each other. Since we imposed $\mathrm{tr}(A) = 1$, we may pick an arbitrary orthonormal basis and sample uniformly from it, and this is equivalent to uniformly generating points on a unit sphere.

On the other hand, the RAGE algorithm is adaptive and proceeds in "phases." In each phase, some elements are removed from the set $\{z_i\}$ based on the most recent estimated regression coefficients, and each of the remaining elements is sampled a certain number of times. The procedure terminates when only one element is left; the screening and sampling steps are constructed to ensure that a desired error probability (given as an input to the algorithm) is achieved at termination. However, the number of phases and samples needed for termination is not known ahead of time.

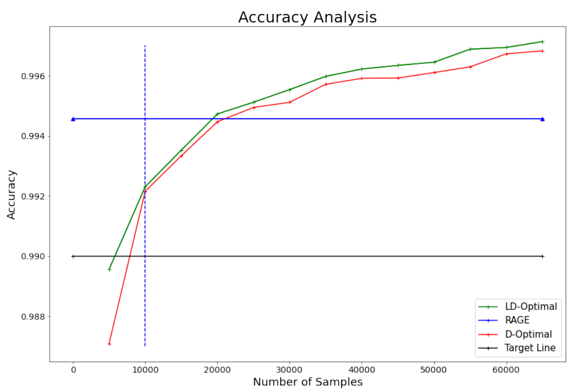
Since our method has no explicit termination criterion (rather, it can be run for as long as our sampling budget allows), we conduct the comparison as follows. First, we run the RAGE algorithm with a desired error probability of $\delta = 0.01$ (this same threshold is also used to set Δ in LD-optimal) to see how many samples it uses. This number is then used as the budget for both D-optimal and LD-optimal. The number of phases can vary widely depending on the test instance, i.e., the set



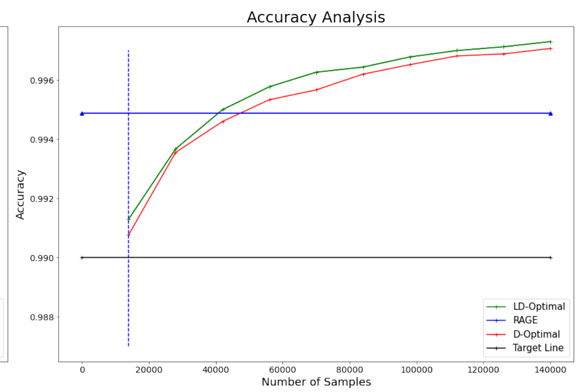
(a) 1 phase



(b) 3 phases



(c) 4 phases

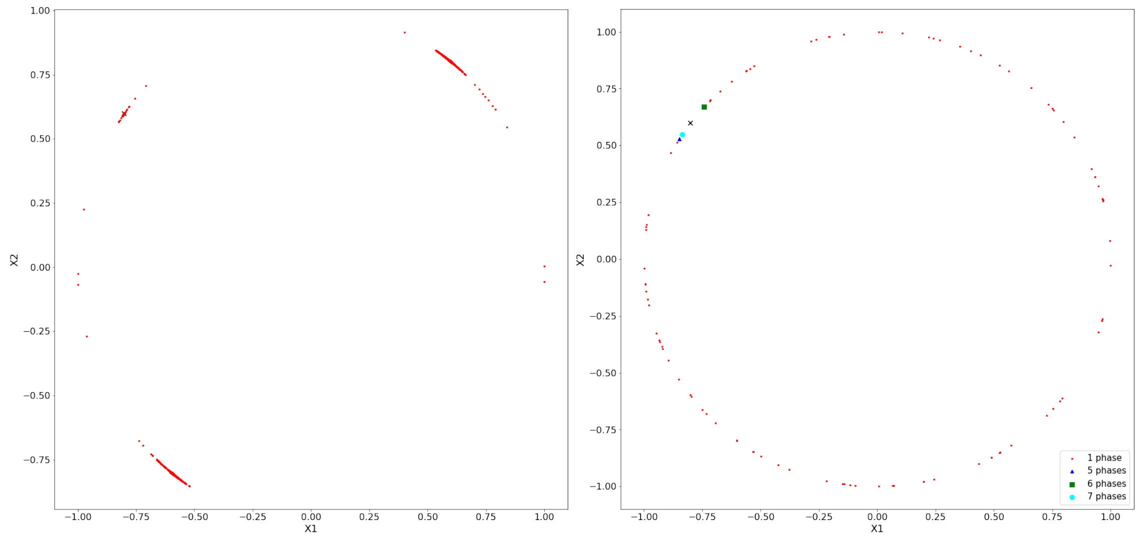


(d) 5 phases

Figure 3.3: Illustration of accuracy.

of vectors $x_i^\star$ that is generated: when there are more of these vectors clustered close together, it is easier to make errors and so more samples are required. Figure 3.3 shows how the accuracy of LD-optimal (averaged over all 100 possible choices of x_i) improves over time for four instances in which RAGE requires 1, 3, 4 and 5 phases, respectively. The performance of LD-optimal is averaged over 100 sample paths to smooth out the trajectory.

We find that, if LD-optimal is allowed to run for as long as RAGE, it achieves very comparable performance, and even outperforms RAGE. We also observe (em-



(a) LD-optimal

(b) RAGE

Figure 3.4: Illustration of empirical sampling distributions.

pirically) that RAGE tends to be somewhat conservative, i.e., its accuracy is higher than the target of 0.99 that was requested. LD-optimal tends to achieve this target accuracy with fewer samples, as indicated by vertical lines in Figure 3.3; for example, in the 5-phase instance, LD-optimal reaches the target with less than 20,000 samples, while RAGE runs for over 140,000. These additional samples only improve the accuracy by $\mathcal{O}(10^{-3})$, which seems to be a classic case of diminishing returns. We acknowledge that RAGE comes with strong guarantees on the error probability at the moment of termination; however, it has often been observed in the past [65] that such "fixed-precision" guarantees often come at the cost of conservativeness. Thus, if the sampling budget is a severe constraint, we believe that LD-optimal offers a powerful alternative.

It is also interesting to consider the empirical distribution of the sampled design points. Figure 3.4 provides an illustration for a different instance in $\mathbb{R}^2$ where the design space can be easily visualized. Again, we discretized the design space into 100 points, generated from a uniform distribution on the unit sphere, in order to run RAGE and D-optimal. In this instance, RAGE ran for 7 phases, which determined the sampling budget for the other two methods. The true value of β is indicated by an X in Figure 3.4a. We see that most of the sampling effort of RAGE is concentrated on a handful of points close to β, with all of the other design points screened out after just one phase. The sampling distribution for D-optimal design is omitted since it is just a uniform distribution.

The LD-optimal method concentrates around β and its orthogonal complement; note that any of the orthonormal basis vectors can be multiplied by -1 without affecting the theory. Furthermore, a majority of the budget is actually assigned to the orthogonal complement, since $p_1^\star$ in (3.26) will be smaller than the other probabilities when Δ is sufficiently small. Perhaps the most interesting insight to be obtained from our work is that sampling the orthogonal complement of β can also be very important for ruling out suboptimal solutions.

3.5 Conclusion

We have derived a new optimal design for linear regression based on a large deviations theoretic analysis of error probability. Our result has several novel characteristics relative to previous work. First, in the linear regression setting, it is not necessary to specify or estimate a particular "optimal" solution that we are trying to select. The asymptotic behaviour of the error probability depends only on the size on the suboptimality gap, so our design simultaneously learns about any gaps, between any two solutions, in excess of a given threshold δ. As a result, the computation of the design becomes exceedingly simple, requiring only estimation of the regression coefficients β. The design thus becomes much easier to implement than those found in [47] and related work, in which it is necessary to make an explicit guess of the optimal solution, creating an additional source of possible error. Thus, our work offers a natural computational benchmark for this problem class, and can perform well under limited sampling budgets.

Chapter 4: Conclusion and future work

4.1 Conclusion

In this dissertation, we focus on both theoretical and applied research of stochastic modeling and optimization. In Chapter 2, we characterize the long-run average queue lengths and choice probabilities for both express and regular service, and then study the dependence of these quantities on the entry fee, which drives the behaviour of various objectives related to revenue and social welfare using an $M/M/\bar{q}$ queueing model. We also include customer choice in our paper, which most of the existing literature does not have.

We note that these findings are obtained in a very general setting that encompasses many possible disutility functions and random choice models. If one makes additional assumptions, it is possible to obtain even more detailed characterizations – for example, under the MNL model, we derive the equilibrium queue lengths in closed form. However, the general setting also applies to, e.g., the exponomial choice model, and all of our general results continue to hold in that context.

In Chapter 3, we derive a new, large deviations theoretic optimality criterion for linear regression, and propose a new design that optimizes this criterion. More specifically, we have derived a static design that optimizes the convergence rate of the probability of error. Unlike all of the existing work on large deviations-based designs, we do not discretize the design space so that our method can be applied for continuous variables. Also, our optimal design is much easier to implement because we do not need to make an explicit guess of the optimal solution. Compared to most of the existing literature on sequential implementation, which first has to guess which alternative is the best and if this guess is incorrect, the estimated proportions will be very inaccurate. For this reason, our approach has considerable practical utility (also illustrated in a numerical example) and can serve as a natural benchmark for continuous optimal design in linear regression.

4.2 Suggestions for future research

In Chapter 2, although we obtain a complete characterization of the social welfare optimization problem, we were not able to characterize the general structure of the revenue curve in detail. A possible solution might be to consider self-adjusting pricing schemes that learn the optimal price dynamically. Instead of fixing the entry fee c, we are working on a dynamic pricing policy, and we believe it may help us to characterize the optimal revenue.

In Chapter 3, although our LD-optimal algorithm has a nice performance, we notice that it doesn't differ from D-optimal design too much for large sample sizes. Thus, we are currently exploring some cases where our algorithm may be better than other algorithms. For example, from Figure 3.3, it seems that our algorithm has a big advantage for small sample sizes. Also while running multiple numerical experiments, we found that if the set of vectors $x_i^\star$ are close to each other, then our algorithm also has a big advantage.

Bibliography

[1] Constantinos Maglaras. Revenue management for a multiclass single-server queue via a fluid model analysis. *Operations Research*, 54(5):914–932, 2006.

[2] Yu Zhang, Jinting Wang, and Fang Wang. Equilibrium pricing strategies in retrial queueing systems with complementary services. *Applied Mathematical Modelling*, 40(11-12):5775–5792, 2016.

[3] R Hassin. Rational queueing. chapman and hall/crc.. 2016.

[4] Jinting Wang, Shiliang Cui, and Zhongbin Wang. Equilibrium strategies in m/m/1 priority queues with balking. *Production and Operations Management*, 28(1):43–62, 2019.

[5] Philipp Afeche and J Michael Pavlin. Optimal price/lead-time menus for queues with customer choice: Segmentation, pooling, and strategic delay. *Management Science*, 62(8):2412–2436, 2016.

[6] Ping Cao, Yaolei Wang, and Jingui Xie. Priority service pricing with heterogeneous customers: Impact of delay cost distribution. *Production and Operations Management*, 28(11):2854–2876, 2019.

[7] Philipp Afèche. Delay performance in stochastic processing networks with priority service. *Operations Research Letters*, 31(5):390–400, 2003.

[8] Philipp Afeche and Vahid Sarhangian. Rational abandonment from priority queues: Equilibrium strategy and pricing implications. *Columbia Business School Research Paper*, (15-93), 2015.

[9] Terry A Taylor. On-demand service platforms. *Manufacturing & Service Operations Management*, 20(4):704–720, 2018.

[10] Philipp Afeche, Opher Baron, Joseph Milner, and Ricky Roet-Green. Pricing and prioritizing time-sensitive customers with heterogeneous demand rates. *Operations Research*, 67(4):1184–1208, 2019.

[11] Jianfu Wang, Opher Baron, and Alan Scheller-Wolf. M/m/c queue with two priority classes. *Operations Research*, 63(3):733–749, 2015.

[12] Philipp Afèche and Barış Ata. Bayesian dynamic pricing in queueing systems with unknown delay cost characteristics. *Manufacturing & Service Operations Management*, 15(2):292–304, 2013.

[13] Ming Hu, Yang Li, and Jianfu Wang. Efficient ignorance: Information heterogeneity in a queue. *Management Science*, 64(6):2650–2671, 2018.

[14] Lillian J Ratliff, Chase Dowling, Eric Mazumdar, and Baosen Zhang. To observe or not to observe: Queuing game framework for urban parking. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 5286–5291. IEEE, 2016.

[15] Refael Hassin and Ricky Roet-Green. The impact of inspection cost on equilibrium, revenue, and social welfare in a single-server queue. *Operations Research*, 65(3):804–820, 2017.

[16] Costis Maglaras, John Yao, and Assaf Zeevi. Optimal price and delay differentiation in large-scale queueing systems. *Management Science*, 64(5):2427–2444, 2018.

[17] Luyi Yang, Laurens G Debo, and Varun Gupta. Search among queues under quality differentiation. *Management Science*, 65(8):3605–3623, 2019.

[18] Rouba Ibrahim. Managing queueing systems where capacity is random and customers are impatient. *Production and Operations Management*, 27(2):234–250, 2018.

[19] Srinagesh Gavirneni and Vidyadhar G Kulkarni. Self-selecting priority queues with burr distributed waiting costs. *Production and Operations Management*, 25(6):979–992, 2016.

[20] Shiliang Cui, Zhongbin Wang, and Luyi Yang. The economics of line-sitting. *Management Science*, 66(1):227–242, 2020.

[21] Mor Armony and Constantinos Maglaras. On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Operations Research*, 52(2):271–292, 2004.

[22] Aydın Alptekinoğlu and John H Semple. The exponomial choice model: A new alternative for assortment and price optimization. *Operations Research*, 64(1):79–93, 2016.

[23] Pengfei Guo and Zhe George Zhang. Strategic queueing behavior and its impact on system performance in service systems with the congestion-based staffing policy. *Manufacturing & Service Operations Management*, 15(1):118–131, 2013.

[24] Fikret Caner Gocmen. *Infrastructure Scaling and Pricing.* PhD thesis, Columbia University, 2014.

[25] Caner Göçmen, Robert Phillips, and Garrett van Ryzin. Revenue maximizing dynamic tolls for managed lanes: A simulation study, 2015.

[26] Jamol Pender, Richard H Rand, and Elizabeth Wesson. Queues with choice via delay differential equations. *International Journal of Bifurcation and Chaos*, 27(04):1730016, 2017.

[27] Jamol Pender, Richard H Rand, and Elizabeth Wesson. An analysis of queues with delayed information and time-varying arrival rates. *Nonlinear Dynamics*, 91(4):2411–2427, 2018.

[28] Jamol Pender, Richard Rand, and Elizabeth Wesson. A stochastic analysis of queues with customer choice and delayed information. *Mathematics of Operations Research*, 45(3):1104–1126, 2020.

[29] Dimitris Bertsimas, Allison O'Hair, Stephen Relyea, and John Silberholz. An analytics approach to designing combination chemotherapy regimens for cancer. *Management Science*, 62(5):1511–1531, 2016.

[30] Holger Dette. Designing experiments with respect to 'standardized'optimality criteria. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(1):97–110, 1997.

[31] Chi Song Wong and Joseph C Masaro. A-optimal design matrices. *Discrete mathematics*, 50:295–318, 1984.

[32] Toby J Mitchell. An algorithm for the construction of "d-optimal" experimental designs. *Technometrics*, 42(1):48–54, 2000.

[33] Victoria Pokhilko, Qiong Zhang, Lulu Kang, and P Mays D'arcy. D-optimal design for network a/b testing. *Journal of Statistical Theory and Practice*, 13(4):1–23, 2019.

[34] Myrta Rodriguez, Bradley Jones, Connie M Borror, and Douglas C Montgomery. Generating and assessing exact g-optimal designs. *Journal of quality technology*, 42(1):3–20, 2010.

[35] Diana M Negoescu, Peter I Frazier, and Warren B Powell. The knowledge-gradient algorithm for sequencing experiments in drug discovery. *INFORMS Journal on Computing*, 23(3):346–363, 2011.

[36] Haihui Shen, L Jeff Hong, and Xiaowei Zhang. Ranking and selection with covariates. In *2017 Winter Simulation Conference (WSC)*, pages 2137–2148. IEEE, 2017.

[37] Siyang Gao, Jianzhong Du, and Chun-Hung Chen. Selecting the optimal system design under covariates. In *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, pages 547–552. IEEE, 2019.

[38] Bin Han, Ilya O Ryzhov, and Boris Defourny. Optimal learning in linear regression with combinatorial feature selection. *INFORMS Journal on Computing*, 28(4):721–735, 2016.

[39] Mark W Brantley, Loo Hay Lee, Chun-Hung Chen, and Argon Chen. Efficient simulation budget allocation with regression. *IIE Transactions*, 45(3):291–308, 2013.

[40] Mark W Brantley, Loo Hay Lee, Chun-Hung Chen, and Jie Xu. An efficient simulation budget allocation method incorporating regression for partitioned domains. *Automatica*, 50(5):1391–1400, 2014.

[41] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

[42] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *NIPS*, volume 11, pages 2312–2320, 2011.

[43] Peter Glynn and Sandeep Juneja. A large deviations perspective on ordinal optimization. In *Proceedings of the 2004 Winter Simulation Conference, 2004.*, volume 1. IEEE, 2004.

[44] Chun-Hung Chen and Loo Hay Lee. *Stochastic simulation optimization: an optimal computing budget allocation*, volume 1. World scientific, 2011.

[45] Chun-Hung Chen, Stephen E Chick, Loo Hay Lee, and Nugroho A Pujowidianto. Ranking and selection: efficient simulation budget allocation. *Handbook of Simulation Optimization*, pages 45–80, 2015.

[46] Chun-Hung Chen, Jianwu Lin, Enver Yücesan, and Stephen E Chick. Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems*, 10(3):251–270, 2000.

[47] Raghu Pasupathy, Susan R Hunter, Nugroho A Pujowidianto, Loo Hay Lee, and Chun-Hung Chen. Stochastically constrained ranking and selection via score. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 25(1):1–26, 2014.

[48] Guy Feldman, Susan R Hunter, and Raghu Pasupathy. Multi-objective simulation optimization on finite sets: Optimal allocation via scalarization. In *2015 Winter Simulation Conference (WSC)*, pages 3610–3621. IEEE, 2015.

[49] Siyang Gao, Weiwei Chen, and Leyuan Shi. A new budget allocation framework for the expected opportunity cost. *Operations Research*, 65(3):787–803, 2017.

[50] Ye Chen and Ilya O Ryzhov. Balancing optimal large deviations in ranking and selection. In *2019 Winter Simulation Conference (WSC)*, pages 3368–3379. IEEE, 2019.

[51] Ye Chen and Ilya O Ryzhov. Complete expected improvement converges to an optimal budget allocation. *Advances in Applied Probability*, 51(1):209–235, 2019.

[52] Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *arXiv preprint arXiv:1409.6110*, 2014.

[53] Tanner Fiez, Lalit Jain, Kevin Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. *arXiv preprint arXiv:1906.08399*, 2019.

[54] Omar Besbes and Costis Maglaras. Revenue optimization for a make-to-order queue in an uncertain market environment. *Operations Research*, 57(6):1438–1450, 2009.

[55] Angelos Aveklouris, Maria Vlasiou, and Bert Zwart. Bounds and limit theorems for a layered queueing model in electric vehicle charging. *Queueing Systems*, 93(1):83–137, 2019.

[56] Thomas G Kurtz. Strong approximation theorems for density dependent markov chains. *Stochastic Processes and their Applications*, 6(3):223–240, 1978.

[57] EL Ince. Ordinary differential equations dover publ. *Inc., New York*, 1956.

[58] Rüdiger Seydel. *Practical bifurcation and stability analysis*, volume 5. Springer Science & Business Media, 2009.

[59] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[60] A Dembod, O Zeltouni, and K Fleischmann. Large deviations techniques and applications. *Jahresbericht der Deutschen Mathematiker Vereinigung*, 98(3):18–18, 1996.

[61] Tze Leung Lai, Ching Zong Wei, et al. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Annals of Statistics*, 10(1):154–166, 1982.

[62] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[63] Haihao Lu, Robert M Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

[64] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[65] Huizhu Wang and Seong-Hee Kim. Reducing the conservativeness of fully sequential indifference-zone procedures. *IEEE Transactions on Automatic Control*, 58(6):1613–1619, 2012.