Data-Driven Percentile Optimization for Multi-Class Queueing Systems with Model Ambiguity: Theory and Application

Faculty Research Working Paper Series

Austin Bren
Arizona State University

Soroush Saghafian
Harvard Kennedy School

February 2018

RWP18-008

The views expressed in the **HKS Faculty Research Working Paper Series** are those of the author(s) and do not necessarily reflect those of the John F. Kennedy School of Government or of Harvard University. Faculty Research Working Papers have not undergone formal review and approval. Such papers are included in this series to elicit feedback and to encourage debate on important public policy challenges. Copyright belongs to the author(s). Papers may be downloaded for personal use only.

Data-Driven Percentile Optimization for Multi-Class Queueing Systems with Model Ambiguity: Theory and Application

Austin Bren[1], Soroush Saghafian[2]

[1]School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ 85287

[2]Harvard Kennedy School, Harvard University, Cambridge, MA 02138

Multi-class queueing systems widely used in operations research and management typically experience ambiguity in real-world settings in the form of unknown parameters. For such systems, we incorporate robustness in the control policies by applying a data-driven percentile optimization technique that allows for (1) expressing a controller's optimism level toward ambiguity, and (2) utilizing incoming data in order to learn the true system parameters. We show that the optimal policy under the percentile optimization objective is related to a closed-form priority-based policy. We also identify connections between the optimal percentile optimization and $c\mu$-like policies, which in turn enables us to establish effective but easy-to-use heuristics for implementation in complex systems. Using real-world data collected from a leading U.S. hospital, we also apply our approach to a hospital Emergency Department (ED) setting, and demonstrate the benefits of using our framework for improving current patient flow policies.

Key words: Model Ambiguity, Data-Driven Optimization, ED Operations

1. Introduction

Multi-class queueing systems require dynamic control in environments where servers must process multiple types of jobs that vary with respect to holding costs, service rates, and other defining characteristics. These types of queueing systems are widely used to model call centers, hospitals, manufacturing lines, and service operations, where elements in the queue can be classified based on differing levels of urgency, processing time, or other attributes. For example, in a hospital Emergency Department (ED), patients are classified through a triage system, which differentiates them based on their severity, medical complexity, or other conditions (see, e.g., Saghafian et al. (2012), Saghafian et al. (2014), and the references therein). Hence, a natural way to analyze ED patient flow is via a multi-class queueing system which separates patients based on their attributes.[1]

In such systems, when all parameters are known, many well-established policies like the $c\mu$ rule have been shown to be optimal for optimizing the system's performance (see, e.g., Van Mieghem (1995) and Buyukkoc et al. (1985)). However, the assumption that all the model parameters are perfectly known is often unrealistic, especially in settings with little supporting data, inaugural system launch, or various other sources of ambiguity. A manager with incorrect parameter specifications may enforce policies that perform poorly, or may not have confidence in using a policy that

[1] See, e.g., Saghafian et al. (2015) for a recent review of various models used to optimize patient flow and improve ED operations.

is obtained from a model with parameters that s/he does not fully trust. In an effort to combat such mistrust, we consider a form of model ambiguity caused by the ambiguity in parameters termed *parameter ambiguity*, and develop strategies that directly take these into account.

Traditionally, robust optimization protects against parameter ambiguity by utilizing a minimax objective on an ambiguity set of parameters which are assumed to contain the true system parameters. However, this type of robustness (1) can result in overly pessimistic policies and (2) ignores the significant potential to learn about the true system parameters from data acquired both before and after system launch. Even when this pessimism is reduced by choosing tighter ambiguity sets, the policies generated are not capable of learning from incoming data. To avoid these deficiencies, we model parameter ambiguity via a Partially Observable Markov Decision Process (POMDP), an extension of Markov Decision Processes (MDPs), which allows for (1) imperfect state knowledge, and (2) learning in a Bayesian manner. A POMDP supports the distribution of the underlying system parameters, known as the belief space, and updates this distribution to reflect received observations. This is ideal from a learning perspective; however, in a POMDP, the decision-maker is assumed to have an initial prior belief which is often a subjective value, guided by scarce data, error-prone expert opinion, intuition, or instinct. For these reasons, Bayesian critics distrust such learning mechanisms, citing the unreliability of the prior specification in real-world applications[2].

To incorporate robustness to such a prior belief (hence gaining robustness to parameter ambiguity), we integrate our POMDP model with a *percentile optimization* approach. Percentile optimization is traditionally used to avoid overly conservative policies by offering a certain level of performance over a percentage of the ambiguity set (see, e.g., Delage and Mannor (2010) and Nemirovski and Shapiro (2006)). We extend percentile optimization in order to incorporate robustness to the belief about the model parameters rather than relying on a robustness generated directly from the parameters themselves. In this way, we investigate strategies where the controller learns the main model parameters (e.g., unknown service rates) while simultaneously controlling the underlying system for superior performance, which contrasts with robust techniques that only focus on parameter ambiguity sets. Thus, our framework allows generating policies that are robust to parameter ambiguities (considering a manager's pessimism level), while simultaneously learning about the true model from data/observation of the system's performance in a Bayesian manner.

Our main contributions stem from extending the robust percentile optimization approach for integration with POMDPs. We find that the percentile optimization objective reduces to the minimax and minimin objectives when the optimism level is set to its lowest and highest values, respectively

[2] Though we mainly focus on a queueing model, our approach can be used for the general class of Bayesian decision-making problems where the decision-maker faces ambiguity with respect to parameters that shape his/her prior (see Corollaries 2 and 3 in Online Appendix B).

and show that the optimal policies under these objectives are myopic $c\mu$ priority policies. Understanding the non-robust problem (which assumes a specified initial belief) proves to be essential in finding robust policies where the belief is subject to ambiguity. We find that optimal robust policies can be formed using specific non-robust policies via a geometric structure known as the *convex floating body*. Therefore, to solve the robust percentile problem, we first solve the non-robust problem that has a known initial belief. As the rate of observations increases, we find that a priority-based policy that acts as as an extension of the well-known $c\mu$ rule becomes asymptotically optimal to the non-robust problem. This policy, which we term E$c\mu$, is myopic and prioritizes the class with the largest expected $c\mu$ value. The proposed E$c\mu$ policy utilizes incoming data for learning (unlike the traditional $c\mu$ rule), and is extremely simple to implement.

Due to its foundation in POMDPs, the robust framework we consider is computationally ambitious and necessitates finding tractable methods for implementation. Using the analytical insights gained from the connection between non-robust and robust policies, constraints via the convex floating body, and the relation of E$c\mu$ to the non-robust objective, we develop a heuristic for the robust problem that (1) is highly scalable to large problem instances, and (2) shows strong performance in extensive simulation experiments. We also develop analytical bounds to the non-robust problem based on queueing systems with fully known parameters. These bounds are (1) tight under a variety of conditions, and (2) can be used to more effectively compute optimal robust policies. Furthermore, since the bounds are based on non-learning policies, they can be computed in an efficient manner.

Finally, we demonstrate the benefits of our approach in a real-world setting by utilizing data that we have collected from a leading U.S. hospital, and by establishing the advantages of using our framework in improving the current ED patient flow policies. Our percentile optimization framework is the first study in the literature to yield data-driven policies for use in EDs that hedge against parameter ambiguity. We find that highly congested EDs are well-suited to our percentile optimization framework, especially in geographical areas with uncertain/unstable patient population characteristics. Additionally, our approach explicitly avoids overly conservative policies that focus only on the "worst-case" scenarios. As a result, we find that percentile optimization performs well over a large spectrum of optimism/pessimism. In particular, our simulations calibrated with hospital data suggest that, by using our approach, an ED manager can typically improve performance by $10\% - 15\%$ regardless of his/her disposition.

The rest of the paper is organized as follows. In Section 2, we provide a literature review of the related studies. Section 3 introduces the non-robust continuous-time formulation of our problem, which is uniformized into a discrete-time problem in Section 3.1, and lays the foundation for the percentile framework developed in Section 3.2. We provide the majority of our analytical insights in Sections 4 and 5, where we establish optimal policies for the non-robust and robust formulations, and identify upper/lower bound results. Section 6 introduces a heuristic to the robust problem that is rooted in the analytical insights generated from Section 4. In Section 7, we present various numerical experiments, discuss the application of our work for improving patient flow in EDs, and use real-world data obtained from a leading U.S. hospital to evaluate the potential benefits of our approach. Finally, in Section 8, we present our concluding remarks.

2. Literature Review

The literature surrounding multi-class queueing systems aims to analyze complex structures and discover their optimal control policies such as the $c\mu$ policy and its variations (see, e.g., Buyukkoc et al. (1985), Van Mieghem (1995), Saghafian and Veatch (2016), and the references therein). A common tool used to analyze and control such systems is Markov Decision Processes (MDPs). However their use is limited to the unrealistic case where the decision-maker is assumed to completely know all the parameters of the model (e.g. service rates). Most notably, this includes a perfect knowledge assumption of the transition matrices that guide a system's state transitions. This assumption can be problematic in various practical applications in which service rates (or other parameters) are not perfectly known. Mannor et al. (2007) and Nilim and El Ghaoui (2005) found that small changes in such parameters can result in significant differences in decision-making strategies. However, a synthesis of most studies on dynamic control in queueing systems indicates the use of tools that heavily rely on a full knowledge about the system's parameters. This is despite the fact that in practice such parameters are typically unknown and often hard to estimate.

Robust methods applied to queueing models are largely involved with reducing the computational burden of characterizing queueing metrics and policies. Su (2006) studies a fluid approximation of a multi-class queueing model's holding cost under a robust paradigm established by Bertsimas et al. (2004) and Bertsimas and Sim (2004). Bertsimas et al. (2011) focuses on finding bounds for performance measures through a method rooted in robust optimization, and studies the performance of this method on tandem and multi-class single server queueing networks. Jain et al. (2010) finds that a queueing network with control over traffic intensities has a simple threshold type policy under a robust objective. For more recent studies on robust techniques used in queueing systems we refer to Pedarsani et al. (2014), Bandi and Bertsimas (2012), Bandi et al. (2015), and the references therein. This stream of research is mainly aimed at increasing tractability by focusing on "worst-case" (i.e., fully pessimistic) scenarios, and establishing related performance metrics. Unlike this stream, our goal is to provide policies that (1) are more optimistic (i.e., less conservative), and (2) incorporate learning from online system-run data/observations.

Adding robustness when facing parameter ambiguity is a topic of significant interest to a variety of fields including economics, operations research/management, computer science, and decision theory among others. Typically, robustness in MDPs is added using "minimax" (or robust optimization) techniques, since this often results in tractable analyses as shown in Nilim and El Ghaoui (2005), Iyengar (2005), and the references therein. Other studies such as Chen and Farias (2013) deal with ambiguities by considering policies that offer guarantees on expected performance. Still other methods of incorporating robustness include regret minimization (Lim et al. (2012)), relative entropy (Bagnell et al. (2001)), and martingale-based approaches (Hansen and Sargent (2007)) that provide less conservative, and hence, potentially more realistic alternatives to minimax techniques. In particular, Delage and Mannor (2010) identify a robust approach applied to MDPs called percentile optimization that effectively avoids over-conservatism (see also Nemirovski and Shapiro (2006) and Wiesemann et al. (2013) for related studies). Instead of finding policies that are tailored to work well in worst-case scenarios, the percentile optimization method finds policies that maximize performance with respect to a level of belief about the true parameters for a given level of optimism.[3]

Chow et al. (2017) also utilize this type of robustness to develop risk-constrained policies for MDPs. However, a significant deficit in current percentile optimization approaches is the lack of ability to *learn* about the true parameters over time. Delage and Mannor (2007) work to fill this gap via a similar formulation to our approach, and find second-order approximations to MDPs that experience transition parameter uncertainty. However, the Dirichlet-type uncertainty assumed in transition parameters does not fit our queueing problem, and in our work, we extend the percentile optimization approach with respect to ambiguity in the initial belief. Thus, system data/observations can be used for learning the true operational model, and as we will show, this ability to learn itself adds a strong layer of robustness for controlling queueing systems (e.g., hospital patient flows) that face parameter ambiguity. Learning to overcome ambiguities are also discussed in Bassamboo and Zeevi (2009), which models a call center application using a data-driven technique. However, their work (1) does not include any notion of robustness, and (2) focuses on near-optimal policies with performance bounds. Our work differs in modeling approach by our joint focus on learning and robustness, and in methodology by our contributions in characterizing the exact optimal policies.

Data-driven parameter learning has been incorporated in POMDPs: Ross et al. (2011) explores a finite-horizon POMDP model that updates a posterior of its parameter belief in a Bayesian manner, and Thrun (1999) investigates a POMDP in continuous action and state spaces that relies on par-

[3] The percentile objective originally arose in single-period contexts (see, e.g., Charnes and Cooper (1959) and Prékopa (1995)).

ticle filtering techniques to determine the belief state. Unlike learning mechanisms, robust methods are almost non-existent in POMDP frameworks. Osogami (2015) shows that traditional minimax approaches with convex ambiguity sets can be extended to POMDPs while still retaining its structural features (such as convexity). In a new approach, Saghafian (2017) extends POMDPs to a new class termed Ambiguous POMDPs (APOMDPs) which incorporates ambiguity in transition and observation probabilities in a robust fashion. The robustness in Saghafian (2017) is achieved by considering α-maximin (α-MEU) preferences, and by incorporating the decision-maker's temperament toward model ambiguity. Different from the APOMDP approach of Saghafian (2017), we utilize a percentile optimization objective to hedge against ambiguities.

3. The Multi-Class Queueing Control Problem with Parameter Ambiguity

We begin by considering a continuous time multi-class queueing control problem with preemption, where a single[4] server is responsible for serving n classes of customers over an infinite time horizon. Unlike the traditional version of this model, we assume the controller does not know the main parameters of the system, and hence, is faced with parameter ambiguity. We focus on the case where the ambiguity is on service rates. To this end, we start by excluding dynamic arrivals to the system, and instead consider a *clearing system*[5] version of the problem. We relax this assumption in Sections 7 and A.4 by allowing for dynamic arrivals, and find that many of our major results are transferable from the clearing system. Our general approach can also be used for systems where arrival rates or other parameters are ambiguous by modifying the underlying dynamic program to include these components along with their learning mechanisms. However, this appears to increases the problem's complexity without providing additional insights.

With $\mathcal{N} = \{1, \ldots, n\}$ denoting the set of customer classes, we assume each customer of class $i \in \mathcal{N}$ accrues a cost $\hat{c}_i > 0$ for each unit of time spent in the system. Let $\hat{\mathbf{c}} = (\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_n)$ be the cost vector, $\alpha \in (0, \infty)$ the discount rate, and $\mathbf{X}(t) = (X_1(t), X_2(t), \ldots, X_n(t))$ the vector of the number of customers in the system, where $X_i(t)$ is number of class i customers in the system at time t. In line with many robust approaches, we begin by outlining an ambiguity set (i.e. a "cloud" of models) that is assumed to include the true model. To this end, and for tractability, we assume service times for each class are i.i.d. exponential[6] random variables with unknown rates for each

[4] For analytical tractability, we restrict our attention to single-server scenarios. Cases with multiple servers may interfere with some of our main analytical results, notably the relation to multi-armed bandit problems and the optimality of $c\mu$-like policies. In Section 7, we investigate the robustness of the insights we gain via simulation experiments.

[5] Clearing systems are typically used to model busy periods by focusing on the customers/jobs already in the system. The goal is then to clear the system with the minimum cost.

[6] In Section 7 we relax the exponential distribution assumption. For instance, our data shows that service times in EDs are close to log-normal. As we will show, our main insights and heuristic control procedures remain effective even when the service times are not exponential.

class. The true service rate for each class $i \in \mathcal{N}$ is chosen by Nature at time $t = 0$, and lies within ambiguity set[7] $\mathcal{M}_i = \{\hat{\mu}_{i,1}, \ldots, \hat{\mu}_{i,m_i}\}$. We further assume that service times for different classes are independent. For future notational convenience, we let $\mathcal{J}_i = \{1, \ldots, m_i\}$. Throughout the paper, we assume $m_i \in \mathbb{N}$, and $\hat{\mu}_{i,j} \neq \hat{\mu}_{i,k}$ for each $i \in \mathcal{N}$ and distinct $j, k \in \mathcal{J}_i$. Though the ambiguity sets $\mathcal{M}_i$ are discrete, the continuous case can be approximated arbitrarily closely by increasing the number of potential service rates m_i to make the mesh size of $\mathcal{M}_i$ close to zero.

Over time, the controller can learn the true service rates by observing the process history which includes all previous service durations, control actions, and observations of service completions. For Markovian systems with incomplete information, it has been shown in Bertsekas (1995) that the Bayesian belief on the unknown parameters with respect to the observed process history is a *sufficient statistic*. We let $\mathcal{B}$ be the set of all such sufficient statistics, i.e., the set of possible belief distributions on the system's service parameters. Letting $m = \sum_{i \in \mathcal{N}} m_i$, each $\mathbf{b} \in \mathcal{B}$ is an m-dimensional vector of the form $\mathbf{b} = (b_{1,1}, b_{1,2}, \ldots, b_{1,m_1}, b_{2,1}, \ldots, b_{n,m_n})$ with the condition that each $b_{i,j} \geq 0$ and that $\sum_{j=1}^{m_i} b_{i,j} = 1$ for each $i \in \mathcal{N}$. In this setting, if $\hat{\mu}_i^* \in \mathcal{M}_i$ is the true (unknown) service rate for class $i \in \mathcal{N}$, $\mathbb{P}\left(\hat{\mu}_{i,j} = \hat{\mu}_i^* | \mathbf{b}\right) = b_{i,j}$. We further assume that the observation made after serving one class does not affect the belief about another. This is aligned with the assumption that service time of one class is independent of that of another class.

To find policies that optimally prescribe which customer class the server should serve at any time, given (1) the available information summarized in the current belief about the service rates, and (2) the number of customers in each queue, it is known that one can restrict attention to policies that are deterministic, stationary, and Markovian (see, e.g., Sondik (1971), Smallwood and Sondik (1973), and Bertsekas (1995)). Consequently, an admissible non-anticipative policy π at each time t maps the current belief and queue length information (information state) to the set of admissible actions: $\pi : \mathbf{X}(t) \times \mathcal{B} \to \mathcal{A}(\mathbf{X}(t)) = \{a \in \mathcal{N} : \mathbf{X}(t) - \mathbf{e}_a \geq 0\}$, where $\mathbf{e}_a$ refers to an n-dimensional vector with a one at the a-th position and zeros elsewhere. Our model described above is schematically illustrated in Figure 1.

We let Π be the set of all admissible policies, and $\mathbf{X}^\pi(t) = (X_1^\pi(t), X_2^\pi(t), \ldots, X_n^\pi(t)) \in \mathbb{Z}_+^n$ be the number of customers in the system under policy $\pi \in \Pi$ at time t. In Appendix B, we show via Lemma 17 shows that idling a server is always suboptimal; hence, we consider only non-idling policies in our analysis. For a given policy π, the expected discounted *true cost* the system experiences is

$$\mathrm{E}_\pi \left[\int_{t=0}^{\infty} e^{-\alpha t} \hat{\mathbf{c}} \mathbf{X}^\pi (t)^{\mathrm{T}} dt | \mathbf{X}(0) \right],$$

[7] The general nature of our ambiguity sets enhances the flexibility of our framework. For ambiguity sets reminiscent of other robust literature, we may choose build each $\mathcal{M}_i$ to surround some nominal value estimated from historical data. This is in fact the strategy we use in our ED application of Section 7.1

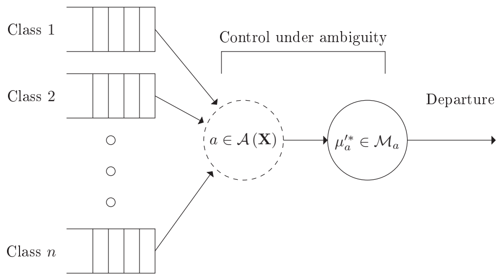
Class 1
Class 2
Class n
Control under ambiguity
$a \in \mathcal{A}(\mathbf{X})$
$\mu_a'^{*} \in \mathcal{M}_a$
Departure

Figure 1 The server serves a class a customer with an unknown rate $\hat{\mu}_a^*$ belonging to ambiguity set $\mathcal{M}_a$.

given the true transition parameters chosen by Nature at time $t = 0$, where the notation "T" represents transpose, and E_π is expectation with respect to the probability measure induced by π. However, since the controller does not know the true transition matrix (as service rates are unknown), we are interested in the expected cost with respect to the controller's belief:

$$\mathrm{J}^{\pi}\left(\mathbf{X}(0), \mathbf{b}(0)\right) = \mathrm{E}_{\pi,\mathbf{b}(0)}\left[\int_{t=0}^{\infty} e^{-\alpha t} \hat{\mathbf{c}} \mathbf{X}^{\pi}\left(t\right)^{\mathrm{T}} dt | \mathbf{X}\left(0\right)\right], \tag{1}$$

where $\mathrm{E}_{\pi,\mathbf{b}(0)}$ denotes expectation with respect to both the initial belief $\mathbf{b}(0)$ and π. We refer to $\mathrm{J}^{\pi}(\mathbf{X}(0), \mathbf{b}(0))$ as the *non-robust cost*, since it assumes a perfectly assigned $\mathbf{b}(0)$ (which is inevitably hard to quantify for any decision-maker who is faced with model ambiguity). The optimal non-robust cost is then given by $\mathrm{J}\left(\mathbf{X}(0), \mathbf{b}(0)\right) = \inf_{\pi \in \Pi} \mathrm{J}^{\pi}\left(\mathbf{X}(0), \mathbf{b}(0)\right)$. In what follows, we first use uniformization to work with the discrete-time model of the non-robust scenario, where the initial belief is given. We then adopt percentile optimization to enable the decision-maker/controller to reduce his/her reliance on $\mathbf{b}(0)$, and thereby make robust decisions.

3.1. A Discrete-Time Non-Robust Framework

The continuous-time Markov chain $\{\mathbf{X}^{\pi}(t) : t \geq 0\}$ can be converted to a discrete-time equivalent using the well-known uniformization technique (Lippman (1975)). Following this method, we first select a *uniformized* exponentially distributed random variable ξ with a rate $\psi > \max_{i \in \mathcal{N}, j \in \mathcal{J}_i} \hat{\mu}_{i,j}$ which serves as our rate of observations made as follows. If the server completes service to a customer of class i a uniformized unit of time (i.e., at the end of each period), an observation indicating the "successful" service to class i is recorded. Otherwise, if no service completion is observed within this time, an observation is recorded indicating an "incomplete" service to class i. We note that this *uniformization rate* ψ may be arbitrarily large so as to approximate continuous observations.

We let σ be the Bayesian learning operator such that $\sigma\left(\mathbf{b}, a, \theta\right)$ is an m-dimensional vector representing the updated belief after taking action a and receiving observation θ, when the prior belief is $\mathbf{b}$. Since there are only two outcomes for observations for any given action, we let "+" signify an observed service completion ("success") during the uniformized time period, and "−" represent an incomplete service ("failure") in that period. In this setting, we use a discrete-time dynamic program with uniformized parameters $\mu_{i,j} = \hat{\mu}_{i,j}/\psi$. For notational convenience, we let $\mathrm{E}\left[\mu_i | \mathbf{b}\right] = \sum_{j=1}^{m_i} \mu_{i,j} b_{i,j}$

be the expected service transition probability of class $i \in \mathcal{N}$ given belief $\mathbf{b}$. In this way, the Bayesian learning operator updates belief $\mathbf{b}$ with components $b_{i,j}$ to belief $\bar{\mathbf{b}} = \sigma(\mathbf{b}, a, \theta)$ with components $\bar{b}_{i,j} = \sigma(\mathbf{b}, a, \theta)_{i,j}$, where $a, i \in \mathcal{N}, j \in \mathcal{J}_i$, and

$$\sigma(\mathbf{b}, a, +)_{i,j} = \begin{cases} \frac{\mu_{a,j} b_{a,j}}{\sum_{k=1}^{m_a} \mu_{a,k} b_{a,k}} = \frac{\mu_{a,j} b_{a,j}}{\mathrm{E}[\mu_a|\mathbf{b}]} & : i = a \\ b_{i,j} & : i \neq a \end{cases} \tag{2}$$

for a successful service observation, and

$$\sigma(\mathbf{b}, a, -)_{i,j} = \begin{cases} \frac{\left(1-\mu_{a,j}\right) b_{a,j}}{\sum_{k=1}^{m_a} \left(1-\mu_{a,k}\right) b_{a,k}} = \frac{\left(1-\mu_{a,j}\right) b_{a,j}}{(1-\mathrm{E}[\mu_a|\mathbf{b}])} & : i = a \\ b_{i,j} & : i \neq a \end{cases} \tag{3}$$

for a failed service observation. Equations (2) and (3) are established due to the fact that under realized parameter $\mu_{a,j}$, the probability of successful service in a given period is $\mu_{a,j}$ and probability of incomplete service is $(1 - \mu_{a,j})$. With this, and defining a discrete-time discounting factor $\beta = \frac{\psi}{\psi+\alpha}$ and instantaneous cost $\mathbf{c}\mathbf{X}^{\mathrm{T}} = \frac{\hat{\mathbf{c}}\mathbf{X}^{\mathrm{T}}}{\psi+\alpha}$, we can identify the non-robust optimal policy and the associated cost via the dynamic program

$$\begin{aligned} \mathrm{V}_{t+1}(\mathbf{X}, \mathbf{b}) = \mathbf{c}\mathbf{X}^{\mathrm{T}} + \beta \Bigg[\min_{a \in \mathcal{A}(\mathbf{X})} \Big\{ & \mathrm{E}[\mu_a|\mathbf{b}] \, \mathrm{V}_t(\mathbf{X} - \mathbf{e}_a, \sigma(\mathbf{b}, a, +)) \\ & + (1 - \mathrm{E}[\mu_a|\mathbf{b}]) \, \mathrm{V}_t(\mathbf{X}, \sigma(\mathbf{b}, a, -)) \Big\} \Bigg], \end{aligned} \tag{4}$$

with the terminal condition $\mathrm{V}_0(\mathbf{X}, \mathbf{b}) = \mathbf{c}\mathbf{X}^{\mathrm{T}}$. In this setting, taking the limit as $t \to \infty$, we define $\mathrm{V}(\mathbf{X}, \mathbf{b}) = \lim_{t\to\infty} \mathrm{V}_t(\mathbf{X}, \mathbf{b})$, and note that $\mathrm{V}(\mathbf{X}, \mathbf{b}) = \inf_{\pi\in\Pi} \mathrm{J}^{\pi}(\mathbf{X}, \mathbf{b})$ (see Lemma 11 in Online Appendix B for a rigorous treatment), where $\mathrm{J}^{\pi}(\mathbf{X}, \mathbf{b})$ is defined in (1). To account for evaluating non-optimal policies, we let $\mathrm{V}^{\pi}_{t+1}(\mathbf{X}, \mathbf{b})$ be a value function similar to that of the dynamic program (4) with minimization operator replaced by serving the class prescribed by policy π. Likewise, we let $\mathrm{V}^{\pi}(\mathbf{X}, \mathbf{b}) = \lim_{t\to\infty} \mathrm{V}^{\pi}_t(\mathbf{X}, \mathbf{b})$ be the infinite-horizon dynamic program value function under policy π.

3.2. Gaining Robustness via Percentile Optimization

Since the controller is facing ambiguity with respect to the true model, s/he may distrust his/her initial prior on the cloud of models, $\mathbf{b}(0)$. The specification of $\mathbf{b}(0)$ is subject to model sensitivities, especially in applications in which there is little or highly variable data to perfectly quantify it. Often, the selection of a prior is a process that requires sussing out probabilities and parameter values from experts in the field, which can be a highly subjective and inaccurate task[8]. In order to achieve robustness to the selection of the initial prior, it is necessary to investigate policies that are not endowed with any particular initial prior, but rather can initialize with any desired prior. We refer to such policies that lie within the set $\Pi_h = \{\pi \in \Pi | \mathbf{b}(0) \in \mathcal{B}\}$ as *prior-flexible* policies.

[8] This is indeed a general criticism to Bayesianism and goes well beyond the queueing setting of this paper.

In traditional robust optimization, one would choose a policy assuming that Nature, being an antagonistic character, picks the worst-case initial belief vector $\mathbf{b}(0)$ for a chosen policy. Hence, the traditional *minimax* robust objective can be defined by first considering the worst-case cost under a policy $\pi_h \in \Pi_h$:

$$\mathrm{R}^{\pi_h}(\mathbf{X}) = \max_{\mathbf{b}\in\mathcal{B}} \mathrm{V}^{\pi_h}(\mathbf{X},\mathbf{b}).$$

The cost under the minimax robust objective is then $\mathrm{R}(\mathbf{X}) = \inf_{\pi_{h\in\Pi_h}} \mathrm{R}^{\pi_h}(\mathbf{X})$. In this setting, the controller assumes that Nature will pick the transition parameters that result in the maximum cost for any given policy, and chooses a policy that minimizes the cost of this worst-case outcome.

In sharp contrast to this type of robustness, which typically yields overly pessimistic control policies, is the overly optimistic *minimin* objective defined by:

$$\mathrm{N}^{\pi_h}(\mathbf{X}) = \min_{\mathbf{b}\in\mathcal{B}} \mathrm{V}^{\pi_h}(\mathbf{X},\mathbf{b}),$$

and $\mathrm{N}(\mathbf{X}) = \inf_{\pi_{h\in\Pi_h}} \mathrm{N}^{\pi_h}(\mathbf{X})$, under which the controller chooses a policy assuming Nature picks the transition parameters resulting in the best-case cost for any given policy. In what follows, we first show that both minimax and minimin optimal policies are within the well-known class of $c\mu$ policies. Thus, they (1) are fully myopic, and (2) have very simple forms.

Proposition 1 **(Minimax/Minimin $c\mu$ Optimal Policies)**. *At any state* $(\mathbf{X},\mathbf{b})$*, optimal policies to the minimax and minimin objectives serve classes* $\arg\max_{a\in\mathcal{A}(\mathbf{X})}\left(\min_{j\in\mathcal{J}_a} c_a\mu_{a,j}\right)$ *and* $\arg\max_{a\in\mathcal{A}(\mathbf{X})}\left(\max_{j\in\mathcal{J}_a} c_a\mu_{a,j}\right)$*, respectively.*

Proposition 1 establishes that optimal policies under both minimax and minimin objectives are myopic priority disciplines (known as the $c\mu$ rule) with respect to the smallest and largest transition rates within the ambiguity set for each class, respectively. However, it should be noted that such policies (1) ignore the potential for learning from the system behavior, and (2) only consider the potentially unrealistic extreme best and worst-case scenarios and can perform poorly in real-world applications. To address this deficit, we next investigate how the percentile optimization approach provides a balancing alternative between these two extreme strategies, while incorporating learning about the hidden probabilities associated with the true transition parameters (i.e., service rates).

To this end, for a given $\epsilon \in [0,1]$, we define the percentile optimization program:

$$\mathrm{Y}^{\pi_h}(\mathbf{X},\epsilon) = \inf_{y_\epsilon \in [\mathrm{N}^{\pi_h}(\mathbf{X}),\mathrm{R}^{\pi_h}(\mathbf{X})]} y_\epsilon \tag{5}$$

$$s.t.\ \ \mathbb{P}_{\mathbf{B}}\left(\mathrm{V}^{\pi_h}(\mathbf{X},\mathbf{B}) \le y_\epsilon\right) \ge 1-\epsilon, \tag{6}$$

and let $\mathrm{Y}(\mathbf{X},\epsilon) = \inf_{\pi_{h\in\Pi_h}} \mathrm{Y}^{\pi_h}(\mathbf{X},\epsilon)$ represent the optimal percentile objective. In (5), we impose that $\mathrm{N}^{\pi_h}(\mathbf{X}) \le y_\epsilon \le \mathrm{R}^{\pi_h}(\mathbf{X})$ so that the value of the objective is within the most optimistic and pessimistic values attainable for any given belief in accordance with the policy, hence enforcing "realizable" expected costs. The probability operator, $\mathbb{P}_{\mathbf{B}}$, in (6) is defined with respect to a specified probability density function over the prior belief space[9], where $\mathbf{B}$ is a random variable whose realization is $\mathbf{b}$. The percentile optimization program (5)-(6) allows us to find a *chance-constrained* policy: it emphasizes policy performance over a portion of the belief space. We thus term the policy that is the solution under the optimal percentile objective as $(1-\epsilon)\%$ *chance-constrained policy.* Intuitively, the smaller the ϵ, the more protection from poor parameter settings since the proportion of the belief space that performs worse than y_ϵ becomes smaller.

It is important to note that the percentile objective acts as a bridge between non-robust and robust objectives; expressing a manager's optimism level is a core ambition of this type of robustness. For instance, the chance-constrained policy reduces to the minimax and minimin policies when ϵ is 0 and 1, respectively.

Proposition 2 **(Percentile/Minimax/Minimin Relationship)**. *The percentile objective, minimax, and minimin policies share the following relation:*

(i) If $\epsilon = 0$ and $\mathbb{P}_{\mathbf{B}}(\mathbf{B}=\mathbf{b}) > 0$ for all $\mathbf{b} \in \mathcal{B}$, then the optimal policy and cost under both minimax and percentile objectives are the same.

(ii) If $\epsilon = 1$, then the optimal policy and cost under the minimin and percentile objectives are the same.

The additional condition $\mathbb{P}_{\mathbf{B}}(\mathbf{B}=\mathbf{b}) > 0$ for all $\mathbf{b} \in \mathcal{B}$ in part (i) is necessary, since $\mathbb{P}_{\mathbf{B}}$ with zeros allows percentile objective to "ignore" certain portions of the belief space while still satisfying constraint (6). For example, if $\mathbb{P}_{\mathbf{B}}$ is the degenerate distribution with respect to a point $\mathbf{b}$, $\mathrm{Y}(\mathbf{X}, 0) = \mathrm{V}(\mathbf{X}, \mathbf{b})$.

4. Structure of Optimal Policies under the Percentile Objective

Analyzing program (5)-(6) is inherently complex both analytically and computationally. However, we find that the solution to this program is linked to solving the non-robust problem. Hence, we first consider the solution of the dynamic program (4), identify important characteristics of these solutions over the belief space, establish the link between non-robust and robust policies, and finally work to characterize optimal percentile policies. In Section 6, we develop an easy-to-use heuristic based on these insights to facilitate tractable solutions.

As the observation rate increases, tending toward continuous observations, the non-robust problem can be transferred to a multi-armed bandit (MAB) problem by noting that (1) under any action, only the belief about transition parameters and number of customers in the served class (the "arms"

[9] One may criticize the use of the percentile objective due to the potential ambiguity of $\mathbb{P}_{\mathbf{B}}$; however, it should be noted that this is a second-order distribution, and perturbations in $\mathbb{P}_{\mathbf{B}}$ result in very similar convex floating bodies, which is the geometric structure investigated in Section 4 that generates our optimal robust policies.

of a MAB) change, and (2) the "discounted cost" can be reinterpreted as "discounted savings" due to our clearing system environment (for further discussion, see Lemma 3 in Online Appendix B). MAB problems are typically solved by indexing policies related to the expected savings in cost experienced through exclusively serving one class over time.

To take advantage of the above-mentioned connection, we term the myopic policy that serves the class $a \in \mathcal{A}(\mathbf{X})$ with largest value of $c_a \mathrm{E}\left[\mu_a|\mathbf{b}\right]$ the "E$c\mu$" policy. Thus, we denote $\pi^{c\mu}$ that serves $\arg\max_{a\in\mathcal{A}(\mathbf{X}(t))} c_a \mathrm{E}\left[\mu_a|\mathbf{b}(t)\right]$ as the E$c\mu$ policy. This policy can be viewed as an extension of the traditional $c\mu$ policy (often seen in the literature surrounding control of multi-class queueing systems) for queueing systems with ambiguous parameters.[10] The expectation operator in this policy dynamically combines all the possible $c\mu$ values for each class based on the belief at time t. In the following theorem, we show that the E$c\mu$ policy is asymptotically optimal for the non-robust problem as the observation rate increases.

Theorem 1 **(E$c\mu$ Asymptotically Optimality)**. *The* E$c\mu$ *policy* $\pi^{c\mu}$ *is asymptotically optimal for the non-robust problem:* $\lim_{\psi\to\infty} \mathrm{V}^{\pi^{c\mu}}(\mathbf{X},\mathbf{b}) = \lim_{\psi\to\infty} \mathrm{V}(\mathbf{X},\mathbf{b})$ *for all* $\mathbf{X} \in \mathbb{Z}_+^n$ *and* $\mathbf{b} \in \mathcal{B}$.

Theorem 1 is surprising in its simplicity since problems based on POMDP formulations typically do not yield closed-form results. In contrast to the usual complexities, the asymptotic optimality of the E$c\mu$ policy implies that the only information necessary to make decisions is the expected transition rates among non-empty queues. Therefore, queue lengths are essentially irrelevant to the decision-maker. Rather, the E$c\mu$ policy features a momentum property; if the current action a prescribed by the policy yields enough successes so that $c_a\mathrm{E}[\mu_a|\mathbf{b}]$ does not fall below the threshold defined by $c_{\hat{a}}\mathrm{E}[\mu_{\hat{a}}|\mathbf{b}]$ of the next highest available class $\hat{a}$, the E$c\mu$ policy will continue to serve class a regardless of the state of other classes. In turn, this means that the policy will not attempt to serve a class with smaller $c_{\hat{a}}\mathrm{E}[\mu_{\hat{a}}|\mathbf{b}]$ until other classes with larger values have experienced a sufficient number of service failures, or have cleared their queue. This property may run counter-intuitive to the exploration-minded individual; even if a class has the potential to be endowed with a very large $c_a\mu_{a,j}$ value (under the realization of system parameters), this potential is only rated on the basis of its contribution to the expected service rate.

Another important property of the E$c\mu$ policy is that under mild conditions, $\mathrm{V}^{\pi^{c\mu}}(\mathbf{X},\mathbf{b})$ is piecewise-linear over the belief space (excluding beliefs near edges and faces of $\mathcal{B}$).[11]

[10] Argon and Ziya (2009) demonstrate the optimality of a similar policy in an average-cost non-learning queueing environment when service rates are known, but customer class is not fully observed.

[11] An infinite horizon POMDP value function is not always guaranteed to be piecewise-linear (see, e.g. White and Harrington (1980)).

Proposition 3 **(Piecewise-Linearity of the Approximate Non-Robust Value Function)**. *Let $\mathcal{B}'$ be any closed subset of $\mathcal{B}$ such that for any $\mathbf{b} \in \mathcal{B}', b_{i,j} > 0$ for all $i \in \mathcal{N}, j \in \mathcal{J}_i$. If $\min_{j \in \mathcal{J}_i} c_i \mu_{i,j} \neq \min_{j \in \mathcal{J}_k} c_k \mu_{k,j}$ for any distinct pair $i, k \in \mathcal{N}$, then $\mathrm{V}^{\pi^{c\mu}}(\mathbf{X}, \mathbf{b})$ is piecewise-linear on $\mathcal{B}'$.*

This result is related to two facts: (*i*) for any given initial prior $\mathbf{b} \in \mathcal{B}'$ (and $\mathbf{X} \in \mathbb{Z}_+^n$), the E$c\mu$ policy is unique, unless $\mathbf{b}$ lies on the break-points of the piecewise-linear function $\mathrm{V}^{\pi^{c\mu}}(\mathbf{X}, \mathbf{b})$ (see Lemma 7 and 3 in Online Appendix B), and (*ii*) policies can be evaluated as linear functions of the belief in any POMDP. Therefore, with respect to closed, non-zero portions of the belief space, the value function $\mathrm{V}^{\pi^{c\mu}}(\mathbf{X}, \mathbf{b})$ is differentiable (except at breakpoints). As we will show in Theorem 2, the differentiability of the value function strongly enhances the relationship between optimal policies of the non-robust problem and those under the robust percentile optimization program (5)-(6). Thus, in identifying an asymptotically optimal policy that exhibits this property enables us to solve the robust percentile optimization program in an efficient way. This is an important insight to our search for robust chance-constrained policies especially since, as Zhang (2010) states, there are no known conditions over which a POMDP value function is differentiable on its entire belief space.

To the purpose of finding robust chance-constrained policies, we introduce the following set of policies. Fix the initial $\mathbf{X}$, and let $\mathcal{K}_{\mathbf{b}} = \left\{\pi_{\mathbf{b}}^1, \pi_{\mathbf{b}}^2, \ldots, \pi_{\mathbf{b}}^k\right\}$ be any finite set of optimal policies to the non-robust problem when the initial prior is $\mathbf{b}$, and $\mathbf{p} = (p_1, p_2, \ldots, p_k)$ be an associated distribution such that $\sum_{i=1}^k p_i = 1$. We define a policy $\pi_{\mathcal{K}_{\mathbf{b}}}^{\mathbf{p}}$ to be a *randomized policy*, if at time 0, an element of $\mathcal{K}_{\mathbf{b}}$, $\pi_{\mathbf{b}}^i$, is chosen with probability p_i, which will dictate all current and future decisions.[12]

Interestingly, we find that there exists a randomized policy that forms an optimal solution to the robust percentile problem. This means that there exists an optimal robust policy that randomizes between optimal non-robust policies obtained for a single belief point $\mathbf{b} \in \mathcal{B}$. Furthermore, we shed light on conditions (associated with the differentiability of $\mathrm{V}(\mathbf{X}, \mathbf{b})$ with respect to the belief space) such that a *deterministic* non-robust policy is optimal even for the robust percentile problem.

Theorem 2 **(Chance-Constrained Policy)**. *For any given $\epsilon \geq 0$, there exists a $\mathbf{b}^* \in \mathcal{B}$ and a distribution $\mathbf{p}^*$ forming a randomized policy $\pi_{\mathcal{K}_{\mathbf{b}^*}}^{\mathbf{p}^*}$ that is optimal under the percentile optimization program* (5)-(6)[13]*: $\mathrm{Y}^{\pi_{\mathcal{K}_{\mathbf{b}^*}}^{\mathbf{p}^*}}(\mathbf{X}, \epsilon) = \mathrm{Y}(\mathbf{X}, \epsilon) = \mathrm{V}(\mathbf{X}, \mathbf{b}^*)$. Furthermore, if $\mathrm{V}^{\pi_{\mathbf{b}}}(\mathbf{X}, \mathbf{b})$ is differentiable at $\mathbf{b}^*$, then $\mathcal{K}_{\mathbf{b}^*}$ consists of a single policy, and hence, $\pi_{\mathcal{K}_{\mathbf{b}^*}}^{\mathbf{p}^*}$ is deterministic.*

The above result significantly reduces the complexity of the search for optimal robust policies. Importantly, it implies that we can combine policies associated with the function $\mathrm{V}(\mathbf{X}, \mathbf{b}^*)$ to find

[12] For these randomized policies, we disallow policies that are not picked at time zero for the purpose of targeting specific contours of the value function.

[13] For notational convenience, we suppress the dependency of $\mathbf{p}^*$ and $\mathbf{b}^*$ on ϵ.

chance-constrained policies. In this way, we no longer need to look at the general space of policies, but rather can focus on the class of non-robust optimal policies. Moreover, Proposition 3 shows that the differentiability condition of Theorem 2 can be met by a surface that converges to the value function. If $\mathbf{b}^*$ lies on a linear segment of the value function that is not a breakpoint, $\mathcal{K}_{\mathbf{b}^*}$ can be composed of a single policy yielding a deterministic chance-constrained policy. Hence, under this assumption, one need not be concerned with finding $\mathbf{p}^*$.

However, Theorem 2 leaves us with an important question: what belief, $\mathbf{b}^*$, should be used to form the chance-constrained policy $\pi^{\mathbf{p}^*}_{\mathcal{K}_{\mathbf{b}^*}}$ for a given percentile problem? If such a $\mathbf{b}^*$ is characterized, then the solution to the percentile problem can easily be found by a randomization of non-robust policies associated with $\mathbf{b}^*$. The answer to this question turns out to be closely related to the geometrical concept of the *convex floating body* first discussed by Dupin (1822). In particular, we utilize the notion of the convex floating body studied by Schutt and Werner (1990) to characterize $\mathbf{b}^*$.

Definition 1 **(Convex Floating Body)**. *Let* $\mathcal{W}_\epsilon = \{(\mathbf{w}, w) \in \mathbb{R}^m \times \mathbb{R} : \mathbb{P}_{\mathbf{B}}(\mathbf{B}\mathbf{w}^{\mathrm{T}} \geq w) \leq \epsilon\}$ *be the set of all half spaces that "cut off"* ϵ *or less volume of the belief space* $\mathcal{B}$ *with respect to* $\mathbb{P}_{\mathbf{B}}$. *An* ϵ*-based convex floating body on* $\mathcal{B}$ *is* $\mathcal{L}_\epsilon = \bigcap_{\{\mathbf{w},w\}\in\mathcal{W}_\epsilon} \{\mathbf{b} \in \mathcal{B} : \mathbf{b}\mathbf{w}^{\mathrm{T}} \leq w\}$. *We let* $\delta\mathcal{L}_\epsilon$ *be the boundary of* $\mathcal{L}_\epsilon$[14].

Based on the above definition, a convex floating body is the region left from hyperplanes "cutting off" a specified volume (ϵ) from an object. For every $\mathbf{b} \in \delta\mathcal{L}_\epsilon$, there exists a hyperplane that divides $\mathcal{B}$ into two pieces, one which has volume less than or equal to ϵ. Figure 2 illustrates the convex floating body of a sphere with uniform density, which is either the empty set or another sphere. We study convex floating bodies with respect to the density measure $\mathbb{P}_{\mathbf{B}}$ on the belief space of our priors to characterize $\mathbf{b}^*$, and thereby find optimal chance-constrained policies as discussed in Theorem 2.

For the purposes of characterizing $\mathbf{b}^*$, it is important that $\mathcal{L}_\epsilon$ is non-empty. Fortunately, Fresen (2013) states that when $\mathbb{P}_{\mathbf{B}}$ is a log-concave probability distribution, $\mathcal{L}_\epsilon$ exists so long as $\epsilon \leq e^{-1}$. Hence, for many robust applications which tend toward pessimism (where ϵ is small), under common distributions, the convex floating body is guaranteed to exist. If $\mathcal{L}_\epsilon$ is nonempty, we find that $\mathbf{b}^*$ (defined in Theorem 2) is found at the largest value of the non-robust problem on the boundary of the convex floating body.

Proposition 4 **(Characterizing $\mathcal{K}_{\mathbf{b}^*}$)**. *If* $\mathcal{L}_\epsilon$ *is nonempty, then* $\mathbf{b}^* = \operatorname{argmax}_{\mathbf{b}\in\delta\mathcal{L}_\epsilon} \mathrm{V}(\mathbf{X}, \mathbf{b})$, *where* $\mathbf{b}^*$ *satisfies* $\mathrm{Y}(\mathbf{X}, \epsilon) = \mathrm{V}(\mathbf{X}, \mathbf{b}^*)$.

Interestingly, Proposition 4 relates percentile optimization to a minimax objective: one can search for a *worst-case* belief within a specified set. Since $\mathrm{V}(\mathbf{X}, \mathbf{b})$ is concave in $\mathbf{b}$ (by the convexity results

[14] We note that if $\mathcal{L}_\epsilon$ is nonempty, $\delta\mathcal{L}_\epsilon$ always exists since closed, convex, and compact sets are equal to the convex hull of their boundary.

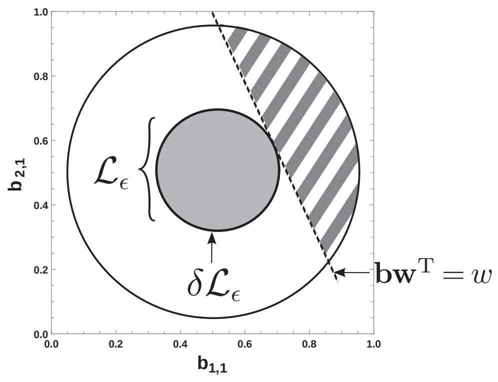

Figure 2 A convex floating body $\mathcal{L}_\epsilon$ when $\mathbb{P}_{\mathbf{B}}$ has uniform density within the circle and is zero elsewhere. It is generated from the intersection of halfspaces $(\mathbf{w}, w) \in \mathcal{W}_\epsilon$, and the striped area must contain less than or equal to ϵ volume. $(n = 2, m_1, m_2 = 2)$

of Sondik (1971) and Smallwood and Sondik (1973)), if $\delta\mathcal{L}_\epsilon$ is easily characterized, we can apply gradient-based optimization to solve the problem rather than evaluating the entire surface which is computationally intractable. Although Theorem 2 states that $\mathcal{K}_{\mathbf{b}^*}$ is a singleton when the value function is differentiable at $\mathbf{b}^*$, the differentiability is not always guaranteed. To this end, in the proof of Proposition 4 (see Online Appendix B), we characterize $\mathbf{p}^*$. We find that the distribution $\mathbf{p}^*$ such that the contour $\{\mathbf{b} \in \mathcal{B} | \mathrm{V}^{\pi^{\mathbf{p}^*}_{\mathcal{K}_{\mathbf{b}^*}}}(\mathbf{X}, \mathbf{b}) = \mathrm{V}(\mathbf{X}, \mathbf{b}^*)\}$ is a subgradient hyperplane to $\mathcal{L}_\epsilon$.

In general, since non-robust policies are only partially characterized (they converge to E$c\mu$ policies asymptotically), it is important to connect the E$c\mu$ policies to the percentile optimization objective. The following corollary is similar to Proposition 4 and shows that there exists a finite randomization of E$c\mu$ policies that are asymptotically optimal as $\psi \to \infty$ to the percentile objective.

Corollary 1 **(Robust** E$c\mu$ **Optimality)**. *If $\mathcal{L}_\epsilon$ is nonempty, then there exists a policy π that is a finite randomization of* E$c\mu$ *policies such that* $\mathrm{Y}^\pi(\mathbf{X}, \epsilon) - \mathrm{Y}(\mathbf{X}, \epsilon) \le \mathrm{V}^{\pi^{c\mu}}(\mathbf{X}, \hat{\mathbf{b}}) - \mathrm{V}(\mathbf{X}, \mathbf{b}^*)$, *where* $\hat{\mathbf{b}} = \arg\max_{\mathbf{b} \in \delta\mathcal{L}_\epsilon} \mathrm{V}^{\pi^{c\mu}}(\mathbf{X}, \mathbf{b})$ *and* $\mathbf{b}^*$ *is defined in Theorem 2.*

This corollary holds despite the fact that $\mathrm{V}^{\pi^{c\mu}}(\mathbf{X}, \mathbf{b})$ is not guaranteed to be concave in $\mathbf{b}$. In fact, if it is concave in $\mathbf{b}$, the randomized policy π can be directly built from non-robust policies. However, if $\mathrm{V}^{\pi^{c\mu}}(\mathbf{X}, \mathbf{b})$ is not concave in $\mathbf{b}$, we can still form the appropriate randomized policy satisfying Corollary 1 via a randomization of policies that satisfy minimax solutions within the set of E$c\mu$ policies on the boundary of the convex floating body, namely $\min_{\mathbf{b}^1 \in \mathcal{B}} \max_{\mathbf{b}^2 \in \delta\mathcal{L}_\epsilon} \mathrm{V}^{\pi^{c\mu}_{\mathbf{b}^1}}(\mathbf{X}, \mathbf{b}^2)$.

With respect to optimal solutions to the percentile objective, additional results can further confine $\mathcal{K}_{\mathbf{b}^*}$ (of Theorem 2) by noting that $\mathbf{b}^*$ must lie near the extreme belief state with worst-case transition parameters. We denote this "worst-case" belief state by $\mathbf{b}_0$, and note that it is composed of components

$$b^0_{i,j} = \begin{cases} 1: & \text{if } \mu_{i,j} = \min_{k \in \mathcal{J}_i} \mu_{i,k}, \\ 0: & \text{otherwise.} \end{cases} \tag{7}$$

It can be shown (see the proof of Proposition 5) that for any prior-flexible policy, $\mathbf{b}_0$ is the worst-case (most expensive) belief state for the system. To further characterize $\mathbf{b}^*$, we define the concept of *visibility* (seen in geometry literature but repurposed for our needs).

Definition 2 **(Visibility)**. *A belief point* $\mathbf{b} \in \mathcal{L}_\epsilon$ *is said to be visible from a reference belief* $\mathbf{b}^1 \in \mathcal{B}$ *if* $\{\mathbf{b}^2 \in \mathcal{B} : \mathbf{b}^2 = \eta \mathbf{b} + (1-\eta)\mathbf{b}^1, \eta \in [0,1]\} \bigcap \mathcal{L}_\epsilon = \mathbf{b}$.

As demonstrated in Figure 3, a belief $\mathbf{b}$ in the convex floating body is visible from a reference belief $\mathbf{b}^1$ if, on the line segment connecting these points, only $\mathbf{b}$ lies within the convex floating body. This implies that if the reference belief point $\mathbf{b}^1$ is distinct from $\mathbf{b}$, and $\mathbf{b}$ is visible from $\mathbf{b}^1$, then $\mathbf{b}$ must lie in the boundary ($\mathbf{b} \in \delta\mathcal{L}_\epsilon$). However, not every point on $\delta\mathcal{L}_\epsilon$ is visible from a reference point $\mathbf{b}^1$. In the following Proposition, we show that the belief $\mathbf{b}^*$ (introduced in Theorem 2) must be visible from the worst-case belief state $\mathbf{b}_0$.

Proposition 5 **(Visibility of $\mathbf{b}^*$)**. *If* $\mathcal{L}_\epsilon$ *is nonempty, then there exists a* $\mathbf{b}^*$ *visible from the worst-case belief* $\mathbf{b}_0$.

Proposition 5 significantly helps us find $\mathbf{b}^*$ (of Theorem 2): we only need to search part of $\delta\mathcal{L}_\epsilon$ which is visible from $\mathbf{b}_0$. Proposition 5 also can facilitate establishing effective heuristics which circumvent the calculation of the non-robust problem. For instance, Figure 4 demonstrates the implications of Proposition 5 for a uniform type $\mathbb{P}_\mathbf{B}$: $\mathbf{b}^*$ lies somewhere on the dashed line.

The set $\mathcal{L}_\epsilon$ (and hence $\delta\mathcal{L}_\epsilon$) typically needs to be estimated by a polytope since most distributions result in convex floating bodies with no easy closed-form representation. However, upper and lower bounds to the percentile objective can be found by optimizing over sets (in the sense of Proposition 4) that contain or are contained by $\mathcal{L}_\epsilon$ which converge to $\Upsilon(\mathbf{X}, \epsilon)$ as the sets converge to $\mathcal{L}_\epsilon$. The details of this are expressed in the proof of Lemma 9 in Online Appendix B. Additionally, with certain $\mathbb{P}_\mathbf{B}$, the problem of estimating $\mathcal{L}_\epsilon$ may be altogether circumvented. This is specifically the case when $\mathbb{P}_\mathbf{B}$ has the form of a spherical-type distribution defined below.

Definition 3 **(Spherical Distribution)**. *We say* $\mathbb{P}_\mathbf{B}$ *is a spherical distribution centered at* $\mathbf{b}_1$*, if, for any* $\epsilon \in \mathbb{R}^+$, $\mathcal{L}_\epsilon = \{\mathbf{b}_2 \in \mathcal{B} : \|\mathbf{b}_2 - \mathbf{b}_1\| \leq d\}$ *for some* $d \in \mathbb{R}^+$*, where* $\|\cdot\|$ *is the* l^2*-norm.*

Remark 1. In cases with spherical distributions, searching for $\mathbf{b}^*$ is simplified even in large dimensional spaces, since we have the expression for $\delta\mathcal{L}_\epsilon$ and bounds based on the visibility from $\mathbf{b}_0$. Thus, the problem is reduced to searching for the maximum of a concave function on a sphere.

Another distribution that features an easily expressible convex floating body is a special case when $\mathbb{P}_\mathbf{B}$ is uniform. If $n = 2, m_1 = 2, m_2 = 2$, and $\mathbb{P}_\mathbf{B}$ is uniform, a small modification of a result by

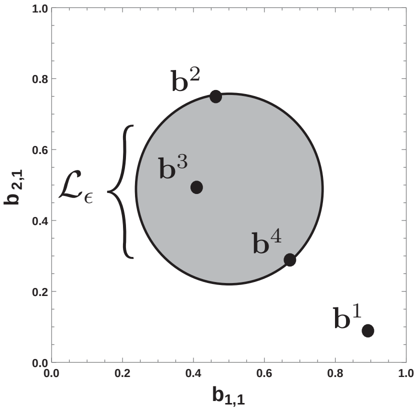

Figure 3 Belief points $\mathbf{b}^2$ and $\mathbf{b}^3$ are *not* visible from reference belief $\mathbf{b}^1$, whereas $\mathbf{b}^4$ is visible from reference belief $\mathbf{b}^1$ $(n=2, m_1, m_2=2)$.

Calgar (2010) shows that $\delta\mathcal{L}_\epsilon$ is given by a curve defined in four quadrants as:

$$b_{2,1} = \begin{cases} \frac{b_{1,1}-1+0.5\epsilon}{b_{1,1}-1} & : 0.5 \leq b_{1,1} \leq 1-\epsilon,\ 0.5 \leq b_{2,1} \leq 1-\epsilon \\ \frac{b_{1,1}-0.5\epsilon}{b_{1,1}} & : \epsilon \leq b_{1,1} \leq 0.5, \quad\ \ 0.5 \leq b_{2,1} \leq 1-\epsilon \\ \frac{0.5\epsilon}{b_{1,1}} & : \epsilon \leq b_{1,1} \leq 0.5, \quad\ \ \epsilon \leq b_{2,1} \leq 0.5 \\ -\frac{0.5\epsilon}{b_{1,1}-1} & : 0.5 \leq b_{1,1} \leq 1-\epsilon,\ \epsilon \leq b_{2,1} \leq 0.5 \end{cases}$$

for $0 < \epsilon < 0.5$ as shown in Figure 4. If we evaluate $\mathrm{V}(\mathbf{X}, \mathbf{b})$ along the quadrant visible to $\mathbf{b}_0$, belief $\mathbf{b}^*$ is revealed as the maximum on this curve. We use this uniform case and various spherical cases in Section 7 to show that since our approach includes learning, it provides robustness to the specification of $\mathbb{P}_\mathbf{B}$. Therefore, even though exact closed-form representations of $\mathcal{L}_\epsilon$ in general are rare, polytope or spherical approximations are sufficient for the purposes of percentile optimization in our framework.

5. Asymptotically Tight Bounds

Although we have characterized the optimal policies of the non-robust and percentile problems, evaluating the non-robust value function $\mathrm{V}(\mathbf{X}, \mathbf{b})$ is still a computationally complex problem (see, e.g., Littman et al. (1998), Mundhenk et al. (2000), and Papadimitriou and Tsitsiklis (1987) for an in-depth discussion regarding the complexity of POMDP programs). If the value function $\mathrm{V}(\mathbf{X}, \mathbf{b})$ and the convex floating body's boundary $\delta\mathcal{L}_\epsilon$ are known, the solution to the percentile optimization is easily characterizable (Theorem 2, Proposition 4, and Proposition 5). Therefore, we provide computationally tractable bounds to the non-robust problem that can be evaluated in closed-form to facilitate the computability of chance-constrained policies.

The bounds we form are based on the performance of (1) queues under no model ambiguity with fixed rate parameters equal to $\mathrm{E}[\mu_i|\mathbf{b}]$, and (2) following a particular server allocation priority rule based on the initial parameter belief. These imply that our bounds rely only on the valuation of fixed priority-based policies that do not change with dynamic observations, significantly reducing the computational complexity of the problem.

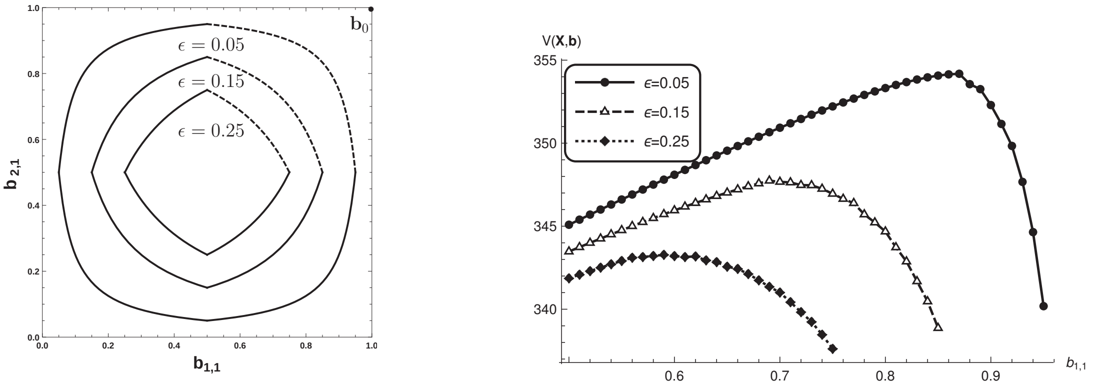

Figure 4 On the left, convex floating bodies $\mathcal{L}_\epsilon$ for $\epsilon = 0.05, 0.15, 0.25$ with $n = 2, m_1, m_2 = 2$, and uniform $\mathbb{P}_\mathbf{B}$. To be visible from $\mathbf{b}_0$, belief $\mathbf{b}^*$ associated with Proposition 4 must lie on the dashed lines assuming $\mu_{1,1} < \mu_{1,2}$ and $\mu_{2,1} < \mu_{2,2}$ (Proposition 5). On the right, $\mathrm{V}((10,10), \mathbf{b})$ is evaluated on these boundaries when $\mu_{1,1} = 0.1, \mu_{1,2} = 0.2, \mu_{2,1} = 0.05, \mu_{2,2} = 0.25$. Belief $\mathbf{b}^*$ lies at the peak of these curves.

For a given belief $\hat{\mathbf{b}} \in \mathcal{B}$, consider a counterpart system identical to our original setting with the exception of the ambiguity sets being $\hat{\mathcal{M}}_i = \left\{\mathrm{E}[\mu_i|\hat{\mathbf{b}}]\right\}$ (analogous to the original ambiguity sets $\mathcal{M}_i$). That is, the counterpart queueing system has fully known service rates that are calculated based on taking an expectation of service rates in $\mathcal{M}_i$ over belief $\hat{\mathbf{b}}$. Obviously, the optimal policy for this system is the traditional $c\mu$ rule, since all of its parameters are fully known. Let $\pi_{\hat{\mathbf{b}}}$ denote this $c\mu$ rule and $\bar{\mathrm{V}}^{\pi_{\hat{\mathbf{b}}}}(\mathbf{X}, \hat{\mathbf{b}})$ be the associated infinite-horizon cost of the counterpart system under $\pi_{\hat{\mathbf{b}}}$. It is important to emphasize that $\pi_{\hat{\mathbf{b}}}$ exhaustively serves class $\arg\max_{a \in \mathcal{A}(\mathbf{X})} c_a \mathrm{E}[\mu_a|\hat{\mathbf{b}}]$ until no customer of that class remains in the system, and acts only as a function of the queue state, not of belief, even when $\pi_{\hat{\mathbf{b}}}$ is implemented in the original system. When $\pi_{\hat{\mathbf{b}}}$ is implemented in the original system, we denote the infinite-horizon cost by $\mathrm{V}^{\pi_{\hat{\mathbf{b}}}}(\mathbf{X}, \hat{\mathbf{b}})$. Using the counterpart system's cost and its associated policy, we can bound the non-robust cost (which is needed to calculate the robust cost; see Theorem 2 and Proposition 4) using the following proposition.

Proposition 6 **(Asymptotically Tight Bounds)**. *For any state* $(\mathbf{X}, \hat{\mathbf{b}})$*, the non-robust cost* $\mathrm{V}(\mathbf{X}, \hat{\mathbf{b}})$ *is bounded as* $\bar{\mathrm{V}}^{\pi_{\hat{\mathbf{b}}}}(\mathbf{X}, \hat{\mathbf{b}}) \leq \mathrm{V}(\mathbf{X}, \hat{\mathbf{b}}) \leq \mathrm{V}^{\pi_{\hat{\mathbf{b}}}}(\mathbf{X}, \hat{\mathbf{b}})$*. Furthermore:*

(i) The gap between the upper and lower bound costs decrease to zero as queue length X_i *increases to infinity, where* $i = \arg\max_{a \in \mathcal{A}(\mathbf{X})} c_a \mathrm{E}[\mu_a|\hat{\mathbf{b}}]$.

(ii) The gap between the upper and lower bound costs monotonically decrease to zero as $\mathrm{Var}[\mu_i|\hat{\mathbf{b}}]$ *decrease to zero (for all* $i \in \mathcal{N}$*).*

Both the upper and lower bounds of Proposition 6 are easily calculable (see Online Appendix B). Furthermore, under the conditions above, these bounds become arbitrarily close approximations, which adds computational tractability to the problem as well as analytical insight to the relationship between our non-robust and traditional $c\mu$ policies. In particular, part *(ii)* of Proposition 6 supports the intuition that gathering more data on unknown service parameters can provide more accurate bound information. Part (*i*) of Proposition 6 provides conditions under which the myopic, non-learning policy's cost converges to that of the optimal policy.

Remark 2. Since the percentile objective relies on the computation of the non-robust problem, the bound results can be easily applied to the percentile formulation as well. For instance, one can refine the search for $\text{argmax}_{\mathbf{b}\in\delta\mathcal{L}_\epsilon} \mathrm{V}(\mathbf{X},\mathbf{b})$ as in Proposition 4: if the upper bound for a $\mathbf{b}\in\delta\mathcal{L}_\epsilon$ is less than the lower bound for $\mathbf{b}'\in\delta\mathcal{L}_\epsilon$, $\mathbf{b}$ must not be the belief point $\mathbf{b}^*$. Since most infinite-horizon POMDPs are calculated by finite-horizon approximations, a second application of the bounds is to use them as the terminal cost used in the finite-horizon dynamic program. That is, when evaluating the finite-horizon approximation, one can replace $\mathrm{V}_0(\mathbf{X},\mathbf{b})$ by lower and upper bounds $\bar{\mathrm{V}}^{\pi_{\hat{\mathbf{b}}}}(\mathbf{X},\hat{\mathbf{b}})$ and $\mathrm{V}^{\pi_{\mathbf{b}}}(\mathbf{X},\mathbf{b})$, respectively. This can provide very tight bounds on the POMDP, since after a certain number of "learning periods," where the POMDP is explicitly evaluated, the controller might have collected enough information to have enough confidence in the true transition parameters.

6. An Analytically-Rooted Heuristic Policy

Chance-constrained policies are inherently difficult to calculate, even given the analytical results established in the previous section. To circumvent complexity arising from (1) the PSPACE-hard problem of evaluating a POMDP over a belief space with high dimensionality, and (2) finding the shape of the convex floating body which requires high-dimensional polytope approximations, we now introduce an effective heuristic policy. This heuristic policy operates by simply choosing the E$c\mu$ policy associated with the belief point on the convex floating body's boundary $\delta\mathcal{L}_\epsilon$ that minimizes the distance from $\mathbf{b}_0$ (the worst-case parameter settings for each class characterized in (7)). This is typically an easy-to-perform task, especially in the cases of uniform and spherical type distributions on the belief space, allowing for managers to benefit from our approach without requiring demanding computations. Moreover, as we will show in Section 7, this heuristic performs extremely well both on randomly generated data and on real-world data that we have collected from a leading U.S. hospital.

We term the E$c\mu$ policy with expectation taken based on belief point $\arg\min_{\mathbf{b}\in\delta\mathcal{L}_\epsilon}\|\mathbf{b}_0-\mathbf{b}\|$, where $\|\cdot\|$ is the l^2-norm, as the $(1-\epsilon)$% E$c\mu$ *heuristic policy*. This heuristic policy takes advantage of three main structural results of the chance-constrained policy (that we established in the previous section), while providing a much simpler version of it:

(1) It assumes that the true optimal policies of the non-robust problem are E$c\mu$, a fact supported by Theorem 1 which shows the asymptotic relationship of the optimal policies to E$c\mu$.

(2) It locates belief $\arg\min_{\mathbf{b}\in\delta\mathcal{L}_\epsilon}\|\mathbf{b}_0-\mathbf{b}\|$ to be near $\mathbf{b}^*$ (of Theorem 2) based on Proposition 4. The worst-case (most expensive) belief state is $\mathbf{b}_0$, and through the proof of Proposition 5 (see Online

Appendix B) the value function is non-increasing in λ with respect to belief $\lambda\mathbf{b}+(1-\lambda)\mathbf{b}_0$ for $\lambda\in[0,1]$. Thus, $\arg\max_{\mathbf{b}\in\delta\mathcal{L}_\epsilon} \mathrm{V}(\mathbf{X},\mathbf{b})$ is expected to be near $\mathbf{b}_0$.[15]

(3) It takes advantage of the fact that $\arg\min_{\mathbf{b}\in\delta\mathcal{L}_\epsilon} \|\mathbf{b}_0-\mathbf{b}\|$ satisfies Proposition 5 (since this belief is visible from $\mathbf{b}_0$).

7. Numerical Experiments

We now perform various numerical experiments in order to (1) identify the advantages of chance-constrained policies in a variety of environments under model ambiguity, (2) demonstrate the sensitivities of the underlying queueing models, (3) study the effectiveness of the proposed E$c\mu$ heuristic in mimicking the optimal chance-constrained policies, and (4) demonstrate the implications of our results in real-world applications. To pursue these goals, we present our analyses in four parts: we (1) investigate how our policies perform over a large parameter suite but in a relatively small queueing system, (2) evaluate our proposed heuristic alongside percentile, minimax, and minimin policies in a larger system, (3) demonstrate the gap between the E$c\mu$ and optimal (non-robust) policies, and (4) apply the E$c\mu$ heuristic to a hospital Emergency Department (ED) setting using real-world data, and discuss its significant implications on improving the current patient flow policies.

To better understand the relative performance of our robust percentile policies, we start by considering a large parameter suite including over 1,000 parameter settings in an $n=2, m_1=2$, and $m_2=2$ setting with four different $\mathbb{P}_\mathbf{B}$ distributions at their 95% chance-constrained policy. We name these $\mathbb{P}_\mathbf{B}$ distributions f_1, f_2, f_3 and f_4 respectively: f_1, f_2, and f_3 are truncated multivariate normal distributions with means $\boldsymbol{\mu}_1=(0.5,0.5)$, $\boldsymbol{\mu}_2=(0.4,0.4)$, $\boldsymbol{\mu}_3=(0.6,0.6)$ and covariance matrices $\boldsymbol{\Sigma}_1=\left(\begin{smallmatrix}1.5 & 0.0\\ 0.0 & 1.5\end{smallmatrix}\right)$, $\boldsymbol{\Sigma}_2=\left(\begin{smallmatrix}0.5 & 0.0\\ 0.0 & 0.5\end{smallmatrix}\right)$, and $\boldsymbol{\Sigma}_3=\left(\begin{smallmatrix}0.5 & 0.0\\ 0.0 & 0.5\end{smallmatrix}\right)$ respectively. Finally, f_4 is the uniform distribution.

We include two non-learning robust policies (minimin and minimax) as benchmarks for the performance of our robust percentile policies and compare the policies by evaluating their total cost when each model (i.e., parameter configuration) is equally likely. That is, we assume that the true (but unknown) prior of our system is $\bar{\mathbf{b}}=(0.5,0.5,0.5,0.5)$, and we evaluate the total cost under 95% chance-constrained, minimax, and minimin policies. Furthermore, we assume $c_1=c_2$. In every problem instance, we assume $\mu_{2,1}<\mu_{1,1}$ and $\mu_{1,2}<\mu_{2,2}$ so that the policy is not uniform throughout the belief space, which provides incentive for gaining additional knowledge. Further detail on this parameter suite is presented in Online Appendix A.1.

We start by investigating whether or not our robust percentile policies overcome the deficiencies we observe for non-robust policies regarding their sensitivity to the selection of initial prior (Observations 8, 9, and 10 established in Online Appendix A.2). Our results presented in Online

[15] This does not imply that $\arg\max_{\mathbf{b}\in\delta\mathcal{L}_\epsilon} \mathrm{V}(\mathbf{X},\mathbf{b}) = \arg\min_{\mathbf{b}\in\delta\mathcal{L}_\epsilon} \|\mathbf{b}_0-\mathbf{b}\|$. $\mathrm{V}(\mathbf{X},\mathbf{b})$ is only assured to be non-increasing on line segments connected to $\mathbf{b}_0$.

Appendix A.5 provide a positive response, showing that chance-constrained policies overcome these issues. We next compare our proposed policies with other non-learning robust policies (minimax and minimin). In Table 1, we present the results of this comparison expressed by the average (among all models) optimality gap percentage under various policies. The optimality gap percentage for policy π at $\mathbf{b}$ is defined as

$$\frac{\mathrm{V}^{\pi}(\mathbf{X},\mathbf{b})-\mathrm{V}(\mathbf{X},\mathbf{b})}{\mathrm{V}(\mathbf{X},\mathbf{b})}\%.$$

From Table 1, we observe that on average, our proposed chance-constrained policies perform much better than the other non-learning policies. Since there is equal chance of every parameter configuration, non-learning policies serve the wrong class for a realized set of parameters 50% of the time, which results in poor performance.

Comparing the chance-constraint policies under f_1, f_2, f_3, and f_4 in Table 1 reveals yet another interesting insight: they exhibit similar performance. The reason behind this is three-fold: (1) as a property of Proposition 5, since we used 95% chance-constrained policies, each $\mathbf{b}^*$ tends to be near $\mathbf{b}_0$, (2) even though the distributions f_1, f_2, f_3, and f_4 are different (e.g., they have differing covariance structures and are centered at different beliefs), their convex floating bodies are quite similar, and (3) the chance-constrained policies we propose exhibit learning. Hence, we can make the following:

Observation 1 **(Sensitivity)**. *The performance of chance-constrained policies is not sensitive to the choice of* $\mathbb{P}_{\mathbf{B}}$.

In Section 6, we introduced the E$c\mu$ heuristic as an easy-to-implement policy that mimics the performance of robust optimal chance-constrained policies. To demonstrate the validity of the first assumption underlying this heuristic – that the optimal policies of the non-robust problem are E$c\mu$ – in Figure 5 we depict the percent optimality gap of the E$c\mu$ heuristic policy by comparing its cost to that of the optimal non-robust policies in a situation where ψ is small. Since we know that E$c\mu$ becomes optimal as ψ becomes large (Theorem 1), this poses a "worst-case" scenario for the performance of the E$c\mu$ policies. From Figure 5, we can make the following:

Observation 2 **(Near Optimality of E$c\mu$)**. *Even when ψ is small, the* E$c\mu$ *performance is close to the non-robust optimal policy, especially when the system is highly congested.*

Observation 2 confirms that the myopic E$c\mu$ policy provides us with a good approximation of the optimal POMDP value function (as we would expect given its asymptotic relationship to the chance-constrained policy; see Theorem 1). However, using such a rule to find the explicit surface of the POMDP value function is computationally challenging, even though the E$c\mu$ policy is simple. This is because policy evaluation (even when a policy is known) in POMDPs is PSPACE complete (see, e.g., Mundhenk et al. (2000)). Hence, the ideal task of searching for the max of the convex floating

			Optimality Gap (%)			
X	Minimax	Minimin	95% Chance Constrained f_1	95% Chance Constrained f_2	95% Chance Constrained f_3	95% Chance Constrained f_4
$(2,2)$	3.17	15.51	1.84	2.07	2.2	1.97
$(2,5)$	2.52	13.58	0.85	0.81	0.86	0.86
$(2,10)$	1.35	8.21	0.52	0.51	0.54	0.49
$(5,2)$	4.48	8.73	2.65	2.74	2.33	2.21
$(5,5)$	5.01	10.3	0.85	0.81	0.75	0.79
$(5,10)$	3.37	7.56	0.61	0.58	0.48	0.57
$(10,2)$	4.14	4.05	1.76	1.93	1.32	1.35
$(10,5)$	5.49	5.79	0.53	0.51	0.54	0.55
$(10,10)$	4.34	5.15	0.33	0.33	0.35	0.35
Ave.	3.76	8.76	1.10	1.14	1.04	1.02

Table 1 Performance of various robust policies over the test suite ($n=2, m_1=2, m_2=2$).

body as in Proposition 4, even with the help of Proposition 5, is highly difficult even in moderate problem instances where $n>3$ and $m>6$. Furthermore, often times the shape of $\mathcal{L}_\epsilon$ is difficult to determine explicitly as is the case even in the simple uniform distributions in more than two dimensions, which further complicates our search. Hence, for implementation in real applications, we turn to our robust heuristic policy.

To gain deeper insights into the performance of our heuristic, we simulate systems with $m_1 = m_2 = m_3 = 3$ with uniform $\mathbb{P}_\mathbf{B}$ in the largest inscribed sphere of the belief space. To also evaluate the robustness of our proposed heuristic vis-a-vis the optimal percentile policy as well as minimin and minimax policies, we use CVar(q), which is the average cost within the most costly q% of our simulated runs. Therefore, if $\mathcal{S} = \{s_1, \ldots, s_r\}$ is the set of the costs from a simulation of r runs ordered from most costly to least costly, then

$$\mathrm{CVar}\,(q) = \frac{\sum_{i=1}^{\lceil (1-q)(r-1)+1\rceil} s_i}{\lceil (1-q)(r-1)+1\rceil}.$$

This statistic may roughly be seen as a function that increases in pessimism, since we use fewer low cost data points in the expectation as q increases.[16]

Using a 95% chance-constrained policy, the E$c\mu$ heuristic, minimin, and minimax policies, Figure 6 illustrates performance over 20,000 simulation runs.[17] The leftmost subfigures display the raw CVar values. However, we direct our attention to the rightmost figures, which display the percentage gap (of CVars) between the four selected polices and "best" policy at a given q. From Figure 6, we observe the following:

Observation 3 **(Heuristic Performance)**. *The* E$c\mu$ *heuristic performs nearly identically to the chance-constrained policy, with a diminishing difference as the system becomes more congested.*

We note that percentile optimization is not concerned about the "worst-case" scenarios, and rather optimizes based on a proportion of the belief space. Hence, being a statistic concerned with the

[16] For instance, one would expect the minimax policy to perform well in comparison to other policies at CVar(1).

[17] The associated confidence intervals are tight, so we only show the averages.

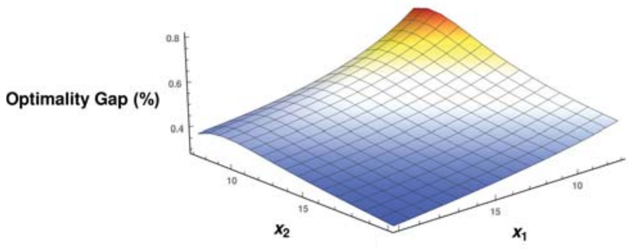

Figure 5 The optimality gap (%) of E$c\mu$ policy when evaluated on the central prior $\bar{\mathbf{b}}$ ($\mu_{1,1}=0.6, \mu_{1,2}=0.7, \mu_{2,1}=0.5, \mu_{2,2}=0.8$).

tail performance of the distribution of costs, CVar (as compared to the expected cost) provides us with a more accurate representation of the value of robustness that percentile optimization offers. Furthermore, Figure 6 demonstrates that the proposed heuristic captures the essence of the chance-constrained policy in that it lies near the optimal policy, mirroring its performance in each simulated run. Overall, our goal to provide an alternative to the over-conservatism and over-optimism of the minimax and minimin policies seems to be met by our percentile optimization technique: it performs well at each level of the CVar statistic. Thus, we make the following:

Observation 4 **(Chance-Constrained vs. Minimin & Minimax)**. *Unlike minimax and minimin policies, chance-constrained policies perform well regardless of the optimism/pessimism level.*

Even in cases where the chance-constrained policy is inferior to other policies with regard to the CVar statistic (e.g., the fourth row of Figure 6 with $\mathbf{X}=(10,10,10)$, where the minimin policy is seen to perform best with regard to CVar(0)), we can see that fixed priority policies (e.g., those obtained under the minimin objective) miss out on the advantages of robustness that the chance-constrained policy offers throughout the optimism spectrum. Furthermore, percentile optimization is flexible: by modifying ϵ, we can change our policy's focus to be more or less optimistic to the point of becoming a minimax and minimin policy itself (Proposition 2). A similar advantage is also gained in the APOMDP framework of Saghafian (2017), where α-maximin expected utility (α-MEU) preferences are used.

7.1. Real-World Application: ED Patient Prioritization

In most hospital Emergency Departments (EDs) in the U.S., patients upon arrival are sorted by means of an urgency-based triage system into one of (typically) five classes known as Emergency Severity Index (ESI) levels. These ESI levels classify patients in descending order of urgency so that a patient of ESI 1, being in dire condition, is immediately treated, whereas patients of levels 4 and 5 are sent to a "fast track" area to be treated. Therefore, the classes served by the main section of the ED (the majority of arrivals) are those with ESI levels 2 and 3 (see, e.g., Saghafian et al. (2012),

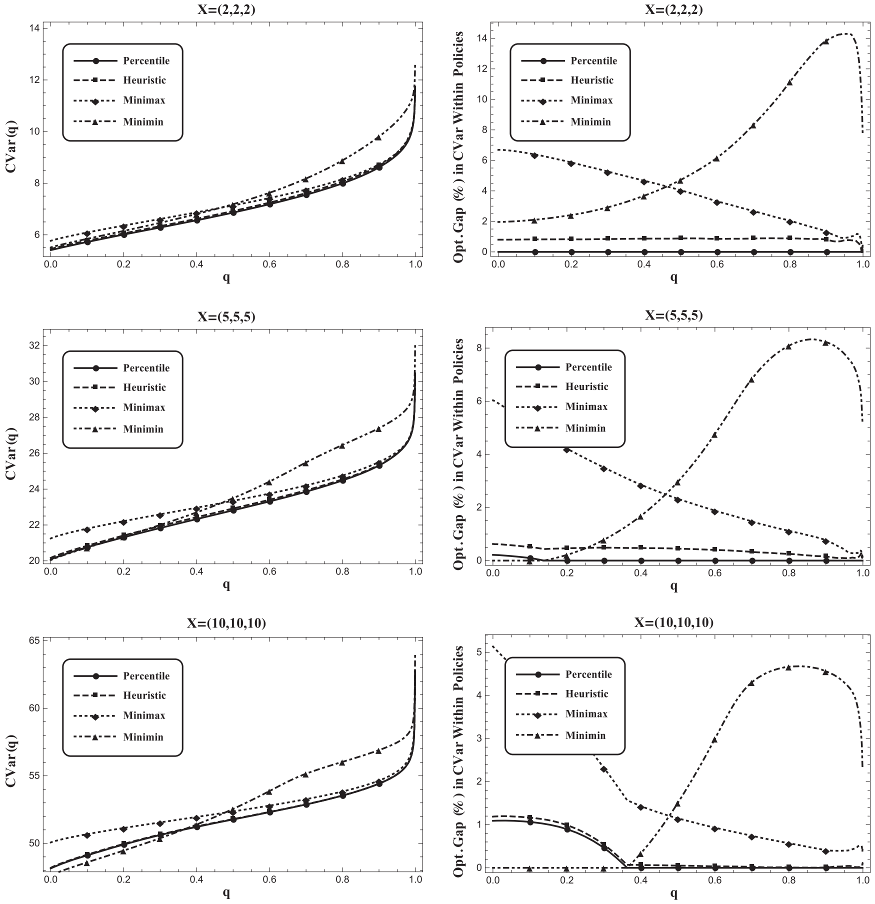

Figure 6 Comparison of policies with respect to CVar (20,000 simulated runs and a uniform $\mathbb{P}_{\mathbf{B}}$ on the largest inscribed sphere of the belief space).

Saghafian et al. (2014), and the references therein). We denote ESI 2 and 3 patients by “Urgent” and “Non-Urgent” patients, respectively.

As patients wait to receive treatment their condition may worsen over time and lead to adverse medical events. Sprivulis et al. (2006) and Plunkett et al. (2011) show that higher patient mortality is associated with longer waiting times prior to seeing a physician. Other research (e.g., an extremely large study on data of nearly 14 million patients by Guttmann et al. (2011)) indicate that the Risk of Adverse Events (ROAE) for patients increases with higher waiting times leading to higher mortality and hospital admission rates. Therefore, with the objective of increasing patient safety, we consider the goal of minimizing average ROAE for ED patients, and investigate optimal prioritization policies. To do so, we assume adverse events occur based on a Poisson process with a higher rate for urgent patients, and note that ROAEs in this setting play the role of holding cost parameters in our multi-class queueing model introduced earlier. The same approach is used in Saghafian et al. (2014), where the benefits of further stratifying these levels in terms of a patient’s *complexity* is discussed. Simple patients are those that experience only a single interaction with the physician, and thus are more quickly treated by the ED than complex patients, whose treatment necessitates several interactions with the physician interspersed with various tests (CT scans, MRI, etc.).

Figure 7 (left) illustrates a schematic flow of patients as a multi-class queueing system. To analyze the multi-class queueing system of Figure 7 (right) in a traditional way, one needs to obtain point estimates of various parameters (e.g., service/treatment rates for each class), a task which is subject to inevitable errors.[18] Furthermore, triaged urgency and complexity levels are subject to misclassifications, which further confuses the true parameter settings of the system. Although misclassifications can be included in the analysis when all of the parameters of the system are known, the misclassification probabilities themselves are also hard to quantify. These create parameter ambiguity, and one needs to use robust analyses to hedge against them. However, current ED patient prioritization policies are based on analyses that ignore such ambiguities.

To demonstrate the benefits of our percentile optimization approach, we now focus on two questions: "how should EDs prioritize their patients given that they are faced with parameter ambiguity?" and "how much benefit can they get by taking ambiguities into consideration?" To answer these questions, we first model the ED from a broad perspective with non-stationary Poisson process arrivals and known service rates for all four classes: Urgent Simple (US), Urgent Complex (UC), Non-Urgent Simple (NS), and Non-Urgent Complex (NC) patients. In this way, we model the ED as a single "super-server" (i.e., with a pooled capacity that we estimate from our data set so as to match the input-output process of the ED as a whole). This allows us to gain insights into the questions we raised above by noting that the ED queueing model of Figure 7 (right) is essentially a special case of our general model depicted in Figure 1 with $n = 4$.

Patient arrivals in an ED fluctuate throughout a given day, so we model these arrivals with a non-stationary Poisson process with hourly rates shown in Figure 15 in Online Appendix A which depicts the actual time-dependent arrival rates to the ED based on our data set. Furthermore, since patient LOS in our data has a lognormal distribution, we fit lognormal service distributions to match the LOS of patients for each class of patients. Next, we design our "cloud of models" by perturbing the fitted rate parameters such that for each class i with fitted rate $\hat{\mu}_{i,3}$, we incorporate four additional possible rate parameters so $\hat{\mu}_{i,1} < \hat{\mu}_{i,2} < \hat{\mu}_{i,3} < \hat{\mu}_{i,4} < \hat{\mu}_{i,5}$. Because patients become fairly stable upon seeing a physician, we focus on adverse events in the waiting area of EDs, and assume ROAE drops to zero once the treatment stage begins. Our model is non-preemptive, which is a reflection of physicians' behavior in EDs: upon initiating treatment to a patient, they rarely

[18] Even after using a large data set that we have collected from a leading U.S. hospital, which includes data about more than 18,000 patient visits, we see that our point estimates are not reliable due to various reasons including the large variation among patient characteristics as well as the need to estimate parameters for each patient class separately.

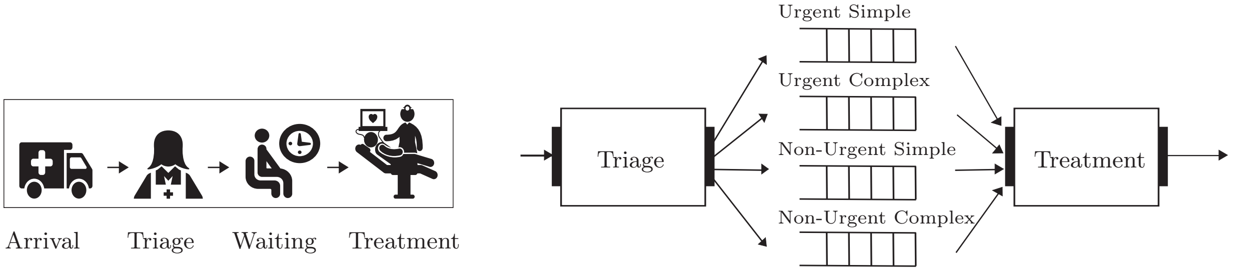

Figure 7 Patient flow in hospital Emergency Departments (left: the overall flow; right: the multi-class flow).

pause treatment to serve a different patient. Since there is a possibility that the ROAE for simple patients differs from that of complex patients, we also consider a variety of such "cost" structures in our study.

Though this model allows for dynamic arrivals (unlike our model introduced in Section 3), we can still incorporate chance-constrained policies through the use of our heuristic, and compare its performance to the complexity-based prioritization policy that serves classes US, UC, NS, and NC in descending priority (demonstrated to be optimal for EDs in Saghafian et al. (2014) when ambiguity is ignored), minimax, and minimin policies. To do so, we simply modify the Bayesian belief to also incorporate arrival data. We simulate these policies, and track the non-discounted ROAE by assuming that $\mathbb{P}_{\mathbf{B}}$ is uniform. The result of 20,000 simulated days expressed in terms of the CVar statistic is reported in Figure 8 (see Online Appendix A.6 for four additional ROAE settings and in depth discussions).

A widely discussed topic in the literature surrounding EDs is the "overcrowding" issue (see e.g. Derlet and Richards (2000), Derlet et al. (2001), and Trzeciak and Rivers (2003)) that stems from high arrival rates and limited resources (such as capacity, physicians, equipment, etc). Overcrowding in EDs results in high ROAE that endangers patients. The third row of Figure 8 demonstrates how policies perform in overcrowded EDs by considering an ambiguity set with smaller service rates (in comparison to the other ambiguity sets). We note that percentile optimization, in comparison with other policies, is especially suited for studying patient prioritization in overcrowded EDs. This is because under heavy congestion, chance-constrained policies learn faster, since more classes are available to serve at any given time. Furthermore, as we show in Corollary 4 in Online Appendix B, the E$c\mu$ policy becomes asymptotically optimal when arrivals occur during intense bursts followed by lull periods. Since hospital EDs typically experience long periods of heavy traffic in the afternoon followed by little traffic after midnight (see the actual arrival pattern depicted in Figure 15 in Online Appendix A), this further establishes our approach in hospital ED applications. Using these results, we can make the following:

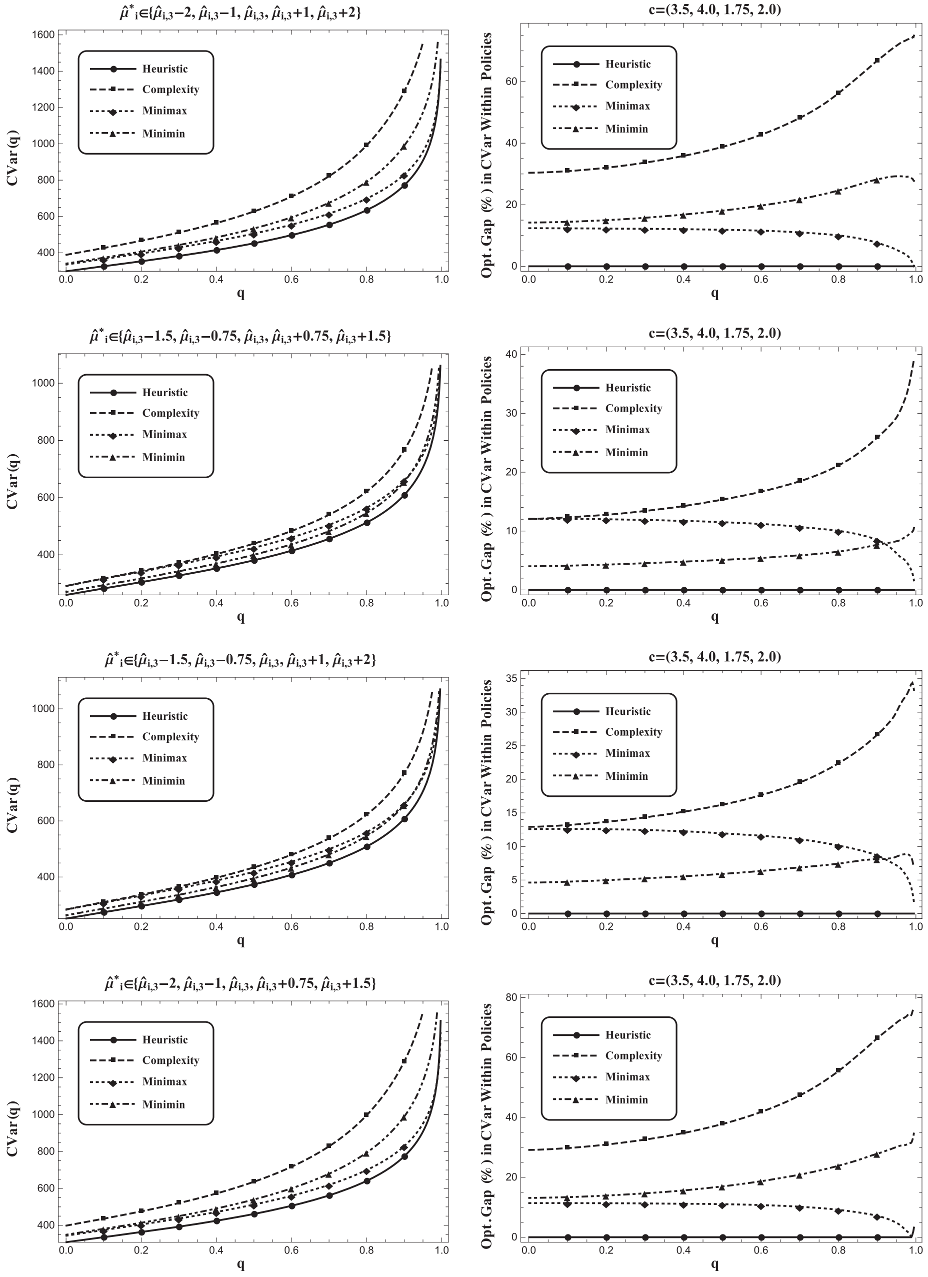

Figure 8 20,000 simulated days in the ED for the complexity-based prioritization, 95% E$c\mu$ heuristic, minimin, and minimax policies, when $\mathbb{P}_{\mathbf{B}}$ is uniform, and the cloud of models perturbs the fitted service rate $\hat{\mu}_{i,3}$ in terms of two-hour time increments with $\mathbf{c} = (3.5, 4.0, 1.75, 2.0)$. (Triage levels US, UC, NS, and NC are denoted 1,2,3, and 4, respectively.)

Observation 5 **(High Traffic)**. *Our percentile optimization approach performs well for prioritizing patients in EDs, especially in highly congested ones (e.g. those in busy research hospitals).*

Also, Figure 8 shows that, once again, the chance-constrained policies nearly dominate the entire spectrum of the CVar statistic since they explicitly incorporate both learning and robustness. The performance advantage over complexity-based prioritization is consistently over $10-15\%$ which suggests implementation regardless of optimism/pessimism levels. Hence, to establish the magnitude of potential benefits percentile optimization can offer to EDs over the current status quo, we make the following:

Observation 6 **(Improved System Performance)**. *Percentile optimization can improve the performance of EDs by* $10\% - 15\%$ *regardless of a manager's disposition.*

In systems with high traffic, learning may occur at an advanced rate, since it has available customers from each class a majority of the time the system is online. Hence, while static priority policies continue to serve the "wrong" classes (due to the underlying parameter ambiguity), the chance-constrained policy quickly identifies the optimal $c\mu$ priority using the observed values. This enhances the quality the robustness percentile optimization offers, especially since one is typically more concerned with overcrowded/busy systems (EDs with low traffic have short patient LOS naturally, and are not in significant need for optimization).

Furthermore, our "clearing" system is a model often used to study queues undergoing overcrowded situations. Therefore, a more congested ED is a better fit to our original model, and in considering dynamic arrivals, we can reconfirm all the previous insights generated in the "clearing" environment. This further confirms the results of Section A.4 (within Online Appendix A), where we show that most of the main insights gained from the "clearing" system holds for systems with dynamic arrivals.

In communities with unstable patient population characteristics, where ED service rates or misclassification probabilities are more ambiguous, ED managers may incorporate percentile optimization to effectively hedge against such ambiguities. Moreover, percentile optimization is well-suited to high levels of ambiguity. In our simulations, this is captured through modifying our cloud of models to incorporate larger differences in the fitted parameters (see the first row of Figure 8 and compare it with the second row). Hence, when patient population characteristics are unstable, percentile optimization stands out as a method that protects from negative consequences of focusing only on extreme outcomes, while simultaneously learning from incoming data. This results in the following:

Observation 7 **(Uncertain Population Characteristics)**. *Percentile optimization can significantly help EDs that are placed in geographical areas with unstable or unknown patient population characteristics to better prioritize their patients.*

8. Conclusion

Multi-class queues are versatile structures widely used in operations management that see a large variety of applications in both service and manufacturing sectors. In such environments, often exact parameter specification is rife with estimation errors that (if ignored) can cause system managers to implement wrong policies. We identify and implement a novel data-driven percentile optimization framework for use in POMDPs. Our method layers chance-constrained optimization on a non-robust learning model, effectively enabling learning of the true system state parameters, and allowing the manager to set an optimism level indicating the extent of protection against poor parameter scenarios s/he desires. We characterize the optimal policies to both the non-robust and percentile problems and find that chance-constrained policies can be established via the non-robust problem.

Since percentile optimization problems are typically computationally difficult, we introduce an analytically-rooted heuristic that can be used to effectively incorporate robustness in managing large and complex service or manufacturing systems. To further improve computational tractability, we find asymptotically tight bounds to the non-robust problem, which can be used to efficiently solve the percentile optimization problem.

Finally, we demonstrate the efficacy of our methods numerically in both stylized and realistic environments. Using real-world data collected from a leading hospital, we observe that our approach provides promising results in improving current patient flow policies, especially for overcrowded EDs, or those facing unknown patient population characteristics. Since ED managers typically do not fully know the service rate parameters, traditional patient flow policies based on queueing models that assume full service rate knowledge subject patients to higher risk than chance-constrained policies. Our work is the first to take into account the inevitable ambiguities in ED operations, and sheds light on the dire consequences of ignoring such ambiguities.

References

Argon, N., S. Ziya. 2009. Priority assignment under imperfect information on customer type identities. *Manufacturing & Service Operations Management* **11**(4) 674–693.

Bagnell, J., A. Y. Ng, J. Schneider. 2001. Solving uncertain Markov decision problems. *Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-01-25* .

Bandi, C., D. Bertsimas. 2012. Tractable stochastic analysis in high dimensions via robust optimization. *Mathematical Programming* **134**(1) 23–70.

Bandi, C., D. Bertsimas, N. Youssef. 2015. Robust queueing theory. *Operations Research* **63**(3) 676–700.

Bassamboo, A., A. Zeevi. 2009. On a data-driven method for staffing large call centers. *Operations Research* **57**(3) 714–726.

Bertsekas, D. 1995. *Dynamic Programming and Optimal Control*, vol. 1. Athena Scientific Belmont, MA.

Bertsimas, D., D. Gamarnik, A. A. Rikun. 2011. Performance analysis of queueing networks via robust optimization. *Operations Research* **59**(2) 455–466.

Bertsimas, D., D. Pachamanova, M. Sim. 2004. Robust linear optimization under general norms. *Operations Research Letters* **32**(6) 510–516.

Bertsimas, D., M. Sim. 2004. The price of robustness. *Operations Research* **52**(1) 35–53.

Buyukkoc, C., P. Varaiya, J. Walrand. 1985. The $c\mu$ rule revisited. *Advances in Applied Probability* **17** 237–238.

Calgar, U. 2010. Floating bodies. Master's thesis, Case Western Reserve University.

Charnes, A., W. W. Cooper. 1959. Chance-constrained programming. *Management Science* **6**(1) 73–79.

Chen, Y., V. Farias. 2013. Simple policies for dynamic pricing with imperfect forecasts. *Operations Research* **61**(3) 612–624.

Chow, Y., M. Ghavamzadeh, L. Janson, M. Pavone. 2017. Risk-constrained reinforcement learning with percentile risk criteria. *arXiv preprint arXiv:1512.01629* .

Delage, E., S. Mannor. 2007. Percentile optimization in uncertain Markov decision processes with application to efficient exploration. *Proceedings of the 24th international conference on Machine learning*. ACM, 225–232.

Delage, E., S. Mannor. 2010. Percentile optimization for Markov decision processes with parameter uncertainty. *Operations Research* **58** 203–213.

Derlet, R. W., J. R. Richards. 2000. Overcrowding in the nation's Emergency Departments: Complex causes and disturbing effects. *Annals of Emergency Medicine* **35**(1) 63–68.

Derlet, R. W., J. R. Richards, R. L. Kravitz. 2001. Frequent overcrowding in U.S. Emergency Departments. *Academic Emergency Medicine* **8**(2) 151–155.

Dupin, C. 1822. *Applications de Géométrie et de Méchanique*. Bachelier, successeur de Mme. Ve. Courcier, libraire.

Fresen, D. 2013. A multivariate Gnedenko law of large numbers. *The Annals of Probability* **41**(5) 3051–3080.

Guttmann, A., M. J. Schull, M. J. Vermeulen, T. A. Stukel. 2011. Association between waiting times and short term mortality and hospital admission after departure from Emergency Department: Population based cohort study from Ontario, Canada. *British Medical Journal* **342**.

Hansen, L., T. Sargent. 2007. Recursive robust estimation and control without commitment. *Journal of Economic Theory* **136**(1) 1–27.

Iyengar, G. N. 2005. Robust dynamic programming. *Mathematics of Operations Research* **30**(2) 257–280.

Jain, A., A. Lim, J. G. Shanthikumar. 2010. On the optimality of threshold control in queues with model uncertainty. *Queueing Systems* **65**(2) 157–174.

Lim, A. E. B., J. G. Shanthikumar, G. Vahn. 2012. Robust portfolio choice with learning in the framework of regret: Single-period case. *Management Science* **58**(9) 1732–1746.

Lippman, S. A. 1975. Applying a new device in the optimization of exponential queuing systems. *Operations Research* **23**(4) 687–710.

Littman, M. L., J. Goldsmith, M. Mundhenk. 1998. The computational complexity of probabilistic planning. *Journal of Artificial Intelligence Research* **9**(1) 1–36.

Mannor, S., D. Simester, P. Sun, J. N. Tsitsiklis. 2007. Bias and variance approximation in value function estimates. *Management Science* **53**(2) 308–322.

Mundhenk, M., J. Goldsmith, C. Lusena, E. Allender. 2000. Complexity of finite-horizon Markov decision process problems. *Journal of the ACM (JACM)* **47**(4) 681–720.

Nemirovski, A., A. Shapiro. 2006. Convex approximations of chance constrained programs. *SIAM Journal on Optimization* **17**(4) 969–996.

Nilim, A., L. El Ghaoui. 2005. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research* **53**(5) 780–798.

Osogami, T. 2015. Robust partially observable Markov decision process. *ICML*. 106–115.

Papadimitriou, C. H., J. N. Tsitsiklis. 1987. The complexity of Markov decision processes. *Mathematics of Operations Research* **12**(3) 441–450.

Pedarsani, R., J. Walrand, Y. Zhong. 2014. Robust scheduling and congestion control for flexible queueing networks. *2014 International Conference on Computing, Networking and Communications (ICNC)*. IEEE, 467–471.

Plunkett, P. K., D. G. Byrne, T. Breslin, K. Bennett, B. Silke. 2011. Increasing wait times predict increasing mortality for emergency medical admissions. *European Journal of Emergency Medicine* **18**(4) 192–196.

Prékopa, A. 1995. *Stochastic Programming*. Klewer Academic Publishers, Dordrecht.

Ross, S., J. Pineau, B. Chaib-draa, P. Kreitmann. 2011. A Bayesian approach for learning and planning in partially observable Markov decision processes. *Journal of Machine Learning Research* **12** 1729–1770.

Saghafian, S. 2017. Ambiguous POMDPs: Structural results and applications. Working Paper, Harvard University.

Saghafian, S., G. Austin, S. J. Traub. 2015. Operations research/management contributions to Emergency Department patient flow optimization: Review and research prospects. *IIE Transactions on Healthcare Systems Engineering* **5**(2).

Saghafian, S., W. J. Hopp, M. P. Van Oyen, J. S. Desmond, S. L. Kronick. 2012. Patient streaming as a mechanism to improve responsiveness in Emergency Departments. *Operations Research* **60**(5) 1080–1097.

Saghafian, S., W. J. Hopp, M. P. Van Oyen, J. S. Desmond, S. L. Kronick. 2014. Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing and Service Operations Management* **16**(3) 329–345.

Saghafian, S., M. H. Veatch. 2016. A $c\mu$ rule for two-tiered parallel servers. *IEEE Transactions on Automatic Control* **61**(4) 1046–1050.

Schutt, C., E. Werner. 1990. The convex floating body. *Mathematica Scandinavica* **66** 275–290.

Smallwood, R., E. J. Sondik. 1973. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research* **21**(5) 1071–1088.

Sondik, E. J. 1971. The optimal control of partially observable Markov processes. Ph.D. thesis, Stanford University.

Sprivulis, P. C., J. Da Silva, I. G. Jacobs, A. Frazer, G. A. Jelinek. 2006. The association between hospital overcrowding and mortality among patients admitted via Western Australian Emergency Departments. *Medical Journal of Australia* **184**(5) 208.

Su, H. 2006. Robust fluid control of multiclass queueing networks. Master's thesis, Massachusetts Institute of Technology.

Thrun, S. 1999. Monte Carlo POMDPs. *Advances in Neural Information Processing Systems* **12** 1064–1070.

Trzeciak, S., E. P. Rivers. 2003. Emergency Department overcrowding in the United States: an emerging threat to patient safety and public health. *Emergency Medicine Journal* **20**(5) 402–405.

Van Mieghem, J. A. 1995. Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *The Annals of Applied Probability* 809–833.

White, C., D. Harrington. 1980. Application of Jensen's inequality to adaptive suboptimal design. *Journal of Optimization Theory and Applications* **32**(1) 89–99.

Wiesemann, W., D. Kuhn, B. Rustem. 2013. Robust Markov decision processes. *Mathematics of Operations Research* **38**(1) 153–183.

Zhang, H. 2010. Partially observable Markov decision processes: A geometric technique and analysis. *Operations Research* **58**(1) 214–228.