

Spatial Service Design for Public Services

by

Sina Hoveida

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Masters of Applied Sciences
in
Management Sciences

Waterloo, Ontario, Canada, 2024

© Sina Hoveida 2024

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

This study aims to design a service network in an urban area with continuous demand to maximize social welfare. We assume that customers are sensitive to travel and wait times, seeking service at a location that maximizes their utility. First, we demonstrate that the problem of spatial service design and pricing for an urban area with strategic customers is equivalent to a bilevel design problem, where customers can be explicitly assigned to service locations. This allows for the explicit assignment of customers to service locations and guarantees the existence of a service fee mechanism that satisfies the optimal assignment. We show that the urban area can be divided into a set of connected and disjoint regions such that, within each region, all customers seek service from the same location. We then derive the relationship between the optimal demand rate served at each service location and the optimal capacity (service rate) of the location. Our findings indicate that the optimal service capacity at each location depends on the capacity cost. For instance, the square root rule is optimal for linear capacity costs, but this does not hold for nonlinear costs. Furthermore, we characterize the relationship between the optimal service fee and observe that when the service capacity is fixed and cannot be adjusted, the optimal service fee is higher at locations with higher optimal demand rates. However, when capacity can be optimally allocated, the optimal service fee depends on the capacity cost. Specifically, if the cost is linear, optimal capacity allocation leads to optimal social welfare, resulting in all locations charging the same service fee. Conversely, if the capacity cost is nonlinear, the optimal fee decreases with the optimal demand rate when the service cost is strictly convex and increases when the cost is strictly concave.

Acknowledgements

I want to express my deepest gratitude to my parents for their unwavering support throughout this journey. Their encouragement and belief in me have been a constant source of strength. I am profoundly thankful to my wife for her endless support, patience, and creative ideas. Her presence has been a cornerstone of my success, and I could not have done this without her. I owe a special thanks to my supervisor, Professor Hossein Abouee-Mehrizi, for constantly pushing me to reach my full potential and guiding me to improve myself every step of the way. His mentorship has been invaluable. Finally, I want to extend my heartfelt thanks to my dear friends Kasra Fallah, Mina Torabi, and Hamid Arzani for their help and support throughout the development of this thesis and my master's journey. Their assistance and friendship have made this journey manageable and enjoyable.

Dedication

To my grandfather, Mohammad Sadegh Farbod, whose patience and dedication to academics have always been an inspiration to me.

Table of Contents

1	Introduction	1
2	Related Literature	5
3	Problem Formulation	8
3.1	The Design Problem	11
4	Design Variables at Optimality	13
4.1	Optimal locations and demand mapping	13
4.2	Optimal Demand Rate for Given Service Capacity	18
4.3	Optimal Demand Rate and Service Capacity	20
5	Optimal Pricing	25
5.1	Optimal Service Fee for Given Service Capacity	25
5.2	Joint Optimization of Service Capacity and Price	28
6	Conclusions and Future Works	32
	References	34
	APPENDICES	39
A	Preliminaries and Notations	40

B	Algorithm Analysis	44
C	Proofs of Results	48
C.1	Existence of service fees (Proof of Theorem 3.1)	49
C.2	Proof of Results on Optimal Design Variables	57
C.3	Proof of Results on Optimal Pricing	79
C.4	Proof of Results on Algorithmic Analysis	84

Chapter 1

Introduction

The service network in the public sector plays a pivotal role in delivering services and fulfilling the needs of citizens and communities. The importance of a well-functioning service network in the public sector cannot be overstated, as it directly impacts the quality of life, economic development, and overall well-being of a society. The location of service facilities directly impacts travel time, accessibility, and the primary service areas they serve. On the other hand, the service capacity is a crucial factor that determines the delay in receiving the service and the wait time that customers need to endure before receiving the service. These factors, in turn, influence customer usage, satisfaction levels, and overall demand for services. In particular, the distance one should travel and the duration of service delay, commonly referred to as wait time, are widely acknowledged as significant factors influencing demand in the public service sector. Within the healthcare sector, for example, patients take into account both travel time and wait time when planning visits to hospitals [42] or when considering services such as elective surgeries [39]. In politics, extensive wait times have been shown to negatively impact voter turnout [5], while research indicates that greater distances to polling locations can also deter participation ([26]; [13]).

Given the substantial impact of travel and wait time on customer behavior, service providers should decide how their service network is designed to maximize social welfare while ensuring access for all customers. This involves determining the location of service stations and the service capacity to serve customers effectively and maximize social welfare. In this problem, the trade-off between travel distance and wait time is pivotal. Centralizing capacity at a single station diminishes average wait times but extends customers' travel distances. Conversely, distributing multiple stations reduces travel time but might result in longer wait times. For instance, stations in densely populated locales, like downtown areas, typically offer shorter travel times for residents. Nevertheless, wait times tend to

be prolonged due to the higher concentration of customers in these areas. Consequently, the service provider should strategically locate service stations to strike a balance between travel and wait times.

While some papers in the literature focus on service network design, they mainly ignore the spatial characteristics of the problem ([8], [9]). In many urban applications, however, thousands of potential customers are distributed over an urban area seeking service where the density of population in some areas might be higher than others. Recently, [14] study social welfare in a spatial service network with a set of service stations where their locations are fixed. They assume that customers are strategic and seek service at a location that minimizes their total wait and travel costs. For a given capacity at service locations, they characterize customer behavior and show that the decisions of self-interested customers may not achieve socially optimal outcomes due to the price of anarchy. Then, they demonstrate that if the service provider can reallocate service capacity between the service locations, it is possible to achieve optimal social welfare when customers act strategically.

Allocating optimal capacity to service locations can eliminate inefficiencies (i.e., the price of anarchy) when capacity costs are linear. However, this approach becomes ineffective when costs are nonlinear, which is common in many services. Marginal costs of service capacity can either increase or decrease depending on the nature of the service. Specifically, marginal costs increase when efforts to raise service rates involve significant additional expenses, such as hiring more staff, investing in advanced technology, or extending operational hours. These efforts can lead to diminishing returns due to resource constraints, coordination challenges, and increased errors, further escalating costs. Conversely, marginal costs decrease in situations where economies of scale, learning effects, technological advancements, and bulk purchasing come into play. Moreover, the strategy of capacity allocation to achieve optimal social welfare fails when service capacity cannot be entirely reallocated among locations.

Therefore, the goal of this paper is to design an incentive mechanism that eliminates system inefficiencies (i.e., the price of anarchy) in a general spatial service network with strategic customers, thereby maximizing social welfare. To achieve this, we focus on an urban area with a population distributed throughout the area. We assume the distribution of customer demand is known over both time and location. Customers, sensitive to travel and wait times, seek service from locations that maximize their utility. The service provider aims to strategically place multiple service locations across the urban area to maximize social welfare, considering service capacity costs and service fees.

Using duality between the space of continuous functions and positive measures, we show that this problem is equivalent to a bilevel problem of minimizing the total cost of

the system, including customer disutility and service capacity costs, where customers can explicitly be assigned to service locations. This result demonstrates that the service fee mechanism can motivate strategic customers to seek service at a location that maximizes social welfare. We further show that the travel distance of customers to service locations is equivalent to the Wasserstein distance between the distribution of the customers' location and their arrival to service locations. These results help us to capture the effect of customers' spatial distributions and their sensitivity to waiting and travel costs on their arrivals to service locations, the optimal capacity of each location, and service fees.

We show that the explicit assignment of customers to locations is equivalent to dividing the urban area into simply connected and disjoint regions, where all customers within each region are assigned to the same service location. When the travel cost is linear, the regions can be characterized as hyperbola, while they are polytope when the travel cost is quadratic. Knowing the existence of regions can significantly simplify decisions regarding the optimal location of service stations. For example, when the cost of travel is quadratic in distance, the optimal location for each region is the *center of mass* of the region. This result, coupled with the relationship between service fees and customer assignments, highlights the connection between optimal pricing and the optimal location of service stations. To the best of our knowledge, such a geometric insight has not been previously discussed in the literature.

Furthermore, we investigate the relationship between the demand served by each service location and the optimal service capacity for that location. We observe that to hedge the system against uncertainties in arrival and service rates, the capacity should be set to the average demand rate plus an additional term that depends on the service cost function. For instance, if the capacity cost at each location is linear, the optimal capacity follows the square-root rule, meaning the additional term in the optimal service rate is the square root of the arrival rate. Conversely, when the service capacity cost is quadratic, the additional term in the optimal service rate is roughly proportional to the cube root of the optimal demand rate, as opposed to the square root in the linear case.

Building on these results, we examine the effect of waiting and travel costs in two settings: one where service capacity is given and another where it is a decision variable. We demonstrate that, in the former scenario, as waiting costs increase, arrival rates converge to a threshold, resulting in almost identical wait times across all service locations. In the latter scenario, the number of service locations depends on the capacity cost function. In particular, if the total waiting cost at the optimal service rate is concave, the optimal number of locations is reduced as the wait cost increases to allow the system to benefit from capacity pooling. However, if this cost is convex, the optimal number of locations remains constant as the wait cost increases while the demand rate served by each location

becomes closer to each other.

Additionally, we investigate how service fees at different locations should be designed to ensure that strategic customers follow the optimal demand assignment. We characterize the optimal service fee and show that, for a given service capacity, as the unit waiting cost increases, the service fees at locations with the same capacity become more similar. In this scenario, the capacity allocation strategy alone cannot be applied, making it necessary to use the service fee mechanism to motivate strategic customers to follow the optimal social behavior. When the service capacity at each location can be determined by the service provider, the optimal service fee depends on the service capacity cost function. For linear costs, a uniform pricing strategy across service locations guarantees optimal social outcomes, meaning the optimal capacity allocation strategy alone ensures the optimal social behavior of customers. However, when capacity costs are nonlinear, service locations need to adjust their pricing to achieve optimal social welfare. We also develop an algorithm to solve this problem efficiently.

The rest of the paper is organized as follows. We review the relevant literature in Section 2. Then, we define the problem in Section 3. We investigate the relation between the optimal demand mapping and service capacity in Section 4, and characterize the optimal pricing strategy in Section 5. We then develop an algorithm to solve the problem in Section B, and conclude the paper in Section 6. All proofs appear in the appendixes.

Chapter 2

Related Literature

Service design networks in the public sector have been studied in the literature since the seminal work of [43], which proposes a framework for establishing service locations in urban areas. Since then, there have been many papers in the literature that focus on determining optimal locations for service facilities where congestion at the service stations is ignored. For a recent comprehensive review of the related literature, readers are referred to [3].

Unlike the above literature, there is a stream of papers that consider congestion at service locations in designing public service networks. For instance, [15] consider a social optimal location problem where the goal is to minimize the total disutility of customers and facility design costs. [49] consider designing a service network so that customer participation in receiving the service is maximized. They assume customers are sensitive to both travel and wait times such that the demand received by each service location is linearly decreasing in the total travel and wait time. [2] also consider designing a service network for the public sector so that the demand served by the network is maximized. [1] proposes an algorithm to design a service network to minimize the fixed establishment and service capacity costs. They assume that customers choose a location where their disutility, measured as the weighted sum of travel and wait time, is minimized. For the relevant literature, readers are directed to the comprehensive review provided by [11] and [3].

While the above-mentioned papers mainly focus on discrete settings, there are also papers in the literature that study facility design and customer equilibrium behavior in two-dimensional settings to model urban applications more accurately. For instance, [16] consider a location-allocation problem in a convex polygon region with uniform demand. They develop an algorithm to locate multiple incapacitated facilities to minimize the total customer travel cost. [48] consider a two-dimensional setting where a set of facilities is

located. Assuming that customers are sensitive to travel and market externality costs, they determine the market share of each service location. [35] consider a facility location problem in a two-dimensional setting where customers are sensitive to travel time and traffic congestion, which is not modeled implicitly. They propose an approximate algorithm to locate multiple facilities so that social welfare is maximized. [8] study a facility location problem in a convex polygon region with uniform demand where customers seek service from the closest service location. Assuming that there are already some locations to serve customers, their goal is to establish a new location so that customers' total travel distance is minimized. [9] present polynomial algorithms to locate a new facility in an urban region where a set of locations already exists, assuming that the demand is uniformly distributed over a convex polygon.

This work is also related to papers that consider service fees as a tool to influence customers' behavior to improve social welfare, which has long been discussed in the literature. The seminal work of [34] was the first to explore customer behavior in a Markovian queueing system, where individuals evaluate the length of the queue before deciding whether to join, and consider a service fee mechanism to encourage customers towards the optimal social setting. In this model, customers act as utility-maximizing decision-makers who are sensitive to wait times. [22] extend the previous work to a setting where customers cannot observe the queue length and base their decision to join the queue on the expected wait time. Readers are referred to [27] for a comprehensive review of papers in this literature.

Empirical evidence indicates that people seeking a service not only prioritize service delay (waiting time) but also factor in the distance they must travel to access the service (see, for example, [4], [39], [13]). Consequently, a stream of papers in the literature investigates pricing and capacity decisions within queueing systems, considering strategic customers who are sensitive to travel and wait time. For instance, [20] explores the optimal placement of a single service location on a line to maximize company profit by determining service capacity and price. In a more recent study, [37] assumes the sensitivity of the customer (patient) to both travel and wait time, deciding the search for services based on a utility function inversely related to travel and waiting costs. Their objective is to determine service capacity and price, assuming service quality increases with service time. [28] postulate strategic customer behavior, considering both travel and wait time sensitivity, with customers only traveling if their expected utility is nonnegative. They assess the impact of queue-length information on customer behavior and system performance metrics. [46] also consider a service system with delays-sensitive customers evaluating travel and wait times in their decision-making process. They propose a distance-based priority that assigns a higher priority to customers traveling a higher distance to enhance system throughput and social welfare.

The previously mentioned papers assume that the system capacity is exogenous and focus on evaluating system performance metrics such as throughput, revenue, and social welfare. A subsequent body of literature deviates from the exogenous capacity assumption and addresses the simultaneous optimization of price and service capacity. For example, [33] studies an unobservable queueing system, assuming that the service provider can regulate both the service price and capacity. In a related paper, [17] investigate pricing and capacity decisions within a single-server Markovian queueing environment with linear capacity costs and a homogeneous customer base. [7] extend this line of work by considering dynamic control of both price and service rate, offering insight into optimal pricing and capacity adjustments in a Markovian-observable queueing system with delay-sensitive customers.

In a recent paper, [14] study social welfare within a spatial service network with strategic customers seeking services at locations that minimize their total waiting and travel costs. They consider a set of service stations with fixed locations, where customers choose the station that minimizes their disutility. They first demonstrate the existence of equilibrium when service rates are fixed. Then, they show that, for a fixed total capacity, optimal capacity allocation to service stations could lead to optimal social welfare. Our work differs from theirs in several ways. First, our goal is to design an incentive mechanism to eliminate inefficiencies created by the price of anarchy in a general spatial service network, where the cost of capacity is nonlinear or the available capacity cannot be completely reallocated. Recognizing that capacity allocation alone cannot achieve this goal, we design a service fee mechanism to ensure strategic customers seek service at locations that maximize social welfare. Second, besides optimizing service capacity, we aim to optimize the locations of service stations, providing unique insights into location analysis. Third, we examine the impact of customers' sensitivity to travel and wait times on the optimal service locations and their capacity.

Chapter 3

Problem Formulation

We consider an urban district area denoted by $X \subset \mathbb{R}^2$, where the demand for a particular service is proportional to the population of the district. We denote the population in the urban area X by $\hat{P}(X)$, where X is a compact set with respect to the Euclidean metric, and consider ω to be the demand per individual per unit of time. Consequently, we can express the demand rate in the urban area X as,

$$P(x) = \omega \hat{P}(x), \forall x \in X. \quad (3.1)$$

Given that the total demand is finite, we may normalize it to one. By adopting the Borel sigma-algebra over X with the previously defined metric, the Radon-Nikodym theorem guarantees the existence of a density function such that the measure of any subset of X is equivalent to the integral of the set's indicator function multiplied by P . Let $\mathbb{P}$ denote the measure defined by this probability distribution. We consider $(X, \mathcal{B}_X, \mathbb{P})$ to be the probability space associated with the demand rate. By construction, the cumulative distribution of $\mathbb{P}$ is strictly increasing and continuous; that is, there are no *atoms* in $\mathbb{P}$. Furthermore, we assume that $\mathbb{P}$ ensures $\mathbb{E}_{\mathbb{P}}[\|x\|_p^p] < +\infty$.

The goal is to establish a finite number of service locations to serve customers in the area, where the maximum number of locations is denoted by N . Each service location has limited capacity such that the station located at y_i serves customers at the rate μ_i per unit of time. We assume that the service duration is exponentially distributed and customers arrive at location y_i with an arrival rate λ_i following a Poisson process, captured by the probability space $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, \mathbb{T}_i)$. To summarize the notation, we let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_N)$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_N)$, and $Y = (y_1, y_2, \dots, y_N)$.

To receive the service, customers should travel to a service location, incurring a travel

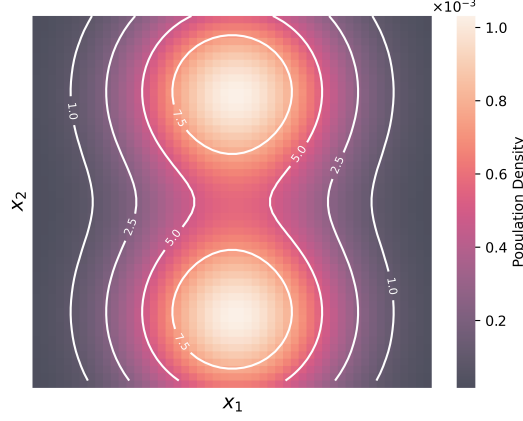


Figure 3.1: An urban area with continuous demand.

cost. Let $c(x, y_i)$ denote the travel cost for an individual from location x to service location i . Then, the total travel cost can be expressed as:

$$C_T = \mathbb{E}_{\mathbb{P}}[c(x, y_{i^*(x)})]. \quad (3.2)$$

To increase the tractability of the problem, we assume that $c(x, y) = \kappa \|x - y\|_p^p$ for $p \in \mathbb{N}$ where $\|\cdot\|_p$ is the l_p norm in $\mathbb{R}^2$. Note that by the equivalence of norms in finite dimensions, one can easily generalize $\|\cdot\|_p^p$ to functions in the form of $\|A(x - y)\|_p^p$ for a suitable choice of A ([40]). For example, if $p = 2$ and A is a semi-positive definite function, $\|A(x - y)\|_2^2$ is a general quadratic cost function.

Once customers arrive at a service location, they may encounter congestion in the system, and thus, they need to wait till customers who arrived before them are served. Let $c_W(i)$ denote the waiting cost for customers seeking service at location i . To express $\mathbb{E}_{\mathbb{T}_i}[c_W(i)]$ explicitly, we assume that the queue length at each station is not completely visible to customers, and they do not leave the queue after joining. Then, the expected waiting cost at each service location can be expressed as

$$\mathbb{E}_{\mathbb{T}_i}[c_W(i)] = \frac{\beta}{\mu_i - \lambda_i} \quad \forall i \in [n],$$

where β denotes the waiting cost of customers per unit of time.

We denote the service fee (price) at service location i by p_i where $\mathbf{p} = (p_1, p_2, \dots, p_N)$

captures the vector of service fees. We also let R be the service reward. Thus, the expected utility of a customer located at x seeking service at location i can be expressed as,

$$U(x, i) = R - p_i - \mathbb{E}_{\mathbb{T}_i}[c_W(i)] - \kappa \|x - y_i\|_p^p, \quad \forall x \in X. \quad (3.3)$$

We assume that customers travel to a service location that maximizes their expected utility to receive the service. Since there are a finite number of service locations, the function of arrival rates can be expressed by $\sum_{i \in [N]} \lambda_i \delta_{\{y_i\}}(x)$ where $\delta_{\{y_i\}}(x)$ is the Dirac measure with density at y_i . Using reasoning similar to that for the demand rate, we can conclude that there exists a density function and probability space, describing the arrival rate as such $(X, \mathcal{B}_X, \sum_{i \in [N]} \lambda_i \delta_{\{y_i\}}(x))$. We denote the set of all such probability measures by $\mathcal{P}(Y)$.

The goal is to design a service network over the urban area so that social welfare is maximized. We assume that service capacity at each location can be determined by the service provider, where the cost of service capacity could be nonlinear. This cost function, denoted by $h(\mu_i)$, is a strictly increasing, twice differentiable function with respect to the service rate such that its derivative is monotone. Let $G(\mathbf{p})$ denote the total revenue obtained by serving customers over the urban area X . Then, we have $G(\mathbf{p}) = \sum_{i \in [N]} \lambda_i p_i$. Moreover, let $i^*(x)$ be the utility-maximizing service location for customers at x . Then, social welfare can be expressed as follows:

$$W_{total} = G(\mathbf{p}) - C_S(\boldsymbol{\mu}) + \mathbb{E}_{\mathbb{P}}[U(x, i^*(x))], \quad (3.4)$$

where $C_S(\boldsymbol{\mu}) = \sum_{i \in [N]} h(\mu_i)$ captures the total service capacity cost. The goal is to maximize social welfare, defined in (3.4), assuming that the service provider determines the number of service stations, their locations, service fee, and capacity.

Based on their information, customers choose a service location that maximizes their expected utility. Thus, we can define the equilibrium of arrival rates in the sense that, given the service rates and fees, customers choose a service location so that there are no improvements in the choice of station for any group of customers with a measure greater than zero. Specifically, a set of arrival rates $\boldsymbol{\lambda}$, where λ_i captures the demand at location

i , is a set of equilibrium arrival rates if $\sum_{i \in [N]} \lambda_i = 1$ and,

$$\lambda_i = \mathbb{P} \left\{ x \in X \mid p_i + \kappa \|x - y_i\|_p^p + \beta \frac{1}{\mu_i - \lambda_i} \leq p_j + \kappa \|x - y_j\|_p^p + \beta \frac{1}{\mu_j - \lambda_j}, \forall i \neq j \right\}, \quad \forall i \in [N]. \quad (3.5)$$

Proposition 3.1 *Given any tuple of $(\mathbf{p}, \boldsymbol{\mu}, Y)$, there exists a unique set of equilibrium demand rate $\bar{\boldsymbol{\lambda}}$ satisfying (3.5).*

Given the above demand equilibrium for a set of service rates and fees, the goal is to build a service network that maximizes social welfare by eliminating inefficiencies resulting from customers' strategic behavior.

3.1 The Design Problem

In this section, using duality results over the set of positive measures, we reformulate the problem of maximizing social welfare with strategic customers, defined in (3.4), as a minimization problem such that customers are assigned optimally to service locations. To introduce the reformulation, we define the function π as the *demand mapping* to service locations. Specifically, let π denote the portion of the population from x that chooses to receive service at station y_i . Then, $\pi : X \times X \rightarrow [0, 1]$ can be defined as $\pi(x, y_i) = \mathbb{P}(\{x \in X \mid y_{i^*(x)} = y_i\})$ for all $x \in X$ and $\{y_i\}_{i \in [N]}$. Note that with this definition of π , the population assigned to the service station located at i should be equal to λ_i ; that is, for a given station located at y_i , we have $\int_X \pi(x, y_i) dx = \lambda_i$. In addition, any infinitesimal portion of the population in any location should be assigned to at least one service station, that is, $\int_Y \pi(x, y) dy = \mathbb{P}(X = x)$. A function with these properties turns out to be a probability measure on $\mathcal{B}_X \times \mathcal{B}_X$ and is a marginal distribution for $\mathbb{P}$ and $\sum_{i \in [N]} \lambda_i \delta_{\{y_i\}}(x)$.

We denote the collection of all such probability measures by Π . Using Kantorovich's duality, in the following theorem, we show that maximizing W_{total} is equivalent to a bilevel optimization that minimizes the total disutility costs of customers and the capacity costs of service locations.

Theorem 3.1 *There exists a set of service fees $\mathbf{p}^*$ such that the problem of maximizing*

social welfare is equivalent to minimizing the total cost of the system, i.e.,

$$\max_{\mathbf{p}, \boldsymbol{\mu}, Y} W_{total}(\mathbf{p}, \boldsymbol{\mu}, Y) = R - \min_{(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \{(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathbb{R}_+^{2N} \mid \|\boldsymbol{\lambda}\|_1 = 1, \lambda_i < \mu_i\}, Y} \sum_{i \in [N]} \left(h(\mu_i) + \beta \frac{\lambda_i}{\mu_i - \lambda_i} \right) + \kappa \min_{\pi \in \Pi} \mathbb{E}_\pi[\|x - y_i\|_p^p]. \quad (3.6)$$

In the problem given on the left side of (3.6), we consider customers who choose their service station so that their expected utility is maximized, which depends on the service fee p_i , the service location y_i , and its capacity μ_i . In the problem on the right-hand side of (3.6), where customers are assigned to service locations, all decision variables reappear except for the service fee. In fact, the pricing is captured by the demand mapping π and the arrival rate at each location λ_i . This means that for any given λ_i , μ_i , and y_i , there exists a set of service fees such that customers follow the optimal demand mapping, and thus, the optimal equilibrium arrival rates are achieved. Therefore, Theorem 3.1 states that, instead of maximizing social welfare, an equivalent problem that minimizes the total disutility costs of customers and the capacity cost of the service provider can be considered to design a service network such that social welfare is maximized.

Note that in the problem defined on the right-hand side of (3.6), we explicitly optimize the value of equilibrium arrival rates; thus, one could consider the arrival rates to be the decision variables that should be determined and the service fees to be the corresponding equilibrium values, similar to [24]. Eventually, the choice of referring to the price or arrival rate as equilibrium depends on considering the right- or left-hand side of (3.6) as the primal problem. To simplify the discussion, henceforth, we shall refer to the equilibrium arrival rates at the optimality of the service fees as the optimal arrival rates.

To summarize, we refer to the problem on the right-hand side of (3.6) (also provided below) as *Design Problem* and call the decision variables involved in this problem, $(\boldsymbol{\lambda}, \boldsymbol{\mu}, Y)$, as the *Design Variables*:

$$\min_{(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \{(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathbb{R}_+^{2N} \mid \|\boldsymbol{\lambda}\|_1 = 1, \lambda_i < \mu_i\}, Y} \left(\sum_{i \in [N]} \left(h(\mu_i) + \beta \frac{\lambda_i}{\mu_i - \lambda_i} \right) + \kappa \min_{\pi \in \Pi} \mathbb{E}_\pi[\|x - y_i\|_p^p] \right). \quad (3.7)$$

Note that the travel cost in the design problem, $\min_{\pi \in \Pi} \mathbb{E}_\pi[\|x - y_i\|_p^p]$, is known to be convex and is, in fact, a distance function between probability measures for any two given probability measures [44].

Chapter 4

Design Variables at Optimality

In this section, we analyze the design problem and derive some properties of the optimal service capacity, arrival rate, and service locations. Then, in Section 5, we show how the optimal service fees can be determined based on the optimal service capacity and demand rate so that all customers follow the optimal mapping of demand assignment to service locations.

The design problem given in (3.7) is related to the bilevel location-allocation problem ([36]). In general, it is not straightforward to tackle such problems, and deriving the optimality conditions for the set of decision variables involved in this problem is a challenging task ([8], [45]). Thus, our goal is to characterize some properties of the problem and then develop an efficient algorithm to solve it. In what follows, we start by describing the optimal demand mapping and location, given the optimal demand rates.

4.1 Optimal locations and demand mapping

From the definition of Problem (3.7), the optimal demand mapping is determined for any feasible set of design variables, including service rates, demand rates, and service locations. Moreover, since the solution to the problem of $\min_{\pi \in \Pi} \mathbb{E}_{\pi}[\|x - y_i\|_p^p]$ is unique for any set of feasible design variables, the total travel cost can be considered as a function of the design variables. Thus, we can, with a slight abuse of notation in comparison to (3.2), define the total travel cost for a given set of design variables (λ, μ, Y) as $C_T(\lambda, \mu, Y) : \mathbb{R}^{3N} \rightarrow \bar{\mathbb{R}}$ where $C_T(\lambda, \mu, Y) = \infty$ for infeasible values of (λ, μ, Y) . Using this function, the design

problem can be rewritten as,

$$\min_{\lambda, \mu, Y} C_T(\lambda, \mu, Y) + \sum_{i \in [N]} \left(h(\mu_i) + \beta \frac{\lambda_i}{\mu_i - \lambda_i} \right).$$

We now focus on the properties of $C_T(\lambda, \mu, Y)$ and demand mapping given a set of design variables. This will help us develop insights into customers' behaviors and serve as stepping stones in the study of optimal design variables. We first show that the urban area defined in the previous section can be divided into regions so that a service station located in each region serves the demand in that region.

Proposition 4.1 *Given a feasible tuple of design variables (λ, μ, Y) , there exists a set of disjoint subsets of X , $\{\chi_i\}_{i \in [N]}$, uniquely determined, such that*

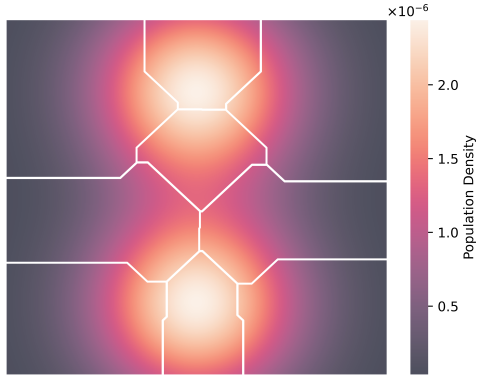
$$\chi_i = \{x \in X \mid \pi^*(x, y_i) > 0\}, \quad \forall i \in [N].$$

Furthermore, $\{\chi_i\}_{i \in [N]}$ are path connected, i.e., any point inside a given region is accessible by all other points within the same region using a continuous path.

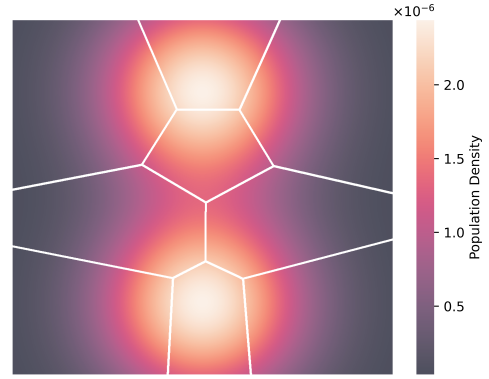
The above proposition indicates that there exists a set of disjoint regions, denoted by χ_i , so that customers in each region are only served by one service location. Proposition 4.1 also shows that the optimal solution can be represented as a set of simply connected regions, χ_i . That is, for any two points in the same region, a continuous path connecting the points exists. Note that Theorem 3.1 suggests that for any optimal mapping, a set of service fees exists such that customers choose the exact service location as assigned. Moreover, the customer utility, defined in (3.3), is continuous and smooth with respect to x for customers with the same $i^*(x)$. Thus, if a point is inside a region, there are always points in its neighborhood that are also inside the region. Using this property, we can always construct a continuous path between any two points in a region.

Corollary 4.1 *Let x_0 be in the interior of a region and $B_\epsilon(x_0)$ be the l_p -ball around x_0 . Then, there exists $\epsilon > 0$ such that the utility-maximizing service location is unique for $B_\epsilon(x_0)$. Furthermore, on the border of two adjacent regions, customers are indifferent towards the choice of service locations.*

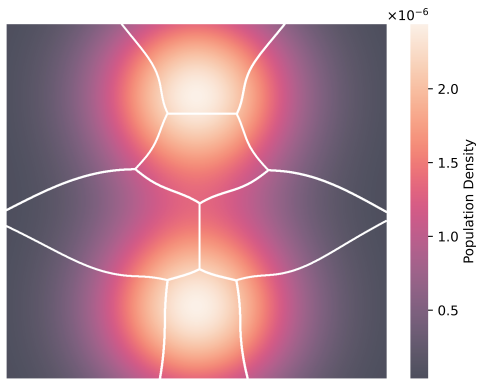
Corollary 4.1 shows that the choice of service location is unique for any group of customers with a positive demand. Furthermore, this corollary suggests that $\{\chi_i\}_{i \in [N]}$ can be interpreted using indifference curves. Precisely, an indifference curve represents demand



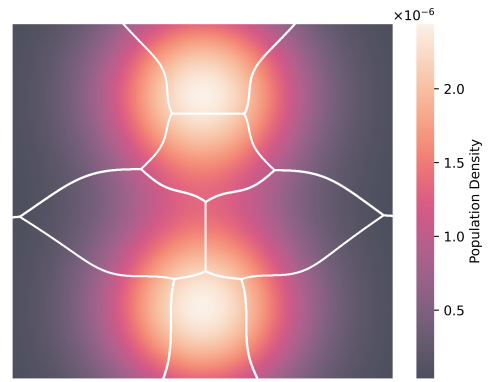
a) l^1 -Norm



b) l^2 -Norm



c) l^3 -Norm



d) l^4 -Norm

Figure 4.1: Optimal regions for different l^p -norm considered for the travel distance.

points in a region that are indifferent between receiving the service from service stations located on either side of the border.

We next characterize the geometric shape of regions characterized in Proposition 4.1. The following proposition shows that when l^2 -norm is considered for the distance between customers and service stations, the shape of the regions is polyhedral.

Proposition 4.2 *Let $\{\chi_i\}_{i \in [N]}$ be the set of optimal regions as defined in Proposition 4.1. Then, the geometric shape of $\{\chi_i\}_{i \in [N]}$ depends only on l^p -norm. Moreover,*

1. *for l^1 -norm, border of regions are piecewise linear functions with slope $\{-1, 0, 1, \infty\}$,*
2. *for l^2 -norm, $\{\chi_i\}_{i \in [N]}$ is a set of polytopes.*

Proposition 4.2 shows, as illustrated in Figure 4.1, that the shape of the regions depends on the l^p -norm considered for the distance between the demand and service locations. For instance, for l^1 -norm, regions' borders are always lines parallel with x axis, y axis, or lines with slope 1 and -1 . However, for l^2 -norm, they can be represented as a set of convex polytopes.

Given the optimal set of regions, the optimal location for the service station in each region can be determined locally, as stated in the following proposition. From henceforth, we will denote any $x \in X$; we denote the first component as x^1 and the second component as x^2 , and $y_{j,i}$ to be the j -th component of location i for $j = 1, 2$.

Proposition 4.3 *The optimal location of the service station in each region can be determined locally and satisfies,*

$$\begin{aligned} \int_{\{\chi_i | x_1 \geq y_{1,i}^*\}} (x_1 - y_{1,i}^*)^{p-1} d\mathbb{P} &= \int_{\{\chi_i | x_1 \leq y_{1,i}^*\}} (y_{1,i}^* - x_1)^{p-1} d\mathbb{P}, \\ \int_{\{\chi_i | x_2 \geq y_{2,i}^*\}} (x_2 - y_{2,i}^*)^{p-1} d\mathbb{P} &= \int_{\{\chi_i | x_2 \leq y_{2,i}^*\}} (y_{2,i}^* - x_2)^{p-1} d\mathbb{P}, \end{aligned}$$

In particular for $p = 2$, we have

$$y_i^* = \frac{1}{\lambda_i^*} \int_{\bar{\chi}_i} x d\mathbb{P}.$$

For $p = 2$, the above proposition states that the optimal location for each region is the center of mass of that region. For instance, for the case where $p = 2$ and $\mathbb{P}$ is uniform, the

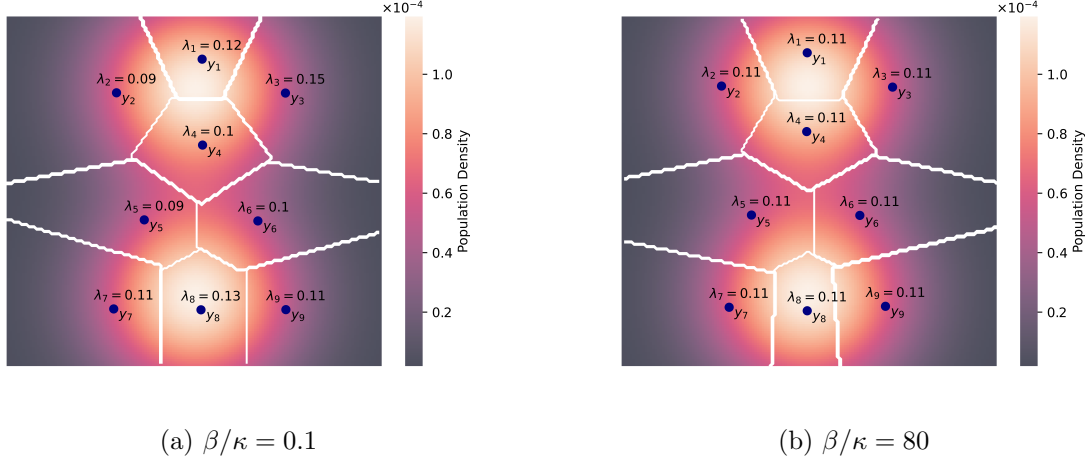


Figure 4.2: Optimal demand rates, regions, and service locations.

shape of the optimal regions is polyhedral, and it is optimal to locate the service station exactly at the center of the polyhedral. In the case where the distribution is uniform and $p \neq 2$, the conditions will melt down to

$$(x_1^+ - y_{i,1}^*)^p = (y_{i,1}^* - x_1^-)^p, \quad \forall x_1^+ \in \text{Bor}(\{x \in \chi_i | x_1 \geq y_{i,1}^*\}), x_1^- \in \text{Bor}(\{x \in \chi_i | x_1 \leq y_{i,1}^*\})$$

and

$$(x_2^+ - y_{i,2}^*)^p = (y_{i,2}^* - x_2^-)^p, \quad \forall x_2^+ \in \text{Bor}(\{x \in \chi_i | x_2 \geq y_{i,2}^*\}), x_2^- \in \text{Bor}(\{x \in \chi_i | x_2 \leq y_{i,2}^*\}),$$

where $\text{Bor}(A)$ is the border of set A . Intuitively, the first condition suggests that the optimal location should be on a curve, where the distance on the first axis is equal to either side of the region. The second condition clarifies the same for the second axis. Thus, the optimal location is at the intersection of these two curves. When $\mathbb{P}$ is not uniform, the distance could be seen as a *weighted* distance where the weight captures the population density.

4.2 Optimal Demand Rate for Given Service Capacity

In this section, we assume that service rates at different locations are given and focus on determining the optimal arrival rate at each station, λ_i^* . The main trade-off in determining the optimal λ_i is between the travel and the waiting costs. This means that the optimal arrival rate to each station depends on the relative sensitivity of customers to travel and waiting costs. In general, customers in dense areas probably have a lower travel cost and a higher waiting cost compared to customers in sparse regions. Thus, if customers are more sensitive to travel costs, the solution would favor a higher arrival rate in dense areas (inside the city) and a lower arrival rate in sparse areas (rural areas). On the other hand, if they are more sensitive to the waiting cost, the optimal arrival rate decreases for service stations located in dense areas.

To examine the impact of waiting and travel costs on the demand served by each service location, we investigate how the gap between the demand rates observed by different locations changes as the waiting cost varies. For a finite set $A \subseteq \mathbb{R}$, we define the function $\Psi(A) : A \rightarrow \mathbb{R}$, as

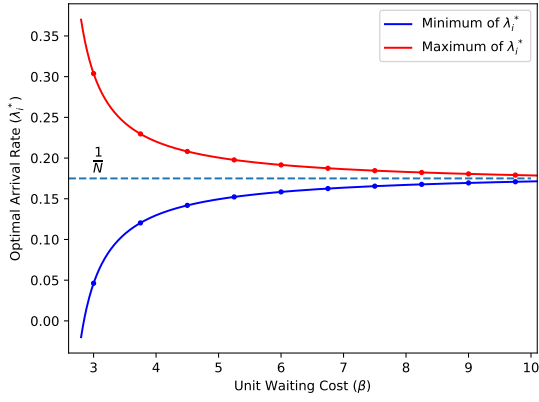
$$\Psi(A) = \frac{1}{|A|} \sum_{a \in A} \left(a - \frac{1}{|A|} \sum_{a \in A} a \right)^2, \quad (4.1)$$

to capture the variation in the set A . Here, $|A|$ is the cardinality of the set A . Note that $\Psi(\cdot)$ can be interpreted as the variance of a random variable $Z : A \rightarrow \mathbb{R}$ with a uniform distribution.

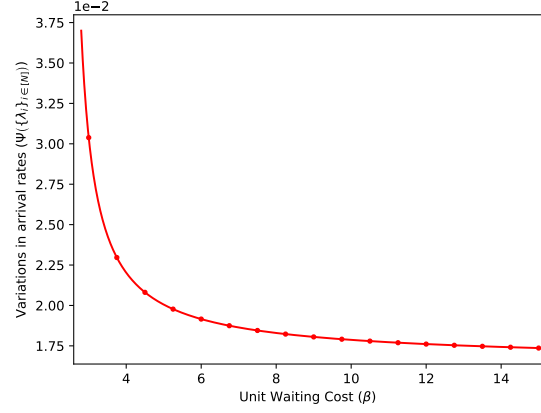
Proposition 4.4 *Suppose that the set of service rates $\{\mu_i\}_{i \in [N]}$ is given and let $\xi = \frac{\sum_{i \in [N]} \mu_i - 1}{\sum_{i \in [N]} \sqrt{\mu_i}}$. Then,*

1. *as β increases, the optimal demand rate for service locations with an arrival rate lower than $\mu_i - \xi\sqrt{\mu_i}$ increases and decreases for service locations with a demand rate higher than $\mu_i - \xi\sqrt{\mu_i}$.*
2. *there exists an $\bar{\beta} \in \mathbb{R}^+$ such that for all $\beta \geq \bar{\beta}$, the optimal demand rate served by location i is $\mu_i - \xi\sqrt{\mu_i}$.*
3. *for all $\beta \geq \bar{\beta}$,*

$$\Psi(\{\lambda_i^*\}_{i \in [N]}) = \frac{1}{N} (\|\boldsymbol{\mu}\|_2^2 + \xi^2 \|\boldsymbol{\mu}\|_{1/2} - (\xi^2 + \|\boldsymbol{\mu}\|_1) \|\boldsymbol{\mu}\|_1).$$



(a) Optimal demand rate λ_i^* as β changes



(b) Variation in λ_i^* ($\Psi(\{\lambda_i\}_{i \in [N]})$) as β changes

Figure 4.3: Impact of the waiting cost on the optimal demand rate.

Proposition 4.4 demonstrates that as the waiting cost increases, when the service capacity is fixed, the arrival rate to locations with high demand (relative to their service capacity) decreases. On the other hand, the arrival rate increases for locations with low demand, as illustrated in Figure 4.2. The reason is that as the waiting cost increases, the impact of the wait time on the objective function increases compared to the effect of the travel time. Thus, to reduce the wait time, the demand served by each location becomes relatively closer to their service capacity, and the wait times at different locations become closer to each other.

The second part of Proposition 4.4 shows that for any service location, there exists a threshold such that as β increases, the optimal arrival rate served by each location becomes closer to that threshold. This value indicates how the total demand should be distributed between service locations so that their expected wait time becomes the same. Thus, service locations can be divided into two groups: overcrowded and uncrowded. Overcrowded service locations serve more customers compared to the case where customers are only sensitive to the waiting cost. On the other hand, uncrowded service locations are probably not chosen frequently due to their inconvenient locations. As β increases, however, the impact of travel cost becomes less significant, and thus, the demand served by each location gets closer to the one where customers are only sensitive to the wait cost.

The third part of Proposition 4.4 shows that as β increases, the variation between λ_i^* becomes more similar to the variation between the service capacity. This result indicates that as the wait cost increases, its impact on the objective function increases, and thus, due

to the convex increasing behavior of the waiting cost, the wait times at different locations become closer to each other.

If the service rates at different stations are the same (i.e., $\mu_i = \mu_0$), the results provided in Proposition 4.4 can be simplified, as shown in Corollary 4.2. Specifically, this corollary shows that the threshold introduced in the above proposition for the demand rate served by each location is the same, i.e., $\lambda_i^* = 1/N$ if all service locations have the same capacity. Thus, the expected wait time for all locations will be the same if customers are very sensitive to the wait time.

Corollary 4.2 *Suppose $\mu_i = \mu_0$ for all $i \in [N]$ and let $\{\lambda_i^*\}_{i \in [N]}$ be the optimal arrival rates. Then, as β increases, the demand served by locations with arrival rates lower than $1/N$ increases, and the demand observed by locations with an arrival rate higher than $1/N$ decreases.*

Figure 4.3 illustrates the results provided in Corollary 4.2. Specifically, part (a) of the figure shows that as the waiting cost increases, the demand served by different locations gets closer to $1/N$. Moreover, part (b) of the figure shows that the variation between the demand served by different service locations decreases as the waiting cost increases. Hence, the gap is almost zero for large waiting costs.

4.3 Optimal Demand Rate and Service Capacity

In this section, we aim to extend the results of the previous section to a setting where both arrival and service rates are determined simultaneously. We first show the necessary optimality condition for the demand rate served by each service location and the capacity of that location.

Proposition 4.5 *Given arrival rates λ_i 's, the optimal service rate at service locations with positive demand can be obtained by,*

$$\frac{\partial h(\mu)}{\partial \mu} \Big|_{\mu=\mu_i^*} = \beta \frac{\lambda_i}{(\mu_i^* - \lambda_i)^2}. \quad (4.2)$$

Furthermore, for any given λ_i , μ_i^* is unique.

Proposition 4.5 provides a necessary optimality condition for the service rate at each location. Since the service capacity cost function $h(\mu_i)$ is smooth, the above proposition specifies the optimal service capacity at each location, given the demand rate at that location. Note that (4.2), in particular, clarifies the relation between the optimal demand rate and service capacity at any location with a positive demand. Due to uniqueness of μ_i^* for any given λ_i , we can define the function $\mu^*(\lambda_i) : (0, 1] \rightarrow \mathbb{R}$ to be the solution of (4.2) for any given λ_i . Then, the following corollary provides the optimal service capacity for three examples of $h(\mu_i)$.

Corollary 4.3 *The optimal service rate $\mu^*(\lambda_i)$, is an increasing function with respect to λ_i and β . Furthermore,*

1. *if $h(\mu_i) = \alpha\mu_i$, the optimal service capacity is*

$$\mu_i^*(\lambda_i) = \lambda_i + \sqrt{\frac{\beta\lambda_i}{\alpha}},$$

2. *if $h(\mu_i) = \log(\gamma_1\mu_i + \gamma_2)$, the optimal service capacity is*

$$\mu^*(\lambda_i) = \lambda_i \left(1 + \frac{\beta}{2\gamma_2}\right) + \sqrt{\frac{\beta}{\gamma_2}\lambda_i} \sqrt{\frac{1}{\gamma_1} + \lambda_i \left(1 + \frac{1}{2\beta\gamma_2}\right)},$$

3. *if $h(\mu_i) = \alpha\mu_i^2$, the optimal service capacity is*

$$\mu^*(\lambda_i) = \frac{2}{3}\lambda_i + \lambda_i^{1/3} \left(c_0 - \lambda_i^2 + c_1\sqrt{c_2 - \lambda_i^2}\right)^{1/3} + \lambda_i^{5/3} \left(c_0 - \lambda_i^2 + c_1\sqrt{c_2 - \lambda_i^2}\right)^{-1/3},$$

where

$$c_0 = \frac{27\beta}{4\alpha}, \quad c_1 = \sqrt{\frac{3^{3/4}}{2}\alpha\beta}, \quad c_2 = \left(\frac{3}{2\alpha}\right)^3\beta.$$

The above corollary shows that the optimal service capacity is increasing in both the demand rate and waiting cost, as illustrated in Figure 4.4. Moreover, it provides the relationship between the optimal service and demand rates for three types of service capacity cost functions, including linear, concave (logarithmic), and convex (quadratic) cost functions. These results show the superlinearity of the optimal service rate as a function of the demand. This indicates that to protect the system against uncertainty, the service capacity should be set to the average demand rate plus an additional term that depends

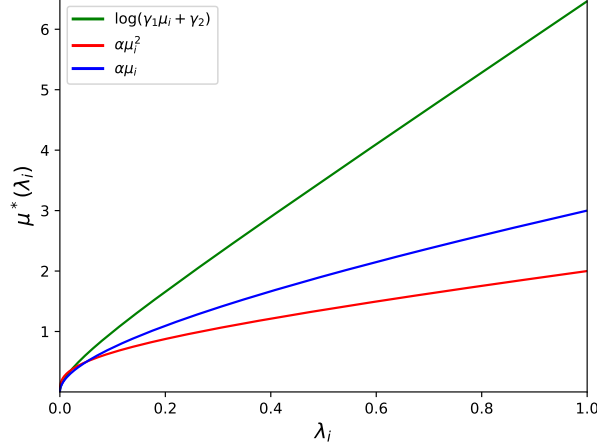


Figure 4.4: Relation between the optimal demand rate and service capacity for different $h(\mu_i)$.

on the demand and the service cost function. Specifically, the additional term appears as a square root when the service capacity cost $h(\mu_i)$ is linear. This is similar to the square root staffing rule suggested for other settings, such as call centers aiming to minimize their total waiting cost (see, e.g., [32]).

For the logarithmic cost function, the increase in the optimal service rate is the largest among the three cost functions because of the concavity of the function. Specifically, compared to the linear cost function, the optimal service rate increases faster with respect to λ_i^* . On the other hand, for the quadratic cost function, the optimal service rate is below the linear cost function, as illustrated in Figure 4.4. For this case, the additional term in the optimal service rate is roughly proportional to the cube root of the optimal demand rate compared to the square root in the linear case.

Next, we examine the impact of waiting and travel costs on the optimal arrival rate. Let $\mathcal{A} = \{\lambda_i^* | \lambda_i^* > 0\}$ denote the set of optimal arrival rates for different service locations. We denote the optimal number of service locations, the cardinality of $\mathcal{A}$, as N^* . In what follows, we focus on the impact of β on the optimal number of service locations and demand rate. It turns out that the curvature of the waiting cost at the optimal service rate has a significant impact on the demand rate served at each location.

Corollary 4.4 *The waiting cost at the optimal service rate at each location rate can be*

determined by

$$C_W^{\mu^*}(\lambda_i) = \sqrt{\beta \lambda_i \frac{\partial h(\mu)}{\partial \mu} \big|_{\mu=\mu^*(\lambda_i)}}. \quad (4.3)$$

Using $C_W^{\mu^*}(\lambda_i)$ provided in Corollary 4.4, we next characterize the impact of the waiting cost on the number of service locations.

Proposition 4.6 *If $C_W^{\mu^*}(\lambda_i)$ is concave, there exists a concave increasing function $W_1(\beta)$ and a convex increasing function $T_1(\beta)$ such that,*

1. *for $\kappa \leq W_1(\beta)$, it is optimal to locate one station and $\lambda_i^* = 1$,*
2. *for $\kappa \geq T_1(\beta)$, λ_i^* can be determined by a distribution clustering with a Wasserstein distance.*

Moreover, if $C_W^{\mu^*}(\lambda_i)$ is convex, there exists a concave increasing function $W_2(\beta)$ and a convex increasing function $T_2(\beta)$ such that,

1. *for $\kappa \leq W_2(\beta)$, it is optimal to locate N station with $\lambda_i^* = \frac{1}{N}$,*
2. *for $\kappa \geq T_2(\beta)$, λ_i^* can be determined by a distribution clustering with a Wasserstein distance.*

Proposition 4.6 shows that, based on the waiting and travel costs (κ, β) , we can have three regions for the optimal number of service locations as depicted in Figure 4.5. First, consider the case where the waiting cost at the optimal service rate $C_W^{\mu^*}(\lambda_i)$ is concave. In this case, when the waiting cost is very high relative to the travel cost, it is optimal to locate only one service station with a high capacity to benefit from the capacity pooling and reduce congestion (Region I in Figure 4.5 (a)). However, as the travel cost κ increases or β decreases, the impact of the travel cost increases, and thus, more service stations should be established to reduce the travel cost. Moreover, when the travel cost is high compared to the waiting cost so that the former dominates the latter, it is optimal to locate many service stations to ensure customers travel a short distance to receive the service. In this case, the optimal demand rate at each location can be determined by finding the best partitions of the space X so that the travel cost is minimized (Region II in Figure 4.5 (a)). This is equivalent to a clustering problem with the goal of minimizing the Wasserstein distance.

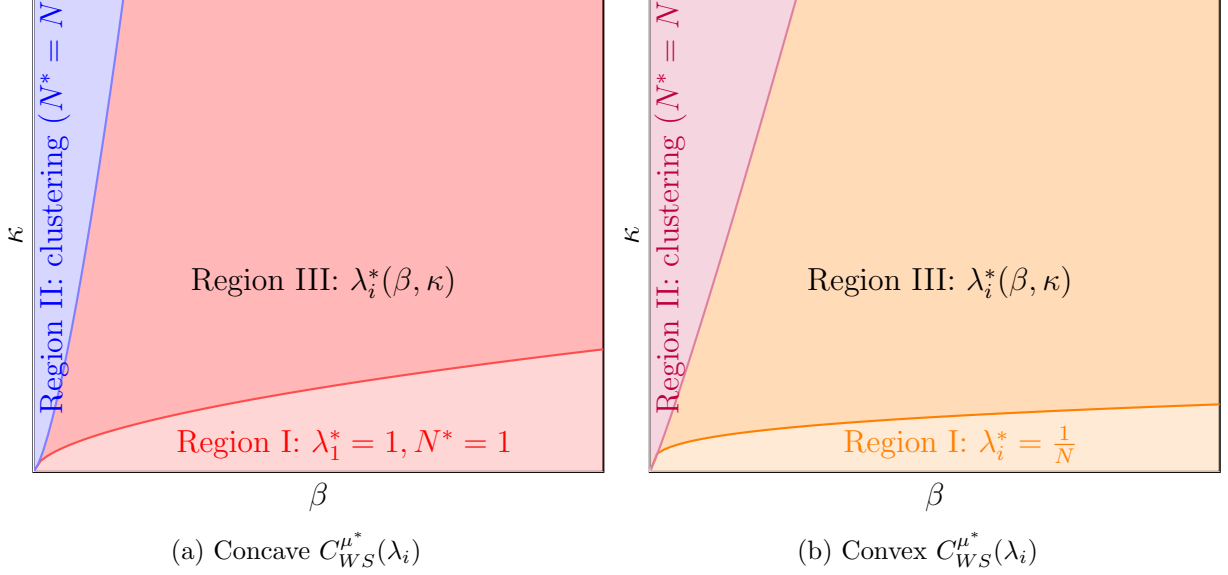


Figure 4.5: Optimal demand rate of service as travel and waiting costs vary.

In the case of $C_W^{\mu^*}(\lambda_i)$ being convex, when κ is high enough such that the travel cost dominates the waiting cost, the optimal demand rate served by each service location is the same as the case with $C_W^{\mu^*}(\lambda_i)$ being concave. Specifically, the optimal demand rate served by each location can be determined so that X is partitioned to minimize the travel cost (Region I in Figure 4.5 (b)). However, as the effect of waiting cost increases, the solution will favor minimizing the total waiting cost. In such a situation, it would be optimal to break down the demand equally between different locations, i.e., $\lambda_i^* = 1/N$ for all locations (Region I in Figure 4.5 (b)).

Chapter 5

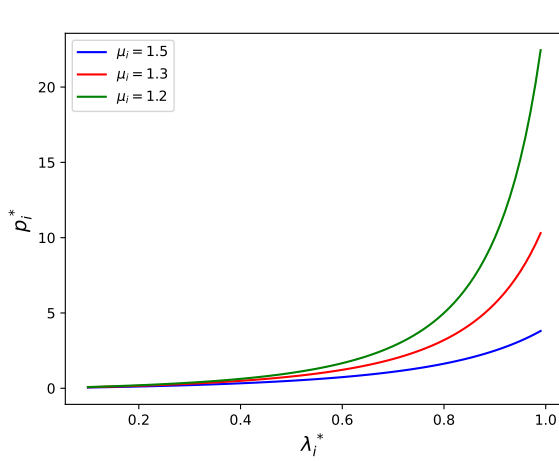
Optimal Pricing

In the previous section, we analyzed the spatial service network problem, relying on Theorem 3.1, and characterized the optimal service and demand rate at each location from a central decision-maker perspective. In this section, we examine how the service fee at different locations can be determined so that utility-maximizing customers follow the optimal demand mapping determined in the previous section. This means that the optimal pricing strategy provided in this section guarantees that all inefficiencies in the system due to the strategic behavior of customers are removed, and thus, social welfare is maximized.

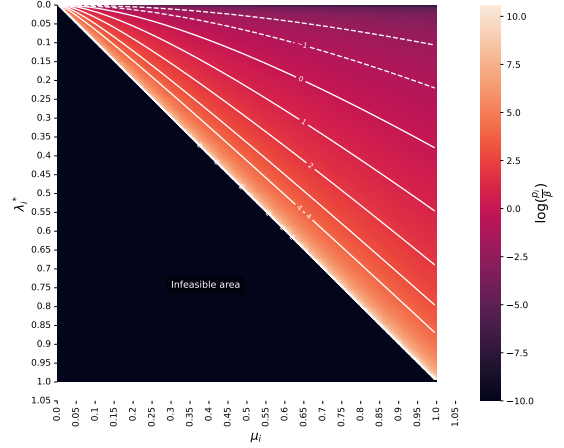
Note that Theorem 3.1 guarantees the existence of a set of service fees such that maximizing social welfare is equivalent to minimizing the total customers' disutility and capacity costs. This means that for a given set of optimal service and arrival rates and service locations, there exists a set of optimal service fees, denoted by p_i^* , that maximizes social welfare when customers are strategic and choose a service location that maximizes their utility. In this section, we investigate how the optimal service fee for each location can be determined using Theorem 3.1 and based on the optimal service and demand rates. We first focus on the case with given service rates in Section 5.1 and then extend the results to the case where optimal service rate and fee are determined simultaneously in Section 5.2.

5.1 Optimal Service Fee for Given Service Capacity

When the service capacity is fixed, demand should be distributed over different locations to balance the waiting and travel costs. To achieve this goal, the service fee should be set such that customers in high-demand areas have an incentive to travel to low-demand regions to receive the service.



(a) Service fee at locations with equal service rate.



(b) Service fees at locations with different service and demand rates.

Figure 5.1: Optimal service fee for given service capacity as demand rate varies.

Proposition 5.1 *Given λ^* and fixed service rates μ ,*

1. *the optimal service fee p_i^* is higher for locations with higher λ_i^* (among service locations with equal service rate),*
2. *there exists a set of optimal service fees such that at location i ,*

$$p_i^* = \frac{\beta \lambda_i^*}{(\mu_i - \lambda_i^*)^2}. \quad (5.1)$$

Proposition 5.1 shows that the optimal service fee at locations with longer waiting times is higher. At first glance, this may seem counterintuitive since customers with higher waiting costs pay higher service fees. However, this result only holds at the optimality of the demand served by each location λ_i^* . At the optimal point, a higher λ_i^* at a service location corresponds to a lower transport cost for customers, making these locations more attractive to strategic customers. Furthermore, although a higher waiting cost may discourage some customers from joining these locations, it is insufficient to achieve social optimality. Thus, the service provider should enforce an additional fee to discourage more customers from joining these locations. Figure 5.1 illustrates how the optimal service fee changes as the

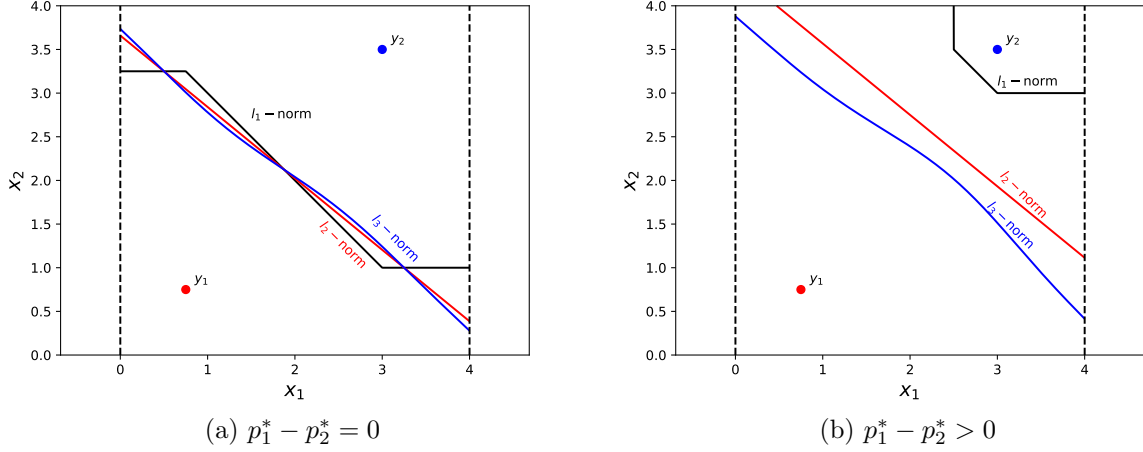


Figure 5.2: Optimal regions for two fixed locations.

optimal demand, represented by λ_i^* , increases. Specifically, part (a) shows that the optimal service fee at service locations with fixed service rates exhibits a convex increasing behavior as the optimal demand rate increases. Part (b) illustrates the changes in service fees as both the service and demand rates vary. Note that the cantors of the graph are in the form of $\mu_i = \sqrt{c\lambda_i^*} + \lambda_i^*$ for a positive c .

In Section 3.1, we discussed how a service fee strategy can be used to encourage strategic customers to follow the socially optimal policy. To understand how pricing affects customer behavior, consider the following simple example. Suppose that the demand distribution is uniform, there are two service locations with fixed locations y_1 and y_2 where $\mu_1 > \mu_2$, and the urban region X is the rectangle $[0, 4] \times [0, 4]$. Figure 5.2 (a) shows customer behavior when the waiting cost is zero ($\beta = 0$), meaning that there is no pricing. In this case, customers choose the location that minimizes their travel costs, and due to the symmetry of X , the arrival rates at the two locations are the same. The black, red, and blue curves illustrate the border for l_1 , l_2 , and l_3 norms, respectively. Note that in the case with $\beta = 0$, the borders are close to each other. However, as β increases, more customers favor the location with the higher service capacity and lower wait time (y_1), as shown in Figure 5.2 (b). To ensure this location is not overcrowded and optimal social welfare is achieved, the service fee at this location should be increased and be higher than the other location to discourage some customers from choosing y_1 . Thus, $p_1^* > p_2^*$. Note that in this case, borders of regions for different norms have a higher distance from each other, with l_1 -norm being the most affected. This is due to the non-linearity of travel cost for $p > 1$; that is,

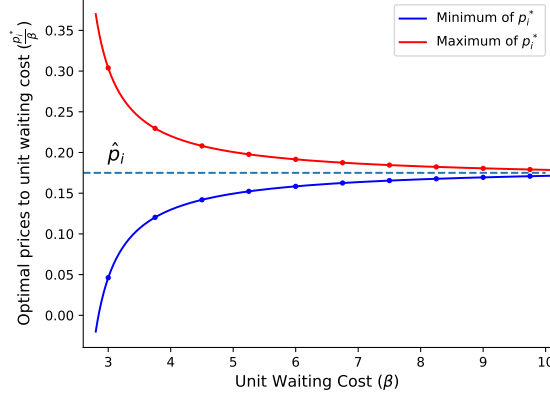


Figure 5.3: Optimal service fee as the waiting cost (β) varies.

the travel cost with higher p has a more significant impact on the customer utility and, thus, their decisions.

Proposition 5.2 *Assume μ is fixed. Then, there exists a set of $\{\hat{p}_i\}_{i \in [N]}$ such that as β increases,*

1. *the optimal price for location i with $\frac{p_i}{\sqrt{\beta}} \leq \hat{p}_i$ increases,*
2. *the optimal price for location i with $\frac{p_i}{\sqrt{\beta}} \geq \hat{p}_i$ decreases.*

Proposition 5.2 states that as β increases, the service fee at different locations approaches $\hat{p}_i$, which is the optimal pricing policy if customers are only sensitive to waiting, as illustrated in Figure 5.3. The reason is that as β increases, the demand served by each location gets closer to each other, as shown in Proposition 4.4, and thus, the gap between the wait times at different locations decreases. This leads to a decrease in the gap between service fees at different locations.

5.2 Joint Optimization of Service Capacity and Price

In this section, we consider the joint optimization of the service fee and capacity, considering the optimal demand rate that should be served by different locations, as discussed in Section

4.3. First, we examine the relation between the optimal service fee at each location and the optimal demand rate served by that location.

Proposition 5.3 *Given the optimal arrival rates served by different locations, there exists a set of optimal service fees such that,*

$$p_i^* = \frac{\partial h(\mu)}{\partial \mu} \Big|_{\mu=\mu^*(\lambda_i^*)}.$$

Proposition 5.3 demonstrates that at the optimality of the demand rate λ_i^* , the optimal service fee can be determined using the marginal service cost at $\mu^*(\lambda_i)$. Note that, since $\mu^*(\lambda_i)$ is increasing in λ_i , the behaviour of the optimal service fee as λ_i^* varies depends on the marginal cost of the service capacity ($\partial h(\mu_i)/\partial \mu_i$) at the optimal service rate. Specifically, the service fees decrease (increase) with the optimal demand rate when the service cost is strictly convex (concave), as illustrated in Figure 5.4.

Corollary 5.1 *Consider the optimal service fee p_i^* . Then,*

1. *for a linear cost of service capacity, the socially optimal demand mapping can be achieved if service locations charge the same service fee,*
2. *for $h(\mu_i) = \log(\gamma_1 \mu_i + \gamma_2)$, the optimal service fee is decreasing in the optimal demand rate and given by,*

$$p_i^* = \left(\lambda_i^* \left(1 + \frac{\beta}{2\gamma_2} \right) + \sqrt{\frac{\beta}{\gamma_2}} \lambda_i^* \sqrt{\frac{1}{\gamma_1} + \lambda_i^* \left(1 + \frac{1}{2\beta\gamma_2} \right)} + \frac{\gamma_1}{\gamma_2} \right)^{-1},$$

3. *for $h(\mu_i) = \alpha \mu_i^2$, the optimal service fee is increasing in optimal demand and given by,*

$$p_i^* = \frac{4}{3} \alpha \lambda_i^* + 2\alpha \lambda_i^{*1/3} \left(c_0 - \lambda_i^{*2} + c_1 \sqrt{c_2 - \lambda_i^{*2}} \right)^{1/3} + 2\alpha \lambda_i^{*5/3} \left(c_0 - \lambda_i^{*2} + c_1 \sqrt{c_2 - \lambda_i^{*2}} \right)^{-1/3},$$

where

$$c_0 = \frac{27\beta}{4\alpha}, \quad c_1 = \sqrt{\frac{3^{3/4}}{2}} \alpha \beta, \quad c_2 = \left(\frac{3}{2\alpha} \right)^3 \beta.$$

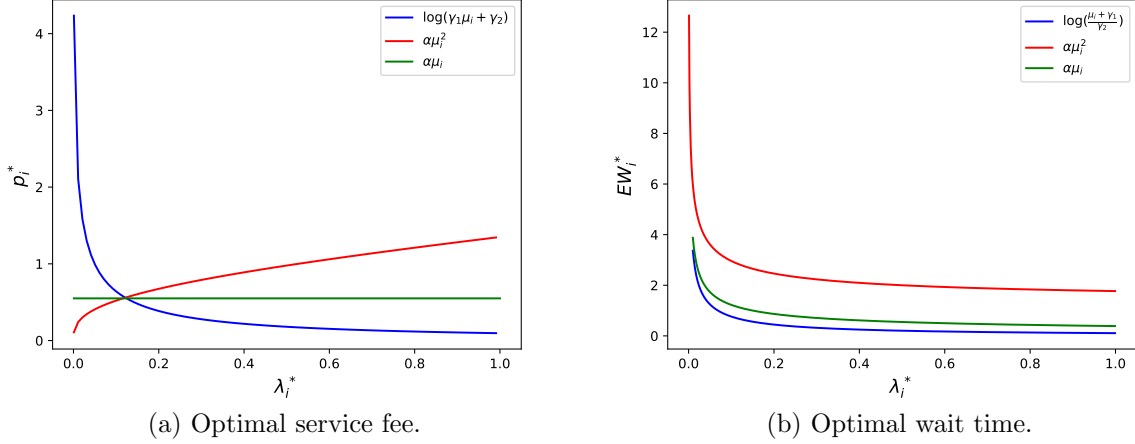


Figure 5.4: Optimal service fee and wait time as a function of the optimal demand rate for different capacity cost functions.

As discussed in Section 5.1, for a group of service locations (A) that charge the same service fee $\bar{p}$, the optimal arrival rates and service rates satisfy,

$$\mu_i = \sqrt{\frac{\beta}{\bar{p}} \lambda_i^* + \lambda_i^*} \quad \forall i \in A. \quad (5.2)$$

Obviously, the above equality does not hold in general when service capacity at each location is given and cannot be optimized, as discussed in Section 5.1. Thus, in such situations, a pricing strategy should be adopted to ensure that optimal mapping is achieved. However, when service capacity at each location can be determined optimally, the equality given in (5.2) holds if the cost of service capacity is linear. Thus, as stated in the first part of Corollary 5.1, with a linear capacity cost, a uniform pricing strategy across different locations guarantees the optimal social behavior of strategic customers. This result can also be inferred from Proposition 5.3, which states that the optimal service fee at each location should be set to the marginal capacity cost at the optimal service rate. For the linear capacity cost, this marginal cost is constant and the same for all locations. Thus, the service fee does not affect customers' decisions, and the optimal capacity allocation leads to optimal social welfare.

Note that the optimal allocation of a fixed capacity between service locations is equivalent to having a linear service cost at each location. Specifically, for any $c > 1$ with

$\sum_{i \in [N]} \mu_i = c$, there exists an α such that the optimal solution of the problem given in (3.7) with $h(\mu_i) = \alpha \mu_i$ is the same as the optimal solution of the following problem,

$$\min_{\{(\lambda, \mu, y) \mid \sum_{i \in [N]} \mu_i = c, \mu_i > \lambda_i\}} \left(\sum_{i \in [N]} \left(\beta \frac{\lambda_i}{\mu_i - \lambda_i} \right) + \kappa \min_{\pi \in \Pi} \mathbb{E}_{\pi} [\|x - y_i\|_p^p] \right).$$

Thus, when there is no cost for service capacity, and the fixed capacity can be optimally allocated between service locations, there is no need to adopt a pricing strategy to make sure the optimal demand mapping is achieved, as discussed in [14].

However, when the capacity cost is not linear, optimal demand mapping cannot be achieved by allocating the service capacity among service locations optimally. Thus, they should charge a service fee to ensure strategic customers follow the optimal demand mapping when they choose a location to receive the service. For the quadratic service capacity cost, the waiting cost changes are lower than in the linear service cost case (Figure 5.4 (b)). Consequently, service fees need to be increased, as illustrated in Figure 5.4 (a) and given in the third part of Corollary 5.1. On the other hand, when the service capacity cost is logarithmic, since the waiting cost for logarithmic service costs changes faster than for linear service costs, as illustrated in Figure 5.4 (b), the service provider needs to decrease the service fee to avoid discouraging too many customers from central locations. This result highlights the impact of the curvature of the service capacity cost on the optimal service fee.

Chapter 6

Conclusions and Future Works

Motivated by the challenges decision-makers face in designing service networks, this paper aims to provide insights into how a service network should be designed in an urban area to maximize social welfare. To achieve this, we consider the main characteristics of an urban area, such as the spatial distribution of customer demand and their sensitivity to travel and wait times. We focus on the trade-off between the cost of service capacity, wait time, and travel distance.

By leveraging the interplay between continuous functions and positive measures, we demonstrated that maximizing social welfare can be rephrased as a two-level optimization problem aimed at minimizing the overall expense of the system, encompassing both customer disutility and service capacity expenses. This setup allows for the explicit assignment of customers to service locations and guarantees the existence of a pricing mechanism such that the optimal assignment is satisfied. Then, by formulating the problem as a bilevel optimization and leveraging concepts from optimal transport theory, we derived insights into the optimal location of service stations, the optimal service capacity, and the optimal pricing strategy.

We showed that optimal customer assignment can be achieved through the division of the urban area into regions and assigning the demand in each region to a service station located in the region. On the basis of these, we were able to provide a geometric identification of the optimal locations. Furthermore, our analysis of the relationship between the demand rates and service capacity has revealed that the shape of the capacity cost plays a significant role in the optimal capacity at each location. For instance, when the capacity cost is linear, the square-root rule is optimal, which indicates that the capacity should be set to the average demand rate plus an additional term that depends on the service cost

function. However, when the service capacity cost is quadratic, the additional term in the optimal service rate is roughly proportional to the cube root of the optimal demand rate, as opposed to the square root in the linear case.

We also explored the connection between optimal demand rates and service fees, considering both predetermined and optimized service capacities. When the service capacity is fixed, the service fees should be higher at locations with higher demand rates to prevent these areas from becoming overly congested. However, this is not the case when both service fees and capacity are optimized simultaneously. Specifically, we demonstrated that service fees decrease, remain constant, or increase with the optimal demand rate when the service cost is strictly convex, linear, or concave, respectively. Additionally, we analyzed how changes in the unit cost of waiting and travel time influence the optimal service fees. Our findings indicate that with a fixed set of capacities, as the unit wait cost rises, the optimal service fees at locations with identical capacities tend to converge.

In the future, we would like to study in what conditions, other than linear service cost, we do not need service fees, especially for specific ranges of travel and waiting costs. Furthermore, we want to study how much we may lose on social welfare if we cannot impose service fees. In another stream of work, we would like to apply our algorithm and methods to real case data. We will check our efficiency compared to other methods in the literature, especially with methods that consider discrete demands points.

In addition to these directions, future research could explore the integration of dynamic pricing models, where service fees might vary depending on real-time factors such as demand fluctuations, traffic conditions, or user behavior patterns. Understanding how these dynamic factors influence the necessity and effectiveness of service fees could provide deeper insights into optimizing system performance while maintaining social welfare. Moreover, investigating the impact of regulatory policies and market conditions on the need for service fees could offer valuable guidelines for policymakers and practitioners.

References

- [1] Robert Aboolian, Oded Berman, and Majid Karimi. Probabilistic set covering location problem in congested networks. *Transportation Science*, 56(2):528–542, 2022.
- [2] Robert Aboolian, Oded Berman, and Vedat Verter. Maximal accessibility network design in the public sector. *Transportation Science*, 50(1):336–347, 2016.
- [3] Robert Aboolian and Majid Karimi. Location of public facilities under congestion. In *Uncertainty in Facility Location Problems*, pages 251–280. Springer, 2023.
- [4] Gad Allon, Awi Federgruen, and Margaret Pierson. How much is a reduction of your customers’ wait worth? an empirical study of the fast-food drive-thru industry based on structural estimation methods. *Manufacturing & Service Operations Management*, 13(4):489–507, 2011.
- [5] R Michael Alvarez, Stephen Ansolabehere, Adam Berinsky, Gabriel Lenz, Charles Stewart III, and Thad Hall. 2008 survey of the performance of american elections. *Caltech/MIT Voting Technology Project*, <http://vote.caltech.edu/drupal/node/231>, 2009.
- [6] Luigi Ambrosio and Aldo Pratelli. *Existence and stability results in the L^1 theory of optimal transportation*, pages 123–160. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.
- [7] Barış Ata and Shiri Shneorson. Dynamic control of an $m/m/1$ service system with adjustable arrival and service rates. *Management Science*, 52(11):1778–1791, 2006.
- [8] Igor Averbakh, Oded Berman, Jörg Kalcsics, and Dmitry Krass. Structural properties of voronoi diagrams in facility location problems with continuous demand. *Operations Research*, 63(2):394–411, 2015.

- [9] Igor Averbakh, Oded Berman, Jörg Kalcsics, and Dmitry Krass. Location problems with continuous demand and unreliable facilities: Applications of families of incremental voronoi diagrams. *Discrete Applied Mathematics*, 300:36–55, 2021.
- [10] Heinz H. Bauschke and Walaa M. Moursi. *Chapter 3: Convex and Lower Semi-continuous Functions*, chapter 3, pages 17–22. Society for Industrial and Applied Mathematics, 2023.
- [11] Oded Berman and Dmitry Krass. Stochastic location models with congestion. In *Location science*, pages 477–535. Springer, 2019.
- [12] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.
- [13] Henry E Brady and John E McNulty. Turning out to vote: The costs of finding and getting to the polling place. *American Political Science Review*, 105(1):115–134, 2011.
- [14] John Gunnar Carlsson, Xiaoshan Peng, and Ilya Ryzhov. Demand equilibria in spatial service systems. *Available at SSRN 4794156*, 2024.
- [15] Ignacio Castillo, Armann Ingolfsson, and Thaddeus Sim. Social optimal location of facilities with fixed servers, stochastic demand, and congestion. *Production and Operations Management*, 18(6):721–736, 2009.
- [16] Tom M. Cavalier and HANIF D. Sherali. Euclidean distance location-allocation problems with uniform demands over convex polygons. *Transportation Science*, 20(2):107–116, 1986.
- [17] Hong Chen and Murray Frank. Monopoly pricing when customers queue. *IIE Transactions*, 36(6):569–581, 2004.
- [18] Donald L. Cohn. *Signed and Complex Measures*, pages 113–141. Springer New York, New York, NY, 2013.
- [19] Raghuveer Devulapalli, Neil Peterson, and John Gunnar Carlsson. Data visualization using weighted voronoi diagrams. In *Geo-Intelligence and Visualization through Big Data Trends*, pages 181–204. IGI Global, 2015.
- [20] Gregory Dobson and Euthemia Stavroulaki. Simultaneous price, location, and capacity decisions on a line of time-sensitive customers. *Naval Research Logistics (NRL)*, 54(1):1–10, 2007.

- [21] Nelson Dunford and Jacob T. Schwartz. *Linear Operators, Part 1: General Theory*. Wiley-Interscience, 1988.
- [22] Noel M. Edelson and David K. Hilderbrand. Congestion tolls for poisson queuing processes. *Econometrica*, 43(1):81–92, 1975.
- [23] Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976.
- [24] Alfred Galichon. *Optimal Transport Methods in Economics*. Princeton University Press, 2016.
- [25] Wilfrid Gangbo and Robert J. McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, January 1996. Publisher: Institut Mittag-Leffler.
- [26] Moshe Haspel and H Gibbs Knotts. Location, location, location: Precinct placement and the costs of voting. *The Journal of Politics*, 67(2):560–573, 2005.
- [27] Refael Hassin. *Rational queueing*. CRC press, New York, NY, 2016.
- [28] Refael Hassin and Ricky Roet-Green. On queue-length information when customers travel to a queue. *Manufacturing & Service Operations Management*, 23(4):989–1004, 2020.
- [29] Mingyi Hong, Zhi-Quan Luo, and Meisam Razaviyayn. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *SIAM Journal on Optimization*, 26(1):337–364, 2016.
- [30] Leonid V Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.
- [31] C.S. Smith M. Knott. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43:39–49, 1984.
- [32] Avishai Mandelbaum and Martin I. Reiman. On pooling in queueing networks. *Management Science*, 44(7):971–981, 1998.
- [33] Haim Mendelson. Pricing computer services: Queueing effects. *Communications of the ACM*, 28(3):312–321, 1985.

- [34] P. Naor. The regulation of queue size by levying tolls. *Econometrica*, 37(1):15–24, 1969.
- [35] Yanfeng Ouyang, Zhaodong Wang, and Hai Yang. Facility location design under continuous traffic equilibrium. *Transportation Research Part B: Methodological*, 81:18–33, 2015.
- [36] Caroline Prodhon and Christian Prins. A survey of recent research on location-routing problems. *European Journal of Operational Research*, 238(1):1–17, 2014.
- [37] Balaraman Rajan, Tolga Tezcan, and Abraham Seidmann. Service systems with heterogeneous customers: Investigating the effect of telemedicine on chronic care. *Management Science*, 65(3):1236–1267, 2019.
- [38] W Rudin. *Signed and Complex Measures*, pages 92–115. Mc-Graw-Hill, New York, NY, 1991.
- [39] Mandy Ryan, E McIntosh, Tara Dean, and P Old. Trade-offs between location and waiting times in the provision of health care: the case of elective surgery on the isle of wight. *Journal of Public Health*, 22(2):202–210, 2000.
- [40] Amol Sasane. *A Friendly Approach to Functional Analysis*. WORLD SCIENTIFIC (EUROPE), 2017.
- [41] A. Shapiro and Y. Wardi. Convergence analysis of gradient descent stochastic algorithms. *Journal of Optimization Theory and Applications*, 91(2):439–454, November 1996.
- [42] Peter Sivey. The effect of waiting time and distance on hospital choice for english cataract patients. *Health economics*, 21(4):444–456, 2012.
- [43] Michael B Teitz. Toward a theory of urban public facility location. *Papers in Regional Science*, 21(1):35–51, 1968.
- [44] Cedric Villani. *Optimal Transport, old and new*. Springer, 2008.
- [45] Zhaodong Wang, Yanfeng Ouyang, and Ruifeng She. On solving a class of continuous traffic equilibrium problems and planning facility location under congestion. *Operations Research*, 70(3):1465–1484, 2022.

- [46] Zhongbin Wang, Shiliang Cui, and Lei Fang. Distance-based service priority: An innovative mechanism to increase system throughput and social welfare. *Manufacturing & Service Operations Management*, 2022.
- [47] R. L. Wilder. Evolution of the topological concept of "connected". *The American Mathematical Monthly*, 85(9):720–726, 1978.
- [48] HAI Yang and S. C. Wong. A continuous equilibrium model for estimating market areas of competitive facilities with elastic demand and market externality. *Transportation Science*, 34(2):216–227, 2000.
- [49] Yue Zhang, Oded Berman, Patrice Marcotte, and Vedat Verter. A bilevel model for preventive healthcare facility network design with congestion. *IIE Transactions*, 42(12):865–880, 2010.

APPENDICES

Appendix A

Prelimiaries and Notations

In this section, we go through some definitions and notations that will be used to prove the results. We start by defining the normed vector space of finite measures. Let $(X, \mathfrak{M})$ be a measurable space and $\mathcal{M}(X)$ the collection of all finite measures defined over this space. Take two measures $\mathbb{F}_1$ and $\mathbb{F}_2$ in $\mathcal{M}(X)$. One could define the sum operators for these measures as $\mathbb{F} = \mathbb{F}_1 + \mathbb{F}_2$ by $\mathbb{F} = \mathbb{F}_1(A) + \mathbb{F}_2(A)$ for all $A \in \mathfrak{M}$, and scalar amplification by $\alpha\mathbb{F} = \alpha.\mathbb{F}$ as $\alpha\mathbb{F}(A) = \alpha.\mathbb{F}(A)$ for all $A \in \mathfrak{M}$. It is easy to see all of the set functions constructed by scalar multiplication or summation are also finite measures. Thus, the space of all finite measures $\mathcal{M}(X)$ is a linear space. Defining $\bar{0}$ as the zero vector, to be zero for all sets in $\mathfrak{M}$, one could show $\mathcal{M}(X)$ is a vector space. Moreover, one could define the total variation norm as

$$\|\mathbb{F}\|_{TV} = \sup\{\mathbb{F}_-(A) | A \in \mathfrak{M}\} + \sup\{\mathbb{F}_+(A) | A \in \mathfrak{M}\},$$

where $\mathbb{F}_+$ and $\mathbb{F}_-$ are the Hahn decompositions of $\mathbb{F}$. Using the monotonicity and positivity of $\mathbb{F}_+$ and $\mathbb{F}_-$, it is easy to see the map $\|\cdot\|_{TV} : \mathcal{M} \rightarrow [0, \infty)$ is indeed a norm. Thus, the space $\mathcal{M}(X)$ normed by $\|\cdot\|_{TV}$ is a normed vector space [18].

Another important normed vector space in our analysis is the space of continuous bounded functions. We denote the set of all bounded and continuous functions by

$$\mathcal{C}_b(X) = \{f : X \rightarrow \mathbb{R} | f \text{ is continuous and bounded}\}.$$

The set $\mathcal{C}_b(X)$ is closed under linear operations, thus $\mathcal{C}_b(X)$ is also a vector space. Moreover, denote

$$\|f\|_\infty = \sup_{x \in X} |f(x)|,$$

and observe the map $\|\cdot\|_\infty : \mathcal{C}_b(X) \rightarrow [0, \infty)$ is a norm. Thus, the space of all bounded and continuous functions with $\|\cdot\|_\infty$ is a normed vector space.

Next, we define the dual space of a normed vector space, as given in [38].

Definition A.1 *Let X be a normed vector space. Then, the space of all continuous linear functions over X is the dual space of X , denoted by $X^\dagger$.*

In this configuration, it is well-known that the dual space of all bounded and continuous functions over a compact space X , with $\|\cdot\|_\infty$ is the set of all Radon measures on the Borel sigma-algebra of X normed with $\|\cdot\|_{TV}$ (see e.g. Theorem IV.6.2 in [21]).

We also need the definition of path connectedness. In an informal sense, a topological space (normed space) is path-connected if there exists a continuous path between every two points in the space. A formal definition is presented below, as given in [47].

Definition A.2 *A space X is path-connected if for all points $x, y \in X$ there exists a path from x to y which is a continuous map $\gamma : [0, 1] \rightarrow X$ with $\gamma(0) = x$ and $\gamma(1) = y$.*

We extend the Real line to $\mathbb{R} \cup \{+\infty\}$, denoted by $\bar{\mathbb{R}}$, and let functions achieve ∞ at some point. This will help us to model the constraints of minimization problems as a part of the objective function. In this context, we define the notion of the domain of a function and indicator functions from convex analysis [10].

Definition A.3 *Let $X \subseteq \mathbb{R}^2$. Then,*

1. *for a function $f(x) : X \rightarrow \bar{\mathbb{R}}$, we define the domain of f as,*

$$\text{dom}(f) = \{x \in X \mid f(x) < \infty\}.$$

2. *for a set $A \subseteq X$, we define the indicator function, $\chi_A : X \rightarrow \bar{\mathbb{R}}$, as*

$$\chi_A(x) = \begin{cases} 0 & x \in A, \\ \infty & x \notin A. \end{cases}$$

It is easy to see that $\chi_A(x)$ is convex if and only if A is convex, and $\text{dom}(f)$ is convex if $f(x)$ is convex [10].

For a convex function with values in $\mathbb{R} \cup \{+\infty\}$, the Fenchel transform of Θ , denoted by $\Theta^\dagger$, defined on the dual space $X^\dagger$ of X is

$$\Theta^\dagger(z^\dagger) = \sup_{z \in X} [\langle z^\dagger, z \rangle - \Theta(z)] .$$

Note that throughout the proofs, we denote probability measures with hollow capital letters, sets with capital letters, and vectors in bold letters. Curvy and bold letters are used to show a collection of sets. We use the notion $\cdot^*$ for any decision variable to show its local (global) optimal value. A summary of notation is presented in Table [A.1](#). To simplify the notation, we will use $\mathbb{Q} = \sum_{i \in [N]} \delta_{\{y_i\}}(x) \lambda_i$.

Variable/Parameter	Notation	Explanation
Set of extended real numbers ($\mathbb{R} \cup \{\infty\}$)	$\bar{\mathbb{R}}$	-
A general metrizable set	A	-
Borel σ - algebra of A with the conventional metric	$\mathcal{B}_A$	subset of 2^A
Set of all probability measures in $(A, \mathcal{B}_A)$	$\mathcal{P}(A)$	-
Set of all continuous bounded functions $A \rightarrow \bar{\mathbb{R}}$	$\mathcal{C}_b(A)$	Set of functions
Order of the l -norm over $\mathbb{R}^n$	p	Positive integer
Maximum number of locations	N	Positive integer
Urban area	X	subset of $\mathbb{R}^2$
Population distribution	$\mathbb{P}$	in $\mathcal{P}(X)$
Arrival rate distribution	$\mathbb{Q}$	in $\mathcal{P}(X)$
Temporal distribution of arrival customers at loc. i	$\mathbb{T}_i$	in $\mathcal{P}(\mathbb{R})$
Set of all marginals of $\mathbb{P}$ and $\mathbb{Q}$	$\Pi(\mathbb{P}, \mathbb{Q})$	in $\mathcal{P}(X \times X)$
Individual transport cost from x to y	$c_T(x, y)$	$\mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$
Individual waiting cost for customers who join loc. i	$c_W(i)$	in $\mathbb{R}$
Utility maximizing service location for customers in x	$i^*(x)$	$\mathbb{R}^2 \rightarrow [N]$
Utility of customers at x that seek service from i	$U(x, i)$	$\mathbb{R}^2 \times [N] \rightarrow \mathbb{R}$
Total social welfare for a tuple of (μ_i, y_i, p_i)	$W_{total}(\mu_i, y_i, p_i)$	$\mathbb{R}^N \times \mathbb{R}^{2 \times N} \times \mathbb{R}^N \rightarrow \mathbb{R}$
Service fee at location i	p_i	Non-negative Real
Arrival rate at location i	λ_i	in $[0, 1]$
Vector representation of arrival rates, $(\lambda_1, \lambda_2, \dots, \lambda_N)$	$\boldsymbol{\lambda}$	in $[0, 1]^N$
Service rate at location i	μ_i	Non-negative Real
Vector representation of service rates, $(\mu_1, \mu_2, \dots, \mu_N)$	$\boldsymbol{\mu}$	in $\mathbb{R}^N$
Location of the i -th service station	y_i	in $\mathbb{R}^2$
Set of locations of all service stations	Y	subset of X
Reward of customers after receiving the service	R	Non-negative Real
Unit cost of waiting	β	Non-negative Real
Unit cost of travel time	κ	Non-negative Real
Cost of service capacity	$h(\mu_i)$	$[0, \infty) \rightarrow [0, \infty)$

Table A.1: Summary of notation

Appendix B

Algorithm Analysis

In this section, we take advantage of the duality of the assignment problem and the alternating-direction method of multipliers to develop an efficient algorithm that reaches a stationary point in the general case. The design problem given in (3.7), in general, is a nonconvex and nonsmooth problem, and it is not straightforward to solve it numerically. However, it precisely consists of two blocks, including the waiting and service capacity costs, which are nonconvex but smooth, and the travel cost, which is nonsmooth but convex.

This block property of the problem motivates us to use block optimization methods, such as the alternating-direction method of multipliers, to solve the problem. The Alternating Direction Method of Multipliers (ADMM) is an algorithm primarily used for solving block optimization problems by breaking them down into smaller sub-problems. Introduced in the 1970s, ADMM combines principles from dual decomposition and augmented Lagrangian methods, making it particularly effective for problems that can be split into separable functions ([23]). The algorithm alternates between optimizing each sub-problem with respect to its variables while keeping the other variables fixed, followed by a step that updates the Lagrange multipliers to enforce consensus among the variables ([12]). This approach is efficient in distributed computing environments, where different sub-problems can be solved in parallel.

We next simplify the design problem using the optimality conditions provided in Section 4 and then develop an ADMM-type algorithm for the problem. Based on Propositions 4.3 and 4.5, we can determine the optimal locations and the service capacity for given λ . We assume that we have $\mu^*(\lambda_i)$ or can approximate it effectively. Thus, in the proposed algorithm, we only need to determine the optimal demand rate and assignment of customers

to locations, i.e., $\boldsymbol{\lambda}^*$ and $\pi^* \in \Pi$. Following the approach introduced by [29], we start by introducing a series of auxiliary variables $\bar{\boldsymbol{\lambda}} \in \mathbb{R}^N$ that is equal to $\boldsymbol{\lambda}$. We will denote the optimal locations for a given λ_i , by $Y^*(\lambda_i)$ and the travel cost for a given set of locations Y , and $\boldsymbol{\lambda}$ as $C_T(\boldsymbol{\lambda}, \{\mu^*(\lambda_i)\}_{i \in [N]}, \{Y^*(\lambda_i)\}_{i \in [N]})$. Thus, the design problem can be rewritten as

$$\min_{\boldsymbol{\lambda}, \bar{\boldsymbol{\lambda}}} C_T(\bar{\boldsymbol{\lambda}}, \{\mu^*(\bar{\lambda}_i)\}_{i \in [N]}, \{Y^*(\bar{\lambda}_i)\}_{i \in [N]}) + \sum_{i \in [N]} \left(h(\mu^*(\lambda_i)) + \sqrt{\lambda_i \frac{\partial h(\mu)}{\partial \mu} \Big|_{\mu=\mu^*(\lambda_i)}} \right)$$

Subjected to:

$$\boldsymbol{\lambda} = \bar{\boldsymbol{\lambda}}, \quad \boldsymbol{\lambda} \in \mathcal{C},$$

where $\mathcal{C}$ is the probability polytope. For this new formulation and for a set of dual variables $v \in \mathbb{R}^n$, we can write the augmented Lagrangian as

$$\begin{aligned} L_\rho(\bar{\boldsymbol{\lambda}}, \boldsymbol{\lambda}, \nu) = & C_T(\boldsymbol{\lambda}, \{\mu^*(\lambda_i)\}_{i \in [N]}, \{Y^*(\lambda_i)\}_{i \in [N]}) + \langle v, \boldsymbol{\lambda} - \bar{\boldsymbol{\lambda}} \rangle \\ & + \frac{\rho}{2} \|\boldsymbol{\lambda} - \bar{\boldsymbol{\lambda}}\|_2^2 + \sum_{i \in [N]} h(\mu^*(\lambda_i)) + \sqrt{\lambda_i \frac{\partial h(\mu)}{\partial \mu} \Big|_{\mu=\mu^*(\lambda_i)}}. \end{aligned}$$

Following the conventional ADMM methods, in each iteration of the algorithm, we update the first set of variables $\bar{\boldsymbol{\lambda}}$, then the second set $\boldsymbol{\lambda}$, and finally take a dual step to improve the feasibility. All the steps are based on minimizing the augmented Lagrangian. However, unlike the conventional ADMM, we update the smooth block only with respect to a given set of indices in each iteration. This leads to greater flexibility and potential numerical improvements. A summary of this algorithm is presented in Algorithm 1.

Algorithm 1 requires solving the problem given in (B.1). Using Kantorovich's duality, one can express the first part of the objective given in (B.1) as follows:

$$C_T(\boldsymbol{\lambda}, \{\mu^*(\lambda_i)\}_{i \in [N]}, \{Y^*(\lambda_i)\}_{i \in [N]}) = \min_{\eta_i \in \mathbb{R}} \sum_{i \in [N]} \lambda_i \eta_i + \mathbb{E}_{\mathbb{P}} \left[\max_{i \in [N]} \{ \kappa \|x - y_i\|_p^p - \bar{p}_i \} \right], \quad (\text{B.2})$$

where $\{\eta_i\}_{i \in [N]}$ is a set of dual variables. Thus, we can write the problem given in (B.1) as a stochastic convex problem on a finite-dimensional space. As a result, we can use convex optimization algorithms, such as stochastic gradient descent, to solve it. Consequently, we can design an algorithm to solve the problem as follows. First, we solve (B.2) to find the dual variables η_i for given $\boldsymbol{\lambda}$ and Y . Then, using dual variables, we determine the optimal regions. Next, the optimal locations are found locally, and we update $\boldsymbol{\lambda}$, taking a gradient

Algorithm 1 An ADMM-type algorithm to solve the design problem

choose y_i^0 and λ_i^0 randomly such that $\sum_{i \in [N]} \lambda_i^0 = 1$

for $t \in \{0, 1, 2, \dots\}$ **do**

$$\bar{\lambda}^{t+1}, Y = \operatorname{argmin}_{\bar{\lambda} \in \mathcal{C}, Y \in X} L_\rho(\bar{\lambda}, \lambda^t, Y, v^t) \quad (\text{B.1})$$

Choose a set of indices $\mathcal{I}^{t+1}$

if $i \in \mathcal{I}^{t+1}$ **then**

solve $\frac{\beta \mu^*(\lambda_i^{t+1})}{(\mu^*(\lambda_i^{t+1}) - \bar{\lambda}_i^{t+1})} + v_i + \rho(\lambda_i^{t+1} - \bar{\lambda}_i^t) + \eta = 0$ for λ_i^{t+1} .

$$v_i^{t+1} = v_i^t + \rho(\lambda_i^{t+1} - \bar{\lambda}_i^{t+1})$$

else

$$\text{let } \lambda_i^{t+1} = \lambda_i^t, v_i^{t+1} = v_i^t$$

end if

end for

descent step with a step size of γ . Algorithm 2 below summarizes these steps to solve the problem given in (B.1).

Algorithm 2 A gradient descent algorithm to solve (B.1)

choose y_i^0 and λ_i^0 randomly such that $\sum_{i \in [N]} \lambda_i^0 = 1$

for $k \in \{0, 1, 2, \dots\}$ **do**

$$\eta_i^k = \operatorname{argmin}_{\bar{p}_i \in \mathbb{R}} \sum_{i \in [N]} \lambda_i \eta_i + \mathbb{E}_{\mathbb{P}}[\max_{i \in [N]} \{\kappa \|x - y_i\|_p^p - \eta_i\}]$$

$$\chi_i^k = \{x \in X \mid \kappa \|x - y_i^t\|_p^p + \eta_i^k \leq \kappa \|x - y_j^t\|_p^p + \eta_j^k, \forall j \in [N]\}$$

Determine y_i^{k+1} using Proposition 4.3.

$$\xi_i^k = \lambda_i^k + \gamma(v_i^t + \rho(\lambda_i - \bar{\lambda}_i^k) - \eta_i^k)$$

$$\bar{\lambda}_i^{k+1} = \frac{\xi_i^k}{\sum_{i \in [N]} \xi_i^k}$$

end for

The first step of Algorithm 2 is a non-constrained stochastic optimization problem that can be efficiently solved using known algorithms in the literature, such as stochastic gradient descent. Note that the iterations in Algorithm 1 follow a gradient descent-type approach, and thus, the number of iterations to achieve the optimal point is polynomially bounded. Furthermore, conventional methods, such as Newton's methods in polynomially bounded steps, can solve the second step of the algorithm. However, the third step generally can not be solved with a polynomial-time algorithm. Nevertheless, it is possible to solve it efficiently for certain $\mathbb{P}$. The following proposition utilizes the results on the convergence of convex functions to prove the convergence of Algorithm 2.

Proposition B.1 *Consider Algorithm 2. Then, it*

1. *converges to the global optimal solution of the problem given in (B.1);*
2. *has a polynomial runtime if and only if $y_i^* = \operatorname{argmin}_{y_i \in \chi_i} \int_{\chi_i} \|x - y_i\|_p^p d\mathbb{P}$ could be solved polynomially for all $i \in \{1, 2, \dots, N\}$;*
3. *has the error of at most $\mathcal{O}(\frac{1}{lk\gamma_g}) + \mathcal{O}(\frac{1}{k}\gamma_g)$ in iteration k where γ_g and l are the stochastic gradient descent step and the number of iterations used to solve the first step, respectively;*

Furthermore, Algorithm 1 converges to a set of stationary points of the design problem given in (3.7).

Appendix C

Proofs of Results

We start by proving Proposition 3.1 and Theorem 3.1, which are stepping stones for the rest of the proofs. Then, we go through the results of the design variables in Section C.2. Finally, we prove the results related to the pricing mechanism.

Proof of Proposition 3.1

Proof. Following a similar reasoning to [14], we show the existence of the equilibrium demand rate for each service location. Note that, for a given tuple of $(\{p_i\}_{i \in [N]}, \{\mu_i\}_{i \in [N]}, \{y_i\}_{i \in [N]})$, Proposition 3.1 is an extension of the existence result in Section 3 of [14] where the utility of strategic customers has an extra term, service fee p_i .

Let $\bar{p}_i = p_i + \beta \frac{1}{\mu_i - \lambda_i}$, and

$$\chi^{\bar{p}_i} = \{x \in X \mid \kappa \|x - y_i\|_p^p + \bar{p}_i \leq \kappa \|x - y_j\|_p^p + \bar{p}_j, \forall j \neq i\}.$$

Furthermore, we define $T_i(\lambda_i) = p_i + \beta \frac{1}{\mu_i - \lambda_i}$ if $\lambda_i < \mu_i$ and $T_i(\lambda_i) = \infty$ if $\mu_i \leq \lambda_i$. Thus, at location i , we have

$$\bar{p}_i = T_i(\mathbb{P}[\chi^{\bar{p}_i}]). \quad (\text{C.1})$$

Our goal is to apply Brouwer's Fixed-Point theorem to show there exists a solution to (C.1). Since the function $T_i(\lambda_i)$ is not bounded, we first reformulate (C.1) to ensure $\bar{p}_i$ is bounded. First, note that since there always exists a set of arrival rates for which the service queues are stable (from the assumption that $\sum_i \mu_i > 1$), we have,

$$\max_{\lambda_i} \min_{i \in [N]} T_i(\lambda_i) = \tilde{p} < \infty.$$

Let $\hat{p} = \max\{1 + \tilde{p} + \hat{x}, \max_{i \in [N]} T_i(0)\}$ where $\hat{x} = \kappa \sup_{x \in X, y_i, y_j} \left| \|x - y_i\|_p^p - \|x - y_j\|_p^p \right|$. Furthermore, let the compact set Δ be

$$\Delta = \{\bar{p}_i \in \mathbb{R}^N | 0 \leq \bar{p}_i \leq \hat{p}, \forall i \in [N]\},$$

and,

$$f_i(\bar{p}_i) = \min\{T_i(\mathbb{P}[\chi^{\bar{p}_i}]), \hat{p}\}. \quad (\text{C.2})$$

Note that $f_i(\bar{p}_i)$ is a function from Δ to Δ . Furthermore, function $T_i(\lambda_i) = p_i + \beta \frac{1}{\mu_i - \lambda_i}$ is continuous in λ_i and $\mathbb{P}[\chi^{\bar{p}_i}]$ is continuous in $\bar{p}_i$ (see [19]). Thus, $f_i(\bar{p}_i)$ is also continuous. Since Δ is a compact convex set, by Brouwer's Fixed-Point Theorem, every continuous function from $\Delta \rightarrow \Delta$ has a fixed point. Thus, there exists a set of $\bar{p}_i \in \Delta$ solving,

$$\bar{p}_i = f_i(\bar{p}_i).$$

However, it can be shown that for a feasible set of $\bar{p}_i$ we have,

$$\min\{T_i(\lambda_i), \hat{p}\} = T_i(\lambda_i) \quad (\text{C.3})$$

For a detailed proof of (C.3) see Theorem 1 in [14]. Thus, there exists a set of $\{\bar{p}_i\}_{i \in [N]}$ satisfying (C.1). Having $\{\bar{p}_i\}_{i \in [N]}$ and $\{p_i\}_{i \in [N]}$, we can uniquely, determine $\{\lambda_i\}_{i \in [N]}$ which completes the proof. $\square$

C.1 Existence of service fees (Proof of Theorem 3.1)

In this section, we first show that for any given λ , μ , and Y , there exists a set of service fees such that,

$$\mathbb{E}_{\mathbb{P}}[U(x, i^*(x))] + \sum_{i \in [N]} \mathbb{Q}[y = y_i] p_i = R - \left(\kappa \min_{\pi \in \Pi} \mathbb{E}_{\pi}[\|x - y_i\|_p^p] + \beta \sum_{i \in [N]} \frac{\lambda_i}{\mu_i - \lambda_i} \right), \quad (\text{C.4})$$

where $\mathbb{Q}[y = y_i]$ is the arrival rate probability measure, i.e., $\mathbb{Q} = \sum_{i \in [N]} \delta_{\{y_i\}}(x) \lambda_i$. Using

(C.4), then, we prove Theorem 1. To show (C.4), we take the following steps: 1) We show that pricing does not directly affect social welfare, i.e., the service fee impact only appears on the service location that each customer chooses. 2) We prove π is a positive measure in the space $(X \times X, \mathcal{B}_X \times \mathcal{B}_X)$ and, it is a marginal for $\mathbb{Q}, \mathbb{P}$. 3) We show the lower-level problem on the right-hand side of (C.4) can be written as Kantorovich's transport problem.

Finally, 4) based on the duality between the space of continuous bounded functions and the space of bounded measures, we construct a set of service fees such that (C.4) holds.

Lemma C.1 *Let $(\boldsymbol{\mu}, Y, \{p_i\}_{i \in [N]})$ be a solution tuple for the left-hand-side problem defined in (C.4). If another set of service prices, $\{\hat{p}_i\}_{i \in [N]}$, leads to the same $i^*(x)$ for all $x \in X$, then $(\boldsymbol{\mu}, Y, \{\hat{p}_i\}_{i \in [N]})$ will lead to the same total social welfare, W_{total} . In particular for any $C \in \mathbb{R}^+$, $(\boldsymbol{\mu}, Y, \{p_i + C\}_{i \in [N]})$ will lead to the same W_{total} .*

Proof. Starting from the definition of social welfare given in (3.4),

$$\begin{aligned}
W_{total}(\boldsymbol{\mu}, Y, \{p_i\}_{i \in [N]}) &= \sum_{i \in [N]} (\mathbb{Q}[Y = y_i] p_i - h(\mu_i)) + \mathbb{E}_P[U(x, i^*(x))] \\
&= \sum_{i \in [N]} (\lambda_i p_i - h(\mu_i)) + \mathbb{E}_{\mathbb{P}} \left[R - \kappa \|x - y_{i^*(x)}\|_p^p - \beta \frac{1}{\mu_{i^*(x)} - \lambda_{i^*(x)}} - p_{i^*(x)} \right] \\
&= \sum_{i \in [N]} (\lambda_i p_i - h(\mu_i)) + R - \mathbb{E}_{\mathbb{P}}[\kappa \|x - y_{i^*(x)}\|_p^p] - \sum_{i \in [N]} \left(\beta \frac{\lambda_i}{\mu_i - \lambda_i} + \lambda_i p_i \right) \\
&= R - \sum_{i \in [N]} h(\mu_i) + \mathbb{E}_{\mathbb{P}}[\kappa \|x - y_{i^*(x)}\|_p^p] - \sum_{i \in [N]} \beta \frac{\mathbb{Q}[Y = y_i]}{\mu_i - \lambda_i}. \\
&= \sum_{i \in [N]} (\lambda_i \hat{p}_i - h(\mu_i)) + R - \mathbb{E}_{\mathbb{P}}[\kappa \|x - y_{i^*(x)}\|_p^p] - \sum_{i \in [N]} \left(\beta \frac{\lambda_i}{\mu_i - \lambda_i} + \lambda_i \hat{p}_i \right) \\
&= \sum_{i \in [N]} (\mathbb{Q}[Y = y_i] \hat{p}_i - h(\mu_i)) + \mathbb{E}_P[U(x, i^*(x))] \\
&= W_{total}(\boldsymbol{\mu}, Y, \{\hat{p}_i\}_{i \in [N]}),
\end{aligned}$$

which concludes the proof of the first part.

For the second part of the lemma, note that, since we are adding a constant value to the service fee in all service locations, $i^*(x)$ is indifferent to the mapping introduced, and thus, W_{total} remains the same. $\square$

Lemma C.2 *Mapping π is a positive measure in space $(X \times X, \mathcal{B}_X \times \mathcal{B}_X)$ and it is a marginal probability measure for $\mathbb{Q}$ and $\mathbb{P}$.*

Proof. First, note that each portion of the population assigned to each location is, by definition, non-negative. Let $\{A_i \times B_j\}_{i,j=1}^\infty$ be an arbitrary mutually disjoint set of subsets of X . Then,

$$\pi(\cup_{i,j=1}^\infty A_i \times B_j) = \sum_{j=1}^\infty \sum_{i=1}^\infty \pi(A_i, B_j)$$

since the demand assigned to each location y_i is the sum of the demand assigned to the location from subsets $\{A_i\}_{i \in \mathbb{N}}$, which is equal to the demand assigned to the location y_i from the union of these sets, $\cup_{i=1}^\infty A_i$. Moreover, for the empty set, π is equal to zero by definition. Thus, π satisfies all the conditions for a set function to be a positive measure. Let $\Pi(\mathbb{P}, \mathbb{Q})$ be the collection of all marginals of $\mathbb{P}$ and $\mathbb{Q}$. Then, by the definition of marginal probabilities,

$$\Pi(\mathbb{P}, \mathbb{Q}) = \left\{ \pi \in \mathcal{P}(X \times X) \mid \pi[A \times Y] = \mathbb{P}[A], \pi[X \times B] = \mathbb{Q}[B], \forall A, B \in \mathcal{B}_X \right\}. \quad (\text{C.5})$$

Note that $\pi \in \Pi$ since we assume that each infinitesimal portion of the population must be assigned to at least one service location and that the total demand assigned to each location is equal to $\mathbb{Q}[y = y_i]$. In other words,

$$\int_X \pi(x, y) dx = \mathbb{Q}[Y = y_i], \quad \int_Y \pi(x, y) dy = \sum_{i \in [N]} \pi(x, y_i) = \mathbb{P}[X = x].$$

□

To proceed, we apply the Fenchel-Rockafellar duality, which is a strong version of the duality rule, or *min-max* principal in the literature [21]. At its very core, it extends the results of Fenchel duality from functions over $\mathbb{R}^N$ to functions over general topological spaces, potentially infinite-dimensional spaces. Using this theorem and the duality between space of continuous bounded functions and bounded measures, Proposition C.1 provides a proof for Kantorovich's duality on compact sets.

Proposition C.1 (Kantorovich's Duality for Compact Sets, Theorem 5.10 in [44])

Let X and Y be two compact sets and $c(x, y)$ be a convex and continuous function. Let Φ_c be,

$$\Phi_c = \{(\varphi, \psi) \in L^1(\mathbb{P}) \times L^1(\mathbb{Q}) \mid \varphi(x) + \psi(y) \leq c(x, y), \quad \forall x \in X, \mathbb{P}\text{-a.s.}, \forall y \in Y, \mathbb{Q}\text{-a.s.}\},$$

where $\Pi(\mathbb{P}, \mathbb{Q})$ is defined in (C.5). Then,

$$\inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \mathbb{E}_\pi[c(x, y)] = \sup_{\varphi, \psi \in \Phi_c} \mathbb{E}_\mathbb{P}[\varphi] + \mathbb{E}_\mathbb{Q}[\psi]. \quad (\text{C.6})$$

Proof. Following the same reasoning as [44], we will prove the duality. We start by proving the following inequality,

$$\sup_{\Phi_c \cap \mathcal{C}_b} \mathbb{E}_\mathbb{P}[\varphi] + \mathbb{E}_\mathbb{Q}[\psi] \leq \sup_{\Phi_c \cap L^1} \mathbb{E}_\mathbb{P}[\varphi] + \mathbb{E}_\mathbb{Q}[\psi] \leq \inf_{\Pi(\mathbb{P}, \mathbb{Q})} \mathbb{E}_\pi[c(x, y)], \quad (\text{C.7})$$

where L^1 is the set of all Lebesgue's' integrable functions. The left inequality is given by the fact that any continuous bounded function is, in particular, integrable and thus, $\mathcal{C}_b(X) \times \mathcal{C}_b(Y) \subset L^1(\mathbb{P}) \times L^1(\mathbb{Q})$. For the right inequality consider $\pi \in \Pi$, then by definition of $\Pi(\mathbb{P}, \mathbb{Q})$,

$$\mathbb{E}_\mathbb{P}[\varphi] + \mathbb{E}_\mathbb{Q}[\psi] = \mathbb{E}_\pi[\varphi(x) + \psi(y)]. \quad (\text{C.8})$$

Now, note that by definition of Φ_c , $\varphi(x) + \psi(y) \leq c(x, y)$ almost everywhere— $\mathbb{P}$ and almost everywhere— $\mathbb{Q}$. Explicitly, let N_x, N_y be sets such that the inequality holds for $(X \setminus N_x) \times (Y \setminus N_y)$ and $\mathbb{P}[N_x] = \mathbb{Q}[N_y] = 0$. Since π has marginals $\mathbb{P}$ and $\mathbb{Q}$, we can write

$$\pi[N_x \times Y] = \mathbb{P}[N_x] = 0, \quad \pi[X \times N_y] = \mathbb{Q}[N_y] = 0.$$

Thus, $\pi[(X \times Y) \setminus (X \setminus N_x) \times (Y \setminus N_y)] = 0$. So,

$$\varphi(x) + \psi(y) \leq c(x, y), \quad \text{almost everywhere—}\pi.$$

Hence,

$$\mathbb{E}_\pi[\varphi(x) + \psi(y)] \leq \mathbb{E}_\pi[c(x, y)].$$

From the above inequality alongside with (C.8), we get

$$\mathbb{E}_\mathbb{P}[\varphi] + \mathbb{E}_\mathbb{Q}[\psi] \leq \mathbb{E}_\pi[c(x, y)].$$

Now, taking the infimum from the right-hand side and supremum from the left-hand side completes the proof of inequality (C.7).

To prove the inverse inequality, we rewrite the constrained infimum problem as an inf sup problem and replace an “inf sup” with a “sup inf.” To do so, we need a rigorous and rather general Min-Max principle. Such a result is provided by Fenchel-Rockafellar duality (theorem IV.6.2 in [21]). Specifically, consider E as a normed vector space, $E^\dagger$ its dual

space and Θ, Ξ two convex functions on E with values in $\bar{\mathbb{R}}$ where there exists z_0 in domain of Θ and Ξ such that Θ is continuous at z_0 . Then, the Fenchel-Rockafellar duality theorem states,

$$\inf_{z \in E} [\Theta + \Xi] = \max_{z^\dagger \in E^\dagger} [-\Theta^\dagger(-z^\dagger) - \Xi^\dagger(z^\dagger)], \quad (\text{C.9})$$

where $\Theta^\dagger$ and $\Xi^\dagger$ are convex conjugate of Θ and Ξ .

For any continuous bounded function $f(x, y) \in \mathcal{C}_b(X \times Y)$, we define

$$\begin{aligned} \Theta(f) &= \begin{cases} 0 & \text{if } f(x, y) \geq -c(x, y), \\ +\infty & \text{else,} \end{cases} \\ \Xi(f) &= \begin{cases} \mathbb{E}_{\mathbb{P}}[\varphi] + \mathbb{E}_{\mathbb{Q}}[\psi] & \text{if } f(x, y) = \varphi(x) + \psi(y) \\ +\infty & \text{else.} \end{cases} \end{aligned} \quad (\text{C.10})$$

Note that Ξ is well-defined since, if we have $\varphi(x) + \psi(y) = \tilde{\varphi}(x) + \tilde{\psi}(y)$ for all x, y , then we could find $s \in \mathbb{R}$ such that $\varphi = \tilde{\varphi} + s$, $\psi = \tilde{\psi} - s$, and thus, $\mathbb{E}_{\mathbb{P}}[\tilde{\varphi}] + \mathbb{E}_{\mathbb{Q}}[\tilde{\psi}] = \mathbb{E}_{\mathbb{P}}[\varphi] + \mathbb{E}_{\mathbb{Q}}[\psi]$. Furthermore, observe that Θ and Ξ are convex since the domains of both are convex (the domains are defined by a linear inequality and a linear equality) and the values in the domain are constant. Also, the continuity assumptions of Fenchel-Rockafellar duality are satisfied with $z_0 = 1$.

The goal is to derive both sides of (C.9) for the function $\Theta + \Xi$ and show that it is equal to (C.6). Recall that the dual space of all bounded and continuous functions over a compact space X endowed with norm $\|\cdot\|_\infty$ is the set of all Radon measures in the Borel sigma-algebra of X endowed with the total variation norm (see Theorem IV.6.2 in [21]).

Recall from Definition A.3 that the indicator function of set A , $\chi_A(f) : \mathcal{C}_b \rightarrow \bar{\mathbb{R}}$ is equal to 0 if $f \in A$ and ∞ if $f \notin A$. The left-hand side of (C.9) is given by,

$$\begin{aligned} \inf_{\varphi, \psi} [\Theta + \Xi] &= \inf_{\varphi, \psi} [\mathbb{E}_{\mathbb{P}}[\varphi] + \mathbb{E}_{\mathbb{Q}}[\psi] + \chi_{\{f(x, y) \geq -c(x, y)\}} + \chi_{\{f(x, y) = \varphi(x) + \psi(y)\}}] \\ &= \inf_{\varphi, \psi} \{\mathbb{E}_{\mathbb{P}}[\varphi] + \mathbb{E}_{\mathbb{Q}}[\psi]; \quad \varphi(x) + \psi(y) \geq -c(x, y)\} \\ &= \inf_{\varphi, \psi} \{-\mathbb{E}_{\mathbb{P}}[\varphi] - \mathbb{E}_{\mathbb{Q}}[\psi]; \quad \varphi(x) + \psi(y) \leq c(x, y)\} \\ &= -\sup_{\varphi, \psi} \{\mathbb{E}_{\mathbb{P}}[\varphi] + \mathbb{E}_{\mathbb{Q}}[\psi]; \quad (\varphi, \psi) \in \Phi_c\} \end{aligned}$$

Next, we derive the convex conjugate of Θ and Ξ :

$$\begin{aligned}
\Theta^\dagger(-\pi) &= \sup_{f \in C_b(X \times Y)} \langle \pi, -f \rangle - \Theta(-f) \\
&= \sup_{f \in C_b(X \times Y)} -\mathbb{E}_\pi[f(x, y)] - \chi_{f(x, y) \geq -c(x, y)} \\
&= \sup_{f \in C_b(X \times Y)} \{-\mathbb{E}_\pi[f(x, y)]; \quad f(x, y) \geq -c(x, y)\} \\
&= \sup_{f \in C_b(X \times Y)} \{\mathbb{E}_\pi[f(x, y)]; \quad f(x, y) \leq c(x, y)\}.
\end{aligned}$$

If π is not a nonnegative measure, then there exists a function $v \in C_b(X \times Y)$ and $v \leq 0$ such that $\mathbb{E}_\pi[v] > 0$. Then, the choice $u = \nu v$, with $\nu \rightarrow +\infty$, shows that the supremum is $+\infty$. On the other hand, if π is a positive measure, then the supremum is achieved by putting $f(x, y) = c(x, y)$. Thus,

$$\Theta^\dagger(-\pi) = \begin{cases} \mathbb{E}_\pi[c(x, y)] & \text{if } \pi \in \mathcal{M}_+(X \times Y) \\ +\infty & \text{else.} \end{cases}$$

For $\Xi^\dagger(\pi)$,

$$\begin{aligned}
\Xi^\dagger(\pi) &= \sup_{f \in \mathcal{C}_b} \langle \pi, f \rangle - \Xi[f] \\
&= \sup_{f \in \mathcal{C}_b} \langle \pi, f \rangle - \mathbb{E}_\mathbb{P}[\varphi] - \mathbb{E}_\mathbb{Q}[\psi] - \chi_{\{f(x, y) = \varphi(x) + \psi(y)\}} \\
&= \sup_{f \in \mathcal{C}_b} \{\langle \pi, f \rangle - \mathbb{E}_\mathbb{P}[\varphi] - \mathbb{E}_\mathbb{Q}[\psi]; f(x, y) = \varphi(x) + \psi(y)\} \\
&= \sup_{f \in \mathcal{C}_b} \{\mathbb{E}_\pi[\varphi(x) + \psi(y)] - (\mathbb{E}_\mathbb{P}[\varphi] + \mathbb{E}_\mathbb{Q}[\psi])\}
\end{aligned}$$

Now, if we could find a function, v such that $\mathbb{E}_\pi[(\varphi(x) + \psi(y))] - (\mathbb{E}_\mathbb{P}[\varphi(x)] + \mathbb{E}_\mathbb{Q}[\psi]) = \epsilon$ then by choosing $u = \nu v$, and $\nu \rightarrow \infty$, the supremum is ∞ . Thus,

$$\Xi^\dagger(\pi) = \begin{cases} 0 & \text{if } \mathbb{E}_\pi[\varphi(x) + \psi(y)] = \mathbb{E}_\mathbb{P}[\varphi] + \mathbb{E}_\mathbb{Q}[\psi] \quad \forall (\varphi, \psi) \in C_b(X) \times C_b(Y) \\ +\infty & \text{else.} \end{cases}$$

Therefore, the right-hand side of (C.9) could be written as

$$\begin{aligned}
\sup_{\pi \in \mathcal{M}} -\Theta^\dagger(-\pi) - \Xi^\dagger(\pi) &= \sup \left\{ -\mathbb{E}_\pi[c(x, y)]; \mathbb{E}_\pi[\varphi(x) + \psi(y)] = \mathbb{E}_\mathbb{P}[\varphi] + \mathbb{E}_\mathbb{Q}[\psi] \quad \forall (\varphi, \psi) \in C_b(X) \times C_b(Y) \right. \\
&\quad \left. \text{and } \pi \in \mathcal{M}_+(X \times Y) \right\} \\
&= \sup \{ -\mathbb{E}_\pi[c(x, y)]; \pi \in \Pi \} \\
&= - \inf_{\pi \in \Pi} \mathbb{E}_\pi[c(x, y)]
\end{aligned}$$

where the second equality is based on the fact that $\pi \in \Pi$ if and only if,

$$\mathbb{E}_\pi[\varphi(x) + \psi(y)] = \mathbb{E}_\mathbb{P}[\varphi] + \mathbb{E}_\mathbb{Q}[\psi]$$

for all continuous bounded functions. Thus, at last

$$\inf_{\Pi(\mathbb{P}, \mathbb{Q})} \mathbb{E}_\pi[c(x, y)] = \sup_{\Phi_c \cap C_b} \mathbb{E}_\mathbb{P}[\varphi] + \mathbb{E}_\mathbb{Q}[\psi]$$

which combines with (C.7) completes the proof. $\square$

Proposition C.2 *For any given (λ, μ, Y) , there exists a set of non-negative optimal service fees such that,*

$$\mathbb{E}_\mathbb{P}[U(x, i^*(x))] + \sum_{i \in [N]} \mathbb{Q}[y = y_i] p_i = R - \left(\min_{\pi \in \Pi} \mathbb{E}_\pi[\kappa \|x - y_i\|_p^p] + \sum_{i \in [N]} \beta \frac{\lambda_i}{\mu_i - \lambda_i} \right).$$

Proof. We start by considering the lower problem on the right-hand side of (C.4),

$$- \min_{\pi \in \Pi} \kappa \mathbb{E}_\pi[\|x - y_i\|_p^p]. \quad (\text{C.11})$$

Since π is marginals of $\mathbb{P}$ and $\mathbb{Q}$, as shown in Lemma C.2, the problem in (C.11) is the Kantorovich transport problem [30]. Thus, by Kantorovich's duality given in Proposition

C.1, the dual problem is

$$\begin{aligned}
& \max_{\psi, \phi \in \mathcal{C}_b \times \mathcal{C}_b} \mathbb{E}_{\mathbb{P}}[\psi(x)] + \mathbb{E}_{\mathbb{Q}}[\phi(y)] \\
& \text{S.t} \\
& \psi(x) + \phi(y) \leq \|x - y_i\|_p^p \quad \forall x \in X \text{ a.s. } \mathbb{P}, \forall y \in X \text{ a.s. } \mathbb{Q}.
\end{aligned}$$

Since $\mathbb{Q}$ is supported in a finite set Y , we have,

$$\begin{aligned}
& \min_{\psi, \phi \in \mathcal{C}_b \times \mathcal{C}_b} \mathbb{E}_{\mathbb{P}}[\psi(x)] + \sum_{i \in [N]} \mathbb{Q}[y = y_i] \phi(y_i) \\
& \text{S.t} \\
& \psi(x) + \phi(y_i) \leq \|x - y_i\|_p^p \quad \forall x \in X \text{ a.s. } \mathbb{P}, \forall i \in [N],
\end{aligned} \tag{C.12}$$

which is known to have a unique solution in a compact X (see Chapter 2 in [44]). Note that the problem in (C.12) can be written as

$$\begin{aligned}
& \min_{\phi \in \mathbb{R}^n} \mathbb{E}_{\mathbb{P}}[\min_{i \in [N]} -\phi(y_i) + \|x - y_i\|_p^p] + \sum_{i \in [N]} \mathbb{Q}[y = y_i] \phi(y_i) \\
& = \min_{\phi \in \mathbb{R}^n} \mathbb{E}_{\mathbb{P}}[\min_{i \in [N]} -\phi(y_i) + \|x - y_i\|_p^p] - \beta \sum_{i \in [N]} \frac{\lambda_i}{\mu_i - \lambda_i} + \beta \sum_{i \in [N]} \frac{\lambda_i}{\mu_i - \lambda_i} + R - R + \sum_{i \in [N]} \mathbb{Q}[y = y_i] \phi(y_i) \\
& = \min_{\phi \in \mathbb{R}^n} \mathbb{E}_{\mathbb{P}} \left[\min_{i \in [N]} \left\{ -R - \phi(y_i) + \|x - y_i\|_p^p \right\} + \beta \frac{1}{\mu_i - \lambda_i} \right] - \beta \sum_{i \in [N]} \frac{\lambda_i}{\mu_i - \lambda_i} + R + \sum_{i \in [N]} \mathbb{Q}[y = y_i] \phi(y_i).
\end{aligned}$$

Thus, by setting $p_i = -\phi_i^* - \beta \frac{1}{\mu_i - \lambda_i}$, where ϕ_i^* is the unique minimizer of the problem defined in (C.12), we have

$$\min_{\phi \in \mathbb{R}^n} \mathbb{E}_{\mathbb{P}}[\min_{i \in [N]} -\phi(y_i) + \|x - y_i\|_p^p] + \sum_{i \in [N]} \mathbb{Q}[y = y_i] \phi(y_i) \tag{C.13}$$

$$= \mathbb{E}_{\mathbb{P}}[\min_{i \in [N]} -R + p_i + \beta \frac{1}{\mu_i - \lambda_i} + \|x - y_i\|_p^p] - \beta \sum_{i \in [N]} \frac{\lambda_i}{\mu_i - \lambda_i} + R + \sum_{i \in [N]} \mathbb{Q}[y = y_i] \phi(y_i) \tag{C.14}$$

$$= -(\mathbb{E}_{\mathbb{P}}[U(x, i^*(x))] + \sum_{i \in [N]} \mathbb{Q}[y = y_i] p_i) - \beta \sum_{i \in [N]} \frac{\lambda_i}{\mu_i - \lambda_i} + R. \tag{C.15}$$

By setting (C.15) equal to (C.11), we get

$$\mathbb{E}_{\mathbb{P}}[U(x, i^*(x))] + \sum_{i \in [N]} \mathbb{Q}[y = y_i] p_i = R - \left(\min_{\pi \in \Pi} \mathbb{E}_{\pi}[\kappa \|x - y_i\|_p^p] + \sum_{i \in [N]} \beta \frac{\lambda_i}{\mu_i - \lambda_i} \right).$$

Finally, using Lemma C.1, we can always find a C such that all service fees are non-negative. This concludes the proof. $\square$

Proof of Theorem 3.1

Proof. Starting from the definition of social welfare given in (3.4),

$$\begin{aligned} \max_{\mathbf{p}, \boldsymbol{\lambda}, Y} W_{total} &= \max_{\mathbf{p}, \boldsymbol{\lambda}, Y} \left(\sum_{i \in [N]} \lambda_i p_i - \sum_{i \in [N]} h(\mu_i) + \mathbb{E}_{\mathbb{P}}[U(x, i^*(x))] \right) \\ &= R - \max_{i^*(x), \boldsymbol{\mu}, Y} \left(-\mathbb{E}_{\mathbb{P}}[\kappa \|x - y_{i^*(x)}\|_p^p] - \sum_{i \in [N]} \left(h(\mu_i) + \beta \frac{\mathbb{Q}[Y = y_i]}{\mu_i - \lambda_i} \right) \right) \quad (\text{by Lemma C.1}) \\ &= R - \max_{\boldsymbol{\lambda}, \boldsymbol{\mu}, Y} \left(-\sum_{i \in [N]} \left(h(\mu_i) + \beta \frac{\mathbb{Q}[Y = y_i]}{\mu_i - \lambda_i} \right) - \min_{\pi \in \Pi} [\kappa \|x - y_i\|_p^p] \right) \quad (\text{by Proposition C.2}) \end{aligned}$$

where the last equality is based on the fact that given $(\boldsymbol{\lambda}, \boldsymbol{\mu}, Y)$ and π , $i^*(x)$ can be determined for any group of customers with a positive measure. $\square$

C.2 Proof of Results on Optimal Design Variables

In this section, we discuss the proof of the results related to three sets of design variables, including mapping (π) , service rates $(\boldsymbol{\mu})$, and arrival rates $(\boldsymbol{\lambda})$.

Optimal mapping and location

In this subsection, we prove the results related to the optimal mapping and service locations. We start by proving that the urban area X can be divided into N regions. Then,

we connect this result to the duality discussions in Proposition C.2, and on the basis of that, prove that the regions are simply connected. Then, we show that optimization for service locations is local, and provide the optimality condition for the locations. Finally, we provide the shape of the optimal regions for two special cases.

Proof of Proposition 4.1

Proof. Given a feasible tuple of design variables $(\boldsymbol{\lambda}, \boldsymbol{\mu}, Y)$, without loss of generality, we reorder the indices of arrival rates, locations, service rates, and regions such that $\mathbb{P}[\chi_1] \geq \mathbb{P}[\chi_2] \geq \dots \geq \mathbb{P}[\chi_N]$. Thus, the regions claimed in Proposition 4.1 are empty for $i \in \{N^* + 1, N^* + 2, \dots, N\}$ and for the rest defined as,

$$\chi_i = \{x \in X \mid \pi^*(x, y_i) > 0\}, \quad \forall i \in [N^*].$$

Thus, χ_i 's are constructed based on the problem $\min_{\pi \in \Pi} \mathbb{E}_\pi[\|x - y_i\|_p^p]$. This problem for $p \geq 1$ turns out to be equivalent to Monge's problem ([6] and [25]) defined as,

$$M_p^p(\mathbb{P}, \mathbb{Q}) = \inf \left\{ \int_X \|x - T(x)\|_p^p d\mathbb{P} \mid \begin{array}{l} T(x) : X \rightarrow X \text{ Borel and } \text{ such that } \mathbb{P}(T^{-1}(A)) = \mathbb{Q}(A) \\ \text{for all } A \in \mathcal{B}_X \end{array} \right\}.$$

In other words, we are looking for a function $T(x)$ to assign each demand point to a service location. It is known that $M_p^p(\mathbb{P}, \mathbb{Q})$ is equivalent to Kantorovich's transportation problem for $p \geq 1$ (see [6] and [25]). In other words, all the customers at point x are assigned to exactly the same location $T(x)$. From the demand mapping perspective, for any set $A \in \mathcal{B}_X$, and location $i \in [N]$, $\pi(A, y_i)$ is either equal to $\mathbb{P}(A)$ or 0. Thus, we can define a group of subsets with cardinality N^* , where only points $x \in X$ are in each subset, that $\pi(x, y_i)$ has a positive value. Since the sum of assigned customers from point x to all locations is equal to $\mathbb{P}(x)$, these sets are disjoint. Thus, X can be divided into N^* regions so that all customers are assigned to the same service location in each region.

It is left to show that the regions are path-connected. Recall from Definition A.2 that a set is path-connected if there exists a continuous path between every two points inside it. Formally, we aim to prove for any arbitrary $a, b \in \chi_i$ there exist a continuous function $\eta(t) : [0, 1] \rightarrow X$ such that $\eta(t) \in \chi_i$ for all $t \in [0, 1]$ and $\eta(0) = a$ and $\eta(1) = b$.

We start by defining,

$$\bar{p}_i = \frac{p_i}{\kappa} + \frac{\beta}{\kappa} \frac{1}{\mu_i - \lambda_i}, \quad \forall i \in [N], \quad (\text{C.16})$$

and,

$$\chi_i^{\bar{p}_i} = \{x \in X \mid \|x - y_i\|_p^p + \bar{p}_i \leq \|x - y_j\|_p^p + \bar{p}_j, \forall i \neq j\}, \quad \forall i \in [N]. \quad (\text{C.17})$$

Fix $i \in [N]$ and recall that from Proposition C.2, given any (μ, λ, Y) , we can always construct a set of service prices $\{p_i\}_{i \in [N]}$, such that π^* is achieved by customers' decisions. Thus, there exists a set of constant $\{\bar{p}_i\}_{i \in [N]}$ such that,

$$\chi_i = \chi_i^{\bar{p}_i} \quad \forall i \in [N], \quad (\text{C.18})$$

where χ_i 's are defined in Proposition 4.1.

Take two points a and b in χ_i and, without loss of generality, assume $\|a - y_i\|_p^p \leq \|b - y_i\|_p^p$. According to (C.18), for a and b we have,

$$\begin{aligned} \|a - y_i\|_p^p + \bar{p}_i &\leq \|a - y_j\|_p^p + \bar{p}_j \quad a \in \chi_i, \forall j \neq i \\ \|b - y_i\|_p^p + \bar{p}_i &\leq \|b - y_j\|_p^p + \bar{p}_j \quad b \in \chi_i, \forall j \neq i. \end{aligned}$$

Now we aim to construct a continuous path, $\gamma(t)$, between a and b that lies completely inside χ_i . Define $D_i : X \rightarrow \mathbb{R}$ as $D_i(x) = \|x - y_i\|_p^p + \bar{p}_i$. At each point, starting from b , we take the direction of the negative gradient of $D_i(x)$. This constructs a path $(x^1, \tau(x^1))$, where $\tau(x) : \mathbb{R} \rightarrow \mathbb{R}$ is the solution of the differential equation,

$$\frac{d\tau(x^1)}{dx^1} \frac{\partial D_i(x)}{\partial x^1} = -1,$$

with boundary conditions $\tau(b^1) = b^2$ and $\tau(y_i^1) = y_i^2$ where $b = (b^1, b^2)$ and $y_i = (y_i^1, y_i^2)$. Note that by infinite differentiability of $D_i(x)$, the solution of the ODE exists uniquely and smoothly. Since, at each point, we take the path of the gradient of $-D_i(x)$, we are taking the direction in which the decrease in $D_i(x)$ is maximized. Note that this results in, for all $y^1 \leq x^1 \leq b^1$, $(x^1, \tau(x^1))$ to be inside χ_i . To see the reason that this holds, assume the contrary: there exists a point e such that $\tau(e^1) = e^2$ and $e \notin \chi_i$. Then, $D_j(e) < D_i(e)$ for some $i \in [N]$, while $D_i(a) \leq D_j(a)$ for all $j \in [N]$. In other words, we move on the path of maximum increment of D_i from a point for which $D_j(\cdot) < D_i(\cdot)$ and end up in a point for which $D_i(\cdot) \leq D_j(\cdot)$ for all $j \in [N]$. This is a contradiction, and thus, for all $y^1 \leq x^1 \leq b^1$, $(x^1, \tau(x^1))$, $(x^1, \tau(x^1))$ is indeed in χ_i .

By the mean value theorem, there exists a point (d^1, d^2) on $(x^1, \tau(x^2))$ where $D_i(d^1, \tau(d^1)) =$

$D_i(a)$. If the point is equal to a , the path is complete; if not, we continue on the path

$$D_i(x^1, (\hat{\tau}(x^1))) = D_i(a),$$

until we reach the point a . This part of the path is also continuous due to the continuity of D_i . In summary, we define $\eta(t)$ as,

$$\eta(t) = \begin{cases} (b^1t + a^1(1-t), \tau(b^1t + a^1(1-t))) & 0 \leq t \leq \frac{d^1-a^1}{b^1-a^1} \\ (b^1t + a^1(1-t), \hat{\tau}(b^1t + a^1(1-t))) & \frac{d^1-a^1}{b^1-a^1} \leq t \leq 1 \end{cases}$$

which is a continuous function in t , and for all $t \in [0, 1]$, $\eta(t) \in \chi_i$. Furthermore, $\eta(0) = b$ and $\eta(1) = a$. Thus, by construction, χ_i is path-connected for all $i \in [N]$. $\square$

Proof of Corollary 4.1

Proof. Let $x_0 \in X$ be an arbitrary point inside the interior of χ_i and for a $\epsilon > 0$ consider

$$B_\epsilon(x_0) = \{x \in X \mid \|x - x_0\|_p \leq \epsilon\},$$

with $\mathbb{P}[B_\epsilon(x_0)] > 0$. For the sake of contradiction, assume for customers inside $B_\epsilon(x_0)$ there exists two locations namely y_1 and y_2 such that $U(x, 1) = U(x, 2)$ for all $x \in B_\epsilon(x_0)$. Now consider $\hat{x} = r\epsilon_0 + x_0$ for an $\epsilon_0 < \epsilon$ and $\|r\|_p = 1$. By definition of l_p -ball, $\hat{x} \in B_\epsilon(x_0)$. However, by the contradiction assumption,

$$\begin{aligned} U(\hat{x}, 1) &= U(\hat{x}, 2) \\ \Rightarrow R - p_1 - \beta \frac{1}{\mu_1 - \lambda_1} - \kappa \|\hat{x} - y_1\|_p^p &= R - p_2 - \beta \frac{1}{\mu_2 - \lambda_2} - \kappa \|\hat{x} - y_2\|_p^p \\ \Rightarrow \kappa \|\hat{x} - y_1\|_p^p &= \kappa \|\hat{x} - y_2\|_p^p \\ \Rightarrow \|r\epsilon_0 + x_0 - y_1\|_p &= \|r\epsilon_0 + x_0 - y_2\|_p. \end{aligned}$$

That is a contradiction for a general $\|r\|_p = 1$. Thus, for any $B_\epsilon(x_0)$ with a positive measure, the utility-maximizing service location is unique.

For the second part of the corollary, in the general form of the transport cost $\|x - y_i\|_p^p$, consider two arbitrary regions, χ_i and χ_j . The border of i and j is defined by $\chi_i \cap \chi_j$; that

is,

$$\begin{aligned}
\chi_i \cap \chi_j &= \{x \in X \mid \kappa \|x - y_i\|_p^p + p_i + \beta \frac{1}{\mu_i - \lambda_i} \leq \kappa \|x - y_k\|_p^p + p_k + \beta \frac{1}{\mu_k - \lambda_k}, \quad \forall k \neq i, \\
&\quad \kappa \|x - y_j\|_p^p + p_j + \beta \frac{1}{\mu_j - \lambda_j} \leq \kappa \|x - y_k\|_p^p + p_k + \beta \frac{1}{\mu_k - \lambda_k}, \quad \forall k \neq j\} \\
&= \{x \in X \mid \kappa \|x - y_i\|_p^p + p_i + \beta \frac{1}{\mu_i - \lambda_i} = \kappa \|x - y_j\|_p^p + p_j + \beta \frac{1}{\mu_j - \lambda_j}, \\
&\quad \kappa \|x - y_i\|_p^p + p_i + \beta \frac{1}{\mu_i - \lambda_i} \leq \kappa \|x - y_k\|_p^p + p_k + \beta \frac{1}{\mu_k - \lambda_k}, \quad \forall k \neq i, k \neq j, \\
&\quad \kappa \|x - y_j\|_p^p + p_j + \beta \frac{1}{\mu_j - \lambda_j} \leq \kappa \|x - y_k\|_p^p + p_k + \beta \frac{1}{\mu_k - \lambda_k}, \quad \forall k \neq j, k \neq i\}.
\end{aligned} \tag{C.19}$$

Thus, customers are indifferent to the choice of adjacent locations on the border of those regions. $\square$

Proof of Proposition 4.2

Proof. In the general form of the transport cost $\|x - y_i\|_p^p$, consider two arbitrary regions, χ_i and χ_j , and let $\bar{p}_i$ to be as defined in (C.16). Then, the border of i and j is given, based on (C.19), as,

$$\begin{aligned}
\chi_i \cap \chi_j &= \{x \in X \mid \|x - y_i\|_p^p + \bar{p}_i = \|x - y_j\|_p^p + \bar{p}_j, \\
&\quad \|x - y_i\|_p^p + \bar{p}_i \leq \|x - y_k\|_p^p + \bar{p}_k, \quad \forall k \neq j, k \neq i \\
&\quad \|x - y_j\|_p^p + \bar{p}_j \leq \|x - y_k\|_p^p + \bar{p}_k, \quad \forall k \neq j, k \neq i\}.
\end{aligned}$$

This means that, if a border exists between two regions χ_i and χ_j , it is a subset of

$$\{x \in X \mid \|x - y_i\|_p^p - \|x - y_j\|_p^p = \bar{p}_j - \bar{p}_i\}.$$

Thus, the shape of borders/regions only depends on the value of p .

For $p = 1$, we have,

$$\Rightarrow \begin{cases} x^1 = \frac{1}{2} (|y_i^2 - x^2| + y_i^1 - |y_j^2 - x^2| + y_j^1 - (\bar{p}_i - \bar{p}_j)) \\ x^1 = \frac{1}{2} (-|y_i^2 - x^2| + y_i^1 + |y_j^2 - x^2| + y_j^1 + (\bar{p}_i - \bar{p}_j)) \end{cases}.$$

By conditioning on the positive and negative values of the absolute function, we have,

$$k_1 x^1 + k_2 x^2 = c,$$

where $c \in \mathbb{R}$ and $k_1, k_2 \in \{-1, 0, 1\}$.

For $p = 2$, recall that,

$$\chi_i = \{x \in X \mid \|x - y_i\|^2 + \bar{p}_i \leq \|x - y_j\|^2 + \bar{p}_j, \quad \forall i \neq j\}.$$

We can rewrite $\|\cdot\|_2^2$ as,

$$\|x - y_i\|_2^2 = \|x\|_2^2 + \|y_i\|_2^2 - 2x^T y_i.$$

Let $\nu_i = \|y_i\|_2^2 + \bar{p}_i$ and note that, for each i , $\|y_i\|_2^2$ is a constant. Then,

$$\chi_i = \{x \in X \mid -\nu_i + 2x^T y_i \geq -\nu_j + 2x^T y_j, \quad \forall i \neq j\}.$$

Hence, the borders of regions are indeed linear. By the Knott-smith optimality condition, regions are convex [31]; thus, regions are polyhedra. Furthermore, since X is bounded, χ_i is bounded for any $i \in [N]$, and regions are polytopes. $\square$

To prove Proposition 4.3, we start by showing that the relation between optimal locations of service stations and the optimal regions is independent for each i .

Lemma C.3 *Let (λ^*, μ^*, Y^*) be an optimal tuple for the design problem with regions $\{\chi_i\}_{i \in [N]}$, as defined in Proposition 4.1. Then, we have,*

$$y_i^* = \operatorname{argmin}_{y_i \in X} \int_{\chi_i} \|x - y_i\|_p^p d\mathbb{P}, \quad \forall i \in [N]. \quad (\text{C.20})$$

Proof. From Proposition 4.1, we have:

$$\mathbb{E}_\pi[\|x - y_i\|_p^p] = \sum_{i \in [N]} \int_{\chi_i} \|x - y_i\|_p^p d\mathbb{P}.$$

To see this, note that π^* is positive if $x \in \chi_i$. Also, by marginality of π , we have $\pi(x, y) = \mathbb{P}$ for $x \in \chi_i$. Thus, we can rewrite the design problem given in (3.7) as

$$\min_{\lambda_i, \mu_i, y_i} \sum_{i \in [N]} \left(h(\mu_i) + \beta \frac{\lambda_i}{\mu_i - \lambda_i} + \kappa \int_{\chi_i} \|x - y_i\|_p^p d\mathbb{P} \right). \quad (\text{C.21})$$

From (C.21), it is easy to see why the statement of the lemma holds. Specifically, let

$$y_i^* = \operatorname{argmin}_{x \in X} \int_{\chi_i} \|x - y_i\|_p^p d\mathbb{P}.$$

Furthermore, assume, for the sake of contradiction, that there exists $\hat{y}_i = y_i^* + y_0$ with $y_0 \neq 0$ such that,

$$\sum_{i \in [N]} \left(h(\mu_i^*) + \beta \frac{\lambda_i^*}{\mu_i^* - \lambda_i^*} + \kappa \int_{\chi_i} \|x - \hat{y}_i\|_p^p d\mathbb{P} \right) \leq \sum_{i \in [N]} \left(h(\mu_i^*) + \beta \frac{\lambda_i^*}{\mu_i^* - \lambda_i^*} + \kappa \int_{\chi_i} \|x - y_i^*\|_p^p d\mathbb{P} \right).$$

Then, for given μ_i^* and λ_i^* ($i \in [N]$), the above inequality leads to

$$\kappa \int_{\chi_i} \|x - (y_i^* + y_0)\|_p^p d\mathbb{P} \leq \kappa \int_{\chi_i} \|x - y_i^*\|_p^p d\mathbb{P}.$$

This contradicts the assumption that (C.20) is satisfied. Note that, due to the strict convexity of the problem, the optimal location is unique. $\square$

Proof of Proposition 4.3

Proof. Based on Lemma C.3 we aim to solve,

$$\min_{y_i \in X} \int_{\chi_i} (x - y_i)^p d\mathbb{P}.$$

This problem is convex and smooth; thus, by the first-order optimality condition, we have:

$$\left(\frac{\partial}{\partial y_i} \int_{\chi_i} (x - y_i)^p d\mathbb{P} \right) \Big|_{y_i = y_i^*} = 0.$$

Since $\|x - y_i\|_p^p$ is bounded due to boundedness of X , using Lebesgue's dominated

convergence theorem, we get

$$\begin{aligned}
\frac{\partial}{\partial y_i^j} \int_{\chi_i} \sum_{j=1}^2 (x - y_i^j)^p d\mathbb{P} &= \lim_{h \rightarrow 0} \frac{1}{h} \left(\int_{\chi_i} \|x - (y_i^j + h)\|_p^p d\mathbb{P} - \int_{\chi_i} \|x - y_i^j\|_p^p d\mathbb{P} \right) \\
&= \lim_{h \rightarrow 0} \int_{\chi_i} \frac{1}{h} (\|x - (y_i^j + h)\|_p^p - \|x - y_i^j\|_p^p) d\mathbb{P} \\
&= \int_{\chi_i} \lim_{h \rightarrow 0} \frac{1}{h} (\|x - (y_i^j + h)\|_p^p - \|x - y_i^j\|_p^p) d\mathbb{P} \\
&= \int_{\chi_i} \frac{\partial}{\partial y_i^j} \|x - y_i^j\|_p^p d\mathbb{P} \\
&= p \int_{\{\chi_i | x^j \geq y_i^{*j}\}} (x - y_i^j)^{p-1} d\mathbb{P} - p \int_{\{\chi_i | x^j \leq y_i^{*j}\}} (x - y_i^j)^{p-1} d\mathbb{P},
\end{aligned}$$

for $j = 1, 2$. Note that to get the last equality, we use the smoothness of $\|x - y_i\|_p^p$ for all but the negligible set $\{x^j = y_i^j\}$. The second part of the proposition can be obtained by setting $p = 2$ and considering that $\int_{\chi_i} d\mathbb{P} = \lambda_i^*$ by definition. $\square$

Optimal demand for given service capacity

The optimal demand for a given set of service capacities, as shown in Proposition 4.4 and Corollary 4.2, depends on the structural results of block optimization. This involves optimizing a function in the form of $f(x) + g(x)$, where each block has some properties such as convexity and differentiability. We first start by showing general results for optimization of functions in the form of,

$$\beta \sum_{i \in [N]} f_i(x_i) + \sum_{i \in [N]} g_i(x), \quad (\text{C.22})$$

where f_i (for $i \in [N]$) is convex, smooth and increasing, and $g(x) = \sum_{i \in [N]} g_i(x_i)$ is convex and Lipschitz continuous. Specifically, we show that as β increases, the optimal solution of (C.22) gets closer to the optimal solution of $\sum_{i \in [N]} f_i(x_i)$. Then, we prove that there exists a value for β such that the optimal solution (C.22) is equal to the optimal solution of $\sum_{i \in [N]} f_i(x_i)$. Using these results, and a result on the optimal solution of $\beta \sum_{i \in [N]} \frac{\lambda_i}{\mu_i - \lambda_i}$, we prove Proposition 4.4 and Corollary 4.2.

Lemma C.4 *Let $f(x, y) : \mathbb{R}^{N \times M} \rightarrow \mathbb{R}$ be a convex and continuous function, and $\mathcal{C}$ a*

convex, compact set. Furthermore, let

$$\begin{aligned}(\bar{x}, \bar{y}) &= \min_{y \in \mathcal{C}} \min_{x \in \mathcal{C}} f(x, y), \\(x^*, y^*) &= \min_{(x, y) \in \mathcal{C}} f(x, y).\end{aligned}$$

Then, if $(\bar{x}, \bar{y}) \in \text{rel}(\mathcal{C})$ where $\text{rel}(\mathcal{C})$ is the relative interior of the set $\mathcal{C}$, $(x^*, y^*) = (\bar{x}, \bar{y})$.

Proof. This result is provided in [10]. □

Lemma C.5 Define $\lambda^{\beta_k}, \lambda^* \in \mathbb{R}^N$ as,

$$\begin{aligned}\lambda^{\beta_k} &= \underset{Y \in X, \lambda_i \in \{\lambda_i \mid \sum_{i \in [N]} \lambda_i = 1\}}{\text{argmin}} \beta_k \sum_{i \in [N]} \frac{\lambda_i}{\mu_i - \lambda_i} + \kappa \min_{\pi \in \Pi} \mathbb{E}_\pi \|x - y_i\|_p^p \\ \lambda^* &= \underset{\lambda_i \in \{\lambda_i \mid \sum_{i \in [N]} \lambda_i = 1\}}{\text{argmin}} \sum_{i \in [N]} \frac{\lambda_i}{\mu_i - \lambda_i}\end{aligned} \tag{C.23}$$

for $k = 1, 2$. Then, for $\beta_1 \leq \beta_2 \in \mathbb{R}^+$,

$$|\lambda_i^{\beta_1} - \lambda_i^*| \geq |\lambda_i^{\beta_2} - \lambda_i^*|, \quad \forall i \in [N]. \tag{C.24}$$

Proof. Let Z be defined as,

$$Z = \min_{Y \in X, \lambda \in \{\lambda \in \mathbb{R}^N \mid \lambda_i < \mu_i, \sum_{i=1}^N \lambda_i = 1\}} \left(\beta \sum_{i=1}^N \frac{\lambda_i}{\mu_i - \lambda_i} + \min_{\pi \in \Pi} \mathbb{E}_\pi [\|x - y_i\|_p^p] \right).$$

From convexity of Z in its arguments,

$$Z = \min_{\lambda_i < \mu_i, \sum_{i=1}^{N-1} \lambda_i \leq 1} \left(\beta \frac{1 - \sum_{i=1}^{N-1} \lambda_i}{\mu_N - (1 - \sum_{i=1}^{N-1} \lambda_i)} + \beta \sum_{i=1}^{N-1} \frac{\lambda_i}{\mu_i - \lambda_i} + \min_{Y \in X} \min_{\pi \in \Pi} \mathbb{E}_\pi [\|x - y_i\|_p^p] \right).$$

Since constraints are not binding, and from Lemma C.4,

$$Z = \min_{\lambda_1 \in \mathbb{R}} \left(\beta \frac{\lambda_1}{\mu_1 - \lambda_1} + \min_{\lambda \in \mathbb{R}^{N-2}} \left(\beta \frac{1 - \sum_{i=1}^{N-1} \lambda_i}{\mu_N - (1 - \sum_{i=1}^{N-1} \lambda_i)} + \beta \sum_{i=2}^{N-1} \frac{\lambda_i}{\mu_i - \lambda_i} + \min_{Y \in X} \min_{\pi \in \Pi} \mathbb{E}_\pi [\|x - y_i\|_p^p] \right) \right)$$

With the same reasoning,

$$Z = \min_{\lambda_1 \in \mathbb{R}^+} \left(\beta \frac{\lambda_1}{\mu_1 - \lambda_2} + \min_{\lambda_2 \in \mathbb{R}^+} \left(\beta \frac{\lambda_2}{\mu_2 - \lambda_2} + \min_{\lambda_3 \in \mathbb{R}^+} \left(\beta \frac{\lambda_3}{\mu_3 - \lambda_3} \right. \right. \right. \\ \left. \left. \left. + \cdots + \underbrace{\left[\min_{\lambda_{N-1} \in \mathbb{R}^+} \beta \frac{1 - \sum_{i=1}^{N-1} \lambda_i}{\mu_N - (1 - \sum_{i=1}^{N-1} \lambda_i)} + \beta \frac{\lambda_{N-1}}{\mu_{N-1} - \lambda_{N-1}} + \min_{Y \in X} \min_{\pi \in \Pi} \mathbb{E}_\pi [\|x - y_i\|_p^p] \right]}_{\text{N-1-th layer level problem}} \right) \right) \right).$$

The above problem is an $N - 1$ layer optimization problem. For the $N - 1$ -th layer problem defined above, the only variable is λ_{N-1} , and by setting $\nu = \lambda_{N-1}$, we can rewrite it as,

$$\min_{\nu \in \mathbb{R}^+} \underbrace{\beta \frac{1 - \sum_{i=1}^{N-1} \lambda_i}{\mu_N - (1 - \sum_{i=1}^{N-1} \lambda_i)} + \frac{\lambda_{N-1}}{\mu_{N-1} - \lambda_{N-1}}}_{F(\nu)} + \underbrace{\min_{Y \in X} \min_{\pi \in \Pi} \mathbb{E}_\pi [\|x - y_i\|_p^p]}_{G(\nu)}. \quad (\text{C.25})$$

We need to prove as β increases, solution of (C.25) gets closer to $\min_{\nu \in \mathbb{R}^+} F(\nu)$. Let, $\nu^*, \nu^{\beta_1}, \nu^{\beta_2} \in \text{dom}(F) \cap \text{dom}(G)$ in the same way as (C.23). Then,

$$\begin{aligned} \beta_1 F(\nu^{\beta_1}) + G(\nu^{\beta_1}) &\leq \beta_1 F(\nu^{\beta_2}) + G(\nu^{\beta_2}) && (\text{From optimality of } \nu^{\beta_1}) \\ \Rightarrow \beta_1 (F(\nu^{\beta_1}) - F(\nu^*)) + G(\nu^{\beta_1}) &\leq \beta_1 (F(\nu^{\beta_2}) - F(\nu^*)) + G(\nu^{\beta_2}) \\ \Rightarrow \beta_1 (F(\nu^{\beta_1}) - F(\nu^*)) &\leq \beta_1 (F(\nu^{\beta_2}) - F(\nu^*)) && (\text{Since } G(\nu^{\beta_2}) - G(\nu^{\beta_1}) \geq 0) \\ \Rightarrow |\nu^{\beta_1} - \nu^*| &\geq |\nu^{\beta_2} - \nu^*| && (\text{By convexity of } F(\nu)) \end{aligned}$$

A similar logic can be utilized for all i -layer problems for $1 \leq i < N - 1$ by defining $G(\nu)$ as the solution of “ $i - 1$ ”-layer problem and $\nu = \lambda_i$. This concludes the proof. $\square$

We next show there exists a value for β such that the optimal solution of (C.22) is equal to the optimal solution of $\sum_{i \in [N]} f_i(x_i)$.

Lemma C.6 *Let $f(x) : \mathbb{R}^N \rightarrow \mathbb{R}$ be a strict convex function such that for any $x \in \text{dom}(f)$, $f(x)$ is smooth, $\nabla f \succ 0$, and $-\infty < \text{argmin } f(x) < \infty$. Also, let $g(x) : \mathbb{R}^N \rightarrow \mathbb{R}$ be a Lipschitz continuous function. Then, for any given κ , there exists a $\bar{\beta}$ such that for all $\beta \geq \bar{\beta}$, we have*

$$\text{argmin } \beta f(x) + \kappa g(x) = \text{argmin } f(x). \quad (\text{C.26})$$

Proof. First, we derive a lower-bound for the inner products of two vectors. Let, $z, y \in \mathbb{R}^N$ be two arbitrary vectors and $Mm \geq 0$ such that,

$$\langle My - z, z - my \rangle \geq 0.$$

Then,

$$\begin{aligned} M \langle y, z \rangle - m \langle y, z \rangle - \|z\|_2^2 - Mm\|y\|_2^2 &\geq 0 \\ \Rightarrow Mm\|y\|_2^2 + \|z\|_2^2 &\leq (M - m) \langle y, z \rangle \\ \Rightarrow \frac{1}{\sqrt{Mm}}\|z\|_2^2 + \sqrt{Mm}\|y\|_2^2 &\leq \frac{M - m}{\sqrt{Mm}} \langle y, z \rangle. \end{aligned}$$

By the elementary inequality, we know that the geometric average is always greater than or equal to the arithmetic average. Thus,

$$\begin{aligned} \|z\|_2\|y\|_2 &\leq \frac{1}{2} \left(\frac{1}{\sqrt{Mm}}\|z\|_2^2 + \sqrt{Mm}\|y\|_2^2 \right) \\ \Rightarrow 2\|z\|_2\|y\|_2 &\leq \frac{1}{\sqrt{Mm}}\|z\|_2^2 + \sqrt{Mm}\|y\|_2^2 \leq \frac{M - m}{\sqrt{Mm}} \langle y, z \rangle. \end{aligned}$$

Therefore,

$$\frac{2\sqrt{Mm}}{M - m} \|z\|_2\|y\|_2 \leq \langle y, z \rangle. \quad (\text{C.27})$$

Now, we prove (C.26). Let $\bar{x} = \operatorname{argmin} f(x)$ and for the sake of contradiction, assume that there exists an $x \in \operatorname{dom}(f(x)) \cap \operatorname{dom}(g(x))$ such that (C.26) does not hold. Then,

$$\begin{aligned} \beta f(x) + \kappa g(x) &< \beta f(\bar{x}) + \kappa g(\bar{x}) \\ \Rightarrow \beta(f(x) - f(\bar{x})) &< \kappa(g(\bar{x}) - g(x)) \\ \Rightarrow \beta \langle \nabla f(\bar{x}), x - \bar{x} \rangle &< \kappa(g(\bar{x}) - g(x)) && (\text{by } f \text{ being convex and smooth at } \bar{x}) \\ \Rightarrow \langle \nabla f(\bar{x}), x - \bar{x} \rangle &< \kappa K_2 \|\bar{x} - x\|_2 && (\text{by Lipschitz continuity of } g(x)) \\ \Rightarrow \beta \frac{2\sqrt{mM}}{M + m} \|\nabla f(\bar{x})\|_2 \|\bar{x} - x\|_2 &< \kappa K_2 \|\bar{x} - x\|_2 && (\text{based on (C.27)}) \\ \Rightarrow \beta &< \frac{\kappa(M + m)}{2\|\nabla f(\bar{x})\|_2 \sqrt{mM}} K_2, \end{aligned}$$

where m and M are positive values satisfying $(M\nabla f - x)^T(x - m\nabla f) \geq 0$. Note that such M and m exist since ∇f and x are positive in all directions. Thus, for the choice of

$\beta > \frac{\kappa(M+m)}{2\|\nabla f(\bar{x})\|_2\sqrt{mM}}K_2$, we have a contradiction. This completes the proof. $\square$

Next, we derive the optimal solution of $\sum_{i \in [N]} f_i(x)$ when $f_i = \frac{\lambda_i}{\mu_i - \lambda_i}$.

Lemma C.7 *The optimal solution of,*

$$\begin{aligned} & \min_{\lambda_i} \sum_{i \in [N]} \left(\beta \frac{\lambda_i}{\mu_i - \lambda_i} \right) \\ & \text{Subject to:} \\ & \lambda_i \geq 0, \quad \forall i \in [N], \\ & \sum_{i \in [N]} \lambda_i = 1, \end{aligned}$$

is given by,

$$\lambda_i = \mu_i - \xi \sqrt{\mu_i}, \quad \forall i \in [N], \quad (\text{C.28})$$

where $\xi = \frac{\sum_{i \in [N]} \mu_i - 1}{\sum_{i \in [N]} \sqrt{\mu_i}}$.

Proof. Since the problem is convex, it is sufficient to show that (C.28) is a KKT point and that a constraint qualification holds at the point. First, note that the Lagrangian function of the problem is given by,

$$L(\lambda_i, \bar{\xi}, \eta_i) = \sum_{i \in [N]} \frac{\lambda_i}{\mu_i - \lambda_i} + \bar{\xi} \left(1 - \sum_{i \in [N]} \lambda_i \right) - \sum_{i \in [N]} \eta_i \lambda_i, \quad (\text{C.29})$$

for $\eta_i \in \mathbb{R}^+$ and $\bar{\xi} \in \mathbb{R}$. Setting the gradient of (C.29) with respect to λ_i to 0 will result in,

$$\begin{aligned} & \nabla_{\lambda_i} L(\lambda_j, \bar{\xi}, \eta_j) = 0 \\ & \Rightarrow \frac{\mu_i}{(\mu_i - \lambda_i)^2} - \bar{\xi} - \eta_i = 0 \\ & \Rightarrow \lambda_i = \mu_i - \sqrt{\frac{\mu_i}{\bar{\xi} - \eta_i}}. \end{aligned}$$

From the constraint $\sum_{i \in [N]} \lambda_i = 1$, we get

$$\begin{aligned} \sum_{i \in [N]} \mu_i - \sum_{i \in [N]} \sqrt{\frac{\mu_i}{\bar{\xi} - \eta_i}} &= 1 \\ \Rightarrow \bar{\xi} - \eta_i &= \frac{(\sum_{i \in [N]} \sqrt{\mu_i})^2}{(\sum_{i \in [N]} \mu_i - 1)^2}. \end{aligned}$$

From the complementary slackness, we know that $\eta_i \neq 0$ if and only if $\lambda_i = 0$. Thus,

$$\lambda_i^* = \mu_i - \sqrt{\frac{\mu_i}{\bar{\xi}}}$$

where $\bar{\xi} = \frac{(\sum_{i \in [N]} \sqrt{\mu_i})^2}{(\sum_{i \in [N]} \mu_i - 1)^2}$. Furthermore, λ_i^* is strictly greater than 0, and thus, Slater's condition holds. This concludes that the KKT condition and constrained qualifications are satisfied. Thus, by the strict convexity of the problem, λ_i^* is the global optimum solution. Setting $\xi = \frac{1}{\sqrt{\bar{\xi}}}$ completes the proof. $\square$

Proof of Proposition 4.4

Proof. First, note that the total travel cost, $C_T(\lambda, Y)$, is Lipschitz continuous and convex because it is a distance over space of all probability measures over X , $P(X)$. Furthermore, $\frac{\lambda_i}{\mu_i - \lambda_i}$ is smooth and convex in λ_i for $\mu_i > \lambda_i$.

Moreover, given the optimal tuple of design variables (λ, μ, Y) and the optimal mapping π^* , from Proposition 4.1, as discussed in the proof of Lemma C.3, we have:

$$\min_{\lambda_i, \mu_i, y_i} \sum_{i \in [N]} \left(h(\mu_i) + \beta \frac{\lambda_i}{\mu_i - \lambda_i} + \kappa \int_{X_i} \|x - y_i\|_p^p d\mathbb{P} \right).$$

Using Lemmas C.5, C.6 and C.7, and considering that $h(\mu_i)$ is a constant for fixed μ_i , we can complete the proof of Proposition 4.4:

1. The first part of the proposition is immediate from Lemmas C.7 and C.5.
2. Setting $f(\lambda_i) = \sum_{i \in [N]} \beta \frac{\lambda_i}{\mu_i - \lambda_i}$ and $g(\lambda_i) = \min_{\pi \in \Pi} \mathbb{E}_\pi \|x - y_i\|_p^p$, the proof is immediate from Lemmas C.6 and C.5.

3. It is easy to see,

$$\Psi(\{\lambda_i\}_{i \in [N]}) = \frac{1}{N} \|\boldsymbol{\lambda}\|_2^2 - \|\boldsymbol{\lambda}_i\|_1^2.$$

Likewise,

$$\Psi(\{\sqrt{\lambda_i}\}_{i \in [N]}) = \frac{1}{N} (\|\boldsymbol{\lambda}\|_1 - \|\boldsymbol{\lambda}_i\|_{1/2}).$$

Thus,

$$\Psi(\{\lambda_i^*\}_{i \in [N]}) = \frac{1}{N} (\|\boldsymbol{\mu}\|_2^2 + \xi^2 \|\boldsymbol{\mu}\|_{1/2} - (\xi^2 + \|\boldsymbol{\mu}\|_1) \|\boldsymbol{\mu}\|_1).$$

□

Proof of Corollary 4.2

Proof. The proof of Corollary 4.2 is immediate from Proposition 4.4 by setting $\mu_i = \mu$.
□

Optimal service rate

In this subsection, similar to the previous subsection, leveraging the block structure of the design problem, we prove a relation between the service rate and the optimal arrival rate. To do this, we start by introducing a necessary optimality condition for the following problem,

$$\min_{x \in \mathbb{R}^N} f(x) + g(x), \tag{C.30}$$

where $f(x)$ is smooth but non-convex and $g(x)$ is convex but non-smooth. Note that the design problem given in (3.7) can be written as (C.30) by setting $f(x) = \sum_{i \in [N]} \left(h(\mu_i) + \frac{\lambda_i}{\mu_i - \lambda_i} \right)$ and $g(x) = \min_{\pi \in \Pi} \|x - y_i\|_p^p$.

Lemma C.8 *Let $f : \mathbb{R}^N \rightarrow \bar{\mathbb{R}}$ and $g : \mathbb{R}^N \rightarrow \bar{\mathbb{R}}$ be two proper, lower semi-continuous functions such that domain of f is a subset of the interior of the domain of g . Furthermore, assume that g is convex and f is smooth, and let x^* be a local minimizer of $f(x) + g(x)$. Then,*

$$-\nabla f(x^*) \in \partial g(x^*).$$

Proof. Let x^* be a local optimal of the problem (C.30). Thus, from convexity of $g(x)$, for a small enough η ,

$$\begin{aligned} f(x^*) + g(x^*) &\leq f(x^*) + \eta \langle x - x^*, \nabla f \rangle + (1 - \eta)g(x^*) + \eta g(x) \\ \Rightarrow \eta(g(x^*) - g(x)) &\leq \eta \langle x - x^*, \nabla f(x^*) \rangle \\ \Rightarrow -\nabla f(x^*) &\in g(x^*), \end{aligned}$$

which concludes the proof. $\square$

Proof of Proposition 4.5

Proof. To prove Proposition 4.5, we define $f(\boldsymbol{\lambda}, \boldsymbol{\mu}, Y)$ and $g(\boldsymbol{\lambda}, \boldsymbol{\mu}, Y)$ as follows:

$$f(\boldsymbol{\lambda}, \boldsymbol{\mu}, Y) = \sum_{i \in [N]} \left(\beta \frac{\lambda_i}{\mu_i - \lambda_i} + h(\mu_i) \right),$$

and,

$$g(\boldsymbol{\lambda}, \boldsymbol{\mu}, Y) = \chi_\Lambda + \min_{\pi \in \Pi} \mathbb{E}_\pi \|x - y_i\|_p^p,$$

where χ_Λ is the indicator function of Λ defined as,

$$\Lambda = \{ \boldsymbol{\lambda} \in \mathbb{R}^N | \lambda_i \geq 0, \sum_{i \in [N]} \lambda_i = 1 \}. \quad (\text{C.31})$$

Note that $g(\boldsymbol{\lambda}, \boldsymbol{\mu}, Y)$ is convex since $\min_{\pi \in \Pi} \mathbb{E}_\pi \|x - y_i\|_p^p$ is convex and Λ is a convex set. Furthermore, $g(\boldsymbol{\lambda}, \boldsymbol{\mu}, Y)$ is continuous and, in particular, lower semi-continuous. Also, $f(\boldsymbol{\lambda}, \boldsymbol{\mu}, Y)$ is everywhere differentiable in its domain which is $\mathbb{R}^2$. Thus, by Lemma C.8, if $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*, Y^*)$ is an optimal tuple of the design problem, it satisfies

$$-\nabla f(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*, Y^*) \in \partial g(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*, Y^*).$$

Since $g(\boldsymbol{\lambda}, \boldsymbol{\mu}, Y)$ is constant in $\boldsymbol{\mu}$, the subgradient of $g(\boldsymbol{\lambda}, \boldsymbol{\mu}, Y)$ has the following structure,

$$\partial g(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*, Y^*) = \{ (z_1, \dots, z_n, 0, \dots, 0, z_{2N+1}, \dots, z_{3N}) | z_i \in \mathbb{R} \}.$$

Thus, we have,

$$\nabla f(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*, Y^*) = 0 \Rightarrow \frac{\partial h(\mu_i)}{\partial \mu_i} \Big|_{\mu=\mu_i^*} - \beta \frac{\lambda_i^*}{(\mu_i^* - \lambda_i^*)^2} = 0, \quad (\text{C.32})$$

which concludes the first part of the proposition.

To prove the uniqueness of the $\boldsymbol{\mu}^*$, note that for any given $t \in \mathbb{R}^+$, the solution of $\beta \frac{\lambda_i^*}{(\mu_i^* - \lambda_i^*)^2} = t$ is given by

$$\mu_i^* = \lambda_i^* \pm \sqrt{\frac{\beta}{t} \lambda_i^*}, \quad \forall i \in [N].$$

Sine μ_i^* should be greater than λ_i^* (due to the stability condition of the queue at location i), we have

$$\mu_i^* = \lambda_i^* + \sqrt{\frac{\beta}{t} \lambda_i^*}, \quad \forall i \in [N].$$

Considering that $\frac{\partial h(\mu_i)}{\partial \mu_i} \Big|_{\mu_i=\mu_i^*}$ is monotone (by assumption), there exists only one value for μ_i^* such that $\frac{\partial h(\mu)}{\partial \mu} \Big|_{\mu=\mu_i^*} = t$. Thus, for any given λ_i^* , μ_i^* is unique. $\square$

Proof Corollary 4.3

Proof. We first show that $\mu_i^*(\lambda_i)$ is increasing in λ_i . Recall that from Proposition 4.5, we have

$$h'(\mu^*(\lambda_i)) = \beta \frac{\lambda_i}{(\mu^*(\lambda_i) - \lambda_i)^2}. \quad (\text{C.33})$$

Taking the implicit derivative with respect to λ_i from (C.33) results in,

$$\begin{aligned} h'(\mu^*(\lambda_i)) \frac{d\mu^*(\lambda_i)}{d\lambda_i} &= \frac{(\mu^*(\lambda_i) - \lambda_i)^2 - 2(\mu^*(\lambda_i) - \lambda_i) \left(\frac{d\mu^*(\lambda_i)}{d\lambda_i} - 1 \right) \lambda_i}{(\mu^*(\lambda_i) - \lambda_i)^4} \\ \Rightarrow \frac{d\mu^*(\lambda_i)}{d\lambda_i} &= \frac{\mu^*(\lambda_i) + \lambda_i}{h'(\mu^*(\lambda_i))(\mu^*(\lambda_i) - \lambda_i)^3 + 2\lambda_i}. \end{aligned}$$

Note that the right-hand side of the last equality is positive by $\mu_i > \lambda_i$ and $h(\mu_i)$ being increasing and positive. Thus, $\mu^*(\lambda_i)$ is increasing in λ_i .

Now, we show Parts 1-3 of Corollary 4.3. From Proposition 4.5, for the linear cost

$h(\mu_i) = \alpha\mu_i$, we have

$$\begin{aligned}\alpha &= \beta \frac{\lambda_i}{(\mu^*(\lambda_i) - \lambda_i)^2} \\ \Rightarrow \mu^*(\lambda_i) &= \lambda_i^* \pm \sqrt{\frac{\beta}{\alpha} \lambda_i} \\ \Rightarrow \mu^*(\lambda_i) &= \lambda_i + \sqrt{\frac{\beta}{\alpha} \lambda_i} \quad (\text{By } \mu_i > \lambda_i).\end{aligned}$$

From Proposition 4.5, for the logarithmic cost function $h(\mu_i) = \log(\gamma_1\mu_i + \gamma_2)$, we get

$$\frac{1}{\mu^*(\lambda_i) + \gamma_1/\gamma_2} = \beta \frac{\lambda_i}{(\mu^*(\lambda_i) - \lambda_i)^2}.$$

Let $\mu^*(\lambda_i) - \lambda_i = \delta_i$, then

$$\begin{aligned}\frac{1}{\delta_i + \lambda_i + \gamma_1/\gamma_2} &= \frac{\beta\lambda_i}{\delta_i^2} \\ \Rightarrow -\frac{\delta_i^2}{\beta\lambda_i} + \delta_i + \lambda_i + \gamma_1/\gamma_2 &= 0 \\ \Rightarrow \mu^*(\lambda_i) &= \lambda_i \left(1 + \frac{\beta}{2\gamma_2}\right) + \sqrt{\frac{\beta}{\gamma_2} \lambda_i} \sqrt{\frac{1}{\gamma_1} + \lambda_i \left(1 + \frac{1}{2\beta\gamma_2}\right)}.\end{aligned}$$

From Proposition 4.5, for the quadratic cost function $h(\mu_i) = \alpha\mu_i^2$, we have

$$\begin{aligned}2\alpha\mu_i &= \beta \frac{\lambda_i}{(\mu^*(\lambda_i) - \lambda_i)^2} \\ \Rightarrow -2\alpha(\mu^*(\lambda_i))^3 + 4\alpha\lambda_i(\mu^*(\lambda_i))^2 - 2\alpha\lambda_i^2\mu^*(\lambda_i) + \beta\lambda_i &= 0.\end{aligned}$$

The only real answer to this third-degree equation is given by,

$$\begin{aligned}
\mu^*(\lambda_i) &= \frac{1}{3} \left(\frac{2^{2/3} \alpha \lambda_i^2}{\sqrt[3]{-4\alpha^3 \lambda_i^3 + 27\alpha^2 \beta \lambda_i + 3\sqrt{3} \sqrt{27\alpha^4 \beta^2 \lambda_i^2 - 8\alpha^5 \beta \lambda_i^4}}} + \right. \\
&\quad \left. \frac{\sqrt[3]{-4\beta^3 \lambda_i^3 + 27\alpha^2 \beta \lambda_i + 3\sqrt{3} \sqrt{27\alpha^4 \beta^2 \lambda_i^2 - 8\alpha^5 \beta \lambda_i^4}}}{2^{2/3} \alpha} + 2\lambda_i \right) \\
&= \frac{2}{3} \lambda_i + \frac{1}{3} \lambda_i^{5/3} \left(\frac{27}{4} \alpha^{-1} \beta - \lambda_i^2 + \sqrt{\frac{3^{15/4}}{16} \alpha^{-2} \beta^2 - \frac{3^{3/4}}{2} \alpha \beta \lambda_i^2} \right)^{-1/3} + \\
&\quad \frac{1}{3} \lambda_i^{1/3} \left(\frac{27}{4} \alpha^{-1} \beta - \lambda_i^2 + \sqrt{\frac{3^{15/4}}{16} \alpha^{-2} \beta^2 - \frac{3^{3/4}}{2} \alpha \beta \lambda_i^2} \right)^{1/3} \\
&= \frac{2}{3} \lambda_i + \frac{1}{3} \lambda_i^{5/3} \left(\frac{27}{4} \alpha^{-1} \beta - \lambda_i^2 + \sqrt{\frac{3^{3/4}}{2} \alpha \beta} \sqrt{\frac{3^3}{8} \alpha^{-3} \beta - \lambda_i^2} \right)^{-1/3} + \\
&\quad \frac{1}{3} \lambda_i^{1/3} \left(\frac{27}{4} \alpha^{-1} \beta - \lambda_i^2 + \sqrt{\frac{3^{3/4}}{2} \alpha \beta} \sqrt{\frac{3^3}{8} \alpha^{-3} \beta - \lambda_i^2} \right)^{1/3},
\end{aligned}$$

which by choosing,

$$c_0 = \frac{27\beta}{4\alpha}, \quad c_1 = \sqrt{\frac{3^{3/4}}{2} \alpha \beta}, \quad c_2 = \left(\frac{3}{2\alpha}\right)^3 \beta, \quad (\text{C.34})$$

completes the proof. $\square$

Proof Corollary 4.4

Proof. From Proposition 4.5, we have

$$\mu_i^* - \lambda_i = \sqrt{\lambda_i \frac{\beta}{\frac{\partial h(\mu_i)}{\partial \mu_i} \big|_{\mu_i = \mu_i^*(\lambda_i)}}}.$$

Thus, the total waiting cost at location i can be written as

$$\beta \frac{\lambda_i}{\mu_i^* - \lambda_i} = \beta \frac{\lambda_i}{\sqrt{\lambda_i \frac{\partial h(\mu_i)}{\partial \mu_i} \Big|_{\mu_i = \mu^*(\lambda_i)}}} = \sqrt{\beta \lambda_i \frac{\partial h(\mu_i)}{\partial \mu_i} \Big|_{\mu_i = \mu^*(\lambda_i)}}.$$

□

Optimal Demand Rate and Service Capacity

From Proposition 4.5, we can rewrite the first two terms of the design problem given in (3.7), namely the capacity and waiting costs, as

$$\sum_{i \in [N]} \left(h(\mu_i^*) + \beta \frac{\lambda_i}{\mu_i^* - \lambda_i} \right) = \sum_{i \in [N]} \left(h(\mu^*(\lambda_i)) + \sqrt{\beta \lambda_i \frac{\partial h(\mu_i)}{\partial \mu_i} \Big|_{\mu_i = \mu^*(\lambda_i)}} \right).$$

Similar to the previous section, we use the block structure of the problem to prove Proposition 4.6.

Now, we determine the optimal value of $\boldsymbol{\lambda}$ in the case that customers are not sensitive to travel costs and the total waiting cost at each location is concave in $\boldsymbol{\lambda}$.

Lemma C.9 *Let $g(\lambda_i)$ be a concave function and Λ be the set defined in (C.31). Then, the global optimal points of the problem,*

$$\min_{\boldsymbol{\lambda} \in \Lambda} \sum_{i \in [N]} g(\lambda_i) \tag{C.35}$$

are given by $\lambda_i = 1$ for an $i \in [N]$ and $\lambda_j = 0$ for all $i \neq j$.

Proof. Let v_i be the vertex of the polytope Λ . Then, any point inside of $\boldsymbol{\lambda} \in \Lambda$ can be written as $\boldsymbol{\lambda} = \sum_{i \in [N]} \eta_i v_i$. Thus, by concavity of $g(\cdot)$,

$$g(\lambda_i) = g \left(\sum_i \eta_i v_i \right) \geq \sum_i \eta_i g(v_i).$$

Hence, the minimum is at one of the vertices. Note that vertices of this polytope are in the form $\lambda_i = 1$ for a fixed i and $\lambda_j = 0$ for all $i \neq j$. Since all the vertices correspond to the same objective value, then global optimal solutions of (C.35) are given by $\lambda_i = 1$ for an $i \in [N]$ and $\lambda_j = 0$ for all $i \neq j$. $\square$

To prove Proposition 4.6, similar to the case where service rates are fixed, we show that there exists a value for β such that the optimal solution of the problem $\beta \sum_{i \in [N]} f_i(x) + \kappa g(x)$ is equal to the optimal solution of $\sum_{i \in [N]} f_i(x_i)$. Moreover, there exists a value for κ such that the optimal solution of the problem $\beta \sum_{i \in [N]} f_i(x) + \kappa g(x)$ is equal to the optimal solution of $g(x)$.

Lemma C.10 *Let $f_0(x) : [0, 1] \rightarrow \mathbb{R}^+$ be a twice differentiable concave function such that $f_0(0) = 0$ and $f'_0(0) \neq 0$. Furthermore, let $g(x) : \mathbb{R}^N \rightarrow \mathbb{R}$ and $h_0(x) : \mathbb{R} \rightarrow \mathbb{R}^+$ be K_1 and K_2 -Lipschitz continuous functions, respectively, such that $\operatorname{argmin}_{x \in \{x \in \mathbb{R}^N \mid \|x\|_1 = 1\}} g(x) > 0$. Furthermore, let $f(x) : [0, 1]^N \rightarrow \mathbb{R}^+$ and $h(x) : [0, 1]^N \rightarrow \mathbb{R}^+$, $f(x) = \sum_{i \in [N]} f_0(x_i)$ and $h(x) = \sum_{i \in [N]} h_0(x_i)$. Then,*

1. *for any given β there exists a $\bar{\beta}$ such that for all $\beta \geq \bar{\beta}$,*

$$\operatorname{argmin}_{x \in \{x \in \mathbb{R}^N \mid \|x\|_1 = 1\}} \sum_{i \in [N]} h_0(x_i) + \beta \sum_{i \in [N]} f_0(x_i) + \kappa g(x) = \operatorname{argmin}_{x \in \{x \in \mathbb{R}^N \mid \|x\|_1 = 1\}} \sum_{i \in [N]} f_0(x_i).$$

2. *for any given β there exists a $\bar{\kappa}$ such that for all $\kappa \geq \bar{\kappa}$,*

$$\operatorname{argmin}_{x \in \{x \in \mathbb{R}^N \mid \|x\|_1 = 1\}} h(x) + \beta f(x) + \kappa g(x) = \operatorname{argmin}_{x \in \{x \in \mathbb{R}^N \mid \|x\|_1 = 1\}} g(x).$$

Proof. Let $\bar{x} = \operatorname{argmin} \sum f_0(x)$ and $\hat{x} = \operatorname{argmin} g(x)$. Note that from Lemma C.9, we know that $\bar{x}_i = 1$ for one $i \in [N]$ and $\bar{x}_j = 0$ for all $j \neq i$.

For the first part of the lemma, we proceed using contradiction. Assume that there exists an $\epsilon \succ 0$ such that for $x = (1 - \sum_{i=2}^N \epsilon_i, \epsilon_1, \epsilon_2, \epsilon_3, \dots, \epsilon_N) = (1, 0, \dots, 0) - \epsilon$ where,

$$\sum_{i \in [N]} h_0(x_i) + \beta \sum_{i \in [N]} f_0(x_i) + \kappa g(x) < \sum_{i \in [N]} h_0(\bar{x}_i) + \beta \sum_{i \in [N]} f_0(\bar{x}_i) + \kappa g(\bar{x}).$$

Then,

$$\begin{aligned}
& h(x) + \beta f(x) + \kappa g(x) < \beta f(\bar{x}) + \kappa g(\bar{x}) + h(\bar{x}) \\
& \Rightarrow \beta(f(x) - f(\bar{x})) < \kappa(g(\bar{x}) - g(x)) + h(\bar{x}) - h(x) \\
& \Rightarrow \beta(f_0(1 - 2 \sum_{i=2}^N \epsilon_i) + \sum_{i=2}^N f_0(\epsilon_i) - f_0(1)) < \kappa(g(\bar{x}) - g(x)) + h(\bar{x}) - h(x) \quad (\text{Definition of } f) \\
& \Rightarrow \beta(f_0(1 - \sum_{i=2}^N \epsilon_i) + \sum_{i=2}^N f_0(\epsilon_i) - f_0(1)) < (\kappa K_2 + K_1) \|\epsilon\|_1 \quad (\text{Lipschitz cont. of } h \text{ and } f) \\
& \Rightarrow \beta \left(\sum_{i=2}^N \left(\frac{\epsilon_i}{1 - \epsilon_i} f_0(\epsilon_i) + \frac{1}{N} (1 - \frac{\epsilon_i}{1 - \epsilon_i}) f_0(1) \right) - f_0(1) \right) < (K_1 + \kappa K_2) \|\epsilon\|_1 \quad (\text{Concavity of } f_0(x_i)) \\
& \Rightarrow \beta \sum_{i=2}^N (f(\epsilon_i) + \frac{\epsilon_i}{\epsilon_i - 1} (f(1) - f(\epsilon_i))) < 2(K_1 + \kappa K_2) \sum_{i=2}^N \epsilon_i \quad (\|\epsilon\|_1 = 2 \sum_{i=2}^N \epsilon_i) \\
& \beta < \frac{\sum_{i=2}^N \epsilon_i}{\sum_{i=2}^N \left(f(\epsilon_i) + \frac{\epsilon_i}{\epsilon_i - 1} (f(1) - f(\epsilon_i)) \right)} (\kappa K_2 + K_1)
\end{aligned}$$

Thus, for any β such that

$$\beta > \frac{\sum_{i=2}^N \epsilon_i}{\sum_{i=2}^N \left(f(\epsilon_i) + \frac{\epsilon_i}{\epsilon_i - 1} (f(1) - f(\epsilon_i)) \right)} (\kappa K_2 + K_1)$$

we have a contradiction. Such a β always exists for a finite ϵ .

For the second part of the lemma, we proceed with contradiction. Let $\hat{x} = \operatorname{argmin} g(x)$. and assume that there exists a $\hat{x} + \epsilon r = x \in \mathbb{R}$ where $\|r\|_2 = 1$ such that the statement does not hold. Then for a small enough ϵ ,

$$\begin{aligned}
& h(x) + \beta f(x) + \kappa g(x) < \beta f(\hat{x}) + \kappa g(\hat{x}) \\
& \Rightarrow \beta(f(x) - f(\hat{x})) < \kappa(g(\hat{x}) - g(x)) + (h(\hat{x}) - h(x)) \\
& \Rightarrow \beta \epsilon \nabla g(\hat{x})^T r < \kappa(g(\hat{x}) - g(x)) + (g(\hat{x}) - g(x)) \quad (\text{Smoothness of } g) \\
& \Rightarrow 2 \frac{\sqrt{mM}}{M + m} \epsilon \beta \|\nabla f(x)\|_2 < (\kappa K_2 + K_1) \epsilon \quad (\text{Lipschitz continuity of } h \text{ and (C.27)}) \\
& \beta < \kappa \frac{M + m}{2\sqrt{mM} \|\nabla f(x)\|_2} (\kappa K_2 + K_1)
\end{aligned}$$

where m and M are any positive value that satisfy $(M\nabla f - x)^T(x - m\nabla f) \geq 0$. Thus, for the choice of $\kappa > (\frac{M+m}{2\sqrt{mM}\|\nabla f(x)\|_1}K_2)^{-1}\beta$, we have a contradiction, considering that $\nabla f(\hat{x})$ is bounded around $\hat{x}$. $\square$

Next, we determine the optimal value of λ in the case that customers are not sensitive to the travel cost, and the total waiting cost at each location is convex in λ .

Lemma C.11 *Let $g(\lambda_i)$ be a convex increasing function and Λ as defined in (C.31). Then, the global optimal solution of*

$$\min_{\lambda \in \Lambda} \sum_{i \in [N]} g(\lambda_i) \quad (\text{C.36})$$

is given by $\lambda_i = 1/N$ for $i \in [N]$ and $\lambda_j = 0$ for all $i \neq j$.

Proof. The Lagrangian function of (C.36) is given by,

$$L(\lambda, \eta) = \sum_{i \in [N]} g(\lambda_i) + \eta(1 - \sum_{i \in [N]} \lambda_i).$$

Thus, the stationary condition is satisfied by,

$$\nabla_{\lambda} L(\lambda, \eta) = 0 \Rightarrow \frac{\partial}{\partial \lambda_i} g(\lambda_i) = \eta. \quad (\text{C.37})$$

Since $g(\lambda_i)$ is convex, $\frac{\partial}{\partial \lambda_i} g(\lambda_i)$ is increasing. Thus, (C.37) has a unique and constant solution for all $i \in [N]$, that is,

$$\lambda_1 = \lambda_2 = \dots = \lambda_N.$$

But, $\sum_{i \in [N]} \lambda_i = 1$. Thus, KKT conditions are satisfied only for $\lambda_i = \frac{1}{N}$. Due to the convexity of the problem, this is enough for the optimality of the solution. $\square$

Proof of Proposition 4.6

Proof. The existence of $W_1(\beta)$ and $T_1(\beta)$ falls under Lemma C.10 and C.11. The existence of $W_2(\beta)$ and $T_2(\beta)$ falls under Lemma C.6 and C.11.

To show T_1 is convex, we show that $R_T = \{(\beta, \kappa) | \kappa \geq T_1(\beta)\}$ is convex, which concludes the convexity of T_1 . To show this, let $(\beta_1, \kappa_1) \in R_T$, $(\beta_2, \kappa_2) \in R_T$ and

$(\bar{\lambda}_i, \bar{y}_i) = \operatorname{argmin}_{\lambda_i, y_i} \mathbb{E}_\pi[\|x - y_i\|_p^p]$, which are the optimal arrival rates and location in Region II of Figure 4.5. Then, for a $\eta \in [0, 1]$, we have

$$\begin{aligned} & (\eta\beta_1 + (1 - \eta)\beta_2)C_W(\lambda_i) + (\eta\kappa_1 + (1 - \eta)\kappa_2)C_T(\lambda_i, y_i) \\ &= \eta(\beta_1 C_W(\lambda_i) + \kappa_1 C_T(\lambda_i, y_i)) + (1 - \eta)(\beta_2 C_W(\lambda_i) + \kappa_2 C_T(\lambda_i, y_i)) \\ &\geq \eta(\beta_1 C_W(\bar{\lambda}_i) + \kappa_1 C_T(\bar{\lambda}_i, \bar{y}_i)) + (1 - \eta)(\beta_2 C_W(\bar{\lambda}_i) + \kappa_2 C_T(\bar{\lambda}_i, \bar{y}_i)) \\ &= (\eta\beta_1 + (1 - \eta)\beta_2)C_W(\bar{\lambda}_i) + (\eta\kappa_1 + (1 - \eta)\kappa_2)C_T(\bar{\lambda}_i, \bar{y}_i). \end{aligned}$$

Thus $(\eta\beta_1 + (1 - \eta)\beta_2, \eta\kappa_1 + (1 - \eta)\kappa_2) \in R_T$ and R_T is convex. Note that if $\beta = 0$, for all $\kappa > 0$, $(0, \kappa) \in R_T$. Thus, the unique minimum of $T_1(\beta)$ is on $\beta = 0$, this, due to convexity of T_1 , concludes that $T_1 : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ is increasing.

To show that W_1 is concave, we show that the area below it, $R_W = \{(\beta, \kappa) | \kappa \leq W_1(\beta)\}$ is convex. Let $(\beta_1, \kappa_1) \in R_W$, $(\beta_2, \kappa_2) \in R_W$ and $(\hat{\lambda}_i, \hat{\mu}_i) = \operatorname{argmin}_{\lambda_i, \mu_i} \sum_{i \in [N]} \frac{\lambda_i}{\mu_i - \lambda_i}$ which are the optimal arrival rates and location in Region I of Figure 4.5. Then, for a $\eta \in [0, 1]$, we have,

$$\begin{aligned} & (\eta\beta_1 + (1 - \eta)\beta_2)C_W(\lambda_i) + (\eta\kappa_1 + (1 - \eta)\kappa_2)C_T(\lambda_i, y_i) \\ &= \eta(\beta_1 C_W(\lambda_i) + \kappa_1 C_T(\lambda_i, y_i)) + (1 - \eta)(\beta_2 C_W(\lambda_i) + \kappa_2 C_T(\lambda_i, y_i)) \\ &\geq \eta(\beta_1 C_W(\hat{\lambda}_i) + \kappa_1 C_T(\hat{\lambda}_i, \hat{y}_i)) + (1 - \eta)(\beta_2 C_W(\hat{\lambda}_i) + \kappa_2 C_T(\hat{\lambda}_i, \hat{y}_i)) \\ &= (\eta\beta_1 + (1 - \eta)\beta_2)C_W(\hat{\lambda}_i) + (\eta\kappa_1 + (1 - \eta)\kappa_2)C_T(\hat{\lambda}_i, \hat{y}_i). \end{aligned}$$

Thus $(\eta\beta_1 + (1 - \eta)\beta_2, \eta\kappa_1 + (1 - \eta)\kappa_2) \in R_W$ and R_W is convex. From Lemma C.10, we know that for any κ there exists a finite $\bar{\beta} \geq \xi\kappa$ for some $\xi \in \mathbb{R}^+$ such that for $\beta \geq \bar{\beta}$, $(\beta, \kappa) \in R_W$. Thus, as $\kappa \rightarrow \infty$, $\bar{\beta} \rightarrow \infty$. Hence, the maximum of W_1 is achieved at $\beta = \infty$; this, due to the concavity of W_1 , concludes that W_1 is increasing.

The same reasoning holds for the convexity of $T_2(\beta)$ and the concavity of $W_2(\beta)$. $\square$

C.3 Proof of Results on Optimal Pricing

We start by providing an optimality condition for the regions, defined in Proposition 4.1.

Theorem C.1 *Consider the regions defined in Proposition 4.1. Then, at any local optimal*

points of λ_i, y_i, μ_i , we have,

$$\chi_i = \left\{ x \in X \left| \|x - y_i^*\|_p^p + \beta \lambda_i^* \frac{\partial EW(\lambda_i^*)}{\partial \lambda_i} \right|_{\lambda_i = \lambda_i^*} + \beta EW(\lambda_i) \leq \right. \\ \left. \|x - y_j^*\|_p^p + \beta \lambda_j^* \frac{\partial EW(\lambda_j^*)}{\partial \lambda_j} \right|_{\lambda_i = \lambda_j^*} + \beta EW(\lambda_j^*), \forall j \neq i \right\}, \forall i \in [N]. \quad (\text{C.38})$$

Proof. As discussed in proof of Proposition 4.3, we can re-write the design problem given in (3.7) as,

$$\min_{\lambda, \mu, Y} \sum_{i \in [N]} \left(h(\mu_i) + \beta \frac{\lambda_i}{\mu_i - \lambda_i} + \kappa \int_{\chi_i} \|x - y_i\|_p^p d\mathbb{P} \right). \quad (\text{C.39})$$

As discussed in the previous sections, Y^* and μ^* can be determined uniquely and locally (in the sense that optimal values for location i , can be determined using λ_i) by λ^* . Thus, we consider an optimal set of λ^* and compare it with another set of optimal demand rates. Note that, changes in λ^* affect $\{\chi_i\}_{i \in [N]}$ implicitly. To capture this, instead of investigating variations in λ^* , we investigate $\{\chi_i\}_{i \in [N]}$ corresponding to λ^* . Note that, by the uniqueness of the solution of the lower-level problem, given λ^* , $\{\chi_i\}_{i \in [N]}$ can be determined uniquely. Thus, consider $\{\chi_i\}_{i \in [N]}$ as discussed and fix two indices j and k in $[N]$. For a $x_0 \in \chi_k$ and $\epsilon > 0$, define $\{\hat{\chi}_i\}$,

$$\begin{cases} \hat{\chi}_k = \chi_k \setminus B_\epsilon(x_0) \\ \hat{\chi}_j = \chi_j \cup (\chi_k \cap B_\epsilon(x_0)) \\ \hat{\chi}_i = \chi_i \text{ for all } i \in \{1, \dots, N\} \setminus \{j, k\}. \end{cases} \quad (\text{C.40})$$

Here, $B_\epsilon(x_0)$ is the l_p -ball around x_0 with radius ϵ . Let $\lambda_\epsilon = \int_{\chi_k \cap B_\epsilon(x_0)} d\mathbb{P}$. From (C.39) and based on the optimality of $\{\chi_i\}_{i \in [N]}$,

$$\begin{aligned} \sum_{i \in [N]} \left(h(\mu_i^*) + \beta \frac{\lambda_i^*}{\mu_i^* - \lambda_i^*} + \kappa \int_{\chi_i} \|x - y_i^*\|_p^p d\mathbb{P} \right) &\leq \sum_{i \in [N]} \left(h(\hat{\mu}_i) + \beta \frac{\hat{\lambda}_i}{\hat{\mu}_i - \hat{\lambda}_i} + \kappa \int_{\hat{\chi}_i} \|x - \hat{y}_i\|_p^p d\mathbb{P} \right) \\ &= \sum_{i \in \{1, \dots, N\} \setminus \{j, k\}} \left(h(\mu_i^*) + \beta \frac{\lambda_i^*}{\mu_i^* - \lambda_i^*} + \kappa \int_{\chi_i} \|x - y_i^*\|_p^p d\mathbb{P} \right) \\ &\quad + h(\hat{\mu}_k) + \beta \frac{\lambda_k^* - \lambda_\epsilon}{\hat{\mu}_k - (\lambda_k^* - \lambda_\epsilon)} + \int_{\chi_k \setminus B_\epsilon(x_0)} \|x - \hat{y}_k\|_p^p d\mathbb{P} \\ &\quad + h(\hat{\mu}_j) + \beta \frac{\lambda_j^* + \lambda_\epsilon}{\hat{\mu}_j + (\lambda_j^* + \lambda_\epsilon)} + \int_{\chi_j \cup (\chi_k \cap B_\epsilon(x_0))} \|x - \hat{y}_j\|_p^p d\mathbb{P}. \end{aligned}$$

By simple algebraic operations, we get

$$\begin{aligned}
& h(\mu_k^*) - h(\hat{\mu}_k) + \beta \left(\frac{\lambda_k^*}{\mu_k^* - \lambda_k^*} - \frac{\lambda_k^* - \lambda_\epsilon}{\hat{\mu}_k - (\lambda_k^* - \lambda_\epsilon)} \right) + \kappa \left(\int_{\chi_k} \|x - y_k^*\|_p^p d\mathbb{P} - \int_{\chi_k \setminus B_\epsilon(x_0)} \|x - \hat{y}_k\|_p^p d\mathbb{P} \right) \\
& \leq h(\hat{\mu}_j) - h(\mu_j^*) + \beta \left(\frac{\lambda_j^* + \lambda_\epsilon}{\hat{\mu}_j - (\lambda_j^* + \lambda_\epsilon)} - \frac{\lambda_j^*}{\mu_j^* - \lambda_j^*} \right) + \kappa \left(\int_{\chi_j \cup (\chi_k \cap B_\epsilon(x_0))} \|x - \hat{y}_j\|_p^p d\mathbb{P} - \int_{\chi_j} \|x - y_j^*\|_p^p d\mathbb{P} \right).
\end{aligned}$$

Dividing both sides by $u_p \epsilon^p$, where u_p is the volume of unit l_p -ball, we get,

$$\begin{aligned}
& \frac{1}{u_p \epsilon^p} (h(\mu_k^*) - h(\hat{\mu}_k)) + \frac{1}{u_p \epsilon^p} \beta \left(\frac{\lambda_k^*}{\mu_k^* - \lambda_k^*} - \frac{\lambda_k^* - \lambda_\epsilon}{\hat{\mu}_k - (\lambda_k^* - \lambda_\epsilon)} \right) \\
& + \frac{1}{u_p \epsilon^p} \kappa \left(\int_{\chi_k} \|x - y_k^*\|_p^p d\mathbb{P} - \int_{\chi_k \setminus B_\epsilon(x_0)} \|x - \hat{y}_k\|_p^p d\mathbb{P} \right) \\
& \leq \frac{1}{u_p \epsilon^p} (h(\hat{\mu}_j) - h(\mu_j^*)) + \frac{1}{u_p \epsilon^p} \beta \left(\frac{\lambda_j^* + \lambda_\epsilon}{\hat{\mu}_j - (\lambda_j^* + \lambda_\epsilon)} - \frac{\lambda_j^*}{\mu_j^* - \lambda_j^*} \right) \\
& + \frac{1}{u_p \epsilon^p} \kappa \left(\int_{\chi_j \cup (\chi_k \cap B_\epsilon(x_0))} \|x - \hat{y}_j\|_p^p d\mathbb{P} - \int_{\chi_j} \|x - y_j^*\|_p^p d\mathbb{P} \right).
\end{aligned}$$

Now, we aim to take the limit $\epsilon \rightarrow 0$ for both sides. First note that,

$$\begin{aligned}
& \int_{\chi_k} \|x - y_k^*\|_p^p d\mathbb{P} - \int_{\chi_k \setminus B_\epsilon(x_0)} \|x - \hat{y}_k\|_p^p d\mathbb{P} = \min_{y_k \in X} \int_{\chi_k} \|x - y_k\|_p^p d\mathbb{P} - \min_{y_k \in X} \int_{\chi_k \setminus B_\epsilon(x_0)} \|x - y_k\|_p^p d\mathbb{P} \\
& = \min_{y_k \in X} \int_{\chi_k} \|x - y_k\|_p^p d\mathbb{P} \\
& - \min_{y_k \in X} \left(\int_{\chi_k} \|x - y_k\|_p^p d\mathbb{P} - \int_{\chi_k \cap B_\epsilon(x_0)} \|x - y_k\|_p^p d\mathbb{P} \right) + \int_{\chi_j \cup (\chi_k \cap B_\epsilon(x_0))} \|x - \hat{y}_k\|_p^p d\mathbb{P} - \int_{\chi_k} \|x - y_k^*\|_p^p d\mathbb{P} \\
& = \min_{y_j \in X} \int_{\chi_j \cup (\chi_k \cap B_\epsilon(x_0))} \|x - y_j\|_p^p d\mathbb{P} - \min_{y_j \in X} \int_{\chi_j} \|x - y_j\|_p^p d\mathbb{P} \\
& = \min_{y_j \in X} \left(\int_{\chi_j} \|x - y_j\|_p^p d\mathbb{P} + \int_{\chi_k \cap B_\epsilon(x_0)} \|x - y_j\|_p^p d\mathbb{P} \right) \\
& - \min_{y_j \in X} \int_{\chi_j} \|x - y_j\|_p^p d\mathbb{P}.
\end{aligned}$$

But for a small enough ϵ , we have,

$$\begin{aligned} \min_{y_k \in X} \int_{\chi_k} \|x - y_k\|_p^p d\mathbb{P} - \min_{y_k \in X} \left(\int_{\chi_k} \|x - y_k\|_p^p d\mathbb{P} - \int_{B_\epsilon(x_0)} \|x - y_k\|_p^p d\mathbb{P} \right) &= \int_{\chi_k \cap B_\epsilon(x_0)} \|x - y_k\|_p^p d\mathbb{P} \\ \min_{y_j \in X} \left(\int_{\chi_j} \|x - y_j\|_p^p d\mathbb{P} + \int_{\chi_k \cap B_\epsilon(x_0)} \|x - y_j\|_p^p d\mathbb{P} \right) - \min_{y_j \in X} \int_{\chi_j} \|x - y_j\|_p^p d\mathbb{P} &= \int_{\chi_k \cap B_\epsilon(x_0)} \|x - y_j\|_p^p d\mathbb{P}. \end{aligned}$$

Thus,

$$\begin{aligned} &\frac{1}{u_p \epsilon^p} (h(\mu_k^*) - h(\hat{\mu}_k)) + \frac{1}{u_p \epsilon^p} \beta \left(\frac{\lambda_k^*}{\mu_k^* - \lambda_k^*} - \frac{\lambda_k^* - \lambda_\epsilon}{\hat{\mu}_k - (\lambda_k^* - \lambda_\epsilon)} \right) + \frac{1}{u_p \epsilon^p} \kappa \left(\int_{\chi_k \cap B_\epsilon(x_0)} \|x - y_k\|_p^p d\mathbb{P} \right) \\ &\leq \frac{1}{u_p \epsilon^p} (h(\hat{\mu}_j) - h(\mu_j^*)) + \frac{1}{u_p \epsilon^p} \beta \left(\frac{\lambda_j^* + \lambda_\epsilon}{\hat{\mu}_j - (\lambda_j^* + \lambda_\epsilon)} - \frac{\lambda_j^*}{\mu_j^* - \lambda_j^*} \right) + \frac{1}{u_p \epsilon^p} \kappa \left(\int_{\chi_k \cap B_\epsilon(x_0)} \|x - y_j\|_p^p d\mathbb{P} \right). \end{aligned}$$

Taking $\epsilon \rightarrow 0$ results in,

$$\frac{\partial}{\partial \lambda_k} h(\mu_k) + \beta \frac{\partial}{\partial \lambda_k} \frac{\lambda_k}{\mu_k - \lambda_k} + \kappa \|x - y_k\|_p^p \leq \frac{\partial}{\partial \lambda_j} h(\mu_j) + \beta \frac{\partial}{\partial \lambda_j} \frac{\lambda_j}{\mu_j - \lambda_j} + \kappa \|x - y_j\|_p^p,$$

which concludes the proof. $\square$

Proof of Proposition 5.1

Proof. Note that, from Theorem 3.1, we know that for the optimal π , there exists a set of service fees $\mathbf{p}$ such that regions defined (C.17) are equal to optimal regions, i.e., they satisfy (C.18). Furthermore, from Theorem C.1, we know that the optimal regions satisfy (C.38). Thus, it suffices to set $\bar{p}_i = \beta \frac{\partial \lambda_i^{EW}(\lambda_i)}{\partial \lambda_i} \big|_{\lambda_i = \lambda_i^*}$. Since p_i are unique up to a constant, and from the definition of utility given in (3.3), the prices are given by,

$$p_i^* = \beta \left(\frac{\partial}{\partial \lambda_i} \frac{\lambda_i}{\mu^*(\lambda_i) - \lambda_i} \right) \big|_{\lambda_i = \lambda_i^*} - \beta \frac{1}{\mu_i - \lambda_i^*} + C. \quad (\text{C.41})$$

From (C.41) by setting $C = 0$, we get

$$\begin{aligned}
p_i^* &= \beta \frac{\partial \lambda_i EW(\lambda_i)}{\partial \lambda_i} \Big|_{\lambda_i=\lambda_i^*} - \beta \frac{1}{\mu_i - \lambda_i} \\
&= \beta \frac{\lambda_i}{(\mu_i - \lambda_i^*)^2} + \beta \frac{1}{\mu_i - \lambda_i^*} - \beta \frac{1}{\mu_i - \lambda_i^*} \\
&= \beta \frac{\lambda_i}{(\mu_i - \lambda_i^*)^2},
\end{aligned}$$

which concludes the second part of the proposition. By strictly increasing of the function in λ_i , the first part of the proposition holds true. $\square$

Proof of Proposition 5.2

Proof. From Proposition 5.1, p_i is monotone in λ_i , and from Proposition 4.4 there exists a value such that for λ_i greater than or equal to that, λ_i decreases as β increases. Thus, the same holds for p_i . $\square$

Proof of Proposition 5.3

Proof. From (C.41), by setting $C = 0$, we get

$$\begin{aligned}
p_i^* &= \beta \frac{\partial \lambda_i EW(\lambda_i)}{\partial \lambda_i} \Big|_{\lambda_i=\lambda_i^*} - \beta \frac{1}{\mu^*(\lambda_i) - \lambda_i} \\
&= \beta \frac{\lambda_i}{(\mu^*(\lambda_i) - \lambda_i^*)^2} + \beta \frac{1}{\mu^*(\lambda_i) - \lambda_i^*} - \beta \frac{1}{\mu^*(\lambda_i) - \lambda_i^*} \\
&= \beta \frac{\lambda_i}{(\mu^*(\lambda_i) - \lambda_i^*)^2}.
\end{aligned}$$

Then, using Proposition 4.5, we get

$$p_i^* = \beta \frac{\lambda_i^*}{(\mu^*(\lambda_i) - \lambda_i^*)^2} = \frac{\partial h(\mu)}{\partial \mu} \Big|_{\mu=\mu^*(\lambda_i^*)},$$

which concludes the proof. $\square$

Proof of Corollary 5.1

Proof. The proof is immediate from Proposition 5.3 and Corollary 4.3. $\square$

C.4 Proof of Results on Algorithmic Analysis

Proof of Proposition B.1

Proof.

1. Algorithm 2 converges since the problem

$$\min_{Y, \lambda} \min_{\pi \in \Pi} \mathbb{E}_{\pi} \|x - y_i\|_p^p$$

can be written as a convex stochastic problem over $\mathbb{R}^N$ using Proposition C.1.

2. Each step of the algorithm is polynomially bounded in time, except for the step that determines the optimal location. Specifically,
 - (a) The first step is a convex stochastic problem over $\mathbb{R}^N$ and can be solved by a stochastic gradient descent algorithm, which is polynomially bounded in time.
 - (b) The second step is a system of equations that can be solved by a Newton-method-based approach in polynomially bounded time.
 - (c) The fourth and fifth steps are simple algebraic updates.
3. The error bounds $\mathcal{O}(\frac{1}{lk\gamma_g})$ and $+\mathcal{O}(\frac{1}{k}\gamma_g)$ are well-known error bounds for stochastic gradient descent and gradient descent algorithms, respectively, (see e.g., [41]). Thus, the combined error is bounded by,

$$\mathcal{O}(\frac{1}{lk\gamma_g}) + \mathcal{O}(\frac{1}{k}\gamma_g).$$

Convergence of Algorithm 1 is provided by convergence of Algorithm 2 and Theorem 2 in [29]. $\square$