Online Learning for Dynamic Service Mode Control*

Wenqian Xing
Stanford University

Yue Hu
Stanford University

Anand Kalvit
Amazon†

Vahid Sarhangian
University of Toronto

Abstract

Modern service systems are increasingly adopting new modalities enabled by emerging technologies, such as AI-assisted services, to better balance quality and efficiency. Motivated by this trend, we study dynamic service mode control in a single-server queue with two switchable modes, each with a distinct service rate and an unknown reward distribution. The objective is to maximize the long-run average of expected cumulative rewards minus holding costs achievable under non-anticipating, state-dependent policies. To address the problem, we first establish the optimality of a threshold policy under full information of the problem primitives. When reward distributions are unknown but samples are observable, we propose an online learning algorithm that uses Upper Confidence Bound (UCB) estimates of the unknown parameters to adaptively learn the optimal threshold. Our algorithm achieves statistically near-optimal regret of $\tilde{O}(\sqrt{T})$ and demonstrates strong numerical performance. Additionally, when additional partial information about the optimal policy is available ex ante (specifically, a non-trivial lower bound on the optimal threshold), we show that an episodic greedy policy achieves constant regret by leveraging a free-exploration property intrinsic to this special setting. Methodologically, we develop a novel regret decomposition and regenerative cycle-based analysis, offering general tools for learning-based queueing control. Lastly, we conduct a healthcare case study on AI-assisted patient messaging demonstrating the practical utility of our approach.

1 Introduction

Service rate control is a fundamental problem in service, manufacturing, and telecommunication systems, where providers adjust service speeds or modes to balance rewards, costs, and congestion. Classical models emphasize cost minimization, with faster service typically reducing congestion as well as operational costs. In contrast, we consider customer-intensive services, e.g., call centers, restaurants, and retail banks, with speed-quality trade-offs where longer service can improve customer satisfaction [23, 27, 14, 5, 3, 16]. More recently, advances in AI-assisted workflows have introduced newer complexities into this landscape: the potential impact of service modes on performance is often unknown and must be learned through experimentation. Examples include hospitals testing generative AI for patient messages and radiology studies on AI-assisted diagnostics, which show higher quality but increased workload [9, 24, 22, 29, 2]. In a similar vein, effective social media content moderation necessitates carefully calibrating AI-based approaches—offering speed and scalability but uncertain efficacy—against slower yet more reliable human review [7]. More broadly, AI-augmented systems exhibit uncertain and noisy payoff structures—such as satisfaction, ratings, and reliability metrics—that complicate the management of speed-quality trade-offs. These challenges call for online learning approaches that dynamically balance exploration of new service modalities to gain valuable information, with exploitation of existing knowledge to optimize performance.

*Full length article is available at: `https://dx.doi.org/10.2139/ssrn.5123355`.

†This work was done while the author was affiliated with Stanford University.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

Motivated by this, we study a service mode control problem in a Markovian single-server queue where the server can dynamically choose between two service modes: a *base* service mode with *faster* service rate and a *premium* service mode with *slower* service rate but potentially higher rewards. Rewards under each mode follow unknown distributions that must be learned through experimentation, and the objective is to maximize the long-run average expected reward (net of holding costs). To address this problem, we first characterize the optimal policy under full information and show that it takes the form of a queue-length-based threshold rule. We then consider the online learning setting where reward distributions are initially unknown but samples are observable, and propose an episodic UCB-based algorithm that learns the optimal threshold policy, adapting to different stability regimes and achieving rate-optimal $\tilde{O}(\sqrt{T})$ regret. Our analysis relies on a novel regret decomposition using regenerative cycles into transient regret, learning regret, residual error, and approximation error, which aggregate to produce the $\tilde{O}(\sqrt{T})$ bound. We further establish that an episodic greedy algorithm achieves $O(1)$ regret under a *free-exploration* property that holds when some side information about the optimal threshold (specifically, a non-trivial lower bound) is available ex ante. Finally, via a healthcare case study calibrated with real-world data, we show that our approach reduces regret by 12%–25% relative to baselines, highlighting its practical value.

Related work. First, our work relates to the literature on dynamic service rate control, which largely focuses on optimizing service speeds under known system parameters, often in conjunction with arrival rate control [5] or admission control [1]. In this context, threshold-type policies are generally known to be optimal in $M/M/1$ queues with finitely many service rates [14], with extensions to general arrival processes [19, 28], and Markov-modulated primitives [4]. Our modified M/M/1 setting, on the other hand, is differentiated by unknown system parameters that must be learned through interaction with the environment. Second, there is growing literature integrating statistical learning with queueing control, e.g., scheduling, admission, and pricing under unknown system parameters [26]. For scheduling, results include constant regret for empirical $c\mu$ rules in stable queues [17], optimal learn-then-schedule methods in overloaded queues [32], and UCB-based algorithms for multi-server networks [18, 30, 13]. For routing and admission control, there are studies on static policy learning informed by linear program relaxations [31, 25], with recent work on threshold-based admission policies [11]. For pricing and capacity sizing, [10, 15] learn optimal fixed prices and/or service rates. Our work contributes to this body of literature by incorporating learning from service mode-dependent stochastic rewards to optimize a weighted sum of competing objectives (rewards vs. congestion). Third, in the reinforcement learning literature, prior work establishes regret guarantees for finite state and action spaces [6], with extensions to countable spaces [21, 20]. Our algorithm achieves rate-optimal regret in a countably infinite state space by leveraging structural properties of the problem, complementing other model-based approaches [12].

2 Model and Problem Formulation

We consider a Markovian single-server queue with Poisson arrivals at rate λ. Customers are served on a first-in-first-out basis and incur a holding cost $c \geqslant 0$ per unit time spent in the system. (Throughout, when $c = 0$, we restrict attention to the regime $\lambda < \mu^p \leqslant \mu^b$, which maintains the optimality of threshold policies under full information; see §3.)

Service modalities. At any time, the server can operate in a *base* mode with exponential service times of rate μ^b, or a *premium* mode with exponential service times of rate $\mu^p \leqslant \mu^b$; the two modes can be switched preemptively. Upon service completion, the system receives a stochastic reward depending on the chosen mode, with mean r^b (base) or r^p (premium). Rewards are σ-sub-Gaussian with mean upper bounded by $R > 0$, and independent across jobs and modes. We are interested in the setting where r^b and r^p are unknown initially and must be estimated online.

Reward-congestion trade-off. Although we do not impose any restriction on the ordering between r^b and r^p, it is natural in many applications to expect $r^b \leqslant r^p$. That is, the base mode reduces congestion but yields lower rewards, while the premium mode offers higher rewards at the cost of slower service. To ensure stabilizability of the system, we assume $\mu^b > \lambda$.

Policies. The horizon of the control problem is $T > 0$. For $t \in [0, T]$, we use $\mathcal{H}_t$ to denote the filtration representing the complete system history up to time t. A policy is a non-anticipatory,

deterministic[‡] mapping $\pi : \mathcal{H}_t \to \{b, p\}$ that at any time selects a mode (base b or premium p) based on current history. We denote the set of all admissible policies by Π.

Objective. We use $R_\pi(t)$ to denote the expected cumulative reward up to time t, and $Q_\pi(t)$ to denote the number in system at time t (which we interchangeably refer to as the queue length in this single-server system) under policy π. Our goal is to design a policy that maximizes the long-run average expected reward (net of holding cost), i.e.,

$$\sup_{\pi\in\Pi} \liminf_{T\to\infty} \frac{1}{T}\left(R_\pi(T) - \mathbb{E}\left[\int_0^T cQ_\pi(t)\,dt\right]\right). \tag{1}$$

We benchmark the performance of π against the optimal value of (1) when the reward parameters are known ex ante, denoted by $\texttt{OPT}$. The regret of policy π up to time t is then defined as

$$\texttt{Regret}_\pi(t) = t\texttt{OPT} - \left(R_\pi(t) - \mathbb{E}\left[\int_0^t cQ_\pi(s)\,ds\right]\right). \tag{2}$$

3 Full Information Setting

We begin by studying the full-information setting where all model parameters are known a priori. We refer to a policy as a threshold policy with threshold $Q \in \mathbb{N}$, denoted by π_Q, if the server deploys the base service when $Q(t) \geqslant Q$ and the premium service otherwise. Building on Theorem 4 in [23], we show that a threshold policy is optimal for (1) when the reward parameters are known. We further establish that for a fixed $Q \in \mathbb{N}$, the long-run average value of π_Q satisfies

$$\frac{1}{T}\left(R_{\pi_Q}(T) - \mathbb{E}\left[\int_0^T c\,\mathbb{Q}_{\pi_Q}(t)\,dt\right]\right) \to \texttt{VAL}(Q, \lambda, \mu^p, \mu^b, c, r^p, r^b) \quad \text{as } T \to \infty,$$

where $\texttt{VAL}(Q, \lambda, \mu^p, \mu^b, c, r^p, r^b) := \mu^p r^p \mathbb{P}_{\mathbb{Q}\sim\mathcal{F}_Q}(1 \leqslant \mathbb{Q} \leqslant Q-1) + \mu^b r^b \mathbb{P}_{\mathbb{Q}\sim\mathcal{F}_Q}(\mathbb{Q} \geqslant Q) - c\mathbb{E}_{\mathbb{Q}\sim\mathcal{F}_Q}[\mathbb{Q}]$. Here, $\mathbb{P}_{\mathbb{Q}\sim\mathcal{F}_Q}(\cdot)$ and $\mathbb{E}_{\mathbb{Q}\sim\mathcal{F}_Q}(\cdot)$ denote the probability and expectation, respectively, under the stationary distribution $\mathcal{F}_Q$ induced by π_Q. The optimal value of (1) under full information is therefore $\texttt{OPT} := \sup_{Q\in\mathbb{N}} \texttt{VAL}(Q, \lambda, \mu^p, \mu^b, c, r^p, r^b)$, achieved at some Q^* (potentially infinite).

4 Online Learning Setting: Towards a Rate-Optimal Adaptive Policy

We now pivot to the case of unknown mean rewards r^p and r^b (and hence Q^*) that must be learned through online experimentation. We begin by establishing a lower bound on achievable regret in (2).

Theorem 1 (Lower bound on achievable regret) *For every policy $\pi \in \Pi$, there exists a problem instance ν_π such that $\texttt{Regret}_\pi(T; \nu_\pi) = O(\sqrt{\lambda T})$ for sufficiently large T.*

We next introduce Algorithm 1 for minimizing regret in (2). The algorithm searches for the optimal threshold (within the class of threshold policies) over a candidate range from 1 up to a pre-specified upper bound $\bar{Q}$. It operates in episodes defined by busy cycles of the queue length process and updates the threshold estimates using the data collected in each cycle. To guide the specification of $\bar{Q}$, we distinguish between degenerate and non-degenerate cases based on system parameters. In the degenerate case, either the holding cost is zero ($c = 0$) or the two service modes have identical service rates ($\mu^p = \mu^b$); in both scenarios, the trade-off between rewards and congestion disappears. In the non-degenerate case, where $c > 0$ and $\mu^p \neq \mu^b$, the system falls into one of three regimes depending on which service modes can stabilize the system: (i) *Bi-stable*: $\lambda < \mu^p < \mu^b$; (ii) *Critically bi-stable*: $\lambda = \mu^p < \mu^b$; (iii) *Uni-stable*: $\mu^p < \lambda < \mu^b$.

In the degenerate case, the problem, in fact, reduces to a simple two-armed bandit (albeit in continuous-time) with unknown means r^p and r^b, and we omit its discussion for brevity. In the non-degenerate case, we show that $Q^* \leqslant \bar{Q} < \infty$, where $\bar{Q}$ is independent of T and in the bi-stable regime given by

$$\bar{Q} = \left\lceil \frac{\mu^b(\mu^p - \lambda)R}{c(\mu^b - \mu^p)} + \frac{\lambda(\mu^b - \mu^p)}{(\mu^b - \lambda)(\mu^p - \lambda)} \right\rceil, \tag{3}$$

whereas in the critically bi-stable and uni-stable regimes, by

$$\bar{Q} = \left\lceil \frac{-(1-A) + \sqrt{(1-A)^2 + 4Ay}}{2} \right\rceil + 1,\ y = \frac{\mu^b}{\mu^b - \lambda},\ A = \frac{2R(\mu^p + \mu^b)}{c} + \frac{2\lambda}{\mu^b - \lambda}. \tag{4}$$

Algorithm 1 [UCB] An episodic UCB algorithm for threshold-triggered service mode control

- **Input:** Arrival/service rates λ, μ^p, μ^b, Holding cost c, Horizon T, Exploration coefficient α, sub-Gaussian parameter σ, Mean reward upper bound R.
- **Compute:** Upper bound $\bar{Q} \in \mathbb{N}$ on the optimal threshold Q^* as per (3) or (4).
- **Initialize:** $\mathbb{Q}(0) = 0$, Threshold $\hat{Q}_1 = 1$.
- **For $k = 1, 2, \cdots$**

  Deploy $\pi_{\hat{Q}_k}$ until the end of episode k. (`episode ends when queue empties again`)

  At the end of episode k, do the following:

  $\hat{r}_k^p \leftarrow$ Empirical mean reward from all premium service completions by the end of episode k.

  $\hat{r}_k^b \leftarrow$ Empirical mean reward from all base service completions by the end of episode k.

  $\hat{Q}_{k+1} \leftarrow \min\Big(\arg\max_{Q\in[\bar{Q}]} \texttt{VAL}\Big(Q, \lambda, \mu^p, \mu^b, c, \hat{r}_k^p + \sqrt{\frac{\alpha\sigma^2 \log T}{2+N_k^p}}, \hat{r}_k^b + \sqrt{\frac{\alpha\sigma^2 \log T}{2+N_k^b}}\Big)\Big)$.

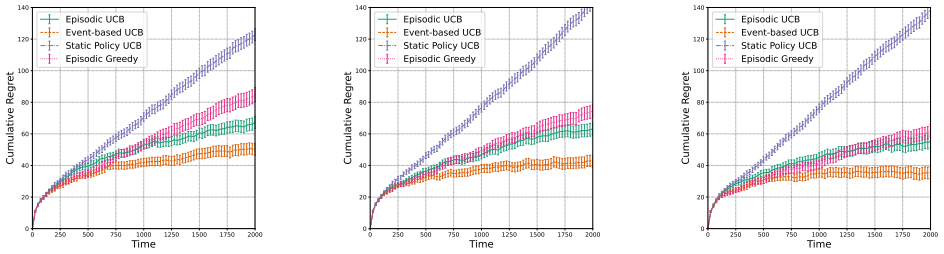
We now present our main regret bound for Algorithm 1. Combined with the lower bound in Theorem 1, Theorem 2 shows that Algorithm 1 achieves the statistically optimal $\sqrt{T}$ rate (up to polylog factors).

Theorem 2 (Upper bound on the regret of Algorithm 1) *The regret of Algorithm 1 initialized with $\alpha \geqslant 30$, remains bounded as $\texttt{Regret}_{\texttt{UCB}}(t) = \tilde{O}(\sqrt{t})$ at all times $t \leqslant T$, where $\tilde{O}(\cdot)$ hides poly-logarithmic factors in T as well as dependencies on $(\lambda, \mu^p, \mu^b, \sigma, R, c)$.*

Free exploration under the greedy policy. We show that if some side information about the optimal threshold Q^* is available ex ante, greedy policies can leverage the free-exploration property of the underlying queueing system to achieve $O(1)$ regret. Specifically, in the non-degenerate case, if $Q^* > 1$ is known, an episodic greedy policy, obtained by removing the UCB factors in Algorithm 1 and restricting $\hat{Q}_k \geqslant 2$ for all k, achieves constant regret. The key observation is that with this modification, there is a positive probability of observing both base and premium service completions in each episode, which eliminates the need for explicit UCB-based exploration.

Theorem 3 (Constant regret under the greedy policy) *In the non-degenerate case, if $Q^* > 1$ is known a priori, then the episodic greedy algorithm achieves $O(1)$, regret.*

Case study on AI-assisted patient message replies. We conduct a numerical case study inspired by recent pilots on AI-assisted patient message replies in healthcare [24, 22, 8, 9, 29]. In this setting, the base service corresponds to providers drafting replies to patient messages independently, while the premium service corresponds to providers collaborating with GenAI chatbots to generate and refine drafts. Rewards and service rates are calibrated using existing studies [22, 24], which find statistically significant evidence of higher ratings for AI-assisted replies ($r^b = 3.38$, $r^p = 3.70$) but slower response speeds ($\mu^b = 0.769$, $\mu^p = 0.659$ per minute). We set the holding cost rate at $c = 0.01$, and vary the arrival rate to mimic different parameter regimes. In Figure 1, we compare the performance of the episodic UCB algorithm and the episodic greedy algorithm against two baseline policies: 1) event-based UCB, which updates estimates upon each service completion, and 2) static policy UCB, which learns the optimal static policy (among "always premium" and "always base"). Our results show that the proposed policies reduce regret by approximately 12%–25% at time $T = 600$ (comparable to the message volume in the pilot study of [22]) relative to static policy UCB.



(a) Bi-stable (b) Critically bi-stable (c) Uni-stable

Figure 1: Case study results under different regimes

[‡]Remarkably, a threshold policy may no longer be optimal if the feasible set is expanded to include randomized policies (and it is unclear what an optimal policy may look like); we leave this case to future work.

References

[1] Kranthi Mitra Adusumilli and John J Hasenbein. Dynamic admission and service rate control of a queue. *Queueing Systems*, 66(2):131–154, 2010.

[2] Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. Combining human expertise with artificial intelligence: Experimental evidence from radiology. Technical report, National Bureau of Economic Research, 2023.

[3] Krishnan S Anand, M Fazıl Paç, and Senthil Veeraraghavan. Quality–speed conundrum: Trade-offs in customer-intensive services. *Management Science*, 57(1):40–56, 2011.

[4] Ari Arapostathis, Anirban Das, Guodong Pang, and Yi Zheng. Optimal control of markov-modulated multiclass many-server queues. *Stochastic Systems*, 9(2):155–181, 2019.

[5] Bariş Ata and Shiri Shneorson. Dynamic control of an m/m/1 service system with adjustable arrival and service rates. *Management Science*, 52(11):1778–1791, 2006.

[6] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in Neural Information Processing Systems*, 21, 2008.

[7] Vashist Avadhanula, Omar Abdul Baki, Hamsa Bastani, Osbert Bastani, Caner Gocmen, Daniel Haimovich, Darren Hwang, Dima Karamshuk, Thomas Leeper, Jiayuan Ma, et al. Bandits for online calibration: An application to content moderation on social media platforms. *arXiv preprint arXiv:2211.06516*, 2022.

[8] John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, 183(6):589–596, 2023.

[9] Shan Chen, Marco Guevara, Shalini Moningi, Frank Hoebers, Hesham Elhalawani, Benjamin H Kann, Fallon E Chipidza, Jonathan Leeman, Hugo JWL Aerts, Timothy Miller, et al. The effect of using a large language model to respond to patient messages. *The Lancet Digital Health*, 6(6):e379–e381, 2024.

[10] Xinyun Chen, Yunan Liu, and Guiyu Hong. An online learning approach to dynamic pricing and capacity sizing in service systems. *Operations Research*, 72(6):2677–2697, 2024.

[11] Asaf Cohen, Vijay Subramanian, and Yili Zhang. Learning-based optimal admission control in a single-server queuing system. *Stochastic systems*, 14(1):69–107, 2024.

[12] Jim G Dai and Mark Gluzman. Queueing network controls via deep reinforcement learning. *Stochastic Systems*, 12(1):30–67, 2022.

[13] Daniel Freund, Thodoris Lykouris, and Wentao Weng. Efficient decentralized multi-agent learning in asymmetric bipartite queueing systems. *Operations Research*, 72(3):1049–1070, 2024.

[14] Jennifer M George and J Michael Harrison. Dynamic control of a queue with adjustable service rate. *Operations research*, 49(5):720–731, 2001.

[15] Huiwen Jia, Cong Shi, and Siqian Shen. Online learning and pricing for service systems with reusable resources. *Operations Research*, 72(3):1203–1241, 2024.

[16] Vasiliki Kostami and Sampath Rajagopalan. Speed–quality trade-offs in a dynamic model. *Manufacturing & Service Operations Management*, 16(1):104–118, 2014.

[17] Subhashini Krishnasamy, Ari Arapostathis, Ramesh Johari, and Sanjay Shakkottai. On learning the cμ rule in single and parallel server networks. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 153–154. IEEE, 2018.

[18] Subhashini Krishnasamy, Rajat Sen, Ramesh Johari, and Sanjay Shakkottai. Learning unknown service rates in queues: A multiarmed bandit approach. *Operations Research*, 69(1):315–330, 2021.

[19] Ravi Kumar, Mark E Lewis, and Huseyin Topaloglu. Dynamic service rate control for a single-server queue with markov-modulated arrivals. *Naval Research Logistics (NRL)*, 60(8):661–677, 2013.

[20] Bai Liu, Qiaomin Xie, and Eytan Modiano. Rl-QN: A reinforcement learning framework for optimal control of queueing systems. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems*, 7(1):1–35, 2022.

[21] Devavrat Shah, Qiaomin Xie, and Zhi Xu. Stable reinforcement learning with unbounded state space. In *Learning for Dynamics and Control*, pages 581–581. PMLR, 2020.

[22] William R Small, Batia Wiesenfeld, Beatrix Brandfield-Harvey, Zoe Jonassen, Soumik Mandal, Elizabeth R Stevens, Vincent J Major, Erin Lostraglio, Adam Szerencsy, Simon Jones, et al. Large language model–based responses to patients' in-basket messages. *JAMA network open*, 7(7):e2422399–e2422399, 2024.

[23] Shaler Stidham Jr and Richard R Weber. Monotonic and insensitive optimal policies for control of queues with undiscounted costs. *Operations Research*, 37(4):611–625, 1989.

[24] Ming Tai-Seale, Sally L Baxter, Florin Vaida, Amanda Walker, Amy M Sitapati, Chad Osborne, Joseph Diaz, Nimit Desai, Sophie Webb, Gregory Polston, et al. Ai-generated draft replies integrated into health records and physicians' electronic communication. *JAMA Network Open*, 7(4):e246565–e246565, 2024.

[25] Sanne van Kempen, Jaron Sanders, Fiona Sloothaak, and Maarten G Wolf. Learning payoffs while routing in skill-based queues. *arXiv preprint arXiv:2412.10168*, 2024.

[26] Neil Walton and Kuang Xu. Learning and information in stochastic networks and queues. In *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications*, pages 161–198. INFORMS, 2021.

[27] Richard R Weber and Shaler Stidham Jr. Optimal control of service rates in networks of queues. *Advances in Applied Probability*, 19(1):202–218, 1987.

[28] Li Xia, Qi-Ming He, and Attahiru Sule Alfa. Optimal control of state-dependent service rates in a map/m/1 queue. *IEEE Transactions on Automatic Control*, 62(10):4965–4979, 2017.

[29] Sherry Yan, Wendi Knapp, Andrew Leong, Sarira Kadkhodazadeh, Souvik Das, Veena G Jones, Robert Clark, David Grattendick, Kevin Chen, Lisa Hladik, et al. Prompt engineering on leveraging large language models in generating response to inbasket messages. *Journal of the American Medical Informatics Association*, page ocae172, 2024.

[30] Zixian Yang, R Srikant, and Lei Ying. Learning while scheduling in multi-server systems with unknown statistics: Maxweight with discounted ucb. In *International Conference on Artificial Intelligence and Statistics*, pages 4275–4312. PMLR, 2023.

[31] Mohammad Zhalechian, Esmaeil Keyvanshokooh, Cong Shi, and Mark P Van Oyen. Data-driven hospital admission control: A learning approach. *Operations Research*, 71(6):2111–2129, 2023.

[32] Yueyang Zhong, John R Birge, and Amy R Ward. Learning to schedule in multiclass many-server queues with abandonment. *Operations Research*, 2024.