Optimal Service Time Windows

Marlin W. Ulmer
Otto-von-Guericke Universität Magdeburg, Magdeburg, Germany, marlin.ulmer@ovgu.de

Justin C. Goodson
Department of Operations & Information Technology Management, Saint Louis University, St. Louis, MO

Barrett W. Thomas
Department of Business Analytics, University of Iowa, Iowa City, IA

Because customers must usually arrange their schedules to be present for home services, they desire an accurate estimate of when the service will take place. However, even when firms quote large service time windows, they are often missed, leading to customer dissatisfaction. Wide time windows and frequent failures occur because time windows must be communicated to customers in the face of several uncertainties: future customer requests are unknown, final service plans are not yet determined, and when fulfillment is outsourced to a third party, the firm has limited control over routing procedures and eventual fulfillment times. Even when routing is performed in-house, time windows often do not receive explicit consideration. In this paper, we show how companies can communicate reliable and narrow time windows to customers in the face of arrival time uncertainty when time window decisions are decoupled from routing procedures. Under assumptions on the shape of arrival time distributions, our main result characterizes the optimal policy, identifying structure that reduces a high-dimensional stochastic non-linear optimization problem to a root-finding problem in one dimension. The result inspires a practice-ready heuristic for the more general case. Relative to the industry standard of communicating uniform time windows to all customers, and to other policies applied in practice, our method of quoting customer-specific time windows yields a substantial increase in customer convenience without sacrificing reliability of service. Our results show that time windows should be tailored to individual customers, time window sizes should be proportional to the service level, larger time windows should be assigned to earlier requests and smaller time windows to later requests, larger time windows should be assigned to customers further from the depot of operation and smaller time windows to closer customers, high quality time windows can be identified even with limited data, and cost savings afforded by routing efficiency should be measured against potential losses to customer convenience.

Key words: time windows, service routing, non-linear optimization

1. Introduction

While one of the authors writes this introduction, he sits at home close to the door waiting for a shipment to arrive. The delivery company Hermes gave him a time window from 3pm to 7pm, but he waits and writes at 9am because experience tells him that even though the communicated time window is already very large, there is no guarantee that it will be satisfied, and he does not want to miss the shipment (again). Waiting is not new for the authors. Recently, Marlin Ulmer scheduled internet installation with Deutsche Telekom. They offered service during a five-hour time window from 8am to 1pm. The wide time window required that he stay at home half the day so that he could receive the service technician. When Justin Goodson purchased an appliance from Samsung, he was given a four-hour delivery time window from noon to 4pm. On the delivery date, Goodson received a phone call at 7am from a third-party logistics company indicating arrival in 20 minutes. If he could not meet the driver, then delivery would need to be rescheduled for a different day. Barrett Thomas paid a premium for FedEx express package delivery. When delivery was attempted outside of the time window, Thomas was not present and the package was not delivered. After two more attempted deliveries, both outside of communicated time windows, FedEx returned the package to the vendor.

Many readers will share the authors' frustrations. Not only do providers of delivery, installation, and in-home repair services typically communicate very large time windows at the time of booking, they often fail to meet those time windows on the day of service. The reasons for wide time windows and frequent failures are twofold. First, the time window is communicated to a customer when they request service. Because requests arrive sequentially, the firm does not know what other customers will be scheduled for service on the same day as the given customer. Consequently, the schedule is set after time windows are communicated. Even then, uncertain travel and service times make it difficult to predict arrival times to customers. Second, because time is money, firms have invested years and millions to develop effective scheduling and routing algorithms. In competitive markets, cost-efficient routing solutions are essential (for example, see Prisco (2017)). Notably, our conversations with service drivers reveal that satisfaction of communicated time windows often does not play a role in routing optimization, though firms will not officially confirm this. Alternatively, whenever delivery operations are outsourced to third-party providers, as

in the case experienced by Justin Goodson, the firm that originates the product or service typically has little say in route design. Further, as highlighted by the recent Amazon routing challenge (Amazon 2021), even when firms employ their own drivers, it is not clear that the drivers follow provided routes. All this helps to explain the practice of communicating very large time windows, as well as their frequent violations (Apte et al. 2007).

While cost-efficient routes save companies money, others pay. People receiving home services have to plan in advance and often take time off work to be home while service occurs. The real-time tracking some companies offer on the day of service does not help much in such cases because adjustments to work and child care schedules, for example, typically require more advance notification. One estimate puts the cost of people waiting for home services at \$38 billion in 2011 (Ellis 2011). With the rise of e-commerce, this number is likely even higher today. Despite their use of routing and scheduling methods that disregard customers' desires for smaller and more reliable time windows, firms stand to benefit from movement in this direction. Missed time windows and customers require costly recourse actions (Lim et al. 2023). Reducing these incidents can benefit both providers and customers.

How can we change the current state of affairs? How can firms keep their carefully developed routing algorithms or outsource logistics to third parties, but at the same time communicate shorter and more reliable time windows to customers? Answering this question is the focus of this paper. We consider the situation in which a firm providing on-location services schedules customers during a booking period and then executes the service at a later date. A time window is communicated when a customer requests service. Because customers randomly request service across the booking period, time windows must be determined under incomplete information about the set of customers that will eventually be served and, consequently, under incomplete information about the order in which the customers will be served. Thus, at the moment the time window is assigned, we know only a probability distribution on the arrival time to the customer. We make no assumption on how customers are routed. Rather, the arrival time distribution accounts for a firm's chosen routing method. For every customer request during the booking period, we set a time window for future service. Because the decision comprises a start and an end time, we also control the size of the time window.

The goal is to minimize the expected time window size across all customers. Since providers draw on their existing scheduling and routing mechanisms or outsource service to third parties, in practice, time window decisions are decoupled from the final routing. Consequently, time window choices do not impact arrival time distributions. This reflects the perspective taken by many firms, that early or late arrivals, no matter their frequency and magnitude, are an acceptable way to reconcile differences in what is initially communicated to customers and an eventual route plan. Thus, we cannot guarantee that a time window communicated during the booking period will be satisfied for every customer. To ensure a measure of reliability, we follow practice and impose a chance constraint to ensure that a high percentage of time windows are satisfied (Keller 2020). For example, OptimoRoute, a routing software for service and goods delivery, measures on-time delivery across customers (OptimoRoute 2022) and Deutsche Telekom aims for 95 percent punctuality across service calls (Schasche 2021). Similarly, Weiss (2012) suggests that Comcast was seeking 98 percent on-time arrivals.

To the best of our knowledge, we are the first to offer high-quality decision support for companies planning to communicate reliable and narrow time windows to customers while keeping their cost-efficient scheduling and routing algorithms or while outsourcing logistics to third parties. Our work addresses a new variant in the class of problems comprising self-imposed time windows and the time-window-assignment vehicle routing problem, introduced by Jabali et al. (2015) and Spliet and Gabor (2015). We model the problem as a stochastic program. Under assumptions on the shape of arrival time distributions to customers (strict monotonicity on either side of a unique mode), we identify structure in an optimal policy that significantly simplifies the optimization. Our main result reduces a high-dimensional stochastic non-linear optimization problem to a root-finding problem in one dimension. For more general distributional forms, the result inspires a practice-ready heuristic. Via a computational study of next-day service operations in a Midwest US city, we show that the heuristic performs competitively relative to a dual bound and significantly outperforms industry standard practices. In short, our work makes complex optimization accessible to large and small firms alike, thus laying the groundwork to improve the customer experience while maintaining a high standard of service.

Through our analytical and computational work, we identify six managerial insights:

1. **Time windows should be customer-specific**, with both size and placement reflecting the shape of the arrival time distribution. Notably, we show that relative to the current practice of communicating uniform time windows to all customers, our method yields a substantial increase in customer convenience without sacrificing reliability of service.

2. We show that **higher service levels require wider time windows** and that **at a given service level, optimization is required to set sizes**.

3. We connect time window size to time of request, showing that in general **firms should assign larger time windows to earlier requests and smaller time windows to later requests**.

4. We tie time window size to location of request, demonstrating that decision makers should **assign larger time windows to customers further from the depot of operation and smaller time windows to closer customers**.

5. Even when firms lack the data necessary to fully characterize arrival time distributions, or the computational resources required to estimate them via simulation, **approximation of the arrival time distributions via mean and variance lead to effective decisions.**

6. We demonstrate that while route optimization can improve efficiency, it may significantly increase arrival time uncertainty, which in turn can increase time window sizes considerably. **Firms should carefully balance cost savings afforded by routing efficiency with the potential loss of customer convenience.**

The paper is organized as follows. In §2, we discuss related literature. In §3, we describe the problem and formulate an optimization model. In §4, we present and analyze our main theoretical result. In §5, we propose a practice-ready heuristic and demonstrate its performance via computational experiments. We conclude the paper in §6.

2.   Related Literature

Our research stems largely from literature on the self-imposed time window problem (SITWP), introduced by Jabali et al. (2015), and the time-window-assignment vehicle routing problem (TWAVRP), introduced by Spliet and Gabor (2015). In these streams, the work of Vareias et al. (2019) is most closely related to this paper.

Vareias et al. present a variant of the SITWP where all customers are known and the objective is to minimize the sum of time window widths for given routes in the face of stochastic travel times. Each customer's time window is chance-constrained, with probability distributions represented by a small set of discrete scenarios. In contrast, we draw on the industry practice of modeling reliability as a global chance constraint across all customers. Further, our model is general enough to incorporate any source of uncertainty that can be captured in the arrival time distribution, including routing decisions. This allows us to focus on the case where not all requests are known when the time window must be communicated to the customer. Instead of relying on information about the arrival time distribution, as we do in this paper, Ulmer and Thomas (2018) propose a supervised learning procedure to predict mean arrival times to customers. These predictions are used to set fixed-width time windows for all customers. We benchmark our customer-specific time windows against those of Ulmer and Thomas. On average, our method provides comparable service levels with significantly narrower time windows.

Our analytical results explain observations noted in the literature. When customer choice is a consideration, Köhler et al. (2020) offer some customers a menu of "short" time windows and other customers a selection of "long" time windows, finding it is best to reserve shorter time windows for customers who request later in the operating horizon. Our analytical results explain why the approach of Köhler et al. works, as well as the observation from Ulmer and Thomas (2018) that time window size tends to decrease with time of request: requests that occur later in the booking period correlate with less arrival time variability, hence these customers can be quoted smaller time windows. However, we note that customers who call later in the booking period in hopes of getting a relatively narrower time window run the risk of receiving no service at all as there may not be capacity to serve them.

The bulk of the SITWP and TWAVRP literature integrates routing and time window assignment in the face of various uncertainties, such as travel time, with the objective of minimizing expected cost. In contrast, our work recognizes that many firms decouple these choices, either because routing efficiency is key or because routing is managed by a third party. This literature includes Madsen et al. (1996), Jabali et al. (2015), Spliet and Desaulniers (2015), Zhang et al. (2015), Dalmeijer and Spliet (2018), Neves-Moreira et al. (2018), Spliet

et al. (2018), Subramanyam et al. (2018), Martins et al. (2019), Jalilvand et al. (2021), and Beskers (2022). Additionally, Flinterman (2022) considers a variant of the TWAVRP in which customer locations are known but randomly require service. Hoogeboom et al. (2021) present a robust optimization approach and hedge against time window violations by incorporating arrival time distributions into the optimization via a risk measure. Our work also gives explicit consideration to variability in arrival times. Yu et al. (2023) model a TWAVRP as a two-stage stochastic program. They test their approach in a dynamic environment where customers request sequentially, but these request uncertainties are not considered in the optimization.

Our work contrasts with business models in which customers choose their own service time windows. Waßmuth et al. (2023) provide an overview of this literature, which includes research in grocery delivery (Campbell and Savelsbergh 2005, Ehmke and Campbell 2014, Yang et al. 2014, Köhler et al. 2020, Agatz et al. 2021). Allowing customers to choose their own time windows creates challenges for firms, often preventing the use of efficient scheduling and routing procedures and significantly increasing cost (Ehmke and Campbell 2014). For this reason, particularly in the service contexts that motivate our work, the business model involves a booking period that allows for the planning of efficient routes before the day of service. Related, many grocery companies now set time windows internally instead of via customer preferences (Visser and Savelsbergh 2019).

Finally, our work is related to research on due date assignment, outlined by Keskinocak and Tayur (2004). In this literature, the key decision is to identify a deadline, rather than a time window, by which a product will be manufactured, with uncertainties in production lead times typically modeled via queuing systems. Similar to our approach, the work of Spearman and Zhang (1999) identifies necessary conditions for optimal deadlines. In the case of setting time windows, we show that necessary conditions are also sufficient. Further, in contrast to Spearman and Zhang, we derive a heuristic from our analytical results and show that it performs well relative to distributions that do not satisfy the assumptions of our analysis. Hafızoğlu et al. (2016), Kahvecioglu and Balcıoglu (2016), and Gel et al. (2020) summarize more recent work on due date management, which considers deadline setting in tandem with pricing decisions and demand conditional on quoted due dates. In contrast, our work focuses on service settings where customers usually are not given a choice regarding the time of service and we consider the more complex task of selecting time windows.

3. Problem Description and Model

In this section, we describe and model the problem of setting service time windows in the face of arrival time uncertainty. We consider the case of a company that provides attended home service to customers. For example, appliance installation and repair, furniture delivery and setup, parcel delivery, utility service, or home health assistance. We assume that the company collects customer requests over a period of time and then provides service on a specified future date. This setting is similar to that of Apte et al. (2007). When a customer requests service, the planner consults the firm's operational state to determine if the request can be accommodated. The state may contain information on customers the firm has already committed to serve, available resource capacities, or a tentative service plan. Before communicating a time window to the customer, the planner considers how arrival time uncertainties may impact service. Between the time of request and the end of the planning horizon, various events may occur, e.g. additional requests, updates to capacity requirements, or handoff to a third party for external logistics planning. In addition, during the execution of service, the technician may encounter other uncertainties such as traffic and unforeseen service challenges. These trajectories of possible events characterize an arrival time distribution, conditional on the operational state. The planner seeks a policy that minimizes expected time window size subject to a constraint on the likelihood of early or late arrivals across possible states. For example, the time windows are such that on average 95 percent of customers are served in their given time window (Schasche 2021). We assume that customers are available for service when the technician arrives, perhaps as a result of a more precise arrival time being communicated on the day of service as was described in the introduction.

Figure 1 illustrates how stochasticity in customer requests and a firm's logistical methods can lead to uncertain arrival times to customers. The left panel shows the operational state at the time of the second request for service on a given day. At this moment in time, the locations of Customers 1 and 2 are known, but the times and locations of any future requests are unknown, as is the sequence of visits in a final routing plan. The middle panel shows several potential route plans after all requests have been realized. The routes vary widely depending on the locations of additional requests. Because Customer 2 is sequenced fourth, fifth, and seventh across the three route plans, the realized arrival time to Customer 2 is different in each plan.
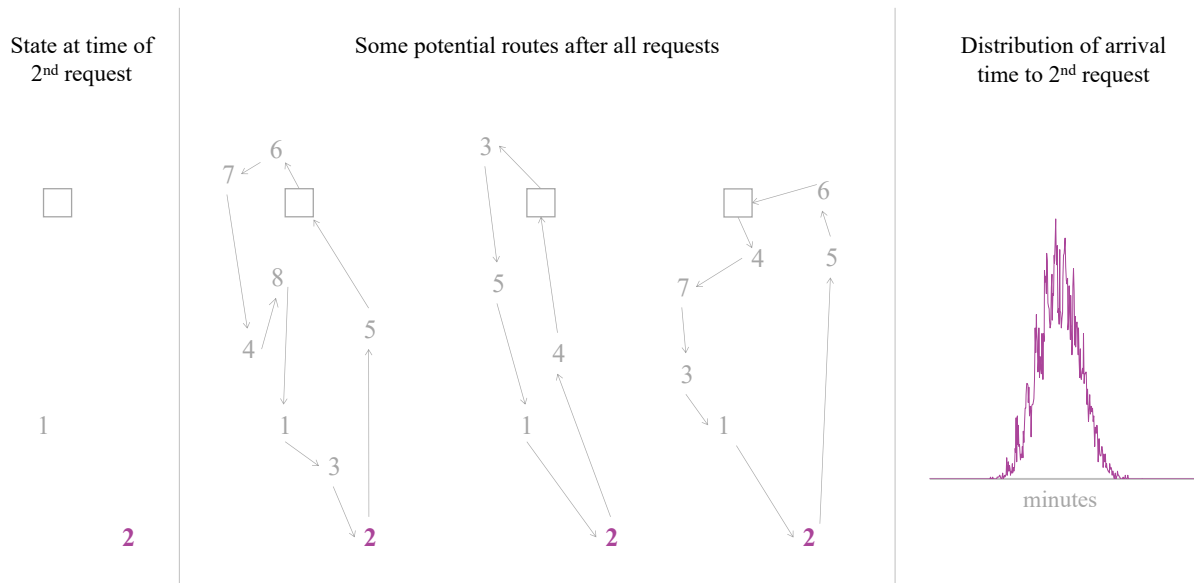
Figure 1 **Impact of Logistics Planning on Arrival Times**

Many other realizations of requests and route plans may be possible. The collection of these realizations, along with their respective likelihoods, account for the bulk of the uncertainty in arrival times (Ulmer and Thomas 2018). The result is a distribution of possible arrival times. The right panel displays the results of a simulation to illustrate a potential distribution of arrival times to Customer 2.

Whatever routing and scheduling method the firm uses, or whatever procedure a third-party logistics provider employs, we take it as exogenous to the model, accounting for them only through the arrival time distributions that they induce. Thus, we assume the final service plan is determined independently from time windows, and thus time window choices do not impact arrival time distributions. This reflects the perspective taken by many firms, that early or late arrivals, no matter their frequency and magnitude, are an acceptable way to reconcile differences in what is initially communicated to customers and an eventual route plan. Because these firms typically view an early or late arrival as a no-cost recourse, time window decisions have no impact on logistical choices. Consequently, whether the routing mechanism is as simple as a cheapest insertion heuristic or as sophisticated as solving a complex mixed integer program, the result is an arrival time distribution associated with the operational state at the time of a customer's request.

Because time window choices do not impact arrival time distributions, the optimization is not a sequence of interdependent decisions. Time window choices are only connected via a global constraint on service

level. However, the optimization is still complex. We must choose a time window for every state-dependent arrival time distribution. Minimizing expected time window duration subject to a service level requirement is a nonlinear stochastic program.

The approach that we take in this paper is different from the approach typical in the routing with time windows literature. In that literature, the time windows are taken as given and then vehicles are routed under the constraint that the time windows must be met. We take the routing policy as given and solve for optimal time windows in an environment of dynamically arriving customer requests.

Formally, let S be the random state of the environment when the customer requests service, with finite support $\mathcal{S}$. Denote by $X(s)$ the random arrival time when the state of the environment is s in $\mathcal{S}$. Let $\mathbb{E}_{\mathcal{S}}[\cdot]$ denote expectation across states and let $\mathbb{P}\{\cdot\}$ denote a probability measure. In this way, we give consideration not only to the distribution of operational states, but also to the arrival time distribution associated with each state. Then, for each state s in $\mathcal{S}$, the firm chooses an interval $[l(s), u(s)]$ as a state-dependent service time window, where $l(s)$ and $u(s)$ represent times during the day of service. The firm's objective is to minimize expected time window size across all states of the environment, $\mathbb{E}_{\mathcal{S}}[u(S) - l(S)]$, thus seeking time windows most convenient for customers. The firm imposes a single service level across all states. The state-wise contribution to service is the probability of $X(s)$ lying within the communicated time window. The service level is $\mathbb{P}\{l(S) \leq X(S) \leq u(S)\}$, which we require to be at least $1 - \alpha$, where the firm chooses α in the half-open interval $(0, 1]$. Rewriting the service level via the law of total probability as $\sum_{s \in \mathcal{S}} \mathbb{P}\{S = s\}\mathbb{P}\{l(s) \leq X(s) \leq u(s)|S = s\}$ clarifies that the metric captures uncertainty across states as well as uncertainty in the arrival time associated with each state. The constraint limits the number of early or late arrivals to be below a threshold, thus protecting the firm's reputation and ensuring a level of reliability. In summary, the problem we seek to solve is

$$
\begin{aligned}
\text{minimize} \quad & \mathbb{E}_{\mathcal{S}}\Big[u(S) - l(S)\Big] \\
\text{subject to} \quad & \mathbb{P}\left\{l(S) \leq X(S) \leq u(S)\right\} \geq 1 - \alpha \\
& u(s) \geq l(s) \ \forall \ s \in \mathcal{S}.
\end{aligned}
\tag{1}
$$

4. Optimization and Analysis

In this section, we characterize the optimal policy under assumptions on the shape of arrival time distributions. Our main result identifies structure that reduces Problem (1) from a high-dimensional stochastic nonlinear optimization problem to a root-finding problem in one dimension. We then discuss managerial implications of the policy structure. In subsequent sections, we use the results presented in this section to inform methods for setting time windows in more general settings.

For the purpose of analysis, we assume that the likelihood of observing state s in $\mathcal{S}$ is strictly positive: $\mathbb{P}\{S = s\} > 0$. If the probability is zero, then the state can simply be removed from the support. We assume for each s in $\mathcal{S}$ that arrival time $X(s)$ is a continuous random variable. We write the cumulative and probability density functions (CDF and PDF) as $F_{X(s)}$ and $f_{X(s)}$, respectively. Both are conditional on $S = s$ to facilitate the service level constraint as expressed by the law of total probability. We require $F_{X(s)}$ to be differentiable and $f_{X(s)}$ to be unimodal. Regarding the latter, each $f_{X(s)}$ must be strictly monotonic on either side of its mode $\bar{x}(s) = \arg\max_x f_{X(s)}(x)$. Denote by $\underline{f}_{X(s)}$ the strictly increasing values of $f_{X(s)}$ left of $\bar{x}(s)$ and by $\overline{f}_{X(s)}$ the strictly decreasing values of $f_{X(s)}$ right of $\bar{x}(s)$. For state s in $\mathcal{S}$, denote by $\mathcal{X}(s)$ the support of $X(s)$. We require $\mathcal{X}(s)$ to be a proper interval on the real line, meaning the interval must contain more than a single point. If $\mathcal{X}(s)$ is left-bounded, meaning the left end of the support interval is finite, then we denote the left antimode by $\tilde{x}(s) = \arg\inf_x \underline{f}_{X(s)}(x)$. If $\mathcal{X}(s)$ is right-bounded, meaning the right end of the support interval is finite, then we denote the right antimode by $\hat{x}(s) = \arg\inf_x \overline{f}_{X(s)}(x)$. If either end of the support interval is unbounded, meaning the left or right end of the interval is infinite, then the corresponding antimode does not exist and we say the support is left-unbounded, right-unbounded, or both. Denote by $\bar{y} = \max\{f_{X(s)}(s) : s \in \mathcal{S}, x \in \mathcal{X}(s)\}$ the largest density value across all PDFs. Denote by $\check{y} = \inf\{f_{X(s)}(x) : s \in \mathcal{S}, x \in \mathcal{X}(s)\}$ the smallest density value across all PDFs.

Theorem 1 shows that time windows for all states may be set as a function of a common density value $y > 0$. Let $\mathcal{X}(s, y) = \{x \in \mathcal{X}(s) : f_{X(s)}(x) \geq y\}$ be all elements of the support of $X(s)$ whose density values are greater than or equal to y. Then, for a given y and for each s in $\mathcal{S}$, Theorem 1 sets time windows as

$$l(s, y) = \begin{cases} \min \mathcal{X}(s, y), & y \leq f_{X(s)}(\bar{x}(s)), \\ \bar{x}(s), & f_{X(s)}(\bar{x}(s)) < y, \end{cases} \tag{2}$$

$$u(s, y) = \begin{cases} \max \mathcal{X}(s, y), & y \le f_{X(s)}(\bar{x}(s)), \\ \\ \bar{x}(s), & f_{X(s)}(\bar{x}(s)) < y. \end{cases} \tag{3}$$

In the first case of Equation (2), when y is at or below the density value of the mode, $l(s, y)$ is the smallest element of the support whose density value meets or exceeds y. If the left antimode exists, and if y lies below $f_{X(s)}(\tilde{x}(s))$, the density value of the left antimode, then Equation (2) sets $l(s, y) = \tilde{x}(s)$. Otherwise, because $f_{X(s)}$ is unimodal and monotonic on either side of its mode, the lower end of the time window is set such that $y = \underline{f}_{X(s)}(l(s, y))$, meaning $l(s, y) = \underline{f}_{X(s)}^{-1}(y)$. In other words, $l(s, y)$ is the point at which y intersects the PDF to the left of the mode. Similarly, in the first case of Equation (3), $u(s, y)$ is the largest element of the support whose density value meets or exceeds y. If the right antimode exists, and if y lies below $f_{X(s)}(\hat{x}(s))$, then Equation (3) sets $u(s, y) = \hat{x}(s)$. Otherwise, the upper end of the time window is set such that $y = \overline{f}_{X(s)}(u(s, y))$, meaning $u(s, y) = \overline{f}_{X(s)}^{-1}(y)$. Here, $u(s, y)$ is the point at which y intersects the PDF to the right of the mode. When y lies above the density value of the mode, the second cases in each equation set both ends of the time window to the mode.

In the proof of Theorem 1 (provided in the Appendix), we show that the there exists a unique common density value $y^\star$ that satisfies Equation (4). This density value establishes a *saddle point*, an optimal solution to Problem (1) and its dual. Further, we show that the left side of Equation (4) strictly decreases across $\check{y} < y^\star \le \bar{y}$. Consequently, Theorem 1 reduces Problem (1) to a line search across an interval.

Theorem 1 (Optimal Time Windows) *There exists an optimal solution to Problem* (1) *corresponding to the unique $y^\star$ in the half-open interval $(\check{y}, \bar{y}]$ satisfying*

$$\mathbb{P}\{l(S, y^\star) \le X(S) \le u(S, y^\star)\} = 1 - \alpha. \tag{4}$$

We graphically describe Theorem 1 via Figure 2, where we illustrate the optimization across two equally likely states s_1 and s_2. The left portion of the figure illustrates how time windows change with the common density value. The figure shows the arrival time distributions for random variables $X(s_1)$ and $X(s_2)$, which we represent with triangular distributions for ease of illustration. PDF $f_{X(s_1)}$ has a left antimode of $\tilde{x}(s_1) = 5$,
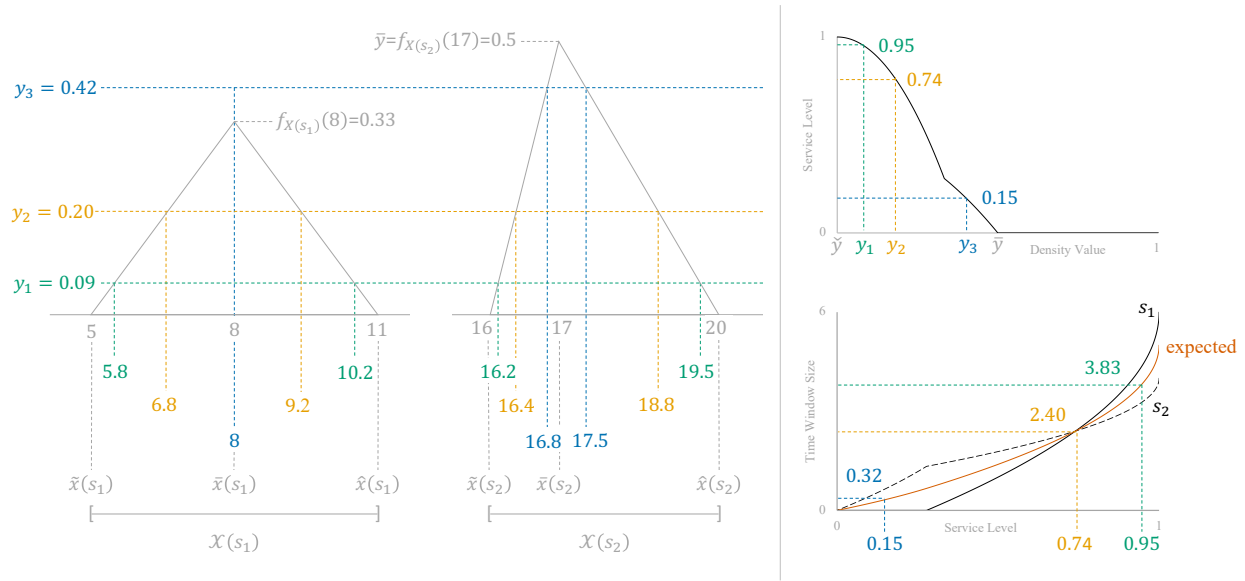
Figure 2   Time Windows, Service Levels, and Common Density Values

Note. As the common density value increases, time windows shrink (left), the service level goes down (top-right), and the minimum expected time window size decreases as well as time window sizes for each state (bottom-right).

a right antimode of $\hat{x}(s_1) = 11$, and is symmetric about its mode $\bar{x}(s_1) = 8$. PDF $f_{X(s_2)}$ is positively skewed about its mode $\bar{x}(s_2) = 17$ with a left antimode of $\tilde{x}(s_2) = 16$ and a right antimode of $\hat{x}(s_2) = 20$. The density value of each antimode is zero: $\tilde{y} = f_{X(s_1)}(5) = f_{X(s_1)}(11) = f_{X(s_2)}(16) = f_{X(s_2)}(20) = 0$. Mode density values are $f_{X(s_1)}(8) = 0.33$ and $f_{X(s_2)}(17) = 0.5$. Supports are $\mathcal{X}(s_1) = [5, 11]$ and $\mathcal{X}(s_2) = [16, 20]$. To facilitate the example, three density values are depicted across the half-open interval $(\tilde{y}, \bar{y}] = (0, 0.5]$.

At density value $y_1 = 0.09$ (which corresponds to an industry-standard service level of 0.95 (Keller 2020)), the time windows $[l(s_1, y_1), u(s_1, y_1)] = [5.8, 10.2]$ and $[l(s_2, y_1), u(s_2, y_1)] = [16.2, 19.5]$ are wide, near the upper and lower boundaries of the supports. Increasing the density level to $y_2 = 0.20$ (which corresponds to a service level of 0.74), time windows $[l(s_1, y_2), u(s_1, y_2)] = [6.8, 9.2]$ and $[l(s_2, y_2), u(s_2, y_2)] = [16.4, 18.8]$ have equal width and are smaller relative to their sizes at y_1. Because density level $y_3 = 0.42$ (corresponding to an impractically low service level of 0.15) is larger than $f_{X(s_1)}(8) = 0.33$, $l(s_1, y_3) = u(s_1, y_3) = 8$ and the time window size is zero. Time window $[l(s_2, y_3), u(s_2, y_3)] = [16.8, 17.5]$ is even narrower than it is at y_2. Although service levels as low as those corresponding to y_2 and y_3 are uncommon in practice, they are helpful in the illustration of Theorem 1.

The top-right portion of Figure 2 graphs Equation (4), depicting the relationship between the common density value and the service level. As the density value increases from y_1 to y_2 to y_3, the service level decreases from 0.95 to 0.74 to 0.15, respectively. The kinks in the graph occur at $f_{X(s_1)}(8) = 0.33$ and $f_{X(s_2)}(17) = 0.5$, the density values of the modes, which push their respective time windows to a size of zero. As stated in the proof of Theorem 1 and reflected in the figure, the left side of Equation (4) is strictly decreasing and bijective from $(\check{y}, \bar{y}]$ to $(0, 1]$. Thus, basic root-finding methods, such as bisection, can be used to identify the point at which the function equals $1 - \alpha$, thereby identifying $y^\star$ and setting optimal time windows per Equations (2) and (3).

Via service level, the bottom-right portion of Figure 2 illustrates the connection between common density values and objective values. As the service level increases from 0.15 to 0.74 to 0.95, the minimum expected time window sizes move from 0.32 to 2.40 to 3.83, respectively, corresponding to smaller common density values y_3, y_2, and y_1. The figure also shows how the time window size for each state contributes to the objective across the same domain of service levels. Again, for a given service level, an optimal solution to Problem (1) can be identified by finding the common density value satisfying Equation (4).

Two managerial insights follow directly from Theorem 1. First, in contrast to both practice and the academic literature, time window size and placement should reflect the shape of the arrival time distribution. The left portion of Figure 2 illustrates this, where the skewness of the arrival time distributions, in conjunction with the common density value, dictate the size and placement of time windows. The popular industry practice of fixed-width time windows suggests customer arrival time distributions are ignored in favor of assigning customers to four-hour morning or afternoon buckets. This practice should be retired in favor of customer-specific time windows tailored to arrival time distributions.

Ulmer and Thomas (2018) consider time-dependent time windows, but center them around expected arrival times. According to Theorem 1, this can only be optimal when the arrival time distribution is symmetric about its mean, as illustrated by the arrival time distribution for state s_1 in Figure 2. However, in state s_2, and in any state whose arrival time distribution is not symmetric about its mean, the time window is not centered about the expected arrival time. Indeed, in the extreme case of a 0.15 service level (corresponding to common

density value y_3), not only is the optimal time window $[l(s_2, y_3), u(s_2, y_3)] = [16.8, 17.5]$ not centered about the mean of 17.7, the time window does not even include the mean. As our numerical experiments in the subsequent section show, customer-specific time windows that follow the shape of the arrival time distribution increase convenience without sacrificing reliability. We summarize these observations in Insight 1.

Insight 1 (Customer-Specific Time Windows) *The beginning and end of a time window should reflect the shape of the arrival time distribution. Consequently, the practices of assigning uniform fixed-width time windows to all customers and of centering time windows around mean arrival times should be replaced by assignment of customer-specific time windows.*

Second, intuition suggests, and Theorem 1 confirms, that higher service levels require wider time windows to better hedge against uncertain arrival times. However, at a given service level, the assignment of time windows is not necessarily dictated by the uncertainties, as one might assume. For example, focusing on the bottom-right of Figure 2, when the service level is 0.74, an optimal policy assigns equally-sized time windows to each state s_1 and s_2: $9.2 - 6.8 = 18.8 - 16.4 = 2.40$. At higher service levels, where most businesses operate, state s_1, whose arrival time standard deviation is higher than that of state s_2, receives the larger time window. This satisfies the intuition that the arrival time distribution with more uncertainty warrants the wider service window, an inclination largely confirmed by our numerical experiments in the subsequent section. However, at lower service levels, the larger time window goes to state s_2. In fact, at service levels less than or equal to 0.28, state s_1 is assigned a zero-width time window, a commitment that almost certainly cannot be kept. Though this scenario is a concern in principle, our experiments in §5 show that such extreme cases are rare when operating at industry-standard service levels and that early or late arrivals are infrequent and small in magnitude.

As discussed in §1, the home services industry assigns fixed-width time windows to customers, often four hours in duration. Further, much of the TWAVRP literature reviewed in §2 assumes fixed-width time windows. Though Köhler et al. (2020) and Ulmer and Thomas (2018) provide computational evidence that differentiated time windows are beneficial, Theorem 1 shows this practice to be optimal and provides guidance on setting time window sizes proportional with the service level. We formalize these observations as Insight 2 and further probe the relationship between arrival time uncertainty and time window size in §5.

Insight 2 (Time Window Size and Service Level) *Higher service levels require wider time windows. At a given service level, optimization is required to size time windows.*

5. Theory to Practice

Though the assumptions leading to Theorem 1 allow for an elegant analysis, they are not always satisfied in practice. For example, in our computational study, arrival times are a result of real customer locations, simulated request frequencies, and common routing heuristics. Consequently, we encounter a variety of empirical PDFs, from approximately unimodal to bimodal with high variance. Examples are given in Figure 3, where the horizontal axes represent arrival time realizations across a 480-minute operating horizon and the vertical axes display relative frequencies. The figure displays distributions for three states. It shows two distributions per state, obtained via two different routing methods discussed later in the paper. The vertical axes vary in scale to aid illustration. The common density values displayed in the figure, $y^\star = 0.0037$ in the top row and $y^\star = 0.0031$ in the bottom row, also aid subsequent discussion. Though many of the empirical PDFs we observe resemble the top left of Figure 3, even in these cases, values do not always strictly decrease left and right of the mode, as assumed in the analysis of §4. Because the proof of Theorem 1 requires this condition, setting time windows via Equations (2) and (3) no longer carries a guarantee of optimality. However, as we show, even when the assumptions leading to Theorem 1 are not met, the resulting time windows are very good. The remainder of the section proceeds as follows. We discuss our heuristic solution approach in §5.1. In §5.2, we describe the simulation designed to demonstrate the quality of the heuristic. We specify benchmarks in §5.3 followed by a discussion in §5.4.

5.1. Heuristic Solution Approach

It is straightforward to use Theorem 1 as a heuristic with general arrival time distributions. Let the mode $\bar{x}(s)$ be any member of the set $\arg\max_x f_{X(s)}(x)$ and use Equations (2) and (3) directly. By the structure of these equations, service level is a non-increasing function of y no matter the shapes of the arrival time distributions. Thus, $y^\star$ is identified by decreasing y from $\bar{y}$ until $\mathbb{P}\{l(S,y) \leq X(S) \leq u(S,y)\}$ rises to $1 - \alpha$.

Algorithm 1 formalizes the heuristic. The procedure takes as input a step size η on Line 1 and initializes the common density value to $y = \bar{y}$ on Line 2. In our computational study, we set $\eta = 0.0001$, which corresponds
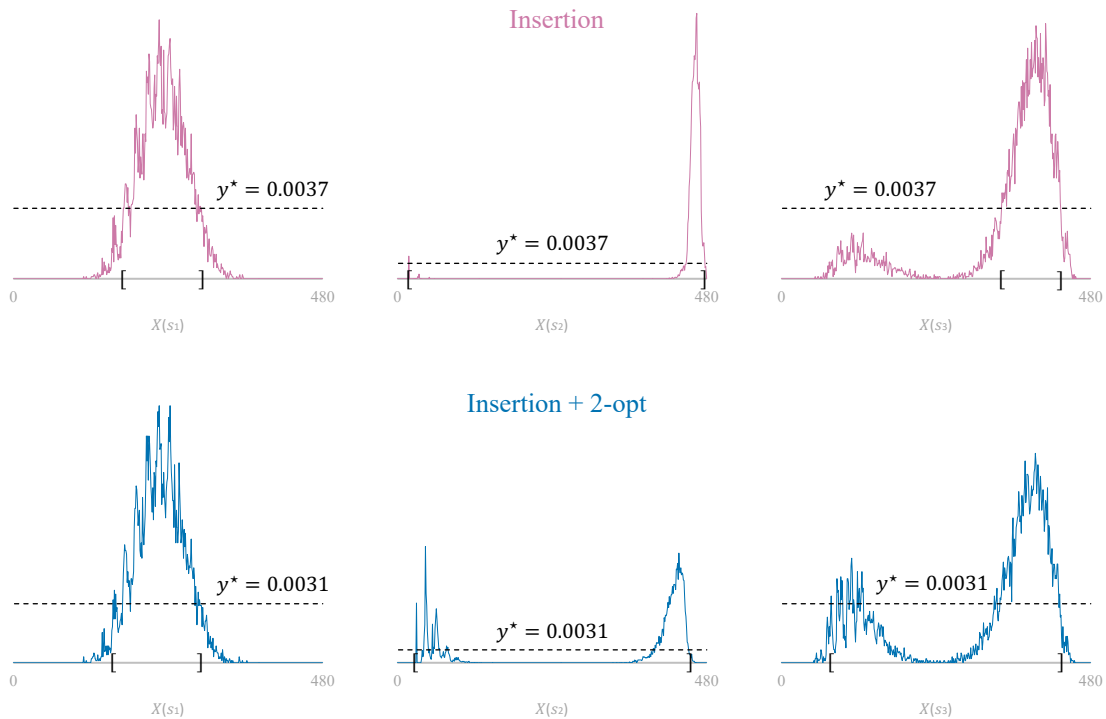
Figure 3   Empirical PDFs and Time Windows

Note. Top row: arrival time distributions when customers are routed via insertion. Top left: approximately unimodal. Top center: low variance, large time window. Top right: high variance, small time window. Bottom row: same distributions when customers are routed via insertion then with 2-opt post-optimization. Vertical scales vary by column to aid illustration.

to the granularity of the empirical PDFs constructed via simulation. Parameter η can also be viewed as a discretization for continuous-time distributions. Each iteration of the loop beginning on Line 3 sets time windows and checks feasibility. For each state s in $\mathcal{S}$, Lines 4-8 map y to a time window beginning at $l(s)$ and ending at $u(s)$ via Equations (2) and (3). Line 9 decrements y by η and Line 10 checks the service level of the current time windows against the threshold $1 - \alpha$. Once the service level is high enough, the procedure terminates on Line 11 with time windows for each state.

Figure 4 depicts the results of the heuristic on the data discussed in §5.2. When the service level is set to $1 - \alpha = 0.95$, Algorithm 1 terminates with common density value $y^\star = 0.0037$, achieving an objective value of 77.3 minutes, as indicated in the right portion of the figure. As we show in §5.4, the time windows selected by Algorithm 1 perform very well. Though Algorithm 1 conducts only a simple line search, this is sufficient to achieve strong performance.
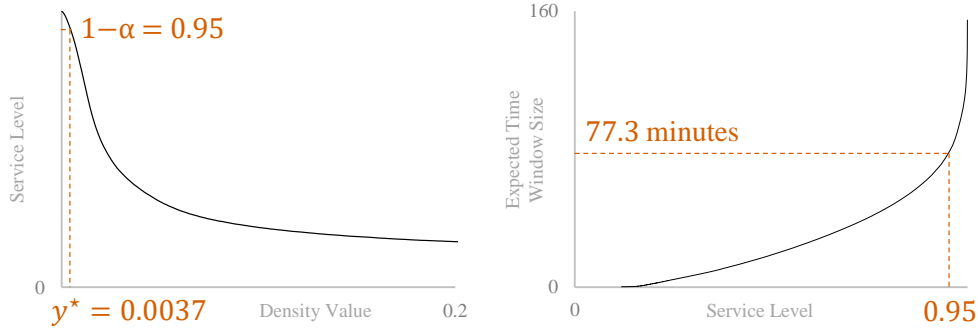

Algorithm 1 Time Window Heuristic

1: **input:** step size η

2: $y \leftarrow \bar{y}$

3: **repeat**

4:     **for all** $s \in \mathcal{S}$ **do**

5:         **if** $y \leq f_{X(s)}(\bar{x}(s))$ **then**

6:             $l(s) \leftarrow \min \mathcal{X}(s,y)$ and $u(s) \leftarrow \max \mathcal{X}(s,y)$

7:         **else**

8:             $l(s) \leftarrow \bar{x}(s)$ and $u(s) \leftarrow \bar{x}(s)$

9:     $y \leftarrow y - \eta$

10: **until** $\mathbb{P}\{l(S) \leq X(S) \leq u(S)\} \geq 1 - \alpha$

11: **output:** time windows $(l(s), u(s))_{s \in \mathcal{S}}$

Figure 4   **Algorithm 1 Mimics Analytical Results**

5.2. Simulation Study

For our computational study, we use the same simulation environment introduced in Ulmer and Thomas (2018). The setting is service routing across the Iowa City area, a region in the Midwest United States with a population of approximately 160,000 and more than 32,000 residential locations. We consider the case of next-day service by a single vehicle with a two-day business model. Requests arrive across an eight-hour capture phase on day one and services are carried out during an eight-hour fulfillment phase on day two. Tours begin and end at the hub of operations, which is located in the northwest corner of the Iowa City area. The experiments of Ulmer and Thomas (2018) reveal that stochastic service times have a relatively small impact on arrival times. Thus, we assume deterministic service times of 15 minutes per customer.

Travel times are based on a constant speed of 30 kilometers per hour and Euclidean distances multiplied by a circuity factor of 1.4 to more closely resemble distances in a road network.

We employ sampling to generate states and arrival time distributions. A capture phase is simulated as follows. Requests arrive randomly across time and space. Request times are uniformly distributed across integers $1, 2, \ldots, 480$ and locations are drawn uniformly from Iowa City location data. Service time windows are communicated at the time of request. A capture phase beings with zero customers accepted for service. As requests arrive, routes are constructed and adjusted via cheapest insertion. This method positions the new customer in the existing route such that the additional travel time is smallest. A request is accepted if the tour duration including service times following cheapest insertion does not exceed the time limit of the fulfillment phase. Otherwise, it is rejected for next-day service, which in practice typically means postponement of service to the following day. Accepted requests constitute a realization s of random state S. For the purposes of the computational study, a state consists of the time of request, the requesting customer's location, the set of customers the firm has already committed to serve, and the route plan. The route plan is the planned visit sequence for the current set of customers, subject to change with new service requests. It is necessary for the implementation of the routing heuristics used in the computational study. Further discussion of route plans can be found in Ulmer et al. (2020).

We repeatedly simulate capture phases to generate $\tilde{n} = 10{,}000$ states $\{s_1, s_2, \ldots, s_{\tilde{n}}\}$. With an average of 20 requests served per day, these states represent approximately 500 days of operation. For each state realization s, we estimate the corresponding arrival time distribution $X(s)$ via an additional $\hat{n} = 10{,}000$ simulations. Each trajectory begins from state s during day one, generates new requests through the end of the capture phase, and routes these requests as described above. In the primary portion of our study, the customer sequence at the end of the capture phase determines routing during the fulfillment phase. In §5.4.5, we allow additional optimization to modify the routing sequence followed during fulfillment. This final routing sequence determines the next-day arrival time for the customer in state s. Denote the simulated arrival times as $\{\dot{x}_1(s), \dot{x}_2(s), \ldots, \dot{x}_{\hat{n}}(s)\}$, which we use to construct an empirical PDF, as in the top row of Figure 3. In our simulation study, arrival time uncertainties stem from route adjustments due to future requests.

Although we do not simulate sources of en-route uncertainty, such as service times or traffic congestion, our methodology can be used to accommodate any source of uncertainty that can be captured in arrival time PDFs. All distributions are publicly available at `https://www.ms.ovgu.de/Research.html`.

We utilize the simulations as input to estimate policy values and service levels. Given time windows $(l(s_i), u(s_i))_{i=1}^{\tilde{n}}$, we estimate the objective value as

$$\frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} u(s_i) - l(s_i). \tag{5}$$

We estimate the corresponding service level as

$$\frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \frac{1}{\hat{n}} \sum_{j=1}^{\hat{n}} 1\Big\{ \dot{x}_j(s_i) \in [l(s_i), u(s_i)] \Big\}, \tag{6}$$

where $1\{\cdot\}$ is an indicator function returning one if the arrival time falls within the service time window and zero otherwise. In our study, we set $\alpha = 0.05$ to correspond with the common industry practice of 95 percent on-time service (Keller 2020, Schasche 2021).

5.3. Benchmarks

We gauge our method for setting service time windows against a dual bound and three policies. We calculate a Lagrangian dual bound by relaxing the service level constraint to the objective. Via standard Lagrangian theory, for any $\gamma \geq 0$, $v(\gamma) = \min\{\mathbb{E}_{\mathcal{S}}[u(S) - l(S)] + \gamma(1 - \alpha - \mathbb{P}\{l(S) \leq X(S) \leq u(S)\}) : u(s) \geq l(s), s \in \mathcal{S}\}$ is less than or equal to the value of an optimal solution to Problem (1). The relaxation is separable by state. Let $v_s(\gamma) = \min\{u(s) - l(s) - \gamma(F_{X(s)}(u(s)) - F_{X(s)}(l(s))) : u(s) \geq l(s)\}$. Then, $v(\gamma) = \gamma(1 - \alpha) + \sum_{s \in \mathcal{S}} \mathbb{P}\{S = s\}v_s(\gamma)$. For a given value of γ, we identify $v_s(\gamma)$ via enumeration. As time is discretized by minute across eight hours, we consider all time windows in the set $\{(l(s), u(s)) \in \{0, 1, \ldots, 480\}^2 : u(s) \geq l(s)\}$. We identify the largest dual bound by enumerating γ. We begin at zero, increase by sufficiently small increments of one, and stop when the bound no longer increases.

Our second benchmark explores the effectiveness of setting time windows via Algorithm 1 when information is limited. For example, when it is not possible to explicitly estimate the arrival time distribution, one may have to rely on moment information to perform the estimation. For each state, we use the first two moments from our arrival time simulations to estimate parameters for a symmetric triangular arrival

time distribution, which we use to set time windows via Equations (2) and (3), estimating objective values and service levels via simulation, as described in §5.2. We refer to this policy as "Limited" and note that experiments with truncated normal distributions yield comparable performance.
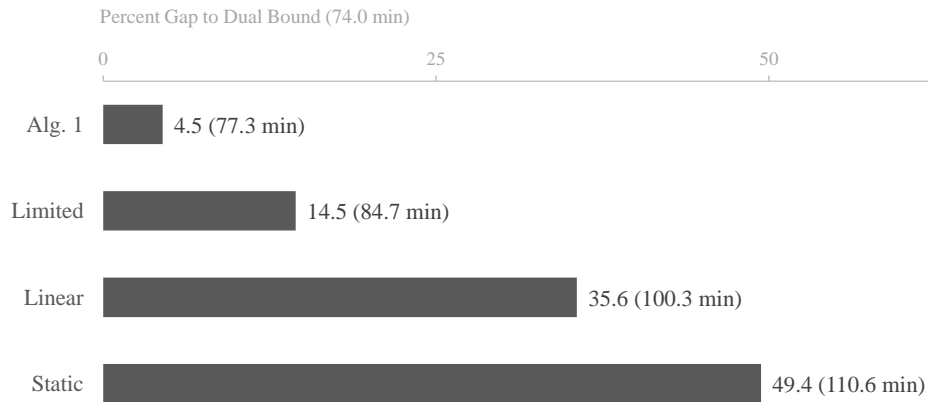
The third benchmark exploits an observation from the literature that time window size tends to decrease with time of request (Ulmer and Thomas 2018, Köhler et al. 2020). Motivated by this, we center a time window around the mean of the empirical PDF, then set the width per a decreasing linear function of request time. Let $t(s)$ be the request time in state s and let $\tau \geq 0$ be a tuning parameter. Then, the time window width for state s is $\tau(480 - t(s))/480$ minutes. This fixes the width to zero at the end of the capture phase when $t(s) = 480$. Otherwise, widths increase with larger choices of τ. We require time windows be no larger than 480 minutes and lie within the 480-minute operating horizon. Thus, some time windows may be truncated. We adjust τ to achieve the smallest expected time window size such that the service level is at least $1 - \alpha$. As before, we estimate objective and constraint values via simulation. We refer to this policy as "Linear."

Finally, the fourth benchmark mirrors current practice by offering a fixed-size service time window to each customer (Weiss 2012). Centering time windows around means, and requiring time windows to begin and end within the 480-minute operating horizon, we identify the smallest range such that the service level is at least $1 - \alpha$, simulating to estimate objective and constraint values. We denote this policy as "Static."

5.4. Discussion

Our discussion examines policy performance across various metrics of managerial interest related to convenience and reliability: time window size, time window violation, and how arrival time variability, time of request, location of request, and routing optimization impact time windows. Our conclusions are presented in the context of arrival time uncertainties due to insertion routing. At the conclusion of this section we demonstrate that our insights are also valid when logistics are managed via more sophisticated optimization.

5.4.1. Time Window Size Figure 5 displays policy quality as measured by the objective and relative to the Lagrangian dual bound. For each policy, the figure shows the gap as a percentage of the dual bound. Notably, setting time windows via Algorithm 1, as depicted in Figure 4, results in a gap of less than five percent, or just over three minutes. Even with limited information, the expected time window size increases

Percent Gap to Dual Bound (74.0 min)

0	25	50

Alg. 1 — 4.5 (77.3 min)

Limited — 14.5 (84.7 min)

Linear — 35.6 (100.3 min)

Static — 49.4 (110.6 min)

Figure 5   Comparison of Policy Values

Note. Each policy value is displayed as a percent gap to the dual bound, with average objective values in minutes displayed in parentheses. The dual bound is 74.0 minutes.

by less than eight additional minutes. Although the requirements for Theorem 1 to yield an optimal policy are not entirely aligned with real-life circumstances, this result demonstrates that the theory transfers well to practice. In contrast, Figure 5 suggests that choosing time windows as a linear function of time of request performs poorly, with a gap of more than 35 percent relative to the dual bound. Further, our methods generate substantial improvement relative to industry practice. Offering all customers a fixed-width time window is far from optimal, with nearly a 50 percent duality gap and an increase of more than 30 minutes in expected time window size relative to our best method.

As all of the policies satisfy the $1 - \alpha$ service-level threshold, these results suggest that firms setting time windows via time of request, as well as firms using static time windows, can significantly increase customer convenience without sacrificing reliability. In particular, the results underscore the importance of Insight 1, that time windows should be customer-specific and tailored to arrival time distributions. Even when firms lack the data necessary or the computational resources required to fully characterize arrival time distributions, the Limited policy demonstrates that effective decisions can still be made. For most firms, estimating means and variances can be accomplished via supervised learning (Ulmer and Thomas 2018), and this will yield significantly better time windows than the industry standard. We highlight this benefit in Insight 3.

Insight 3 (Arrival Time Distributions) *Even rough estimates of arrival time distributions can lead to significant improvements in expected time window size.*

5.4.2.   Time Window Violation In addition to expected time window size, service quality is an important managerial concern. Much like the classical inventory service level metric captures stockout occurrence without regard to stockout quantity, the service level constraint in Problem (1) measures whether or not a time window violation occurs, but not the magnitude of the violation. Because the severity of an early or late arrival has managerial importance, Figure 6 characterizes for each policy the distribution of expected time window violations. Each horizontal histogram displays the number of states, across all simulations, for which the expected violation was five minutes or larger. The figure also displays the average and maximum expected violations. For reasons of scale, we do not display frequencies for expected violations of five minutes and less. For each policy, these cases account for at least 93 percent of the simulations.

Figure 6 reveals that the average expected violation is similar for each policy, between 72 and 90 seconds. The policies are more readily differentiated when we look to the maximum expected violations. At worst, Algorithm 1 is 48 minutes outside the communicated time window. In contrast, worst-case values for the other policies range from nearly two hours to nearly three hours. One explanation for the difference in performance is that Algorithm 1 accounts for the skewness of the arrival time distribution, whereas the other policies center time windows about the mean. Even though the average time window sizes of Algorithm 1 are smaller than those of the other policies, deviations in the case of missed time windows are not as severe. Thus, our method not only increases convenience with reduced time window sizes, it also posts favorable performance when a commitment is not kept.

5.4.3.   Arrival Time Variability Beyond policy performance with respect to convenience and reliability, we seek to better understand the characteristics of the time windows set by Algorithm 1. Figure 7 examines the relationship between variability in arrival time and time window width. The heat map displays results across all simulations, marking areas of high concentration with darker colors and areas of lower concentration with lighter colors. The overall trend suggests that as standard deviation of arrival time increases, time windows widen, generally aligning with the intuition that more uncertainty warrants a larger range for service.
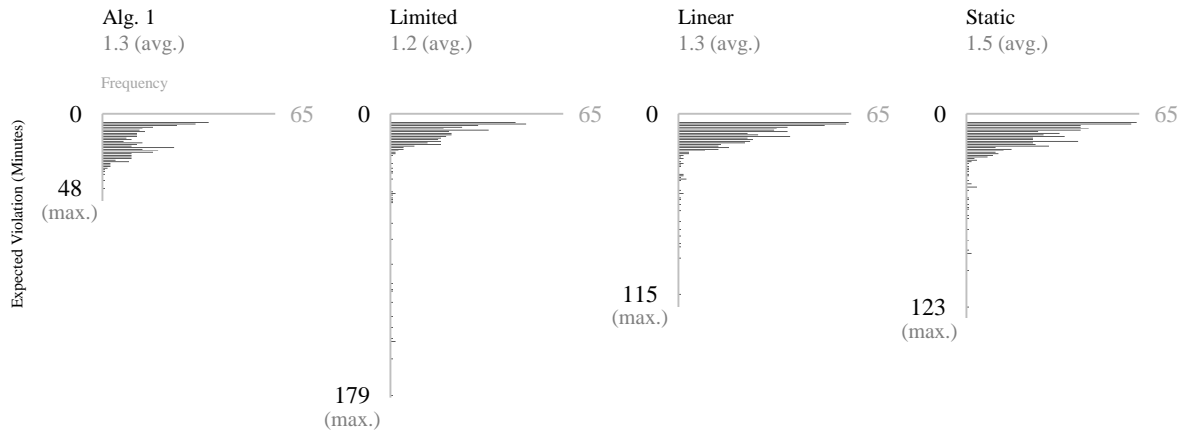
Figure 6   Magnitude of Time Window Violations
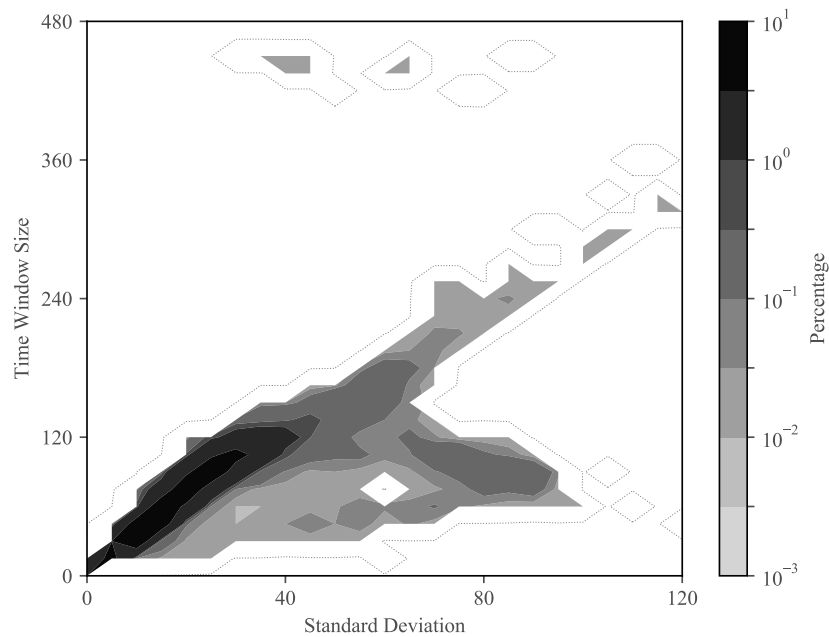
**Figure 7   Impact of Arrival Time Variability on Time Window Size**

Although Figure 7 shows strong positive correlation between arrival time variability and time window size, a closer look at outliers is instructive. In the bottom-right portion of the figure, we find high-variance distributions assigned small time windows, while in the top-left we find low-variance distributions with large time windows. Circumstances leading to these cases are displayed in the top row of Figure 3, the center and right PDFs, respectively. Though the bulk of the probability mass in the top center distribution is

concentrated at the end of the service horizon, a small portion of the mass is situated toward the beginning. This reflects the strong likelihood of receiving service toward the end of a route, with a small chance of being among the first in the delivery sequence. Despite the top center distribution's heavy right skewness, the common density value of $y^\star = 0.0037$ is small enough to capture the left-most mass, thus assigning a nearly eight-hour time window to a low-variance customer. In contrast, though the top right PDF in Figure 3 is also right-skewed, it is appreciably more bimodal than the top center PDF. However, its probability mass is spread such that the chances of early arrival are all less than $y^\star = 0.0037$, thus leading to a much smaller time window for a high-variance customer.

Each PDF presents difficulties: either the time window is unnecessarily large or the likelihood of meeting the time window is relatively small, both leading to potential customer dissatisfaction. Following this line of thinking to an extreme case, where arrival time variability is so high that the distribution approaches a uniform PDF, a large enough $y^\star$ will assign a time window width of zero to this ultra-high-variance customer, a commitment that cannot be met. Though this scenario is a concern in principle, our results suggest it occurs primarily when service levels are impractically low, in which case $y^\star$ may even dominate distributions with moderate variability. In the context of lead time quotation, Spearman and Zhang (1999) raise a similar concern, suggesting that different service metrics may be warranted. However, as Figure 7 illustrates, when the $1 - \alpha$ threshold is set high per industry standards, the vast majority of customers receive time windows proportional to their arrival time variability. Further, we know from Figure 6 that large time window violations are rare. Thus, these extreme cases are unlikely to lead to poor management practices.

5.4.4. Time and Location of Requests It is instructive to consider the impact of variability on time window size in conjunction with time of request and the likelihood of the customer's request to be accepted for day-two service. The left side of Figure 8 displays the average time window size across the first six hours of the capture phase (the point at which routes are typically full) when setting time windows via Algorithm 1. The right side shows the percentage of requests accepted by time of request across the same period. Taken together, Figure 7 and the left portion of Figure 8 suggest variability in arrival time is negatively correlated with time of request. In other words, customers who request service later in the capture phase often receive
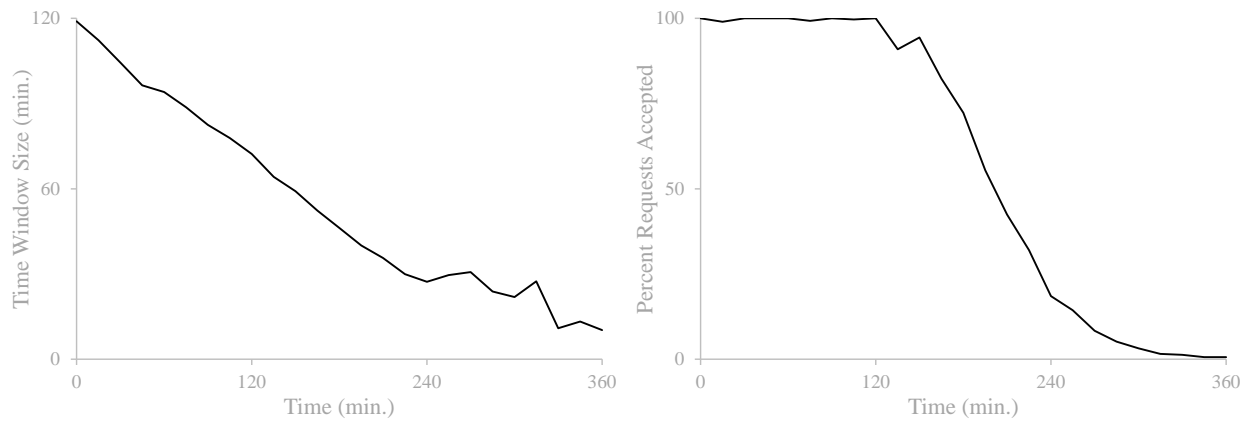
Figure 8   Time Window Sizes and Accepted Requests Across Capture Phase

a smaller time window than customers who request earlier, largely because routing uncertainty is smaller when the potential for additional customers and further route adjustments is lower. As we will show, request location together with temporal considerations gives a more complete picture of factors influencing time window size.

Although on the surface it may appear there is convenience to be had by requesting later, customers who follow such a strategy may not receive next-day service. The right side of Figure 8 shows that over the first two hours of capture, most requests are accepted. Then, as routes fill and the driver's temporal capacity is consumed, later-in-the-day requests cannot always be feasibly serviced, leading to fewer and fewer acceptances as the capture phase progresses. In practice, requests rejected for next-day service would likely be considered early requests for service two days out, and consequently receive a large time window. Thus, though earlier service requests tend to receive larger time windows, likelihood of next-day service can be significantly higher relative to customers who request later.

Initially, differences in time window size by time of request may be cause for concern because customers are not treated equitably with respect to convenience. However, we see similar situations in other aspects of business. Just as late arrival to a store could result in marked down leftovers or a stockout, a later request could lead to a smaller time window, postponement of service, or no inventory left at all. Further, compared to the industry standard static policy, which according to Figure 5 gives each customer a time window of just over 110 minutes, the left side of Figure 8 shows that when time windows are chosen via Algorithm 1,

the vast majority of customers receive smaller time windows. Consequently, even when time windows are tailored to individual customers, collective convenience is not only improved, but individual convenience is very often higher as well, all without compromising service. We summarize the managerial implications for setting time window size based on time of request in Insight 4.

Insight 4 (Time of Request) *Assigning larger time windows to earlier requests and smaller time windows to later requests might reduce expected time-window violations.*

Next, we examine the role of location on time window size. Figure 9 displays the Iowa City area on an 10x10 grid with the delivery depot represented as a red circle in the upper-left. The heat map displays time window sizes chosen by Algorithm 1, with lighter colors representing regions with smaller time windows, on average, and darker colors areas with larger time windows, on average.

The pattern in Figure 9 is clear: locations closest to the depot enjoy smaller time windows relative to locations further away. Because of their proximity to the depot, locations in the upper-left are typically sequenced toward the beginning or end of a service route. Thus, arrival at the start or finish of a workday is predictable and time windows are consequently narrower. In contrast, locations positioned beyond to/from trips to the depot experience more arrival time volatility. These customers often fall in the middle of a route and may appear earlier or later in the delivery sequence depending on where and when other service requests are realized. Because arrival times at these locations are less predictable, the assigned service time windows are therefore larger. We summarize the managerial implications for setting time window size based on location of request in Insight 5.

Insight 5 (Location of Request) *Assigning larger time windows to customers further from the depot of operation and smaller time windows to closer customers can reduce time-window violations.*

We end this part of the discussion by exploring the joint impact of time and location of request on time window size. The heat map in Figure 10 illustrates the relationship, where the horizontal axis tracks time of request in minutes and the vertical axis measures travel distance from the depot in kilometers. The shade of each tile represents the average time window size selected by Algorithm 1 across all simulations with
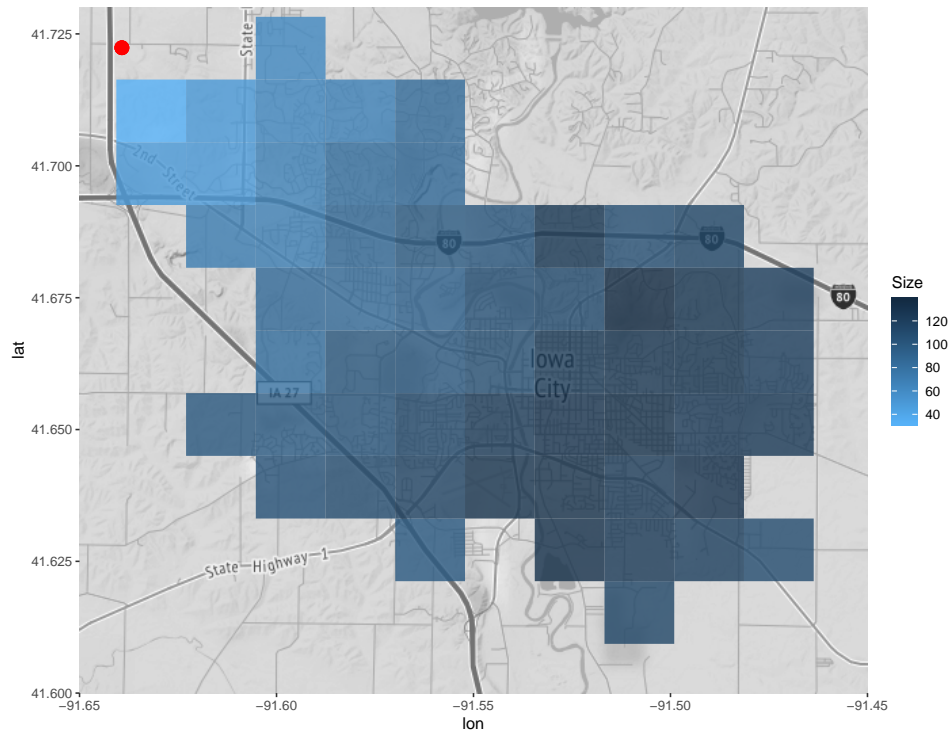
Figure 9   Average Time Window Size by Location

the specified time and location. Darker areas mark larger time windows and lighter areas mark smaller time

windows. Unmarked tiles contain fewer than 10 observations and are omitted from the analysis.

Figure 10 largely confirms previous observations, that time window size decreases with time of request

and increases with distance from the depot. The notable exception occurs in the bottom-left, where larger

time windows are assigned to customers in close proximity to the depot who also place early requests. Such

cases are illustrated by the top center and top right PDFs in Figure 3, where the probability masses are pushed

to the beginning and end of the operating horizon. In these scenarios, because the time of request is early, the

state does not contain enough information about other customer requests to reliably predict a position toward

the beginning or end of the delivery sequence, hence the unusually large time window. However, as seen in

the bottom row of tiles, customers with close proximity to the depot who make service requests later in the
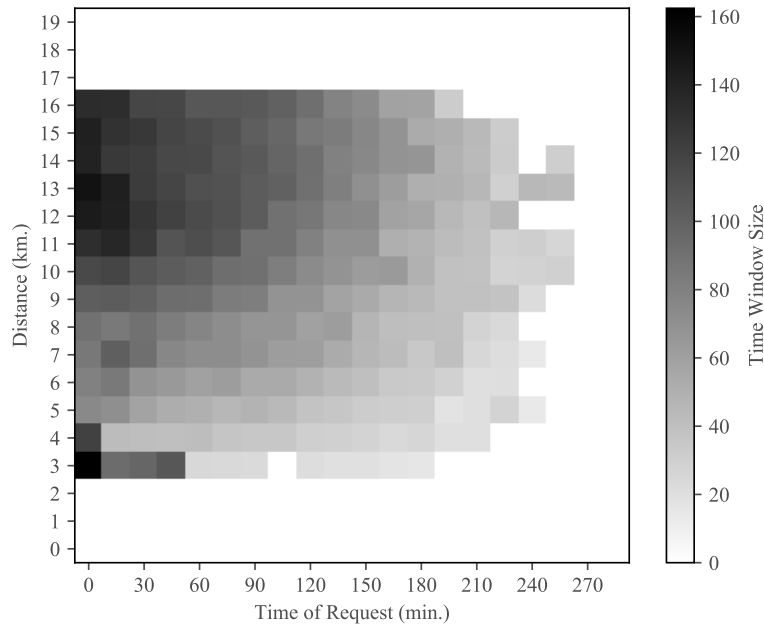
Figure 10   Average Time Window Size By Time of Request and Distance From Depot

day receive smaller time windows. In these cases, position in the delivery sequence is more certain and the decrease in arrival time variability yields a smaller window for service.

5.4.5.   Routing Optimization Arrival time distributions are influenced not only by customer characteristics such as time and location of request, but also by how customers are routed by the firm or by a third party. Up to this point in our analysis, requests are sequenced via cheapest insertion. Because every minute may count, firms may employ more sophisticated procedures to sequence requests. To examine the impact of more complex routing optimization, we perform a two-step procedure to route requests. As before, we employ cheapest insertion to build routes as requests are realized. Then, after all requests are made, we seek to improve the routing sequence via 2-opt local search.

Post-optimization in this fashion yields minor reductions in route duration, on average one to two minutes. Arrival time distributions, however, exhibit major changes. Under insertion, the average standard deviation across arrival time distributions is 23.7 minutes. When 2-opt is employed to improve routes, the average standard deviation increases to 55.2 minutes. For three states, Figure 3 contrasts arrival time PDFs under each routing scheme. PDFs on the top row are tied to insertion routing and PDFs on the bottom row follow

from insertion plus 2-opt improvement. From top to bottom, the standard deviation of arrival time increases one percent (left column), 286 percent (center column), and 31 percent (right column). Collectively, post-optimization of routes doubles arrival time variability in exchange for a relatively small improvement in route duration. The additional arrival time uncertainty stems from how the two routing methods treat the relative order of customer visits. Insertion preserves the relative order. If some request is sequenced before another in a particular iteration of the insertion procedure, that sequence will be maintained in subsequent iterations. Additional customers may be inserted between the two requests, but one will always follow the other. In contrast, because the 2-opt algorithm seeks improvement by swapping customer positions, the relative order established by insertion may be changed by the post-optimization. Consequently, requests that tend to appear in a particular part of a sequence under insertion routing may now appear more frequently in other parts of the sequence. The result is less predictable sequences and thereby more uncertain arrival times. Though we focus on 2-opt improvement for computational tractability (we consider 10,000 arrival time realizations in each of 10,000 states), we would expect to see similar results for any optimization method that has the potential to make large changes to the customer sequence.

Increases in arrival time variability in turn impact time window sizes. The quantities in Figure 5 update as follows: Lagrangian dual bound 120.8 minutes, Algorithm 1 37.3 percent gap (165.8 minutes), Limited 56.8 percent gap (189.4 minutes), Linear 115.3 percent gap (260.1 minutes), and Static 125.0 percent gap (271.8 minutes). Even though the heuristic motivated by Theorem 1 substantially outperforms the other methods, its average time window size more than doubles compared to that of Figure 5. Further, in stark contrast to the quantities in Figure 5, the dual bound with 2-opt local search exceeds the average time window size of all methods when customers are sequenced by insertion alone. Thus, sequence changes due to 2-opt make it impossible to realize even the poor performance of the fixed-size time windows under insertion routing. More generally, to achieve a given service level, expected time window size must increase substantially to accommodate the added arrival time variability associated with more complex routing optimization. While optimization may improve routing efficiency, it does so by forfeiting customer convenience. We summarize this tradeoff in Insight 6.

Insight 6 (Routing Optimization and Convenience) *Cost savings afforded by routing efficiency can lead to a potential loss of customer convenience.*

Although Insights 4 and 5 are presented in the context of insertion routing, they remain valid when routes are improved via 2-opt local search. Despite the increase in arrival time volatility and the jump in objective value, time and location of requests still impact time window sizes as described in §5.4.4. Further, the maximum expected time window violation reported for Algorithm 1 in Figure 6 does not increase.

6. Conclusion

We study the optimal setting of service time windows. Under assumptions on the shape of arrival time distributions, we show how to identify minimum expected-width time windows subject to a constraint on service level. Our analysis demonstrates that, under the assumptions of our problem description, time windows should be set via a common density value across arrival time distributions. For the case of general distributional forms, our analytical results motivate a heuristic that significantly outperforms benchmarks, including the industry standard. We glean six managerial insights that might inform practice: (i) The beginning and end of a time window should reflect the shape of the arrival time distribution. Consequently, industry should revisit the practice of assigning uniform fixed-width time windows to all customers and consider replacing them with the assignment of customer-specific time windows. (ii) Higher service levels likely require wider time windows. At a given service level, optimization is required to size time windows. (iii) Even rough estimates of arrival time distributions can lead to significant improvements in expected time window size. (iv) Assigning larger time windows to earlier requests and smaller time windows to later requests may improve performance. (v) Assigning larger time windows to customers further from the depot of operation and smaller time windows to closer customers may improve performance. (vi) Routing efficiency might lead to a loss of customer convenience.

We suggest several directions for future research. First, as is still common in many service routing applications, our work assumes customer acceptance and routing decisions are exogenous to the process of setting time windows. Future work could consider the integration of these decisions, drawing on our research as a starting point. Second, Vareias et al. (2019) allows for customer-specific service levels. Future work

might consider an extension of our results to that case. Third, we do not consider the possibility of revising time windows after they have been set. Real-time adjustments are already used in practice and are the subject of recent research (Dalmeijer et al. 2019). However, there does not seem to be academic work examining the setting of time windows at the time of the request with the possibility of future updates. Our results in this paper may provide insight to such an analysis as well as a benchmark. Fourth, future work might also examine the conundrum posed by the bimodal arrival time distributions discussed in §5.4.3. Such research could consider the possibility of communicating two time windows to the customer, one toward the beginning of the operating horizon and a second toward the end. In some circumstances, this may be preferable to service during one much larger time window, or initial communication of a small time window followed by a drastic correction. Fifth, our methodology relies on knowledge of arrival time distributions. Although we show that approximating PDFs via only two moments leads to competitive results, selecting time windows via the full shape of distributions is better. In practice, however, the data needed to construct a fully-featured PDF may be lacking. Future research might explore procedures to build detailed distributions when input data is limited. Sixth, our experiments reveal that routing optimization has a significant role on expected time window size. Future work should aim to identify routing methods that yield efficient routes and low arrival time volatilities. Seventh, as our work has shown, even without integrated routing and time window decisions, our policy can achieve solutions with small time window violations. Future work may build on this, optimizing both decisions together. A first step might be to partition the fulfillment into smaller portions, determine the assignment of customers to these partitions based on the previous time window decisions, and use our policy for each portion. Finally, future research might consider the role of pricing and incentives. For instance, while we focus on minimizing expected time window width to achieve a particular service level, one might also consider offering a customer a lower price to encourage the customer to choose a wider time window, potentially leveraging our analysis surrounding arrival time distributions. Larger windows would allow for more customers to be served on a given day without decreasing the firm's likelihood of meeting service commitments.

Acknowledgments

References

Agatz, Niels, Yingjie Fan, Daan Stam. 2021. The impact of green labels on time slot choice and operational sustainability. *Production and Operations Management* **30**(7) 2285–2303.

Amazon. 2021. Amazon last mile routing research challenge. `https://routingchallenge.mit.edu/`, Accessed on July 27, 2023.

Apte, Aruna, Uday M Apte, Nandagopal Venugopal. 2007. Focusing on customer time in field service: A normative approach. *Production and Operations Management* **16**(2) 189–202.

Bazaraa, M., H. Sherali, Shetty C. 2006. *Nonlinear Programming Theory and Algorithms*. 3rd ed. John Wiley & Sons, Inc., Hoboken, NJ.

Beskers, E. 2022. Time window assignment: A case study in retail distribution. Master's thesis, Eindhoven University of Technology, Eindhoven, the Netherlands.

Campbell, Ann Melissa, Martin WP Savelsbergh. 2005. Decision support for consumer direct grocery initiatives. *Transportation Science* **39**(3) 313–327.

Dalmeijer, Kevin, Remy Spliet. 2018. A branch-and-cut algorithm for the time window assignment vehicle routing problem. *Computers & Operations Research* **89** 140–152.

Dalmeijer, Kevin, Remy Spliet, Albert Wagelmans. 2019. Dynamic time window adjustment. Tech. Rep. EI2019-22, Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam. URL `http://hdl.handle.net/1765/116533`.

Ehmke, Jan F, Ann M Campbell. 2014. Customer acceptance mechanisms for home deliveries in metropolitan areas. *European Journal of Operational Research* **233**(1) 193–207.

Ellis, Blake. 2011. Waiting for the cable guy is costing us $38 billion. URL `http://money.cnn.com/2011/11/03/pf/cost_of_waiting/`. [Online; accessed 12-December-2022].

Flinterman, J. B. 2022. Assigning delivery time windows before all customers are known. Master's thesis, Department of Industrial Engineering & Innovation Sciences, Eindhoven University of Technology, Eindhoven, the Netherlands.

Gel, Esma S, Pinar Keskinocak, Tuba Yilmaz. 2020. Dynamic price and lead time quotation strategies to match demand and supply in make-to-order manufacturing environments. *Women in Industrial and Systems Engineering*. Springer, 541–560.

Hafızoğlu, A Baykal, Esma S Gel, Pınar Keskinocak. 2016. Price and lead time quotation for contract and spot customers. *Operations Research* **64**(2) 406–415.

Hoogeboom, Maaike, Yossiri Adulyasak, Wout Dullaert, Patrick Jaillet. 2021. The robust vehicle routing problem with time window assignments. *Transportation Science* **55**(2) 395–413.

Jabali, Ola, Roel Leus, Tom Van Woensel, Ton de Kok. 2015. Self-imposed time windows in vehicle routing problems. *OR Spectrum* **37**(2) 331–352.

Jalilvand, Mahdi, Mahdi Bashiri, Erfaneh Nikzad. 2021. An effective progressive hedging algorithm for the two-layers time window assignment vehicle routing problem in a stochastic environment. *Expert Systems with Applications* **165** 113877.

Kahvecioglu, Gökçe, Barış Balcıoglu. 2016. Coping with production time variability via dynamic lead-time quotation. *OR Spectrum* **38**(4) 877–898.

Keller, Aaron. 2020. Importance of industry standards & on-time delivery. URL `https://blog.kellerlogistics.com/importance-of-industry-standards-on-time-delivery`. [Online; accessed 04-July-2023].

Keskinocak, Pinar, Sridhar Tayur. 2004. Due date management policies. D. Simchi-Levi, S.D. Wu, Z.J. Shen, eds., *Handbook of quantitative supply chain analysis*, *International Series in Operations Research & Management Science*, vol. 74. Springer, Boston, 485–554.

Köhler, Charlotte, Jan Fabian Ehmke, Ann Melissa Campbell. 2020. Flexible time window management for attended home deliveries. *Omega* **91** 102023.

Lim, Stanley Frederick WT, Qingchen Wang, Scott Webster. 2023. Do it right the first time: Vehicle routing with home delivery attempt predictors. *Production and Operations Management* **32**(4) 1262–1284.

Madsen, Oli B. G., Kim Tosti, Jan Vælds. 1996. A heuristic method for dispatching repair men. *Annals of Operations Research* **61**(1) 213–226.

Martins, Sara, Manuel Ostermeier, Pedro Amorim, Alexander Hübner, Bernardo Almada-Lobo. 2019. Product-oriented time window assignment for a multi-compartment vehicle routing problem. *European Journal of Operational Research* **276**(3) 893–909.

Neves-Moreira, Fábio, Diogo Pereira Da Silva, Luís Guimarães, Pedro Amorim, Bernardo Almada-Lobo. 2018. The time window assignment vehicle routing problem with product dependent deliveries. *Transportation Research Part E: Logistics and Transportation Review* **116** 163–183.

OptimoRoute. 2022. Use on-time delivery metrics to improve customer satisfaction. URL `https://optimoroute.com/on-time-delivery-metric/`. [Online; accessed 21-August-2023].

Prisco, Jacopo. 2017. Why UPS trucks (almost) never turn left. URL `https://www.cnn.com/2017/02/16/world/ups-trucks-no-left-turns/index.html`. [Online; accessed 12-December-2022].

Schasche, Stefan. 2021. Die Telekom stellt den Service in den Mittelpunkt. URL `https://www.wuv.de/Archiv/Die-Telekom-stellt-den-Service-in-den-Mittelpunkt`. [Online; accessed 04-July-2023].

Spearman, Mark L, Rachel Q Zhang. 1999. Optimal lead time policies. *Management Science* **45**(2) 290–295.

Spliet, Remy, Said Dabia, Tom Van Woensel. 2018. The time window assignment vehicle routing problem with time-dependent travel times. *Transportation Science* **52**(2) 261–276.

Spliet, Remy, Guy Desaulniers. 2015. The discrete time window assignment vehicle routing problem. *European Journal of Operational Research* **244**(2) 379–391.

Spliet, Remy, Adriana F. Gabor. 2015. The time window assignment vehicle routing problem. *Transportation Science* **49**(4) 721–731.

Subramanyam, Anirudh, Akang Wang, Chrysanthos E Gounaris. 2018. A scenario decomposition algorithm for strategic time window assignment vehicle routing problems. *Transportation Research Part B: Methodological* **117** 296–317.

Ulmer, Marlin W, Justin C Goodson, Dirk C Mattfeld, Barrett W Thomas. 2020. On modeling stochastic dynamic vehicle routing problems. *EURO Journal on Transportation and Logistics* **9**(2) 100008.

Ulmer, Marlin W, Barrett W Thomas. 2018. Enough waiting for the cable guy - estimating arrival times for service vehicle routing. *Transportation Science* **53**(3) 897–916.

Vareias, Anastasios D., Panagiotis P. Repoussis, Christos D. Tarantilis. 2019. Assessing customer service reliability in route planning with self-imposed time windows and stochastic travel times. *Transportation Science* **53**(1) 256–281.

Visser, Thomas, Martin Savelsbergh. 2019. Strategic time slot management: A priori routing for online grocery retailing. Tech. rep.

Waßmuth, Katrin, Charlotte Köhler, Niels Agatz, Moritz Fleischmann. 2023. Demand management for attended home delivery–a literature review. *European Journal of Operational Research* .

Weiss, Darren. 2012. How Comcast killed the four-hour service window. URL `https://fsd.servicemax.com/2012/01/18/how-comcast-killed-the-four-hour-service-window/`. [Online; accessed 15-June-2020].

Yang, Xinan, Arne K Strauss, Christine SM Currie, Richard Eglese. 2014. Choice-based demand management and vehicle routing in e-fulfillment. *Transportation Science* **50**(2) 473–488.

Yu, Xian, Siqian Shen, Babak Badri-Koohi, Haitham Seada. 2023. Time window optimization for attended home service delivery under multiple sources of uncertainties. *Computers & Operations Research* **150** 106045.

Zhang, Zhenzhen, Mengyang Liu, Andrew Lim. 2015. A memetic algorithm for the patient transportation problem. *Omega* **54** 60–71.

Appendix

Proof Summary. The proof of Theorem 1 identifies a saddle point of Problem (1), a necessary and sufficient condition for an optimal solution. We proceed by stating the Lagrangian dual of Problem (1), by identifying candidate solutions for Problem (1) and the Lagrangian dual, then by demonstrating that these solutions constitute a saddle point.

Lagrangian Dual Problem. Let $\boldsymbol{l} = (l(s))_{s \in \mathcal{S}}$ and $\boldsymbol{u} = (u(s))_{s \in \mathcal{S}}$ be time window vectors across all state realizations. Let λ be a penalty multiplier. Formulate a Lagrangian of Problem (1) by relaxing the service level constraint to the objective. Using the law of total probability and arrival time CDFs, we have

$$L(\boldsymbol{l}, \boldsymbol{u}, \lambda) = \sum_{s \in \mathcal{S}} \mathbb{P}\{S = s\}(u(s) - l(s)) + \lambda \left[1 - \alpha - \left(\sum_{s \in \mathcal{S}} \mathbb{P}\{S = s\} \left[F_{X(s)}(u(s)) - F_{X(s)}(l(s)) \right] \right) \right]. \quad \text{(A1)}$$

Rewrite the Lagrangian by collecting terms around states. Let

$$L_s(l(s), u(s), \lambda) = u(s) - l(s) - \lambda \left(F_{X(s)}(u(s)) - F_{X(s)}(l(s)) \right). \quad \text{(A2)}$$

Then,

$$L(\boldsymbol{l}, \boldsymbol{u}, \lambda) = \lambda(1 - \alpha) + \sum_{s \in \mathcal{S}} \mathbb{P}\{S = s\} L_s(l(s), u(s), \lambda). \quad \text{(A3)}$$

The Lagrangian dual problem seeks a penalty multiplier that maximizes the minimum value of the Lagrangian:

$$\max_{\lambda \geq 0} \left\{ \min_{\boldsymbol{l} \leq \boldsymbol{u}} \{ L(\boldsymbol{l}, \boldsymbol{u}, \lambda) \} \right\} = \max_{\lambda \geq 0} \left\{ \lambda(1 - \alpha) + \sum_{s \in \mathcal{S}} \mathbb{P}\{S = s\} \left[\min_{l(s) \leq u(s)} \{ L_s(l(s), u(s), \lambda) \} \right] \right\}. \quad \text{(A4)}$$

Thus, for a given value of λ, the inner minimization may be carried out separately for each state realization.

Candidate Solutions. For any $\lambda > 0$, select time windows as

$$l(s, 1/\lambda) = \begin{cases} \min \mathcal{X}(s, 1/\lambda), & 1/\lambda \leq f_{X(s)}(\bar{x}(s)), \\ \\ \bar{x}(s), & f_{X(s)}(\bar{x}(s)) < 1/\lambda, \end{cases} \quad \text{(A5)}$$

$$u(s, 1/\lambda) = \begin{cases} \max \mathcal{X}(s, 1/\lambda), & 1/\lambda \leq f_{X(s)}(\bar{x}(s)), \\ \\ \bar{x}(s), & f_{X(s)}(\bar{x}(s)) < 1/\lambda. \end{cases} \quad \text{(A6)}$$

We leverage Lemma 1, below, to set λ. Recall that $\bar{y}$ is the largest density value across all PDFs and that $\check{y}$ is the smallest. For $\check{y} < 1/\lambda \leq \bar{y}$, Lemma 1 shows that $\mathbb{P}\{l(S, 1/\lambda) \leq X(S) \leq u(S, 1/\lambda)\}$ is strictly decreasing in $1/\lambda$ and a bijection from $(\check{y}, \bar{y}]$ to $[0, 1)$. Consequently, for any $0 < \alpha \leq 1$, there is a unique $\lambda_\alpha > 0$ such that $\mathbb{P}\{l(S, 1/\lambda_\alpha) \leq X(S) \leq u(S, 1/\lambda_\alpha)\} = 1 - \alpha$. Letting $\boldsymbol{l}(1/\lambda_\alpha) = (l(s, 1/\lambda_\alpha))_{s \in \mathcal{S}}$ and $\boldsymbol{u}(1/\lambda_\alpha) = (u(s, 1/\lambda_\alpha))_{s \in \mathcal{S}}$, the candidate solutions for the Lagrangian dual and Problem (1) are λ_α and $(\boldsymbol{l}(1/\lambda_\alpha), \boldsymbol{u}(1/\lambda_\alpha))$, respectively.

Lemma 1 (Shape of Service Level) *Let $\lambda > 0$. $\mathbb{P}\{l(S, 1/\lambda) \leq X(S) \leq u(S, 1/\lambda)\}$ is strictly decreasing in $1/\lambda$ and a bijection from $(\check{y}, \bar{y}]$ to $[0, 1)$. For all $0 < 1/\lambda \leq \check{y}$, $\mathbb{P}\{l(S, 1/\lambda) \leq X(S) \leq u(S, 1/\lambda)\} = 1$. For all $1/\lambda > \bar{y}$, $\mathbb{P}\{l(S, 1/\lambda) \leq X(S) \leq u(S, 1/\lambda)\} = 0$.*

Proof. Rewriting $\mathbb{P}\{l(S, 1/\lambda) \leq X(S) \leq u(S, 1/\lambda)\}$ yields

$$\sum_{s \in \mathcal{S}} \mathbb{P}\{S = s\} \left[F_{X(s)}(u(s, 1/\lambda)) - F_{X(s)}(l(s, 1/\lambda)) \right]. \tag{A7}$$

We begin by analyzing $F_{X(s)}(l(s, 1/\lambda))$ and $F_{X(s)}(u(s, 1/\lambda))$ for a given state s. Suppose $\mathcal{X}(s)$ is left- and right-bounded and that $\check{y} < f_{X(s)}(\tilde{x}(s))$ and $\check{y} < f_{X(s)}(\hat{x}(s))$. By assumption, $\underline{f}_{X(s)}$ is strictly increasing and continuous from $[\tilde{x}(s), \bar{x}(s)]$ to $[f_{X(s)}(\tilde{x}(s)), f_{X(s)}(\bar{x}(s))]$. Thus, $\underline{f}_{X(s)}^{-1}$ is strictly increasing and a bijection from $[f_{X(s)}(\tilde{x}(s)), f_{X(s)}(\bar{x}(s))]$ to $[\tilde{x}(s), \bar{x}(s)]$. Then, $l(s, 1/\lambda) = \min \mathcal{X}(s, 1/\lambda) = \underline{f}_{X(s)}^{-1}(1/\lambda)$ is strictly increasing in $1/\lambda$ and a bijection from $[f_{X(s)}(\tilde{x}(s)), f_{X(s)}(\bar{x}(s))]$ to $[\tilde{x}(s), \bar{x}(s)]$. It follows that $F_{X(s)}(l(s, 1/\lambda))$ is strictly increasing in $1/\lambda$ from $[f_{X(s)}(\tilde{x}(s)), f_{X(s)}(\bar{x}(s))]$ to $[0, F_{X(s)}(\bar{x}(s))]$ and is therefore an injection from $[f_{X(s)}(\tilde{x}(s)), f_{X(s)}(\bar{x}(s))]$ to $[\tilde{x}(s), \bar{x}(s)]$. Since $F_{X(s)}(l(s, 1/\lambda))$ is strictly increasing, $\min_{1/\lambda} F_{X(s)}(l(s, 1/\lambda)) = F_{X(s)}(l(s, f_{X(s)}(\tilde{x}(s)))) = F_{X(s)}(\tilde{x}(s))$ and $\max_{1/\lambda} F_{X(s)}(l(s, 1/\lambda)) = F_{X(s)}(l(s, f_{X(s)}(\bar{x}(s)))) = F_{X(s)}(\bar{x}(s))$. Since $F_{X(s)}(l(s, 1/\lambda))$ is continuous, its image is $[F_{X(s)}(\tilde{x}(s)), F_{X(s)}(\bar{x}(s))] = [0, F_{X(s)}(\bar{x}(s))]$, which implies that $F_{X(s)}(l(s, 1/\lambda))$ is a surjection. Because $F_{X(s)}(l(s, 1/\lambda))$ is an injection and a surjection, it is a bijection. By construction, $F_{X(s)}(l(s, 1/\lambda)) = F_{X(s)}(\tilde{x}(s)) = 0$ when $1/\lambda < f_{X(s)}(\tilde{x}(s))$ and $F_{X(s)}(l(s, 1/\lambda)) = F_{X(s)}(\bar{x}(s))$ when $f_{X(s)}(\bar{x}(s)) < 1/\lambda$. Recognizing that $u(s, 1/\lambda) = \max \mathcal{X}(s, 1/\lambda) = \overline{f}_{X(s)}^{-1}(1/\lambda)$ when $1/\lambda$ lies between $f_{X(s)}(\hat{x}(s))$ and $f_{X(s)}(\bar{x}(s))$, then analogous arguments show that $F_{X(s)}(u(s, 1/\lambda))$ is strictly decreasing in $1/\lambda$ and a bijection from $[f_{X(s)}(\hat{x}(s)), f_{X(s)}(\bar{x}(s))]$ to $[0, F_{X(s)}(\bar{x}(s))]$. By construction, $F_{X(s)}(u(s, 1/\lambda)) = F_{X(s)}(\hat{x}(s)) = 1$ when $1/\lambda < f_{X(s)}(\hat{x}(s))$ and $F_{X(s)}(u(s, 1/\lambda)) = F_{X(s)}(\bar{x}(s))$ when $f_{X(s)}(\bar{x}(s)) < 1/\lambda$.

Next, examine the difference of CDFs $F_{X(s)}(u(s, 1/\lambda)) - F_{X(s)}(l(s, 1/\lambda))$. Suppose $\check{y} < f_{X(s)}(\tilde{x}(s)) \leq f_{X(s)}(\hat{x}(s))$. For $1/\lambda < f_{X(s)}(\tilde{x}(s))$, the difference is $F_{X(s)}(u(s, 1/\lambda)) - F_{X(s)}(l(s, 1/\lambda)) = 1 - 0 = 1$. For $f_{X(s)}(\tilde{x}(s)) \leq 1/\lambda < f_{X(s)}(\hat{x}(s))$, the difference is $F_{X(s)}(u(s, 1/\lambda)) - F_{X(s)}(l(s, 1/\lambda)) = 1 - F_{X(s)}(l(s, 1/\lambda))$, which is strictly decreasing in $1/\lambda$ and a bijection from $[f_{X(s)}(\tilde{x}(s)), f_{X(s)}(\hat{x}(s))]$ to $[1 - F_{X(s)}(\hat{x}(s)), 1]$. For $f_{X(s)}(\hat{x}(s)) \leq 1/\lambda \leq f_{X(s)}(\bar{x}(s))$, because the composition of bijective functions is bijective, $F_{X(s)}(u(s, 1/\lambda)) - F_{X(s)}(l(s, 1/\lambda))$ is a bijection from $[f_{X(s)}(\hat{x}(s)), f_{X(s)}(\bar{x}(s))]$ to $[0, F_{X(s)}(\hat{x}(s)) - F_{X(s)}(\underline{f}_{X(s)}^{-1}(f_{X(s)}(\hat{x}(s))))]$. Further, because $F_{X(s)}(l(s, 1/\lambda))$ strictly increases in $1/\lambda$ and $F_{X(s)}(u(s, 1/\lambda))$ strictly decreases in $1/\lambda$, $F_{X(s)}(u(s, 1/\lambda)) - F_{X(s)}(l(s, 1/\lambda))$ strictly decreases in $1/\lambda$. For $f_{X(s)}(\bar{x}(s)) < 1/\lambda$, the difference is $F_{X(s)}(u(s, 1/\lambda)) - F_{X(s)}(l(s, 1/\lambda)) = F_{X(s)}(\bar{x}(s)) - F_{X(s)}(\bar{x}(s)) = 0$.

Thus, $F_{X(s)}(u(s, 1/\lambda)) - F_{X(s)}(l(s, 1/\lambda))$ is strictly decreasing in $1/\lambda$ and a bijection from $[f_{X(s)}(\tilde{x}(s)), f_{X(s)}(\bar{x}(s))]$ to $[0,1]$. If $\check{y} < f_{X(s)}(\hat{x}(s)) \leq f_{X(s)}(\tilde{x}(s))$, then analogous arguments show that $F_{X(s)}(u(s, 1/\lambda)) - F_{X(s)}(l(s, 1/\lambda))$ is strictly decreasing in $1/\lambda$ and a bijection from $[f_{X(s)}(\hat{x}(s)), f_{X(s)}(\bar{x}(s))]$ to $[0,1]$. It follows that $F_{X(s)}(u(s, 1/\lambda)) - F_{X(s)}(l(s, 1/\lambda))$ is a bijection from $[\min\{f_{X(s)}(\tilde{x}(s)), f_{X(s)}(\hat{x}(s))\}, f_{X(s)}(\bar{x}(s))]$ to $[0,1]$. If $\check{y} = f_{X(s)}(\tilde{x}(s))$ or $\check{y} = f_{X(s)}(\hat{x}(s))$, then analogous arguments demonstrate that $F_{X(s)}(u(s, 1/\lambda)) - F_{X(s)}(l(s, 1/\lambda))$ is strictly decreasing in $1/\lambda$ and a bijection from $(\check{y}, f_{X(s)}(\bar{x}(s))]$ to $[0,1)$. Further, for $f_{X(s)}(\bar{x}(s)) < 1/\lambda$, the difference is $F_{X(s)}(u(s, 1/\lambda)) - F_{X(s)}(l(s, 1/\lambda)) = 0$.

Finally, we consider the entire function. By construction, for at least one state s in $\mathcal{S}$, $\check{y} = f_{X(s)}(\tilde{x}(s))$ or $\check{y} = f_{X(s)}(\hat{x}(s))$. Similarly, for at least one state s in $\mathcal{S}$, $\bar{y} = f_{X(s)}(\bar{x}(s))$. Because a composition of bijective functions is itself bijective, it follows that Equation (A7) is strictly decreasing in $1/\lambda$ and a bijection from $(\check{y}, \bar{y}]$ to $[0,1)$. Further, Equation (A7) is equal to one for all $0 < 1/\lambda \leq \check{y}$ and is equal to zero for all $1/\lambda > \bar{y}$. Analogous arguments apply when $\mathcal{X}(s)$ is left- or right-unbounded. □

Saddle Point Criteria. Theorem 6.2.5 of Bazaraa et al. (2006) shows that the candidate solutions constitute a saddle point if and only if

1. $\mathbb{P}\{l(S, 1/\lambda_\alpha) \leq X(S) \leq u(S, 1/\lambda_\alpha)\} \geq 1 - \alpha$,

2. $\lambda(1 - \alpha - \mathbb{P}\{l(S, 1/\lambda_\alpha) \leq X(S) \leq u(S, 1/\lambda_\alpha)\}) = 0$,

3. $L(\boldsymbol{l}(1/\lambda_\alpha), \boldsymbol{u}(1/\lambda_\alpha), \lambda_\alpha) = \min\{L(\boldsymbol{l}, \boldsymbol{u}, \lambda_\alpha) : \boldsymbol{l} \leq \boldsymbol{u}\}$, and

4. there is no duality gap.

The first and second conditions hold by the construction of λ_α. Because $\mathbb{P}\{l(S, 1/\lambda_\alpha) \leq X(S) \leq u(S, 1/\lambda_\alpha)\} = 1 - \alpha$, the inequality in the first condition holds at equality and the parenthetical term in the second condition is zero, thus satisfying both conditions. The remainder of this subsection addresses the third and fourth conditions.

Because L decomposes by state, we address the minimization in the third condition separately for each s in $\mathcal{S}$. For all possible values of $\lambda_\alpha > 0$, we identify a closed form solution of $\min\{L_s(l(s), u(s), \lambda_\alpha) : l(s) \leq u(s)\}$. We show that these solutions reduce to Equations (A5) and (A6). Figure A1 provides a visual aid for the analysis. The figure displays $1/\lambda_\alpha$ in relation to $f_{X(s)}$, which we depict via truncated normal PDFs. For each of five cases, the figure indicates how the signs of the partial derivatives change across the critical points of L_s. Critical points occur where the partial derivatives are zero or undefined. Derivatives ar zero where $1/\lambda_\alpha$ intersects $f_{X(s)}$ and undefined at support boundaries (if they exist), where L_s has kinks. By inspecting the signs of the partial derivatives in each case, we identify the
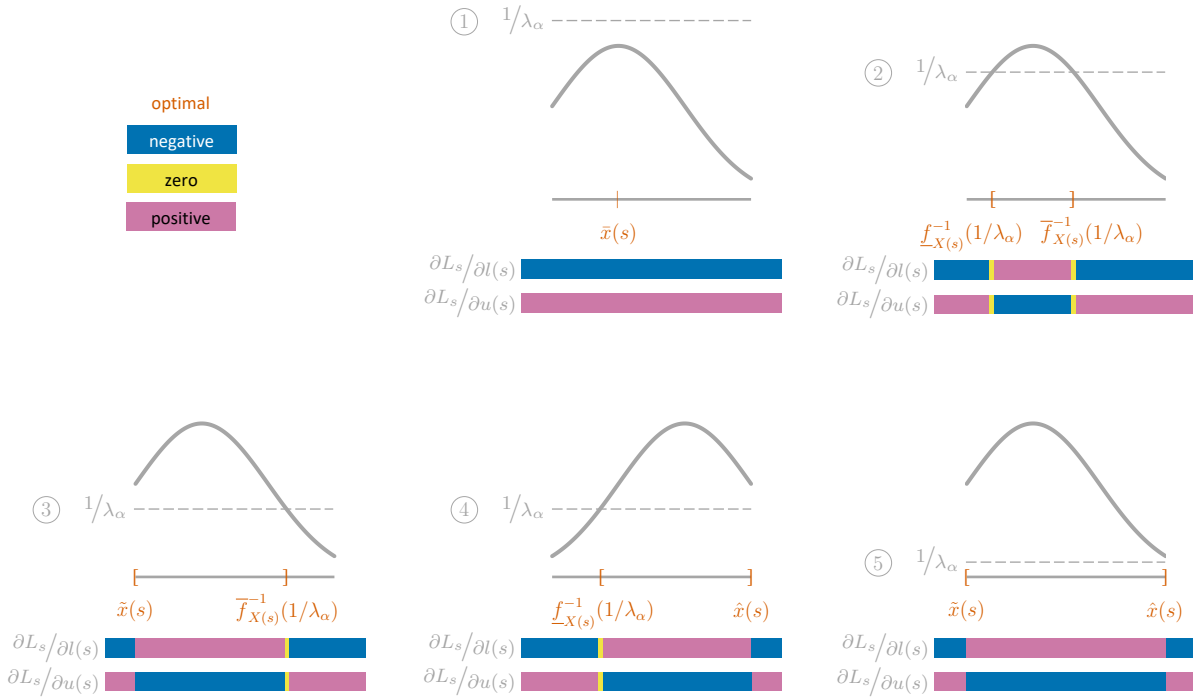
Figure A1   PDFs, Partial Derivatives, and Optimal Time Windows

minimizers of L_s, which are also displayed in the figure. The partial derivatives of L_s with respect to $l(s)$ and $u(s)$ are

$\partial L_s/\partial l(s) = -1 + \lambda_\alpha f_{X(s)}(l(s))$ and $\partial L_s/\partial u(s) = 1 - \lambda_\alpha f_{X(s)}(u(s))$.

Case ① : $f_{X(s)}(\bar{x}(s)) < {}^1\!/_{\lambda_\alpha}$. In this case, ${}^1\!/_{\lambda_\alpha}$ sits above the density value of the mode. The partial derivatives are strictly negative and positive, respectively:

$$\frac{\partial L_s}{\partial l(s)} = -1 + \lambda_\alpha f_{X(s)}(l(s)) < -1 + \frac{f_{X(s)}(l(s))}{f_{X(s)}(\bar{x}(s))} \leq -1 + \frac{f_{X(s)}(\bar{x}(s))}{f_{X(s)}(\bar{x}(s))} = 0 \tag{A8}$$

$$\frac{\partial L_s}{\partial u(s)} = 1 - \lambda_\alpha f_{X(s)}(u(s)) > 1 - \frac{f_{X(s)}(u(s))}{f_{X(s)}(\bar{x}(s))} \geq 1 - \frac{f_{X(s)}(\bar{x}(s))}{f_{X(s)}(\bar{x}(s))} = 0. \tag{A9}$$

The first inequality in each equation follows from the case condition that $\lambda_\alpha < {}^1\!/_{f_{X(s)}(\bar{x}(s))}$. The second inequalities follow from the assumption that the maximum density occurs at the mode. Thus, $f_{X(s)}(l(s)) \leq f_{X(s)}(\bar{x}(s))$ and $f_{X(s)}(u(s)) \leq f_{X(s)}(\bar{x}(s))$. Because $\partial L_s/\partial l(s)$ is everywhere negative, L_s decreases with $l(s)$. Similarly, because $\partial L_s/\partial u(s)$ is everywhere positive, L_s increases with $u(s)$. Thus, extrema occur wherever $l(s)$ is increased to a boundary and $u(s)$ is decreased to a boundary. It follows that all $l(s) = u(s)$ are extrema. For all $l(s) = u(s)$, L_s is zero. Thus, all $l(s) = u(s)$ minimize L_s. In this case, we arbitrarily set $l(s) = u(s) = \bar{x}(s)$ to the mode.

Case ②: $f_{X(s)}(\tilde{x}(s)) \leq 1/\lambda_\alpha$, $f_{X(s)}(\hat{x}(s)) \leq 1/\lambda_\alpha$, *and* $1/\lambda_\alpha \leq f_{X(s)}(\bar{x}(s))$. In this case, $1/\lambda_\alpha$ lies at or between the density values of the antimodes and the density value of the mode. When $l(s) < \underline{f}_{X(s)}^{-1}(1/\lambda_\alpha)$, because $\underline{f}_{X(s)}$ strictly increases, $f_{X(s)}(l(s)) < 1/\lambda_\alpha$. It follows that

$$\frac{\partial L_s}{\partial l(s)} = -1 + \lambda_\alpha f_{X(s)}(l(s)) < -1 + \lambda_\alpha \frac{1}{\lambda_\alpha} = 0. \tag{A10}$$

Similarly, when $u(s) < \underline{f}_{X(s)}^{-1}(1/\lambda_\alpha)$,

$$\frac{\partial L_s}{\partial u(s)} = 1 - \lambda_\alpha f_{X(s)}(u(s)) > 1 - \lambda_\alpha \frac{1}{\lambda_\alpha} = 0. \tag{A11}$$

When $l(s) = \underline{f}_{X(s)}^{-1}(1/\lambda_\alpha)$, or when $l(s) = \overline{f}_{X(s)}^{-1}(1/\lambda_\alpha)$, because $f_{X(s)}(l(s)) = 1/\lambda_\alpha$, it follows that

$$\frac{\partial L_s}{\partial l(s)} = -1 + \lambda_\alpha f_{X(s)}(l(s)) = -1 + \lambda_\alpha \frac{1}{\lambda_\alpha} = 0. \tag{A12}$$

Similarly, when $u(s) = \underline{f}_{X(s)}^{-1}(1/\lambda_\alpha)$, or when $u(s) = \overline{f}_{X(s)}^{-1}(1/\lambda_\alpha)$,

$$\frac{\partial L_s}{\partial u(s)} = 1 - \lambda_\alpha f_{X(s)}(u(s)) = 1 - \lambda_\alpha \frac{1}{\lambda_\alpha} = 0. \tag{A13}$$

When $\underline{f}_{X(s)}^{-1}(1/\lambda_\alpha) < l(s) < \overline{f}_{X(s)}^{-1}(1/\lambda_\alpha)$, because $f_{X(s)}$ strictly increases left of the mode and strictly decreases right of the mode, $1/\lambda_\alpha < f_{X(s)}(l(s))$. It follows that

$$\frac{\partial L_s}{\partial l(s)} = -1 + \lambda_\alpha f_{X(s)}(l(s)) > -1 + \lambda_\alpha \frac{1}{\lambda_\alpha} = 0. \tag{A14}$$

Similarly, when $\underline{f}_{X(s)}^{-1}(1/\lambda_\alpha) < u(s) < \overline{f}_{X(s)}^{-1}(1/\lambda_\alpha)$,

$$\frac{\partial L_s}{\partial u(s)} = 1 - \lambda_\alpha f_{X(s)}(u(s)) < 1 - \lambda_\alpha \frac{1}{\lambda_\alpha} = 0. \tag{A15}$$

When $\overline{f}_{X(s)}^{-1}(1/\lambda_\alpha) < l(s)$, because $\overline{f}_{X(s)}$ strictly decreases, $f_{X(s)}(l(s)) < 1/\lambda_\alpha$. It follows that

$$\frac{\partial L_s}{\partial l(s)} = -1 + \lambda_\alpha f_{X(s)}(l(s)) < -1 + \lambda_\alpha \frac{1}{\lambda_\alpha} = 0. \tag{A16}$$

Similarly, when $\overline{f}_{X(s)}^{-1}(1/\lambda_\alpha) < u(s)$,

$$\frac{\partial L_s}{\partial u(s)} = 1 - \lambda_\alpha f_{X(s)}(u(s)) > 1 - \lambda_\alpha \frac{1}{\lambda_\alpha} = 0. \tag{A17}$$

Thus, L_s has extrema in both $l(s)$ and $u(s)$ at $\underline{f}_{X(s)}^{-1}(1/\lambda_\alpha)$ and $\overline{f}_{X(s)}^{-1}(1/\lambda_\alpha)$. Because a change in the sign of the partial derivative from negative to positive indicates a minimum, we set $l(s) = \underline{f}_{X(s)}^{-1}(1/\lambda_\alpha)$ and $u(s) = \overline{f}_{X(s)}^{-1}(1/\lambda_\alpha)$.

Case ③: $1/\lambda_\alpha < f_{X(s)}(\tilde{x}(s))$ *and* $f_{X(s)}(\hat{x}(s)) \leq 1/\lambda_\alpha$. In this case, $1/\lambda_\alpha$ sits below the density value of the left antimode and at or above the density value of the right antimode. This case only applies if $\mathcal{X}(s)$ is left-bounded. When $l(s) < \tilde{x}(s)$, because $f_{X(s)}(l(s)) = 0$, it follows that

$$\frac{\partial L_s}{\partial l(s)} = -1 + \lambda_\alpha f_{X(s)}(l(s)) = -1. \tag{A18}$$

Similarly, when $u(s) < \tilde{x}(s)$,

$$\frac{\partial L_s}{\partial u(s)} = 1 - \lambda_\alpha f_{X(s)}(u(s)) = 1. \tag{A19}$$

When $\tilde{x}(s) \le l(s) < \overline{f}_{X(s)}^{-1}(1/\lambda_\alpha)$, because $f_{X(s)}$ strictly increases left of the mode and strictly decreases right of the mode, $1/\lambda_\alpha < f_{X(s)}(l(s))$. It follows that

$$\frac{\partial L_s}{\partial l(s)} = -1 + \lambda_\alpha f_{X(s)}(l(s)) > -1 + \lambda_\alpha \frac{1}{\lambda_\alpha} = 0. \tag{A20}$$

Similarly, when $\tilde{x}(s) \le u(s) < \overline{f}_{X(s)}^{-1}(1/\lambda_\alpha)$,

$$\frac{\partial L_s}{\partial u(s)} = 1 - \lambda_\alpha f_{X(s)}(u(s)) < 1 - \lambda_\alpha \frac{1}{\lambda_\alpha} = 0. \tag{A21}$$

When $l(s) = \overline{f}_{X(s)}^{-1}(1/\lambda_\alpha)$, because $f_{X(s)}(l(s)) = 1/\lambda_\alpha$, it follows that

$$\frac{\partial L_s}{\partial l(s)} = -1 + \lambda_\alpha f_{X(s)}(l(s)) = -1 + \lambda_\alpha \frac{1}{\lambda_\alpha} = 0. \tag{A22}$$

Similarly, when $u(s) = \overline{f}_{X(s)}^{-1}(1/\lambda_\alpha)$,

$$\frac{\partial L_s}{\partial u(s)} = 1 - \lambda_\alpha f_{X(s)}(u(s)) = 1 - \lambda_\alpha \frac{1}{\lambda_\alpha} = 0. \tag{A23}$$

When $\overline{f}_{X(s)}^{-1}(1/\lambda_\alpha) < l(s)$, because $\overline{f}_{X(s)}$ strictly decreases, $f_{X(s)}(l(s)) < 1/\lambda_\alpha$. It follows that

$$\frac{\partial L_s}{\partial l(s)} = -1 + \lambda_\alpha f_{X(s)}(l(s)) < -1 + \lambda_\alpha \frac{1}{\lambda_\alpha} = 0. \tag{A24}$$

Similarly, when $\overline{f}_{X(s)}^{-1}(1/\lambda_\alpha) < u(s)$,

$$\frac{\partial L_s}{\partial u(s)} = 1 - \lambda_\alpha f_{X(s)}(u(s)) > 1 - \lambda_\alpha \frac{1}{\lambda_\alpha} = 0. \tag{A25}$$

Thus, L_s has extrema in both $l(s)$ and $u(s)$ at $\tilde{x}(s)$ and $\overline{f}_{X(s)}^{-1}(1/\lambda_\alpha)$. Because a change in the sign of the partial derivative from negative to positive indicates a minimum, we set $l(s) = \tilde{x}(s)$ and $u(s) = \overline{f}_{X(s)}^{-1}(1/\lambda_\alpha)$.

Case ④ : $f_{X(s)}(\tilde{x}(s)) \le 1/\lambda_\alpha$ *and* $1/\lambda_\alpha < f_{X(s)}(\hat{x}(s))$. In this case, $1/\lambda_\alpha$ sits below the density value of the right antimode and at or above the density value of the left antimode. This case only applies if $\mathcal{X}(s)$ is right-bounded. When $l(s) < \underline{f}_{X(s)}^{-1}(1/\lambda_\alpha)$, because $\underline{f}_{X(s)}$ strictly increases, $f_{X(s)}(l(s)) < 1/\lambda_\alpha$. It follows that

$$\frac{\partial L_s}{\partial l(s)} = -1 + \lambda_\alpha f_{X(s)}(l(s)) < -1 + \lambda_\alpha \frac{1}{\lambda_\alpha} = 0. \tag{A26}$$

Similarly, when $u(s) < \underline{f}_{X(s)}^{-1}(1/\lambda_\alpha)$,

$$\frac{\partial L_s}{\partial u(s)} = 1 - \lambda_\alpha f_{X(s)}(u(s)) > 1 - \lambda_\alpha \frac{1}{\lambda_\alpha} = 0. \tag{A27}$$

When $l(s) = \underline{f}_{X(s)}^{-1}(^1/\lambda_\alpha)$, because $f_{X(s)}(l(s)) = {}^1/\lambda_\alpha$, it follows that

$$\frac{\partial L_s}{\partial l(s)} = -1 + \lambda_\alpha f_{X(s)}(l(s)) = -1 + \lambda_\alpha \frac{1}{\lambda_\alpha} = 0. \tag{A28}$$

Similarly, when $u(s) = \underline{f}_{X(s)}^{-1}(^1/\lambda_\alpha)$,

$$\frac{\partial L_s}{\partial u(s)} = 1 - \lambda_\alpha f_{X(s)}(u(s)) = 1 - \lambda_\alpha \frac{1}{\lambda_\alpha} = 0. \tag{A29}$$

When $\underline{f}_{X(s)}^{-1}(^1/\lambda_\alpha) \le l(s) < \hat{x}(s)$, because $f_{X(s)}$ strictly increases left of the mode and strictly decreases right of the mode, $^1/\lambda_\alpha < f_{X(s)}(l(s))$. It follows that

$$\frac{\partial L_s}{\partial l(s)} = -1 + \lambda_\alpha f_{X(s)}(l(s)) > -1 + \lambda_\alpha \frac{1}{\lambda_\alpha} = 0. \tag{A30}$$

Similarly, when $\underline{f}_{X(s)}^{-1}(^1/\lambda_\alpha) \le u(s) < \hat{x}(s)$,

$$\frac{\partial L_s}{\partial u(s)} = 1 - \lambda_\alpha f_{X(s)}(u(s)) < 1 - \lambda_\alpha \frac{1}{\lambda_\alpha} = 0. \tag{A31}$$

When $\hat{x}(s) < l(s)$, because $f_{X(s)}(l(s)) = 0$, it follows that

$$\frac{\partial L_s}{\partial l(s)} = -1 + \lambda_\alpha f_{X(s)}(l(s)) = -1. \tag{A32}$$

Similarly, when $\hat{x}(s) < u(s)$,

$$\frac{\partial L_s}{\partial u(s)} = 1 - \lambda_\alpha f_{X(s)}(u(s)) = 1. \tag{A33}$$

Thus, L_s has extrema in both $l(s)$ and $u(s)$ at $\underline{f}_{X(s)}^{-1}(^1/\lambda_\alpha)$ and $\hat{x}(s)$. Because a change in the sign of the partial derivative from negative to positive indicates a minimum, we set $l(s) = \underline{f}_{X(s)}^{-1}(^1/\lambda_\alpha)$ and $u(s) = \hat{x}(s)$.

Case $\textcircled{5}$: $^1/\lambda_\alpha < f_{X(s)}(\tilde{x}(s))$ *and* $^1/\lambda_\alpha < f_{X(s)}(\hat{x}(s))$. In this case, $^1/\lambda_\alpha$ sits below the density values of the left antimode and the density value of the right antimode. This case only applies if $\mathcal{X}(s)$ is both left- and right-bounded. When $l(s) < \tilde{x}(s)$, because $f_{X(s)}(l(s)) = 0$, it follows that

$$\frac{\partial L_s}{\partial l(s)} = -1 + \lambda_\alpha f_{X(s)}(l(s)) = -1. \tag{A34}$$

Similarly, when $u(s) < \tilde{x}(s)$,

$$\frac{\partial L_s}{\partial u(s)} = 1 - \lambda_\alpha f_{X(s)}(u(s)) = 1. \tag{A35}$$

When $\tilde{x}(s) \le l(s) \le \hat{x}(s)$, because $^1/\lambda_\alpha < f_{X(s)}(l(s))$, it follows that

$$\frac{\partial L_s}{\partial l(s)} = -1 + \lambda_\alpha f_{X(s)}(l(s)) > -1 + \lambda_\alpha \frac{1}{\lambda_\alpha} = 0. \tag{A36}$$

Similarly, when $\tilde{x}(s) \leq u(s) \leq \hat{x}(s)$,

$$\frac{\partial L_s}{\partial u(s)} = 1 - \lambda_\alpha f_{X(s)}(l(s)) < 1 - \lambda_\alpha \frac{1}{\lambda_\alpha} = 0. \tag{A37}$$

When $\hat{x}(s) < l(s)$, because $f_{X(s)}(l(s)) = 0$, it follows that

$$\frac{\partial L_s}{\partial l(s)} = -1 + \lambda_\alpha f_{X(s)}(l(s)) = -1. \tag{A38}$$

Similarly, when $\hat{x}(s) < u(s)$,

$$\frac{\partial L_s}{\partial u(s)} = 1 - \lambda_\alpha f_{X(s)}(u(s)) = 1. \tag{A39}$$

Thus, L_s has extrema in both $l(s)$ and $u(s)$ at $\tilde{x}(s)$ and $\hat{x}(s)$. Because a change in the sign of the partial derivative from negative to positive indicates a minimum, we set $l(s) = \tilde{x}(s)$ and $u(s) = \hat{x}(s)$.

The time windows that follow from Cases $\textcircled{1}$–$\textcircled{5}$ can be summarized as follows:

$$l(s) = \begin{cases} \tilde{x}(s), & \frac{1}{\lambda_\alpha} < f_{X(s)}(\tilde{x}(s)), \\[2mm] \underline{f}_{X(s)}^{-1}(\frac{1}{\lambda_\alpha}), & f_{X(s)}(\tilde{x}(s)) \leq \frac{1}{\lambda_\alpha} \leq f_{X(s)}(\bar{x}(s)), \\[2mm] \bar{x}(s), & f_{X(s)}(\bar{x}(s)) < \frac{1}{\lambda_\alpha}, \end{cases} \tag{A40}$$

$$u(s) = \begin{cases} \hat{x}(s), & \frac{1}{\lambda_\alpha} < f_{X(s)}(\hat{x}(s)), \\[2mm] \overline{f}_{X(s)}^{-1}(\frac{1}{\lambda_\alpha}), & f_{X(s)}(\hat{x}(s)) \leq \frac{1}{\lambda_\alpha} \leq f_{X(s)}(\bar{x}(s)), \\[2mm] \bar{x}(s), & f_{X(s)}(\bar{x}(s)) < \frac{1}{\lambda_\alpha}. \end{cases} \tag{A41}$$

The first cases in each equation apply only to states with left- and right-bounded supports, respectively. The second and third cases apply to all states. In the second cases, if the support is left- or right-unbounded, respectively, then because the corresponding antimode does not exist, any value of $\frac{1}{\lambda_\alpha}$ less than or equal to the upper limit satisfies the criteria.

With two adjustments, Equations (A5) and (A6) follow directly from Equations (A40) and (A41). Consolidate the first two cases of Equation (A40) to set $l(s) = \min \mathcal{X}(s, \frac{1}{\lambda_\alpha})$ when $\frac{1}{\lambda_\alpha} \leq f_{X(s)}(\bar{x}(s))$. If the left antimode exists, then if $\frac{1}{\lambda_\alpha} < f_{X(s)}(\tilde{x}(s))$, $\min \mathcal{X}(s, \frac{1}{\lambda_\alpha}) = \tilde{x}(s)$. Otherwise, and when the left antimode does not exist, $\min \mathcal{X}(s, \frac{1}{\lambda_\alpha}) = \underline{f}_{X(s)}^{-1}(\frac{1}{\lambda_\alpha})$ for all $\frac{1}{\lambda_\alpha}$ at or below the density value of the mode. Second, consolidate the first two cases of Equation (A41) to set $u(s, \frac{1}{\lambda_\alpha}) = \max \mathcal{X}(s, \frac{1}{\lambda_\alpha})$ when $\frac{1}{\lambda_\alpha} \leq f_{X(s)}(\bar{x}(s))$. If the right antimode exists, then if $\frac{1}{\lambda_\alpha} < f_{X(s)}(\hat{x}(s))$, $\max \mathcal{X}(s, \frac{1}{\lambda_\alpha}) = \hat{x}(s)$. Otherwise, and when the right antimode does not exist, $\max \mathcal{X}(s, \frac{1}{\lambda_\alpha}) = \overline{f}_{X(s)}^{-1}(\frac{1}{\lambda_\alpha})$ for all $\frac{1}{\lambda_\alpha}$ at or below the density value of the mode. Thus, $L(l(\frac{1}{\lambda_\alpha}), u(\frac{1}{\lambda_\alpha}), \lambda_\alpha) = \min\{L(l, u, \lambda_\alpha) : l \leq u\}$, which completes the proof for the third saddle point condition.

Finally, to prove the fourth saddle point condition, we show that the duality gap at the candidate solutions is zero:

$$L(\boldsymbol{l}(^1/_{\lambda_\alpha}), \boldsymbol{u}(^1/_{\lambda_\alpha}), \lambda) = \sum_{s \in \mathcal{S}} \mathbb{P}\{S = s\} \left(u(s, ^1/_{\lambda_\alpha}) - l(s, ^1/_{\lambda_\alpha})\right)$$

$$+ \lambda \left[1 - \alpha - \left(\sum_{s \in \mathcal{S}} \mathbb{P}\{S = s\} \left[F_{X(s)}(u(s, ^1/_{\lambda_\alpha})) - F_{X(s)}(l(s, ^1/_{\lambda_\alpha}))\right]\right)\right] \quad \text{(A42)}$$

$$= \sum_{s \in \mathcal{S}} \mathbb{P}\{S = s\} \left(u(s, ^1/_{\lambda_\alpha}) - l(s, ^1/_{\lambda_\alpha})\right) \quad \text{(A43)}$$

$$= \mathbb{E}_{\mathcal{S}}\left[u(S, ^1/_{\lambda_\alpha}) - l(S, ^1/_{\lambda_\alpha})\right]. \quad \text{(A44)}$$

The first equation is the Lagrangian evaluated at the candidate solutions. The second equation follows from the second saddle point condition. The third equation is the objective value of Problem (1) at the candidate solution. This completes the proof establishing the candidate solutions as a saddle point. It follows that the candidate solutions are optimal solutions to the Lagrangian dual problem and to Problem (1).

Proof Conclusion. The candidate solutions reduce Problem (1) to a root finding problem in one dimension. Specifically, identify λ_α such that $\mathbb{P}\{l(S, ^1/_{\lambda_\alpha}) \leq X(S) \leq u(S, ^1/_{\lambda_\alpha})\} = 1 - \alpha$. By Lemma 1, because $\mathbb{P}\{l(S, ^1/_{\lambda_\alpha}) \leq X(S) \leq u(S, ^1/_{\lambda_\alpha})\}$ strictly decreases with $^1/_{\lambda_\alpha}$, any line search method will quickly identify λ_α. In the body of the paper, the theorem is presented in terms of $y > 0$ instead of $^1/_{\lambda_\alpha} > 0$. The purpose of the substitution is to ease the presentation. The formulation presented in the proof is recovered by setting $y = ^1/_{\lambda_\alpha}$.